Forest Land Cover Classification

Rui Yang

***Abstract*—This paper describes a classification system that classifies the forest based on the remote sensing signals from NASA' Landsat satellites. In this paper, I represent two machine learning models by using pyspark to train and test on the large dataset from the remote sensing dataset. The logistic regression and the random forest algorithms had been applied to the data. Both models give us exciting results on predicting forests. The result shows the forest land covers on the earth surface.**

I. Introduction

Forests are vital to life on Earth. They purify the air we breathe, filter the water we drink, prevent erosion, and act as an important buffer against climate change. Understanding how the forests work in our earth environment system and monitoring them are important topics for our human and other lives on our earth. One of the benefits that advanced aerospace technologies provide us is that it can send us the images of land covers from the satellites in their orbits. Classifying the forest area and monitoring the change of it from the space now is achievable. NASA' Landsat satellites, both Landsat 7 and Landsat 8 are in a near-polar orbit of our planet. Each satellite repeats its orbital pattern every 16 days, with the two spacecraft offset so that each spot on Earth is measured by one or the other every eight days. As the Landsat satellites orbit, the instruments capture scenes across a swath of the planet that is 185 kilometers (115 miles) wide. Each pixel in these images is 30-meters across, which is about the size of a baseball infield, or — more important for resource management — an average U.S. crop field. This project used a land cover remote sensing dataset of forest from the Global Land Analysis Discovery at University of Maryland. And developed two machine learning models to identify the forest cover based on the remote sensing data from the satellite. The results are the forest cover or not on one pixel of a 1000 by 1000 pixels satellite sensing image. Which one pixel by one pixel represents the 30 by 30 square meters area on the actual earth surface.

II. Data

A. Dataset

The dataset came from the Global Land Analysis Discovery at University of Maryland (https://glad.umd.edu/). The dataset comes from the satellites' remote sensing signals of a year. GLAD's product will produce an image of 1000 by 1000 pixels based on remote sensing. And split 70 percent for training and 30 percent for testing. Each 1 by 1 pixel contains the raw remote sensing data and their features of a 30 meters by 30 meters area on the earth surface. In total, this dataset will cover 900,000,000 square meters area. The remote sensing data contains visible light color bands: blue, green, red. And invisible light: near-infrared regions light, short-wave infrared light. For each of the light bands, there are several statistical features: minimum, maximum, median, average of minimum and maximum and the difference between max amplitude and min amplitude. The dataset's shape will be a P by N 2D array in the .csv file. Where p is the number of pixels (1000 by 1000), and n is the number of features in each pixel. The total sample points will be 1000 x 1000, which is 10e6 sample points. For each sample, it contains 35 features. The detail about feature metric list is show in the Figure 1:

Index	The index of sample points, not use in this project
label,	The label of forest/not forest,(1/0)
blue_min,	The minimum blue light signal in amplitude of a year
blue_max,	The maximum blue light signal in amplitude of a year
blue_median,	The median blue light signal in amplitude of a year
blue_avminmax,	The average of max and min amplitude
AMP_blue_max_blue_min,	The difference between max amplitude and min amplitude
green_min,	The minimum green light signal in amplitude of a year
green_max,	The maximum green light signal in amplitude of a year
green_median,	The median green light signal in amplitude of a year
green_avminmax,	The average of max and min amplitude
AMP_green_max_green_min,	The difference between max amplitude and min amplitude
red_min,	The minimum green light signal in amplitude of a year
red_max,	The maximum green light signal in amplitude of a year
red_median,	The median green light signal in amplitude of a year
red_avminmax,	The average of max and min amplitude
AMP_red_max_red_min,	The difference between max amplitude and min amplitude
nir_min,	The minimum near-infrared regions light signal in amplitude of a year
nir_max,	The maximum near-infrared regions light signal in amplitude of a year
nir_median,	The median near-infrared regions light signal in amplitude of a year
nir_avminmax,	The average of max and min amplitude
AMP_nir_max_nir_min,	The difference between max amplitude and min amplitude
swir1_min,	The minimum Short-wave infrared light signal in amplitude of a year
swir1_max,	The maximum Short-wave infrared light signal in amplitude of a year
swir1_median,	The median Short-wave infrared light signal in amplitude of a year
swir1_avminmax,	The average of max and min amplitude
AMP_swir1_max_swir1_min,	The difference between max amplitude and min amplitude
swir2_min,	The minimum Short-wave infrared light signal in amplitude of a year
swir2_max,	The maximum Short-wave infrared light signal in amplitude of a year
swir2_median,	The median Short-wave infrared light signal in amplitude of a year
swir2_avminmax,	The average of max and min amplitude
AMP_swir2_max_swir2_min,	The difference between max amplitude and min amplitude
RN_min,	The minimum normalized difference vegetation index (Red and NIR) of a year
RN_max,	The maximum normalized difference vegetation index (Red and NIR) of a year
RN_median,	The median normalized difference vegetation index (Red and NIR) of a year
RN_avminmax,	The average of max and min amplitude
AMP_RN_max_RN_min,	The difference between max amplitude and min amplitude
flag	Not use in this project

Fig. 1. Features List of Dataset

The remote sensors are mounted on the international space station. Theoretically the satellites will revisit the same point about every 16 days. The sensors keep collecting the remote sensing data for one point every 16 days. Annually, for one

30 meters by 30 meters on the earth surface, it should have 24 sample points containing all remote signals raw data. However, the satellites will not revisit the same point accurately. The satellites will go back to a range of the point it was 16 days ago. Which will give a remote sensing signal of a different place. And also the weather conditions between 16 days could be very different. Heavy clouds can cover the surface and the remote sensing signals could have major changes because of the clouds. Preprocessing the raw data is a huge amount of work. Doing signal processing itself could be a challenging topic. Lucky, the GLAD group at University of Maryland has an industry-level progress to pre-process the raw data of remote signals. This dataset came from one of their projects, Global Forest Watch. The GLAD group applied a cloud mask, which masks off all the clouds on the image, and also does the pre-processing to get the statistical features of the whole year's raw data.

III. Experiment

In this project, there are two machine learning models that have been used to predict on the satellite remote sensing signals data as a forest classifier by using pyspark.

The pyspark workflow of this project is:

1. Creating a spark session, loading the large .csv file.
2. Using *vectorAssembler()* to set up the Features by creating a features array
3. Splitting the training/testing data by 70/30
4. Training the machine learning algorithm
5. K-fold cross validation
6. Training on the best performance model
7. Using the best performance model to predict on all data, plot the predicting forest map

The features will be used as the dataset contains. The satellite has remote statistical features of sensing signals in different color bands (both visible light and invisible light) and the normalized difference vegetation index of Red and NIR light.

Condensing the size of the dataset and the limitation of the performance on the personal device, choosing algorithms from simple to complex is a good strategy. The algorithms used for this project are logistic regression and random forest. The neural network was considered. However, due to the limitation of the performance of the personal device, and high accuracy of random forest model given, the neural network model has not been developed for this project. Choosing logistic regression as the baseline of this project will give a standard of the performance of the model.

And the random forest is used by GLAD, they have a very successful model on predicting other land covers by using random forest. Choosing the algorithm they have had successful experiences on is a good strategy to run the models in this project. However, they have the most advanced computational powder in their research area. Testing on the performance of random forest in this project on a personal device can be a good comparison with them on computational power.

A. Baseline

Logistic Regression:

Logistic regression has been used as the baseline machine learning classification algorithm. Considering the size of the dataset, first check if the simple machine learning algorithm could be able to handle this job. The data set has been randomly split to 70 percent of training data and 30 percent of testing data. The default parameters are regularization parameter equal to 0.5 and elastic net parameter equal to 0.2.

The normalized logistic regression confusion matrix is shown in figure 2. And the classification report table is shown in figure 3. The logistic regression model has a very good performance on the forest classification. The overall f1 score is 0.903. The true positive is less than the true negative, it could be caused by the randomly split dataset with an imbalance between true labels and false labels. And the dataset itself is imbalanced, there are about 50,000 more sample points label as tree than non-tree. Since the performance is too good on a very simple model, a cross validation is needed to check if the model is overfitting on the given dataset.

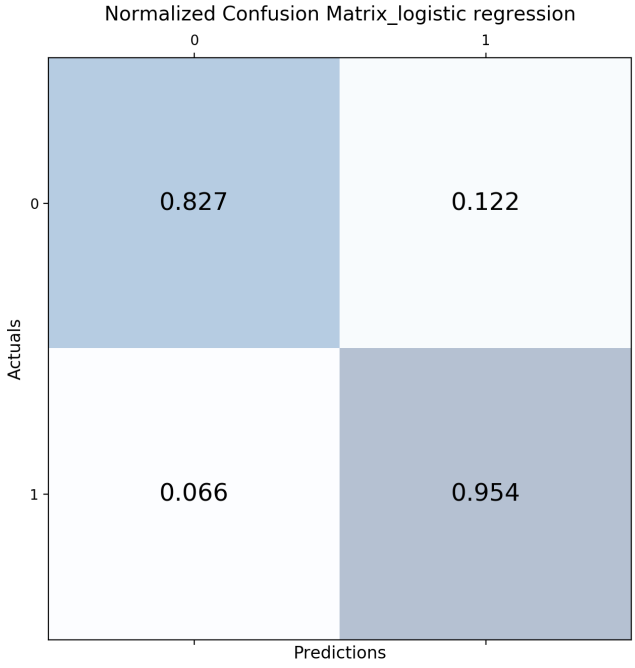

Fig. 2. Logistic Regression Confusion Matrix

Logistic Regression with K-fold Cross Validation:

By checking the baseline performance in the previous section. A cross validation is needed to check if there are over-fitting issues. And the cross validation will also help to train the model with an imbalance dataset. However, due to the limitation of the personal computational device. The largest number of K I can run on my personal device is 5. The 5-fold cross validation score is shown in figure 5. There is an

Label	Precision	Recall	f1-score	support
0	0.93	0.83	0.87	124357
1	0.89	0.95	0.92	175984
Accuracy			0.9	300341
Macro Avg	0.91	0.89	0.9	300341
Weighted Avg	0.9	0.9	0.9	300341

Fig. 3. Logistic Regression scores

improvement on the overall accuracy, from 0.901 to 0.911. The prediction of the true positive has increased. Since the 5 fold validation score does not have a huge difference with the baseline model. The logistic regression model has been proved that it has a good performance on this dataset with an average accuracy score of 0.91.

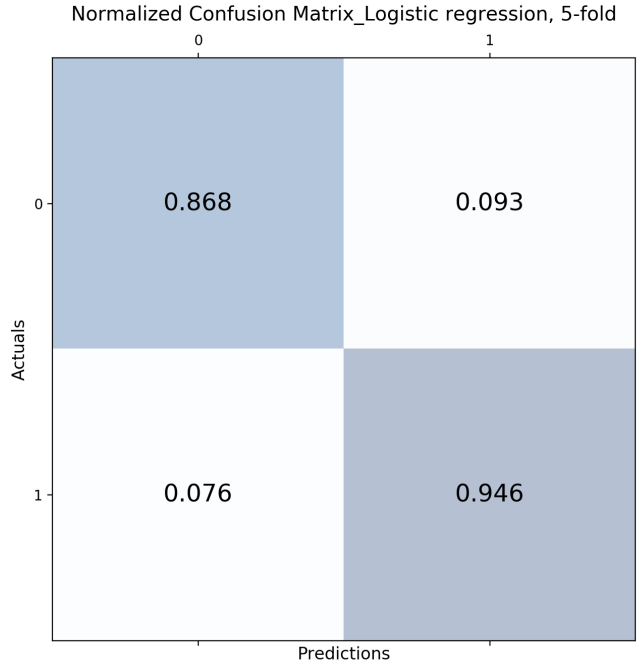

Fig. 4. Logistic Regression with 5-fold Validation Confusion Matrix

Label	Precision	Recall	f1-score	support
0	0.94	0.91	0.92	124590
1	0.96	0.96	0.95	175912
Accuracy			0.94	300502
Macro Avg	0.94	0.93	0.94	300502
Weighted Avg	0.94	0.94	0.94	300502

Fig. 5. Logistic Regression with 5-fold Cross Validation Scores

Logistic Regression with Parameter Combinations: There are different regularization parameters and elastic net parameter combinations have been tested for the logistic regression model. The final best performance logistic regression model is the regularization parameter of 0.3 and elastic net parameter of 0.2. The average accuracy of 5-fold cross validation is 0.977.

B. Advanced model

Random Forest:

In this project, the logistic regression model already gives us exciting results on forest classification. However, this project still needs a comparison with the logistic regression model. The GLAD group at UMD has a very successful experience with building land cover classification models with random forest. According to their experiences, building and training a random forest model and testing on personal devices can be a good comparison with logistic regression model's performances.

Similar preparation as logistic regression, the dataset has been split into training and testing parts. The training part takes 70 percentages and the testing part takes 30 percentages. The default number of trees is 10, the maximum depth is 5 and the maximum bins is 20.

The normalized random forest confusion matrix is shown in figure 6. And the classification report table is shown in figure 7. The random forest model has a very good performance on the forest classification as expected. The overall f1 score is 0.91. The true positive is again less than the true negative, it could be caused by the randomly split dataset with an imbalance between true labels and false labels. And the dataset itself is imbalanced, there are about 50,000 more sample points labeled as tree than non-tree. The overall performance is good enough. And cross validation is needed to check if the model is overfitting on the given dataset.

Random Forest with k-fold Cross Validation:

In order to increase the confidence on the random forest model, a k-fold cross validation is needed in this project. Due to the limitation on my personal laptop, the largest K value I can run on my laptop without crashing is 5. The 5-fold cross validation score is shown in figure 8. There is an improvement on the overall accuracy, from 0.91 to 0.953. The prediction of the true positive has increased. Since the 5 fold cross validation score has increased, the performance of the random forest model has been improved by using 5-fold cross validation. The final overall 5-fold cross validation score is 0.953.

Random Forest with Parameter Combinations:

There are three different parameters that can be tuned in the random forest model, number of trees, maximum depth and maximum bins. There are different combinations that have been tested for the random forest model. The final best performance model is the number of trees equals to 5, maximum depth equals to 7 and maximum bins equals 30. The average accuracy of 5-fold cross validation is 0.989 on the best model. The performance of the random forest is very good on the given dataset.

Predicted Forest Map VS Real Forest Map:

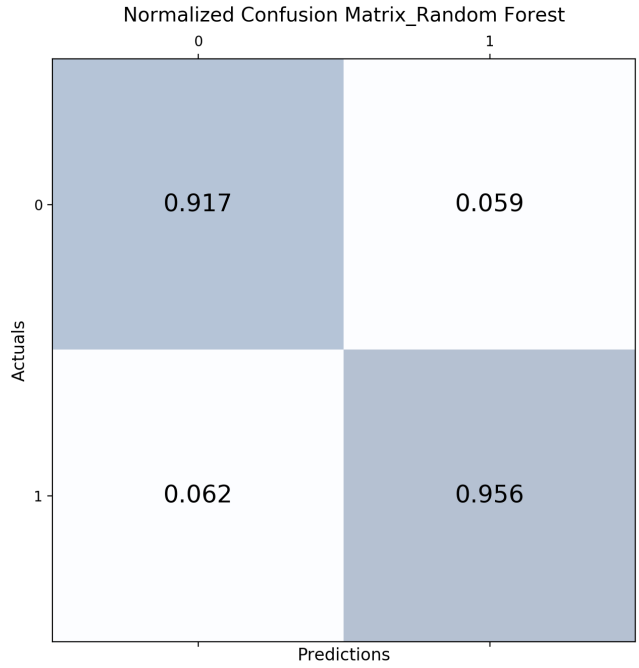

Fig. 6. Random Forest Confusion Matrix

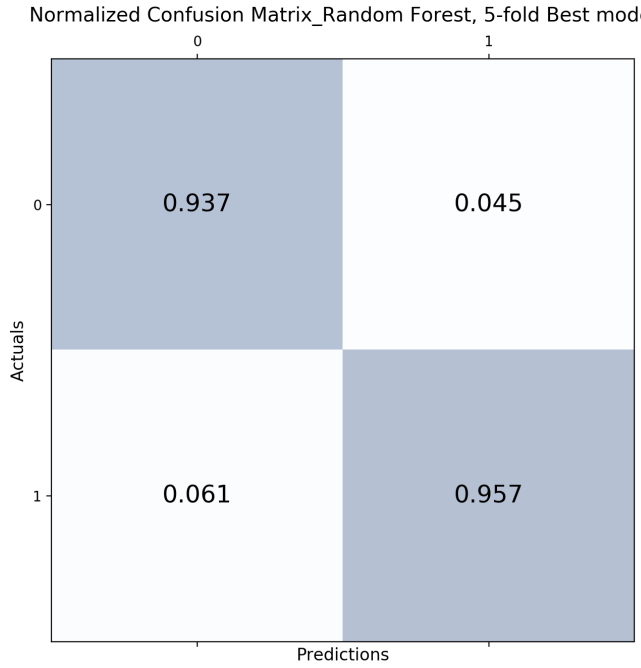

Fig. 8. Random Forest with 5-fold Cross Validation Confusion Matrix

Label	Precision	Recall	f1-score	support
0	0.92	0.87	0.89	124301
1	0.91	0.95	0.93	175660
Accuracy			0.91	299961
Macro Avg	0.91	0.91	0.91	299961
Weighted Avg	0.91	0.91	0.91	299961

Fig. 7. Random Forest scores

Label	Precision	Recall	f1-score	support
0	0.94	0.94	0.94	124202
1	0.96	0.96	0.96	175373
Accuracy			0.91	299575
Macro Avg	0.95	0.95	0.95	299575
Weighted Avg	0.95	0.95	0.95	299575

Fig. 9. Random Forest scores with 5-fold cross validation

The figure 10. shows the predicted forest map generated by the best performance model, the random forest classifier with 0.977 average score. The forest pixels are marked by yellow and non-forest pixels are marked by purple. The size of this forest image is 1000 by 1000 pixels, which is about 900 square kilometers on the earth surface. Compared with the actual forest cover map, the predicted forest map almost has no difference with the real signal generated forest map. On both the actual and predicted forest map, there are some horizontal lines over the map. Those horizontal lines are caused by the off orbit of the satellites. When the satellites revisit the same place, the orbit can be different from 16 days ago, which could cause some edge cases to mixed regions.

IV. Conclusion

The project takes a very large dataset (contains 1 million rows) of remote sensing signals from the satellites. And building, training and testing by using pyspark. Both logistic regression model and random forest model give us a good prediction on the dataset. The results of the logistic regression model bring a lot of surprise. Thanks to the high-accuracy dataset, the logistic regression model gives an average score of 0.977 on the best model with 5-fold cross validation. The random forest model, which the GLAD group already had some successful experiences on this classification algorithm, gives us an average score of 0.989 on the best model with 5-fold cross validation. Due to the limitation of the personal computational device, the size of the data would crash the memory. However, with Spark's help, handling this large amount of dataset on a personal device is possible.

V. Discussion

Data preprocessing is a very important part of using remote sensing signals. The remote sensors on the satellites can be able to receive the light signals in different bands. The sonser will collect many channels of light bands of a year. For one location on the earth,theoretically the satellite will revisit it every 16 days. For one 30 by 30 square meters location on the

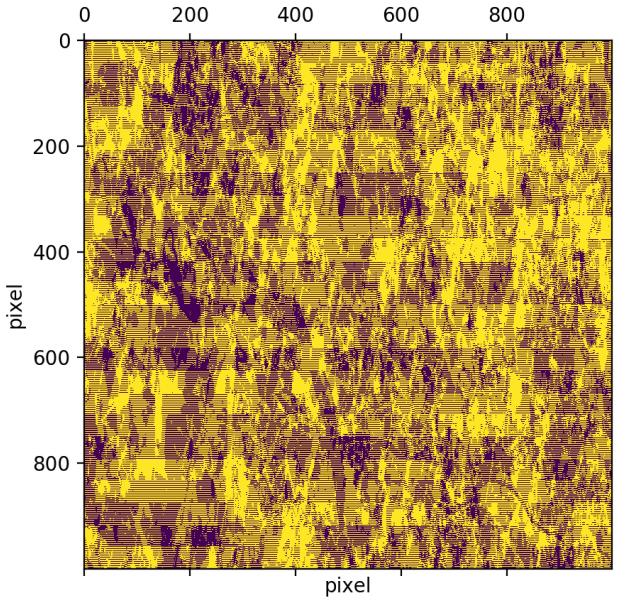

Fig. 10. The Predicted Forest Map

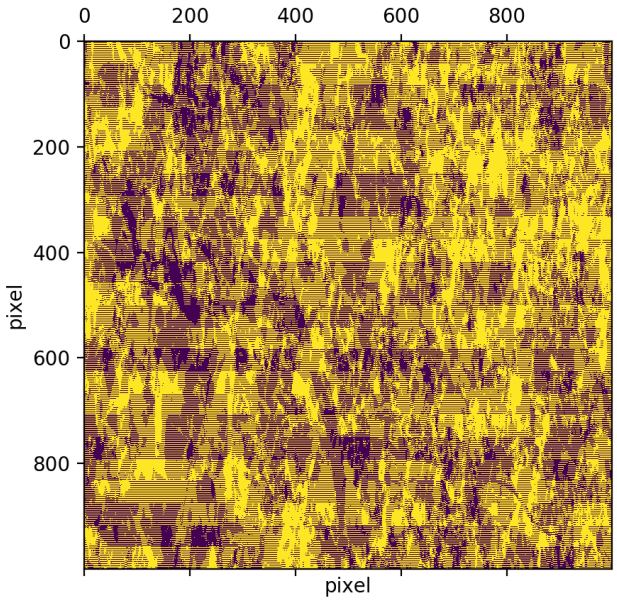

Fig. 11. The Real Forest Map

earth, one year of data on one pixel is about 25 different days' raw data. However, the earth's surface changes during the year. For example, deciduous forests start to fall their leaves during the autumn and do not have leaves during winter. This characteristic of deciduous forests will reflect on the remote sensing signals. During the summer there should be a larger green light channel amplitude than the winter on the deciduous forest area. The other difficulty of preprocessing the remote sensing signals from space is weather. A heavy cloudy day will give the sensor very different signals, since the cloud can cover up the signals from the earth surface.

Finding the characteristic of each different land cover and preprocessing the signal to handle the season changing could be another very challenging and important topic in remote sensing areas. The GLAD group at UMD have their successful products on pre-processing the raw remote sensing signals from NASA's Landsat satellites, which save a lot of time and work for other researchers who want to study and use the remote sensing data. For this project, because of the high-accurate preprocessed data, even the simplest logistic regression model can have a surprising high performance on forest classification.

The satellite will have some deviations on one pixel when it revisits after 15 days. For the large scale land covers, like global forest land cover, this deviation may not cause a huge issue. Since one pixel reflects on the actual earth surface is only about 30 by 30 square meters. These deviations are acceptable for large forest land cover. However, for any other land cover smaller than 30 by 30 square meters, the accuracy of this dataset can be very off. For this project and also the GLAD group's products all focus on the very large scale of land covers, like forest, water, crops and cities. More smaller scale items classification, like vehicles or humans on this dataset should not be considered.

VI. Future Works

1. One pixel is 30 by 30 square meters, where 1000 by 1000 pixel is 900 square kilometers, which is about 347 square miles. Compare with the area size of the state of Maryland, which is over 12,400 square miles. The map scale on this project is very small compared to the global scale. For this project the computational power was limited by the personal device. In the future, running this forest classification model on a powerful machine to get the global forest map should be achievable.

2. Feature extension. The GLAD group at UMD has a very efficient industry-level progress to preprocess the raw data from remote sensors. However, the feature extension on the raw data is not very abundant. There can be more statistical feature extensions from the raw remote sensing data.

3.Deciduous forests vs evergreens classifications by using this dataset. Deciduous forests have the characteristic of changing leaves during a year, which will reflect on the remote sensing signals.

VII. Reference

NASA. 2021. Landsat Overview. [online] Available at: ¡https://www.nasa.gov/mission$_{p}ages/landsat/overview/index.html >$ $[Accessed 11 December 2021]$.

Glad.umd.edu. 2021. Dataset — GLAD. [online] Available at: ¡https://glad.umd.edu/dataset¿ [Accessed 11 December 2021].