K-MEANS CLUSTERING AND ADAPTIVE CLUSTERING

By Eric Z. Chen

University of Pennsylvania

1. Introduction. Clustering analysis is a common unsupervised approach often used to identify the internal structure of the data and can be used to extract useful information from the data without labels. K means clustering is one of the most commonly used approaches. For a given number of clusters, this method identifies the cluster membership for each data point such that the distance between each data point and the cluster centroid within the cluster is minimized. Although K means clustering is widely used, it has several limitations. It requires pre-specified cluster numbers and the algorithm does not guarantee to find the global optimum. It often fails if the true cluster pattern is complex or the size of the data in each cluster is unbalanced. To identify more complex cluster patterns, one can map the data from data space to the feature space with kernels, which is called kernel K means clustering. To overcome limitations of k means clustering, Efimov et al. recently proposed an adaptive nonparametric clustering approach. This approach begins with a set of seed points defined by small radii and then performs clustering in an adaptive way with increased radius. In each iteration, it performs a statistical test on two points for significance of overlapping between their neighbor points, where the neighbor points is defined by the radii. The two points can be assigned into one cluster if the overlap of their neighbor points is significant.

2. Method. Three clustering algorithms were implemented in Python, namely K means clustering, kernel K means clustering and adaptive clustering. Those functions were wrapped into a Python package. The whole analysis presented in this report is also available as a Jupyter Notebook and can be download from the same github repo.

Features of real data were normalized to [0,1] if they are in different scale. The Euclidean distance measure was used in three clustering algorithm. For kernel K means, the Gaussian kernel was used. The parameter σ was tuned by searching in $(1.0, 2.0, 3.0, 5.0)$ and the best result was reported. In order to find the best cluster numbers for K means and kernel K means clustering, values from $[max(2, true\_n\_clusters - 5), true\_n\_clusters + 5]$ were tested and the Silhouette scores were calculate. The best cluster number was selected with the largest average Silhouette score. Since the true cluster labels are available, the Rand index was used to compare the performance of the three clustering algorithms. Check the notebook for details about data analysis.

3. Discussion. The K means and kernel K means clustering do not guarantee to find the global optimum. Thus it is usually repeated with different random centroid initializations to avoid local optimum. Some methods have also been proposed

TABLE 1
The Rand scores of K means clustering, kernel K means clustering and adaptive clustering on simulated and real data.

Dataset	K means	Kernel K means	Adaptive clustering
aggregation	0.762	0.785	0.809
compound	0.437	0.729	0.803
D31	0.883	0.909	0.517
flame	0.430	0.550	0.950
jain	0.324	0.261	1.000
pathbased	0.461	0.521	0.970
R15	0.792	0.892	0.928
spiral	-0.006	0.051	1.000
wine	0.869	0.899	0.268
glass	0.252	0.263	0.024
thyroid	0.232	0.247	0.252
iris	0.568	0.568	0.568
lwdbc	0.730	0.730	0.003

to find a good initializations. K means clustering often fails to identify complex cluster patterns although using kernel can help to alleviate this problem. Another drawback of K means is that it often to identify the correct structure if the size of the data in each cluster is unbalanced. Outliers in the data can also affect the performance of K means clustering. K means clustering also requires user to specify the number of clusters. One great advantage of K means clustering is that this method can handle very large sample size and therefore is widely used in the industry.

Another popular clustering algorithm I often use is DBSCAN. It can find very complex patterns in the data and can identify outliers. For text data, I often use LDA (topic models) to cluster the documents. One major difference of such model from K means clustering is that topic models assign a probability for cluster membership, rather than 0/1. So it is often called soft clustering and it is especially useful in some scenario.

Biomedical data are usually high dimensional. Any clustering method based on distance does not work well in high dimensional data. This is because in high dimension space, the distance between any two date points are almost the same and thus it is difficult to define the nearest neighbors.

The recently proposed adaptive clustering approach by Efimov et al. addresses some of the limitations of K means clustering such as unbalance size in different clusters, complex cluster pattern, outliers, unknown number of clusters. Since this approach is still a distance based approach, it is expected to not work well for the high dimensional data. Actually, I found that for the real data with dimensions > 10, this method fails to identify the underlying cluster patterns. One thing I noticed is that the author proposed to use $n_0 = 2p + 2$ as the initial number of neighbors for each data points. When dimension is relatively large, n_0 gets large value and thus h_0 is large. This causes the weigh matrix fail to initialize properly. Also, when dimension is relatively large, it is difficult to find the neighbors since almost all points are equally distanced. This caused the h_k sequence difficult to generate properly. Therefore, I tried PCA on those data first and perform the adaptive clustering on first three PCA components.