Ensembling Incremental Decision Trees

Özge Nur Ergün
Computer Engineering
Yildiz Technical University
Istanbul, Turkey
0000-0002-9997-0853

Mehmet Fatih Amasyalı
Computer Engineering
Yildiz Technical University
Istanbul, Turkey
amasyali@yildiz.edu.tr

***Abstract*—Decision tree is a method to create rules for given data to predict unseen instances. Previous research state that ensembling trees gives more accurate results than a single decision tree. In this research, standard decision tree implementation is modified so that train data is processed iteratively to grow the tree. 10 different trees are created by shuffling the train data and their results are averaged. Obtained results are compared to other ensemble tree methods such as random forest, bagging, boosting and random subspace.**

Keywords—Ensemble Learning, Online Decision Tree, Incremental Decision Tree

I. INTRODUCTION

Decision tree is a well known machine learning method. It takes all the train data to create rules for predicting results. These rules are created by checking which information is more valuable to divide data into subsets.

Creating nodes by checking these rules may cause overfitting. To avoid overfitting, some ensemble methods are developed. These methods aim to train multiple trees and make a final prediction by checking all of the predictions. Each of these multiple trees are called base learners. Base learners should be accurate and different than each other. In this way, one wrong prediction of a base learner can be tolerated by others.

Random forest is one of the mostly used ensemble tree method. Each base learner of a random forest uses a different subset of dataset features and a subset of dataset samples. Results of each base learner are averaged for regression problems, and mostly voted class is chosen for classification problems.

Bagging is another ensemble tree method. For creating different trees, random samples from dataset are chosen. Like random forest, results of base learners are evaluated equally to make a final prediction.

Boosting is also a mostly used ensemble tree method. While creating base learners, each base learner focuses on the mistakes of the previous base learner. Probability of choosing wrong predicted samples are increased but sample choices are still made randomly. Unlike random forest and bagging, predictions of more accurate base learners effect final prediction more than other base learners.

Random subspace is also an ensemble tree method Each base learner uses a different subset of dataset features. Results of each base learner are evaluated equally to make final prediction.

II. METHODOLOGY

Decision tree is modified to take all of the dataset iteratively. At first, given sample is used to create a leaf and initiate a tree.

After creating that leaf, for each sample at dataset, tree checks if the sample is classified correctly. If yes, then nothing is applied and next instance is fed to the tree. If no, the leaf that current instance is placed is changed into a node and the created node is splitted according to the information gain of each attribute.

For building an ensemble, datased is shuffled before creating a new base learner. Each base learner is fed with same data but in different order. So, each base learner has a different rule set according to the order of given samples.

10 base learners are created and their predictions are counted. Most voted label is returned as the output of the ensemble. Ensemble performance is measured by using 36 UCI classification datasets. 25% of each dataset is splitted for testing. For making final prediction, predictions of base learners are counted and mostly voted class is assigned as the output of the ensemble. Flow chart of the work is illustrated in Figure 1. Source code can be found at github [1].

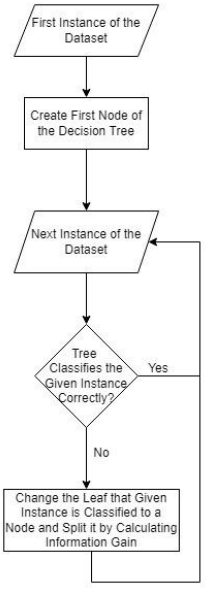

Fig. 1. Flowchart of the Study

III. RESULTS AND DISCUSSION

Performance of the ensemble of online decision trees is compared to decision tree, random forest, bagging, gradient boosting and random subspace at 36 UCI classification datasets. Obtained accuracy results are presented in Table 1.

TABLE I. ACCURACIES OF THE MODELS

Dataset	Decision Tree	Online Tree	Online Tree Ensemble	Random Forest	Bagging	Gradient Boosting	Random Subspace
ringnorm	89.40	58.77	75.43	**95.19**	93.62	92.48	90.54
mushroom	**100.00**	67.02	88.41	**100.00**	**100.00**	99.75	99.51
hepatitis	64.10	65.56	67.22	**79.48**	76.92	71.80	76.92
letter	87.28	48.02	79.70	**93.32**	92.22	73.84	75.40
audiology	83.72	72.84	77.99	88.37	**93.02**	**93.02**	88.37
balance-scale	75.79	68.74	72.83	78.98	78.98	**83.43**	82.80
autos	76.47	56.29	60.64	**78.43**	72.55	66.66	72.54
iris	**100.00**	66.67	68.13	**100.00**	**100.00**	**100.00**	**100.00**
breast-cancer	59.72	63.88	69.44	75.00	69.00	**77.77**	73.61
diabetes	71.35	63.59	67.46	69.79	72.36	**75.00**	**75.00**
colic	75.00	62.43	70.98	84.78	82.61	**86.96**	**86.96**
credit-a	75.72	52.38	54.83	**85.55**	84.93	84.97	83.24
anneal	**100.00**	98.78	99.29	**100.00**	99.55	99.10	**100.00**
breast-w	93.71	92.07	92.83	**96.57**	95.43	94.29	94.86
labor	66.66	72.27	77.51	**93.33**	80.00	66.66	73.33
credit-g	69.60	63.59	70.48	**77.60**	73.60	73.60	73.20
abalone	20.21	17.43	21.13	24.06	24.45	**25.12**	24.83
hypothyroid	**99.79**	54.47	89.67	98.51	99.68	99.68	99.36
glass	65.38	44.98	47.95	**78.85**	71.15	67.31	75.00
kr-vs-kp	**99.12**	77.31	84.43	97.87	98.99	92.99	92.99
ionosphere	86.36	74.17	80.54	**94.31**	**94.31**	92.00	92.04
heart-statlog	75.00	60.04	62.67	**80.88**	73.53	76.47	**80.88**
lymph	86.11	65.63	72.75	**94.44**	86.11	86.11	88.88
primary-tumor	34.21	35.66	40.78	44.74	39.47	**48.68**	47.34
splice	93.23	78.21	84.70	93.23	**95.99**	94.48	94.61
vowel	77.01	27.17	28.43	**86.69**	85.08	69.76	77.82
waveform	72.88	58.72	70.48	81.20	82.08	82.72	**83.52**
segment	94.63	30.82	39.02	96.19	**96.36**	93.77	94.81
zoo	**100.00**	94.40	98.69	**100.00**	**100.00**	**100.00**	**100.00**
vehicle	69.34	31.48	49.73	73.58	71.22	73.11	**74.53**
sick	98.51	85.27	93.95	97.98	**99.04**	98.19	97.56
vote	92.66	68.81	81.65	92.66	**96.33**	**96.33**	**96.33**
sonar	71.15	50.29	52.60	**80.77**	76.92	76.92	80.76
soybean	92.89	35.14	69.70	94.67	**95.86**	94.08	93.49
d159	98.44	49.80	67.73	99.05	**99.44**	93.37	94.93
col10	70.49	36.47	55.92	**74.25**	73.46	72.67	71.28

To understand model performances, the count of each model to be most successful model compared to other models at datasets is presented in Table 2 and illustrated in Figure 2.

TABLE II. COUNT OF BEING THE BEST MODEL FOR EACH CLASSIFIER

Model	Count of Being the Best Model
Decision Tree	6
Online Tree	0
Online Tree Based Ensemble	0
Random Forest	19
Bagging	11
Boosting	10
Random Subspace	9

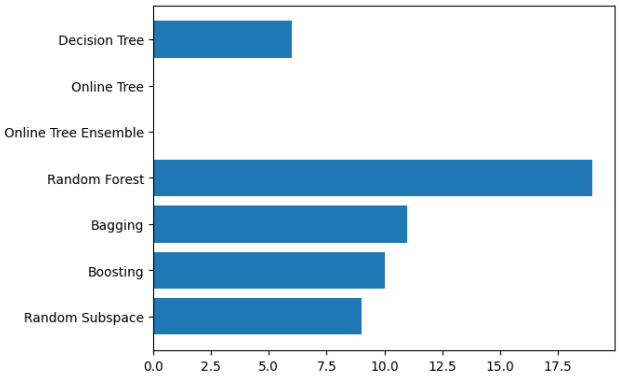

Fig. 2. COUNT OF BEING THE BEST MODEL FOR EACH CLASSIFIER

Average performance rank of each model is presented in Table 3 and illustrated in Figure 3.

TABLE III. AVERAGE RANKS OF THE MODELS

Model	Average Ranks of the Models
Random Forest	69
Bagging	80
Random Subspace	91
Gradient Boosting	99
Decision Tree	134
Online Tree Based Ensemble	172
Online Tree	216

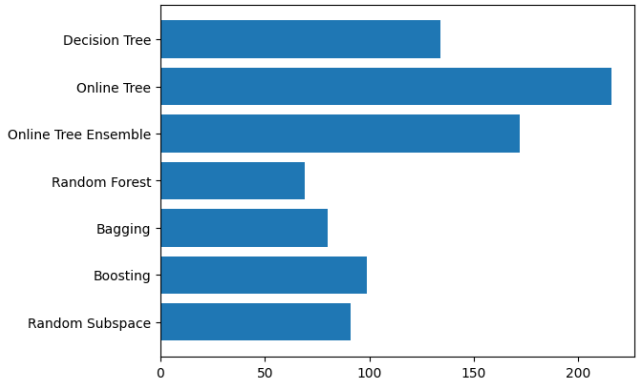

Fig. 3. AVERAGE RANKS OF THE MODELS

IV. CONCLUSIONS

In this study, standard decision tree is modified in an online way to create rules by considering instances iteratively. An ensemble of online decision tree is created to make more accurate predictions. Obtained results from single online tree and ensemble of online trees are compared to decision tree, random forest, bagging, gradient boosting and random subspace methods.

Since online trees are not fed with all of the dataset at the beginning of training, first nodes of these trees does not generalize data and does not create meaningful rules. They also aim to fit the tree all of the dataset to make correct predictions, this causes overfitting. Test results show that online trees grow much bigger than standard decision trees to fit the train data. Outliers cause the tree become less accurate. If test data is similar to train data, online trees may perform good; if not, they would not be able to predict test labels.

Obtained results show that online tree base learners does not generalize the data well and does not achieve good accuracy results. Base learner predictions may be different, however the performance of the base learners are low. Ensembling them gives better results but they are not as accurate as standard ensembling methods.

Changing data order while training online trees cause big differences of the tree performance. If first samples are not outlier, tree may be accurate. Also there it is possible to create very different online trees by shuffling the dataset. This can be an advatage of the online trees and shuffling the dataset may give good results. Another advantage of online trees is, if new data is added to original dataset, there is no need to rebuild the tree, existing tree can be updated.

REFERENCES

[1] https://github.com/ozgeergun/Ensemble-of-Online-Decision-Trees