IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 10, Issue 5, May 2020, ||Series -I|| PP 58-63

Theoretical Framework of Ensemble Methods and Applications

Barkha Bhardwaj[1*], Shivam Tiwari[2*], Deepshikha Agarwal[3], Garima Srivastava[4]

[1234]*Department of Computer Science & Engineering, Amity School of Engineering & Technology*
Amity University, Uttar Pradesh, India
Received 9 May 2020; Accepted 22 May 2020

Abstract: Ensemble methods is a strategy based algorithm. It is the combination of best suited machine learning algorithms to get more reliable, accruable result, ensemble methods came in to picture. Troupes (ensemble) are sets of learning machines that consolidate their choices or their learning calculations, or different perspectives on information, or other specific attributes to acquire increasingly solid and progressively precise expectations in regulated and also learning issues. In this paper the previous study related to Ensemble methods is discussed and gives a future aspect of Ensemble methods. Here also discussed about the categorization of Ensemble methods, As well as the comparison between single machine learning model and combination of multiple models is given. These days outfit strategies speak to one of the fundamental momentum research lines in Artificial Intelligence. The main purpose of this study is to spread out the importance of Ensemble Methods and prediction of future aspect of it.
Keywords: Ensemble Methods, Combination of classifiers.

I. INTRODUCTION

Outfit strategies is an Artificial Intelligence procedure that joins a few base models so as to deliver one ideal prescient model. Numerous specialists have examined the method of joining the expectations of different classifiers to deliver a solitary classifier. Gathering learning (combination of classifiers) improves Artificial Intelligence results by joining a few models. This methodology permits the creation of better prescient execution contrasted with a solitary model. That is the reason outfit techniques set first in numerous renowned Artificial Intelligence rivalries, for example - The Netflix Competition, KDD 2009 and kaggle Ensemble methods can be isolated into two gatherings:

1. Sequential Ensemble method (Dependencies between base learners)
2. Parallel Ensemble method (No dependencies between base learners)

Popular Ensemble methods are:

- Bagging
- Boosting

Bagging

Bagging, it is also known as bootstrap aggregating bagging gets its name since its joins bootstrapping and aggregation to frame one outfit model. Given an example of information, different bootstrapped subsamples are considered. A decision tree is shaped on each of the bootstrapped considered sample data. For each consideration decision tree has been formed. After that an aggregation takes place for most effective prediction.

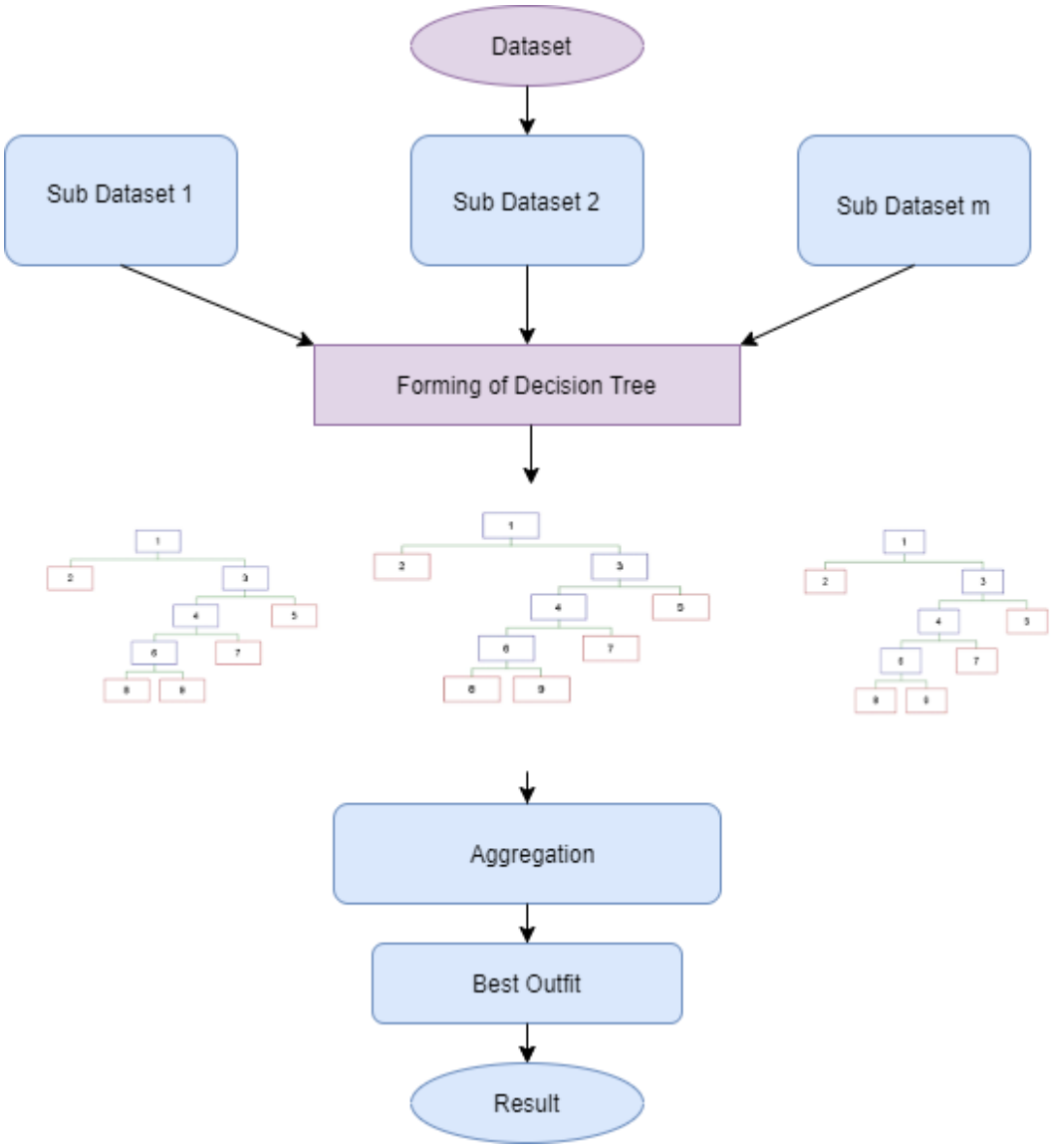

Figure 1: Systematic Bagging Model for Data Analysis

We can complete the ensemble by given formula. As per consideration of 'm' sub dataset, there will be 'm' decision tree:

$$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

Bagging is an approach to diminish the change in the forecast by creating extra information for preparing from dataset utilizing blends with reiterations to deliver multi-sets of the first information.

Boosting

Boosting is an iterative procedure which changes the heaviness of a perception dependent on the last grouping. On the off chance that a perception was characterized erroneously, it attempts to expand the heaviness of this perception. Boosting all in all forms solid prescient models.

II. DIFFERENCE BETWEEN BAGGING AND BOOSTING

Table 1: Bagging and Boosting

Parameters	**Bagging**	**Boosting**
Data	Random	Higher priority based samples
Goal	Low variance value	Boosting accuracy
Model	Random subspace model	Gradient descent approach
Example	Random forest	Ada boost

III. REASON BEHIND ERROR TAKES PLACE IN MACHINE LEARNING MODEL

There are three basic factors, their names are given below:

1. Variance
2. Bias
3. Noise in data

We are using ensemble methods to decreasing the error ratio. Many researches found that instead of using single classifier combination of multiple classifiers are used then the predictive result is more reliable with more accuracy.

Model Name	**Explanation**
Bagging	Decreasing variance properly
Boosting	Decrease the bias of the model
Stacking	Increment in predictive parameters of the classification algorithms of machine learning

IV. LITERATURE REVIEW AND DISCUSSION

In[1], Lei Su and Zhengta Yu explained the details of how a question answering system is prepared and how the accuracy of the system can be increased. By using ensemble methods, they observed the better results instead of using single classifier. The paper shows that experiment is done on Chinese question answer system for tourism domain. By using ensemble methods bagging and boosting the performance can be increased. Here Chinese question data is used. This data set having 5 coarse type and around 23 fine types for example:

Coarse type: Place

Fine type: Place synopsis, traffic, climate, parking etc.

Data set containing 5 coarse names are given below:

1. Science spot
2. Hotel
3. Place
4. Local customers
5. other

there are 2 sub data sets are used SSID1 and SSID2. In result, applying Bagging on SSID1 then found the accuracy 91.43% which is increased by 1.36 percentage from single classifier used. The second result for the second data set SSID2, by applying Bagging on this, the result is 82.98% accuracy. Which is better by 2.02 percentage than the single classifier machine learning algorithm.

In[2], this paper a new learning algorithm is introduced names as Lasso-Bagging. It is a modified ensemble method. This algorithm gives a better accuracy results as compared to single use of decision tree. In order to the introduction of this new method, in simple words it is modified Bagging tool and it improves the learning capacity of Bagging and it is faster to build a decision tree for sub data set.

The efficiency of this Lasso-Bagging is much better than the efficiency of Bagging tool of ensemble methods. This Lasso-Bagging methods containing selective ensembles.

In this paper UCI data sets are used. There are 6 data sets used for comparing the performance .

Datasets:

1. Boston housing
2. Algae UCI data
3. Ozone data set
4. Friedman dataset
5. Simulated friedman
6. O3 DMEF data

- Boston housing UCI data set having 501 samples and 11 variables.
- Ozone data set having 330 samples and 7 variables.
- Algae UCI data having 338 samples and 11 variables.
- Friedman data set having 1200 samples and 9 variables.
- Simulated Friedman having 1200 samples and 3 variables.
- O3 DMEF data having 7926 samples and 55 variables.

In order to results, comparison chart is given for each data set through error factor.

UCI Data	**Error by using Simple Bagging**	**Error by using Lasso-Bagging**
Boston Housing	20.0	11.9
Ozone data	22.6	19.1
Algae UCI data	56.4	48.7
Friedman	11.4	5.3
Simulated friedman	28.3	19.2
O3 DMEF	25.1	16.3

so it is proved that Lasso algorithm gives better accuracy in prediction.

In[3], Ensemble methods are used in pattern recognition machine learning field to improve the accuracy, performance of the algorithm. Here the comparison of popular ensemble methods is given like Bagging, Boosting and Random forest. All the performance is measured on the UCI machine learning data sets. The importance of ensemble methods are well explained by concluding the paper. Experiments on 3 UCI data sets are done individually using Bagging, Boosting and Random forest.

The meaning of Random forest is randomly selected the features of given data set and make a new sub data sets and then apply the algorithm. Whereas Bagging is different, Bagging containing sub data sets with some feature values. Here Pattern classification data set is trained first then apply 4 algorithms, logistic regression, decision tree, artificial neural network and support vector machine then applying Bagging, Boosting and Random forest then analyzing the result and then gives the final output.

In order to result for 1st data set:

Algorithms	**Single use**	**Bagging**	**Boosting**	**Random forest**
Logistic Regression	86.5	87	86.56	100
Decision Tree	84	86.3	85.2	94.3
Artificial Neural Network	83.2	85.01	83.65	93.12
Support Vector Machine	85.6	85.7	84.19	97

As per table given, Random forest gives better accuracy as compared to single, Bagging and Boosting methods. In[4], this paper the majority voting model is improved by using ensemble method and also discussed the importance of ensemble method in different areas. Here a concept is proposed, that is named M- ensemble learning. This model is used in majority voting calculations. It improves the performance of classification's known algorithms. This method introduced a new way of combining the algorithms. The model is divided into two parts,

In first part odd numbers algorithms (for example: Naive Bayes, Perceptron method and Decision tree) are used. In second part even number algorithms are used. In order to result first part gives 83.13 percentage accuracy, second part gives 81.86 percentage accuracy which is better than the accuracy of 80.67 percentage by using Multilayer perceptron method.
For better understanding of result, take a UCI data set "Urban":
The results are given below in table:

Algorithm Used	**Accuracy In Percentage**
Naive Bayes	77.91
Multilayer Perceptron	76.92
Decision Tree	67.65
K-NN	70.02
Based Model	76.92
M-Ensemble	82.05

This shows M- Ensemble model gives better accuracy.

In[5], this paper comparison of different Ensemble methods is discussed. In this paper individual Ensemble methods' result is calculated. Here a new model is introduced that is Ensembles of Ensembles that means combination of Ensemble methods. Here basically 4 models, results are compared, that are Bagging, Boosting, Stacking and Random forest. By combining all 4 models calculated the time taken in training and other complexities. All performance is done on the UCI data sets. Here the meaning of Bagging, Boosting, Stacking and Random forest are explained and then calculating the results by taking UCI data sets. Here 31 data sets are used. For better understanding the performance taking Pima- Diabetes data set.

Model	**Accuracy in percentage**
Boosting	74.3
Bagging	74.6
Stacking	65.1
Random forest	73.83

For Pima-Diabetes data set Bagging gives better results

V. FUTURE OF ENSEMBLE METHOD

One fascinating inquiry we intend to examine is the manner by which successful a solitary classifier approach may be in the event that it was permitted to utilize the time it takes the Ensemble Methods to prepare various classifiers to investigate its idea space. For instance, a neural system approach could perform pilot studies utilizing the preparation set to choose proper estimation of parameters, for example, shrouded units, learning rate, and so on. Boosting techniques are incredibly effective in numerous spaces, we intend to examine novel methodologies that will hold the advantages of Boosting. The objective will be to make a student where you can basically push a begin catch and let it run. To do this we would attempt to protect the advantages of Boosting while at the same time anticipating overfitting on uproarious informational indexes. Ensemble methods are the combination of independent classifiers. The goal of this technology is to achieve a better result. Now a days Ensemble methods are used in most of the machine learning fields for example: Ensemble methods are used to predict diabetes patient, predict the student academic performance, cancer patient, pilot's behavior, sentiment analysis etc. Combination of algorithms gives more accurate result as compared to single use of algorithm. It is beneficial to study of Ensemble methods because this field has a bright future aspect in every field like robotics, pattern recognition etc

VI. CONCLUSION

In spite of the fact that Ensemble methods can enable you to win AI rivalries by contriving complex calculations and creating results with high precision, it is regularly not favored in the ventures where interpretability is increasingly significant. In any case, the viability of these techniques are irrefutable, and their advantages in suitable applications can be enormous. In fields, for example, human services, even the littlest measure of progress in the exactness of AI calculations can be something really profitable. Ensemble Methods have been fruitful in setting record execution on testing datasets and are among the top champs of Kaggle information science rivalries. Ensemble methods have a better future aspect in each prediction field. We see in many cases Ensemble methods gives more accurate results than the single use of machine learning models but in some cases we see that Neural Network gives same accuracy. So It is dependent on the data sets that which method is much better.

REFERENCES

[1]. Alejandro Pena-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Expert Systems with Applications, Pp.1432-1462,(2014).
[2]. C.Romeroand S.Ventura. Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications,Vol.33,(2007),135-146.
[3]. Han, J., Kamber, M., and Pei, J. (2013). Data Mining Concepts and Techniques, third edition. Morgan Kaufmann.
[4]. Hashmia Hamsa, Simi Indiradevi, Jubilant J.Kizhakkethottam. Student academic performance Prediction Model Using Decision Tree and Fuzzy genetic algorithm, Procedia Technology, Vol. 25 (2016), 326 332.
[5]. Evandro B. Costa *, Baldoino Fonseca, Marcelo Almeida Santana, Fabrsia Ferreira de Arajo, Joilson Rego., Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, Computers in Human Behavior 73 (2017) 247-256.
[6]. ElGamal,A.F.(2013).Aneducationaldataminingmodelforpredicting student performance in the programming course. International Journal of Computer Applications, 70(17). 4th Int'l Conf. on Recent Advances in Information Technology | RAIT-2018 |.
[7]. Romero, C., Lopez, M. I., Luna, J. M.,& Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 68, 458e472.
[8]. Amirah Mohamed Shahiria,, Wahidah Husaina, Nuraini Abdul Rashida. "A Review on Predicting Students Performance using Data Mining Techniques", The Third Information Systems International Conference, Procedia Computer Science, Vol. 72 (2015), 414 – 422.
[9]. Shovon, M., and Haque, M., An approach of Improving Student Academic Performance by using K-means clustering Algorithm and Decision tree, 2012, 3:8 .
[10]. C.Romero and S.Ventura, "Educational Data Mining: a review of the state art, Systems, Man, and Cybernetics", Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, (2010).pp,601-618.
[11]. Juhanak, L., et al., Using process mining to analyze students' quiz- taking behavior patterns in a learning management system, Computers in Human Behavior (2017), https://doi.org/10.1016/j.chb.2017.12.015

Barkha Bhardwaj, et. al. "Theoretical Framework of Ensemble Methods and Applications." *IOSR Journal of Engineering (IOSRJEN)*, 10(5), 2020, pp. 58-63.