IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

# Enhancement of Data Classification Accuracy using Bagging Technique in Random Forest

**Vikas S.[1*], Thimmaraju S.N.[2]**

[1,2]VTU PG Centre, Mysuru, Karnataka, India

*Corresponding Author: vikas.smg@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i8.185188 | Available online at: www.ijcseonline.org

Accepted: 14/Aug/2019, Published: 31/Aug/2019

***Abstract-*** Random forest are able to do classification on high performance through a classification ensemble with a decision trees that grow mistreatment at random elect subspaces of information. The performance of associate degree ensemble learner is very obsessed on the accuracy of every element learner and also the diversity among these parts. In random forest, organisation would cause incidence of unhealthy trees and should embrace related trees. This ends up in inappropriate and poor ensemble classification call. During this paper a shot has been created to enhance the performance of the model by applying material technique in a very random forest. Experimental results have shown that, the random forest are often more increased in terms of the classification accuracy.

***Keywords-*** Random forest, Classification Accuracy, Bagging.

## I. INTRODUCTION

Random forest (RF) meth0dology may be a machine learning technique helpful for prediction issues. The RF algorithmic rule, developed by Leo Breiman [1], applies bootstrap aggregation (bagging) [2] and random feature choice [3, 4] to individual classification or regression trees for prediction. There are several studies showing that RFs have spectacular prophetic performance in regression and classification issues in varied fields, together with monetary statement, remote sensing, and genetic and medicine analysis [5 -12]. Random Forest classifiers [1] attract increasing attention inside the pc vision community. Variants like Random Ferns [08] and intensely randomised trees [12] are renowned. The analysis add the world of random forest aims at either up accuracy, or reducing time needed for learning and classification or each.

This analysis work aims to enhance the accuracy of random forest. Random forest is currently notable to be one in all the foremost economical classification ways [16 -18]. However, attributable to the complexness of information distribution in high dimensional future area, a random forest may embrace dangerous tree classifiers which might lead to wrong classification results. The vote of all the trees to create associate degree ensemble classification call, it'll create a wrong call once there are an oversized proportion of dangerous trees enclosed in random forest. to create optimisation within the random forest deduct and exclude dangerous trees thus on cut back their negative effects on the performance of the random forest [19].

Random forest has, organisation would cause incidence of the related decision trees which can have an effect on the performance of a random forest. By minimizing the correlation among these trees, the classification accuracy of the random forest are often improved. This paper aims to optimize, sizable amount of call trees in an exceedingly random forest through the choice of solely unrelated and smart trees with high classification accuracies. Decision trees include the voracious of best split point from dataset at each stage. Here we eliminate the bad trees and selecting only the good tress and bagging them to get high classification accuracy. Decision trees helpless to high difference in the event that they are not trimmed. This high fluctuation can be bridled and decreased by using various trees with different examples of the training dataset and combining their expectations is a technique called bagging.

Numerous trials are performed on medical datasets utilizing varied classifiers and future selection ways. a good live of analysis on breast malignant growth is found in writing. a substantial ton of them show nice classification accuracy. Angeline and Dr. Sivaprakasam demonstrates the foremost precise one with the accuracy of ninety six.99% mistreatment the C4.5, Naïve Bayes support vector machine (SVM) and K-Nearest Neighbor. Guo and Nandi projected a Multilayer Perceptron (MLP) with unfold of blunder calculation and purchased ninety six.21% accuracy. within the year 2017 IEEE Region ten Humanitarian Technology Conference (R10-HTC) twenty one – twenty three Dec 2017, Dhaka, bangladesh a team shown ninety eight.06% accuracy with three hundred feature maps and tenfold cross validation.

 

Fernndez-Delgado tried "one hundred and seventy nine" machine learning algorithms "one hundred and twenty one" UCI datasets and the outcomes were astounding. Random forest was positioned first mirroring the points of interest ensemble methods in machine learning. with conclusion to medicinal field an enormous number of issues can be credited to grouping, including explicit cases, frequently reflecting countless case qualities. Each of the eigenvalues in general for conducting and testing is not really ready to get accurate outcomes. For extending the accuracy of the model ensemble technique is to be utilised, ideally ensemble technique supported accord the model utilised is boosted classification tree and random forest, call tree. a method like neural system used for cluster due to that potency of a model is to be extended.

In our paper we have got 99.40% accuracy on the classification on multiple datasets using bagging technique. Ensemble strategy dependent on agreement the model utilized is boosted classification tree, decision tree. A technique such as neural network is used for ensemble because of which efficiency of model to be explained. A classification dataset is used that contains discrete qualities. We have taken sample dataset of breast cancer and fruits datasets.

## II. METHODS AND DATASETS

In our paper we have proposed classification using bagging technique with breast cancer datasets and fruits datasets. It is designed supported the mix of call trees r=I' wherever x is that the input vector and eight k denoted to random split freelance vector with equal distribution of trees within the forest, 81, ... ,8 k-l . Meanwhile, T is that the ensemble bootstrap sample drawn from coaching information. every tree is made on a distinct bootstrap sample, consisting of N samples drawn indiscriminately, and with replacement from the N samples of the coaching set. At every node variety m of the entire number of predictors M is chosen indiscriminately. within the method of model coaching, every call tree is made on a bootstrap sample of the coaching information employing a willy-nilly hand-picked set of variables[7]. Therefore, some samples that don't seem to be utilized in the coaching method are referred to as the "Out-of-Bag Samples (OOB)".

1. At node n, at random sample m of the M predictor variables.
2. for every of the m sampled variables Vk, whereby k = 1…., m fmd the simplest split Sk among all attainable splits.
3. Then, choose the simplest split s* among the k = one, ..., m splits Sk so as to separate the node. This variable Vbes! is known on that cut purpose c* is employed to separate the node.
4. Split all the info entries i = one..., n that's gift within the parent node, by causation the observations with Vibes! < c* to the left descendant node and every one observations vibest >c* to the correct descendant node.
5. Repeat steps 1- four on all descendant nodes to grow a maximally sized tree Tb•.

***Bagging technique***

It is a straight forward and extremely ground-breaking gathering strategy. A group strategy is a method that consolidates the expectations from numerous AI calculations together to make more exact forecasts than any individual model. Bootstrap Aggregation is a general method that can be utilized to lessen the fluctuation for those of calculation that have high change. A calculation that has high difference are choice trees, similar to characterization and relapse trees. Choice trees are touchy to the particular information on which they are prepared On the off chance that the preparation information is changed (for example a tree is prepared on a subset of the preparation information) the subsequent decision tree can be very unique and thus the forecasts can be very extraordinary. Stowing is the utilization of the Bootstrap methodology to a high-fluctuation AI calculation, regularly decision trees. How about we expect we have an example dataset of 1000 cases (x) and we are utilizing the characterization and relapse trees CART calculation. Stowing of the CART calculation would fill in as pursues.[07]

Make many (for example 100) irregular sub-tests of our dataset with substitution. The main parameters when sacking choice trees is the quantity of tests and thus the quantity of trees to incorporate. This can be picked by expanding the quantity of trees on pursue keep running until the exactness starts to quit appearing (for example on a cross approval test outfit). Large quantities of models may set aside a long effort to get ready, however won't overfit the preparation information. Much the same as the choice trees themselves, Bagging can be utilized for arrangement and relapse issues.

**Classification Functions**

Support vector machine[10], It is a machine learning language that is utilized for two class assignments and has shown sublime execution. We give our consideration to parallel characterization wherein subjects are named that are having a place with one of two classes. In the preparation information as n properties and a moment classification banner, we can think about the information in n-dimensional space.

**Decision Tree**

It is a structure where each inner hub speak to a judgment of a characteristic each branch speaks to the yield of a judgment result lastly each leaf speaks to an order result, it need regulated figuring out how to assemble a choice tree the part of the tree is set up dependent on the various estimations of

 

record field and over and over structure lower hubs and branches in each branch subset. The key of structure choice tree lies in the decision of various qualities for the record field when setting up a branch.[06]

**AdaBoost Classifier**

Is an iterative calculation and its center thought is to prepare various classifiers for a similar preparing informational collection and after that gathering these feeble classifiers to shape a solid classifier. the calculation itself is executed by changing information circulation. it decides the heaviness of each example dependent on whether the arrangement of each example in each preparation set is right and exactness of the last generally speaking characterization. the new informational index with altered weight is sent to the lower classifier for preparing. at last it combined the classifier in each preparation as an official conclusion classifier.

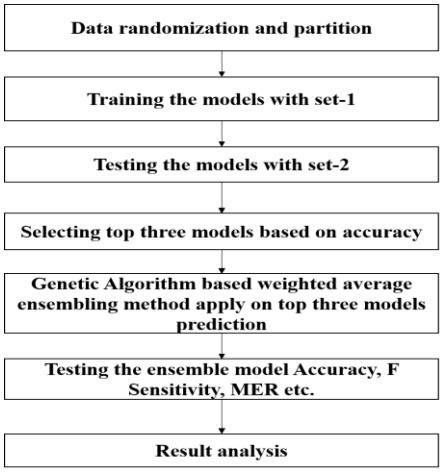


Fig: 1 Data flow

*(MER = Minimum error rate)*

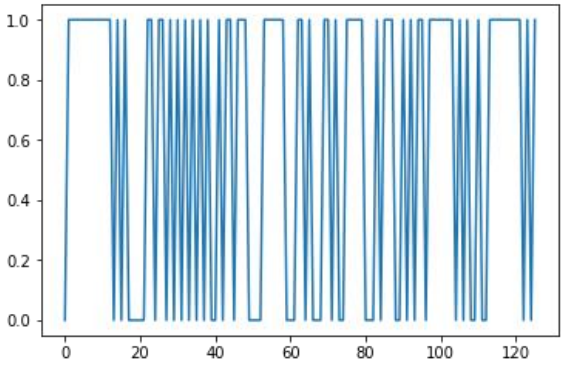


Fig: 2 Breast cancer dataset samples

*Picture Segmentation:* means discovering area of enthusiasm from the image. K-implies is the bunching calculation found by MacQueen in 1967. K-implies is for the most part conveyed to locate the tainted areas in a picture.

*Highlight Extraction*: Features are removed from the organic product. In light of the highlights of the natural product the characterization of organic product infections is finished. *Concentrate different* - highlights from natural product like shade of organic product, surface and shape.

In the proposed methodology some shading and surface highlights are utilized. Color Features. The shading highlights used as a piece of the leafy foods arrangement natural product illnesses recognizable proof are Global Color Histogram, Color Coherence Vector. Global Color Histogram: It is the basic technique to interpret the data accessible in a picture. For each of the different shading GCH is a gathering of qualities which demonstrates the likelihood. For declines the diverse shading and scaling inclination consistent standardization and quantization are used. It is the Simple methodology for independent shading.

*Color Coherence Vector*: Some sizable bordering space has sound pixels, whereas scattered pixels aren't. The shading area and also the varieties between the neighboring pixels are dispense with to establish ccvs. afterward the order of that pixels either sound or indiscernible by the associated elements within the image it's adept and invariant to minor changes. Texture options The surface highlights used as a bit of the foods grownup from the bottom order/natural product infections identifying proof are neighborhood twofold example, complete near paired example, and neighborhood ternary example. native Binary Pattern: it's determined by different the image pixels and its neighbor.[09]

| | fruit_label | fruit_name | fruit_subtype | mass | width | height | color_score |
|---|---|---|---|---|---|---|---|
| 0 | 1 | apple | granny_smith | 192 | 8.4 | 7.3 | 0.55 |
| 1 | 1 | apple | granny_smith | 180 | 8.0 | 6.8 | 0.59 |
| 2 | 1 | apple | granny_smith | 176 | 7.4 | 7.2 | 0.60 |
| 3 | 2 | mandarin | mandarin | 86 | 6.2 | 4.7 | 0.80 |
| 4 | 2 | mandarin | mandarin | 84 | 6.0 | 4.6 | 0.79 |

Samples of fruits Datasets

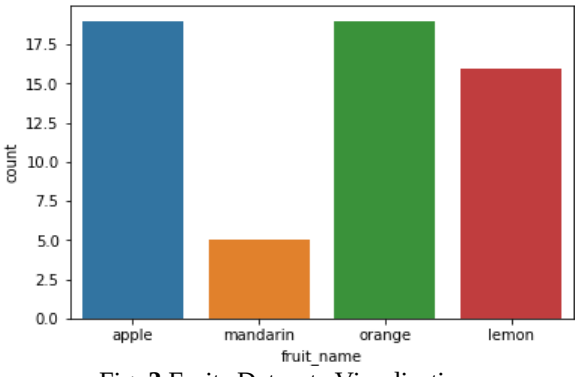


Fig: **3** Fruits Datasets Visualization

 

**Ensemble Algorithm**

Algorithm: Our proposed approach is delineated within the sort of formula a pair of. a unique section of information process and prediction are mathematically explained during this formula. The formula shows the projected model algorithm. formula a pair of GA primarily based Weighted Average Ensemble methodology

whereas S != zero do
section one : DATARANDOMIZATION AND PARTITION
S1 = random (S,frac=0.50)
S = Dataset when process
S2 = SS1
S1→ Trainingset[set-1]
S2→ Testingset[set-2]
section a pair of : coaching MODELS
Strain = random (S1, frac=0.50)
Stest = S1- Strain
/****TRAINING MODELS **** /
Trained all models Mk on an equivalent dataset
/****TESTING MODELS IN coaching **** /
Apply check cases on completely different models (Mk)
/****ERROR **** /
MER, Acc%, Spec, Sens.
Result: M(x) = completely different models prediction.
Section three: ENSEMBLE AND TESTING section Mk for k top(Mk $ intensify,3)
W1, W2, W3 = Weight optimize by GA formula.
P1, P2, P3 = Prediction chance of prime 3 models.
P=P1 W1+ P2W2+P3W3;
P = Weighted Average Ensemble methodology Result victimization GA. MER, Acc%, Spec, Sens.
end while

## III. CONSLUSION

Our paper contrasts a few machine learning strategies with anticipate high accuracy of the considerate and dangerous of bosom malignant breast cancer and fruits datasets. The informational collection figured from a digitized picture of a fine needle suction of a bosom mass. We utilized bagging strategy to improve the accuracy of models. Each method as its very own restrictions and qualities explicit to the kind of utilization. We have found high classification accuracy with help of bagging technique. Were bosom malignant breast cancer has achieved 98.4% Accuracy and for fruit Data set for the trained data set it is 100% Accuracy and for test data set is 92.31% Accuracy has been achieved.

## REFERENCES


[1] Breiman, L. 2001. Random Forests. Machine Learning, Vol. 45 Issue 1, pp. 5-32.
[2] Breiman, L. 1996. Heuristics of instability and stabilization in model selection. The Annals of Statistics, Vol.24 Issue 6, pp. 2350–2383.
[3] P. Boyle and B. Levin, "World Cancer Report,"World Health Organization, Geneva, Switzerland 2008.
[4]J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, vol. 2, pp. 59- 78, 2006
[5]G. Richards, V. J. Rayward-Smith, P. Sönksen, S. Carey, and C. Weng, "Data mining for indicators of early mortality in a database of clinical records," *Artificial intelligence in medicine*, vol. 22, no. 3, pp. 215-231, 2001.
[6] L. Breiman, "Random forests," *UC Berkeley TR567*, 1999.
[7] (2016, 2016-12-10). *Random Forest – RapidMiner Documentation*.
[8] V Adegoke, Daqing Chen, Ebad Banissi, and Safia Barikzai. Prediction of breast cancer survivability using ensemble algorithms. 2017.
[9] Abdulsalam Alarabeyyat, Mohannad Alhanahnah, et al. Breast cancer detection using k-nearest neighbor machine learning algorithm. In *Developments in eSystems Engineering (DeSE), 2016 9th International Conference on*, pages 35–39. IEEE, 2016.
[10] Dania Abed Aljawad, Ebtesam Alqahtani, AL-Kuhaili Ghaidaa, Nada Qamhan, Noof Alghamdi, Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji. Breast cancer surgery survivability prediction using Bayesian network and support vector machines. In *Informatics, Health & Technology (ICIHT), International Conference on*, pages 1–6. IEEE, 2017.
[11] Bhavini J. Samajpati, Sheshang D. Degadwala,"A Survey on apple Fruit Diseases Detection and Classification" International Journal of Computer Applications (0975 –8887) Volume 130 – No.13, November 2015.
[12] Wang, Bo, et al. "Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine." *IEEE Transactions on Smart Grid* (2016):1-1.


## AUTHORS PROFILE

**Mr.Vikas S**. received M.Phil degree in Computer Science in the year 2009 and Master of computer Applicaions (MCA) in the year 2007 from Visvesvaraya Technological University and Bachelors Degree in Computer Science in the year 2004 from kuvempu University. He is currently working as Assistant Professor in the **Department of MCA, Visvesvaraya Technological University,** PG Center, Mysore, Karnataka, where he is involved in research and teaching activities. He is having 11 years of teaching experience and 02 years of Industrial experience. He is a Life member of India Society for Technical Education (LMISTE), Computer Society of India (CSI) and Doing Research work on the Area **Big data Analytics**.

**Dr. Thimmaraju S N**, he is presently a professor and heading the **Department of Master of Computer Application, Visvesvaraya Technological University,** PG Center, Mysore, Karnataka, he has received his Ph.D degree from **Visvesvaraya Technological University(VTU)**, Belgaum in the year 2010, M.E., degree in Computer Science and Engineering from University Visvesvaraya College of Engineering (UVCE), Bangalore in 2002 and Bachelors Degree in Computer Science and Engineering from PESCE, Mandya in the year 1999. He is involved in research and teaching activities. His major areas of research are Computer Networks, WSN's and Graph theory. He is having 17 years of teaching experience. He has published around 15 research papers which include International Journals, International Conferences and Notional Conferences.

 