Project Proposal

borui sun

1/24/2020

MACS Introduction to Machine Learning

Project Proposal: Predictive Modeling of Wildfire

Borui Sun, Bohan Yin, Shuai Yuan

Introduction Wildfire is a natural phenomenon, and it has many undesirable impacts on human safety, health and regional economic development. In October 2019, the news that Amazon rainforest fires have been burning over 3 three weeks rapidly occupied almost all of the press main pages overnight. The long-lasting fire is considered highly destructive and harmful for both human beings and the environment. In addition, due to elusive meteorological and weather scenarios, complicated topographical nature of the Earth, mixed soil mixture and fuel types, the difficulty in predicting wildfire will normally make loss greater once it breaks out. Indeed, the elusivity of wildfire has posed serious challenges for prevention, detection and monitoring. Therefore, early detection and prevention will be of utmost importance in reducing the loss triggered by wildfire. To achieve this, this research project aims to leverage multiple machine learning methods to develop a predictive modeling of wildfires by using data collected from satellite images and weather stations over the past five years. Above all, when selecting relevant parameters, cross validation or penalized regression will be applied. As for model building, relevant machine learning methods that build different models include GLM, random forest, KNN (K-nearest neighbor), SVM (support vector machines), etc. Also, this project will present the most accurate one after making a comparison across all model results. Given the significance of wildfire prevention, the authors believe that the success of this project can provide valuable insights into the early detection of wildfires and also relative mechanism for future wildfire prevention and management.

Literature Review Wildfires occur in different climatic zones and across different land use types. Mario et.al (2015) summarized different methodologies and indexes that can determine the likelihood of wildfire. This research identifies potential 28 factors corresponding with wildfire ignition and categorizes them into climatic (precipitation, temperature, wind, etc), topographic (slope, aspect, altitude), in-situ (fuel type, soil mixture, etc), historical and anthropogenic factors. Although it barely touches any machine learning method, this paper offers solid theoretical guideline for this project, points clear the direction in which the datasets can be found, as well as leads to the accurate identification of potential variables in causing a wildfire.

Castelli et al. (2015) discuss the possibility of introducing machine learning techniques to modeling and estimating burned areas in a wildfire. The major machine learning techniques used are SVM with a polynomial kernel, random forests, radial basis function network, linear regression, isotonic regression, and neural networks, which in combination helped to ensure a highly-reliable predictive method for predicting how large the burned area could be in a wildfire.

Garzón et al. (2006) once used machine learning to predict species habitats in forests. Though this research didn't provide any framework for analyzing wildfire problems, one predictive machine learning framework for habitat modeling is established through training, test and cross-validation. All the machine technique applications are insightful for predictive models: neural networks, random forests, and tree-based classification.

Rishickesh et al (2019) combined data mining and machine learning techniques to predict the eventuality of forest fires. Using the dataset from UCI machine learning repository, this paper uses physical factors and climatic conditions of the Montesinho park situated in Portugal to generate forest fire prediction. Relevant machine learning techniques include Logistic regression, Support Vector Machine, Random forest, K-Nearest neighbors in addition to Bagging and Boosting predictors, both with and without Principal Component Analysis (PCA). Among the models in which PCA was applied, Logistic Regression performs the best result and among the models where PCA was absent, Gradient boosting performs the best.

Sayad et al.(2019) built a predictive modeling of wildfires using several Artificial Intelligence techniques and strategies such as Big Data, Machine Learning, and Remote Sensing. They collected from satellite images over large areas and extract insights from them to predict the occurrence of wildfires and avoid such disasters. When applying big data strategies, they used supervised machine learning algorithms to label the data. All these researches have provided great insights in this projects, and therefore this project will adopt some relevant machine learning techniques used by previous researches but create new rastors with the latest data collected. The details will be discussed in the methodology and data section.

Data NASA's Fire Information for Resource Management System (FIRMS), which distributes near real time active fire data within 3 hours of satellite observation from as early as 2000 to present, serves as the primary source of collecting data on the target variable in this research project. A common challenge that machine learning researchers and data scientists often encountered is the lack of sufficient training and testing datasets when evaluate the accuracy of their model predictions. Because the NASA FRIMS active fire data is updated within every 3 hours, it offers an abundant amount of data that can be used to perform model evaluation and adjust accordingly.

Despite its exhaustiveness in fire location and date, it provides little information on the climate, topography, human activity and other important risk factors. Such information will be supplemented by borrowing data from other sources. To prevent loss of valuable information, this research project will start broadly in their data collection process to include as much as relevant factors as possible and then narrow it down by using penalized regressions to carefully select final parameters. Mhawej, Faour and Adjizian-Gerard (2015) identifies five categories risk factors of wildfire, which are climatic, topographic, in-situ, historical and anthropogenic factors: specifically, climatic factors consist of precipitation, temperature, wild speed, wind direction, etc; topographic factors include slope, altitude; in-situ factors refer to fuel type, fuel density, soil moisture, etc; historical factor means the probability of occurrence of a wildfire in the past; arthrographic factors include proximity to agriculture land, road, urban areas and exploitation zones.

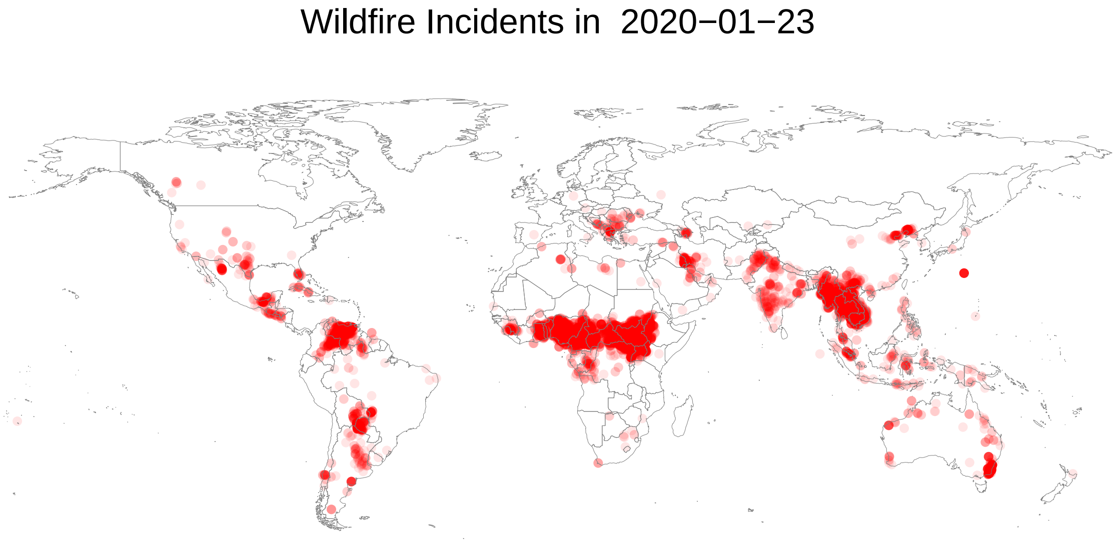
Research Methodology Forsell et al. (2009) argues that the major problems in dealing with an incredibly large wildfire point dataset are how to explore, analyze, and visualize in a proper and understandable manner. There are several important applications of machine learning for geospatial data worth mentioning: regional classification of environmental data, mapping of continuous environmental data, and optimization of monitoring networks.

Literature in the past also provides useful insights in deciding which machine learning algorithm to adopt. Logistic regression, Support Vector Machine, K-Nearest Neighbor, Random Forest are some of the most well-known, commonly used methods that have been proved efficient in modeling dependent variables with binomial distribution and in the domain of fire occurrence prediction or spatial distribution modeling (Saya, Mousannif and Moatassime, 2019;Rishickesh, Shahina and Khan, 2019; Vilar et al, 2016; Mi et al 2017), while one might be more appropriate than the other depending on the formation of final data set. These methods are also within the range of this course. Since NASA FIRMS only provides presence data, therefore, generating absence data through random sampling is required in order to properly apply these methods. Other methods may also be considered if allows. For example, Maximum Entropy Models is generally robust when using presence-only data but relies on heavy assumptions.

Regarding model selection and evaluation, another issue must be addressed is spatial autocorrelation. The presence of spatial autocorrelation-adjacent locations may exhibit more similar values than those further apart-has posed serious concerns for scholars in the past, as it violates the i.i.d assumption (independent and identically distributed) of standard statistical analysis and hence increases type I error rates (Dormann et al, 2007). Unfortunately, spatial autocorrelation is frequently ignored in the application of machine learning in geographic data. Therefore, in the process of model selection and evaluation, the authors aim to address this particular issue to their best efforts.

Preliminary Result

```
## Reading layer `MODIS_C6_Global_24h' from data source `C:\Users\sunsh\Desktop\ML\wildfire project\modi
## Simple feature collection with 14476 features and 13 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -175.39 ymin: -49.022 xmax: 177.183 ymax: 56.807
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
```



Wildfire Incidents in 2020−01−23

Bibliography *1. Castelli, M., et al. "Predicting Burned Areas of Forest Fires: An Artificial Intelligence Approach." Fire Ecol. 11, 2015, pp. 106–118. doi:10.4996/fireecology.1101106*

2. Garzón, M. B., et al. "Predicting Habitat Suitability with Machine Learning Models: the Potential Area of Pinus Sylvestris l." In the Iberian Peninsula. Ecol. Model. 197, 2006, pp. 383–393. doi:10.1016/j.ecolmodel.2006.03.015

3. Forsell, N., et al. "Reinforcement Learning for Spatial Processes." in 18th World IMACS/MODSIM Congress, Cairns, Australia, 2009, pp. 755–761.

4. Mhawej, Mario, et al. "Wildfire Likelihood's Elements: A Literature Review." MDPI, Multidisciplinary Digital Publishing Institute, 8 Dec. 2015, www.mdpi.com/2078-1547/6/2/282.

5. Rishickesh, R., et al. "Predicting Forest Fires Using Supervised and Ensemble Machine Learning Algorithms." International Journal of Recent Technology and Engineering 2, vol. 8, no. 2, 2019, pp. 3697–3705., doi:10.35940/ijrte.b2878.078219.

6. Sayad, Younes Oulad, et al. "Predictive Modeling of Wildfires: A New Dataset and Machine Learning Approach." Fire Safety Journal, vol. 104, 2019, pp. 130–146., doi:10.1016/j.firesaf.2019.01.006.

7. Dormann, Carsten F., et al. "Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: a Review." Wiley Online Library, John Wiley & Sons, Ltd, 27 Sept. 2007, onlinelibrary.wiley.com/doi/full/10.1111/j.2007.0906-7590.05171.x.

8. "Fire Information for Resource Management System (FIRMS)." NASA, NASA, 24 Jan. 2020, earthdata.nasa.gov/earth-observation-data/near-real-time/firms.