Summary of Algorithm

**A hybridized K-means clustering approach for high dimensional dataset**

**In Brief...**

Due to incredible growth of high dimensional dataset, conventional data base querying methods are inadequate to extract useful information, so researchers nowadays is forced to develop new techniques to meet the raised requirements. Such large expression data gives rise to a number of new computational challenges not only due to the increase in number of data objects but also due to the increase in number of features/attributes. Hence, to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by an efficient dimensionality reduction method

The Principal Component Analysis (PCA) method is used as a first phase for K-means clustering which will simplify the analysis and visualization of multi-dimensional data set. Also, a new method is proposed to find the initial centroids to make the algorithm more effective and efficient.

As original K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, PCA is applied on original data set, to obtain a reduced dataset containing possibly uncorrelated variables. Then the resulting reduced data set will be applied to the K-means clustering algorithm to determine the precise no. of clusters. As quality of the final clusters heavily depends on the selection of the initial centroids, a new method is proposed to choose such data objects as initial centroids whose squared Euclidian distance is maximum among all the data objects, to make the algorithm more effective and efficient.

.

The steps of the hybridized k-means clustering algorithm are as follows.

Input: *X* = {*d1*, *d2*,……..,*dn*} // set of n data items.

*K* // Number of desired clusters.

An array *Cen* [ ] having size *k* initially being empty.

Output: A set of *k* clusters

// **Phase-1: Apply PCA to reduce the dimension of the data set**

1. Organize the dataset in a matrix *X*.

2. Normalize the data set using Z-score.

3. Calculate the singular value decomposition of the data matrix.

4. Calculate the variance using the diagonal elements of *D*.

5. Sort variances in decreasing order.

6. Choose the *p* principal components from *V* with largest variances.

7. Form the transformation matrix *W* consisting of those *p* PCs.

8. Find the reduced projected dataset *Y* in a new coordinate axis by applying *W* to *X*.

//**Phase-2: Find the initial centroids**

9. Set *m*=1.

10. Compute the distance between each data points in the set *Y*.

11. Choose the two data points *yi* and *yj* such that distance (*yi, yj* ) is maximum.

12. *Cen*[*m*] = *yi* ; *Cen*[*m+1*] = *yj* ; *m*=*m*+2 ;

13. Remove the two objects *yi* , *y*j from *Y.*

14. While (*m* <= *k*)

1. Find the distance of each object in *Y* to *Cen*[*i*], for *i* = 1 to *m*-1.

2. Find the average of all the distances to the centroid for each object in *Y*.

3. Choose the data object *yo* having maximum average distance from previous centroids.

4. *Cen*[*m*] = *yo* ; *m* = *m*+1;

5. Remove the object *yo* from *Y*.

// Phase-3: Apply the K-means clustering with the initial centroids given in array *Cen*.