STA 521 Project 2 Cloud Data Report

Yicheng Shen (yicheng.shen@duke.edu) & Yunhong Bao (yunhong.bao@duke.edu)

25 November, 2022

1 Data Collection and Exploration

1.1 Background & Motivation

With the global climates getting increasingly extreme, humans are making the best use of sciences and technologies to understand the environment, especially in the Arctic region. The detection of clouds in satellite images has become an important task, as cloud coverage is closely related to the surface air temperatures and atmospheric carbon dioxide levels. Yet it is a challenging problem since clouds are similar on snow- and ice-covered surfaces. In this study, we are going to examine various classification methods, build reliable models that distinguish the presence of cloud from Arctic satellite images using available features and evaluate our models' performance.

The data is obtained from a study by Yu et al. (2008). This team of researchers collected the data via the camera of Multiangle Imaging SpectroRadiometer (MISR) launched by the NASA. The data are in the forms of image pixels, with each MISR pixel covering a 275 m by 275 m region on the ground. Since standard classification frameworks of clouds were not readily applicable, their goal was also to build operational cloud detection algorithms that can efficiently process the massive MISR data set one data unit at a time without requiring human intervention or expert labeling.

Yu et al. (2008) proposed two algorithms, an enhanced linear correlation matching (ELCM) algorithm based on thresholding the three features with values either fixed or data-adaptive, and an ELCM algorithm with Fisher's quadratic discriminant analysis (ELCM-QDA).

Their results suggest that both proposed algorithms are computationally efficient for operational processing of the massive MISR data sets. The accuracy and coverage of ELCM are better and more informative compared with conventional MISR operational algorithms. This findings provides important implications and foundations for further analysis of MISR data.

1.2 Data Description

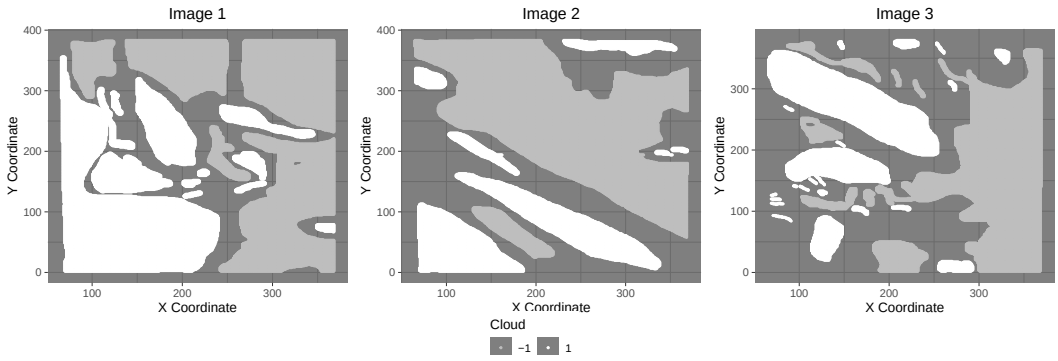

Figure 1: Maps of three images with expert labels. White represents high confidence of cloud; gray represents high confidence of clear; and dark represents unlabeled pixels.

In this study, we primarily focus on three of the MISR images. These three images contain 115110, 115229 and 115217 pixels respectively. However, not all pixels are labeled with confident experts' classification. As shown in in Figure 1, there are considerable portions of images not labeled. We notice the pattern that usually experts have difficulties distinguishing the areas around what they recognize as clouds. It is understandable that the borderlands between cloudy and clear surfaces are more challenging to determine.

We can also observe that labeled clouds are often clustered in chunks. Therefore, adjacent pixels' labels are not independent and instead seem to have very high positive correlations. For example, if a pixel's neighbors are all labeled as clouds, it is highly likely that it is also labeled as part of the cloud.

In Table 1, we present the percentages of expert labels in each image. Since we have no confident expert opinion on unlabeled pixels, they are viewed as missing / unknown values. After removing unlabeled pixels, we have 82148, 70917 and 54996 pixels in each image, with available information (labels and features).

Table 1: Proportions of Cloudy and Clear Surfaces by Expert Labeling

Cloud Labels	Clear Labels	Unknown	Image
0.3411172	0.3725306	0.2863522	1
0.1776549	0.4377891	0.3845560	2
0.1843825	0.2929429	0.5226746	3

In specific, each pixel has eight features. The first three are physically useful features for characterizing the scattering properties of ice and snow covered surfaces: the correlation (`CORR`) of MISR images of the same scene from different MISR viewing directions, the standard deviation (`SD`) of MISR nadir camera pixel values across a scene, and a normalized difference angular index (`NDAI`) that characterizes the changes in a scene with changes in the MISR view direction.

The latter five are five view zenith angles of the cameras. 70.5° (DF), 60.0° (CF), 45.6° (BF), and 26.1° (AF) in the forward direction and 0.0° (AN) in the nadir direction.

1.3 Explortary Data Analysis

Conducting a EDA on the available features gives a preliminary picture of the relationship within potential predictors as well as with the response. First of all, we notice in Figure 2 that there is positive correlation between the features. The correlation is particularly strong and positive among the five radiance angles. For example, `BF`, `AF` and `AN` are very strongly correlated. We also see positive correlations between `NADI`, `CORR` and `SD`.

However, the correlations between the first three features (`NADI`, `CORR` and `SD`) and five radiance angles (`Df`, `CF`, `BF`, `AF`, `AN`) are mostly negative.

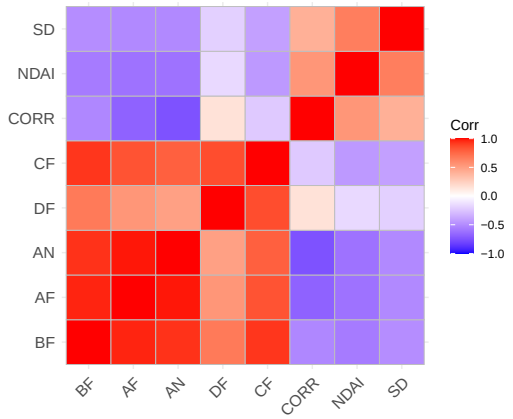

Figure 2: Pair-wise correlation between available features.

Figure 3 shows that for pixels labeled as cloud, their `NADI`, `CORR` and `SD` are likely to be higher than those labeled as no cloud. This distinct pattern provides strong support to use these features as predictors of our classification models. It also seems reasonable to have a log transformation of `SD` so that it is in a similar scale as other features.

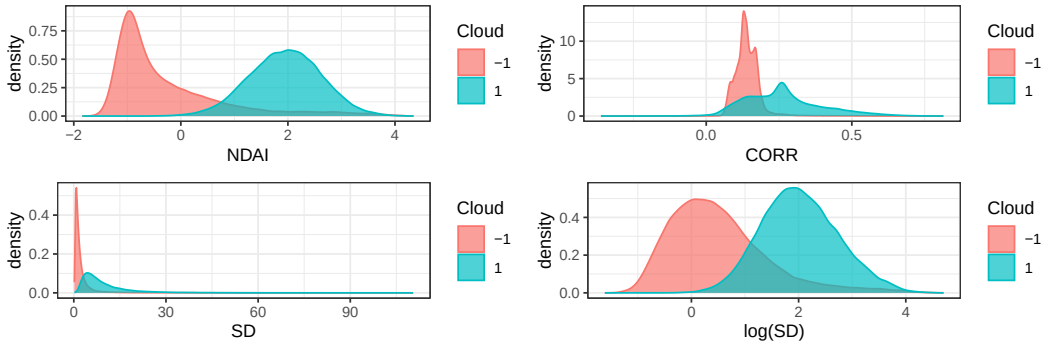

Figure 3: Density distributions of three features that describe cloud and clear pixels.

In terms of other features, we can see that the density distributions of radiance angles of cloud and cloud-free pixels are pretty consistent across angles in Figure 4. The radiance angles of cloud pixels are usually wider and right skewed, whereas the radiance angles of cloud-free pixels are usually higher and distributed in a bimodal shape. The distributions between cloud and cloud-free pixels are not as separable and distinct as the first three features.

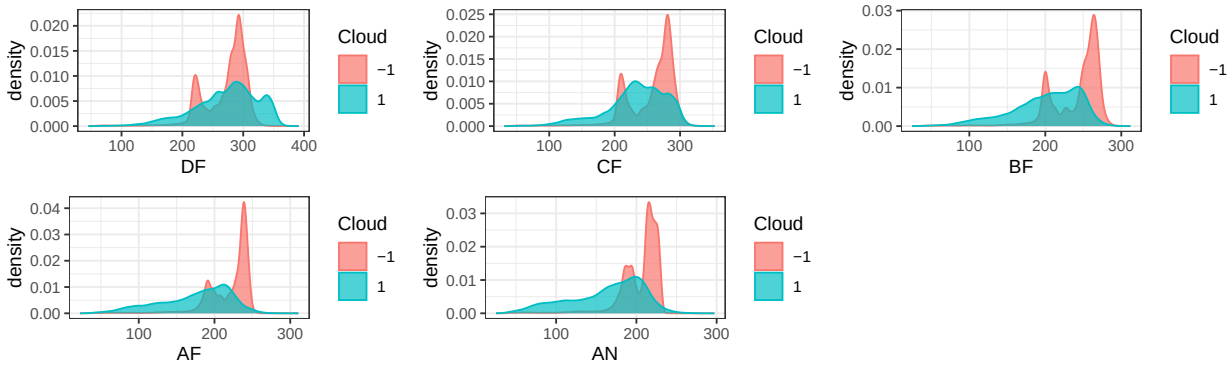

Figure 4: Density distributions of radiance angles from cloud and clear pixels.

2 Preparation

2.1 Data Splitting

Before building models, it is important to prepare the data as training, validation and testing sets so that we can make the best use of the data and evaluate our models. The key idea of our data splitting is to take into account the fact that this data set is not i.i.d. Therefore, we propose the two following ways of dividing data into blocks.

A. Horizontal Cuts: The first method cuts each image horizontally in order to ensure each resulting block has a reasonable portion of clouds and clear surfaces. Basically, each image is cut into five horizontal blocks by evenly dividing Y coordinates (shown in Figure 5), and three of them would be used as training data, the rest two blocks are used as validation and testing respectively. This methods splits the data into 58.59% training data, 19.03% validation data and 22.38% testing data (roughly 3:1:1).

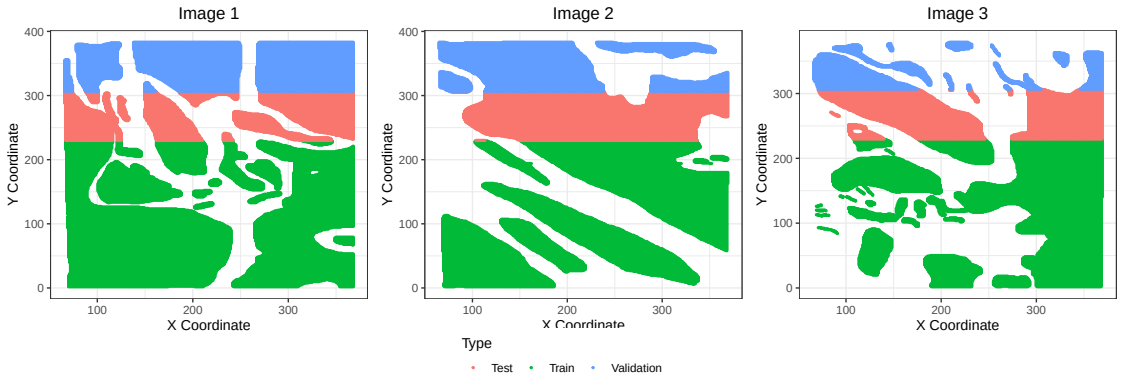

Figure 5: The first data splitting method is to divide each image horizontally.

B. K-means Clusters: The second method of blocked data splitting is to use the K-means algorithm. By selecting a cluster size of five, we can divide each image's datapoints into five distinct groups (according to X-Y coordinates). Again, shown in Figure 6, three of these are used for training data, one is for validation and the last one is for testing. The K-means method splits all datapoints into 60.72% training data, 20.04% validation data and 19.24% testing data.

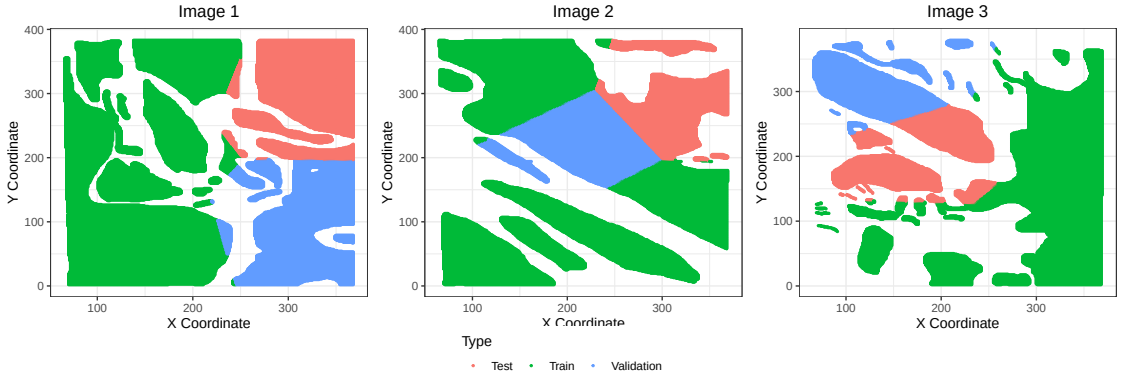

Figure 6: The second data splitting method is to divide based on the K-means algorithm.

The table below describes how much cloud pixels there are in each set after two ways of data splitting. We think they are both reasonable in terms of having both cloud and clear pixels in every subset of data.

Table 2: Proportions of cloud pixels in each set

Type	Cloud Prop	Method
Test	0.3384662	Horizontal Cuts
Train	0.4778361	Horizontal Cuts
Validation	0.1760129	Horizontal Cuts
Test	0.3296281	K-means
Train	0.4479412	K-means
Validation	0.2684817	K-means

2.2 Baseline Accuracy

We can examine the accuracy of a trivial classifier which sets all labels to -1 on the validation set and on the test set as shown in the table below. Logically, the accuracy of this trivial classifier depends entirely

on the percentage of cloud free pixels labeled in the data. If the image is mostly cloud free, then labeling all points as -1 would easily achieve a high accuracy.

Table 3: Accuracy of a trivial classifier that sets all labels to -1

Data Type	Accuracy	Method
Validation	0.8239871	Horizontal Cuts
Test	0.6615338	Horizontal Cuts
Validation	0.7315183	K-means
Test	0.6703719	K-means

(c) (First order importance) Assuming the expert labels as the truth, and without using fancy classification methods, suggest three of the "best" features, using quantitative and visual justification. Define your "best" feature criteria clearly. Only the relevant plots are necessary. Be sure to give this careful consideration, as it relates to subsequent problems.

(d) Write a generic cross validation (CV) function CVmaster in R that takes a generic classifier, training features, training labels, number of folds K and a loss function (at least classification accuracy should be there) as inputs and outputs the K-fold CV loss on the training set. Please remember to put it in your github folder in Section 5.

3 Modeling

(a) Try several classification methods and assess their fit using cross-validation (CV). Provide a commentary on the assumptions for the methods you tried and if they are satisfied in this case. Since CV does not have a validation set, you can merge your training and validation set to fit your CV model. Report the accuracies across folds (and not just the average across folds) and the test accuracy. CV-results for both the ways of creating folds (as answered in part 2(a)) should be reported. Provide a brief commentary on the results. Make sure you honestly mention all the classification methods you have tried.

In the modeling section, we examine a number of classification methods using cross-validation, including logistic regression, LDA, QDA, Naive Bayes and boosting trees. The results are shown in the table below.

Table 4: CV Results and Test Accuracy based on Two ways of Data Splitting

classifier	data	fold.1	fold.2	fold.3	fold.4	fold.5	fold.6	fold.7	fold.8	fold.9	fold.10	Cv.mean	Test
Logistic	Horizontal Cut	0.6989	0.7931	0.7025	0.9983	0.7148	0.9810	0.9018	0.9190	0.9984	0.9788	0.8687	0.8705
Logistic	K-means	0.9621	0.9950	0.7803	0.9982	0.7681	0.9983	0.8340	0.6920	0.7516	0.8660	0.8646	0.9065
LDA	Horizontal Cut	0.6830	0.8055	0.7082	0.9981	0.8046	0.9783	0.9012	0.9272	0.9983	0.9780	0.8783	0.8779
LDA	K-means	0.9619	0.9932	0.7916	0.9980	0.8274	0.9964	0.8431	0.6935	0.7418	0.8649	0.8712	0.9089
QDA	Horizontal Cut	0.6891	0.8274	0.6518	0.9989	0.8559	0.9926	0.8818	0.8983	0.9969	0.9863	0.8779	0.8957
QDA	K-means	0.7512	0.6402	0.8823	0.9978	0.8942	0.9990	0.8449	0.9705	0.8197	0.9966	0.8796	0.9031
Naive Bayes	Horizontal Cut	0.5574	0.8075	0.6151	0.9990	0.9580	0.9396	0.8071	0.8654	0.9877	0.9713	0.8508	0.8455
Naive Bayes	K-means	0.9340	0.9204	0.8083	0.9986	0.9618	0.9912	0.7678	0.6008	0.6447	0.8075	0.8435	0.8974
Boosting Trees	Horizontal Cut	0.9958	0.7299	0.9925	0.5908	0.9662	0.9596	0.8636	0.9685	0.9286	0.9981	0.8994	0.9393
Boosting Trees	K-means	0.9573	0.9996	0.9085	0.9852	0.5777	0.9966	0.8511	0.9811	0.8618	0.7360	0.8855	0.9499

(b) Use ROC curves to compare the different methods. Choose a cutoff value and highlight it on the ROC curve. Explain your choice of the cutoff value.

(c) (Bonus) Assess the fit using other relevant metrics. Use quantitative measures and show clean and interpretable figures!

4 Diagnostics

(a) Do an in-depth analysis of a good classification model of your choice by showing some diagnostic plots or information related to convergence or parameter estimation.

(b) For your best classification model(s), do you notice any patterns in the misclassification errors? Again, use quantitative and visual methods of analysis. Do you notice problems in particular regions, or in specific ranges of feature values?

(c) Based on parts 4(a) and 4(b), can you think of a better classifier? How well do you think your model will work on future data without expert labels?

(d) Do your results in parts 4(a) and 4(b) change as you modify the way of splitting the data?

(e) Write a paragraph for your conclusion.

5 Reproducibility

In addition to a writeup of the above results, please submit a zip file containing everything necessary to reproduce your writeup to Gradescope "PROJ2 code". Specifically, imagine that at some point an error is discovered in the three image files, and a future researcher wants to check whether your results hold up with the new, corrected image files. This researcher should be able to easily re-run all your code and produce all your figures and tables. This zip file should contain:

(i) the raw Latex, Rnw, Qmd or Word used to generate your report,

(ii) your R code (with CVmaster function in a separate R file),

(iii) a README.md file describing, in detail, how to reproduce your paper from scratch (assume researcher has access to the images).

6 Reference

Yu, Bin, Tao Shi, Eugene E Clothiaux, and Amy J Braverman. 2008. "Daytime Arctic Cloud Detection Based on Multi-Angle Satellite Data with Case Studies." *Journal of the American Statistical Association* 103 (482): 584–93.