Prediction of Forest Fires using Data Mining Methods

By Hye Rin Kim

Supervised by Dr. McLeod

Master's Project

July 2009

Contents

1 Executive Summary

Forest fires are a major environmental issue, creating economical and ecological damage while dangering human lives.

In this research paper, we will have two major parts. The first part will reproduce the experiments/results conducted in the paper and the second part will take a different approach as a classfication problem to better predict the burned area of forest fires. We perform a Data Mining (DM) approach to predict the burned area of forest fires. In first part, four different DM techniques such as Naive, Multiple Regression, Neural Nets, and Support Vector Machine and four distinct feature selection setups (using spatial, temporal, FWI components and weather attributes) will be applied on recent real-world data collected from the northeast region of Portugal.

In second part, the classfication methods such as (Binary) logstic regression and multiple logistic regression will be applied on the same data as the first part. We will later compare two parts and see if the second part provides improvements on predicting the burned area of forest fires.

2 Introduction

2.1 Forest Fire Data

This problem will consider forest fire data from the Montesinho natural park, from the Trás-os-Montes Northeast region of Portugal. Satellite-based and infrared/smoke scanners have high costs. However, weather conditions, such as temperature and air humidity, are known to affect fire occurence, automatic meteorological satations are often available, and such data can be collected in real-time with low costs. We will present a Data Mining forest fire approach with emphasis on the use of real-time and non-costly meteorological data to predict the burned area of forest fires.

The data was collected from January 2000 to December 2003 with a total of 517 entries. In the dataset, there are 13 attributes that are the spatial and temporal attributes, four FWI components that are affected directly by the weather conditions, four meteorological attributes, and the response variable, the burned **area**. The data is consist of 12 input variables that are X, Y, month, day, FFMC, DMC, DC, ISI, temp, RH, wind, and rain and the response variable that is area.

The first four attributes are the spatial and temporal attributes. The first two attributes are the X and Y axis values where the fire occured within a 9*9 grid and the third and fourth attributes are the month and day of the week temporal variables. The next four FWI components are affected directly by the weather conditions. The forest Fire Weather Index(FWI) is the Canadian System for rating fire danger and it includes six components. Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and FWI. The first three are related to fuel codes: the FFMC denotes the moisture content surface litter and influences ignition and fire spread, while the DMC and DC represent the moisture content of shallow and deep organic layers, which affect fire intensity. The ISI is a score that correlates with fire velocity spread, while BUI represents the amount of

Attributes	Description
X	x-axis coordinate (from 1 to 9)
Y	y-axis coordinate (from 1 to 9)
month	Month of the year (January to December)
day	Day of the week (Monday to Sunday)
FFMC	FFMC code
DMC	DMC code
DC	DC code
ISI	ISI index
temp	Outside temperature (in Celsius)
RH	Outside relative humidity (in percentage)
wind	Outside wind speed (in kilometer per hour)
rain	Outside rain (in millimeter per square meter)
area	Total burned area (in ha)

available fuel. The FWI index is an indicator of fire intensity and it combines the two previous components. Different scales are used for each of the FWI elements, but high values suggest more severe burning conditions. The BUI and FWI were discarded since they are dependent of the previous values. The next four weather attributes are used by the FWI system and from the meteorological station database. In this case the values denote instant records, as given by the station sensors when the fire was detected. The rain variable represents the accumulated precipitation within the previous 30 minutes. The area variable represents the total burned area in hectares (ha). In the dataset, there are 247 samples with a zero value. All entries denote fire occurrences and zero value means that an area lower than $1ha/100 = 100m^2$ was burned. The burned area denoted a positive skew and we applied the logarithm transformation, y = ln(x + 1), to reduce the skewness and improve symmetry.

We'd like to examine the impact of the input variables and four distinct feature selection setups were tested for each DM algorithm. The four feature selection setups are as follows.

```
STFWI   using spatial, temporal and the four FWI components
STM     with the spatial, temporal and four weather variables
FWI     using only the four FWI components
M       with the four weather conditions
```

3 Analysis I

3.1 Multiple Regression

We are going to fit Multiple Regression model on forest fires data. Before fitting the model, some preprocessing was required. The nominal variables such as the month and day were transformed into indicator variables and all attributes except the indicator variables and the response variable were standardized to a zero mean and one standard deviation. Next, the regression models were fitted using lm function in R. The MR parameters were optimized using a least squares algorithm. For each feature selection setup, STFWI, STM, FWI, and

M, we applied Multiple Regression (MR), predicted the burned area of forest fires using all data, and calculated the overall performace by using the Mean Absolute Deviation (MAD) and Root Mean Squared (RMSE) below. Since the area variable is log transformed, the outputs were postprocessed using the inverse of the logarithm transform and then the MAD and RMSE are calculated.

$$MAD = 1/N * \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N}$$

The MAD and RMSE errors for the MR are obtained in Table 1.

	STFWI	STM	FWI	M
MR.MAD.Error	12.80	12.79	12.96	12.97
MR.RMSE.Error	64.35	64.34	64.48	64.47

Table 1: The MAD errors and RMSE errors for Multiple Regression

The errors are similar to the paper. In paper, the MR with the feature selection FWI provided the lowest MAD error, but we obtained that the MR with the feature selection STM provided the lowest MAD error. For RMSE, we have obtained very similar results with the paper and the MR with the feature selection STM provided the lowest RMSE for both cases.

3.2 Neural Networks

Next, we fitted Neural Netsworks on forest fires data. The same preprocessing was used as the previous method (MR) such as using indicator variables and standardization. The neural network is fitted using the nnet function in library nnet in R. We will fit the feed-forward neural networks in which inputs are connected to one or more nodes in the input layer, and these nodes are connected forward to further layers until they reach the output layer. The input nodes are used to represent the input attributes and an ouput node is used to represent the model output. The input nodes are connected forward to each and every node in the hidden layer, and these hidden nodes are conneted to the single node in the output layer. We will consider multilayer perceptrons with one hidden layer of H hidden nodes and logistic activation functions and one output node with a linear function. The jth node of the hidden layer of the feed-forwad network is

$$h_j = f_j(\alpha_{0j} + \sum_{i->j} w_{ij} x_i)$$

where x_i is the value of the ith input node, $f_j(.)$ is an activation function which is logistic function in here $f_j(z) = exp(z)/1 + exp(z)$. α_{0j} is called the bias, the summation i->j means summing over all input nodes feedting to j, and w_{ij} are the weights.

Hidden Nodes	STFWI	STM	FWI	M
NN	4	6	4	4

If the output activation function is linear, then the output of a feed-forward neural network (FFNN) can be written as

$$o = \alpha_{0o} + \sum_{j=1}^{k} w_{jo} h_j,$$

If the output activation function is linear, then the output of a skip-layer feed-forward neural network can be written as

$$o = \alpha_{0o} + \sum_{i=1}^{l} \alpha_{io} x_i + \sum_{j=1}^{k} w_{jo} h_j,$$

where the first summation is summing over the input nodes, l is the number of input nodes, k is the number of nodes in the hidden layer and h_j is given above. The second equation allows the direct connections from the input layer to the output layer which is referred as a skip-layer feed-forward network.

The NN performance will depend on the value of H. The best hidden nodes for four feature selection setups were found as above in the paper. Using the best hidden node, the Neural Net model is fitted for each feature selection setups with all training data in R. E = 100 epochs is used.

We first fitted a FFNN for each feature selection setups using all data. For the feature selection, STFWI, we obtained a 23-4-1 network for the data with 124 weights. For STM, we obtained a 23-6-1 network with 174 weights. For FWI, a 4-4-1 network is obtained with 29 weights. For M, a 4-4-1 network is obtained with 29 weights.

We obtained the estimates of their biases and weights using BFGS algorithm using nnet function in R. After a feed forward neural network is built, it is used to compute forecasts to predict the burned area of forest fires for each feature selection setups using all data. Then the MAD and RMSE are computed. The output is transformed using inverse of logarithm transform same as the previous method (MR) and the MAD and RMSE errors for Neural Network are obtained as follows. Since the NN cost function is nonconvex (with multiple minima), NR=3 runs were applied and the NN with the lowest penalized errors were selected for each feature selections in Table 2.

	STFWI	STM	FWI	M
NN.MAD.Error	11.73	11.14	12.81	12.69
NN.RMSE.Error	62.80	61.86	64.31	64.29

Table 2: The MAD errors and RMSE errors for FFNN

Next, a skip-layer FFNN is fitted for each feature selection setups using all data.

As you can see from the Table 3, a skip layer feed forward neural network showed a little improvements on the MAD and RMSE. The errors for a skip-

	STFWI	STM	FWI	M
NN.Skip.MAD.Error	11.51	11.03	12.82	12.65
NN.Skip.RMSE.Error	62.12	61.18	64.28	64.19

Table 3: The MAD errors and RMSE errors for a skip-layer FFNN

layer FFNN are a little smaller than the erorrs for the FFNN. Therefore, we choose a skip layer FFNN over the FFNN.

3.3 Support Vector Machine

Next, we fitted Support Vector Machine on forest fires data. The Support Vector Machine (SVM) is fitted using svm function in e1071 library in R. This library uses LIBSVM version 2.88. We used the same indicator variables for month and day variables and per default, data are scaled internally (both x and y variables) to zero mean and unit variance by using svm function in R. The scale changed between Neural Networks and Support Vector Machine because these methods are nonlinear. Given a training set of instance-label pairs$(x_i, y_i), i = 1, ..., l$, the SVM require the solution of the following optimization problem.

$$\min\ (1/2)w^T w + C \sum_{i=1}^{l} \xi_i$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0.$$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function ϕ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space:

$$\hat{y} = b + \sum_{i=1}^{m} w_i \phi_i(x)$$

where $\phi_i(x)$ represents a nonlinear transformation, according to the kernel function $K(x, x') = \sum_{i=1}^{m} \phi_i(x)\phi_i(x')$. The e-insensitive loss function is used. We used the popular Radial Basis Function kernel, which presents less hyperparameters and numerical difficulties than other kernels (e.g. polynomial or sigmoid), $K(x, x') = exp(-\gamma||x - x'||^2), \gamma > 0$.

The SVM is fitted using three parameters, C which is user-specified parameter representing the penalty of misclassifying the training instances, ε which is the width of the ε-insensitive zone, and γ which is the parameter of the kernel and using the Sequential Minimal Optimization algorithm. C=3 and $\varepsilon = 3\hat{\sigma}\sqrt{ln(N)/N}$ are used as heuristics proposed in [4], where $\hat{\sigma}$ is the standard deviation as predicted by 3-nearest neighbour algorithm. For fitting 3-NN, we standardized all the variables. The thirty runs of a 10-fold (in a total of 300 simulations) were applied to each feature setups in order to find the best $\gamma \in \{2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$ and the median values of the selected γ are as below as per the paper [1].

γ	STFWI	STM	FWI	M
SVM	2^5	2^3	2^3	2^3

We applied SVM for each feature selections using those parameters, predicted the burned area of forest fires using all data, and calculated the overall performace by using MAD and RMSE. The outputs were postprocessed using the inverse of the logarithm transform as before and then the MAD and RMSE are computed.

The MAD and RMSE for the SVM for each feature selections are shown below in Table 4.

	STFWI	STM	FWI	M
SVM.MAD.Error	12.29	10.96	12.76	12.61
SVM.RMSE.Error	64.48	63.38	64.71	64.64

Table 4: The MAD errors and RMSE errors for Support Vector Machine

3.4 Conclusion of Analysis I

The MAD and RMSE for all the methods are shown below in Table 5 and Table 6. The naive average predictor was also added in the first row of the table. The Naive predictor was computed by averaging the response variable area, and then MAD and RMSE were computed.

	STFWI	STM	FWI	M
NAIVE.MAD.Error	18.57	18.57	18.57	18.57
MR.MAD.Error	12.80	12.79	12.96	12.97
NN.MAD.Error	11.73	11.14	12.81	12.69
NN.Skip.MAD.Error	11.51	11.03	12.82	12.65
SVM.MAD.Error	12.29	10.96	12.76	12.61

Table 5: The MAD errors for all methods

The results we have found are similar with the result in the paper, but they are a little bit different. For MAD error, the paper mentioned that the SVM with the feature selection M provided the smallest error and it predicted the burned are of forest firest the best. However, we found that the SVM with the feature selection STM produced the smallest error. The errors for SVM and NN are a little bit different. The SVM with the feature selection M still can be used for prediction with little more error of 1.65 than the SVM with STM. However, the SVM with the feature selection STM will make a better prediction. Thus, the SVM method with the spatial, temporal and four weather variables produced the best predictions and it was found that the spatial, temporal and four weather variables are important for predicting the burned area of forest fires. For RMSE, the paper mentioned that the best option is the naive average predictor. However, in here, it was found that the skip-layer FFNN with the

	STFWI	STM	FWI	M
NAIVE.RMSE.Error	63.59	63.59	63.59	63.59
MR.RMSE.Error	64.35	64.34	64.48	64.47
NN.RMSE.Error	62.80	61.86	64.31	64.29
NN.Skip.RMSE.Error	62.12	61.18	64.28	64.19
SVM.RMSE.Error	64.48	63.38	64.71	64.64

Table 6: The RMSE for all methods

STM setup produced the best prediction. The NN and a skip layer NN with STFWI and STM and the SVM with the STM provided the better predictions than the naive average predictor. The SVM provided the biggiest error among all other methods except the SVM with the STM setup for the RMSE. It is due to the nature of each error criteria, i.e. the RMSE is more sensitive to outliers than the MAD metric.

The REC curves are provided for a more detailed analysis to the quality of the predictive errors. The SVM with the the feature selection STM, the skip-layer NN with STM, the MR with STM and Naive models are plotted in Figure 1. The SVM with STM is the best MAD configuration and the skip-layer NN is the best RMSE configuration. For SVM, 50.7% of the examples are accurately predicted if an error of $1ha$ is accepted and this value increases to 65.6% when the admissible error is $2ha$. Regarding the native predictor, it is the worst method, surpassing the SVM and MR methods only after an absolute error of 12.85. SVM and NN methods are similar but SVM is little more better. NN is surpassing SVM if the absolute errror of 0.67 is allowed and surpassing after absolute error of 9 in Figure 1. MR is below the NN curve. Therefore, the SVM method is the best if absolute error of between 0.68 and 8.9 are allowed. Otherwise, NN is better than SVM. The SVM and NN provide similar predictions. The SVM with the feature selection STM provides the best prediction.

Also, we have provided another plot to complement the REC curve for the SVM-STM configuration in Figure 2. From the Figure 2, we can see how the errors are distributed along the output range. The x-axis ranges from 1 to 517*30 runs and the y-axis was set within the range [0, $20ha$]. The black dots denote the real values and and the grey dots denote predictions within a relative error from 10% to 50% as shown in Figure 2. The real values and the prediction values are ordered according their burned area. We can see from the plot that the SVM method with STM is better at predicting small fires.

In the paper, it says that the solution is based in a SVM and requires only four direct weather inputs such as temperature, rain, relative humidity and wind speed and it is capable of predicting the burned area small fires, which are the majority of the fire occurences. There are 247 samples with a zero value in the dataset. The SVM performace is better when predicting small fires, but has lower predictive accuracy for large fires. In here, we have obtained that the SVM with the spatial, temporal, and four weather variables is the best solution and it is capable of predicting the burned area small fires. Therefore, we are going to fit this problem in a different approach as a classfication problems in the next section (Analysis II) for predicting the burned ara of forest fires and

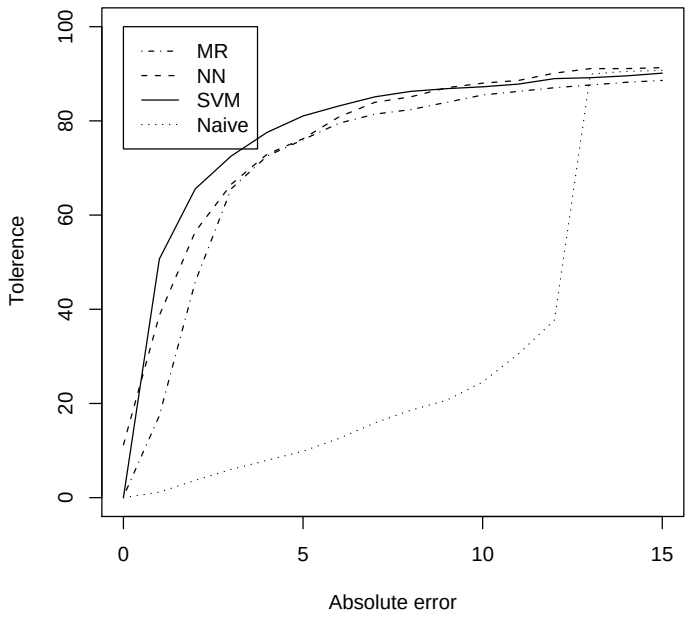

Figure 1: The REC curve for the skip layer NN-STM, SVM-STM, MR-STM and Naive Models

see if they provide better predictions on the burned area of forest fires. Also, we are going to see which variables are impacting importantly on predicting the burned area of forest fires.

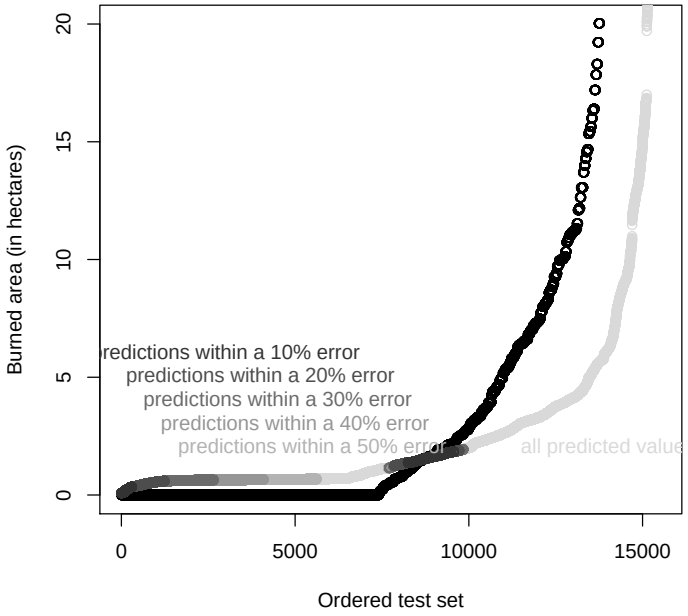

Figure 2: SVM-STM predictions (gray dots) along the y-axis output range

4 Analysis II

4.1 Logistic Regression

Logistic Regression using all variables.

	0	1
0	114	133
1	75	195

Table 7: Binary Logistic Regression using all variables

[1] 0.4023211
Using Step function

	0	1
0	104	143
1	77	193

Table 8: Binary Logistic Regression using step function

[1] 0.4255319

4.2 Multiple Logistic Regression

5 Conclusion

6 References

1. Paulo Cortez and Aníbal Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. http://www.dsi.uminho.pt/Ÿpcortez
2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. Support Vector Machine.
3. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A Practical Guide to Support Vector Classfication.
4. V. Cherkassy and Y. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. Neural Networks
5. Ruey S. Tsay. Analysis of Financial Time Series.
6. W. N. Venables, B. D. Ripley. Modern Applied Statistics with S.