Hashing using Hierarchical K-means

Bharat Mohan
bmohan@iitg.ac.in

March 15, 2020

1 Hierarchical K Means Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a ”bottom-up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a ”top-down” approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Here the divisive method of Hierarchical 2-Means clustering (ball tree) is made use of, to generate a Hash for each data in the dataset.

In order to decide where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Here as a measure of dissimilarity Euclidean distance is used to compute distance of each data from the centroid of the clusters which forms the crux of K-Means algorithm.

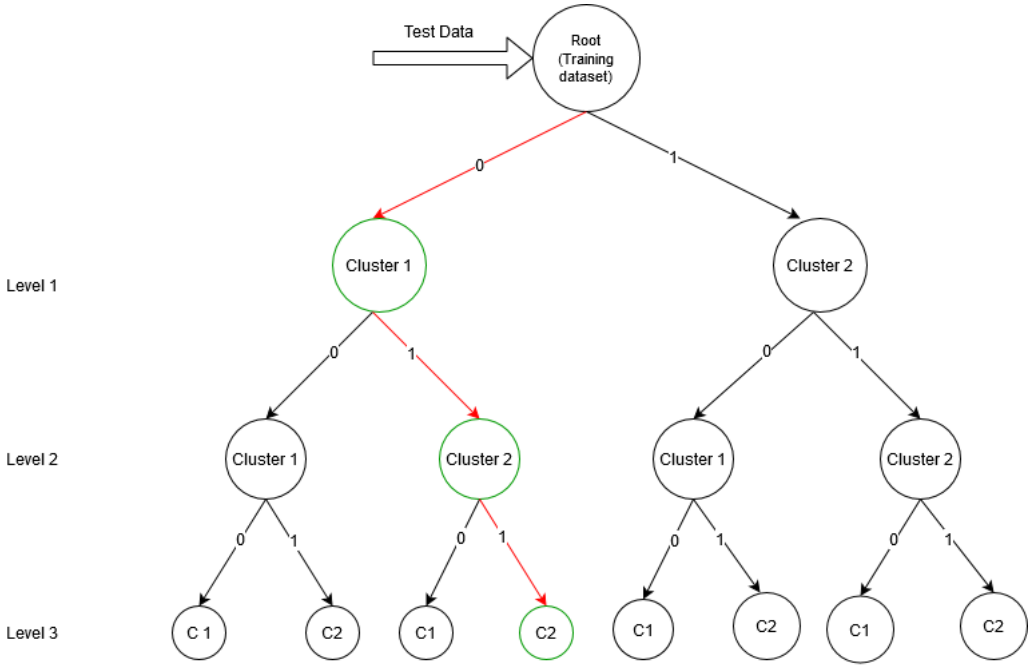
$$\|a - b\| = \left(\sum_i (a_i - b_i)^2\right) \tag{1}$$

2 Hashing Using Hierarchical 2-Means

Hashing is a method of representing a collection of higher dimensional data by a lower dimensional hash. The primary motivation behind is the reduction of time complexity involved in searching and finding the nearest neighbors of a particular data, for example an image. The image (usually a very high dimensional data) can be represented by its hash value and similar images can be found out by comparing the hash values associated with the images in the database.
One of the methods of generating hash involves hierarchical clustering. Here hashing using Hierarchical 2-means is illustrated.The training dataset is employed to generate a tree,which involves clustering the data into 2 clusters at each node, the cluster mean and

the associated data (stored as additional info) are stored in the respective node.The complete tree is thus generated whose number of levels are determined by required dimension of the hash value. To determine the hash value for a particular test data, a complete tree traversal upto the leaf node is carried out,of which the path traversed determines the hash value.

3 Illustration



Hash generated: 011