Data Analysis for Prediction of Forest Fires

Mervenur Gulel(✉) and Ali Serdar Tasan

Department of Industrial Engineering, Dokuz Eylül University,
35397 Izmir, Turkey
gulelmervenur@gmail.com, serdar.tasan@deu.edu.tr

Abstract. The earth offers various resources for humanity but these resources are limited. Somethings have to be done in order to save this resources and also nature within sustainability perspective. Nature is essential for life by offering water resources, clean air and various food resources, etc. Forests have a great impact on these essential resources besides aesthetics. However, many forests were lost due to unexpected and uncontrolled forest fires especially at recent years. The main idea of this study is, an effective forest fire prediction system can help us to save forests. In this study, several basic methodologies are studied to predict the forest fires with a dataset containing meteorological data, from the literature. Some preventive actions can be taken with a successful prediction information. Also this prediction facilitates the planning of resources to cope with forest fires. For prediction study, neural networks, linear regression, decision tree and random forest methodologies are used within Python and MATLAB environment. The dataset from the literature which contains meteorological data (temperature, rain, humidity, wind) and also size of burned area data, is used in this study. Results are compared and it is realized that neural networks performs better than the other in means of accuracy value, for forest fire prediction with meteorological data.

Keywords: Data analysis · Forest fires · Prediction

1 Introduction

Forests, which are part of nature and can be described as the antidote of global warming, greatly affect ecosystems and therefore all assets. One of the major threats to forests is the forest fires. The control of the fires' results and negative effects is very important. Therefore, the early estimation and determination of the forest fires and appropriate planning activities must be done for the protection of the forests. The spread and effects of forests fires depends on appropriate meteorological conditions such as air temperature, relative humidity, and wind. These meteorological data and some indexes such as DMC (calculates with temperature, rain, relative humidity data and some equations [1]) which is used in some studies to ensures the accuracy of the prediction.

If planning and prevention work can be done correctly with an accurate estimation, use of available but limited resources could be more efficient and effective. And eventually, this reduces the forest fires and their negative impacts.

Z. Anisic et al. (Eds.): IJCIEOM 2019, LNMUINEN, pp. 86–93, 2020.
https://doi.org/10.1007/978-3-030-43616-2_10

2 Prediction of Forest Fires

Several methods were used in the literature for prediction of forest fires. According to [2], some sensors are used for the detection of forest fires and it is known that meteorological values have an effect on forest fires and some fire indices. In the study, Data Mining (DM) method was investigated for estimation of burned areas. The actual data for a region in Portugal was tested with a choice of five different DM techniques and four different features. This study is important in terms of ensuring the most effective, correct use of the available resources and making improvements [2].

Sakr et al. [3] used artificial intelligence for prediction and presented a new forest fire risk prediction algorithm based on support vector machines. The algorithm depends on the previous weather conditions to estimate the fire hazard level of a day. Implementing the algorithm using data from Lebanon has demonstrated the ability to accurately predict the danger of a fire [3].

Rajasekaran et al. [4] collected and analysed data by using Hadoop tool to predict forest fire before it occurs. There is a machine learning tool called Mahout, which is used to aggregate and filter data sets and can estimate the current output. People can be alerted via GSM when a fire occurs. Signal and infrared image processing is used to track signals and images throughout the forest every 30 min, and this data are stored in datasets. Thus, forest fire can be predicted by using these data [4].

Lin et al. [5] was found that observing local weather and human behaviours was the most important factors related to forest fire, since the high incidence and destruction of forest fire determined the importance of forest fire prediction or early detection. Therefore, a fuzzy inference and big data analysis algorithm were proposed to evaluate fire risk and quantitative potential fire risk was calculated. The rechargeable wireless sensor network collects the 24-h continuous weather information. The risk of high potential forest fires can be measured with this algorithm to prevent forest fire situation [5].

In this study, we decided to use linear regression, decision tree, random forest and neural network models for prediction of forest fires and finally compare the results of all these methodologies. Random forest is a machine learning algorithm that is flexible and easy to use, it is used for classification and also regression. Random forest is basically a controlled learning algorithm. According to the working principle, it actually creates a group of decision "forest" structure by bagging method. It creates multiple decision trees and then combines these decision trees for accurate estimation. The algorithm also adds additional randomness to the model when growing trees [6].

Decision Tree is basically a tree-based learning algorithm and frequently used. It can be integrated into the solution of all problems. For a large data set, it is a method used to divide it into smaller groups by applying certain rules of decision. In other words, it applies some procedures when making decisions, and thus uses the data that has a large number of records in smaller groups. This structure, which is easy to understand, has the advantages of being able to operate on different data types and being easy to interpret [7].

Regression is basically one of the methods used statistically. The relationships between variables in a dataset can be analysed with this method. Therefore, it is generally preferred by many disciplines. Linear regression, which is the first step for

the analysis which can be called as complex, is the most basic technique used to model the relationship between different variables. This method is divided into two sections as simple regression and multiple regression. The main difference between them is the number of independent variables. In simple regression this number is one, but in multiple regression it can be more than one [8]. For multiple regression the general equation is (1). So, in this study it was decided to use multiple linear regression in Python.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u \tag{1}$$

For neural networks, the structure and working methods of the brain and nerve cells should be taken into consideration. Because it seems to be similar in principle. There are many different kinds of artificial neural networks. It can be divided into different classes as single-layer, multi-layer or feedforward, feedback. There are different advantages of this method. Unlike other methods, artificial neural networks, maintain operation even if there are incomplete data (like missing data or information) [9].

In this paper; linear regression, decision tree, random forest and neural network models were used to estimate the forest fire risk level based on four main meteorological data: temperature, humidity, wind and rain. The first three methodologies were coded in Python environment, and neural network model was developed in MATLAB. The dataset of Montesinho Natural Park [2] from literature was used in this study. This dataset contains temperature (in °C), relative humidity (in %), wind (in km/h), rain (in mm/m^2), burned area (in ha) data with some index and other data such as month, day.

Three different scenarios were defined for prediction of the burned area. After each scenario, RMSE (Root Mean Squared Error) (2), MAE (Mean Absolute Error) (3) and MAPE (Mean Absolute Percentage Error) (4) are calculated in order to evaluate the performance of the prediction model and scenario and choose the best way. Before the prediction, we normalize all dataset with D Min Max normalization with the equation of (5). Because D Min Max normalization method gives realistic results [10]. (For Eqs. (2)–(5), n: number of data, y: real data, y′: prediction, e = y′−y).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{3}$$

$$MAPE = \frac{\%100}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \tag{4}$$

$$x' = 0.8\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} + 0.1 \tag{5}$$

In addition, another criterion as the percentage of accuracy was used and calculated as follows: if the absolute value of difference between denormalized test and train data is less than 1, then it accepted as true.

For each method a number of experiments were done to decide the parameter values. After experimental study, parameter values were decided as follows: for random forest, n estimator is 500 (the number of trees in the forest is 500), the random state value is 100. Also for all methods, 80% of whole dataset were used for training, 20% for testing.

2.1 Scenario – 1

At Scenario – 1, four meteorological data - temperature, rain, relative humidity and wind - were used for predict the area's value. Before the prediction to reduce skewness and improve symmetry, the logarithm function (6) was applied to the area attribute, then the dataset was normalized with D Min Max normalization.

$$y = \ln(x+1) \tag{6}$$

The neural network model was developed in MATLAB environment. Different structures for prediction were studied, it was decided to construct the neural network with feedforward backpropagation, multilayer perceptron and has three hidden layers (each layer has ten neurons) with four input (temperature, relative humidity, wind and rain) and one output (area – prediction value) for good prediction. The constructed network was trained and then retrained for good prediction. Results for Scenario – 1, error values and the percentage of accuracy are given in Table 1.

Table 1. The results for Scenario – 1.

	Random forest	Decision tree	Linear regression	Neural networks
RMSE	1.2973	1.9022	1.4512	1.3056
MAE	1.0229	1.45631	1.1082	1.0389
MAPE (%)	67.1546	90.8090	66.8717	60.1874
Accuracy (%)	57	50	51	52

When all the solution methods applied for Scenario – 1 are examined, it is seen that neural networks gives the low MAPE value among the all methods. As a result, it can be said that, neural networks is the best method for estimating with meteorological data according to Scenario – 1.

2.2 Scenario – 2

At the second scenario, it was thought that, the dataset has different attributes and if these attributes can be combined and some of them is used in prediction models for input, less error value could be observed. However, the selection of the best attributes

for prediction is important. Therefore, the feature selection in Python was used for choose the best attributes for prediction methodologies. Feature selection is the process of selecting such features that do not reduce accuracy and create subsets of features. Thus, it is possible to move away from the data which has low level of interest. This allows access to a data set free of unnecessary features and models with higher accuracy [11].

The result of feature selection is duff moisture code (DMC). DMC is calculated with many equations and it contains rain, temperature and relative humidity attributes. So we decided to use DMC with wind for predict the area. The logarithm function (6) was applied to the area attribute, then the all dataset was normalized with D Min Max normalization. The structure of neural network was same as Scenario – 1's but only one difference: with two inputs that is DMC and wind. Results for Scenario – 2, error values and the percentage of accuracy are given in Table 2.

Table 2. The results for Scenario – 2.

	Random forest	Decision tree	Linear regression	Neural networks
RMSE	1.3117	1.5650	1.2559	1.3009
MAE	1.039421417	1.2333	1.0093	1.0065
MAPE (%)	57.1830	64.2007	62.0969	56.3935
Accuracy (%)	57	56	54	57

Neural networks model gave better results than the others as in Scenario – 1. In addition, there is a difference for MAPE value. It means that if inputs predicted with more related attributes were selected, it would give better results than usage of the meteorological data directly and generally.

2.3 Scenario – 3

Four meteorological data - temperature, rain, relative humidity and wind were considered again, in addition area data converted into classes in the last scenario. First the area data were converted from hectares to acres and then classified as Class 1 is one acre or less, Class 2 is more than one acre but less than 10 acres, Class 3 is 10 acres or more but less than 100 acres, Class 4 is 100 acres or more but less than 300 acres, Class 5 is 300 acres or more but less than 1000 acres and Class 6 is 1000 acres or more but less than 5000 acres. Then the dataset was normalized with D Min Max normalization. The structure of Neural Network was same as Scenario – 1's. Results for Scenario – 3, error values and the percentage of accuracy are given in Table 3.

Table 3. The results for Scenario – 3.

	Random forest	Decision tree	Linear regression	Neural networks
RMSE	0.8819	1.4466	0.7201	0.7071
MAE	0.5926	1.0556	0.4444	0.4259
MAPE (%)	25.1852	46.3580	21.5741	20.6482
Accuracy (%)	87	28	90	91

Neural networks model gave better results than the others as Scenario – 1 & 2. In addition, there is a big difference between all scenarios about error values. It means that if the area data were classified for prediction, better error values and percentage of accuracy would be observed.

3 Prediction with a New Dataset

Another dataset was obtained and studied with previously defined Scenario – 3 in order to evaluate the performance of proposed prediction models. New dataset for Washington area was prepared by using data from DNR Fire Statistics 2008 – Present [12] and Weather Underground [13]. Again same attributes (temperature, wind, rain, humidity and burned area) for year 2011 to 2019 were used.

First the area data were classified as Class 1 is one acre or less, Class 2 is more than one acre but less than 10 acres, Class 3 is 10 acres or more but less than 100 acres, Class 4 is 100 acres or more but less than 300 acres, Class 5 is 300 acres or more but less than 1000 acres and Class 6 is 1000 acres or more but less than 5000 acres. Then the dataset was normalized with D Min Max normalization. The structure of Neural Network was determined with feedforward backpropagation and multilayer perceptron. Error values and the percentage of accuracy are given in Table 4.

Table 4. The results for the new dataset.

	Random forest	Decision tree	Linear regression	Neural networks
RMSE	0.7698	0.9027	0.6666	0.70719
MAE	0.4444	0.5926	0.2963	0.3519
MAPE (%)	28.09	42.9012	13.272	16.975
Accuracy (%)	89	63	92	93

According to the results, neural network model outperforms among others. Also, there is a big difference between all scenarios' error values. It means that if the method of Scenario – 3 (classified the area data for predict) was used, we could get better error values and percentage of accuracy, usage of this method with a different dataset validate the methods usefulness and effectiveness.

4 Conclusion

In the study, it was realized that different methods, techniques and scenarios can be used in the prediction process and they produce different results and accuracy levels. The aim of this study is to find both the most effective and efficient technique among the different techniques for prediction of the forest fires. Thus, both effective and efficient resource utilization can be ensured and both the costs and losses for the planning and preventive activities due to inaccurate prediction process can be minimized.

The motivation of this study on the forest fires, is the importance of this issue to the whole world, because it is an environmental issue that has continuously increases its importance from the past to the present, and the opportunity to make improvements in this field.

In this study, historical forest fire data and meteorological data were used to make estimations for future fires and this provide an input for effective preventive actions. It can be possible to make an effective planning and prevent possible forest fires with the improved accuracy of this estimation procedure.

Linear regression, decision tree, random forest and neural network models were utilized in this study to estimate the forest fire risk level based on four main meteorological data: temperature, humidity, wind and rain. These methods were applied to two dataset. As we can see from the results, neural networks performed well in all scenarios. The Scenario – 3 (class the area data) is seen more appropriate to predict forest fires as the risk class can be predicted before the fire comes out. Therefore, the methodology of Scenario – 3 was used for second dataset and it gave realistic estimates. By the way, neural networks outperformed among the other methods in all of the scenarios.

References

1. Van Wagner, C.E., Pickett, T.L.: Equations and FORTRAN Program for the Canadian Forest Fire Weather Index System. Canadian Forest Service, Ottowa (1985)
2. Cortez, P., Morais, A.A.: Data mining approach to predict forest fires using meteorological data. In: Neves, J., Santos, M.F., Machado, J. (eds.) New Trends in Artificial Intelligence: EPIA 2007, pp. 512–523. APPIA, Portugal (2007)
3. Sakr, G., Elhajj, I., Mitri, G., Wejinya, U.: Artificial intelligence for forest fire prediction. In: Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM 2010, pp. 1311–1316. IEEE Press (2010)
4. Rajasekaran, T., Sruthi, J., Revathi, S., Raveena, N.: Forest fire prediction and alert system using big data technology. In: Proceedings of the International Conference on Information Engineering, Management and Security, ICIEMS 2015, pp. 23–26. ICIEMS, India (2015)
5. Lin, H., Liu, X., Wang, X., Liu, Y.A.: Fuzzy inference and big data analysis algorithm for the prediction of forest fire based on rechargeable wireless sensor networks. Sustain. Comput.-Inform. **18**(1), 101–111 (2018)
6. Donges, N.: The random forest algorithm. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. Accessed 05 May 2019

7. Makine Öğrenimi, Bölüm-5. https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-5-karar-ağaçları-c90bd7593010. Accessed 24 Nov 2019
8. Makine Öğrenimi, Bölüm-6. https://medium.com/kodcular/makine-öğrenimi-bölüm-6-regresyon-3d837236eb6b. Accessed 24 Nov 2019
9. Makine Öğrenimi, Bölüm-3. https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-3-4b160df1f4c8. Accessed 24 Nov 2019
10. Yavuz, S., Deveci, M.: İstatistiksel normalizasyon tekniklerinin yapay sinir ağın performansına etkisi. ERU IIBFD **40**(1), 167–187 (2012)
11. Feature Selection. https://gurcanyavuz.wordpress.com/2014/12/16/feature-selection/. Accessed 25 Nov 2019
12. DNR Fire Statistics 2008 – Present. http://geo.wa.gov/datasets/dabefcb8f03549b49bee7564d4c3c4b5_8. Accessed 25 Nov 2019
13. Seattle, WA Weather History. https://www.wunderground.com/history/monthly/us/wa/seattle/KSEA/date/2017-5. Accessed 25 Nov 2019