STA 521 Project 2 Cloud Data Report

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

06 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 surface 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 is in the forms of image pixels, with each MISR pixel covers 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 algorithm, 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 further analysis of MISR data.

1.2 Data Description

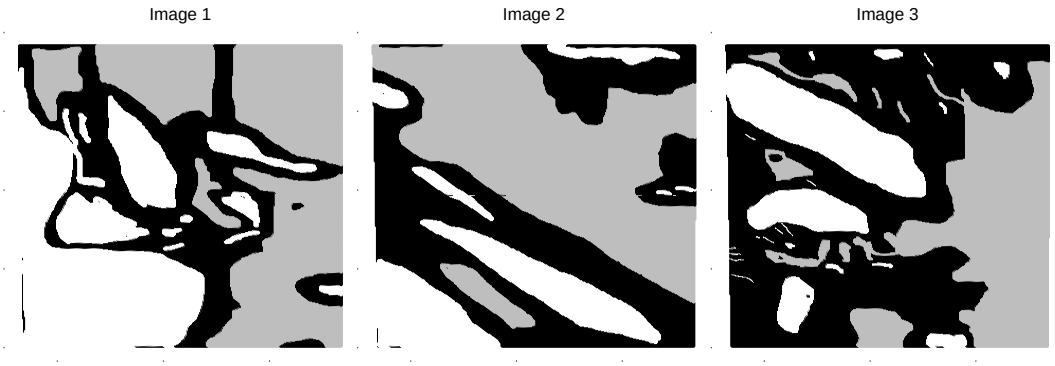

Figure 1: Maps of three images with expert labels. White represents high confidence cloudy; gray, high confidence clear; and black, unlabeled pixels.

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. The borderlands between cloudy and clear surfaces are certainly 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 labelled as clouds, it is highly likely that it is also 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 values. After removing unlabeled pixels, we have 82148, 70917 and 54996 pixels in each image, with available information.

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

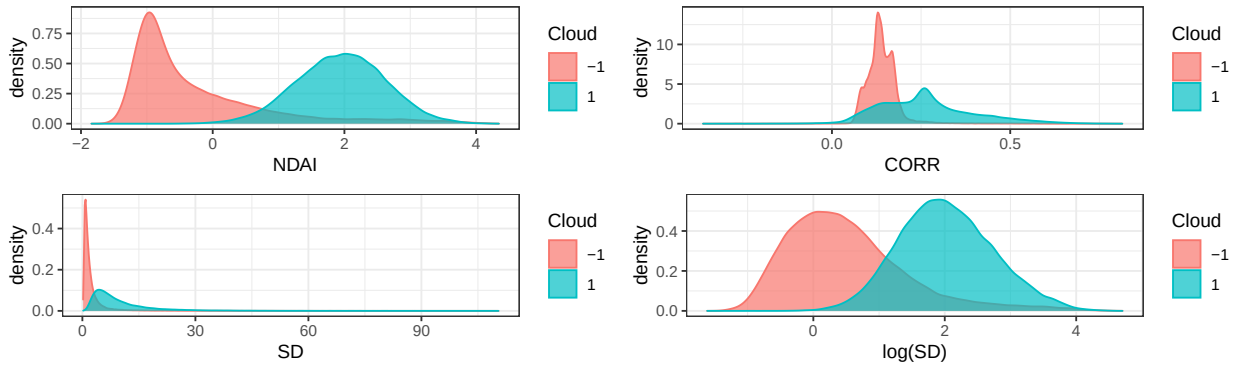

Figure 2: Density distributions of features that describe cloud and clear pixels.

2 Preparation

2.1 Data Spliting

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

Horizontal Cuts: The first method cuts each image horizontally in order to ensure each block has a reasonable portion of clouds and clear surfaces. Bacially, each image is cut into five horizontal blocks,

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 61.27% training data, 18.66% validation data and 20.07% testing data.

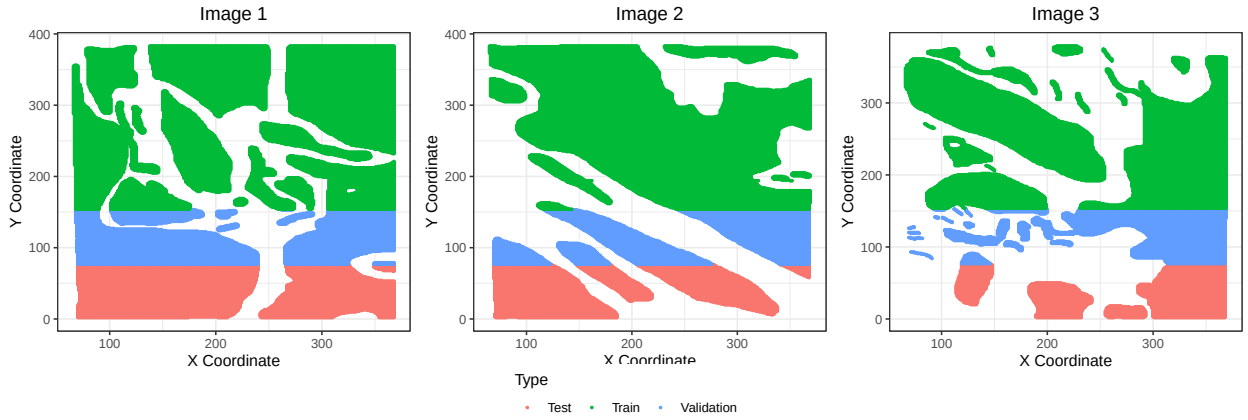

Figure 3: The first data spliting method is to divide each image horizontally.

K-means Clusters: The second method of blocked data spliting 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. Again three of these are used for training data, one is for validation and the last one is for testing. The K-means method splits the data into 60.72% training data, 20.04% validation data and 19.24% testing data.

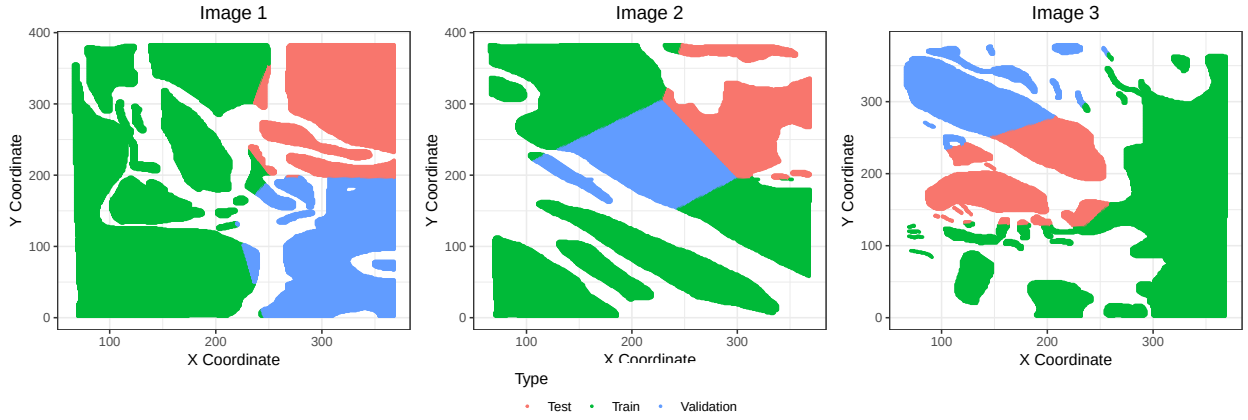

Figure 4: The second data spliting method is to divide based on the K-means algorithm.

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	Perct Cloud	Method
Test	0.6074003	Horizontal Cuts
Train	0.3059055	Horizontal Cuts
Validation	0.4281005	Horizontal Cuts
Test	0.3296281	K-means
Train	0.4479412	K-means
Validation	0.2684817	K-means

(b) (Baseline) Report the accuracy of a trivial classifier which sets all labels to -1 (cloud-free) on the validation set and on the test set. In what scenarios will such a classifier have high average accuracy? Hint: Such a step provides a baseline to ensure that the classification problems at hand is not trivial.

(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.

(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).

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.