Anomaly Detection Challenge 2

Vishal Bhalla Matriculation No: 03662226
Nithish Raghunandanan Matriculation No: 03667351

Satellite Image Dataset Classification

November 29, 2015

1 Abstract

The goal of this challenge was to implement a machine learning algorithm (Multi-class Classification) to classify the given the multi-spectral values in a satellite image. The database consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. In the sample database, the class of a pixel is coded as a number.

We found that K Nearest Neighbours (with k=3) and Random Forest as the best models for this classification.

2 Data Pre-processing

2.1 Rearranging the input data for multi spectral values

In each line of data the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom.
We rearranged the input data so that the all the data relating to a spectrum are grouped together from top-left followed by the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom.

2.2 Features

There were no qualitative or categorical features in the dataset. Moreover, as all features were numerical and within the range of 0-255, the data need not be normalized at all.

2.3 Handling Missing Values

We spent most of our time in pre-processing the data to handle missing values. We used various replacement techniques like replacing by zeros, column mean, column median, row mean, row median, row spectral mean, row spectral median, column minimum, column maximum, row minimum, row maximum, middle value, row spectral minimum, row spectral maximum, row mode, column mode, minimum, maximum and interpolated values.

3 Model

In particular we tried fitting different models to our data viz. One Vs One Classifier, Random Forest Classifier, Gaussian Naive Bayes Classifier, Bernoulli Naive Bayes Classifier and Multinomial Naive Bayes Classifier, Multi-class SVM Classifier, KNN Classifier amongst other models. Some of the concerns and issues faced during Model Selection were as below:

1. We used Stratified K-Fold Cross-Validation as a model evaluation metric. We split the training samples in the ratio of 4:1.

2. The below graph depicts the performance of various models on the different replacement techniques we used for handling Missing Values.

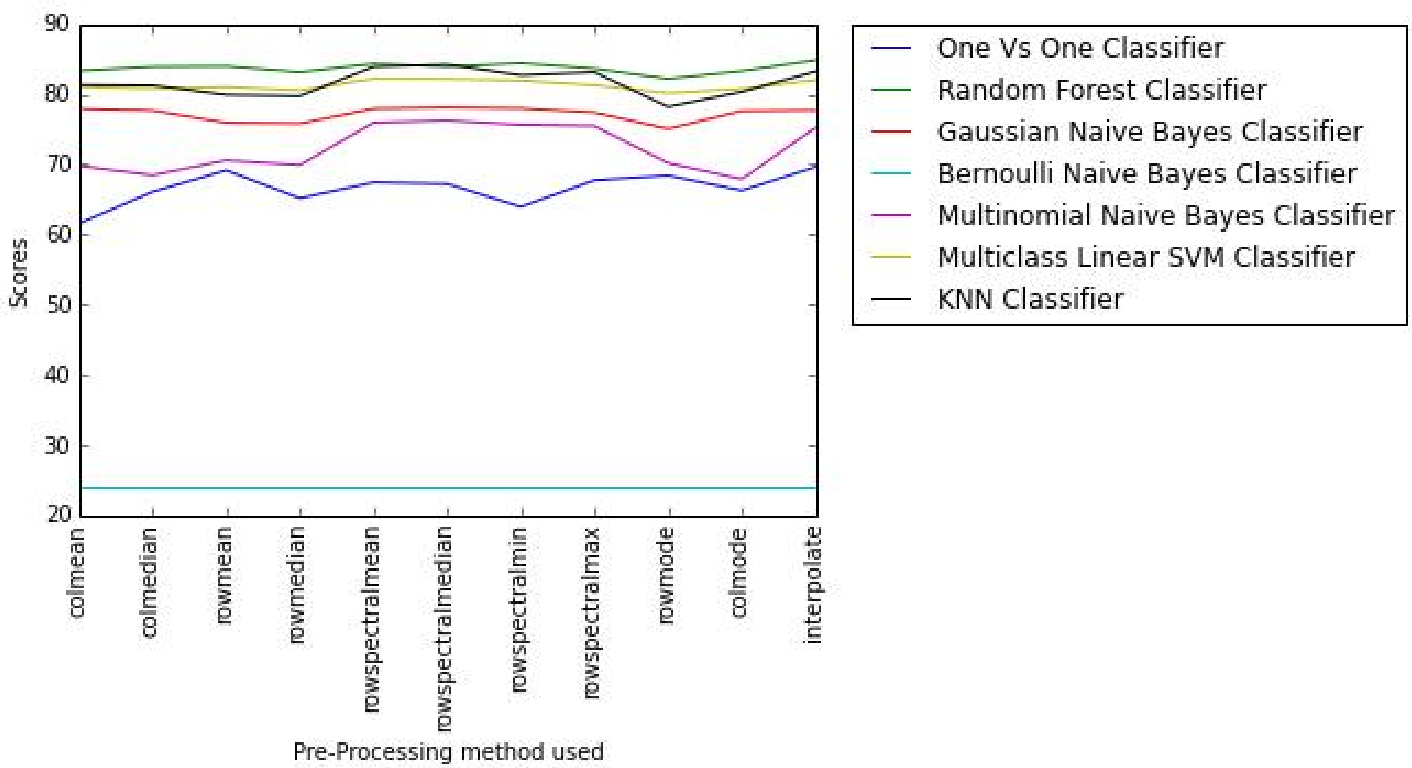

Figure 1: Scores on different models by trying out different techniques to replace missing values.

Replacement by row spectral median values gave us the best results over all models in comparison to all other replacement methods that we tried.

4 Results

The performance is evaluated by computing the Categorization Accuracy i.e. the percentage of correct predictions. The results of the different models on the Satellite Image Dataset are depicted as below:

Table 1: Summary by Categorization Accuracy of different Models

Model	Missing values replacement	Accuracy
Random Forest (50 trees with Entropy)	Row Spectral Median values.	0.89800
Random Forest (50 trees with Gini)	Row Spectral Median values.	0.89300
KNN with 3 neighbors	Row Spectral Median	0.89200
Random Forest	Interpolation	0.89000
KNN with 3 neighbors	Row Spectral Mean	0.89000
KNN with 3 neighbors	Column Mean	0.87800
Multi-class Linear SVM	Row Spectral Median values.	0.84000

5 Conclusion

By trying out different models on our training set, using rearranged and normal data, replacing missing values with row median, column median, interpolated and other such replacement techniques we infer that the best model is the one incorporating Random Forest (50 trees with Entropy as a criterion) on rearranged data and using row spectral median values for missing values that gave 89.80% accuracy on the test set.

Regularization would help reduce the over-fitting, but we could not find the best penalty factor for any of our models.