Analysis of MLP algorithm on large dataset

Deepanjali Sharma

Integrated M.Sc Computer Science

2013IMSCS005

IXth semester

INTRODUCTION

Neural Network:- It is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

Random Forest:- A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

Goal-

- To select a subset of relevant features for model construction.
- To classify an instance based on a set of predictors and also analyse the performance of CART and MLP algorithm on large dataset.

Dataset-

All the analysis has been performed on wave500k dataset. It is a well known dataset described in the CART book(Breiman et al., 1984) . Baisc Statistics of the dataset can be seen in **Table 1**

Total Instances	65536
No. of features	21
Class labels	A , B , C

Table1:- Basic Statistics of Dataset

Method-

The impurity decrease from each feature can be calculcted using Random Forest Classifier and the features are ranked according to this measure.

A multilayer perceptron (MLP) is a class of feedforward Artificial Neural Network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. The MLP consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes making it a deep neural network. Since MLPs are fully connected, each node in one layer connects with a certain weight to every node in the following layer.

Some of the parameters of MLP classifier are as follows:-

Solver:- The solver for weight optimization

Hidden_layer_sizes :- tuple, length = n_layers - 2, default (100,) The ith element represents the number of neurons in the ith hidden layer.

*Activatio*n:- {'identity', 'logistic', 'tanh', 'relu'}, default 'relu'. Activation function for the hidden layer.

Parameter Tuning

Parameter Tuning of *MLP* is done i.e, parameters are given different value and corresponding accuracy and time complexity is evaluated for each value. Eventually the parameters with value which gives the best precision compared to the other values is selected as the result for that *MLP* algorithm. Parameter Tuning for MLP can be clearly visualised from **Table 2**

SOLVER	ACTIVAT	HIDDEN LAYERS	ALPHA	Precision	Time
lbfgs	identity	5,2	1.00E-05	86.86%	19.71
sgd	weighted	5,8	0.008	86.62%	4.4
adam	tanh	516	0.092	86.68%	7.5

Table 2:- Parameter Tuning for MLP

Results

MLP algorithm gives the almost same precision with different hiidden layers, activation function and solver but, the complexity of the algorithm varies. From the above table we can observe that the complexity of the algorithm increases when alpha is very small and the hidden layers are also less. Also MLP proves to be much better than other algorithms ,one of them is CART as it gives much higher precison than CART with reduced Time complexity.

Conclusion & Discussion

MLP algorithm proves to be much better for the large dataset than other algorithms. Also the feature selection technique using random forest classier helps a lot for greater performance of the model. As it reduces the large dimension of the dataset by avoiding the irrelevant features of the dataset. Further in future we can even include the Backpropogation algorithm and also more efficient

techniques for feature selection to achieve higher accuracy.