CropHarvest

January 3, 2022

1 Student Information

Full Name: Reza Bayat
Kaggle Username: rezabyt

2 Introduction

The problem is classifying each land data point into two main categories, crop and non-crop type. The category of the problems is a classification since we know that we only have two classes to predict, and the dataset is labeled. The random forest algorithm has been used to tackle this problem, which basically is a set of decision trees trained on the proportion of dataset or features. The best hyper-parameters choice for the random-forest model, selected based on the Grid Searching on hyper-parameters and cross-validation, had the f1_score of 0.99757 on the held-out test dataset.

3 Methods

3.1 Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset or features and uses averaging or voting to improve the predictive accuracy and control over-fitting [1]; it can be used for many tasks; especially for classification; the output of the random forest is the class selected by most trees. As an another advantage, random forests are great with high-dimensional data since it works with subsets of data or features, and can be used to tackle our problem with many features.

3.2 Grid Searching

Grid Search is a tuning technique that attempts to compute the optimum values of hyper-parameters. It is an exhaustive search that is performed on the specific parameter values of a model. For choosing the best set of hyper-parameters, the grid search approach is used to find the best model with the best performance on the validation set, defined by the cross-validation approach.
For this problem, I did grid-search on the "n_estimators", "min_samples_leaf", and "bootstrap" hyper-parameters of the random-forest model; the HPs choices and the best setup are presented in the following table.

	bootstrap	min_samples_leaf	n_estimators
Choices	True or False	3, 5, 11, 30	100, 1000
Best	False	3	1000

4 Results

Before training and doing a hyper-parameters search on any model, the dataset is split into two datasets, train and validation sets since I wanted to check the performance of the model on the unseen dataset, the validation set in this section has not been used for training or choosing a model, so the results are unbiased; 80% of the dataset is used for training and the rest of it is sued for the final validation set.
Note: for final evaluation, model has been trained on the whole dataset; and the f1_score of the model on test set is 0.99757.

4.1 Metrics

Three metrics have been used to monitor the performance of the model; the total accuracy of the model on the validation set is 85.03%; the balanced accuracy is 83.97%, and fr_score is 88.96%.

Total Accuracy	Balanced Accuracy	F1_Score
85.03%	83.97%	88.96%

4.2 Confusion Matrix

The Confusion Matrix of the model on the validation set is presented in the figure 1.

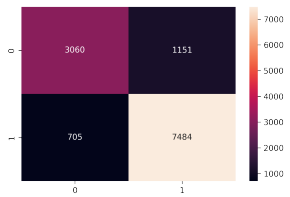

Figure 1: Confusion Matrix

4.3 ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters; i) True Positive Rate and ii) False Positive Rate: [2]
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

The ROC curve of the model on the validation set is presented in the figure 2, which is very close to the ideal ROC curve.

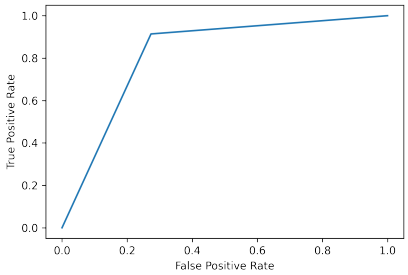

Figure 2: ROC Curve

5 Discussion

First of all, since hyper-parameters searching needs a lot of computing resources, I only could do it on the Random Forest; therefore, if more resources are available, we can try more hyper-parameters searching on different kinds of algorithms such as XGBoost, etc.
In the following section, some suggestions for future works and improving our approach are also presented:

- Considering the nature of features and using feature selection techniques:
  - Since the nature of the dataset is time series we also can use other models such as RNN.
  - Creating new features: using existing features to create new features that are meaningful.
- Customizing algorithms for imbalance datasets by using some methods such as weighted loss or re-sampling data using methods like Under-Sampling and Over-Sampling.
- Trying other kinds of ensemble methods; for example, a voting method using different types of algorithms including decision tree, SVM, etc.
- Using a large portion of the original CropHarvest dataset. We can apply much more complex models and features analyses; in this case, we also can use pre-trained models.

6 Statement of Contributions

"I hereby state that all the work presented in this report is that of the author"

References

[1] RandomForest, `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`, 19 10 2021.

[2] ROC, `https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc`, 19 10 2021.