

Predicting Forest Fire with Linear Regression and Random Forest

Ruide Guan*

Virginia Polytechnic Institute and State University, 240 W Campus DR, Harper Hall 0509,
Blacksburg, VA, 24061

*Corresponding author: gruide@vt.edu

Abstract. Forest fire is a global problem and it bothers a lot of countries. It not only threatens human life, but also causes serious damage to the environment and great economic losses. Therefore, forest fire is particularly important for forest fire prevention. and economic loss. However there many methods can be used to predict the forest fires then the forest fires could be stopped even before it happened. So, in the research will be talking about two methods in analyzing forest fire data set in order to predict the forest fires, they are linear regression and random forest. And before analyzing the data, the data will be pre-processed by PCA in order to get a more accurate result. Finally compare the result from linear regression and random forest to compare which method has a higher accuracy and better performance. And from the result and the comparison linear regression showed to be perform better in predicting forest fire.

Keywords: Forest fires, PCA, linear regression, random forest.

1. Introduction

Forest fire is a severe type of nature disaster. It has many serious consequences once it happened, it burns the trees, which is very important to earth's environment, causes the reduction of forest area, damage the forest environment, that is devastating challenge to the local ecosphere. It also pollutes the air, emits pm 2.5 particles that reduces air quality, it also causes the deaths of the animals in the forest and even human. According to the data form U.S. News [1], there are over 33,000 people around the world were killed by forest fire each year. The economy loss caused by forest fire is also gigantic. But what most important is, it is extremely hard to control or eliminate it once it happened. That's why the prediction of the forest fire becomes very important. After all, doing the prediction and preventing it from happening is a more efficient way, because all the loss caused by the forest fire cannot be brought back. So, in this research, researchers are doing analysis of the forest fires that already happened and use its data, to find out the most related factors that could cause the forest fire [2].

This research is using multiple ways to analysis the forest fire data. To find out the most related factor among positions of forest fire, the dates the forest fire happened, the temperature, the relative humidity, wind speed, rain and the forest fire burned area. Then get the R2 score of each method used in analyzing, which represents the accuracy of the result, to find out the most accurate way in analyzing the data of forest fire. Compare to many other researches only used one method in analyze forest fires, this research is using three different ways to analysis the forest fires and it will be able to find out the method that fit the situation the most, the people can get the result from the most fit method, so in that case that result can be the very accurate after doing many comparisons.

Now there are 3 most used and very grown-up forest fire level system which are CFFDRS used in Canada, NFFDRS used in United States and AFFDRS used in Australia [2] [3].

The first one to talk about is CFFDRS. This system has two subsystems which are FWI and FBP. FWI has 6 components which are Fine Fuel Moisture Code(FFMC), Duff Moisture Code(DMC), Drought Code(DC), Initial Spread Index(ISI), Build UP Index(BUI) and Forest Weather Index(FWI). In these components, FFMC, DMC and DC are the moisture rates of the burnable items, because of the different rate of the burnable to dry, when weather changes, the moisture rates will change significantly. DMC and DC refer to spreading rate and reduction rate while spreading of the fire. And

FWI is calculated by the daily measured moisture rate, relative moisture rate, wind speed and rain. FBP is a different system from FWI. FBP is mainly focus on the relationship of the fire spreading rate with the type of the burnable and the landform. FBP is formed by the spreading rate of fire, the amount of burnable, the magnitude of the fire and the type of the fire, including the distance of spreading, the area of spreading and the perimeter of the fire area as supporting. The system covers the burnable type, weather, topographic slope, slope direction, altitude, longitude, latitude, the dates of firing and lasting time as variable to do the prediction [4].

The US built another system called NFDRS base on its local situation. The NFDRS system is more complicated. This system assumes forest fire is depend on the status of the burnable, including the water containing rate, the amount of burnable and the physical temperature status. This system is able to calculate the water containing rate of the burnable under many complicate environment, including the calculation of the spreading rate of forest fire, the flame height, reaction strength, energy release and burning index, and the input data includes burnable model, humidity, temperature, weather condition and rain amount. Then after complex calculation, the index to be get are Ignition Components(IC), Spreading Components(SC), Burning Index(BI) and Energy Release Components(ERC). And NFDRS also differentiate the burnable into 21 exact types and have their own calculate index and model [5-6].

And next, AFFDRS is built base on multiple forest fire experiments. There were over 800 experiments made to build this system. The principle is to measure the water containment level in the burnable to predict the forest fire risk level. Because of the large area in relatively fewer people in Australia, the human related reason is smaller impact on forest fire. Forest fire depends more on exterior weather reasons. Base on that specialty, AFFDRS used all the experiments to get enough data sets and model [7]. And because Australia's local environment, most of the forest is formed by eucalyptus trees. Eucalyptus trees are easier to burn than other trees and harder to put off, the Australia government made strict standard for the experiments such as the burning area, the smoke release amount and the standard for data collection. The input index of AFFDRS includes the rain amount over a year, dates from last rain, wind speed, highest daily temperature, relative humidity. The result is mainly focus on the fire strength, spreading speed and the difficulty to put of the fire. So, this system is pretty limited and it suits eucalyptus forest better, not well enough when doing prediction on other burnable [8-9].

In this research, the methods used for the forest fire prediction are random forest and linear regression. Each method has their own advantage. For random forest, it's an integrated method so it has better accuracy than a normal method, so the result can be very accurate. And linear regression is a highly efficient method as for some small sized data and easy relationship, the model is easy to understand so it helps to make decisions based on the result.

2. Method

The method in this research is using PCA to do some pre-process for the data set, that would make the next steps become easier. After the pre-process, find out the most related characters that causes the forest fire, then use linear regression to process the data, to get the first result, and then use random forest as the second method, to get the second result. Then compare each method and their result, to see which method would be more accurate in predicting forest fires. The processes are shown as Figure 1.

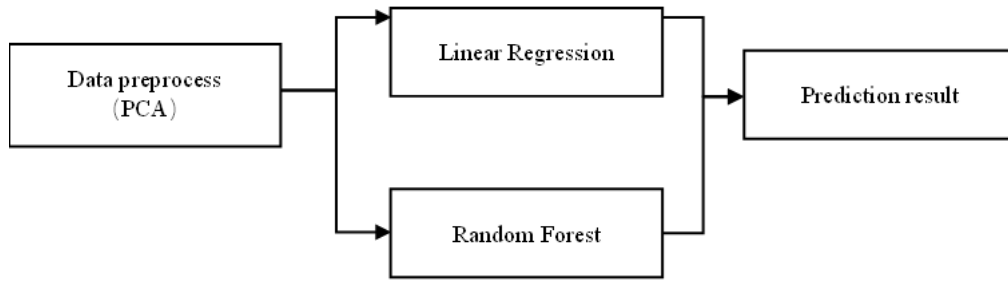


Figure 1. Data process procedure.

2.1. PCA Data Process

PCA (Principal Component Analysis) is a widely used data condense method. In this process, it drops off some irrelevant characters, and keep all the important characters that have large impact to the result. And that would help to do the analysis much easier. In this case, finding the character that have the most impact to the forest fire will become the job to do [10]. So among all the 8 characters in the data set (FFMC, DMC, DC, ISI, temp, RH, wind, rain), it will find the 4 characters that have the most impact on forest fire. So in this process, first find the explained variance ratio, that is the number stands for the relativity of the character to the result, the closer the result number to 0, more related the characters are. After getting all the explained variance ratio, compare them to each other and then the relativity, or the impact of the characters to the result can be get. The 4 characters has the largest impact are shown in the Table1.

Table 1 Characters has the largest Impact on forest fires

Variable order	Character code	Code meaning
0	PC0	Temp
1	PC1	RH
2	PC2	Wind
3	PC3	Rain

The 4 characters has the largest impact are temp, RH, wind and rain. And the data of these four characters will be used for the next step of linear regression.

2.2. Linear Regression

Linear regression is a method to analysis the relationship between one or more independent variable and dependent variable using regression formula. It's the linear combination of one or more returning model [11]. Linear regression can be used to make predictions by using data sets to make a prediction model, then if there is a new variable X and there is no existing Y to fit that X , that model can be used to predict a new Y . Or giving a variable Y and some variable $X_1, \dots, X_p$, all those X can have relationship to Y , and linear regression can be used to evaluate the strength of relationship between Y and any random X_r , and recognize which X has the unnecessary information about Y .

The basic principle of linear regression is to find a line that could best fit all the data points in a data set. But since all the data point cannot perfectly fit the line, so there must be error during the linear regression. In order to find out which line can represent all the data point the best, here's the formula of the loss function.

$$\sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2 \quad (1)$$

In this formula, $y^{(i)}$ represents the actual data point value and $\hat{y}^{(i)}$ represents the data point value after the linear regression. When the result of the loss function gets smaller, that means the line fit all the data point better.

And from the loss function, $\hat{y}^{(i)}$ can be written as:

$$\hat{y}^{(i)} = \theta_0 + \theta_1 X_1^{(i)} + \dots + \theta_n X_n^{(i)} \quad (2)$$

where θ is the final formula to be get, it can be shortened as:

$$\hat{y}^{(i)} = X_1^{(i)} \cdot \theta \quad (3)$$

after the evaluation:

$$\sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2 = (y - X_b \cdot \theta)^T (y - X_b \cdot \theta) \quad (4)$$

the final formula can be got as:

$$\theta = (X_B^T X_B)^{-1} X_B^T y \quad (5)$$

2.3. Random Forest

Random forest method also has a very wide application prospect. It can be used from marketing to medical insurance. It is an integrated method and has the base unit called decision tree [12]. This method integrates more than one decision tree by using ensemble learning. So basically, is to input a simple, and get a result from each decision tree and integrate all of the results from every one of the decision trees. Then find the final result. And because the simple is processed for many times from each decision trees, so the result will have excellent accuracy. And because it doesn't need variable deletion, it will be able to process high dimensional data sets and will be very efficient. In that case using random forest in forest fire prediction would a properly choice.

The basic rule of random forest is to set there are n forest fire data sets, and m weather causes. Then use bootstrap to draw the random data sets from n forest fire data sets, in order to get a n_t as the model collection of n , then to get the random forest tree of n_t , next draw m_{test} ($m_{test} \leq m$) weather causes, and choose the one with most classifications to do the subfield, let each tree grow as they can and do not do any cut off. Then all the n_t trees become the random forest and choose the mode of the n_t trees to be the result of the random forest. All the data sets not drawn from the n_t can be formed to be the out-of-bag(OOB) data sets, those can be used as the test sets of the random forest model.

In building up the random forest, n_t and m_{test} are the two most important custom parameters. There are experiments proved that $m_{test} = \sqrt{m}$ would be a better choice, and for n_t , as long as the n_t value is large enough, the random forest will become accurate.

And random forest can make judgement on the importance of each character variable. The main idea is for the variable X_f , first calculate all error rate for the out-of-bag(OOB) data as $errOOB$ for each tree t , then make randomly change the sequence of the variable X_f from the OOB. Keep all the other variable unchanged and re-calculate the OOB data after the random sequence change and error rate as OOB_t^f and $errOOB_t^f$. By analyzing the change of error rate can estimate the importance of each character variable. The importance of variable X_f can be written as

$$VI(X^f) = \frac{1}{n_t} \sum_t errOOB_t^f - errOOB_t \quad (6)$$

And in order to reduce the influence of the training sets to the test result, all the data sets will be divided into 60% as training sets to build the model, and the rest 40% will be used as test sets to test the model.

3. Result

This research will start with how the data processed, and compare the results to discuss and find out the best way to predict the forest fires.

3.1. Data sets

The data set used in this research of forest fire prediction is the Forest Fires Data Set form UCI Machine Learning Repository. This is a data set that aims for the prediction of the burned area of forest fires in the northeast region of Portugal. In this data set there are many characters about the forest fire from many perspectives. There are the X and Y coordinates within the Montesinho park map, the month of the year and the date of the week when the forest fire happened, the FFMCI index, DMC index, DC index and ISI index from the FWI system, the temperature, relative humidity, rain and burned area of the forest fire. Each character has their own measurement unit and method, but all the character records some perspective of the forest fire.

3.2. Data process

During the data process, first set a new character called area class. This new character area class is related to the character area which represents the burning area. When the character area is 0, area class is also equal to 0, that means there is no fire happening in that area. But when area is not 0, area class will become 1, which represents there is fire happening in that area. Find all the points of area to find all the observed points, then set all the points that with area class equal to 0 into green, and all the points with area class equal to 1 into red. Then all the fire area can be seen very clearly. Like the figure below. The X and Y axis are the coordinates of the Montesinho park map, and all the points represents all the locations been observed and the red points means there are forest fire happened and green points means there are not forest fire, as shown in Figure 2.

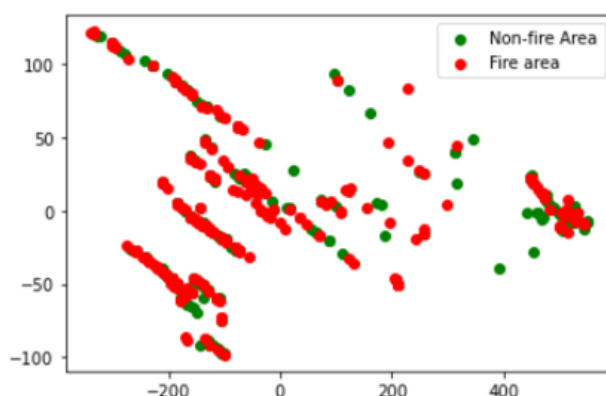


Figure 2. Fire and non-fire area

And in order to find the character has the most influence on forest fire, explained variance ratio can be used to check the relativity for each character. Here's the explained variance ratio of each character (Table 2).

Table 2 Character explained variance ratio

Var1	FFMC	0.35
Var2	DMC	0.16
Var3	DC	0.15
Var4	ISI	0.11
Var5	temp	0.08
Var6	RH	0.06
Var7	wind	0.03
Var8	rain	0.02

And put the explained variance ratio into a chart, both X and Y axis is the explained variance ratio, closer the point to the central point, the more related the characters to the results are (Figure 3).

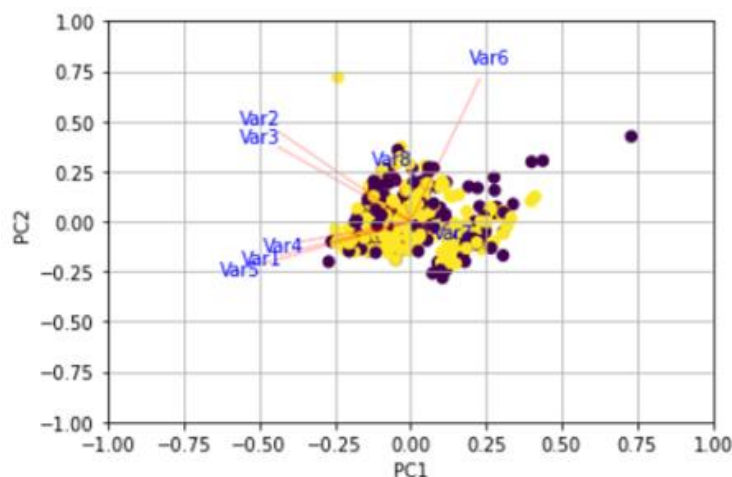


Figure 3. Character explained variance ratio spreading situation.

3.3. Performance and accuracy compare

In order to compare the performance of the two methods used in this research, there are several index can be used for the checking and the compare (Table 3). The first one is Mean Absolute Error(MAE), it's the mean value of the difference between every single observed value and the arithmetic mean. It can be used to accurately reflect the error of the prediction. Larger the MAE value is, the bigger the errors are. The second one is Mean Squared Error(MSE). It's a measurement of the difference between the estimator and the estimated. MSE is less stable than MAE, but in some situations, MSE can get a more accurate result than MAE. And the third one is R2 score. R2 score is the ratio of the square of the regression and the total square. Normally larger the R2 score is, means the independent variable has a more explain rate to the dependent variable. And smaller the R2 score is, means the method better fits model. And the result is shown in the Table 3.

Table 3 The results of different methods compared.

Method	MAE	MSE	R2
Linear Regression	31.69	17008.71	-1.60
Random Forest	35.13	14844.89	-9.69

The MAE value, MSE value and R2 score can be seen in this table. Compared to linear regression, random forest has a relatively smaller MSE value, but the MAE value is much larger and also the R2 score is further to 0. Which means in comparison of MSE, random forest has a better performance than linear regression, but linear regression has a better performance in the comparison of MAE and R2.

3.4. Discussion

In general, linear regression method has higher accuracy and better performance than random forest method in predicting forest fire. Because in MAE value and MSE value, both linear regression and random forest has their own advantage and disadvantage. Linear regression has a better performance in MAE and random forest has a better performance in MSE, and the difference of the two method in each comparison is not very obvious. But talking about R2 score, linear regression has a very large lead than random forest because the result of R2 score is a lot closer to 0. That means the results of forest fires prediction get by linear regression will be more accurate.

4. Conclusion

In order to predict forest fires, first this research used PCA to pre-process the data, removed irrelevant data and made the following analysis easier. Then two method linear regression and random forest are used to do the analysis and the prediction of the data sets after the PCA pre-process. And

during the analysis to the data sets, the relationship of each character of forest fire causes to the happening of the forest fire were found, temperature, relative humidity, wind and rain were the four characters has the largest impact on forest fires. And after getting the result of predicting forest fires by using linear regression and random forest, the two method were compared to each other using MAE and MSE index as long as R2 score. After the comparison, linear regression has a small lead in MAE and a large lead in R2 score. So linear regression performs better and can get a more accurate result in predicting forest fires. This method is more appropriate in predicting forest fires.

References

- [1] Reinberg, S. "Wildfires Cause More Than 33,000 Deaths Globally Each Year". US.News, (2021).
- [2] Jia N., Chen Y., Kang K., Li J., "Improved Forest fire risk assessment model and its application based on the RF algorithm", Journal of Safety and Environment 220(4) (2020).
- [3] Wotton B M., "Interpreting and using outputs from the Canadian forest fire danger rating system in research applications" Environmental & Ecological Statistics, 16(2): 107-131(2009)
- [4] Bai H., "Forest fires prevention and control capabilities in Sanming region, Fujian province" Beijing Forestry University (2012).
- [5] Deng P., Yu P., Wu Q., "Application of random forest algorithm on the forest fire prediction based on meteorological factors in the hilly area, central Hunan Province" Journal of Northwest Forestry University (2018).
- [6] Breiham L., "Random forests machine learning" Scientific Reports, 24(2):123-140(2021).
- [7] Cortez P. and Morais A., "A data mining approach to predict forest fires using meteorological data" Proceedings of the 13th Portuguese Conference on Artificial Intelligence 1121-1128, (2007).
- [8] Guo F., Hu H., Ma Z., "Applicability of different models in simulating the relationships between forest fire occurrence and weather factors in Daxing'an Mountains" Chinese Journal of Applied Ecology 21(1): 159-164 (2010).
- [9] Li X., "Using random forest for classification and regression" Chinese Journal of Applied Entomology (4): 159-164 (2013).
- [10] Mao G. "The relationship between forest fire and weather conditions and forecast" Atmosphere 14(9): 52-54 (1988).
- [11] Yao D., Yang J., Zhan X., "Feature selection algorithm based on random forest" Journal of Jilin University 52, 41-49, (2014).
- [12] Chuvieco E., Giglio L., Justice C., "Global characterization of fire activity: towards defining fire regimes from earth observation data" Global Change Biology 14(7): 1488-1502(2018).