Jilin Cao

Sizhuo (Cindy) Liu

Statistics 154

Spring 2019

Project 2 Final Report

Introduction

The purpose of this report is to conduct detailed data exploration and modeling of cloud detection based on data collected through MISR sensor. The primary goal of such analysis is to develop a classification model that can effectively distinguish the clouds and ice / snow surfaces using given features.

The paper consists of four major components: data collection and exploration, data cleaning, modeling, and diagnostics. In data collection the background and source of data was explained, and explorative data analysis was carried out. In data cleaning, the information was split into multiple sets and particular features were chosen for later analysis. In modeling, different classification methods were used and their fit was assessed through cross validation and other techniques. In the diagnostics section, the best classifier from the previous part was assessed for its accuracy in predicting the response variables.

1. Data Collection and Exploration

1.1 Summary of Paper

The purpose of this study is to develop a new operational Arctic cloud detection algorithm using Multiangle Imaging Spectroradiometer (MISR) imagery. In the contemporary world, the effect of the amount of atmospheric carbon dioxide is of great scientific interest to researchers. Arctic appears to be an ideal area for investigations in this field as it displays the strongest dependences of surface air temperatures on increasing atmospheric carbon dioxide levels. However, a major issue is that accurate Arctic-wide measurements are often difficult to obtain--it is hard to detect clouds over other surface types, such as liquid and ice-water cloud particles that have similar scattering properties. Therefore, it is crucial to develop a set of algorithms that accurately characterize the properties of clouds over the daylight Arctic.

The data used throughout the study were collected from 10 MISR orbits of path 26 over the Arctic, northern Greenland, and Baffin Bay through a time span of approximately 144 days. Each MISR pixel covers a 275 m x 275 m region on the ground. Six data units from each orbit are included in the observations. In addition to the key parameters such as correlation, NDAI index, standard deviation, and information from the six data units, the dataset also includes a column of expert label values, which represents the experts’ hand-labeled data indicating whether a particular patch of image pixels signify cloudy or clear scenes: all of which that are based on highly-confident domain knowledge were marked as clear or cloudy and those with ambiguity were unlabeled.

The data also suggests that the ELCM algorithm is more accurate and provides more comprehensive spatial information than the existing MISR operational algorithm. The conclusions from this research is significant in that it provides a better understanding of the polar cloud properties and its potential effect to the changes in the Arctic region brought about by increasing concentrations of atmospheric carbon dioxide.  Another key aspect of this research is that it shows the power of statistical thinking and the key role that statistics play in contributing to modern scientific problems.

A screenshot of a cell phone

Description automatically generated

1.2 Data Summary

A close up of text on a black background

Description automatically generatedWhen exploring the relationships between different variables, we purposely combined rows of datasets from all three images so that we could observe the overall relationships across all images. In addition, we stored a data frame in which we omitted all data with an expert label of zero (no information) as they, at this stage, does not contribute to the investigation. It will also make the classification process later easier as most of the classification methods used only support binary response variables. The original data is still stored under a different name just in case it is needed in later analysis.

A screenshot of a cell phone

Description automatically generatedAs shown by the plots of the three images on the left, the region of ice/snow surfaces and cloudy areas are clustered without obvious patterns. For this reason, we deduce that there are spatial correlations between pixels. Therefore, the independent and identical distribution assumption is not valid in this dataset.

The top left figure on the following page displays the pairwise plots for NDAI, SD, and CORR, which allows us to observe the pairwise relationship between the three variables. However, no significant pattern was found between the features except for NDAI plotted against SD, which seems to suggest a wider range of SD values as the value of NDAI index increases. Nevertheless, such correlation does not appear to be very strong. Such observations are similar across all three images. The top right figure on the following page shows the correlation values between different radiance angles and all values are highly correlated with each other. This may be because the data for the radiance angles are collected at around the same region with the same set of tools, just with slightly different angles. Relationships between NDAI, CORR, and the different radiance angle were observed by plotting the angles against NDAI or CORR. It was discovered that when CORR was plotted against the radiance angles AF, AN, and BF, there exists a negative, approximately linear relationship, as displayed in the three figures in the following page. This confirms the calculation of CORR used AF, AN, and BF in Bin’s paper (2008).

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A picture containing text

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1.3 Data Exploration

Histograms displaying the distributions of SD, CORR, NDAI, and a selection of radiance angles were also plotted, color-coded by the expert label. It is obvious from such histograms that the distributions of SD and CORR for both categories (cloudy and clear) are approximately the same, while the count for clear is significantly higher than the count for cloudy. The histogram for NDAI (shown on the right above) for all three images suggest that the lower range of NDAI index corresponds to a high number of “clear” labels, while all the “cloudy” labels were made at higher NDAI indices: the distributions for the two labels are relatively distinct from each other. This indicates that NDAI may be a good feature for predicting if a particular part of the images is cloudy or not. The other histograms, on the other hand, show that the distributions of the cloudy and clear labeled data have very similar distributions. The abundance of overlap suggest that it can be difficult to distinguish the two labels using the features.

1. Data Cleaning and Preparation

2.1 Data Splitting

In order to build and evaluate the model, we decided to separate data for each image into training, validation, and test set independently and then combine the sets for each image.

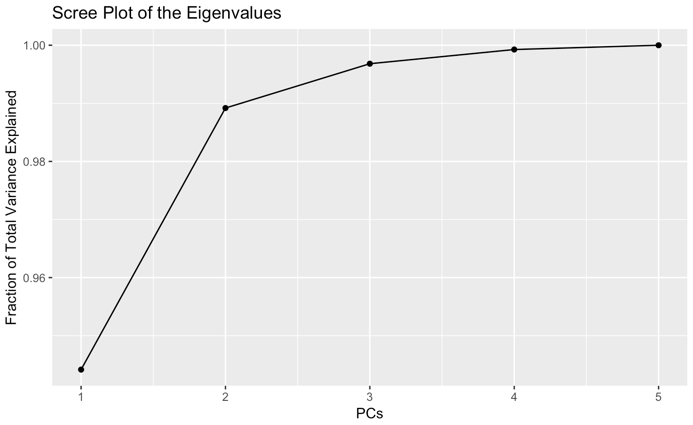
Taking the spatial relationship between adjacent pixels into account, we know that the data points are not independent and identically distributed. A naive splitting approach might cause the training dataset to consist of all cloudy or clear pixels. This will lead to a misrepresentation of the original data and thus negative affect the model performance. For example, if our training dataset has 99 out of 100 pixels as cloudy, then a trivial classifier will get 99% training accuracy.  In order to keep the balance in our three datasets as well as preserve local spatial correlation, we proposed two ways of splitting to avoid the issues. In the first method, we partition each image into n x n rectangular boxes, and then, viewing each box as an independent entity, we randomly sample training, validation and testing set among the boxes.

The second method is similar to the first one but slightly different.  Instead of choosing an entire box, we randomly sample from the data within each box into training, validation and testing sets. Then, we combine all training, validation and testing from each box into our final training, validation and testing sets. The second method may damage spatial relationship but can still manage to ensure a balance. It can also serve as a comparison against the first method. Our conjecture at this point is that our first method can do better as it also preserves the spatial relationships. However, such conjecture can only be assessed in the later modeling and diagnostics sections.

2.2 Baseline Accuracy

A trivial classifier that automatically sets all labels to be -1 was created. The accuracies of such classifier on the validation set and test set were then calculated in comparison to the actual expert label: they are 60.85% and 60.82%, respectively. This sets a minimal requirement for our research by implying that any model created should obtain an accuracy that is at least as high as the baseline accuracy, as such baseline model is the most basic and simple model one can possibly create.

2.3 Variable Selection

To investigate on the predictive power of the raw features, we implemented PCA on all five radiance angle features. As shown by the scree plot below, the first three PCs capture about 99.7% of the overall variance, which means the first three features are almost capable of explaining the entire dataset. Therefore, we decided to use the first three principal components as our features.

To see if our choice is reasonable, we performed a quick and simple logistic regression and tested it on validation set. We got an accuracy around 84.5%, which is higher than SDCM, ASCM, and Offline SVM methods mentioned in Bin’s paper (shown on the left). The coverage is also 100%.

A screenshot of a cell phone

Description automatically generatedHowever, this is not as powerful as sophisticated features: NDAI, CORR, and SD in Bin’s paper. The modeling and diagnostics will be done in next sections.

1. Data Modeling

3.1 Classification Methods and Fit Assessment

Several classification methods were implemented, and their fit was assessed using cross-validation.

The logistic regression model assumes that the predictor variables are all independent from each other, so that the problem of multicollinearity will not arise. From the pairwise relationship analysis from an earlier section, it was observed that the five radiance angles are highly correlated because their values were measured in very similar ways. However, from the variable selection section, it was concluded that NDAI, SD, and CORR do not display any obvious pattern or trend with each other. Since the actual model used only involves these three features, the assumption of non-multicollinearity is satisfied.

The LDA model makes two crucial assumptions: that the data under each label is multivariate normal and has the same variance. We can make a simple check if it is Gaussian by plotting a histogram each individual feature in the model. The histograms indicate that none of the features display a bell curve that is characteristic of a normal distribution. To check the variances across different features, two covariance matrices for the three variables were created, one for a set that contains only “cloudy” data, and another including only “clear” data. The final matrices are clearly different from each other. Therefore, neither of the assumptions of LDA are satisfied in our case.

The QDA model holds the same assumptions as LDA except that it does not require the variables to have equal variance. Since neither of the assumptions for LDA was fulfilled, so is the case for QDA.

SVM is very computationally expensive and so we decided not to implement for our model.

The k-nearest neighbor algorithm, decision tree, and random forest are all nonparametric methods. Therefore, their models make no assumptions over the data.

The accuracies of both methods of creating folds are as the following:

The tables above indicate that for all classification methods implemented, the accuracy is relatively consistent between different folds.

3.2 ROC curves

ROC curves for all fitted models are plotted here. The x-axis of ROC curve is false positive rate -- the type 1 error. The y-axis is the true positive rate also known as sensitivity or recall. Sometimes we may want to change the cutoff depends on our tasks, but in general we want to pick a cutoff value for which we can get the highest true positive rate and lowest false positive rate. For example, in the plot “Logistic model ROC curve”, we would want to choose a cutoff that is to the top left, which is showed in green and 0.5 is a good choice. For LDA, QDA and KNN, a higher cutoff value is preferred as shown in blue in the graph and the cutoff value is 0.8 or above.

3.3 Precision Recall curves

We also plotted precision recall curves here for different models. The ideas are similar to ROC curves. The x-axis is precision -- true positive / (true positive + false positive) and the y-axis is recall -- true positive / (true positive + false negative). Let’s say 1 is the label that we want to predict, in our case is cloudy. The precision tells us among pixels that are predicted as cloudy, how many of them are truly cloudy.  The recall tells us how many pixels are predicted as cloudy among cloudy pixels in the original dataset. For a good model we want both recall and precision to be high. Therefore, we want cutoff value to be close to top right corner. We observed similar cutoff as we get from ROC curves.

1. Diagnostics

[plots the misclassification maps]

c) We see that for some particular areas, our chosen model did very poorly. It fails to distinguish the cloudy and clear areas especially the borderline between clear areas and cloudy areas as we see in the plots. Therefore, we propose an ensemble learning idea. A single decision tree is high variance and not stable, while random forests train a lot of different trees with randomly sampled subsets of the data (bagging), and randomly sampled subsets of the features to decorrelate the trees. We combine a bunch of weaker learners and hope some weaker learners could specialize in hardly distinguishing areas and then we can use their knowledge. Since the random Forest is doing bagging on data, it’s sort of like feeding into new data. We tested it on our validation and testing datasets and achieved 97.35% accuracy.

1. [reference]

https://www.lexjansen.com/wuss/2018/130\_Final\_Paper\_PDF.pdf

https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated