Final

Clustering is a very intriguing problem that is still popular these days. **Clustering problem** can be defined as a task of grouping a set of objects in such a way that objects in the same group are more similar to each other than objects in other groups. Clustering, due to its nature, is a typical unsupervised learning method. Because clustering can be used in many fields such as pattern recognition, data compression, machine learning and so on, there exists numerous algorithms to tackle problems of many kinds.

Hierarchical clustering is one of the most common algorithms in solving clustering problem. This type of clustering, also known as connectivity-based clustering, is based on the notion that objects typically relate to their nearest objects than those faraway. Implementations for hierarchical clustering can be divided into two groups: agglomerative (bottom-up approach) and divisive (top-down approach). The results of clustering can be shown in a dendrogram. The standard algorithm for hierarchical agglomerative clustering can reach $O(N^3)$, which is unbearable in most case. And divisive clustering with exhaustive search typically consume $O(2^N)$ and cannot be accelerated by heuristics methods.

K-means is also a widely used clustering algorithm. The basic idea is to partition N instances into K clusters with their closest mean. **The distinctions between two algorithms** are three-fold.

First, the computational complexity of two algorithms vary. K-means is a NP-hard problem, which means it is hard to give a precise solution, but we can use heuristic algorithms to quickly converge to a local optimum. Hierarchical clustering, however, uses greedy method. But the time complexity of this algorithm is too high to accept when N is sufficiently large. They both cannot give a precise solution, but K-means perform much better in finding a converge solution when the input is large enough.

Second, two algorithms use different cluster models, varying significantly in property. K-means use centroid models, meaning each cluster can be represented by a single mean vector. Hierarchical clustering, on the other hand, uses connectivity model that is based on distance connectivity.

Finally, their applications differ. K-means cannot handle irregular-shaped dataset and is dependent on the choice of k. Therefore, K-means can process datasets consist of uniform size of instances. Hierarchical clustering can deal with data points in complex space and can detect noise. So, Hierarchical clustering can be implemented in datasets with small but irregular-shaped instances.