Project 2 Cloud Data

Caiyi Deng 3033303379, Winnie Gao 3031884025

April 30, 2019

1   Data Collection and Exploration

a. Summary of Paper

Climate changes has been a popular topic in scientific research. Particularly, the warming Arctic is one of the biggest stories in our times. Global climate models predict that the increasing atmospheric carbon dioxide levels is strongly related to the increasing surface air temperatures in the Arctic, where cloud plays an important role in producing more atmospheric carbon dioxide. In this paper, scientists use the Multiangle Imaging SpectroRadiometer (MISR) imagery to perform a cloud detection to ascertain whether cloud can potentially lead to further warming in the Arctic.

MISR collects a massive amount of data from its nine cameras viewing at a different angel in four spectral. It covers the daylight side of the Earth from the Arctic down to Antarctica in 45 minutes and completes all paths in 16 days of a cycle. Each path is subdivided into blocks, with the block numbers increasing from the North Pole to South Pole, and each complete trip of MISR around the Earth is counted as a unique orbit. However, due to the transmission channel constraints, only the red radiances and all channels from the nadir camera are transmitted at full 275m * 275m resolution. The remaining blue, green and near-infrared radiances from the non-nadir cameras are aggregated to a lower resolution before transmission.

Scientists utilizes correlations in brightness among multiple MISR views of the same scene under cloud-free conditions to model the surface. This new algorithm, enhances linear correlation matching (ELCM), is based on thresholding three features: the correlation (CORR) of MISR images, the standard deviation (SD_An) of the MISR nadir camera pixel values, and a normalized difference angular index (NDAI) to create labels for classification. Then, the resulting labels are used in the second algorithm, ELCM-QDA, to produce more informative probability prediction.

The results suggest that the ELCM algorithm based on the three features outperform those existing algorithms based only on the radiation measurement, where it provides better spatial coverage for cloud detection in the Arctic. Moreover, the ELCM algorithm combines classification and clustering framework to fit the MISR data processing, which help improve the computational speed online effectively.

This research not only creates a better algorithm to implement cloud detection but also encourages further study on the changing cloud properties to the warming Arctic. In addition, it demonstrates the significant impact of statistics in successfully solving a modern scientific problem. Statisticians are now directly involved in the data processing and use powerful statistical thinking to help tackle challenges.

b. Summery of Data

For image1 data, there are 115229 data points. We calculate % of pixels for the different classes: 17.77% classified as cloud, 43.78% as not cloud and 38.46% as unlabeled. For image2 data, there are 115110 data points. We calculate % of pixels for the different classes: 34.11% classified as cloud, 37.25% as not cloud and 28.64% as unlabeled. For image3, there are 115217 data points. We calculate % of pixels for the different classes: 18.44% classified as cloud, 29.29% as not cloud and 52.27% as unlabeled. After we combined three image data sets, there are 345556 data points. We calculate % of pixels for the different classes: 23.43% classified as cloud, 36.78% as not cloud and 39.79% as unlabeled. In order to view the pattern or trend for the data points, we plot scatter plots and color each data point according to their labels.
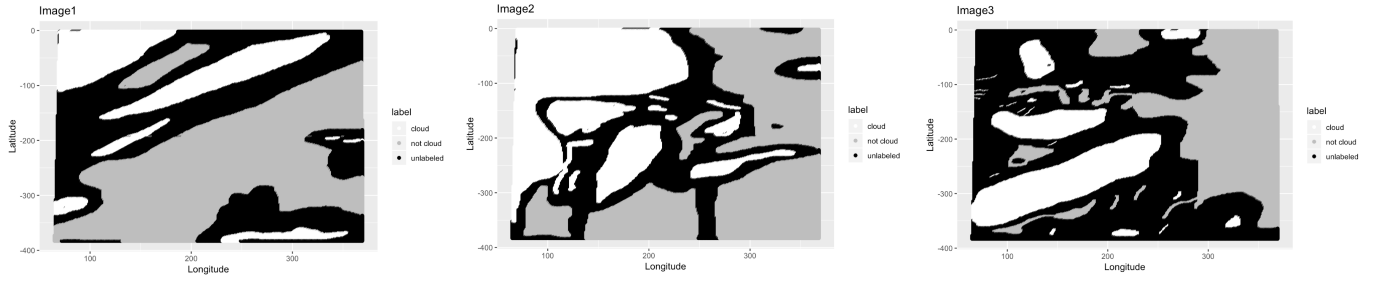
From the three plots, we can see the pattern that the pixels with same labels are connected to each other, and unmarked pixels stay around those marked as clouds. This pattern contradicts to our assumption that the samples are i.i.d., because the pixels adjcent to pixels marked as clouds are more likely to be marked as clouds.

c. EDA

To further explore our data, we also summarize the pairwise relationship between the features themselves.
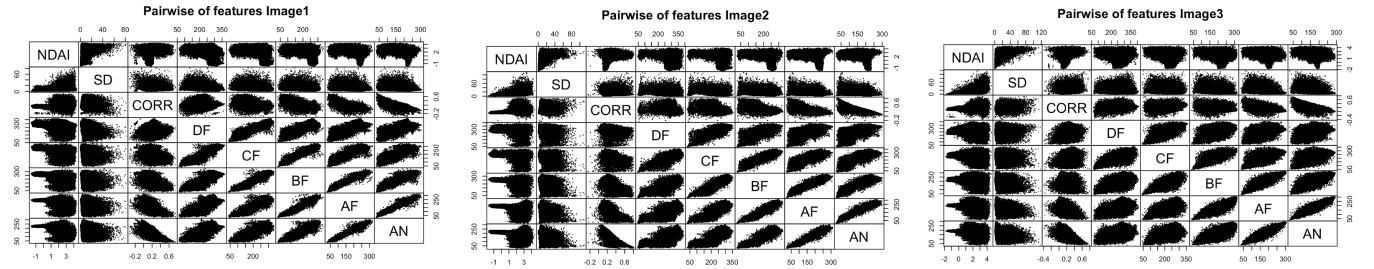
The three pairwise scatterplots all show that DF, CF, BF, AF and AN shows a linear relationship between each other, especially for AF and AN, AF and BF. However, we observe that data points in the third pairwise scatterplot spread more widely than in the first two pairwise sctterplots. The differences between three graphs can also show that the (linear) relationships decrease over time.

After checking pairwise relationship plots, we study the relationship between each independent featues and their expert labels for each image and all data combined.

(relationship between expert label and NDAI)
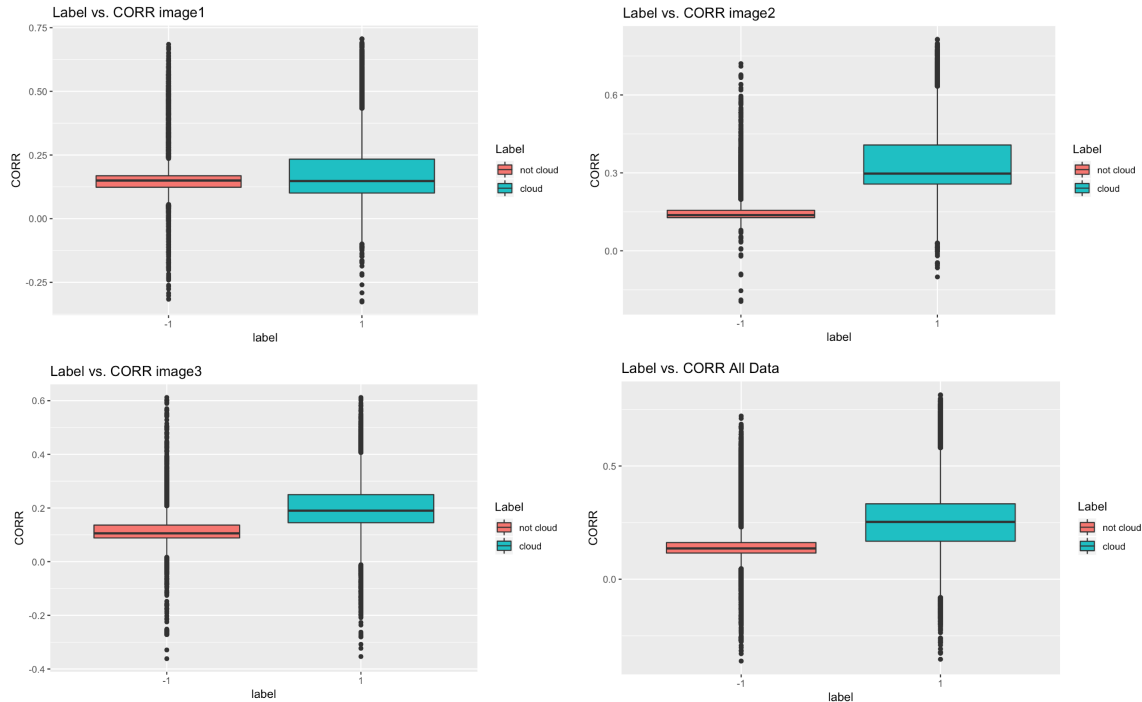
The boxplots for CORR based on different labels show that CORR is on average higher for pixels marked as cloud than as not cloud. CORR also has a higher variance for those marked as cloud.

This conclusion contradicts to the statement in the paper that high correlations over cloud-free areas or low clouds are expected. This is because high correlations for clouds also occur under rare circumstances. More importantly, recklessly declare clear for high CORR pixels and cloudy for low CORR pixels will produce errors due to smoothness of surface terrain and the difference of attitudes of clouds.

Therefore, we should also involve the feature SD to identify surfaces into our investigation. We first plot SD vs. Label independently for each image and all data.
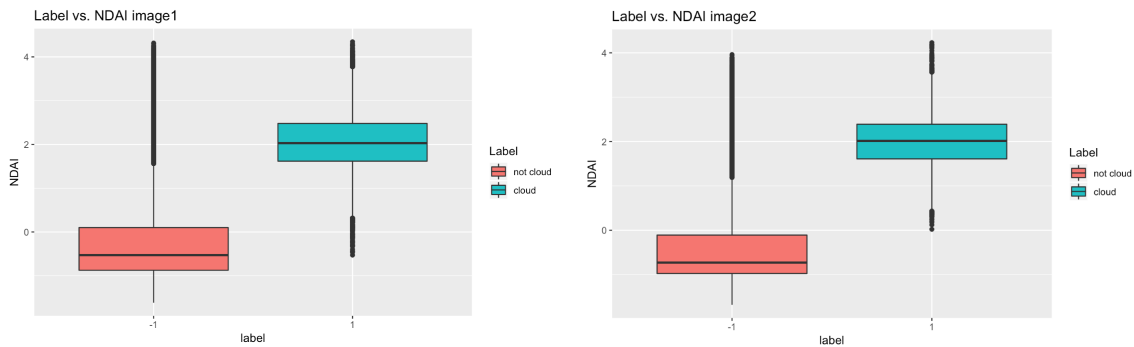
(relationship between expert label and SD)
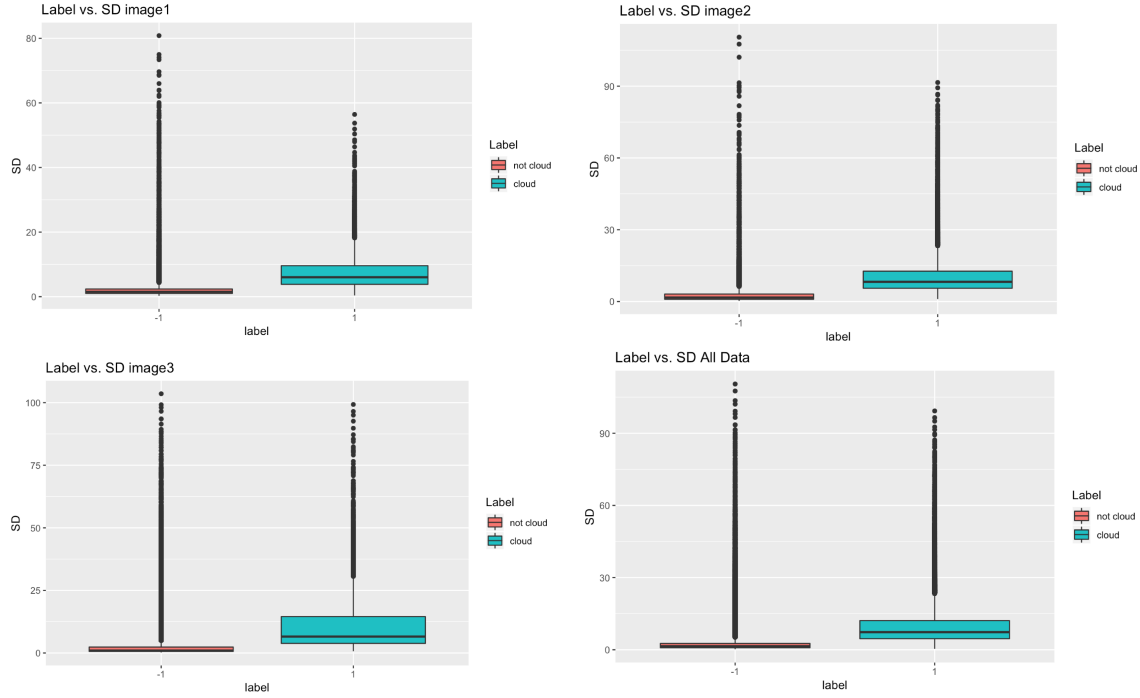
From the boxplots for SD based on different labels, we can clearly see that the SD is higher for cloud pixels than those cloud-free ones. But the cloud-free pixels spread more widely. It can be explained that SD are usually small for radiation emanating from smooth surfaces.

Finally, the thrid feature NDAI relates to the differences for isoreopic level of surface-leaving radiation between low-attitude clouds and snow-coved surfaces.
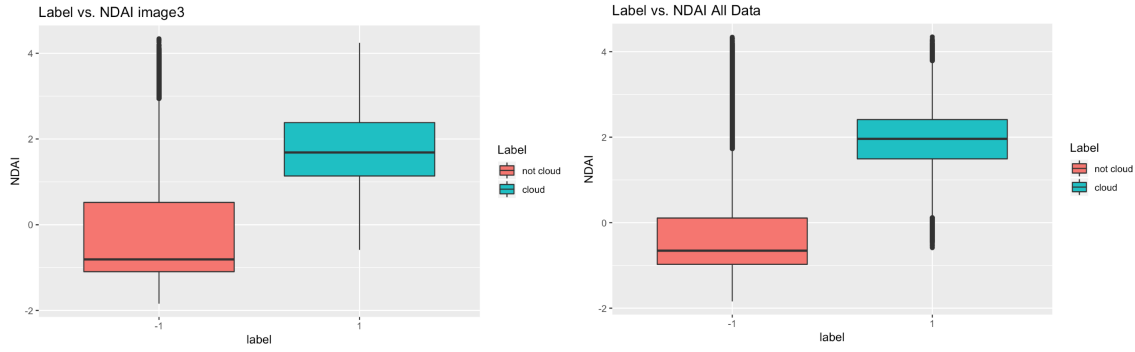
(relationship between expert label and NDAI)

From the boxplots for NDAI based on different labels, we can clearly see that the NDAI is higher for cloud pixels than those cloud-free ones. The distrubition for cloud-free pixels are left skewed, while the distribution for pixels marked as cloud are roughly symmetric. The distribution for three different image data are roughly the same as well.

2 Preparation

a. Data Split

Even though the three image data set represent the cloud dsitribution at different times at the same place, we still decide to combine the data into one to have more data for training. We also clear out the unlabeled data here because they are not helpful enough for classification problem. Since the data are not i.i.d., we cannot simply split the data by random.

First method: We decided to divide the data into 25 groups by cutting the image horizontally. In this case, data with y_coord from 2.0-78.2 and the corresponding x-coordinate are in the same group. We randomly select 23 groups from these 25 groups as training/validation data and 2 groups as test data. In the 23 groups, we randomly sampled 18 groups as validation data. Splitting the data in this way looks like a one-stage cluster sampling. Every time we sample a group wihout replacement and include all data points in that group. Data in the same groups tend to be correlated to each other, which fit the property of the non i.i.d. data.

Our Second method is, for each image data set, to divide all the data into 25 groups by cutting the image into 5*5 small images. For instance, data with y_coord from 2.0-78.2 and x_coord from 65.0-125.8 are in the same group. We randomly select 23 groups from these 25 groups as training/validation data and 2 groups as test data. In the 23 groups, we randomly sampled 18 groups as validation data. Splitting the data in this way looks like stratified sampling. Every time we sample a group wihout replacement and include all data points in that group. We can also see this method as combining small pixels into a much larger one in a visual way. This method can guarantee pixels that have high correlations (adjacent pixels) can be sampled at the same time (except those on the boundary). And the random sampling on the 25 groups can make sure that we are not too sticking to the patterns for the training data so that overfitting can be somewhat prevented.

b. Baseline

Setting all labels to -1, the accuracy for test data is 0.520 and that for validation data is 0.270, which are relatively low. If the test data and validation data contain mostly -1, the trivial classifier will have a high average accuracy.

c. First Order Importance

To find three best features, we first caculate and plot the correlation between those features and the expert labels. The magnitudes of the absolute values of the correlations represent how close the relationships are

between features and expert labels. Larger correlation value means better features. We found that NDAI and CORR are better than others on average.

Correlation for Image 1

|	[,1]
NDAI	0.7697836
SD	0.4423335
CORR	0.1965444
DF	-0.5845621
CF	-0.5856738
BF	-0.5761878
AF	-0.5485383
AN	-0.5089311

Corrplot For Image 1

Correlation for Image 2

	[,1]
NDAI	0.8193795
SD	0.4736005
CORR	0.7500088
DF	0.3090899
CF	-0.2534827
BF	-0.5154060
AF	-0.5842979
AN	-0.5726166

Corrplot For Image 2

Correlation for Image 3

	[,1]
NDAI	0.49887924
SD	0.23597170
CORR	0.34274487
DF	0.14241384
CF	0.02168176
BF	-0.05722629
AF	-0.12841936
AN	-0.17290429

Corrplot for Image 3

Then, we plot biplots and loadings on first two PCs to see how much each feature contributes to these PCs. We choose to use two PCs because they capture almost 90% of the data which is enough to represent the entire data set. The loading plots can illustrate the association between features and PCs. The larger the loading of a feature in a given PC, the more associated the feature is with that PC. After viewing the four loading plots, we observe that SD and DF have larger loadings than other features (expect NDAI and CORR). Considering there are other reasons that is more associated with scientific explanation, we decide to be consistent with the three features chosen in the paper, which are CORR, NDAI and SD.
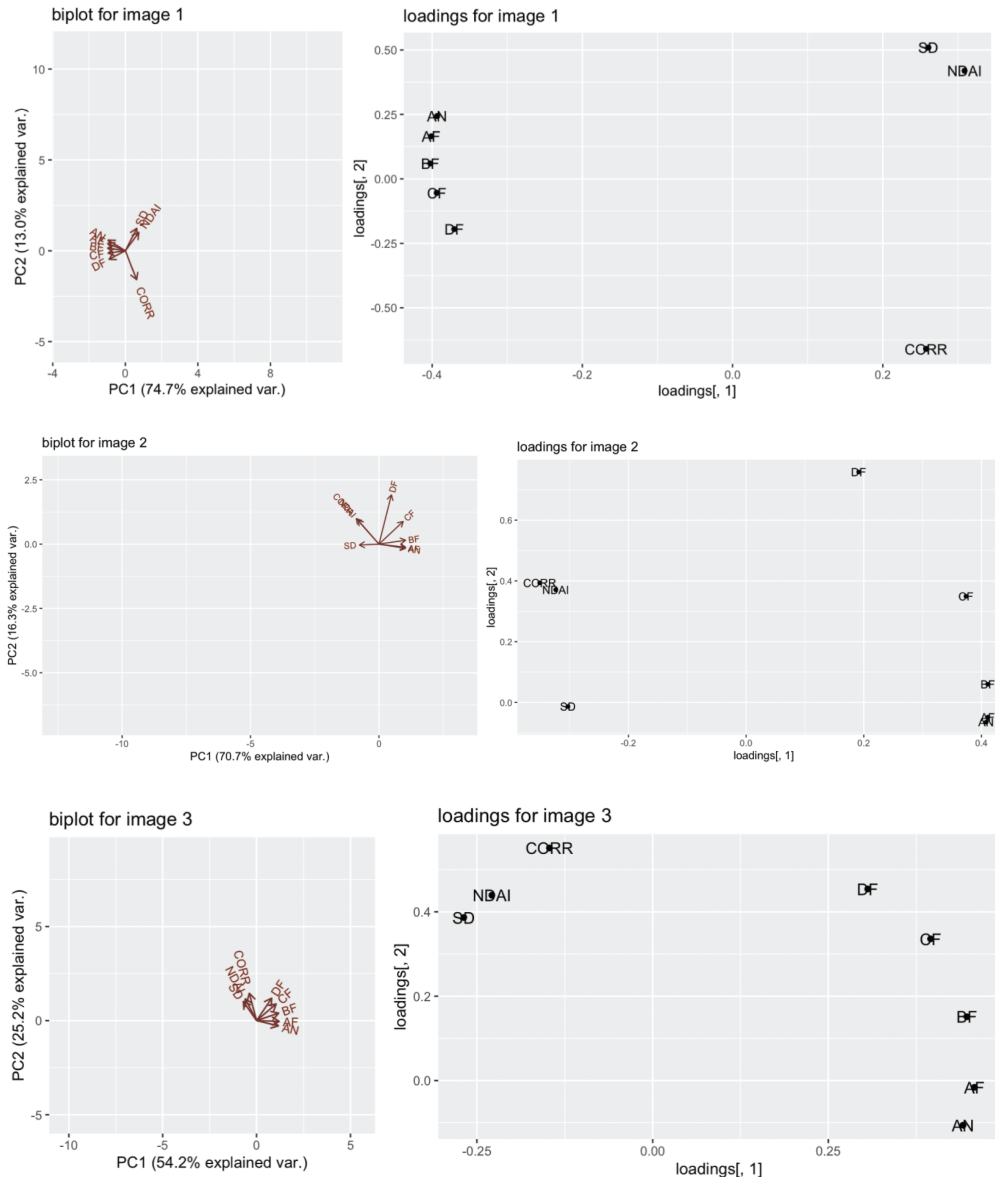
d. CVgeneric Function
See `https://github.com/WinnieGao/Project2_Git` for more detailed explanation and code.

3   Modeling

a. Different Classification Models
To train the data, we try several different classification methods: Generalized Linear Model, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-nearest Neighbor and Support Vector Machine. For GLM, LDA and QDA, since, in our natural world, most phenomena can be approximated using Gaussion distribution, we assume our data sets are under Gaussian distribution, which means LDA and QDA are appliable. Since Gaussian distribution is exponential family, we can also apply GLM in our case. Since there is no specific assumptions for using KNN nad SVM, they are also satisfied. We used cross-validation with 8 folds to train the models, and get the accuracies across all folds and the averaged one. We also apply the model on the test set and get the accuracy on the test set.

For our first splitting method,....

For the second one, the cross-validation average accuracies demonstrate that QDA have the best performance on validation data. QDA has a high accuracy on test data, but LDA performs better. GLM has lowest average cross-validation accuracy on validation set and svm performs worst on test data. Considering both average validation set accuracy and test set accuracy, we believe that QDA is the best model for the second splitting method.

	glm <dbl>	lda <dbl>	qda <dbl>	knn <dbl>	svm <dbl>
1	0.7767825	0.8990213	0.9047321	0.8906669	0.9003490
2	0.8290906	0.9018136	0.9040138	0.9026772	0.9028212
3	0.8000246	0.8915289	0.8905115	0.8976084	0.8932436
4	0.7853162	0.8939197	0.8973964	0.9009526	0.8948832
5	0.8250315	0.9015435	0.9011351	0.9038857	0.9030449
6	0.8625074	0.9406148	0.9410797	0.9480536	0.9450137
7	0.8657387	0.9409888	0.9516311	0.9452806	0.9414238
8	0.9226541	0.9798456	0.9945033	0.9807617	0.9790603
average	0.8333932	0.9186595	0.9231254	0.9212358	0.9199800

glm <dbl>	lda <dbl>	qda <dbl>	knn <dbl>	svm <dbl>
0.9599973	0.9921065	0.9892969	0.9845475	0.908403