International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies

Syeda Sarah Azmi[1], Shwetha Baliga[2]

[1]*Student, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59, Karnataka, India,*
syedasarahazmi.ec16@rvce.edu.in, 9742854662
[2]*Assistant Professor, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59, Karnataka, India, shwethaprabhun@rvce.edu.in, 8892303813*

Abstract - *Decision tree is a well-known Machine Learning algorithm that is used for regression and classification. It is an inductive inference algorithm mainly known for its interpretability and representability. The internal nodes of a decision tree act like tests whereas the leaf nodes give the final category. Decision trees are known to suffer from bias variance. There is a huge predisposition related with straightforward trees and a huge difference with complex trees. To overcome this drawback Ensemble methods are adopted, it combines the performance of several decision trees to produce a more accurate prediction than a single tree. Boosting is one such ensemble technique. In this paper AdaBoosting (Adaptive Boosting) and XGBoosting (Extreme Gradient Boosting), two famous boosting techniques are discussed.*

Key Words: **Decision tree, AdaBoost, XGBoost, Boosting, Ensemble techniques.**

1. INTRODUCTION

There are different decisions one needs to make regularly. The way toward choosing an alternative from the huge number of decisions is the embodiment of dynamic. Decision tree is a calculation that encourages us settle on choices by taking in all the significant qualities into thought. Decision trees are a profoundly well-known AI (Artificial Intelligence) procedure that has been effectively applied to take care of various functional issues in zones, for example, regular language handling, data extraction, information mining, and example acknowledgment. Decision trees are additionally the establishment of further developed AI strategies, for example, arbitrary backwoods or helped trees [1].

Decision trees experience the ill effects of predisposition difference. To beat this Ensemble methods are utilized. Two such strategies are Bagging and Boosting. Here, boosting procedures are examined AdaBoost (Adaptive Boosting) and XGBoost (Extreme Gradient Boost). AdaBoost is a Boosting group of calculations. This kind of student is portrayed by more consideration on the wrongly arranged examples during preparing, modifying the example dispersion, and rehashing this activity until the quantity of trainings of the powerless classifier arrives at a pre-indicated esteem, finishing the learning [2].

XGBoost is an effective and versatile execution of the Gradient Boosting Machine (GBM), which has been a serious device among man-made brainpower techniques because of its highlights, for example, simple parallelism and high forecast exactness [3].

The structure of this paper is outlined as follows; in Section2, this paper discusses about the different Machine Learning algorithms starting from simple decision tree progressing to boosting algorithms XGBoost and AdaBoost. Section 3 concludes the paper discussing the various factors that will determine the selection of a particular Machine learning algorithm for classification.

2. MACHINE LEARNING ALGORITHMS

The different Machine learning algorithms are discussed in the following sub-sections.

2.1 DECISION TREE ALGORITHM

Decision Tree is a tree structure similar to the root of a tree. The ID3 algorithm proposed R. Quinlan is known as the most classical one [4]. The root node of the decision tree poses the first most important question depending on the answer the next node is selected from the child nodes, the child node now poses a new question and this process repeats until the leaf node is reached and the data point is finally categorized by the result the leaf node holds. Fig -1 shows the basic structure of a decision tree, each non-leaf node represents an input attribute of data set, “Attribute value” represent the value of this attribute, and leaf node represent value of output attribute.

From the point of view of the entire tree, Decision Tree speak to disjunction of characteristic worth restriction's combination. Each way from the root to the leaves relate to a lot of characteristics of conjunctions, disjunction of these conjunctions compares to the tree itself. So it is anything but difficult to change over into IF-THEN type of arrangement

rules. As indicated by the arrangement rules obscure information item can be effortlessly characterized and anticipated [5].

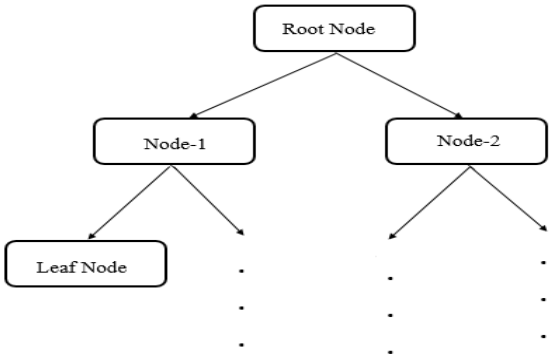

Fig -1 : Basic Structure of a Decision tree

2.2 XGBOOST ALGORITHM

XGBoost stands for Extreme Gradient Boost. It is the streamlined group calculation dependent on GBDT (Gradient Boosting Decision Tree). The principle thought of the boosting calculation is that numerous decision trees perform superior to a solitary one. Every decision tree may make an awful showing. At the point when numerous trees are incorporated, the presentation shows signs of improvement.

Fig -2 shows the work-flow of XGBoost algorithm when used to create an ML (Machine Learning) model. The principle steps of the calculation are recorded as follows:

1. Generate an emphatically related standard set, discretize the component vectors if the element in preparing set is constant.

2. Choose the principles whose result is the forecast mark, at that point structure the new standard set. Ascertain the lift of each standard and evacuate the principles with lift<1.0. In this way, we can get the compelling guideline set.

3. Sort the principles in the successful standard set by most brief length, greatest lift, and greatest help.

4. Print the component in the list of capabilities and quit if cycle record is more prominent than the given number. Else, we get the primary guideline in the successful standard set and add its resulting to the list of capabilities, expel the standard from the viable principle set. We don't have to include the component which as of now exists on the list of capabilities.

5. Delete examples that fulfill the standard, ascertain lift and backing in the rest of the preparation set.

6. Sort the standards in the viable principle set by most extreme lift, greatest help, go to Step 4 [6].

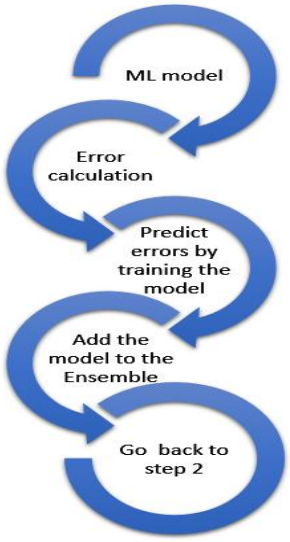

Fig -2: Work-flow of XGBoost

XGBoost is a decision tree-based gathering Machine Learning calculation that utilizes an inclination boosting system. In forecast issues including unstructured data (pictures, content, and so on.) artificial neural systems will in general beat every other calculation or structures. Be that as it may, with regards to little to-medium organized/forbidden data, decision tree based calculations are viewed as best.

2.3 ADABOOST ALGORITHM

AdaBoost stands for Adaptive Boost. An AdaBoost classifier begins with fitting a classifier on the original data set and later starts fitting a copy of the same classifier on an updated dataset which the inaccurate and error prone data points have been balanced out in such a way that the proceeding classifiers can focus on the cases that cause more inaccuracy [7].

Fig -3 is the sample representation of how AdaBoost functions. It keeps magnifying the weight of the misclassified data points in order to correct the error and finally adds up the output of the different classification models to create a more accurate combined classifier.

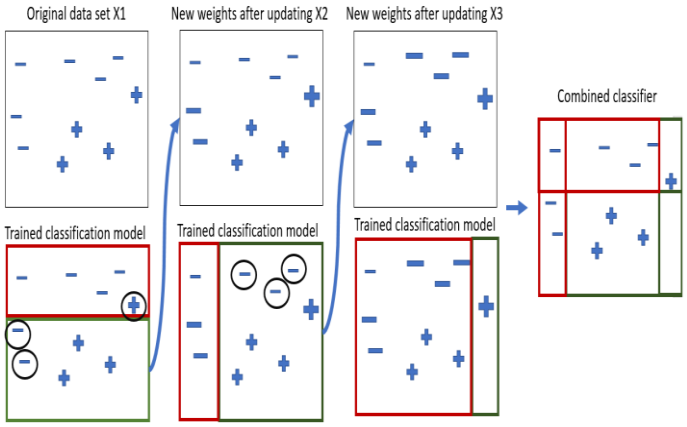

Fig -3: Sample representation of the workflow of AdaBoost

AdaBoost is a sort of iterative calculation, the principle thought of the calculation is preparing various classifiers (powerless classifiers) with a preparation set, at that point utilizing a few different ways to coordinate them to develop a more grounded classifier. The calculation itself is accomplished by changing the information circulation, which is as indicated by the order amendment of test of the preparation set, just as the last precision of the general arrangement to decide the heaviness of each example. At that point send the loads of the changed new information to the lower classifier for preparing, lastly each preparation classifier is melded as an official conclusion classifier. Calculation key points are as per the following:

1. Each emphasis of the change is the conveyance of the example, as opposed to the resampling;

2. The difference in test dispersion relies upon whether the examples were accurately characterized, and the arranged effectively test weight low, the misclassified test weight high (generally the example close to the limit), this will empower the following classifier center around the current misclassified tests;

3. Combine the entirety of the powerless classifiers to get the outcome [8].

Hence, it is observed that AdaBoost focusses on the wrongly identified points by increasing the weights of the misclassified data thereby decreasing the error and increasing the accuracy.

3. CONCLUSIONS AND FUTURE SCOPE

Decision tree is a broadly utilized calculation yet it experiences certain disadvantages, for example, inclination change. There is a huge inclination related with straightforward trees and a huge change with complex trees. To conquer this disadvantage Ensemble techniques are embraced, it consolidates the presentation of a few decision trees to deliver a more precise expectation than a solitary tree. Boosting is one such group strategy. The AdaBoost and XGBoost boosting algorithms were reviewed in this paper.

This paper focusses on the different ensemble boosting techniques used to improve the result obtained by the classical decision tree algorithm. It is observed that the XGBoost algorithm performs well with large amount of unstructured data but when it comes to small to medium data in a structured form decision tree is the best model to use. AdaBoost on the other hand focusses on the wrongly identified points by increasing the weights of the misclassified data.

AdaBoost is a consecutive gathering procedure though XGBoost is an equal troupe system. Models are free in XGBoost, while in AdaBoost attempts to include new models that do well when past models come up short. Weighted Average is equivalent in XGBoost, while in AdaBoost more weight is given to the models with better execution on preparing information. XGBoost is inclined to over-fitting, yet AdaBoost attempts to decrease inclination. Thus, it is concluded that it is not necessary that the boosting algorithms always outperform the simple decision tree model. The selection of the algorithm to be used it depends on the type of data that needs to be classified.

REFERENCES

[1] R. García Leiva, A. Fernández Anta, V. Mancuso and P. Casari, "A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design," in IEEE Access, vol. 7, pp. 99978-99987, 2019.

[2] Y. Zhang et al., "Research and Application of AdaBoost Algorithm Based on SVM," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 662-666.

[3] M. Chen, Q. Liu, S. Chen, Y. Liu, C. Zhang and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," in IEEE Access, vol. 7, pp. 13149-13158, 2019.

[4] L. Fang, H. Jiang and S. Cui, "An improved decision tree algorithm based on mutual information," 2017 13th

International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 1615-1620.

[5] V. Hable-Khandekar and P. Srinath, "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-6.

[6] X. Shi, Q. Li, Y. Qi, T. Huang and J. Li, "An accident prediction approach based on XGBoost," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, pp. 1-7.

[7] V. Bobkov, A. Bobkova, S. Porshnev and V. Zuzin, "The application of ensemble learning for delineation of the left ventricle on echocardiographic records," 2016 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, 2016, pp. 1-5.

[8] X. Shu and P. Wang, "An Improved Adaboost Algorithm Based on Uncertain Functions," 2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, Wuhan, 2015, pp. 136-139