STAT 6500

Statistical Machine Learning

Land Use Cover

Kendall Byrd - Atitarn Dechasuravanit - Alexys Rodriguez - Abdulaziz Alsugair

Project Proposal

Spring 2022
Wednesday, March 2

Contents

1 Introduction

It is often stated that currently the world is in the era of 'Big Data'. This is because there is an overwhelming amount of data generated everyday by both single individuals and large corporations. For example, there are 277,000 tweets per minute, 2 million queries are searched on Google every minute, 72 hours of video are uploaded to YouTube every minute, 100 million emails are sent, and more than 570 websites are created every minute Gaurav et al. (2018). The massive amounts of data being generated everyday provide an ocean of information explaining individuals' habits, corporation logistics, global trends, and much more. Naturally, many individuals and corporations wanted methods to analyze and handle these large amounts of unorthodox data. This desire led to new discoveries in statistics and paved the way for a technique referred to as statistical machine learning or simply, machine learning.

Machine learning is the study of computer algorithms that can improve or 'learn' automatically by using information from past experiences and using data. Thanks to advancements in computational power, large datasets can be analyzed, and important information extracted. This can provide powerful insight for many areas of research. In fact, machine learning techniques can allow individuals to predict the outcome of events before they happen. Some applications of machine learning are autonomous vehicles, voice recognition, 3-D modeling, and image information extraction. For this study, we propose implementing machine learning techniques on satellite imagery to determine the effect scale has on accurately classifying features extracted from the images. The dataset that will be analyzed in this project is the Urban Land Cover Data Set. This data can be found in the UCI Machine Learning Repository and was originally sourced by Brian Johnson, who is a Research Manager at the Institute for Global Environmental Strategies.

The Urban Land Cover Data Set is a multivariate data set with dimensions of 168 rows and 148 columns. It has twenty-two attributes, that are repeated for seven different coarser scales. This data set contains training and testing data for classifying a high-resolution aerial image into nine classes (target classification variable) of urban land cover. The nine land cover classes are concrete, trees, soil, grass, buildings, cars, asphalt, pools, and shadows. There are a low number of training samples for each class (14-30) and a high number of classification variables (148), so testing different feature selection methods will be interesting. The testing data set was generated from random sampling of the image. All attribute abbreviations and brief explanations can be seen in the Table 1.

Table 1: Variables Description

Class..Land.cover.class..nominal.	SD_R..Standard.deviation.of.Red..texture.variable.
BrdIndx: Border Index (shape variable)	SD_NIR: Standard deviation of Near Infrared (texture variable)
Area: Area in m2 (size variable)	LW: Length/Width (shape variable)
Round: Roundness (shape variable)	GLCM1: Gray-Level Co-occurrence Matrix (texture variable)
Bright: Brightness (spectral variable)	Rect: Rectangularity (shape variable)
Compact: Compactness (shape variable)	GLCM2: Another Gray-Level Co-occurrence Matrix attribute (texture variable)
ShpIndx: Shape Index (shape variable)	Dens: Density (shape variable)
Mean_G: Green (spectral variable)	Assym: Assymetry (shape variable)
Mean_R: Red (spectral variable)	NDVI: Normalized Difference Vegetation Index (spectral variable)
Mean_NIR: Near Infrared (spectral variable)	BordLngth: Border Length (shape variable)
SD_G: Standard deviation of Green (texture variable)	GLCM3: Another Gray-Level Co-occurrence Matrix attribute (texture variable)

2 EDA

From the source Johnson (2018) the dataset is divided into training (168 instances) and testing (507 instances). The number of attributes for both datasets is 148. The first attribute is `class` which contains the target (y) variable and is detailed (training and testing) in Table 2 for training data.

Table 2: Target Attribute 'class'

	asphalt	building	car	concrete	grass	pool	shadow	soil	tree
trainClass	14	25	15	23	29	15	16	14	17
testClass	45	97	21	93	83	14	45	20	89

The same group of variables is repeated for different image segmentation scales (40, 60, . . .), the Table 3 shows a summary of the descriptive statistics for the original feature set applied to the images without any scaling. In addition, scatter-plots, box-plots and correlations values is displayed in Figure **??**.

As the main purpose of this study is to analyze the effect of the scale in the prediction, the Figures 2 and 3 show detailed comparison of feature correlation inter and between different scales.

Table 3: Descriptive Statistics

	var	mean	sd	se	md	range	iqr	skew
4	BrdIndx	2.01	0.63	0.05	1.92	3.19 (1-4.19)	0.84	0.70
1	Area	565.87	679.85	52.45	315.00	3649 (10-3659)	489.00	2.68
17	Round	1.13	0.49	0.04	1.08	2.87 (0.02-2.89)	0.62	0.49
5	Bright	165.57	61.88	4.77	164.49	207.07 (37.67-244.74)	87.92	-0.52
6	Compact	2.08	0.70	0.05	1.94	3.7 (1-4.7)	0.91	1.04
21	ShpIndx	2.23	0.70	0.05	2.13	3.24 (1.06-4.3)	0.98	0.58
12	Mean_G	161.58	63.41	4.89	187.56	215.67 (30.68-246.35)	119.90	-0.77
14	Mean_R	163.67	71.31	5.50	160.62	220.87 (32.21-253.08)	133.63	-0.20
13	Mean_NIR	171.46	67.97	5.24	178.34	213.2 (40.12-253.32)	115.84	-0.36
18	SD_G	10.13	5.18	0.40	8.01	32.07 (4.33-36.4)	4.73	2.18
20	SD_R	9.35	5.00	0.39	7.93	34.23 (3.22-37.45)	4.72	2.47
19	SD_NIR	9.31	4.96	0.38	7.77	33.13 (2.72-35.85)	4.28	2.34
11	LW	2.21	1.76	0.14	1.79	15.23 (1-16.23)	1.02	4.89
8	GLCM1	0.54	0.14	0.01	0.54	0.76 (0.09-0.85)	0.19	-0.51
16	Rect	0.75	0.13	0.01	0.78	0.78 (0.22-1)	0.17	-0.97
9	GLCM2	6.47	0.43	0.03	6.51	3.03 (4.34-7.37)	0.44	-1.39
7	Dens	1.65	0.32	0.02	1.64	1.68 (0.62-2.3)	0.42	-0.45
2	Assym	0.58	0.24	0.02	0.62	0.98 (0.02-1)	0.36	-0.31
15	NDVI	0.00	0.18	0.01	-0.06	0.75 (-0.36-0.39)	0.20	0.60
3	BordLngth	188.11	108.43	8.37	176.00	546 (14-560)	154.00	0.79
10	GLCM3	3064.53	940.01	72.52	2978.36	6766.83 (1225.78-7992.61)	1089.03	1.45

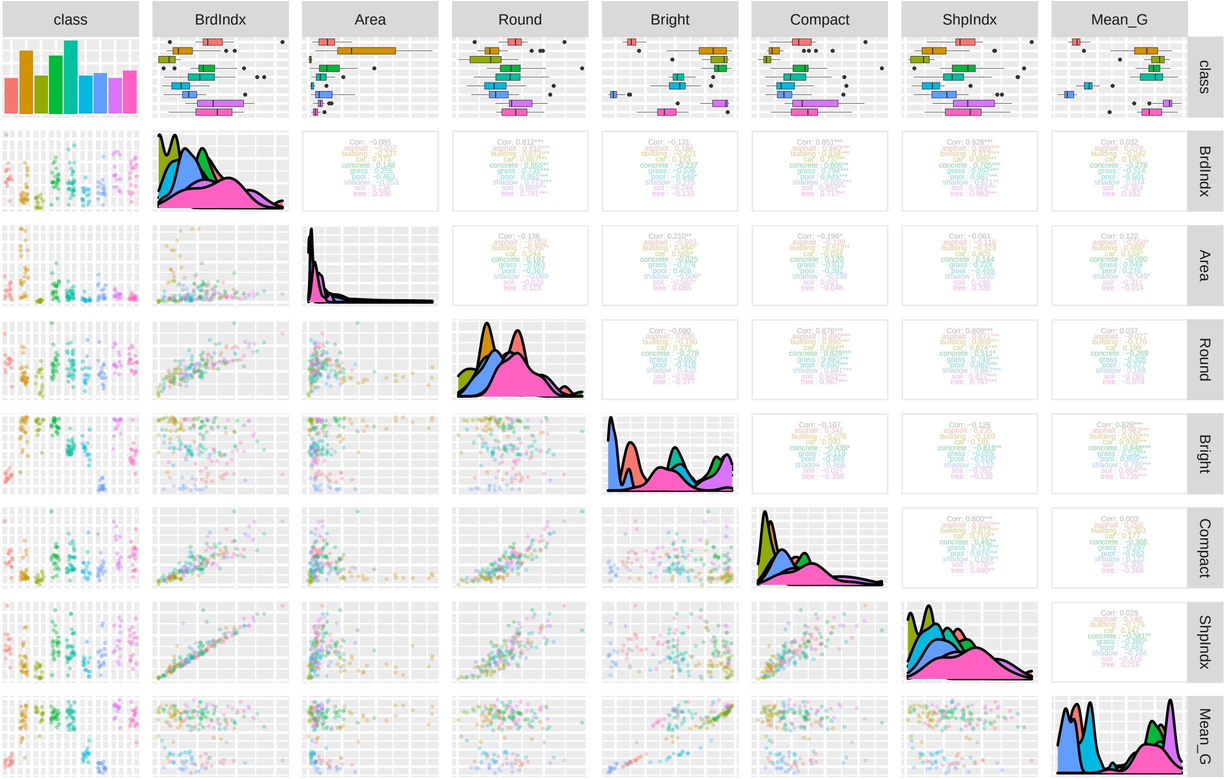

Figure 1: Desciptive graphics for main set of features (not scaling)

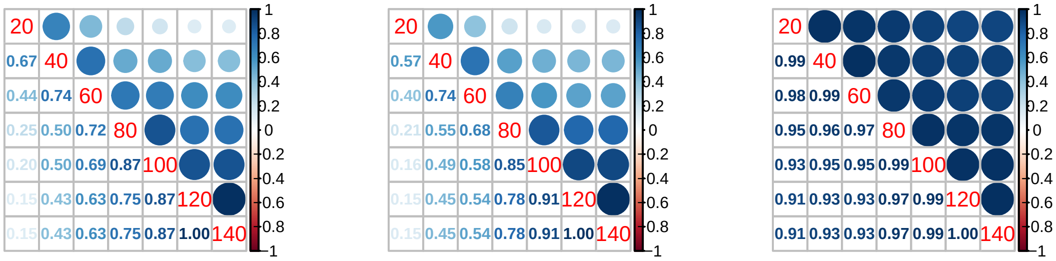

Figure 2: Multicolinearity. Features (Area, Round, Brigh) at different scales

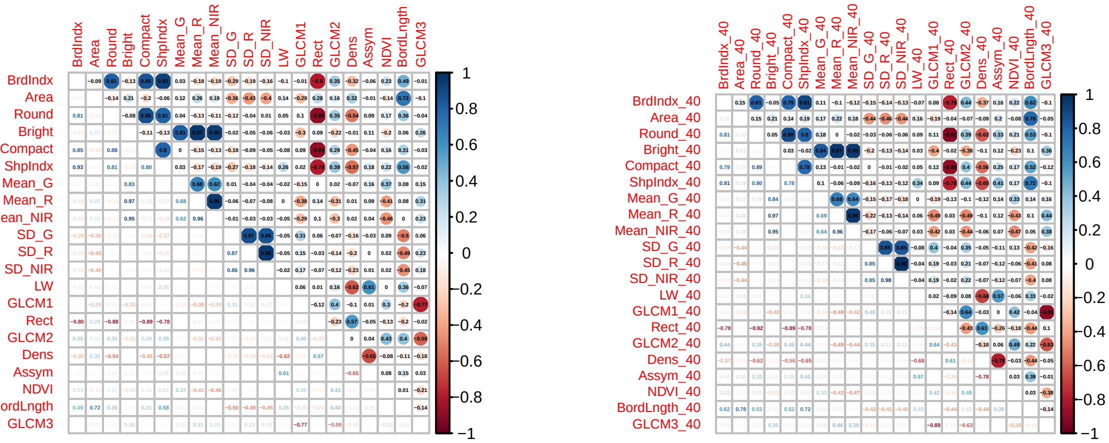

Figure 3: Multicolinearity between features at the same scales (left: no scale, right: coarseness level 40)

References

Gaurav, Devottam, Jay Kant Pratap Singh Yadav, Rohit Kumar Kaliyar, and Ayush Goyal. 2018. “An Outline on Big Data and Big Data Analytics.” In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE. https://doi.org/10.1109/icacccn.2018.8748683.

Johnson, Brian. 2018. “Urban Land Cover Data Set.” Kamiyamaguchi, Hayama, Kanagawa,240-0115 Japan: Institute for Global Environmental Strategies; Digital Repository. November 2018. https://archive.ics.uci.edu/ml/datasets/Urban Land Cover#.