Anomaly Detection Challenge
Challenge 1: Cotton Crops Classification In Satellite Images

Mustika Rizki Fitriyanti (03667399)

Ishmeet Kaur (03677735)

November 2016

1 Introduction

The aim of this challenge is to use supervised anomaly detection in order to detect cotton soil on satellite image data. The data that is used in this challenge is a sub-area of a scene that consisting of 82 x 100 pixels. It has been split into training and test sets with a total of 36 features. Each line contains the pixel values in the four spectral bands of each of the 9 pixels in the 3x3 neighbourhood and each of it has been labeled with 0 as Normal Soil and 1 as Cotton Soil.

2 Data Preprocessing

Initially, we rearranged the data based on the spectral value of every image. Since the satellite data contained missing values , handling those values was integral .We replaced the missing values based on the other dataset . The following methods were used for data replacement.

- Mean

  Replacement of the missing values was done both row and column wise.

- Median

  The missing values were replaced by the median value both along the row and column.

- Mode

  The missing values were replaced by the most frequent occurring element in both row and column wise.

- Min/Max

  Minimum and maximum values for both row and column were used to replace missing values.

- Interpolation

  The missing values were estimated by interpolation.

- Per Spectral Values

  The mean , median and mode values for every spectral of the image were used to replace the missing values per spectral.

3 Create and Evaluate The Model

After the pre-processing step, which consisted of rearranging the data and replace the missing values, we created the model using some algorithms and evaluated the model.

3.1 Using Stratified k- Cross Validation

Based on information that we read on scikit learn website [1] and a discussion on stackoverflow [2], we decided to use Stratified Cross Validation (4 fold) since there is an unbalanced proportion between the cotton and normal soil in the training set.

3.2 Result 1: Implementation of Classification Algorithms

During the model creation, we implemented every result of the replacement method into some algorithms such as Naive Bayes Classifier, Logistic Regression, Decision Tree Classifier, K Nearest Neighbors and Random Forest, to find the best classifier model to fit in the data set. Some additional evaluations made during model selection are summarized below:

- By Calculating the mis classification error, we found that the best accuracy for K-Nearest Neighbors could be achieved when the k=5.

- Initially,we tried with number of estimators as 50 on Random Forest Classifier.

After implementing all the algorithms and looking at the accuracy using the cross validation, we found that the best accuracy was generated by Random Forest Classifiers as seen in the Figure 1.

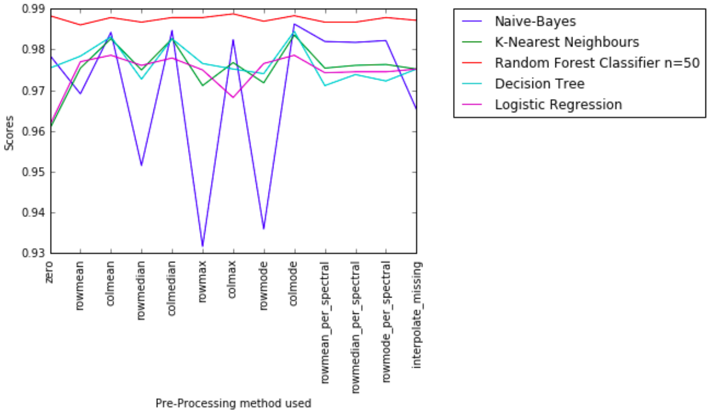

Figure 1. First Result of Accuracy score with a different replacement method

3.3 Result 2: Enhancing Accuracy and Tuning Parameters

As the Random Forest proved to be the best classifier out of the classifiers after comparing the performance of various classification algorithms on the processed dataset,we tried to enhance the accurracy by tuning the parameters of the Random Forest Algorithm.)After detailed analysis of the random forest algorithm,we observed that higher the number of tress used for iteration ,higher is the accuracy. Using OOB(Out of Bag) Error Rate, we make a list number of maximum estimators for every replacement method, then we process in Random Forest Classifier with cross validation in order to make it comparable to other algorithms as seen in Figure 2.

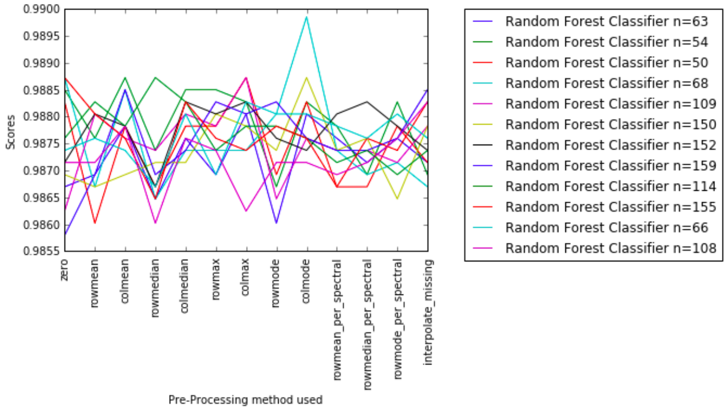

Figure 2.Result of Accuracy score with a different number of estimators

Based on the figure, we tried to set the the number of tree to be 68 instead of 50.

Here is the result of private and public score that we got on Kaggle:

Method	Public Score	Private Score
Random Forest Colmode n=68	0.9812	0.98953
Random Forest Colmedian n=68	0.97693	0.99421
Random Forest Colmean n=50	0.97693	0.99009

4 Result Summary

After evaluating the performance of the Random Forest Algorithm with the number of estimators as 50, we concluded that it worked comparatively better as compared to other algorithms as seen in Figure 1 in section 3.2.

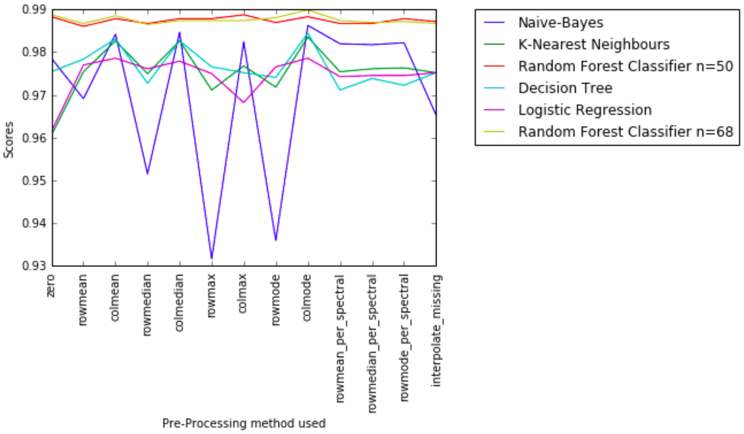

Figure 3. Comparison the Random Forest Classifier n=68 with other algorithms

After increasing the number of estimators to be 68, the accuracy increased. From Figure 3, we can see the maximum accuracy is achieved by Random Forest Classifier with n=68 and using colmode as replacement method. However, according to the private scoring, the maximum accuracy on the testing data is achieved by Random Forest Classifier with n=68 using colmedian as replacement method.

References

[1] $http : //scikit - learn.org/stable/modules/cross_validation.html$

[2] $http : //stackoverflow.com/questions/32615429/k - fold - stratified - cross - validation - with - imbalanced - classes$