Random Forest Classifier For Classifying Birds Species using Scikit-learn

Dr B. Bhoomeshwar
Associate Professor
Dept. Computer Science & Engineering
Mettu university, Mettu
Ethiopia.
eshwarmsis@gmail.com

Dr. Y. Nagesh
Associate Professor
Dept Information Technology
Mettu university, Mettu
Ethiopia.
nageshyagnam1@gmail.com

Dr. K. Raja Shekar
Assistant Professor
Dept. Computer Science & Engineering
Mettu University, Mettu
Ethiopia.
rajashekar562@gmail.com

***Abstract*— A random forest classifier (RFC) is a collection or ensemble of decision trees. Each tree is trained on a random subset of the attributes. We propose a classification technique using voting method with random forests. Random forests are extensions of decision trees and it is a kind of ensemble method. Our proposed method can achieve high accuracy by building several classifiers and running each classifier independently. Accuracy of our proposed method is high compared with other traditional classification algorithms. Voting technique takes outcome from each decision tree and based on the majority of votes it decides which is the actual outcome. Using Scikit-learn tool we evaluated the efficiency of our proposed method. Scikit- learn is a machine learning tool which is extremely used in various machine learning applications for predicting the behavior of data**

***Keywords*— Random Forest, Decision tree, Ensemble method and accuracy**

I. INTRODUCTION

Random Forest is a large collection of decision trees to classify any given instance by using majority votes. The decision trees used in Random Forest Algorithm are typical decision trees. The difference between a single decision tree, particularly in a random forest, is that each Tree is only allowed to look at some of the attributes, typically a small number relative to the total number of attributes available. Each tree is specialized to just those attributes. These Specialized trees are collected and each offers a vote for its prediction. Whichever outcome gets the most votes from the ensemble of specialized trees is the winner. That is the final Prediction of the random forest.

We consider a random forest when there is a sufficient number of attributes to make trees and the accuracy is paramount. When there are fewer trees, the interpretability is difficult compared to a single decision tree. We have to avoid using random forests if interpretability is important because if there are too many trees, the models are quite large and can take a lot of memory during training and prediction. Hence, resource-limited environments may not be able to use random forests.

In existing research with Random forest, they tried to choose training set randomly and predicted the outcome. Then they replaced existing training set with new training set by choosing new feature. By using this new feature selection technique, they have obtained better performance [1]. In their research they improved the performance of Random Forests algorithm by replacing ordinary voting technique with weighted voting technique and they have ontained fast, robust to noise and not overfit Random Forest[3].Then in continuation to this, next researchers have shown great enhancement in action recognition by mapping a 3D video patch with 4D Hough space. Using this mapping method, they even obtained a class label also [3]. In their research work they used SVMs, Decision trees, Bagging, Boosting and Random Forest algorithms for conducting experimental comparison of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5 and Random Forest on seven Microarray cancer data sets. They used two statistical tests to validate their performance and they also observed that Wilcoxon signed rank test is better than sign test.The Experimental results shown that all ensemble methods perform C4.5[4]. In their proposed method they tried to capture the various sounds produced by human activities by using Non-Markovian Ensemble voting technique. Using this technique, a robot can able to infer the corresponding actions. These Non–Markovian Ensemble Voting can classify multiple human activities in online without need for silence detection or audio stream segmentation [6]. In their research they combine the results of multiple classifiers to achieve improved prediction compared to the optimal single classifier. By combining classifiers built from each sub space, the proposed method successfully tackles the dimensionality problem. They have performed a experiment on four data sets from different areas. Their proposed method performed well compared to widely used classification methods.[7].Next technique continued with predictive mapping technique by using Regression Tree Analysis (RTA), Bagging Trees (BT),Random Forests (RF), and Multivariate Adaptive Regression Splines (MARS) .They applied these techniques to four species in USA for mapping current and future climate using kappa and fuzzy kappa statistics. RF and BT produced good results for these four species for future estimates of suitable habitat after climate change [8]. Moreover, the margin theory is being implemented as a confidence phenomenon measuring aspects of the classifier, and to confirm the relevance of input related features for rural classification are discussed the quantitative technique results confirmed the importance of the required joint use of optical analysis multispectral and lidar data are discussed [9]. Here the single vote per tree as per need gives good results, and is faster than alternative approaches. The sampling step can mostly

IJSER

be adjusted to moderate balance between speed and accuracy. Consequently, overall method is very fast but allowing the track of faces need at required frame-rate [11]. In their research they discussed the applications of ensemble methods to microarray and MS-based proteomics. They also discussed the role of these ensemble methods for gene expression, mass spectrometry-based proteomics, gene-gene interaction identification from genome-wide association studies and prediction of regulatory elements from DNA and protein sequences [13].

II. Random Forest Classifier

In this paper Random forest classifier (RFC) used for predicting Birds species. The following species are available in the dataset as shown in figure 1:

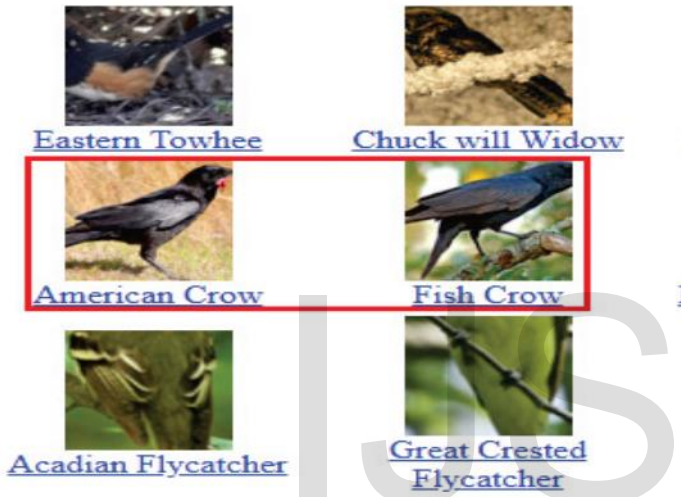


Fig. 1: Various Identical Birds Species

American Crow and the Fish Crow are almost indistinguishable, at least visually. The attributes for each picture, such as color and size, have actually been labeled manually as shown in figure 2.

Fig. 2 : Various Identification attributes of sample

We can observe that the Summer Tanager is marked as having a red throat, a solid belly pattern, a perching-like shape, and so on. The dataset includes information about how long it take to decide it manually on the labels and how confident the person is with their decisions, but we never consider manual decisions.

Fig. 3 : Output result of the text files

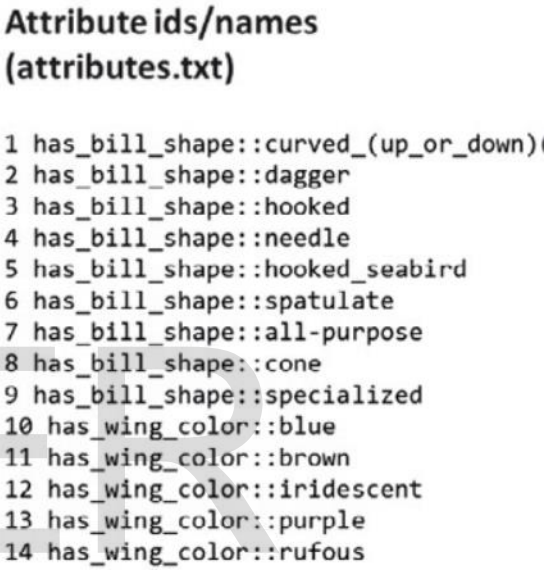


Fig. 4 : Output result of attribute names

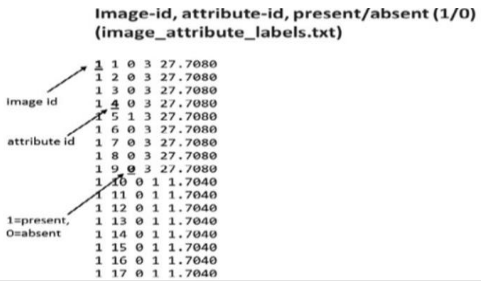
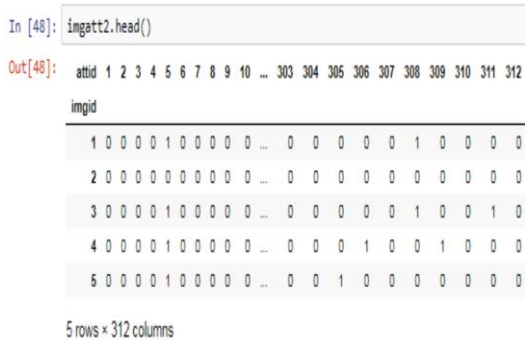


Fig. 5: Output Discrete image result

As shown in figure 3, figure 4 and figure 5 the classes.txt file shows class IDs with the bird species names. The images.txt file shows image IDs and filenames. The species for each picture is given in the image_class_labels.txt file, which connects the class IDs with the image IDs. The attributes.txt file gives the name of each attribute figure 5. connects each image with its attributes in a binary value that's either present or absent for that attribute

III. Experimental Result of RFC



```
In [48]: imgatt2.head()
```

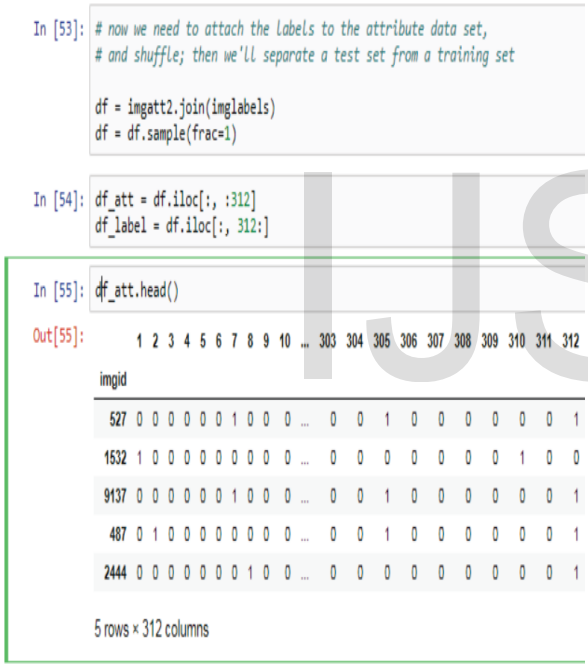
Out[48]:

imgid	attid 1	2	3	4	5	6	7	8	9	10	...	303	304	305	306	307	308	309	310	311	312
1	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	1	0	0	1	0
4	0	0	0	0	1	0	0	0	0	0	...	0	0	0	1	0	0	1	0	0	0
5	0	0	0	0	1	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0

5 rows × 312 columns

Fig. 6: Result of Birds attributes

We feed this data into a random forest. We observed in the above experiment that, we have 312 columns and 312 attributes, which is ultimately about 12,000 images or 12,000 different examples of birds as shown in figure 6.



```
In [53]: # now we need to attach the labels to the attribute data set,
         # and shuffle; then we'll separate a test set from a training set

         df = imgatt2.join(imglabels)
         df = df.sample(frac=1)

In [54]: df_att = df.iloc[:, :312]
         df_label = df.iloc[:, 312:]

In [55]: df_att.head()
```

Out[55]:

imgid	1	2	3	4	5	6	7	8	9	10	...	303	304	305	306	307	308	309	310	311	312
527	0	0	0	0	0	0	1	0	0	0	...	0	0	1	0	0	0	0	0	0	1
1532	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
9137	0	0	0	0	0	0	1	0	0	0	...	0	0	1	0	0	0	0	0	0	1
487	0	1	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	1
2444	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	1

5 rows × 312 columns

Fig. 7: Result of Birds attributes

After shuffling, we have the first row as image 527, the second row as image 1532, and so forth. The attributes in the label data are in agreement. On the first row, it's image 527, which is the number 10. You will not know which bird it is, but it's of the kind, and these are its attributes. But it is finally in the right form. We need to do a training test split. There were 12,000 rows, so let's take the first 8,000 and call them training, and the call rest of them testing (4,000). We'll get the answers using Random Forest Classifier as shown in figure.7.

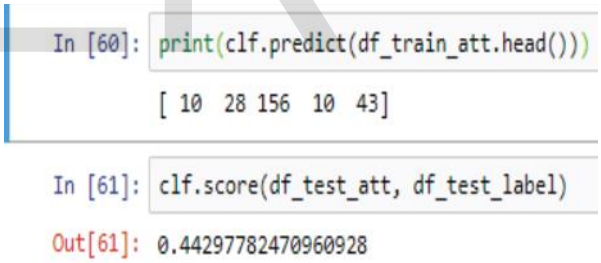
```
In [57]: df_train_att = df_att[:8000]
         df_train_label = df_label[:8000]
         df_test_att = df_att[8000:]
         df_test_label = df_label[8000:]

         df_train_label = df_train_label['label']
         df_test_label = df_test_label['label']

In [58]: from sklearn.ensemble import RandomForestClassifier
         clf = RandomForestClassifier(max_features=50, random_state=0, n_estimators=100)
```

Fig. 8: Dividing data set into training and test tuples

Max features show the number of different columns each tree can look at. For an instance, if we say something like, look at two attributes, that's probably not enough to actually figure out which bird it is. Some birds are unique, so you might need a lot more attributes. Later if we say max_features =50 and the number of estimators denote the number of trees created. The fit actually builds it as shown in figure.8. Let's proceed for a prediction of few cases and using attributes from the very first five rows of the available training set, thereby it predicts species 10, 28, 156, 10, and 43. Testing performance arrives the reliability of 44% accuracy as shown in figure 9.



```
In [60]: print(clf.predict(df_train_att.head()))

         [ 10  28 156  10  43]

In [61]: clf.score(df_test_att, df_test_label)

Out[61]: 0.44297782470960928
```

Fig. 9: Accuracy measure with RFC

IV. Birds species Identification Technique

Let's make a confusion matrix to verify which birds the dataset confuses. The ratio of two hundred by two hundred is very difficult to understand in a numeric form the above output as shown in figure 10.

IJSER

```
In [62]: from sklearn.metrics import confusion_matrix
         pred_labels = clf.predict(df_test_att)
         cm = confusion_matrix(df_test_label, pred_labels)

In [63]: cm

Out[63]: array([[ 5,  1,  6, ...,  0,  1,  0],
                [ 0, 12,  0, ...,  0,  0,  0],
                [ 0,  0,  8, ...,  0,  0,  0],
                ...,
                [ 0,  0,  0, ...,  6,  0,  0],
                [ 0,  0,  0, ...,  0, 11,  0],
                [ 0,  0,  0, ...,  0,  0, 13]])
```

Fig. 10: Result of confusion matrix

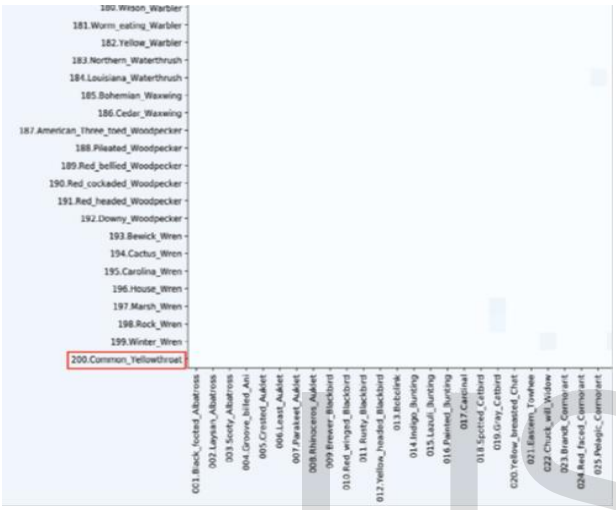


Fig. 11: Result of actual bird names with predicted bird names

The figure 11 shows output which gives on the y axis we represent the actual birds, and on the x-axis represents the predicted birds. The common yellow throat is the true one. After analyzing the following graph, finally arrived to conclude that the common yellow throat is confused with the black-footed albatross. This is the set regarding sparrows. The confusion matrix gives us the things that we expect, that is, birds that look similar are confused with each other. There are little squares of confusion, as we seen in the previous screenshot. For the most part, you don't want to confuse an albatross with a common yellow throat because this means that the dataset doesn't know with what it's doing. Since the bird's names are sorted, lesser is the square of confusion.

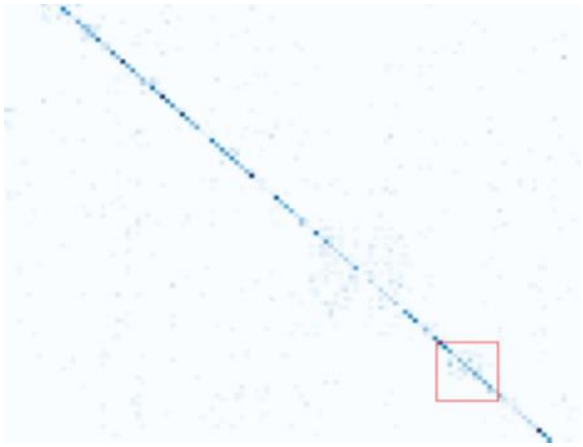


Fig. 12: Result of Birds often confused

As shown in figure 12, it's like a square of confusion that is there between the common yellow throat and the black-footed albatross. Some features are terns, such as the arctic tern, black tern, Caspian tern, and the common tern. Terns are apparently easy to confuse because they look similar.

IV. Comparative Analysis of RFC with SVM and Decision Tree

```
In [67]: from sklearn import tree
         clftree = tree.DecisionTreeClassifier()
         clftree.fit(df_train_att, df_train_label)
         clftree.score(df_test_att, df_test_label)

Out[67]: 0.26953537486800422
```

Fig. 13: Result of accuracy measure of decision tree

```
In [68]: from sklearn import svm
         clfsvm = svm.SVC()
         clfsvm.fit(df_train_att, df_train_label)
         clfsvm.score(df_test_att, df_test_label)

Out[68]: 0.28616684266103487
```

Fig. 14: Result of accuracy measure of SVM

Experimental results show in figure 13 that the accuracy with Decision tree classifier is 27%, which is less than the Random Forest Classifier (RFC) which has shown 44% accuracy in figure 9. Therefore, the decision tree is worse. If we use a Support Vector Machine (SVM), which is the neural network approach, the accuracy observed as 29% in figure 13, which is even less than 44% obtained by RFC which is shown in figure 9.

IJSER

```
In [90]: import matplotlib.pyplot as plt
         from mpl_toolkits.mplot3d import Axes3D
         from matplotlib import cm
         fig = plt.figure()
         fig.clf()
         ax = fig.gca(projection='3d')
         x = rf_params[:,0]
         y = rf_params[:,1]
         z = rf_params[:,2]
         ax.scatter(x, y, z)
         ax.set_zlim(0.2, 0.5)
         ax.set_xlabel('Max features')
         ax.set_ylabel('Num estimators')
         ax.set_zlabel('Avg accuracy')
         plt.show()
```

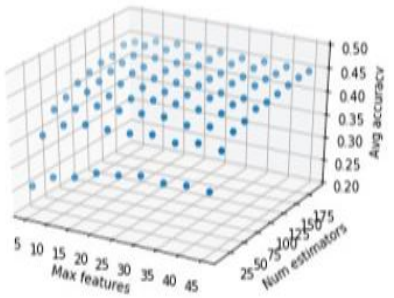


Fig. 15: Accuracy measure of RFC

In the figure 15 it represents Average accuracy of RFC technique in multi-dimensional cuboid. In x-axis it takes number of features randomly chosen, y-axis represents numeric estimators and z-axis represents average accuracy. When performing the same experiment many times by randomly changing features always it has shown the 44% accuracy. Hence it proved that average accuracy of RFC (Random Forest Classifier) is 44% which better than that of Support vector machine and Decision Tree.

V. CONCLUSION

We show that accuracy of our proposed method is high compared with other traditional classification algorithms. Voting technique takes outcome from each decision tree and based on the majority of votes it decides which is the actual outcome. Using Scikit learn tool we evaluated the efficiency of our proposed method. Scikit learn is a machine learning tool which is extremely used elaborately in various machine learning applications for predicting the behavior of data. Also, we can prove that increasing the number of trees produces a better outcome. Also, increasing the number of features produces better outcomes if we are able to observe more features, but ultimately, when we consider at about 20 to 30 features and we have about 75 to 100 trees, that's about as good as you're going to get an accuracy of 45%. We have focused on comparing the voting method with different feature detectors, rather than on producing the best possible facial feature finder aspects. In future the approach being more accurate in identifying bird's species by considering more number of features equally.

References

[1] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[2] Andy Liaw and Matthew Wiener. Classification and Regression by RandomForest, Vol.2(1), ISSN 1609-3631 ,2002

[3] Marko Robnik-Sikonja. Improving Random Forests, 359–370, 2004.

[4] Hong Hu Jiuyong Li, Ashley Plank,Hua Wang,Grant Daggard. A Comparative Study of Classification Methods for Microarray Data Analysis,2005

[5] Anantha M. Prasad, Louis R. Iverson, and Andy Liaw. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction, 9: 181–199,2006

[6] Johannes A. Stork Luciano, Spinello Jens, Silva Kai O. Arras. Audio-Based Human Activity Recognition Using Non-Markovian Ensemble Voting,2006

[7] Hongshik Ahn, Hojin Moon, Melissa J. Fazzari, Noha Lima,James J. Chen, Ralph L. Kodell. Classification by ensembles from random partitions of high-dimensional data, 51 (2007) 6166 – 6179

[8] Angela Yao, Juerge Gall & Luc Van Gool. ". Hough Transform-Based Voting Framework for Action Recognition ",2010

[9] Li Guo, Nesrine Chehata , Clément Mallet , Samia Boukir. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests,66:56-66,2011

[10] Gang Yu, Norberto A. Goussies, Junsong Yuan, and Zicheng Liu. Fast Action Detection via Discriminative Random Forest Voting and Top-K Subvolume Search, VOL.13, NO.3,507-517, JUNE2011

[11] T.F.Cootes, M.C.Ionita, C.Lindner and P.Sauer. Robust and Accurate Shape Model Fitting using Random Forest Regression Voting, Conference Paper • October 2012

[12] Vrushali Y Kulkarni, Dr Pradeep K Sinha. Random Forest Classifiers: A Survey and Future Research Directions, ISSN:2051-0845, Vol.36, Issue.1,1144-1153,2013

[13] Pengyi Yang, Yee Hwa Yang, Bing B. Zhou and Albert Y. Zomaya. A review of ensemble methods in bioinformatics: Including stability of feature selection and ensemble feature selection methods, 5, (4):296-308,2016

[14] WEIWEI LIN1, ZIMING WU, LONGXIN LIN, ANGZHAN WEN, AND JIN LI. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis, VOLUME 5,16568-16574,2017

IJSER