CMPE 462 - Project 3
Implementing K-Means & PCA

Task 1: K-Means Clustering

First part of the assignment is the implementation of k-means algorithm which will be applied on the given test data illustrated in Figure 1.

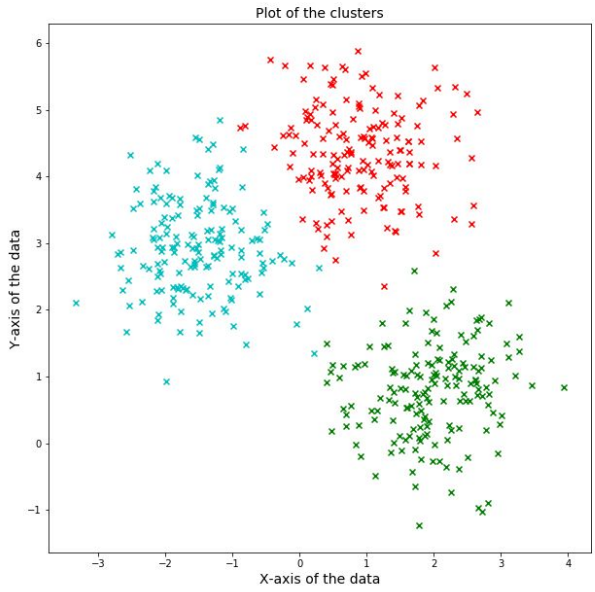

Figure 1: Clusters of given data

There are three clusters given in the label set. Therefore, we firstly applied the k-means algorithm on this data by selecting k as 3. The resulting plots can be seen in Figure 2 after 1 to 9 iterations. Here, the data points of clusters are illustrated in different colors that are chosen from random RGB values. Points illustrated with black stars represent the center of the clusters. For the calculation of accuracy, we detected the layer with the biggest amount of data points in each of our clusters. Then, the data points that are not labelled as the most populated layer in each cluster are considered as the error of clustering.

By evaluation of the accuracy of each iteration, it can be said that the algorithm converges after 4 iterations.

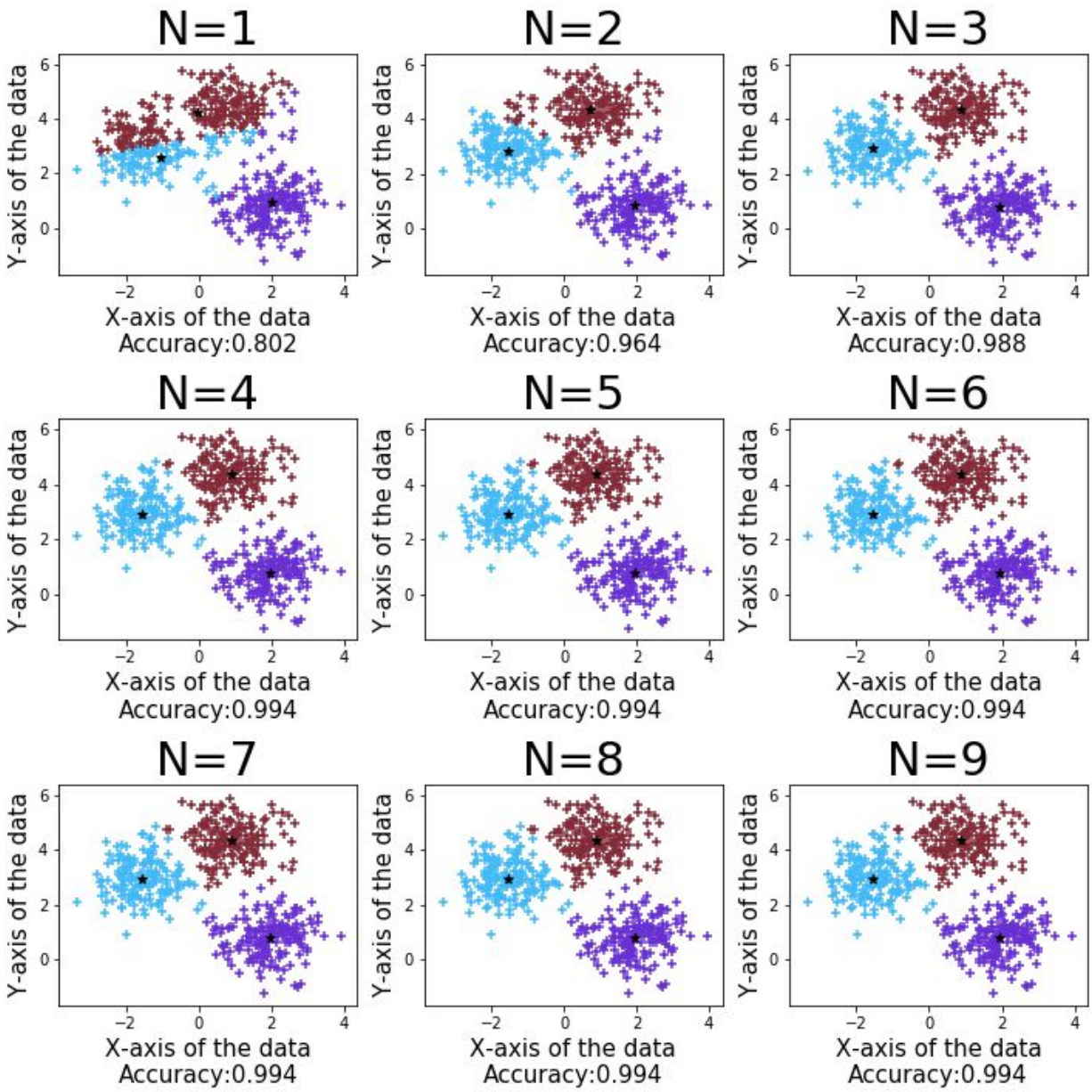

Figure 2: K-Means Implementation of k=3 with accuracy

For test purposes we also tested our k-means algorithm with k=5. The results are illustrated in Figure 3. Here, additional clusters created an intercluster within a given correct cluster.

For 5-means clustering, the accuracy shows a fluctuant performance. After 4 iterations, the accuracy shows a slight decrease then in 8th iteration a slight increase. The fact that the data originally consists of 3 labels while trying to classify it by 5 clusters causes this performance problem.

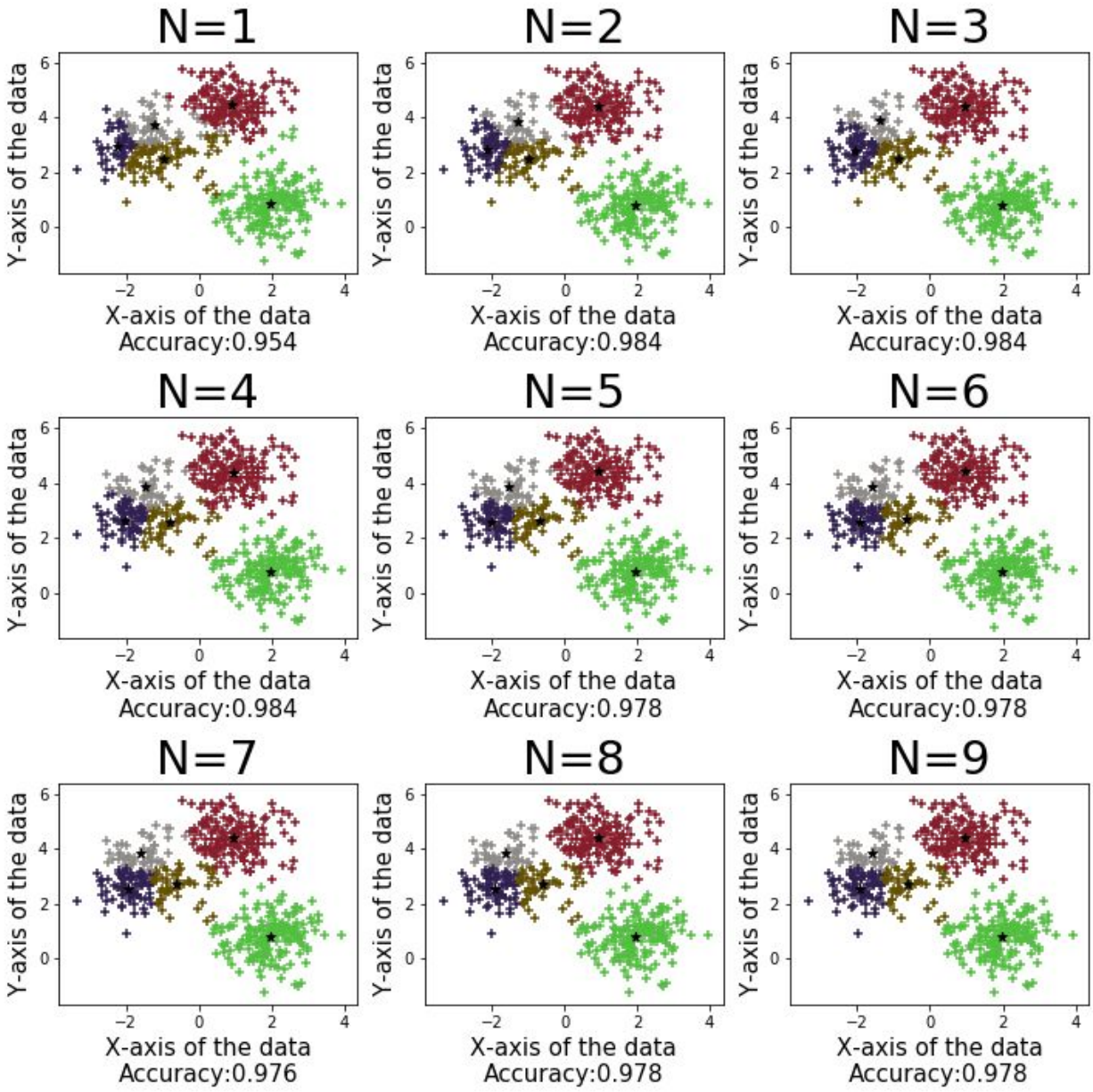

Figure 3: K-Means Implementation of k=5 with accuracy

Task 2: Principal Component Analysis (PCA)

This part is about the implementation of principal component analysis on an image set with the size of 3000x256. We applied PCA to this set with different d values that is used as the dimensionality reduction value. Then, we found the transformation matrix G by combining the eigenvectors of the covariance matrix with greatest d eigenvalues. The covariance matrix is created with $\frac{X^T X}{n}$ where X is the standardized data matrix.

We reconstructed images with original size boing th operation $G(G^T . X^T)$. The reconstructed images with the reduced pixel sizes (d) of 50,100,200 and 256, and the corresponding original images are illustrated in Figure 4. Here, we can see that there is a reconstruction error for all d values even if it is selected as the original pixels value. Moreover, this error increases for smaller d values since the number of principal components in the transformation matrix decreases.

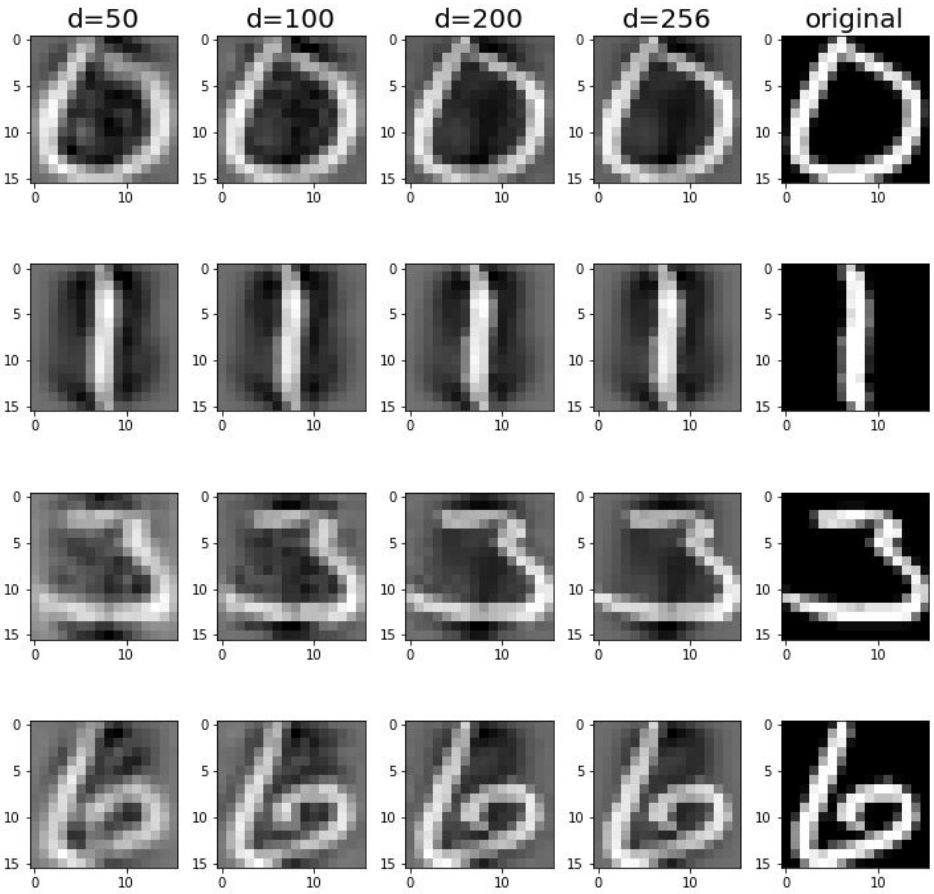

Figure 4: Reconstructed images for d = 50,100,200,256 and original image