HYPER PARAMETER OPTIMIZATION IN GRADIENT BOOSTING MACHINE MODEL

Dr.K.Chitra Assistant Professor, New Horizon College, Bangalore

chitranatarajan1979@yahoo.in

Abstract

Boosting is one of the techniques in Machine learning that uses the ensemble learning concept. Ensemble method is used to combine several base models to produce one optimal predictive model. Boosting is applied to the models with high bias and low variance. Boosting is often applied to decision trees. It is used to convert weak learners into strong learners. It is used to add the models to the ensemble sequentially. Gradient Boosting Machine or GBM is one of the popular Boosting algorithms. GBM combines the predictions from multiple decision trees to generate the final output. Every successive decision tree is built on the errors of the previous trees. The trees in a gradient boosting machine algorithm are built in sequential manner. The three elements of GBM are – loss function, weak learner and additive model. The loss function is used depends upon the problem. The decision trees are used as weak learners. Trees are added one by one and the existing one are not changed in additive model. ***Hyperparameters*** are adjustable parameters that should be tuned to obtain optimal output for the model. In the research paper, the various types of hyperparameters of GBM is discussed and the performance analysis of Boosting parameters like Learning rate and n_estimators are presented with example.

Keywords :

Learning rate, n_estimators Accuracy, parameters, datasets, validation, testing

I.INTRODUCTION

Gradient Boosting is the machine learning technique which uses ensemble techniques for producing the prediction. The GBM have three categories of parameters. They are Tree-specific Parameters, Boosting Parameters and Miscellaneous Parameters. Tree specific parameters affects each individual tree in the model. Boosting parameters affects the boosting operation in the model. Miscellaneous parameters are used for overall functioning. The tree specific parameters are min_samples_split,_min_samples_leaf, min_weight_fraction_leaf,_max_depth,_max_leaf_nodes,_max_features. The boosting parameters are learning_rate,_n_estimators,_subsample. The Miscellaneous parameters are loss,_init,_random_state, **verbose,_warm_start,,_presort.**

Hyperparameter tuning is used to control the overall behavior of a machine learning model.. Every machine learning model will have different types of hyperparameters.A hyperparameter value is set before the learning process starts. In this research paper , the hyper parameters in GBM model are discussed and the various values for boosting parameters like learning_rate and n_estimators are explained with example dataset. It provides several hyperparameter tuning options that make the function fit and very flexible. GBM provides predictive accuracy always. No data pre-processing is required for GBM model. It will work great with categorical and numerical values.

II.LITERATURE SURVEY

This section presents some of the existing research works related to decision trees and Gradient Boosting model. M Yao et al. [1] proposed an approach based on a gradient boosting decision tree (GBDT), to predict line loss rate. GBDT inherits the advantages of both statistical models and AI approaches, and can identify the complex and nonlinear relationship while computing the relative importance among variables. An empirical study on a data set in a city demonstrates that their proposed approach performs well in predicting line loss rate, given a large number of unlabeled examples. Experiments and analysis also confirmed the effectiveness of their proposed approach in anomaly detection and practical project management. F Wang, G Song [2] proposed a rapid and non-invasive structural health monitoring approach, their method can be applied to detect damages in other structures and thus guides future investigations.

L Kou et al.,[3] proposed ADASYN–GBDT method satisfied the requirements of real-time performance and accuracy for online fault detection. It might therefore aid in the fault detection of bogies. H Rao et al. [4] initialized the feature space spanned by the dataset. Less relevant features are suppressed according to the information they contribute to the decision making using an artificial bee colony algorithm. Experiments are conducted with two breast cancer datasets and six datasets from the public data repository. Experimental results demonstrate that the proposed method effectively reduces the dimensions of the dataset and achieves superior classification accuracy using the selected features.

Y Lee et al. [5] proposed algorithm that are assessed using the ERA-Interim and radiosonde observations (RAOB) as the reference data. The results show that the DNN model performs better than RF and XGB with a correlation coefficient of 0.96, a mean bias of 0.90 mm, and a root mean square error (RMSE) of 4.65 mm when compared to the ERA-Interim. Contributing variables to retrieve the TPW in each model and the spatial and temporal analysis of the retrieved TPW are carefully examined and discussed. W Alajali et al. [6] proposed three popular ensemble decision tree models that are used in the batch learning scheme, including Gradient Boosting Regression Trees (GBRT), Random Forest (RF) and Extreme Gradient Boosting Trees (XGBoost), while the Fast Incremental Model Trees with Drift Detection (FIMT-DD) model is adopted for the online learning scheme. The proposed approach is evaluated using public data sets released by the Victorian Government of Australia. The results indicate that the accuracy of intersection traffic prediction can be improved by incorporating nearby accidents and roadworks information

G Rong et al., [7] proposed the models that all have high enough model accuracy to be applied to produce LSM, the performance of the RF is better than the GBDT model without BO, while after adopting the Bayesian optimized hyperparameters, the prediction accuracy of the RF and GBDT models is improved by 1% and 7%, respectively and the Bayesian optimized GBDT model is the best for LSM in this four models. I Babajide Mustapha, F Saeed [8] investigated for the prediction of biological activity based on quantitative description of the compound's molecular structure using Gradient Boosting Machine. Seven datasets, well known in the literature were used in this paper and experimental results show that Xgboost can outperform machine learning algorithms like Random Forest (RF), Support Vector Machines (LSVM), Radial Basis Function Neural Network (RBFN) and Naïve Bayes (NB) for the prediction of biological activities. In addition to its ability to detect minority activity classes in highly imbalanced datasets, it showed remarkable performance on both high and low diversity datasets

III. HYPER PARAMETERS IN GRADIENT BOOSTING MACHINE

The hyperparameters of this model can be divided into 3 categories:

1. **Tree-Specific Parameters:** These parameters are used to affect each individual tree in the model.
2. **Boosting Parameters:** These parameters are used to affect the boosting operation in the model.
3. **Miscellaneous Parameters:** These parameters are used to affect the overall functioning.

Tree-Specific Parameters

1. **min_samples_split**
   - It defines the minimum number of observations that are used in a node to be considered for splitting.

2. **min_samples_leaf**
    - It defines the minimum observations needed in a leaf node.
3. **max_depth**
    - It defines the maximum depth of a tree.
4. **max_leaf_nodes**
    - It defines the maximum number of terminal nodes in a tree.
5. **max_features**
    - It defines the number of features to consider while searching for a best split. These will be selected randomly

Boosting Parameters

1. **learning_rate**

- It determines the contribution of each tree on the final outcome
- These Values ranges from 0–1.
- Lower values are generally preferred as they make the model robust to the specific characteristics of tree.

2. **n_estimators**

- The number of sequential trees to be modeled
- GBM is robust at higher number of trees. This should be tuned using CV for a particular learning rate.

3. **subsample**

- The fraction of observations to be selected for each tree. Selection is done by random sampling.
- Values slightly less than 1 make the model robust by reducing the variance.

Miscellaneous Parameters

1. **loss**

- It refers to the loss function to be minimized in each split.
    - It can have various values for classification and regression case.

2. **init**

- This affects initialization of the output.
- This can be used if we have made another model whose outcome is to be used as the initial estimates for GBM.

3. **random_state**
   - The random number seed so that same random numbers are generated every time.
   - It can potentially result in overfitting to a particular random sample selected.

- **verbose**
   - The type of output to be printed when the model fits. The different values can be:
     - 0: no output generated
     - 1: output generated for trees in certain intervals
     - >1: output generated for all trees

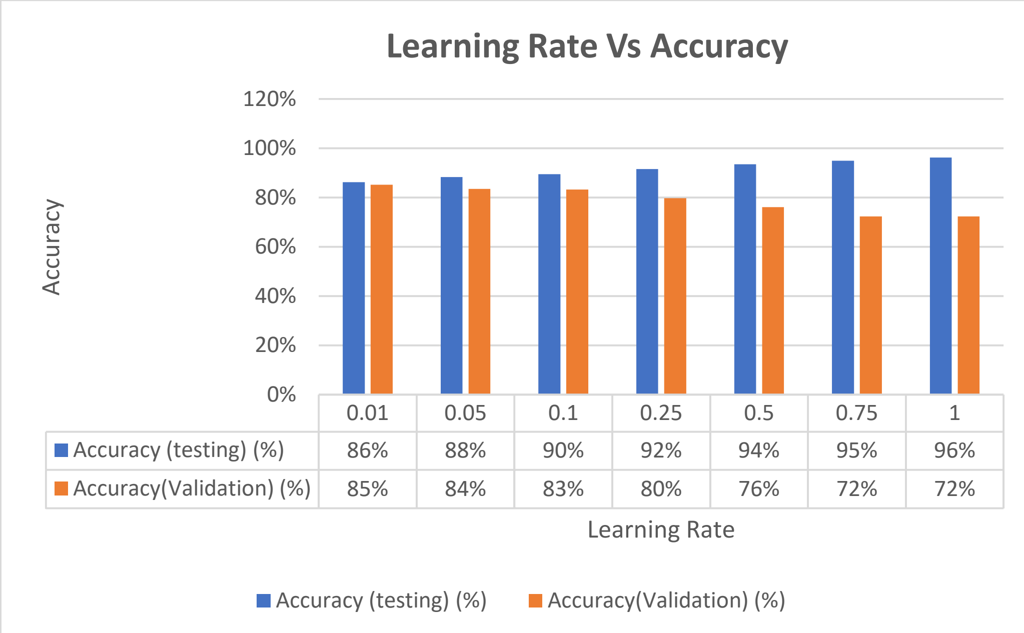
4. **presort**
   - Select whether to presort data for faster splits.
   - It makes the selection automatically by default but it can be changed if needed.

IV.PERFORMANCE ANALYSIS

This section presents the performance analysis of Learning rate and n_estimators parameters for finding the accuracy of the prediction. The main objectives of the study is to analyze the hyper parameters of GBM model by considering the various values of hyper parameters like learning rate and n_estimators in the temperature dataset. Eight variables were analyzed, namely: day, month, year, Week, temp1, temp2, average, actual. The dataset were separated into a testing dataset (70%) and a validation dataset (30%). The temperature is predicted in the dataset. The comparison of accuracy of testing and validation datasets for various values of hyper parameters were achieved.

A) Impact of Learning rate

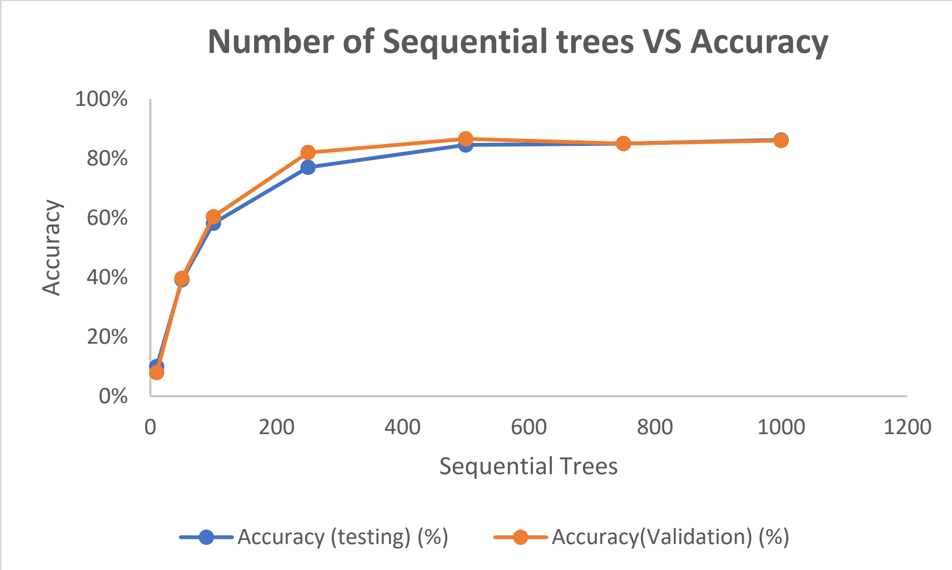
Learning Rate	Accuracy (testing) (%)	Accuracy(Validation) (%)
0.01	86%	85%
0.05	88%	84%
0.1	90%	83%
0.25	92%	80%
0.5	94%	76%
0.75	95%	72%
1	96%	72%



In this research. the range of the learning rate is considered from 0.01 to 1. If the learning rate value increases, accuracy also increases in testing dataset and the accuracy decreases in validation dataset.

B) Impact of n_estimators Parameter

Number of Sequential Trees	Accuracy (testing) (%)	Accuracy(Validation) (%)
10	10%	8%
50	39%	40%
100	58%	60%
250	77%	82%
500	85%	87%
750	85%	85%
1000	86%	86%



In this research, the range of n_estimators parameter is considered from 10 to 1000. If the n_estimators value increases, accuracy increases in testing dataset and validation dataset.

V.CONCLUSION

Gradient boosting is a one of the machine learning technique for regression and classification problems, which gives prediction model in the form of an ensemble of weak prediction models. A **hyperparameter** is a parameter whose value is **used** to control the learning process. Hyper parameters are used to control the behavior of the training algorithm and have a impact on the performance of the training model. The benefits of good **hyperparameters** are 1) Optimal result of the model 2) to manage a large set of experiments for **hyperparameter** tuning.

The main objective of this research work is to analyze the performance of hyper parameters like learning_rate and n_estimators in testing and validation dataset. If the learning rate value increases, accuracy also increases in testing dataset and the accuracy decreases in validation dataset. If the n_estimators value increases, accuracy increases in testing dataset and validation dataset.

VI.REFERENCES

[1] M Yao, *et al.*, "Research on Predicting Line Loss Rate in Low Voltage Distribution Network Based on Gradient Boosting Decision Tree", *Energies* **2019**, *12*(13), 2522; **https://doi.org/10.3390/en12132522**

[2] F Wang, G Song ., "**Bolt-looseness detection by a new percussion-based method using multifractal analysis and** gradient boosting decision tree", Structural Health Monitoring, 2020 - journals.sagepub.com, Volume: 19 issue: 6, page(s): 2023-2032

[3] L Kou, Y Qin, X Zhao, Y Fu "**Integrating synthetic minority oversampling and** gradient boosting decision tree **for bogie fault diagnosis in rail vehicles**",- journals.sagepub.com

Volume: 233 issue: 3, page(s): 312-325

[4] H Rao et al., "**Feature selection based on artificial bee colony and** gradient boosting decision tree",Elsevier, Available online 5 November 2018.

[5] Y Lee et al., "**Retrieval of Total Precipitable Water from Himawari-8 AHI Data: A Comparison of Random Forest, Extreme Gradient Boosting, and Deep Neural Network**",- Remote Sensing, 2019 - mdpi.com ,*Remote Sens.* **2019**, *11*(15), 1741; **https://doi.org/10.3390/rs11151741**

[6] **Walaa Alajali**, *et al.*, "**Intersection Traffic Prediction Using Decision Tree Models**" ", September 2018,Symmetry 10(9):386,DOI:10.3390/sym10090386

[7] G Rong, *et al.*, "**Rainfall Induced Landslide Susceptibility Mapping Based on Bayesian Optimized Random Forest and Gradient Boosting Decision Tree Models—A Case Study of Shuicheng County, China**", *Water* **2020**, *12*(11), 3066; **https://doi.org/10.3390/w12113066**

[8] I Babajide Mustapha, F Saeed., "**Bioactive Molecule Prediction Using Extreme Gradient Boosting**", Molecules, 2016 - mdpi.comAcademic Editor: Leif A. Eriksson,*Molecules* **2016**, *21*(8), 983; **https://doi.org/10.3390/molecules21080983**

