**Progress Report**

Hyperspectral data classification

Jun Chen, Dan Wang, Zhou Zhang and Penghao Wang

chen1241@purdue.edu, wang1766@purdue.edu

zhan1553@purdue.edu, wang1128@purdue.edu

1. **Data Preprocessing**

The data we are using for this project is a real hyperspectral image which contains 200 features and 16 different labels. There are 21025 samples in the whole image and 10366 of them are labeled samples. Although we have a large number of labeled samples in total, some classes have very few labeled samples, i.e. Oats (class 12) has only 20 labeled samples. To achieve a balance class distribution, we ignored the classes with fewer than 200 labeled samples. Then, we only focus on 12 out of the 16 classes for the classification problem. So for each class, we randomly selected 100 labeled samples as the training set, and other 100 labeled sample as the test set. After that, we also normalized the data and the data is ready to use.

1. **Future wok**

We will use three different classifiers for this hyperspectral image classification problem, which are SVM, Adaboot, and Random forest. More specifically, for SVM, we will use the RBF kernel as it is reported to have better performance for high dimensional data. In Adaboost, we will use the decision stump as the fundamental classifier. All these three models have parameters, such as the band width in the RBF kernel, the number of trees, and maximum iteration, and so on. We will use the cross validation approach to choose these hyper-parameters.

To evaluate the models, we will calculate overall accuracy, kappa statistics, and average accuracy. We will also plot the confusion matrix to see which classes are most confused. Since we know the physical meaning of each class, the confusion matrix can help us test if the classification result is reasonable.