Part 2 – Your Own Dateset

We choose a very typical dataset, CIFAR-10 dataset to perform a clustering analysis based on work mentioned above. CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. Inspired by part 1, which we achieve data importing, Kmeans clustering, contigency table of two classes clustering, 5 measures, namely homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score, 2 ways of dimension reduction, namely principle component analysis(Truncated SVD specifically), and non-negative matrix factorization, visualization, and different methods for performance improvement.

With regard to this pipeline, we achieve a realization for the whole clustering process for CIFAR-10. Since we do 10 classes clustering, it is more convenient to use 5 measures, homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score, to measure the performance than contigency tables.

Because we are doing 10 classes clustering, the 5 measures we achieve are much lower than those in the Part 1. However, the improvement still holds, which is that scaling featuring can enhance the performance. For example, the homogeneity of 10 classes clustering after PCA(Truncated SVD) is 0.078, and the homogeneity after scaling is 0.105, which gains an increase of about 35%. For more details, please refer to our jupyter notebook document.

Part 3 – Color Clustering

In this part, we use an image of G.E.M., a famous singer in Hong Kong, China, as shown in Fig., which is found in <http://j.17qq.com/article/hcsutuax.html>. We do a 3 classes clustering using Kmeans to every pixel in this image and obtain a new image, as shown in Fig. There are exactly 3 colors in this new image, corresponding to (255,0,0), (0,255,0), and (0,0,255) in RGB vector space.

QUESTION 11: BONUS - can you suggest a methodology to make an appropriate choice of k

and initial seeds of cluster centers?

There are three main factors we need to consider when constructing the methodology. One is the 5 measures mentioned, namely homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score. The second one is time t that each epoch consumes. And the last one is the number of classes we want to get, namely k. We denote the sum of 5 measures as s. And we can construct a grade as s/(k\*log(t+1)). The bigger the grade is, the more appropriate k and initial seeds are.

