Sakshi Jain(B20ME065)

Spectral Relaxation for K-means Clustering

By - Hongyuan Zha & Xiaofeng He,
Chris Ding & Horst Simon, Ming Gu

The approach used in this paper is to minimize the sum-of-squares cost function using the coordinate descend method. Minimizing the cost function is equivalent to trace maximization formulation and assigning the cluster membership using pivoted QR decomposition, by taking into account the special structure of the partial eigenvector matrix. Then taking the performance of the clustering algorithms using document clustering as an example.

Possible Updates In The Approach:

Here in this paper sum-of-square cost function and Kmeans are used, and using these two in a different sense can work the clustering in a more efficient manner. For this, one needs to first assign the Kmeans centroid points and then calculate the distance of each point of the dataset from the assigned centroid points. Then calculate the minimum distance/maximum similarity between the centroid point and the data point which is closest to that centroid point. Then find the closest point of the data point other than the centroid point from all other data points, this process continues until all the data points are being assigned to any of the clusters. Then calculate the centroid of the formed clusters by taking the mean value of each coordinate point in that cluster. Continue this process until the same cluster set is obtained after doing the same procedure. This is an alternative way to use the method of Kmeans clustering with the use of the minimum cost function.

Another way to find the minimum distance/maximum similarity point to each centroid point and then form the clusters based on the distance of data points from centroid points. Then assign the next cluster centers as the boundary points of the previous clusters.

Implementation:

The implementation of the updated approach leads to increasing in the accuracy by Kmeans clustering. The benefits of using this method are that it is relatively simple to implement and it scales to large data sets. It applies to unlabeled data sets and for linearly separable data and this algorithm works quickly. It can be used with large datasets conveniently.

Ordering points to identify the clustering structure is an algorithm for finding density-based clusters in spatial data. The basic approach of OPTICS is similar to DBSCANbut instead of maintaining known, but so far unprocessed cluster members in a set, they are maintained in a priority queue. Extracting clusters from this plot can be done manually by selecting a range on the x-axis after visual inspection, by selecting a threshold on the y-axis. Even when no spatial index is available, this comes at an additional cost in managing the heap. It can simply be set to the maximum possible value. When a spatial index is available, however, it does play a practical role with regard to complexity. optics abstracts from dbscan by removing this parameter, at least to the extent of only having to give the maximum value.

Advantages of this approach:

It can find arbitrarily shaped clusters. The order of the point in the database is insensitive. Handles noise and outliers. It does not require density parameters. The clustering order is useful to extract the basic clustering information. Produce nonspherical clusters which actually occur only on spatial data. The term spatial data is used to express points, lines, and polygons. An interesting property of density-based clustering is that these algorithms do not assume clusters to have a particular shape. Furthermore, the algorithms allow “noise” objects that do *not* belong to any of the clusters. Internal measures for cluster evaluation also usually assume the clusters to be well-separated spheres (and do not allow noise/outlier objects) - not surprisingly, as we tend to experiment with artificial data generated by a number of Gaussian distributions. This process classifies the whole dataset into heterogeneous clusters. Across the years, many versions of this algorithm were developed to enhance its performance, such as k-medoids, kernel k-means, and k-harmonic-means.

Disadvantages of the previous approach, which are being overcome:

It Can not perform well with large differences in densities and is not suitable when various density involve. It only produces a cluster ordering. It can’t handle high-dimensional data. Clusters formed in spatial data clusters may have arbitrary shapes. The need for subsequent full efficiency in large-size databases. The ability to detect and reduce noise and outliers. OPTICS and DBSCAN are not the only algorithms that implement the density-based method.