Ahmet Süheyl Kiriş
509211101

1 Introduction

It was studied over Botswana's Okavango Delta with a dataset taken by NASA's Earth Observation-1 satellite between 2001 and 2004. The data is taken over an area of 7.7 km and has a resolution of 30m/pixel. The noisy 97 bands were removed and the data received over 145 bands were studied. The data analyzed in this study consists of observations from 14 defined classes that represent land cover types in seasonal marshes, sparse marshes and drier woodlands etc. This data has been downloaded from RSLAB.

2 Data & Data Prepation

The data is in the form of a 3D matrix and the size of the data is 1450x256x145. If we consider the received data as a photograph, it is a photograph with 1450x256 pixels and each pixel is a vector with 145 elements. However, the groundtruth data is only a 1450x256 matrix.

Table 1: Class Numbers & Features

Class No	Class Feature
0	Unlabeled
1	Water
2	Hippo Grass
3	Floodplain grasses 1
4	Floodplain grasses 2
5	Reeds
6	Riparian
7	Firescar
8	Island interior
9	Acacia woodlands
10	Acacia shrublands
11	Acacia grasslands
12	Short mopane
13	Mixed mopane
14	Exposed soils

When we consider the groundtruth data, there are 14 different classes in total, as seen in Table 1. However, 99.3% of the data, which do not belong to any class, are labeled with the value "0". Other pixels are almost evenly distributed as seen in Figure 1. Without preparing the groundtruth data, no algorithm will give a proper result. Therefore, the classified pixels on the groundtruth data and their boundaries should be selected very well and the groundtruth data should be made effective. Then supervised learning algorithms will become applicable on top of the underlying data.

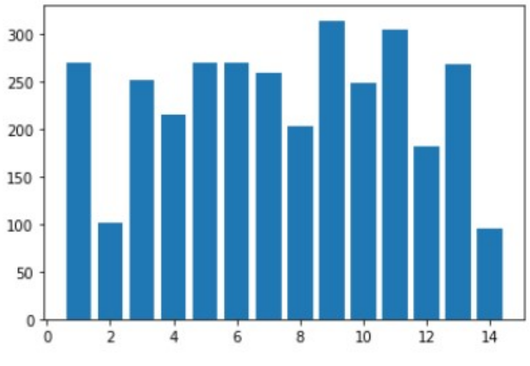

Figure 1: Distribution of Labeled Pixels

Data was prepared with two different methods. First, if all elements of a row or column are ”0”, that row or column number is deleted from both groundtruth and satellite data. In this way, it is ensured that each row or column contains at least one labeled pixel. Despite this, labeled pixels only make up 4% of all pixels.

In the second method, only labeled pixels and corresponding elements in satellite data (along all bands) are selected. In this way, all unlabeled pixels are ignored.

3 Methods & Results

Only machine learning algorithms have been tested on this data. Deep learning, reinforcement learning etc. were not used. In this context, Logistic Regression, Support Vector Machine, KNN, LDA, K-Means algorithms were used with different parameters. Hierarchical Clustering algorithm has occured memory error. It is applicable for smaller data. PCA and DBSCAN algorithms are not suitable for this data.

Considering previous studies, 3 different “solvers” were used for the logistic regression model: “lbfgs”, “newton-cg” and “sag”. For the SVM model, the ”solver” parameters are: ”linear” and ”rbf”. However, “C” and “tol” parameters are kept large enough for both models. For KNN and K-Means algorithms, the k value is chosen as the number of classes, “k=14”. For LDA, only ”n_components ¡= n classes-1” condition is provided. Except LDA, all calculations were made using multiple cores by UHEM.

Since there is no meaningful groundtruth data, it is not possible to calculate any accuracy score or confusion matrix. Looking at the previous studies [1], only the image of the estimated data in Figure 2 is available.

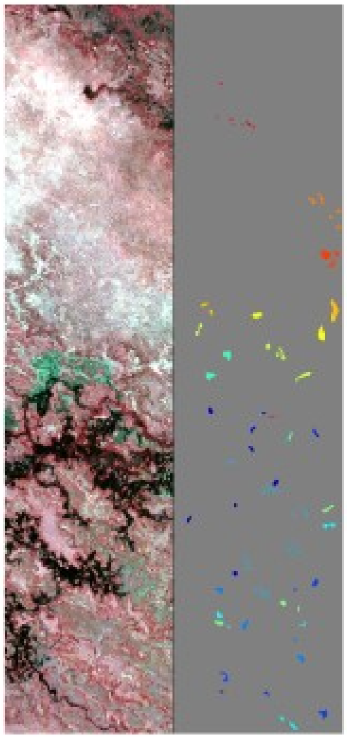

Figure 2: Botswana

First, a color palette was created to visualize the data. Then, all algorithms were tested with both prepared data. The printouts are plotted with the same color palettes. The K-Means algorithm plots in a different color because it is unsupervised.

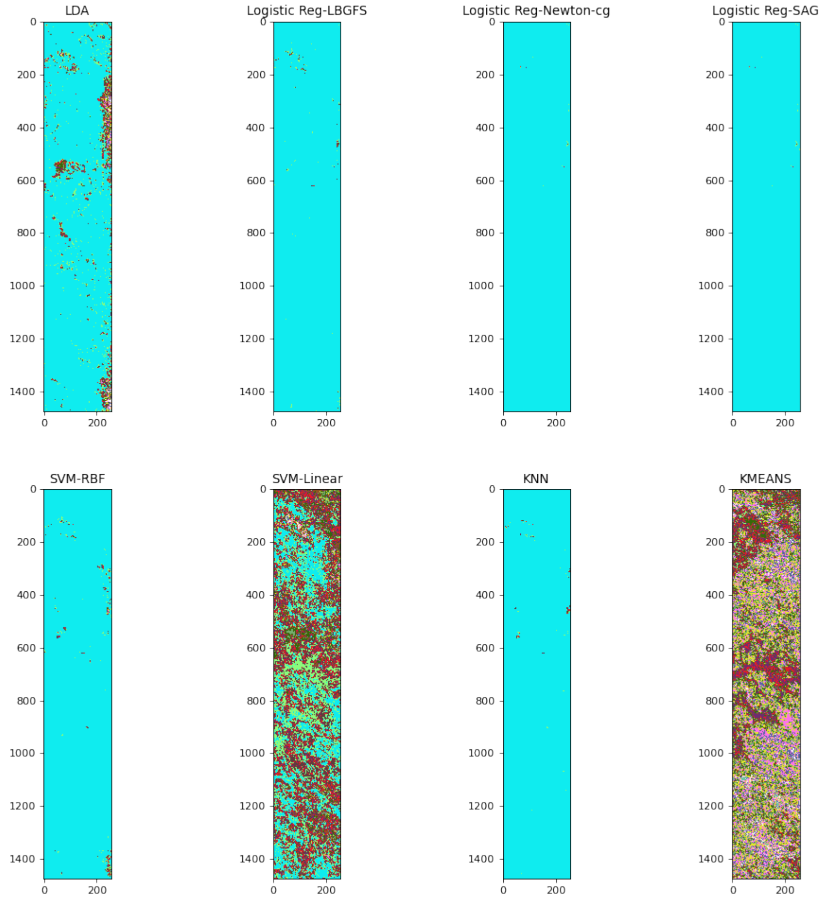

Figure 3: Results for Data Prepared by First Method

As expected, the data set prepared by the first method could not make a satisfactory prediction because it contained too many unlabeled pixels.

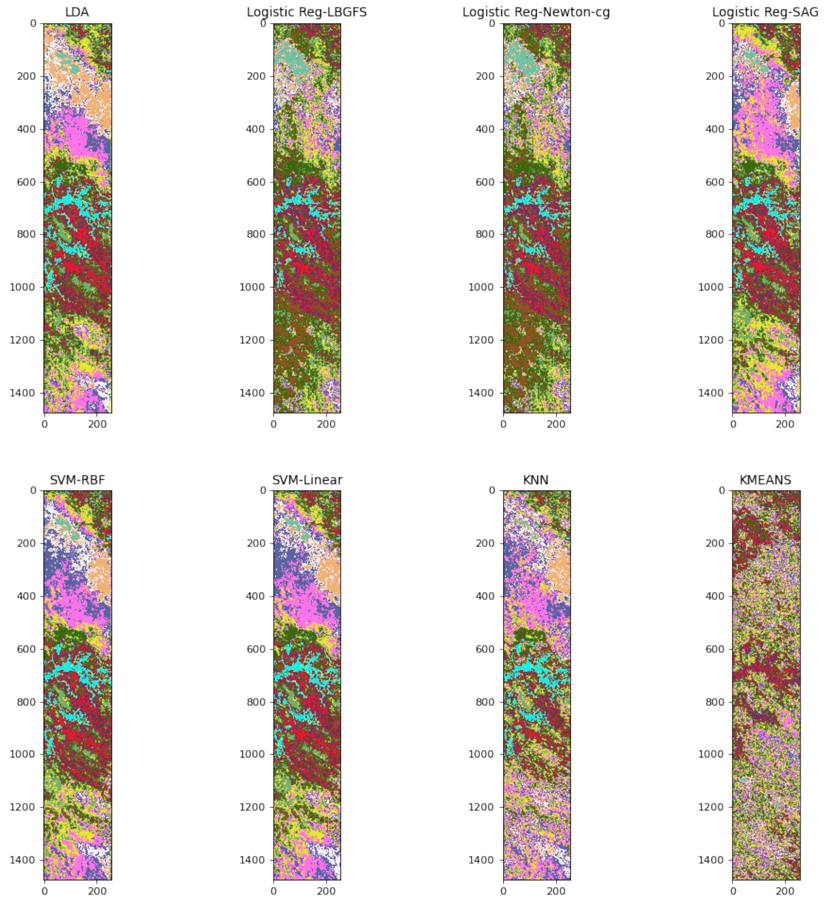

Figure 4: Results for Data Prepared by Second Method

The second method used looks much better than the first. Aqueous regions have different color and light reflectivity were easily detected by almost all algorithms. However, since the refractive indices for other regions were very similar, these regions were more difficult to distinguish. In general, the distinguished regions are similar to each other.

Python Codes: `https://github.com/Lambd4Velorum/StatisticalDataAnalysis`

4 Conclusion

It may be possible to obtain the most realistic result by using different methods such as multi-input CNN method or RGB bands filtering method on the data and evaluating the results of all of them. However, the results obtained at the moment show that the machine learning algorithms are seriously correct for some regions.

References

[1] Alexander Liu, Goo Jun, and Joydeep Ghosh. “Spatially Cost-Sensitive Active Learning.” In: Apr. 2009, pp. 814–825. DOI: 10.1137/1.9781611972795.70.