Demo Of Selected
Data Analysis Techniques

By
Satish Kumar

Note:

1. Codes are written in MATLAB
2. Data files are contained in the respective code folder
3. This pdf contains program descriptions and results

I. Clustering

NOTES:

1. For kmeans, To generate initial random cluster centers, I used 'RandStream' functionality of MATLAB. A distribution of random data, which results in least cumulative distance of data from there cluster centers, has been picked for the further result and analysis.

2. I have used ground truth table data to create true cluster centers. For assignment of cluster centers obtained, with true cluster centers, I have used hungarian algorithm.

3. For Mean-Shift, I experimented with different kernal functions and windows size(h). A result with better accuracy is being reported here.

4. For Classification accuracy, I have used Euclidean distance between true center and the obtained center. Note that, It can be also choosen as fraction of data having correct labels when compared with true cluster easily.

Results: Classification accuracy, for kmeans with different measure of distances are as follows:

	L1	L2	L3	L-inf
Accuracy	101.4544	17.2031	18.7708	24.6618

Plots obtained are as follows:

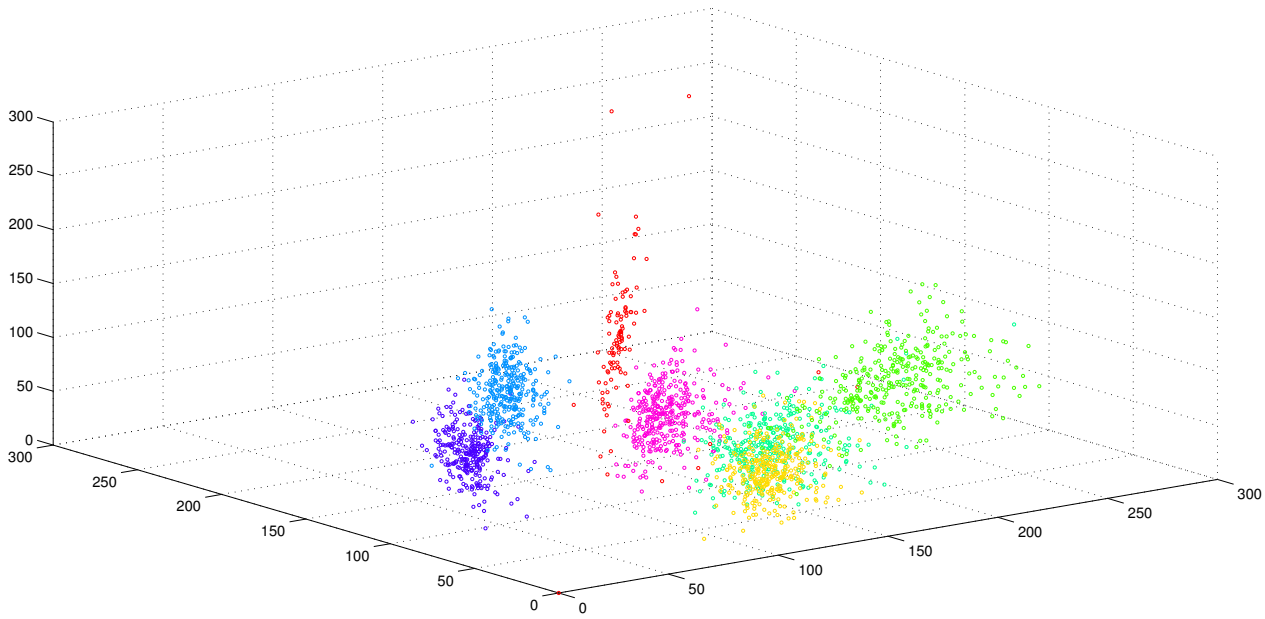

Fig. 1. Image of cluster of data, based on ground truth

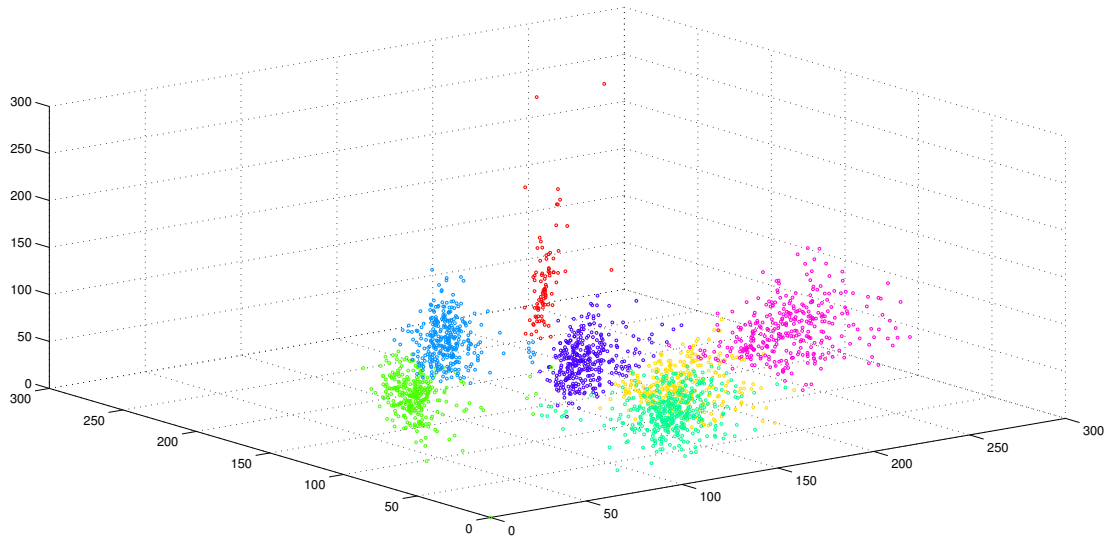

Fig. 2. Image of cluster of data, With kmeans (k = 7)

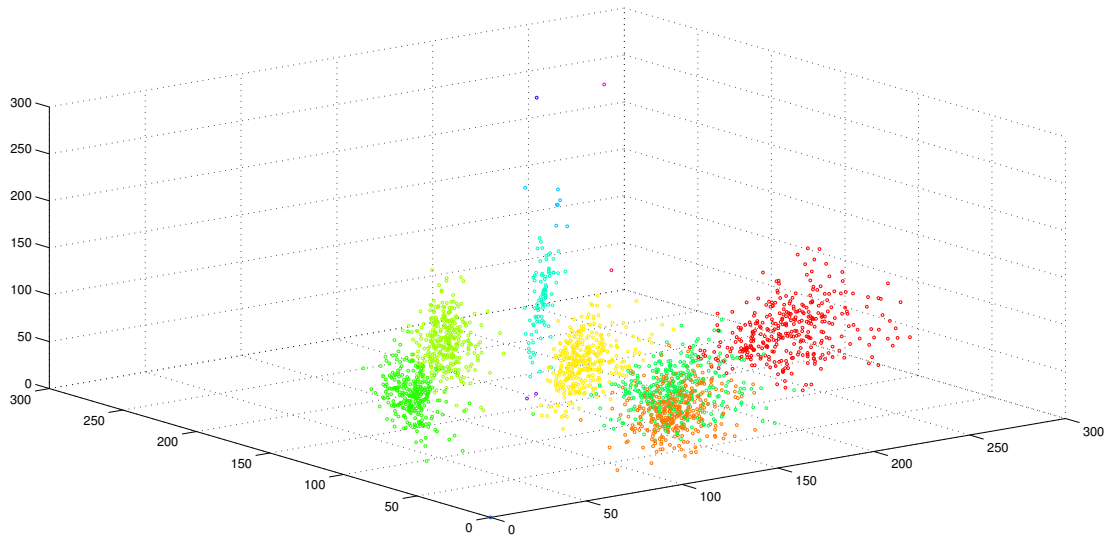

Fig. 3. Image of cluster of data, With Mean Shift (Normal Kernal)

Analysis:

- L2 distance major is giving better accuracy for k-means. Reason could be that, distribution from which the data set has been obtained supports uniform weight to each of the dimensions. Since, Euclidean(L2) norm gives equal weight to each element in the vector.
- Normal kernal function gives better accuracy because its gradient data distribution which is close to the distribution of given data set.

Part i continued on next page...

- kmeans gives better accuracy. Reason is because the parameter chosen is close enough to the true no. of clusters. In contrary, mean shift requires an appropriate choice of windows size (h). In some case when we have outliers type data, It will lead to generate/merge modes despite of choosing h carefully.
- Mean-Shift is computationally much more expensive than kmeans. But, Advantage is that you dont need no. of cluster which is generally not known.

II. Principal Component Analysis (PCA)

Results: Plots obtained are as follows:

Fig. 4: 16 Eigen Faces obtained for 80 sample faces

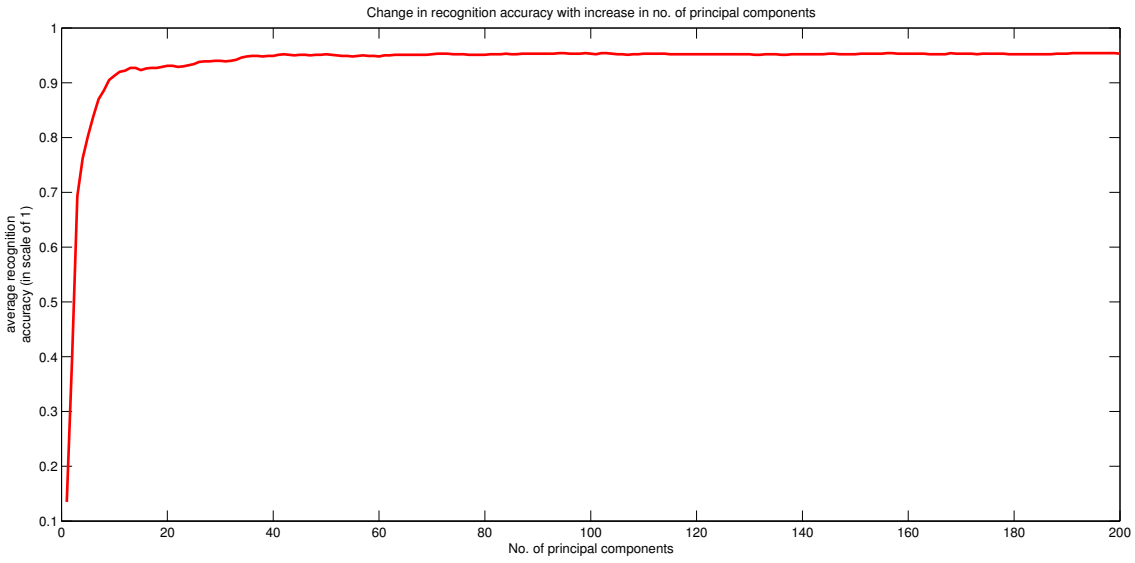

Fig. 5: Average Recognition Accuracy with 5 experiments for 200 train-test random split

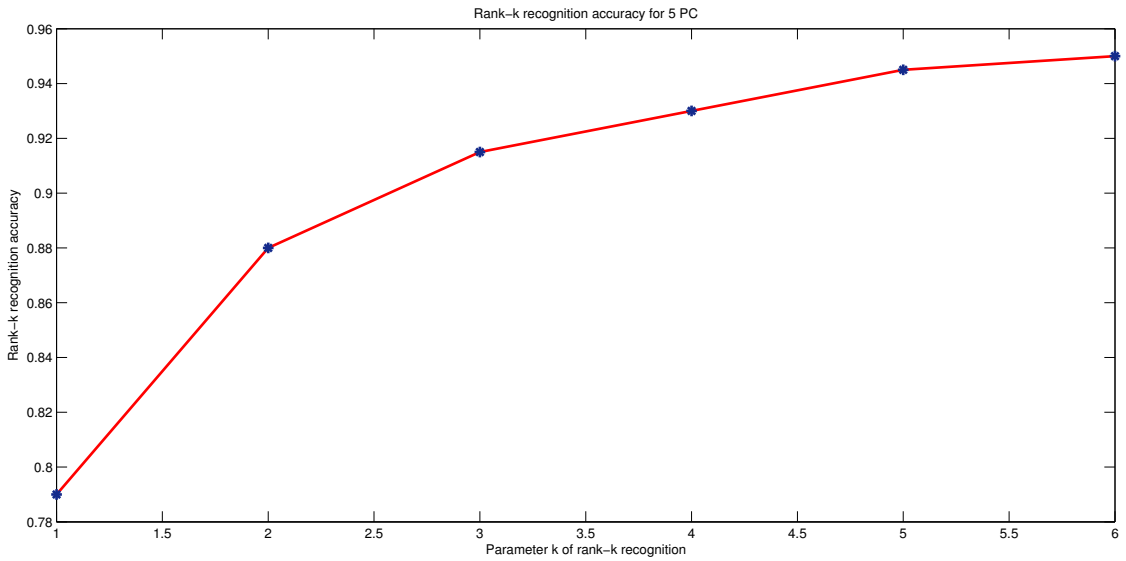

Fig. 6: Rank-K Recognition accuracy for 5 Principle Components with 200 train-test split

Observations:

1. **Fig. 5** shows that, We are getting about 95% accuracy even for about 16 principle components. Hence, We are able to reduce the dimension (From 10304 to 16) very significantly, without loosing much of information. It can be noted that choosing more pca components beyond this value may not be a good idea, since It will lead to much more computation while doing data analysis, but very less improvement in accuracy of result.

2. **Fig. 6** shows that, with increase in k-value, rank-k accuracy is increasing. This is obviously expected. But, we see that rate of increase of accuracy is very high for first few k values (We are getting 93% accuracy for k = 4, even for 5 PCA Component). Since rank-k accuracy can be a good measure of accuracy for some practical application (e.g. google search), This shows advantages of dimensionality reduction even with simple PCA technique.

III. Compact Image Representation

Results: mAP obtained for BoW and VLAD compact representation are as follows:

	BoW	VLAD
mAP	0.712	0.7612

Analysis:

1. Most computationally expensive step of compact representation transformation of image is to apply kmeans to find codebook. Also, BoW representation of image is much bigger (10k) compared to VLAD(64), BoW representation is computationally expensive to deal with compared to VLAD.

2. VLAD, Despite of being much more compact compared to BoW Representation, results in very good mAP.

3. But, good thing about BoW is, it's simple to represent and keeps most of the information (Only order of patches/words neglected). Hence, PCA followed by BoW could be a good choice.