Project report

Compare the accuracy of UAV images using different classifiers

Task 2

Background

The reason we come up with this idea is based on the requirement of the client which in our case is State Forest Management Centre (RMK). The other partner of this project is university of Tartu department of geography. Our clients are seeking the enhanced classification for monitoring the changes that has been occurred recently in some specific sites near Tartu, Estonia. There is a great deal of importance to assess the quantity of the different types of vegetation. In more detail, it can encompasses finding the best methods to recognize the landscape indices which can be tied to different species of animals. Depends on the client request we can offer them biomass calculation and greenhouse gas emission as additional outcomes.

Business goal

In our project, there is no specific business goal due to the fact that this partnership is between university and forest management. So, our aim is more inclined to academic goals rather than business aims. Our aim background is quiet simple. There is necessity of information to measure what is happening in the ecosystem. Due to drainage, natural habitat and its balance is changed. That being said, couple of restoration projects have been conducted. The point here is investigation of those changes and monitoring if nature is coming back to its natural conditions or not. By only using UAV images and looking at them the estimation of the amount of changes is not feasible. In doing so, we will be training a classifier all over the area and then we are able to monitor the changes. The other important fact about this project and why it has been conducted is that all these changes must be reported to EU institutions, therefore there should be scientific analyzation and conclusion.

Inventory of resources

This project includes experts for taking samples from various vegetation in project sites. Also, the images derived from UAV have been calibrated by photogrammetry methods by specialist.

There is ArcGIS software involved in this project. Other software is Jupiter notebook (python) in order to analyse the data and receiving the results.

Requirement, assumptions, and constraints

These data have been provided to our team from the University of Tartu geography lab and there is no limitation in employ and edit it. The schedule for project termination is mid-December.

Risks and contingencies

The only case can trigger the project to end up in delay completion is not receiving the result that has been asked by client and not fulfilling the basic requirements.

Defining data mining goals

Data-mining goals

The provided data needs to be analysed via different algorithm in order to reach the final goal. The algorithm is going to work on the various features of the provided dataset. It might also need more than two algorithms to reach the preferable accuracy.

Data-mining success criteria

The most important criteria for reaching the goal of the project is to be focused on the analyzation and interpretation of the data which encompasses the answers to our project. That being said the visualization of data and the distribution and finding the important features which have vital effect on the accuracy is in our priority.

Task 3

Data understanding

As it mentioned above, the data has been provided to us from Forest management and university of Tartu geography department. The total amount of the data is roughly around 20G. This project might only zoom into not the whole area and just some small areas due to time schedule. The dataset includes different types of data types such as shape file and calibrated UAV images (raster files), DTM, DSM model of the same dataset is also available. The point shape files include the categories of vegetation which has taken as sample by an expert from the project sites. Some sample of these vegetation has been shown in figure 1.

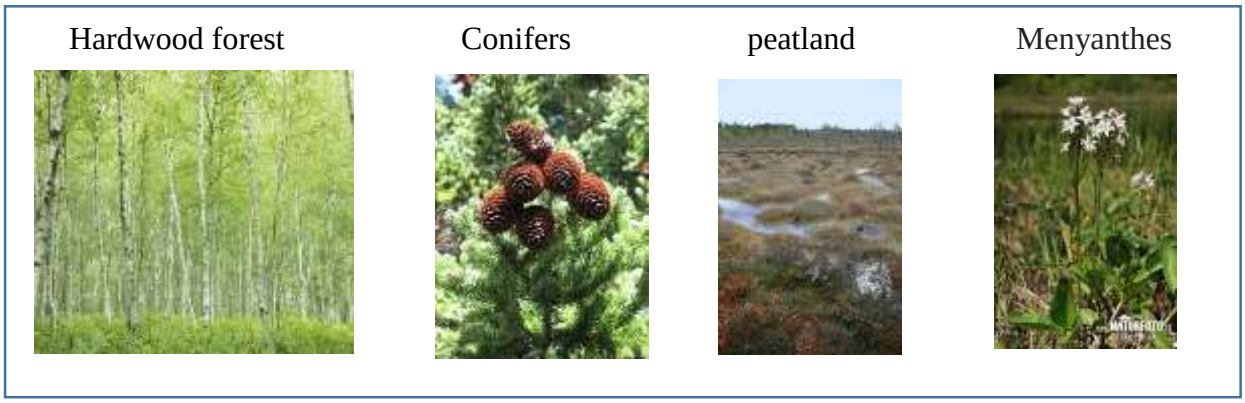

Figure1: different types of vegetation in dataset

Outline data requirements

Here is a list of available data and their format for the project.

1st project site: SOOSAARE

- opialad_Soosaare_180529__modif_180703_punkt (.shp)
- Laukasoo_180607_L1_stacked_data.tif (raster file,stacked RGB)
- Laukasoo_180607_L1_canopy_h.tif (raster file,Canopy Height)

Verify data availability

The data has been handed to our team project on mid-November and the data distribution has been visualized in Jupiter notebook.

Define selection criteria

First of all, it is necessary to just look at the created DataFrame from the point shape file of vegetation samples and raster image of the project site Figure2. The number of columns and features for define training is apparent. By looking through the dataset, we come up with various ideas to go step by step in our project planning section. The dataset including the kr-id which it has given each vegetation type a special code and these codes for our project is specified in the (kr-id column) of the opialad_Soosaare_180529__modif_180703_punkt (.shp).

	kl_id	geometry	band_1	band_2	band_3	band_4	band_5	band_6	band_7	band_8	band_9
0	3	POINT (610639.1267079123 6496934.632328085)	1273.0	2169.0	1560.0	5096.0	9935.0	57.341755	11.574537	-0.728578	-0.531250
1	3	POINT (610589.044893882 6496933.520281457)	1235.0	1839.0	1524.0	4225.0	8060.0	57.003460	10.797261	-0.681970	-0.469821
2	3	POINT (610556.2395184744 6496932.487666734)	620.0	1030.0	855.0	2705.0	6541.0	47.260578	0.618205	-0.768794	-0.519663
3	3	POINT (610561.1642963806 6496916.760150196)	843.0	1560.0	985.0	3511.0	7927.0	56.456062	10.077833	-0.778950	-0.561833
4	3	POINT (610627.748087992 6496941.443613655)	1007.0	1637.0	1220.0	3829.0	8910.0	54.656975	8.553057	-0.759131	-0.516736

Figure 2: project DataFrame

The most important feature of this dataset criteria selection is kr-id column because our training and test splits is literally based on this column. There are other important criteria that has direct effect on the accuracy of our selected algorithm. In our case it might be around 9 criteria or 9 bands of the raster file.

Describing the data

It is apparent that before starting to analyse the available data in order to become familiar with dataset is to know what it encompasses and what needs to be deleted and is not applicable for our aim which it will trigger to waste less time on it. Our provided data has shown a great deal of sustainability in order to fulfil our requirements for next part of the project which is employing the data for classification and reaching appropriate accuracy. As it has mentioned in above sections the dataset is including the shape files and raster images. For applying the various algorithm on the dataset there is no need for a considerable amount of memory, it may take a bit longer with a small memory capacity. Due to the defined goal the amount of the data is adequate and there is a chance if everything goes well, more data can be inserted into the algorithms for better understanding of the results.

Exploring the data

By looking closely to the data, range of variables have been distinguished and showed in figure 3. It was mentioned before, that the most important criteria for training the data is the labels (kr-id) column that encompasses the various type of vegetation. The distribution of the vegetation will make it clear for our client what the provided dataset encompasses in first place.

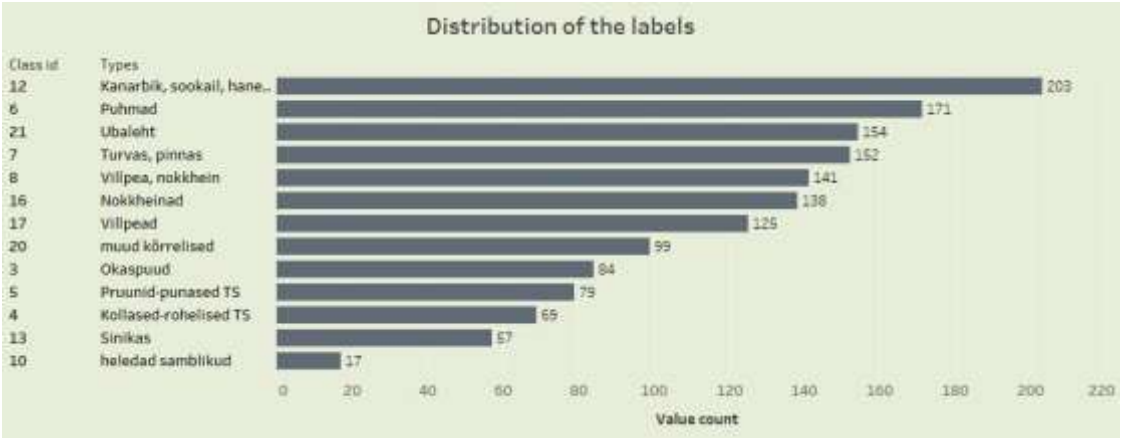

Figure3: Distribution of the labels

There are also some problems with the dataset which it can prone to disrupting problems in the analyzation part. In doing so the data cleaning part is significantly important before starting to employ the data. One of the main issue we faced up to now is lack of correlation between some features of the dataset. They have been figured out and has been deleted from raw dataset.The important point which has been considered before going to the next step is finding the not available and non-values of the dataset. In figure 4, the result of analyzation is apparent that there is no non value and not available data in the selected dataset.

```
Data columns (total 11 columns):
kl_id       1489 non-null int64
geometry    1489 non-null object
band_1      1489 non-null float64
band_2      1489 non-null float64
band_3      1489 non-null float64
band_4      1489 non-null float64
band_5      1489 non-null float64
band_6      1489 non-null float64
band_7      1489 non-null float64
band_8      1489 non-null float64
band_9      1489 non-null float64
dtypes: float64(9), int64(1), object(1)
memory usage: 128.0+ KB

<matplotlib.axes._subplots.AxesSubplot at 0x19313c68320>
```

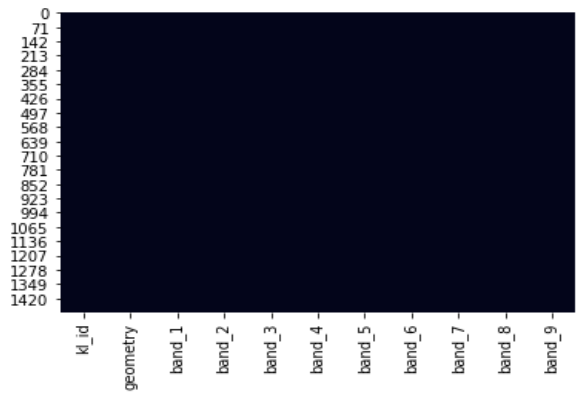

Figure4: Finding the null values in datasets

Verifying the quality of the data

It can be confirmed that the proper amount and quality of the data is provided by the client to reach the required results. There is no lack of data. In addition to the necessary data there are some alternative data in order to compare the result of our methodology on other part of the project sites.

Task4

Producing your project plan

Our project plan can be summarized in more than 5 steps up to now. First of all, it is important to visualize the data to see how features are distributed and to distinguish which of them have more effective role in our final goal and which doesn't. Also it can give us information related to the correlation between the features of the datasets. In doing so it is easier for us to find the appropriate algorithm. In second step, based on the fact that there is a need to clean the dataset before doing analyzation which it will make it easier and faster and eventually is easier to interpret it. The third and fourth part are interrelated to each other. By applying the first classification method Random Forest and receiving the accuracy, it is time to plot the most important features which will help us to gain the vision of our methodology that if we are going to apply it in another area which features can be more effective and this will inevitably save time and effort. By going through the analyzation part we can apply another machine learning algorithm such as SVM and the following step is applying another machine learning algorithm which it definitely depends on the data visualization. It is important to mention that in each part for applying the classification, rescaling or normalizing the data has been done. In addition, the analyzation step, the conclusion part will compare the various methods on classification of the data and the optimal option will be chosen as a result for further steps.The visualization of the classification also can be shown in arcmap.

Initial assessment of tools and techniques

The most conspicuous method for this project is using the machine learning algorithm and data visualization.

Our team project consists of two team members

Marjansadat barekaty

Jeonghwan choi

The estimated time spending for completing this project for both member is 30hrs.