Artificial neural network approach to predict forest fire with data encoding technology

Yiwen Wang

Research School of Computer Science,
Australian National University, Canberra ACT 0200, Australia

Abstract. In the area of machine learning, input data processing is not a easy thing of real world, especially when the data structure is unknown. Choosing the proper data encoding methods is a must in most of the cases in the beginning of learning process.

This article predicts the area of forest fire in Montesinho national park based on a reliable data set[1] with artificial neural network. In this paper, we mainly focused on finding a proper method to encode input data to balance the information loss and simplification. The result was worse than a published research paper for the same data set.

Keywords: Forest fire, back propagation, neural network, input coding technology, data analysis, cross validation

1 Introduction

Forest fire is one of the worst environment disasters in the world which causes fatal damage to local environment and endanger the property and life of nearby residents. Once it starts without any notice, it would spread to a very large region and take a lot of time and effect to control. Hence, it is reasonable to find a way to do the prediction before it starts. Artificial neural network is one approach to proceed such prediction.

The raw data is from Forest Fire data set from UCI Machine Learning Repository which created by: Paulo Cortez and Anbal Morais[1]. The reason for me to choose this data set is that it is a very complex and challenging data set, since there is no clear outliner and the number of large area fire burn samples is very small. Another reason is that it is one of the most popular data on UCI since 2007[2] which indicates I could find many articles working on the same data set to compare my results.

The data from the data set includes 517 instances which has 12 attributes as inputs and 1 output. I separate raw into two sets -- 450 data for training and model choosing, 56 data for testing. Several input encoding mechanism is applied in the data pre-processing stage. After that I use 10-fold cross validation mechanism to select one proper neural network model. Then, let inputs go through the model and use SGD mechanism and back-propagation mechanism to adjust the weight and train the model. Finally, use the 67 test data set to test the accuracy of model.

2 Data analysis and encoding

2.1 Input data encoding

According to forest fire data set, 12 input attributes are listed as following:

- **X** x-axis coordinate (from 1 to 9)
- **Y** y-axis coordinate (from 1 to 9)
- **month** Month of the year (Jan to Dec)
- **day** Day of the week (Mon to Sun)
- **FFMC** FFMC code(unknown encoded)
- **DMC** DMC code(unknown encoded)
- **DC** DC code(unknown encoded)
- **ISI** ISI index(unknown encoded)
- **temp** Outside temperature (in ℃)
- **RH** Outside relative humidity (in %)
- **wind** Outside wind speed (in km/h)
- **rain** Outside rain (in mm/m^2)

Data analysis and encoding decision

X Y

- Since all inputs should have similar scale
- Hence normalize by divided by 10

Month

- Is the circular information
- Nearby month should have similar input
- Cannot simply use 1 to 12 to encode
- Encode this attribute into 4 units(M1 to M4) to keep the information(**Table 1**)

Month	M1	M2	M3	M4
Jan	1	0	0	2
Feb	2	0	0	1
Mar	3	0	0	0
Apr	2	1	0	0
May	1	2	0	0
Jun	0	3	0	0
Jul	0	2	1	0
Aug	0	1	2	0
Sep	0	0	3	0
Oct	0	0	2	1
Nov	0	0	1	2
Dec	0	0	0	3

Table 1. Month encoded into 4 input attributes

FFMC	F1	F2	F3	F4
0-80	0.9	0.1	0.1	0.1
80-85	0.1	0.9	0.1	0.1
85-90	0.1	0.1	0.9	0.1
90-100	0.1	0.1	0.1	0.9

Table 2. FFMC encoded into 4 input attributes

Day

- Most of forest fire is caused by human
- Day of the week influence the frequency of human activities
- Simplify day into two class: weekday and weekend

FFMC

- Unknown encoded
- According to data frequency analysis, most of data lays between 80 and 100(**Figure 1**)
- Encoded into 4 different classes(**Table 2**)
- F1 from 0 to 80
- F2 from 80 to 85
- F3 from 85 to 90
- F4 from 90 to 100

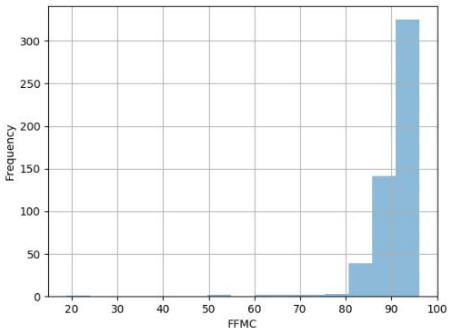

Figure 1. FFMC frequency diagram

DMC

- Unknown encoded
- According to data frequency analysis, the probability density just like the mixture of two Gaussian(**Figure 2**)
- Keep the raw data to avoid information lose
- Scale down the value of data by divided by 300

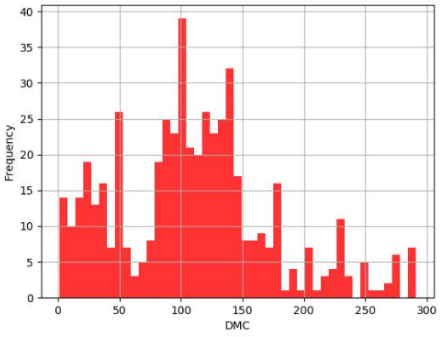

Figure 2. DMC frequency diagram

DC ISI temp RH

- According to data frequency analysis, all their probability densities broadly distributed(**Figure 3**)
- Keep the raw data to avoid information lose just like **DMC**
- Scale down the value of data by divided by their maximum values

Rain

- Important input attribute, since obviously fire area would be greatly influenced if it is raining
- According to data frequency analysis, the probability density tell me, in most case, the value is zero or in other word no rain(**Figure 4**)
- Simplify day into two class: not rain(R1) and rain(R2)

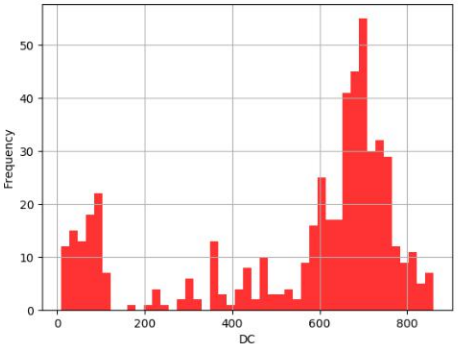

Figure 3. DC frequency diagram

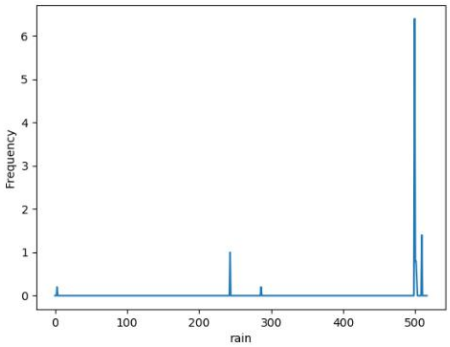

Figure 4. Rain frequency diagram

Rain	R1	R2
0-0.5	0.9	0.1
0.5-600	0.1	0.9

Table 3. Rain encoded into 2 input attributes

2.2 output encoding

Area

- The output of data set
- According to data frequency analysis, the probability density tell me, in most case, the value is zero and it is hard to tell any information from rest of area because their value is too low(**Figure 5**)
- Apply ln(area+1) to data set to to reduce skewness and improve symmetry[3](**Figure 6**)
- Because this is a classification problem, encode output into 4 different classes to classify the area of forest fire

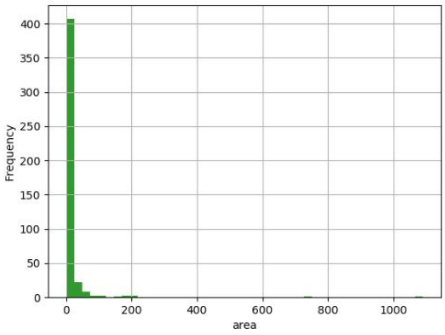

Figure 5. Area frequency diagram

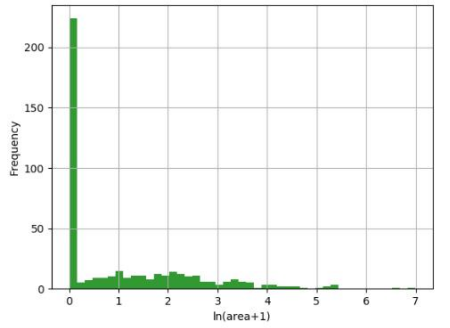

Figure 6. ln(Area+1) frequency diagram

3 Artificial neural network model

Cross validation

10-fold cross validation is applied to find the most suitable training model. 450 input data are separated into 10 folds and each time I pick up first of them as the rest of them to serve as the training set and. After that train the specified network and test the accuracy. Then I pick up the second validation set and second training set to do the same thing. After 10 loop, I average the test accuracy to evaluate the performance of the network model. The results of validation is as following:

Average accuracy	One hidden layer network	Two hidden layer network
40 nodes in HL1	50.82%	NA
80 nodes in HL1	50.82%	NA
200 nodes in HL1	50.82%	NA
80 nodes in HL1 20 nodes in HL2	NA	50.82%
40 nodes in HL1 10 nodes in HL2	NA	50.82%

Table 4. Average accuracy of 10-fold cross validation

I should have selected the model with the highest average accuracy, but the results are surprisingly identical. In the following work, I choose the two hidden layer neural network with 80 nodes in hidden layer 1 and 20 nodes in hidden layer 2. The learning rate is 0.01, number of inputs are 19, number of outputs are 4, number of epochs are 500.

4 Results and Discussion

After training, the accuracy of my model on the test data is 44.07%, which is a poor accuracy comparing with other researcher's work. For instance, according to "Prediction of Forest Fires Using Articial Neural Networks" by Youssef Sa and Abdelaziz Bouroumi[4], they worked on the same data set with similar approach of me and achieve over 95% of the accuracy. The detail of their coding is not published yet, expect for their hidden layer model, number of epochs and learning rate.

5 Conclusion and Future Work

In this paper an artificial neural network is implemented to predict forest fire. The input data pre-processing is the main focus of this article. Yet the final accuracy is no optimistic. Raw data using in the implementation can be found from UCI machine learning repository.

A lot of future work is waiting for me to do, especially when the current model has very poor performance. One sample of future work is to do the network pruning. Eliminate the unit with no function, the inverse unit and similar unit. Another idea is to change the network model and apply a better model, ie: Cascade Correlation, to improve the performance rather than using basic artificial neural network.

References

[1] Paulo Corte and Aníbal Morais, Forest Fires Data Set, Available: http://archive.ics.uci.edu/ml/datasets/Forest+Fires

[2] UCI Machine Learning Repository, Available: http://archive.ics.uci.edu/ml

[3] P. Cortez and A. Morais. "A Data Mining Approach to Predict Forest Fires using Meteorological Data", Department of Information Systems/R&D Algoritmi Centre, University of Minho, 4800-058 Guimar˜aes, Portugal, 2007

[4] Youssef Saand Abdelaziz Bouroumi, "Prediction of Forest Fires Using Articial Neural Networks", Modeling and Instrumentation Laboratory, Ben Msik Faculty of Sciences Hassan II Mohammedia-Casablanca University, BP.7955 Sidi Othmane Casablanca, 20702, Morocco, 2013