CS 532 Project Proposal

Topic: Machine learning approaches to predict forest fires

Name: Sicheng Fu

Email: sfu42@wisc.edu

Project Dataset

The data we processed in this project is from **UCI machine learning repository** (http://archive.ics.uci.edu/ml/datasets/Forest+Fires), which is a forest fire data collected from the Montesinho natural park. The data in the project using two sources, one was collected by the park inspectors, another was collected by meteorological station in the park.

In this dataset, we have a total of 517 entries, 12 features and 1 label. The first two columns denote the **X and Y axis values** where the fire occured, ranges from 1 to 9. The following two columns represent the **month and day of the week** when the fire burned. Next column come the **four FWI components** (FFMC, DMC, DC, ISI) that are affected directly by the weather conditions. Those four components value represents different moisture content of layers, both of them affect the forest fire velocity spread. Then some surrounding condition factors were considered to cause the forest fire. **Temperature**, **humidity**, **wind speed** and **rain** data were also listed in set. The **burned area** will be our output label, which represent the area size of the fire burned. We will classify the area data into 3 levels, when burned area smaller than 1 hectare, it labeled as fire danger level 1. When burned area greater than 1 hectare but smaller than 100 hectare, it labeled as fire danger level 2. And area greater than 100 hectare, labeled as fire danger level 3. So we will use 12 related features to predict the forest fire danger levels.

Therefore, all the data can gather into a 517*13 (one bias) dimensional matrix. Moreover, we will split our data into 70% training data and 30% testing data, and develop our model by using Linear regression, BPNN and SVM.

Algorithms:

Our model will applied three method: Linear regression, neural network and support vector machine.

Linear regression is a classical approach which has been widely used. The weighted vectors and parameters will be optimized using a least square method. And K-fold cross validation will be applied to derive a more accurate estimate of model performance.

For Neural Network and SVM, all attributes data will standardize to a zero mean and one standard deviation in the data pro-processing.

In our study, the multilayer perceptron back propagation neural network will be considered. Some parameters number of hidden layers **N**, number of hidden neurons at each layer **H**, and other hyper -parameters include learning rate, momentum, batch size and epoch size will be adjusted based on our test result. Similarly, holdout validation will be used to test the error rate.

For SVM model, the input data is transformed into a high dimensional feature space, by kernel function. Then, SVM finds the best linear separating hyper plane in feature space. Three parameters will affect our model, the width of the intensive zone **d**, the parameter of the kernel function **Gamma** and a regularization parameter **C.**

And the overall performance of each model is computed by mean squared error (MSE). $MSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N$

y_i represents the actual label values, and the $\hat{y}_i$ is the predictive values. Minimize the MSE is our objective for prediction. In addition, we will evaluate and compare different algorithm based on MSE values of testing data set.

Above all, we will use those three methods to predict the fire forest danger levels.

Project Github:

https://github.com/SichengIce/CS532_project

Project timeline:

Time	Tasks
10/23-11/1	Literature Review and Investigating Algorithms
11/1-11/5	Dataset processing
11/6-11/17	Dataset processing, First learning algorithm design
11/17-11/22	Second learning algorithm design
11/23-11/31	Third learning algorithm design
12/1-12/7	Evaluation and validation
12/7-12/15	Writing Final report