Unsupervised Hyperspectral Image Segmentation of Indian Pines (April 2017)

Jason Bonvie, Michael Lowe, Ruoyu Wang

University of Florida Department of Electrical and Computer Engineering

***Abstract*—Currently hyperspectral imaging is used to obtain a spectrum for each given pixel in an image by a remote sensing device. This large multi-dimensional information allows researchers to analyze an image to detect, identify, and classify objects in application. Experimentation reflects on the testing, abstraction, and implementation of the K-Means Algorithm, Principle Component Analysis, Neighborhood Bias and the Gaussian Mixture Model to accurately generate a classified hyperspectral image.**

INTRODUCTION

Hyperspectral imaging (HSI) is performed using remote sensing devices. These devices obtain information about a specific object area in observance. By analyzing electromagnetic radiation transmitted through, reflected from, or absorbed by a local observed medium; i.e. Land; imaging information can be gathered and applied to obtain specific physical information about said medium [1]. The overall objective in applying hyperspectral image segmentation is to correctly cluster and classify the data provided to our system in comparison to the actual overlay of the geographic landscape.

I. PCA EXPERIMENTATION

Principle Component Analysis(PCA) is a procedure used to reduce datasets by identifying a small group of uncorrelated data with the largest variance—effectively reducing the observed feature space. This method allows for the given Indian Pine data to be reduced from a feature space of 200 components to a 20—Essentially providing 97 percent of the variance in from the original dataset with 10 percent of the total components.

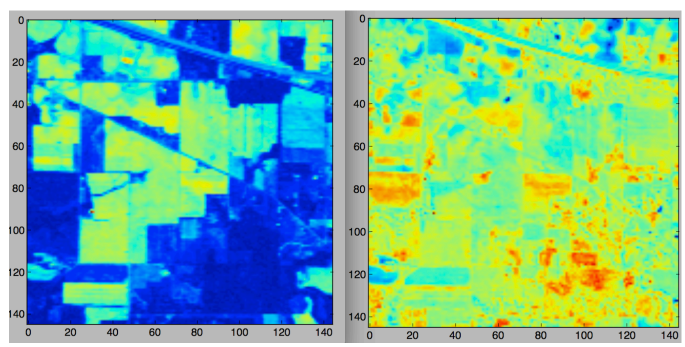

Figure 1.0 Useful hyperspectral band(right) vs Useless hyperspectral band(left).

Because there were several bands in the dataset that provided little to no information, PCA allows one to narrow the feature set for a smaller set of data with higher variance. The motivation behind doing so is to weed out what would be 23 percent of components that provided useless features and unneeded noise to the dataset (Bands 30-100). The following figure illustrate PCA as performed on the 200-component and after reduction.

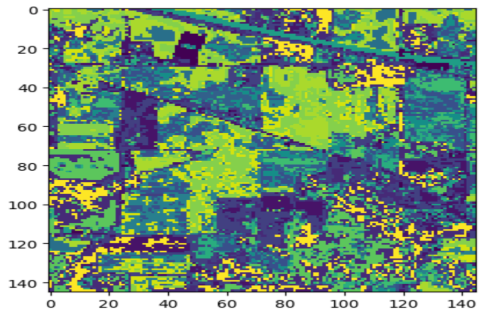

Figure 1.1 Reduced 20 component HSI after PCA performed.

II. K-MEANS EXPERIMENTATION

The K-Means Algorithm (see figures 1.3 and 1.4) is an iterative clustering technique used to classify data points in an unsupervised environment. The Algorithm compares the distance of each data point to the center of a preset number N of select centroids. The centroid is then updated at the end of each traversal until the centroid values become static [2].

Algorithm 2 *K*-Means Algorithm

```
Input: data points, number of clusters(k)
Output: centroids, centroid membership
if centroids are not initialized then
    Randomly select k points as centroids
end if
for each cluster center do
    Calculate the distance from each centroid to each data point
    Assign each data point to the closed centroid
    Recalculate centroid by calculating the mean of a cluster
    if centroids do not change and each data point maintains centroid classification then
        end for
    else
        Continue loop
    end if
end for
```

Figure 1.3 K-means algorithm implementation structure [2].

After applying PCA on the sample bands and reducing the feature space, an N-Clusters vs Adjusted Rand Score comparison can be made. The experimentation process can is mapped out by **figure 1.3** By adjusting the number of clusters used to classify the data points in the spectral bands, an optimal Rand Score is found at 9 clusters before applying neighborhood biasing.

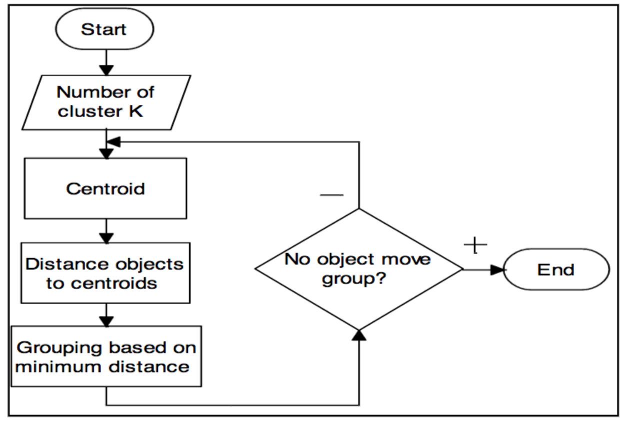

Figure 1.4 State Diagram for K-Means process [1]

Though the ground truth provided implies that the number of clusters used for classification cannot exceed 16, there was a significant drop off in classification accuracy the further the clusters were moved away from its local maximum value of 9(As observed in **figures 1.5** and **1.6**). However, the problem encountered from this inference is that the closer the number of clusters become to the ground truth, the less accurate the model becomes.

K-Clusters	Adj. Rand(%)
3	19.99
5	20.29
6	24.1
9	26.93
11	23.67
13	23.333
15	20.67

Figure 1.5 Adjusted Clusters Tested by K-Means Algorithm.

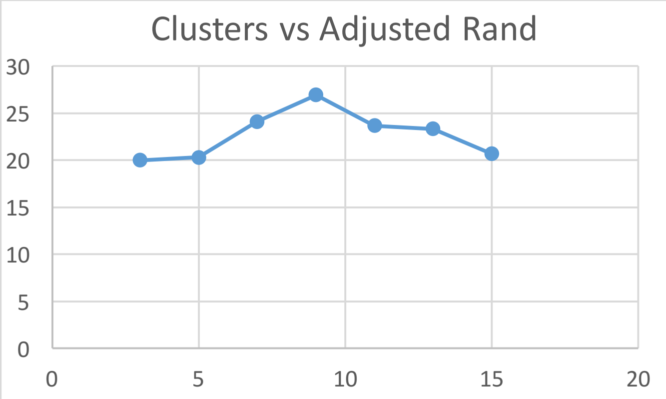

Figure 1.6 Plot of Adjusted Clusters Tested by K-Means Algorithm.

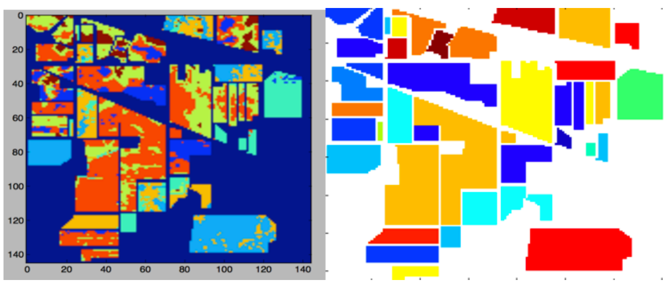

Figure 1.7 Experimental Yield(left) vs Ground Truth[3](right)

III. NEIGHBORHOOD BIASING

After testing different cluster models, we had hypothesized that it would be useful to implement some type of neighborhood biasing to change outlaying pixels inside of other clusters and reduce the noise in the clusters.[4] This algorithm takes a radius around each pixel to determine what most of the pixels around it are classified as then changes that pixel accordingly. Ideally this would allow us to achieve a better accuracy while increasing the number of clusters to get better adjusted rand index results. The neighborhood bias algorithm uses the 9 neighboring pixels from every point to determine what the general color of that pixel should be. Upon implementing this and testing different numbers of clusters we found the new most optimum cluster to be 15 clusters yielding and adjusted rand index score of .335.

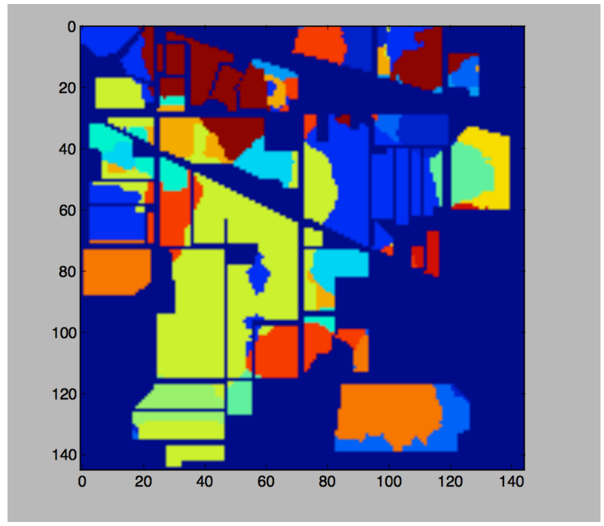

Figure 1.8 Final experimental results using PCA down to Twenty features, K-mean and a neighborhood biasing algorithm[4].

CONCLUSIONS AND DRAWBACKS

To more effectively reduce the feature space used in analyzing the hyperspectral bands, KPCA was applied to the data; however, given that the KPCA algorithm has a time complexity of $\mathbf{O(p^2n+p^3)}$, this process took too long to draw kind of inference simply because of dimensionality of the data, and local computing power.

Given the dimensionality of the data, and the ground truth supplied, the optimal clustering classification levels were surprisingly far away from the ground truth. Given that the data was not uniformly distributed across all 16 classes, the classification algorithm without neighborhood bias is not sensitive enough to recognized the existence of the smaller classes, nor is able to accurately separate them enough to generate a fine grain classification. After applying neighborhood bias a more defined classification was obtained through a larger number of optimal clusters.

References:

[1] Sahar A.El_Rahman, "Hyperspectral Image Classification Using Unsupervised Algorithms", Benha University, Cairo, Egypt, 2016.

[2] Hao Sun, "MAP-GUIDED HYPERSPECTRAL IMAGE SUPERPIXEL SEGMENTATION USING SEMI-SUPERVISED PARTIAL MEMBERSHIP LATENT DIRICHLET ALLOCATION", M.S. thesis, Graduate School at the University of Missouri, Columbia, Missouri, year 2016.

[3] University of the Basque Country, "Hyperspectral Remote Sensing Scenes", Available: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[4] Spectral Weighting and Spatial Biasing for Hyperspectral K-Means Clustering , Daniel Hanson, Sam Kreter, Brendan Marsh, and Christina R.S. Mosnick , Available: https://github.com/samkreter/kmeans-clustering-with-spatial-bias/blob/master/Report.pdf, 2016.