Final Draft for CMPUT 466 Mini-Project

Jinzhu Li
1461696

Introduction:

One major environment concern nowadays is the occurrence of the forest fires (wildfires), which is a threat to the forest preservation, economical and ecological environment, and human activities. Such phenomenon is caused by multiple reasons such as lighting and human negligence, and each year millions of forest hectares are destroyed worldwide. The accuracy of the estimated burned areas has great importance in the real life. For example, underestimating the size of a forest fire can be an extreme bad outcome since nearby homes may not be evacuated safely. Whereas overestimating forest fires may not cause extra damage to nearby homes, it may cost the fire station to invoke more firemen than needed, which is a waste of resource.

Therefore, this study is going to predict the burned area of forest fires using meteorological data. Three different machine learning algorithms, e.g. Neural Network (NN), Support Vector Machine (SVM), and Random Forest (RF) were tested on the real-world data collected from the northeast region of Portugal.

Finally, the result is going to answer the question that which algorithm works the best, that has the smallest error.

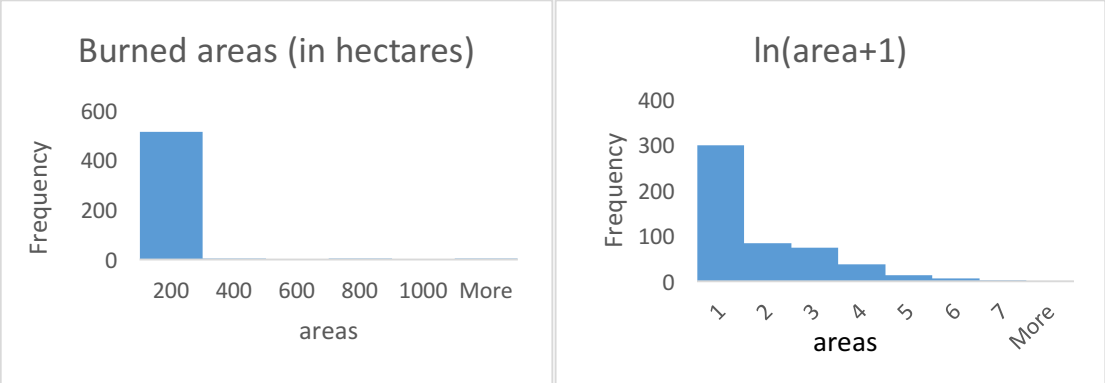
Description of the Dataset:

The data is the real-world data, collected from the Montesinho natural park, from the northeast region of Portugal, with the aim of predicting the burned area (or size) of the forest fires. The data used in the experiment were collected from January 2000 to December 2003.

There are total of 517 instances and 13 attributes: spatial location X, spatial location Y, Month of the year, Day of the week, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), temperature, relative humidity, wind speed, rain, total burned area. The target value is the total burned areas of the forest fires. All these variables are numerical, and Mouth and Day have been converted to numbers accordingly.

The burned area is shown in the figure below, denoting a positive skew, with the majority of fires presenting in a small size. X-axis is the burned areas; Y-axis is the frequency of that range of area occur. To reduce the skewness and improve the symmetry, the logarithm function $y = \ln(x+1)$, which is a common

transformation that tends to improve regression results for right-skewed targets was applied to area attribute. The final transformed variable will be the output of this work.



The histogram for burned area (left) and respective logarithm transform (right)

Learning Methods:

Three learning methods, each one has its own purposes and capabilities, have been chosen for this project: Neural Network (NN), Support Vector Machines (SVM), and Random Forest (RF).

- *Neural Network (NN):* This study will consider multilayer perceptron with one hidden layer of 100 hidden nodes and the rectified linear unit function (f(x)=max(0,x)) as the activation function and one output node with a linear function. L2 penalty (regularization term) was incorporated and the parameter is 0.0001. The solver for weight optimization is stochastic gradient descent, with cross-validation incorporated since the dataset is small. The proportion of the training data set aside as validation is 10%. Also, "early stopping" is used to terminate the training when validation score is not improving by at least the tolerance (set as 1e-4) for two consecutive epochs.
- *Support Vector Machines (SVM):* In SVM regression, the input is transformed into a high dimensional feature space, by using a nonlinear mapping. The SVM finds the best linear separating hyperplane in the feature space. SVM presents theoretical advantages than NN, such as the absence of local minima in the model optimization phase, that's why it was chosen. The popular Radical Basis Function Kernel was used, which presents less hyperparameters and numerical difficulties than other kernels, for example polynomial and sigmoid.
- *Random Forest (RF):* Random Forest is easy to interpret and this approach has been widely used. Random Forest is a method based on tree search. It is

a resemble of T unpruned Decision Tree (DT), using random feature selections from bootstrap training samples. The RF predictor is built by averaging the outputs of the T tree. This method was chosen because it solves the problem that multiple regression can only learn linear mappings, and it has a substantial improvement over a single DT.

Design of experiment:

All the experiment reported in this study were conducted using the scikit-learn, an open source library for Python that facilitates data mining and data analysis. Within the total of 517 entries, 300 were used as training set, and 217 were used for predictions. Cross-validation was implemented.

Before fitting the model, some preprocessing was required by NN, SVM and RF models. First, month and day were transformed into numerical attributes. Also, for NN and SVM methods, all attributes were standardized to a zero mean and one standard deviation. Next, the regression model was fitted. The default parameters (number of trees in the forest) were adopted for RF (e.g. T=500), the NN was adjusted using E=100 epochs until the tolerance less than 1e-4, and the Sequential Minimal Optimization was used for SVM. After fitting the models, the outputs were post-processed using the inverse of the algorithm transform. In few cases, this transformation may lead to negative numbers, and such negative outputs were set to zero.

To infer about the impact of the input variables, four distinct feature selection setups were tested for each machine learning algorithms: STFWI – using spatial, temporal, and the four FWI components; STM – with spatial, temporal and four weather variables; FWI – using the only four FWI components; and M – with the four weather conditions.

All the methods are using the squared loss as a cost function. Since this is a linear regression problem and I would like to measure how the predictions overshoot the target y. Therefore, trying to minimize the distance between the predictions and the target will be the most convenient way.

A paired t-test (95% confidence intervals, p-value=0.05) will be used for the statistical significance of the result.

Final Results:

Overall, among three algorithms, with all features included, SVM works the best, according to the calculated cost function (C=$(\mathrm{y\_hat} - \mathrm{y})^2$). The proposed solution includes only four weather variables (i.e. rain, wind, temperature and

humidity) in conjecture with SVM and it is capable of predicting the burned area of small fires, which constitutes the majority of the fire occurrence.

The results are shown in table in terms of mean and respective t-test 95% confidence interval. An interesting result is the non relevance of the spatial and temporal variables, since when removed the SVM performance improves. In effect, the best configuration is given by FMI setup and SVM model and paired t-test against all other models confirmed the statistical significance of the result. For the SVM, it is better to use FWI rather than weather conditions variables (the four FWI components are affected directly by the weather conditions). This is interesting outcome, since the metrological variables can be acquired directly from the weather sensor, with no need for accumulated calculations.

ML Models	Feature Selection Setup			
	STFWI	STM	FWI	M
NN	**17.58**±0.04 (7.0±0.0)	34.48±0.60 (14.4±8.5)	27.40±0.05 (7.05±0.1)	31.35±0.69 (8.5±3.4)
SVM	0.20±0.04 (0.09±0.0)	0.20±0.02 (0.1±0.0)	<u>0.11</u>±0.00 (0.0±0.0)	0.12±0.01 (0.0+0.0)
RF	**96.26**±0.02 (17.0±0.0)	178.78±0.01 (76.5±0.0)	277.09±0.05 (127.5±0.1)	118.13±0.01 (22.7±0.0)

Table. The predictive results in terms of mean absolute errors (Root Mean Squared error in the parentheses; <u>underline</u> – best model; **bold** – best within the feature selection)

Conclusion:

Forest fires cause a significant environmental damage while threatening human lives. In the last few decades, a substantial effort was made to build automatic detection tools that could assist Fire Management Systems to predict fire in advance. In this study, a Machine Learning approach that uses meteorological data, as detected by local sensors in the weather stations, and that is known to influence the forest fires was proposed. The advantage is that such data can be obtained in real-time with very low cost, when compared with other satellite and scanner approaches. Recent real-world data, from the the northeast region of Portugal was used in the experiment. This problem was modeled as a regression task, where the target was the prediction of the burned areas. Three different machine learning algorithms, including Neural Network, Support Vector Machines, and Random Forest were tested.

The proposed solution, which was based on SVM, is capable of predicting small fires, which constitute the majority of the fire occurrences. The drawback was the possibility of the lower predictive accuracy for large fires.