University of Jean Monnet

Course: Data Mining and Knowledge Discover

Project

FOREST COVER TYPE CLASSIFICATION

Author:
Tsolakidis Dimitrios

Student Number:
16007165

March 24, 2017

1 Problem Understanding

Natural resource managers responsible for developing ecosystem management strategies require basic descriptive information including inventory data for forested lands to support their decision-making processes. However, managers generally do not have this type of data for in-holdings or neighboring lands that are outside their immediate jurisdiction. One method of obtaining this information is through the use of predictive models.
The task is to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (as opposed to remotely sensed data). The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS.

2 Data understanding

The data consist of 55 attributes. All the attributes are quantitative and the last attribute is the class labels corresponding to each cover type. The Wilderness Area and the Soil Types attribute consists of 4 and 40 binary columns respectively.

Table 1: Attribute Information

Name	Data Type	Measurement	Description
Elevation	quantitative	meters	Elevation in meters
Aspect	quantitative	azimuth	Aspect in degrees azimuth
Slope	quantitative	degrees	Slope in degrees
Horizontal_Distance_To_Hydrology	quantitative	meters	Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology	quantitative	meters	Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways	quantitative	meters	Horz Dist to nearest roadway
Hillshade_9am	quantitative	0 to 255 index	Hillshade index at 9am, summer solstice
Hillshade_Noon	quantitative	0 to 255 index	Hillshade index at noon, summer soltice
Hillshade_3pm	quantitative	0 to 255 index	Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points	quantitative	meters	Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns)	qualitative	0 (absence) or 1 (presence)	Wilderness area designation
Soil_Type (40 binary columns)	qualitative	0 (absence) or 1 (presence)	Soil Type designation
Cover_Type (7 types)	integer	1 to 7	Forest Cover Type designation

There are four different wilderness areas listed belows, each one represented by a binary row of size 4.

Table 2: Attribute Information

Wilderness Area
1 – Rawah
2 – Neota W
3 – Comanche Peak
4 – Cache la Poudre

Seven different cover types:

Table 3: Attribute Information

Wilderness Area
1 – Spruce/Fir
2 – Lodgepole Pine
3 – Ponderosa Pine
4 – Cottonwood/Willow
5 – Aspen
6 – Douglas-fir
7 – Krummholz

3 Data Preperation

The dataset consists of 581012 observations along with 55 attributes, in which one of them is the class label vector. There are no missing values in the dataset, so we do not have to bother with this. A good idea is to convert the binary vectors which correspond to wilderness areas and soil types by a numerical value. With this method the dimensionality of the the dataset is reduced.
The 4 different wilderness areas can be represented by numbers from 1 to 4, corresponding to each wilderness area (Rawah = 1, Neota = 2, Comanche_Peak = 3, Cache_la_Poudre = 4). The same method is also applied for the soil type attribute. After that, the binary columns in the dataset are removed.
A histogram is produced in order to observe the distribution of cover types:

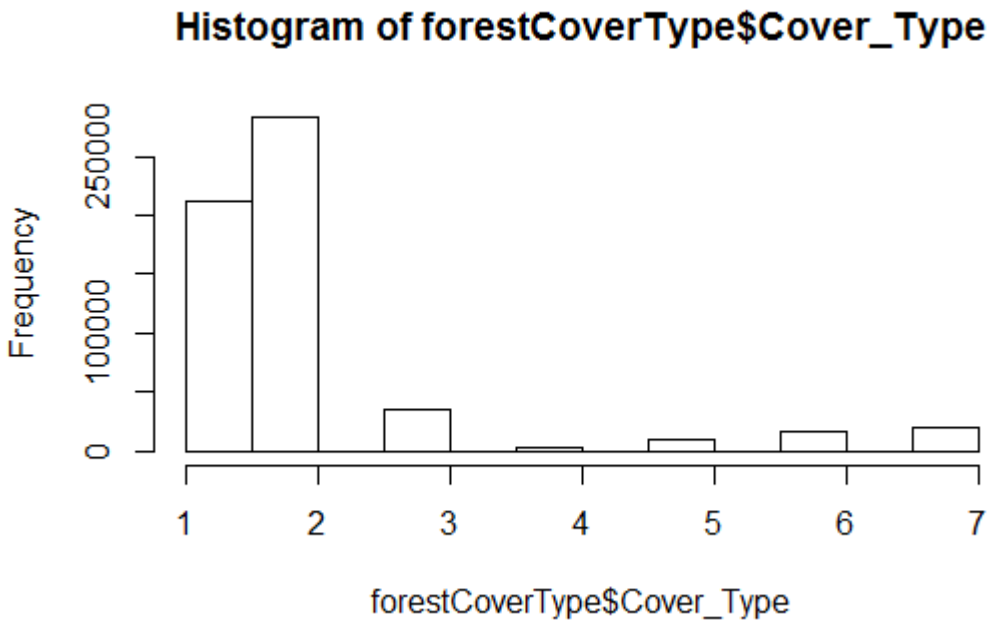

Figure 1: Cover Type Distribution

We observe that the classes are way imbalanced and class 1 (Spruce/Fir), and Class 2 (Lodgepole Pine) are dominating.

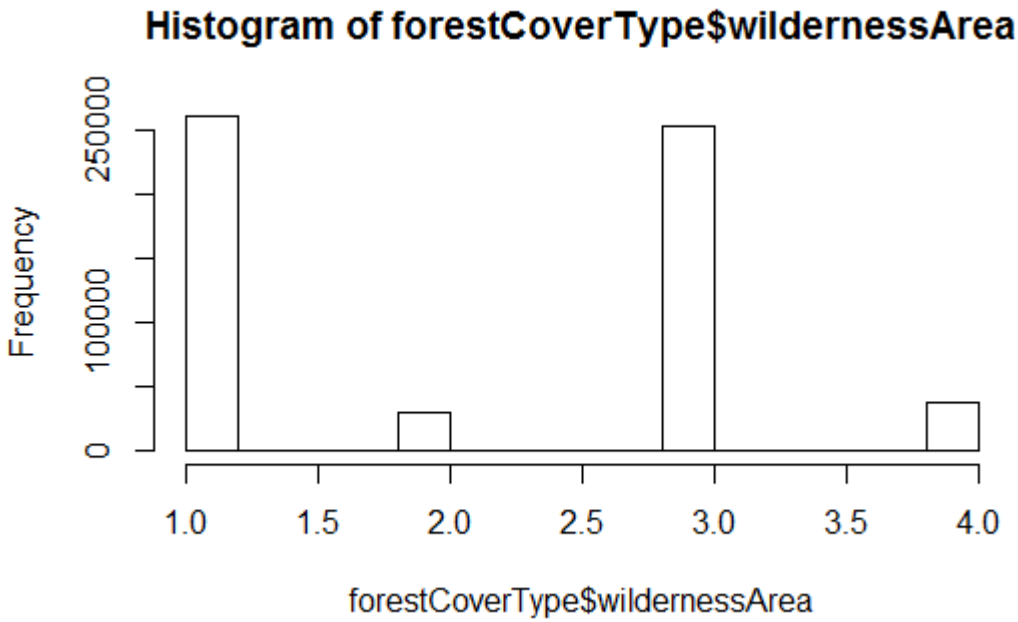

Figure 2: Wilderness Areas Distribution

The most cover types exist at Rawah and Comanche_Peak wilderness areas, whereas much fewer belongs to Neota and Cache_la_Poudre.
Due to the fact that the dataset has more than a million observations and considering the unbalanced dataset, we selected 2000 observations from each class, leading to a subset with 14000 observations. Next step is to see the correlation between the variables.

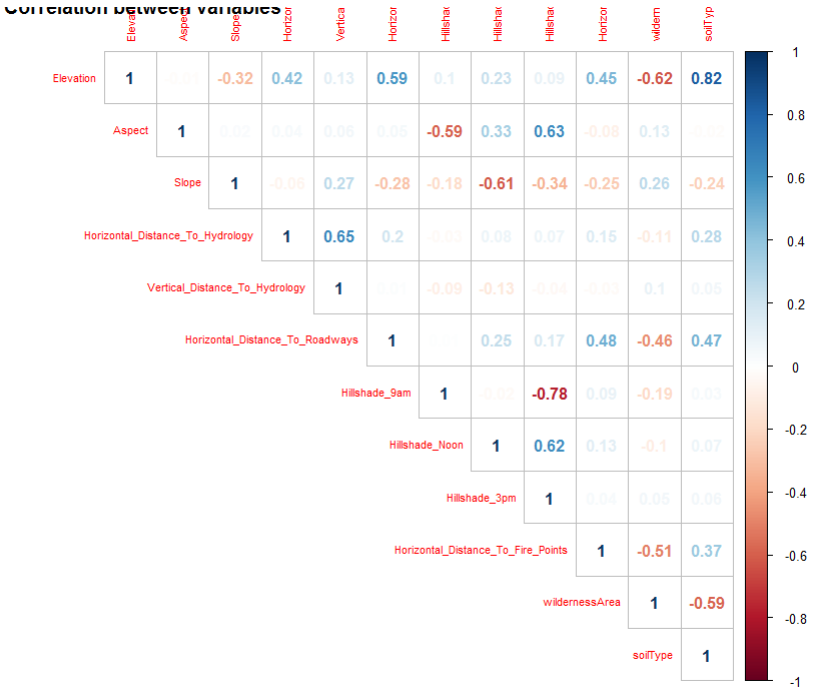

Figure 3: Correlation between variables

High correlation between Elevation and Soil_type at 0.83. Also there exists a correlation between Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology at 0.65 and a high negative of -.78 between Hillshad_9am and Hillshade_3pm. Furthermore, data normalization was employed, by changing the range of the values at [0, 1] and last feature selection was conducted using random forests to take into account only the most important attributes.

4 Modelling

First the data was split by selecting 70% of each class observations as training set and 30% as testing set. Three different supervised algorithms were employed; SVM (Linear and RBF kernel), KNN and Naive Bayes.
A 5-Fold cross-validation was used taking into consideration the training data in order to tune the parameters of SVM and discovering the best parameters of the training model. Similar, cross-validation was used in Naive Bayes and in KNN in order to find out the best k.

5 Evaluation

The testing data was used in order to see the accuracy of the three different models. The accuracy of each classifier is shown below:

Table 4: Classification Results

Classifier	Accuracy
SVM (Linear Kernel)	68.6%
SVM (RBF Kernel)	83.52%
KNN (best K = 5)	77.75%
Naive Bayes	65.89%

Using SVM classifier with RBF kernel, yields a high accuracy and much higher than the other methods.