Boosting Accuracy of Supervised Algorithm with the Introduction of Helper Constant

Saravanakumar C Shanmugam [1], Sivakartik Sreedhara [2]

[1] *Software Architect, Bosch Global Software Technologies, India*

[2] *Engineering Manager, Bosch Global Software Technologies, India*

Abstract: If any machine learning algorithm the success of it is based on the accuracy. If the algorithm accuracy decreases due to the increase in the size of the feature vector then Boosting techniques help the algorithms (in this case it is Regression algorithms) to maintain or improve the accuracy. In this paper we have dealt with the supervised algorithms. The accuracy of the supervised algorithms is improved with the introduction of wrapper constants so that the accuracy is improved with the large dataset as well as with increase in features.

Keywords: Boosting, Feature vector space, Machine Learning, Supervised learning.

1. INTRODUCTION

In the modern world the impact of Artificial Intelligence and Machine learning is huge, such that it plays a vital role in bringing the automation in our daily life. Examples of the same is in various applications such as autonomous car driving, image processing, Forecast etc., One should be aware of the difference between Artificial Intelligence and Machine Learning. If a human wants to be intelligent the first step is to learn. Likewise if a machine need to be artificially intelligent then the machine has to learn and update the knowledge through Machine Learning. Therefore Artificial Intelligence is mirroring human intelligence with the learning process through Machine Learning.

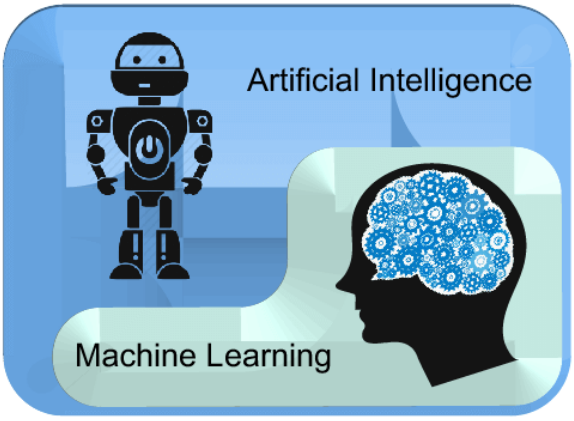


Fig 1. Artificial Intelligence Vs Machine Learning

Machine Learning is broadly classified into three important types namely Supervised Learning (SL), Unsupervised Learning (UL) and Reinforcement Learning. Reinforcement learning works on the feedback it receives, whereas unsupervised learning does not need any labelled output defined. Supervised Machine learning algorithms is one of the important categories of machine learning algorithms where it works on the labelled output and is used mainly for Classification and regression problems. In the classification problems for a new unseen data we will be classifying it based on the labelled dataset. In the regression problem we will predict the value of a continuous variable. Supervised learning algorithms land up in the decrease of accuracy when there is more number of dataset as well as when there are more dimensions involved. Through this paper we present a boosting concept to increase accuracy of the supervised machine learning algorithms irrespective of the scenarios. Section 2 walks through on different boosting

algorithms available for Supervised Machine Learning algorithms, Section 3 describes about our own approach of boosting technique, Section 4 briefs on Results and finally Section 5 provides the conclusion.

2. BOOSTING TECHNIQUES

Boosting is a technique used in Machine learning algorithms to solve complex data driven problems. Boosting is a technique used to convert a weak model into a strong model so that the accuracy of the model will be improved.

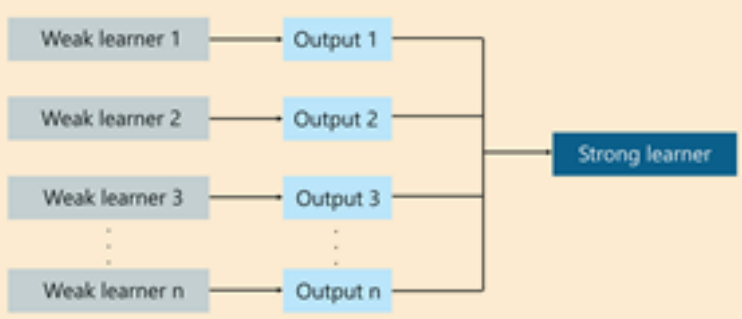


Fig 2. Boosting mechanism

There are different kind of boosting techniques which use different kind of engines and the important ones are as follows,

AdaBoost (Adaptive Boosting):

For different weighted training data AdaBoost fits a sequence of weak learners. The action starts by predicting the original dataset where each observation has been given equal weightage. If the first weak learner predicts wrongly then higher weight is given to the observation which has been predicted wrongly. In an incremental way more learners are added until the threshold value is reached with respect to models or accuracy. Basically decision stamps are used along with AdaBoost but we can also any machine learning algorithms as a base which accepts weight on dataset. AdaBoost is used in Classification and Regression problems.

Gradient Boosting

Gradient Boosting trains many models sequentially. Each model tries to minimize the loss function. The learning procedure consecutively fit new models to provide a more accurate estimate of the response variable. The principle idea is to construct new base learners so that they can correlate with the negative gradient of the loss function, associated with the whole ensemble. Gradient Boosting can be used for regression and classification problems.

3. BOOSTING WITH HELPER CONSTANT

When we train the dataset with the supervised machine learning algorithms in many scenarios the accuracy of the model decreases when the dataset size increases. Also supervised learning face the problem of over-trained and due to these issues the model loses its strength and it becomes weak. To prevent the model losing its originality in accuracy and boost its accuracy strength we have come up with the helper constant. The helper constant algorithm is explained in steps as follows,

STEP 1: Each attribute is extracted and fit for different model running with different supervised learning algorithms.

STEP 2: For each model its accuracy is weighed and the best accuracy model is selected.

STEP 3: With the best accuracy model selected add the average value of the objective class to the attributes weightage and divide it with the booster constant value ranging from 1 to 2.

$$\mathbf{BHC = (Attr_{(i)} + Avg_{OC}) / B_C} \quad \textbf{(1)}$$

Where $Attr_{(i)}$ is collection of attributes

Avg_{OC} is average value of Objective class.

B_C is the booster constant value which can be chosen between the range 1 to 2.

Identification of C value for the booster constant is done with the help of elbow diagram.

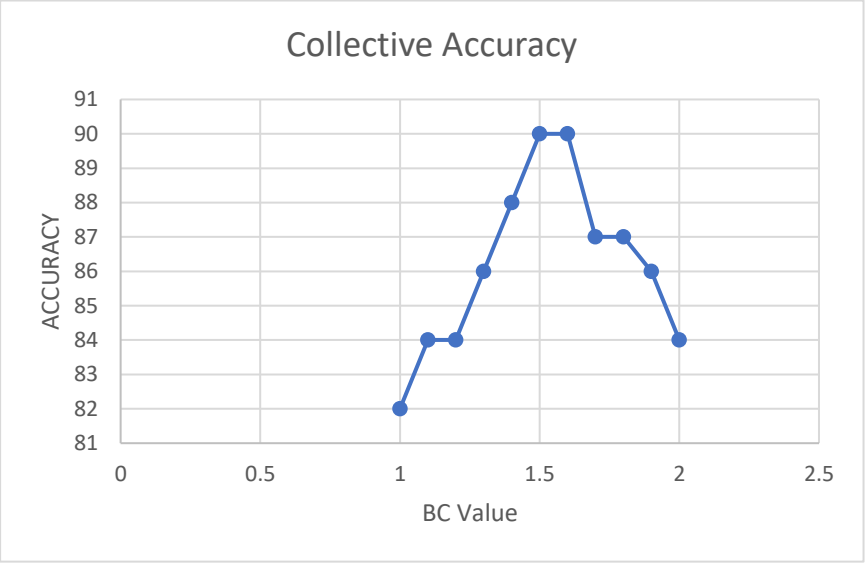
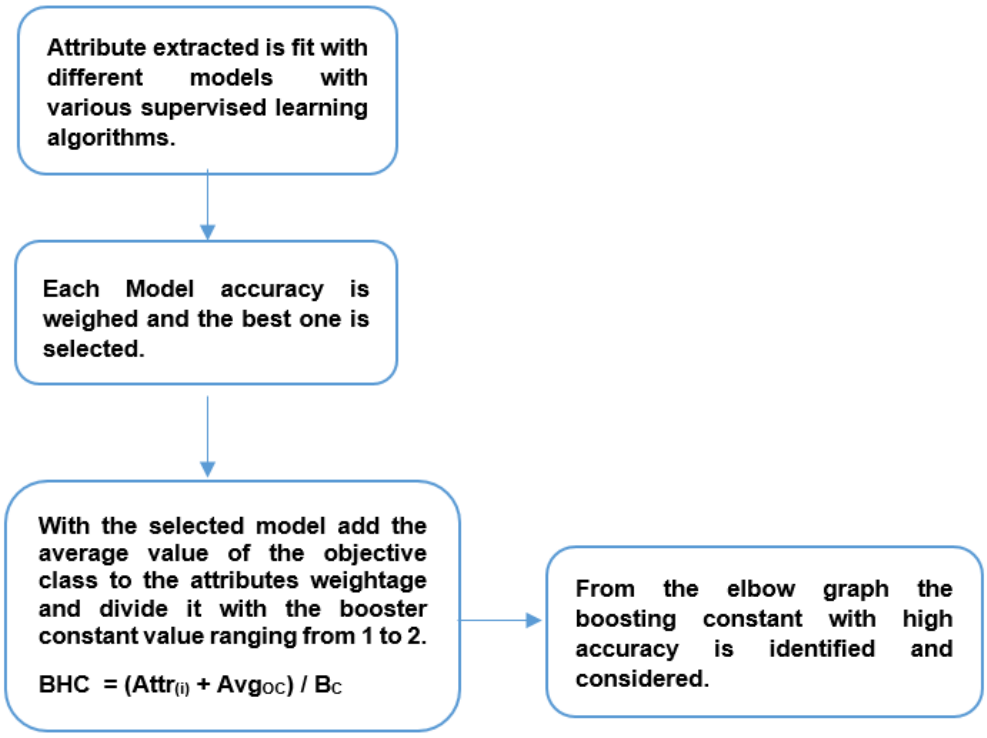


Fig 3. Identifying the best BC value

STEP 4: Booster constant value is identified with the help of elbow diagram where the value of the constant ranging between 1 to 2 is tested for various trials and the collective accuracy is noted. The accuracy stands best is where the Booster constant is finalized.



4. RESULTS AND DISCUSSIONS

We have taken Microsoft paraphrase corpus to validate our concepts. From the Microsoft corpus required attributes are extracted and the objective is text classification to conclude whether the given texts are paraphrase or not. The first model that has been chosen is Logistic Regression.

The accuracy of the Logistic regression in classifying the paraphrase corpus before applying the booster helper constant is as follows,

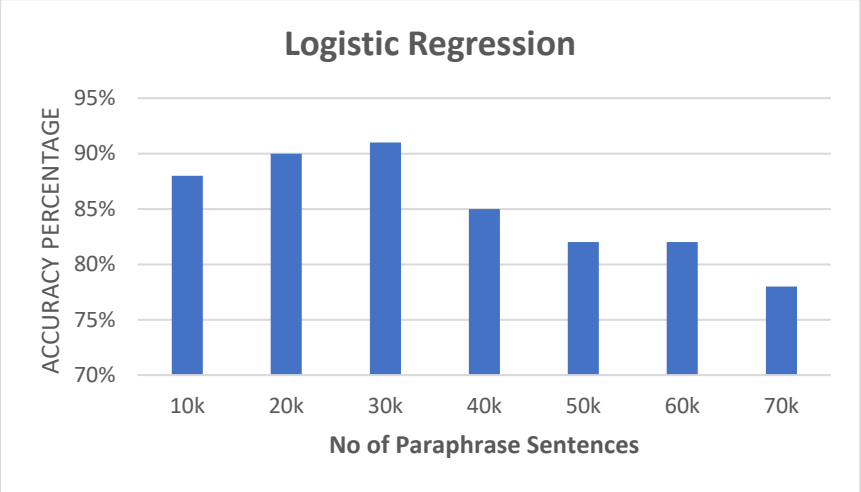


Fig 4. Logistic Regression without BHC

Now involving our Boosting concept the formula is applied and also the best helper constant value is identified with the help of elbow diagram where the value are tried from 1 to 2.

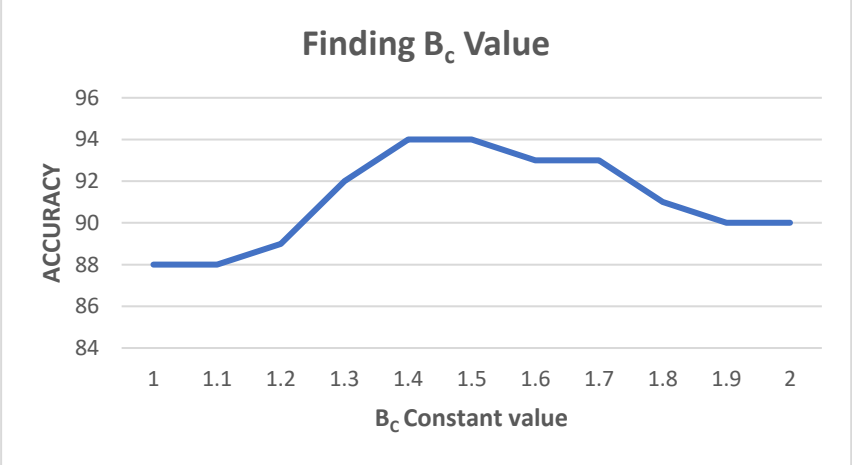


Fig 5. Identifying the best B_C Value

By choosing Booster constant value as 1.4 and by using the formula the accuracy of the Logistic Regression is calculated.

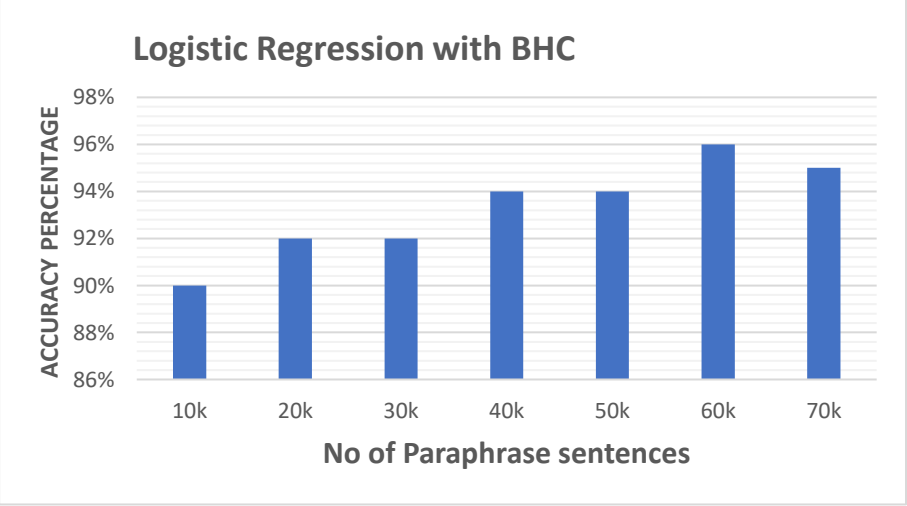


Fig 6. Logistic Regression with BHC

Now taking another regression technique, Linear Regression the same comparison is listed as follows,

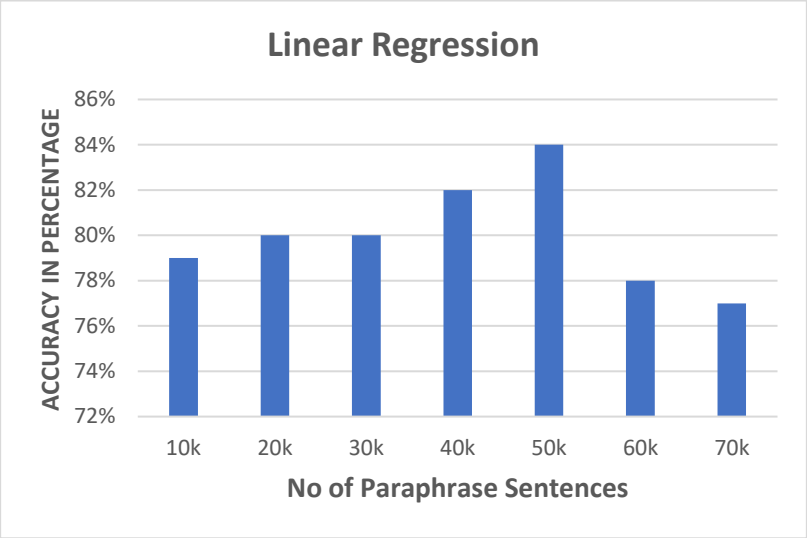


Fig 7. Linear Regression without BHC

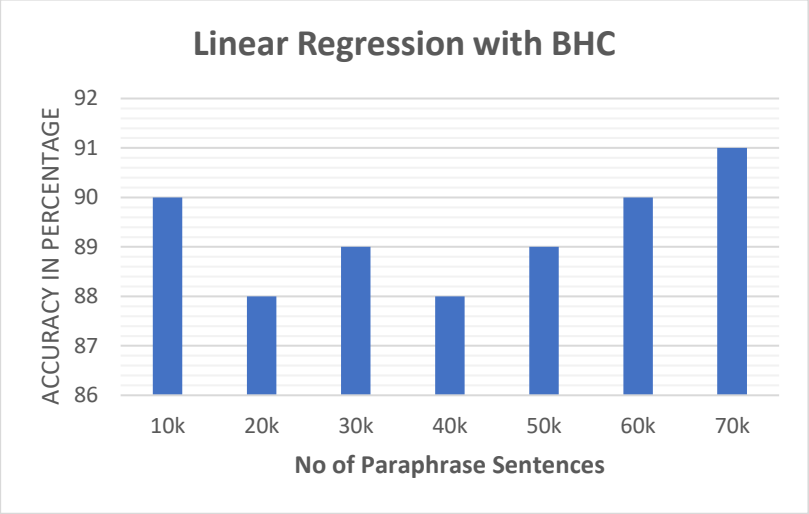


Fig 8. Linear Regression with BHC

5. CONCLUSION

According to ensemble equivalence, through this paper we have analyzed the existing boosting techniques and for various supervised machine learning models we have been able to fit our boosting technique and observed that the accuracy of the supervised learning algorithms which was decreased when huge volume of data is involved looks to be stable or with the improved accuracy with our approach of boosting technique. Going forward, to ensure on the improvement of our technique with respect to compatibility we will further focus on Unsupervised and Semi supervised algorithms.

6. REFERENCES

[1] C S SaravanaKumar, Dr. R.Santhosh, "Effective information retrieval and feature minimization technique for Semantic Web Data", Computers and Electrical Engineering, 2019.

[2] Nathan Neuhaus, Boris Kovalerchuk, “Interpretable Machine Learning with Boosting by Boolean Algorithm”, Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV), 2019

[3] C.S.Saravana Kumar, Dr.M.Mohanapriya, Dr.C.Kalaiarasan, “A New Approach for Information Retrieval in Semantic Web Mining Involving Weighted Relationship”, International conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, 2017.

[4] Pavan Kumar, Mallapragada, Rong Jin, Anil K. Jain, Yi Liu, “SemiBoost:Boosting for Semi-Supervised Learning”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009

[5] C S Saravana Kumar, Sivakartik Sreedhara, “A Noval Approach to identify Build Failure Causes involving Machine Learning Tehcniques”, International Conference Automatics and Informatics (ICAI), 2021.

[6] Hui Zhang, Yi Liu, Bojun Xie, Jian Yu, “A boosting approach to learning receptive fields for scene categorization”, IEEE International conference on Image processing, 2013.

DOI: https://doi.org/10.15379/ijmst.v10i2.2962