Modify Random Forest Algorithm Using Hybrid Feature Selection Method

Prof. Dr. Ahmed T. Sadiq, Karrar Shareef Mohsen

Dept of Compute Science, University of Technology, Baghdad, Iraq.
drahmaed_tark@yahoo.com & Karrarmuswi1@gmail.com

Abstract— The Importance of Random Forrest(RF) is one of the most powerful methods of machine learning in Decision Tree. The Proposed hybrid feature selection for Random Forest depend on two measure Information Gain and Gini Index in varying percentages based on weight. In this paper, we tend to propose a modify Random Forrest algorithm named Random Forest algorithm using hybrid feature selection that uses hybrid feature selection instead of using one feature selection. The main plan is to computation the Information Gain for all random selection feature then search for the best split point in the node that gives the best value for a hybrid equation with Gini Index. The experimental results on the dataset showed that the proposed modification is better than the classic Random Forest compared to the standard static Random Forest the hybrid feature selection Random Forrest shows significant improvement in accuracy measure.

Keywords— Random Forest algorithm, Feature selection, Ensemble of Classifiers.

I. INTRODUCTION

The most important subfield of artificial intelligence (AI) is a Machine learning. The objective of machine adapting, for the most part, is to comprehend the structure of information and fit that information into models that can be comprehended and used by individuals. Even though machine learning is a field inside computer science, it contrasts from customary computational methodologies. In traditional computing, algorithms are sets of instruction customized guidelines utilized by computers to calculate or issue explain. Machine learning calculations rather consider computers to prepare information sources of info and utilize use statistical analysis keeping in mind the end goal to yield esteems that fall within a particular range. in [1] Machine Learning has become one of the powers of IT and with that, a rather central, albeit regularly hidden, component of our development. With the ever-increasing amounts of data becoming available there is a real reason smart data analysis will become even more pervasive as a necessary part to technological progress. Along these lines, machine learning helps computers in building models from unit data keeping in mind the end goal to decision-making forms in manners based on data inputs. Machine learning examines computers calculations for figuring out how to do stuff. We may, for example, be occupied with figuring out how to finish a job, or to make exact expectations, or to behave wisely. The discovering that is being done is constantly considering a type of perceptions or information, for example, illustrations… coordinate understanding, or direction. So as a rule, machine learning is tied in with figuring out how to improve the situation later in light of what was knowledgeable about the past.

In [2] the building of predictor done with multiple versions, doing these to make a group of the predictor. Classifier or predictor giving an equal vote when predicting a class, the majority vote will take to representing predict class.

In [3] The construction of trees can be done with different partitioning mechanisms, these being more dependent on the data than in building trees (accuracy follows the same trend as the number of trees in all dividing mechanisms rises) .

In [4] Not all feature can be used to construct trees, determine which feature are most important and associate to the class or carry a lot of information. Inductive learning is used in synchronization with tree construction. Trees affect the selection process for trees that are under construction.

In [5] The algorithms Random Forest strategies type of classification that has reliance the compound of numerous predicator call trees. The characteristic of such Ensemble of Classifier (EoC) is that their tree depends on randomness to growth. Approved this thought, Random Forrest is described as a general standard of randomized aggregates

call trees. Despite the concept has earlier done utilized during the 90's.

In [6] The normal definition and the use of the expression "Random Forest ", become implied at 2001 of an exceeding introduction paper written by Leo Breiman, the description random forest may be a classifier compound of a group of tree-organized classifiers the random forest is a set of a decision tree model based on random data taken from the database. It is analysed by drawing on each tree within the forest to build the trees. After the construction of the models within the forest, to any class and then vote among tree models for final decision.

We propose modify in feature selection of the Random Forest by using hybrid feature selection for Random Forest depend on two measure Information Gain and Gini Index in varying percentages based on weight.

The rest of this paper content Related work in section 2, Random Forest Algorithm in section 3, Modified Random Forest in section 4, Experimental Result in section 5, and Conclusion in section 6.

II. Related work

In [7] the high-dimensional genetic data implementation is incapable to identify the related predictions. In the state of greater impact dimension, it is possible to apply for this extension. The researchers suggested a weighted random forest, unlike a Random Forrest that relies on all trees and all trees of the same weight. In this research, they suggested a weighted Random Forrest, an extension of Random Forrest that increased the performance of the original algorithm to detect interactions in high-dimensional genetic data.

In [8] The Dynamic Random Forest algorithm, unlike the standard Random forest algorithm, shows an enhancement in accuracy. The process of working the Random Forest is to move away from building trees that can make the level of forest performance low in terms of correctness in decision-making, by guiding the algorithm to create trees that will only handle the errors of a set of trees already built. They present in this paper that guidance in the forest is random can help from the adoption of a dynamic method.

In [9] With motivation from Random Forests with regards to classification, another bunching troupe method Cluster Forests (CF) is proposed. Geometrically, CF arbitrarily tests a high-dimensional information cloud to acquire great nearby clustering and after that totals by means of spectral bunching to get group assignments for the entire dataset. The hypothetical investigation uncovers that the kappa measure makes it conceivable to develop the neighborhood grouping in an attractive way it is commotion safe.

In [10] the researchers adjust the possibility of irregular projections connected to the yield space, to upgrade tree-based gathering strategies with regards to multi-mark characterization. They likewise demonstrate that irregular yield space projections might be utilized as a part of a request to achieve distinctive inclination change trade-offs, over a wide board of benchmark issues, and this may prompt enhanced exactness while lessening essentially.

In [11] The approach taken here depends on nonparametric scoring and ROC curve enhancement in the feeling of the AUC standard. In this issue, aggregation is utilized to expand the execution of scoring rules produced by positioning trees, as those created in [12]. ROC investigation is a well-known method for assessing the capacity of a given scoring standard to separate between two populaces. ROC bends and related execution measures such as the AUC have now happened to standard use for surveying the quality of positioning strategies in a bipartite framework.

III. Random Forest Algorithm

Randomized forests are a group of regression trees and a classification that trains on the data set as a training Dataset, which is created by random selection on the original Dataset. When constructing a tree, a set of randomized test Data as well, which does not contain any Record constraint from the training dataset, the researchers use this as a group to test trees within the forest. And to determine the accuracy of each tree by obtaining the error rate in the determination of the type of input from the test group for the tree and to obtain the error rate of the forest is by taking the error rate for all the trees within the forest. Berman (1996) showed that the rule can don't separate the data is in two parts because of the strength of the algorithm so that, according to the out of bag scale, it can be tested with the same data and given the same accuracy as laboratory experiments. In order to classify a new entry, these values are applied to all the trees within the random forest. Each tree is predicted the

class of this entry Record and after that, the decision is taken by the forest by the majority in the vote by the trees.

Random Forests are one of the most powerful algorithms, and they follow specific methods in building and assembling trees. Even if there are outliers in large dimensional databases, Random Forrest are more powerful than other algorithms in the field of machine learning. To determine the importance of the attribute in Random Forrest, follow the method of random selection and then use the Impurity measure to determine the degree of association of the attribute in the class, and by means of the Impurity measure information Gini Index, the property is evaluated. When using an information Gini Index scale, the lowest values from the scale results should be selected on the attribute to obtain the best binary split in the node. The main stages of Random Forrest algorithm as below

Algorithm: Random forest
Input: (train: Training Dataset, test: Testing Dataset, number of trees in forest: n trees, number of selected feature: n features)
Output: (tree: learned tree, accuracy)

Step 1: for I from 1 to n_trees
Step 2: get random features from train number of features equal n features
Step 3: get sample train put sample in t_ sample the size of sample equal size sample
Step 4: build tree by using t_ sample and features that chose random
Step 5: append the tree to the forest
Step 6: end for
Step 7: for each tree in forest
Step 8: get the vote of tree for test date
Step 9: end for
Step 10: Get most frequent class in voting for each class in Forest
Step 11: end

Algorithm: Decision Tree
Input: (train: Training Dataset, n_features: number of selected feature, depth: the depth of tree level
Output: Tree
Step 1: for index from 1 to the number of n_features
Step 2: for each row in train
Step 3: split train dataset to groups based on value of row(index)
Step 4: calculus Gini Index for groups
Step 5: if Gini Index less than old value of Gini
Step 6: p_index and p_groups = index and groups
Step 7: end if
Step 8: end for
Step 9: end for
Step 10: if don't reach depth of tree to zero then
Step 11: call Decision Tree with new value parameter for left side and right side
Step 12: else
return root equal p_groups
Step 14: end if
Step 15: end

IV. Modified Random Forest

The selection of features in the Random Forest algorithm plays an important role in determining the efficiency and performance of the algorithm. The traditional methods used in the selection of features give uneven results and are often good. This gives the Random Forest algorithm strength and superiority over other classification algorithms. But that is noticeable the features may be important in a particular method and the same features are not important in another method. This is what we have found in the two methods (Information Gain, Gini Index). These are good methods of selection, but there are important features that appear in one of them (filtered and may not be filtered in the other) We suggested that we present an important equation between the two methods.

$$Value(Fi) = \left(w * info\ gain(Fi)\right) + \left((1 - w) * \left(1 - Gini(Fi)\right)\right) \quad (1)$$

Thus, the Random Forest algorithm developed as follows
Propose algorithm

Algorithm Decision Tree
Input: (train: Training Dataset, n_features: number of selected feature, depth: the depth of tree level)
Output: Tree

Begin
build tree (train, n_features, depth)
for index=1 to n_features:
gain val = gain (dataset, index)
for each row in train:
groups = split train Dataset on row [index] value

```
Gini = Gini index (groups)
Value(Fi)= (w * info gain(Fi)) + ((1 - w) * (1 – Gini(Fi)))
if Gini < old Gini then p_index, p_groups = index, groups
end for
end for
if (depth not equal zero) then
build tree (left p_groups, n_features)
build tree (right p_groups, n_features)
else
return {root= (left p_groups, right p_groups )
        end if
End
Algorithm Modified Random Forest
Input: (train: Training Dataset, test: Testing Dataset, n
trees: number of trees in forest
        number of selected feature n features)
Output: (tree: learned tree, accuracy)
Begin

for I =0 to n_trees
features =Random (train, n features)
   train = sub Train (Train, size sample)
                tree = build  tree (train, feature)
trees. append(tree)
  end for
 for I=0 to n trees
correct = get vote (test, trees(I)))
 end for
 Accurse = correct divide on number of class in test data
  End
```

The importance of Feature selection of a Random forest plays important role because it selects the attribute, the current methods can't reflect the nature of the real dataset.

This depends on the tow of the methods of Feature selection. To explore the power of attribute.

V. Experimental Result

To evaluate the performance of Hybrid Features Selection Random Forest in a real dataset, we implemented both the regular Random Forest that uses the method Gini Index measures and regular Random Forest that uses the method information Gain measures along with Hybrid Features Selection Random Forest on real dataset available on Database of UC Irvine Machine Learning Repository [13] the first dataset is Connectionist Bench (Sonar, Mines vs. Rocks) Data Set The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock see Table(1).

The second dataset is SPECTF Heart Data Set Data on cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient classified into two categories: normal and abnormal see Table (2).

The third dataset is Phishing Websites Data Set This dataset collected mainly from: Phish Tank archive, Miller Smiles archive, Google searching operators see Table (3).

The forth dataset is Breast Cancer Wisconsin (Original) Data Set Original Wisconsin Breast Cancer Database see Table (4).

The fifth dataset is Ionosphere Data Set Classification of radar returns from the ionosphere see Table (5).

According to the results in Figure (1), it is clear that the proposed equation is better than the measurements found in the standard or the Classics of the Random Forest

Table (1): Connectionist Bench (Sonar, Mines vs. Rocks) Data Set Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	208	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	60	**Date Donated**	N/A
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	101397

Table (2): SPECTF Heart Data Set Description

Data Set Characteristics:	Multivariate	**Number of Instances:**	267	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	44	**Date Donated**	2001-10-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	65255

Table (3): Phishing Websites Data Set Description

Data Set Characteristics:	N/A	**Number of Instances:**	2456	**Area:**	Computer Security
Attribute Characteristics:	Integer	**Number of Attributes:**	30	**Date Donated**	2015-03-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	67649

Table (4): Breast Cancer Wisconsin (Original) Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life

Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	343906

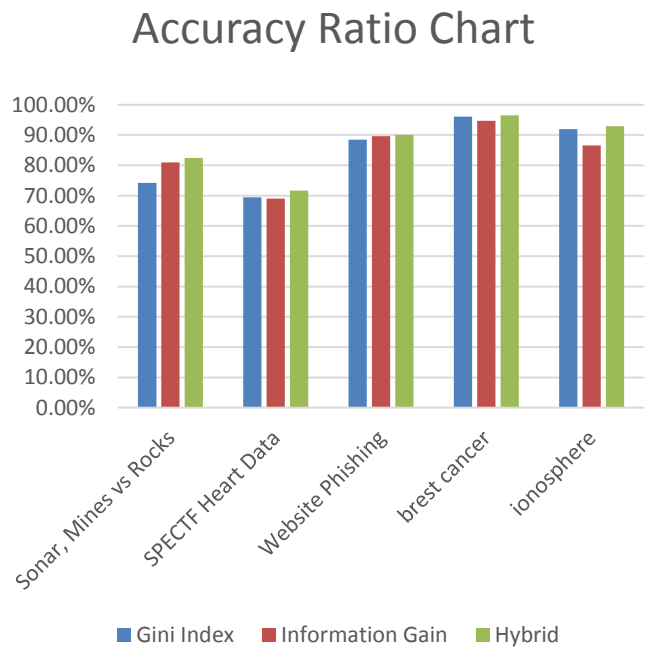
Table (5): Ionosphere Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	351	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	34	**Date Donated**	1989-01-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	149905

Table (6): Accuracy Ratio Using Three Feature Selection Algorithms for Three Dataset (where W=0.7)

	Gini Index	**Information Gain**	**Hybrid**
Sonar, Mines vs Rocks	74.146%	80.976%	82.439%
SPECTF Heart Data	69.434%	69.057%	71.698%
Website Phishing	88.444%	89.630%	90.074%
Brest cancer	96.106%	94.690%	96.460%
ionosphere	92.000%	86.571%	92.857%

Figure (1) Accuracy Ratio Chart



VI. Conclusions

A hybrid method was proposed for the Random Forest algorithm, which is important algorithm for machine learning strategies. The proposed model provided a good diversity in the selection of attributes based on two measures instead of one (Information Gain, Gini Index), which caused the selection of attributes that did not exist in them if used one. In all experiments on different dataset the result of the proposed method was better than the two measures mentioned. The best value for the weight mentioned in Equation (1) is (0.7). Experiments on different values (0.1 < Wight < 0.9) and on the other hand were increased at the time less than the total time of the measurements. The equation was proposed which is useful in features selection of the Random Forest. The equation was based on the measurements of the Information Gain and Gini Index to give the opportunity to explore the power of the attribute to the class in the Dataset. According to the apparent results, it is clear that the proposed equation is better than the measurements found in the standard or the Classics of the Random Forest.

References

[1] Alex Smola, S.V.N. Vishwanathan. "Introduction to Machine Learning". Book, (2008)، the press syndicate of the university of Cambridge , Cambridge .

[2] Leo Breiman, "Bagging predictors"، Machine Learning 24 ,2, (1996) 123–140.

[3] T. Ho, "The random subspace method for constructing decision forests", IEEE Transactions on Pattern Analysis and Machine Intelligence 20 ,8, (1998) 832–844.

[4] Y. Amit, D. Geman, "Shape quantization and recognition with randomized Trees", Neural Computation 9, 7 (1996) 1545–1588.

[5] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization", Machine Learning 40,2, (1998) 139–157.

[6] Leo Breiman, "Random forests", Machine Learning, Volume 45, Issue 1, (2001), pp 5–32.

[7] Stacey J Winham1, Robert R Freimuth1, and Joanna M Biernacka1,2, "A Weighted Random Forests Approach to Improve Predictive Performance"، Stat Anal Data Min. , 6, 6, (2013) 1-17.

[8] Simon Bernard, Sébastien Adam, Laurent Heutte. "Dynamic Random Forests". Pattern Recognition Letters, Elsevier, 33 12, (2012)، pp.1580-1586.

[9] Yan D, Chen A, Jordan MI," Cluster forests", Computational Statistics and Data Analysis, 66, (2013), 178–192.

[10] Arnaud Joly, Pierre Geurts, Louis Wehenkel "Random forests with random projections of the output space for high dimensional multi-label classification", Machine Learning and Knowledge Discovery in Databases, 3, (2014), pp 607-622.

[11] Clémençon S, Depecker M, Vayatis N "Ranking forests", The Journal of Machine Learning Research, 14, 1, (2013), 39-73.

[12] Cl´emenc¸on, M. Depecker, and N. Vayatis. "Bagging ranking trees". Proceedings of ICMLA'09,(2009), pages 658–663,

[13] Dua, D. and Karra Taniskidou, E. (2017)." UCI Machine Learning Repository "[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.