Project2: Cloud Image

Tianyu Wu(. . .), Shiqi Liu(2752089)

1 Data Collection and Exploration

(a)

The key purpose of the paper *Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data With Case Studies* is to build an algorithm in order to ascertain whether a pixel of satellite images of Arctic is covered by cloud or not. Since the surfaces of cloud in Arctic are similar to those of Arctic surface, it is challenging to tell the difference between cloud and Arctic surface. The satellite images or the data used by the authors are shot by NASA Terra Satellite. The satellite comprises nine cameras with nine different angles in four spectral bands.

The data are collected from 10 MISR (Multiangle Imaging Spectro Radiometer) orbits of path 26 over the Arctic, northern Greenland, and Baffin Bay. This path is selected due to its richness of surface features. Six data units from each orbit are included in this study, and three of the total 60 units are excluded since the sea ice melts in the summer and affect the MISR operational algorithm. To evaluate the performance of the study, one of the authors hand-labels the image pixels as either clear or cloudy. Around 71.5% (5086002) of the pixels are labeled in total; the others are left unlabeled due to ambiguity.

The authors develop a classifying algorithm using enhances linear correlation matching (ELCM) and quadratic discriminant analysis (QDA) with three features (the linear correlation of radiation measurements from different MISR view directions, the standard deviation of MISR red radiation measurements, and the normalized difference angular index). As a result, the algorithm performs much better than other existing MISR operational algorithms. The study itself is also significant since the whole study only included three features and relatively simple classifying methods to separate clear and cloud regions. Potentially, a more efficient and accurate classifying algorithm for Arctic cloud will eventually enable the scientific community to have more accurate global climate model simulations.

(b)

Percentages of pixels for difference classes for each image are shown in Figure 1. In Image1, the percentage of each class is relatively even. In Image2, the distribution of pixels on three classes is imbalanced. 43.78% of pixels in Image2 are not cloud, while only 17.77% pixels are cloud. In image3, though it has a more imbalanced distribution compared with Image3, most of the imbalance comes from unlabeled class, which is irrelevant to the training process.

image	not cloud	unlabeled	cloud	Total
Image1	37.25%	28.64%	34.11%	100%
Image2	43.78%	38.46%	17.77%	100%
Image3	29.29%	52.27%	18.44%	100%

Figure 1: Percentage of Percentage of not cloud, unlabeled and cloud for each image.

Well-labeled maps are shown in Figure 2. Based on the labeled map, the pixels show a sticky pattern in all three images. Meanwhile, cloud regions and clear regions are separated by unlabeled pixels. Therefore, i.i.d. assumption can't be justified for this data set.

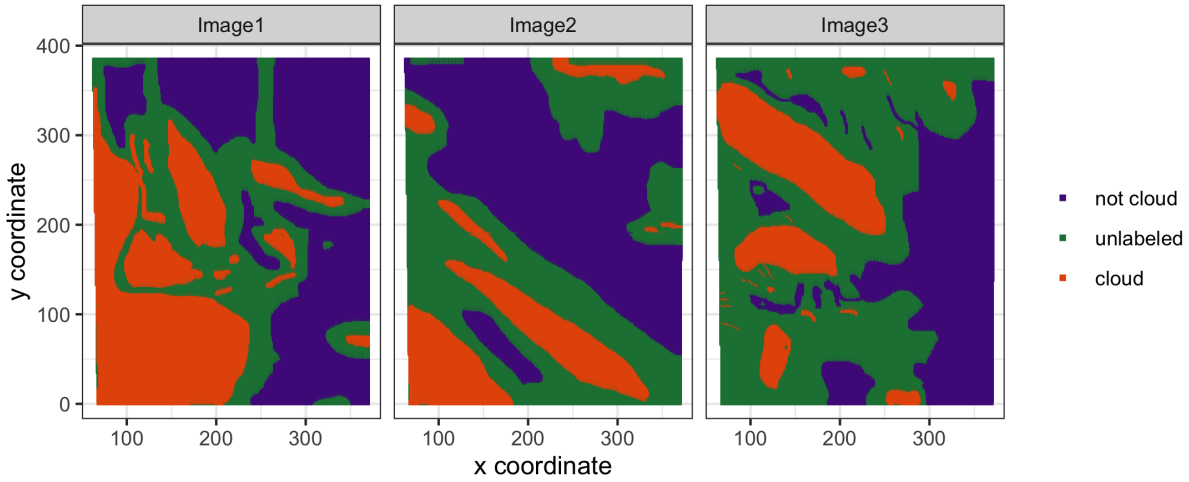

Figure 2: Image1-3 with expect labelled.

(c)

See Figure 3.

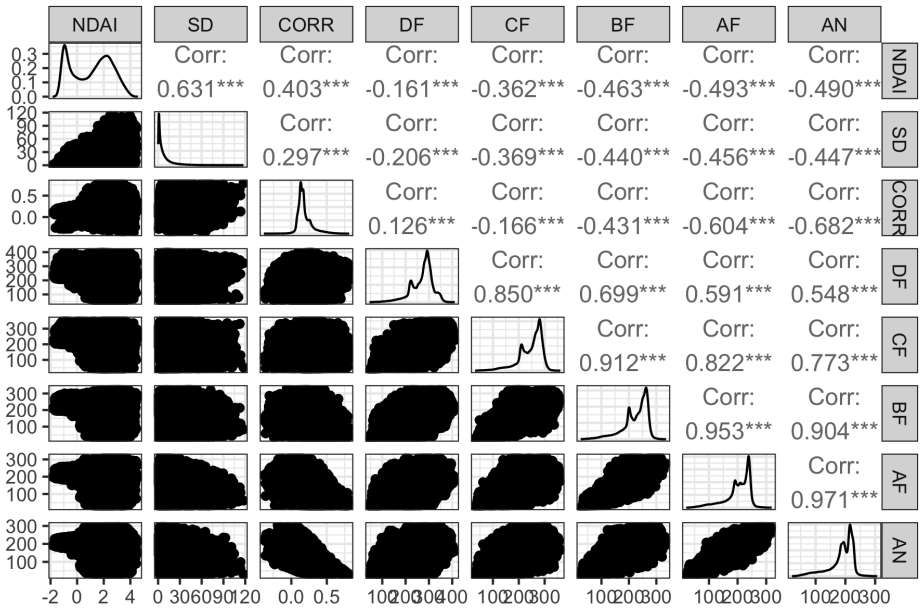

Figure 3: Pairwise relationships between the three features

See Figure 4.

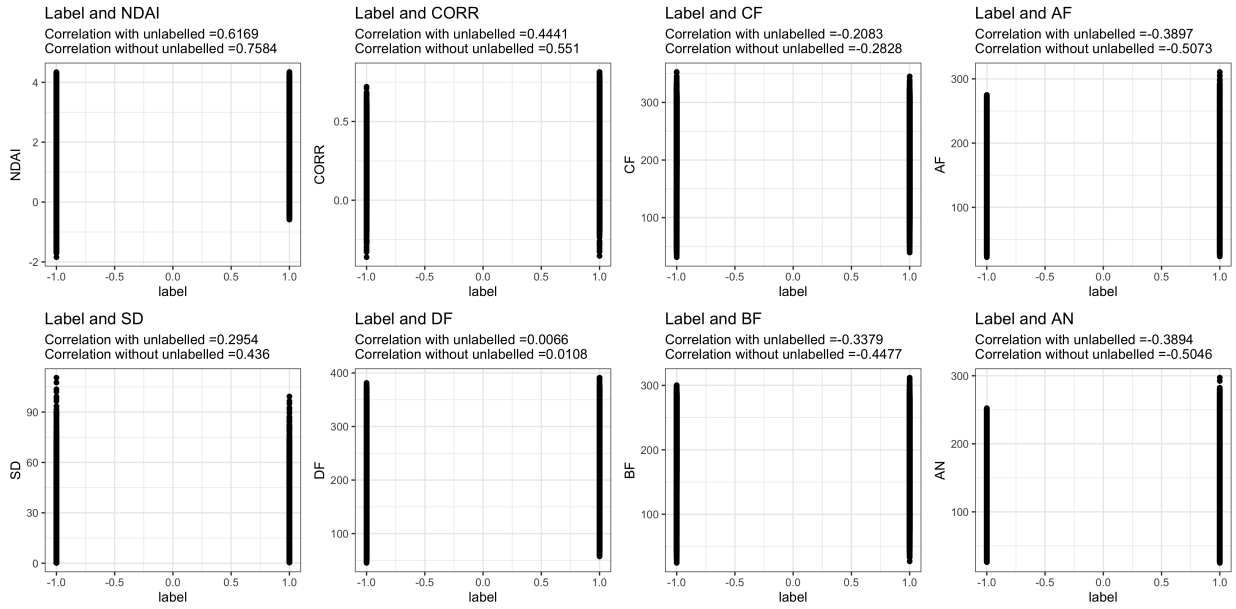

Figure 4: Scattor-plots of label verse each feature, with unlabelled data removed. Correlations of label and each feature with and without unlabeled data.

2 Preparation

(a)

For the three images in Figure 2, each of them may have unique pattern. In case of over-fitting to one particular image, the combination of the three images is necessary before training. We also notice that there are some strong dependencies between one pixel and its neighbors. Therefore, we choose to split the data set by block splitting, to avoid breaking its spatial structure. To be specific, we first divide the data set into numerous 8×8 pixel blocks and for pixels in the same 8×8 block, they are all assigned to one fold randomly chosen from K folds, and the default value of K is 6, and unlabeled points are removed. Also, we don't desire that blocks at same location of different images are in the same fold, so the three data sets are grouped separately before combination. As a result of large number of blocks, percentages of all three labels (in Figure 5) are nearly even among different folds. After all, all blocks in fold 1 are used for testing data, all blocks in fold 2 are used for validation, and the rest are used for training data.

To enlarge the gap between the features of not cloud data and features of cloud data, we drop blocks which are not purely not cloud or cloud. it means that we drop almost all data points around the boundaries (in Figure 2) between different labels, and everything else is the same as the first method.

fold	not cloud	cloud	Total
1	62.48%	37.52%	100%
2	57.17%	42.83%	100%
3	59.10%	40.90%	100%
4	64.20%	35.80%	100%
5	63.35%	36.65%	100%
6	59.72%	40.28%	100%

Figure 5: Percentage of not cloud and cloud for fold1-6.

(b)

data	split method	not cloud	cloud	Total
test	block split	62.48%	37.52%	100%
valid	block split	57.17%	42.83%	100%
test	remove margin	64.79%	35.21%	100%
valid	remove margin	64.15%	35.85%	100%

Figure 6: Percentage of not cloud and cloud for validation set and test set of two split method.

(c)

By performing the logistic regression forward selection on all features (in Figure 7), NDAI, CORR, and SD are three of the "best" features. With adding NDAI, CORR, and SD to the training model, AIC drop significantly.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	208060	278134.6	278136.6
+ NDAI	-1	147904.4698	208059	130230.2	130234.2
+ CORR	-1	9675.0264	208058	120555.1	120561.1
+ SD	-1	4225.1203	208057	116330.0	116338.0
+ DF	-1	2551.9709	208056	113778.0	113788.0
+ BF	-1	1703.2342	208055	112074.8	112086.8
+ AN	-1	853.0864	208054	111221.7	111235.7
+ CF	-1	153.4786	208053	111068.2	111084.2
+ AF	-1	172.2860	208052	110895.9	110913.9

Figure 7: Result of logistic regression forward selection on all features.

###(d)

The CVmaster function is in CVmaster.R.

3 Modeling

###(a)