Random Forests

Subham Roy
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260

1 Overview

An ensemble method is an approach that combines many simple "building block" models in order to obtain a single and potentially very powerful model. These simple building block models are sometimes known as "weak learners" since they may lead to mediocre predictions on their own. Random forests are one type of ensemble methods for which the simple building block is either regression or a classification tree.

2 About Random Forests

Ensemble learning methods are made up of a set of classifiers like decision trees—and their predictions are aggregated to identify the most popular result. The most well-known ensemble methods are bagging, also known as bootstrap aggregation, and boosting. In the bagging method, a random sample of data in a training set is selected with replacement. After several data samples are generated, these models are then trained independently, and depending on the type of task— say for classification, the majority of those predictions yield a more accurate estimate. This approach is commonly used to reduce variance. The predictions in this case from the bagged trees will be highly correlated. Unfortunately, averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities. In particular, this means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. Random forests overcome this problem by forcing each split to consider only a subset of the predictors. This is a key difference among decision trees, bagging and random forest. Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled which is generally $(p - m)/p$ (Here, "p" is the set of all predictors and "m" is a subset of predictors). From there, the random forest classifier can be used to solve regression or classification problems. The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (OOB) sample. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Finally, the OOB sample is then used for cross-validation and finalizing that prediction. For instance, if a random forest is built using m = p, then this amounts simply to bagging.

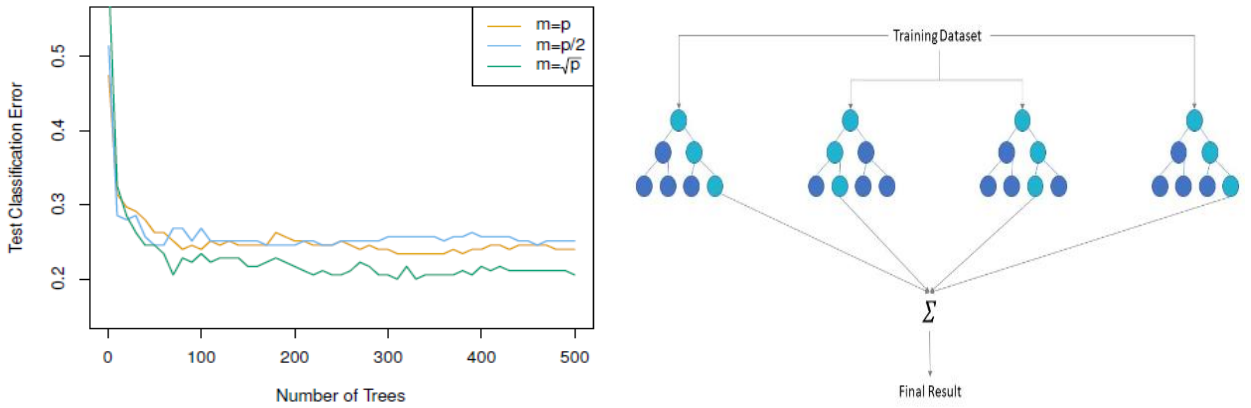


Fig: (Left) In this plot we see how increasing the number of trees reduces the Test error. Also, comparison between bagging(m=p), and random forest(m=p/2 and m=√p). (Right) The figure shows how Random Forest works with multiple decision trees and produces the final prediction.

Key benefits of Random forests are it reduces the risk of overfitting, provides flexibility as it can handle both regression and classification, and it helps to determine important features among all the predictors. This method has its challenges as well, some of which are that it is a time-consuming process, it requires more resources and it's complicated.

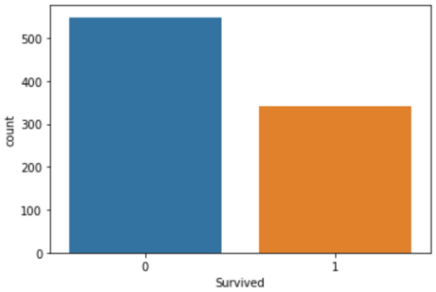
3 Experiment

The dataset used here is the Titanic dataset. In this problem we predict if the passenger of the ship survived or not. Therefore, the 'survived' column would be our response column. The dataset has 891 rows and 12 columns. Out of these 12 columns, 'survived' column is the response and other features are the independent variables. This dataset has null values in 3 of the independent columns (Age, Cabin, Embarked).
We see the columns Age, Cabin, and Embarked have less non-null values compared to others. Furthermore, we create the dummy variables for useful string type column as our model cannot accept the string type column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

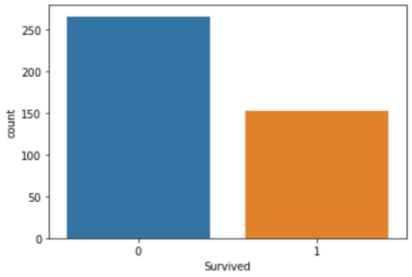
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

We can see that Pclass-1 passenger has a higher survival rate than passengers in lower PClass like 2 and 3. It means mainly the passengers who were travelling in lower class (poor people or workers) died and first-class passengers were given more preference. Also, we can see that the females are highly correlated with survival, so it means females survived more compared to the number of survived males. Lastly, from the bar histogram plot, we can see that around 350 passengers had survived(Orange) and around 600 had died(Blue).



4 Analysis

Below plot shows the number predicted survived and not survived passengers. The accuracy of the model is close to 85%. The result is shown for Test data.



References

[1] Gareth James [at], Daniela Witten [aut], Trevor Hastie [aut, cre], Rob Tibshirani [aut], Balasubramanian Narasimhan [ctb], *Introduction to Statistical Learning, Second Edition*.

[2] Tom M. Mitchell, *Machine Learning: A multistrategy approach, 1997 Edition*.

[3] URL - *https://github.com/cchangyou/100-Days-Of-ML-Code/blob/master/Code/Day%2034%20Random_Forest.md*