Spectral Relaxation for K-means Clustering

Hongyuan Zha, Xiaofeng He
{zha,xhe}@cse.psu.edu
Dept. of Compo Sci. and Eng.The Pennsylvania State University, University Park, PA 16802

Chris Ding, Horst Simon
{chqding,hdsimon}@lbl.gov
NERSC Division Lawrence Berkeley National Lab. UC Berkeley, Berkeley, CA 94720

Ming Gu
mgu@math.berkeley.edu
Dept. of Mathematics UC Berkeley, Berkeley, CA 95472

In Kmeans, clusters are represented by the centre of mass of their members. It uses an algorithm in which each data point is assigned to the nearest cluster centre and then computing the centroid of its member data vectors and assigning that point as the cluster centre. Another approach 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.

1 Spectral Relaxation

The data-vectors are m dimentional and there are n such data-vectors. The m-by-n data vector matrix

$$A = [a_1, a_2, ..., a_n]$$

A_i is m-by-s_i matrix,i.e., the ith cluster contains the data vectors in A_i,

$$A_i = [a_1^{(i)}, a_2^{(i)}, ..., a_{s_i}^{(i)}]$$

The sum-of-squares cost function is defined as:

$$ss(\Pi) = \sum_{i=1}^{k} \sum_{s=1}^{s_i} (||a_s^{(i)} - m_i||)^2, \qquad m_i = \sum_{s=1}^{s_i} \frac{a_s^{(i)}}{s_i}$$

$$ss(\Pi) = \sum_{i=1}^{k} \left(trace(A_i^T A) - \left(\frac{e^t}{\sqrt{s_i}} \right) A_i^T A \left(\frac{e}{\sqrt{s_i}} \right) \right)$$

$$ss(\Pi) = trace(A^T A) - trace(X^T A^T A X)$$

where X is n-by-k orthonormal matrix.

$$X = \begin{pmatrix} s_1 \\ s_2 \\ . \\ s_k \end{pmatrix} \begin{pmatrix} \frac{e}{\sqrt{s_1}} & & \\ & \frac{e}{\sqrt{s_2}} & \\ . & . & . \\ & & \frac{e}{\sqrt{s_k}} \end{pmatrix} \qquad (1)$$

We need to minimize the cost function which is equivalent to

$$max\{trace(X^T A^T A X)\}$$

And it is easy to see that

$$trace(X^T A^T A X) = \sum_{i=1}^{k} \frac{x_i^T A^T A x_i}{x_i^T x_i} = \sum_{i=1}^{k} \frac{(||A x_i||)^2}{(||x_i||)^2}$$

Left-hand-side can be easily written as the weighted sum of the squared Euclidean norm of the mean vector of each cluster.

$$min\{ss(\Pi)\} \geq trace(A^T A) - max\{trace(X^T A^T A X)\} = \sum_{i=k+1}^{min\{m,n\}} \sigma_i^2(A)$$

where $\sigma_i(A)$ is the i largest singular value of A.

Let H be a symmetric matrix with eigenvalues as:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$$

and the corresponding eigenvectors $U = [u_1, u_2, ..., u_n]$. Then

$$\lambda_1 + \lambda_2 + ... + \lambda_k = max_{X^T X = I_k} trace(X^T H X)$$

Finally,

$$min_{\Pi} ss(\Pi) \geq max_a \sum_{i=k+1}^{min\{m,n\}} \sigma_i^2(A - ae^T)$$

$$ss(\Pi) = \frac{1}{2} \sum_{i=n-k+1}^{n} \lambda_i(W)$$

where $W = (||a_i - a_j||)_{i,j=1}^n$ and λ_i is the eigenvalue.

2 Cluster Assignment Using Pivoted QR Decomposition

The gram matrix of A is:

$$A^T A = \begin{pmatrix} A_1^T A_1 & 0 & ... & 0 \\ 0 & A_2^T A_2 & ... & 0 \\ . & . & . & . \\ 0 & & 0 & ...A_k^T A_k \end{pmatrix} + E = B + E$$

If the overlaps among the clusters represented by the submatrices A_i are small, then the norm of E will be small as compare with the block diagonal matrix B in the above equation. Let the largest eigenvector of $A_i^T A_i$ be y_i , and $A_i^T A_i y_i = \mu_i y_i, \qquad ||y_i|| = 1,$ i = 1,2,...,k.

Let the eigenvalues and eigenvectors of $A^T A$ be:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n, \qquad A^T A x_i = \lambda_i x_i$$

Assume that there is a gap between the two eigenvalue sets $\{\mu_1, \mu_2, ..., \mu_k\}$ and $\{\lambda_{k+1}, ..., \lambda_n\}$, i.e.,

$$0 < \delta = min\{|\mu_i - \lambda_i|, |i = 1, ..., k, j = k+1, ..., n\}$$

By Davis-Kahan sin(θ) theorem,

$$X_k^T = [x_1, x_2, ..., x_k] = Y_k V + O(||E||)$$

where V is an k-by-k orthogonal matrix. Ignoring the $O(||E||)$ term, we get that,

$$X_k^T = [y_{11} v_1, ..., y_{1s_1} v_1, ..., y_{k1} v_k, ..., y_{ks_k} v_k]$$

where $y_i^T = [y_{i1}, ..., y_{is_i}]$ and $V^T = [v_1, ..., v_k]$

When QR decomposition is applied with column pivoting to X_k^T, with a permutation matrix P, such that

$$X_k^T P = QR = Q[R_{11}, R_{12}],$$

where Q is a k-by-k orthogonal matrix, and R_{11} is a k-by-k upper triangular matrix.We then compute the matrix

$$\hat{R} = R_{11}^{-1}[R_{11}, R_{12}]P^T = [I_k, R_{11}^{-1} R_{12}]P^T$$

Then the cluster membership of each data vector is determined by the row index of the largest element in absolute value of the corresponding column of $\hat{R}$

3 Experimental Results

From experiments, the following conclusions are made:

- The two clustering algorithms p-QR and p-Kmeans are comparable to each other, and both are better and sometimes substantially better than K-means in case of binary clustering.
- Both p-QR and p-Kmeans perform better than Kmeans.
- For data sets with small overlaps, p-QR performs better than p-Kmeans