Paper one

New _1-Norm Relaxations and Optimizations for Graph Clustering

Motivation

For graph clustering, traditional methods often use spectral relaxation to solve it. But the approximation solutions may possibly be continuous values while the ideal solution should be discrete values. Thus the authors want to propose a new relaxation in order to get spectral solutions. What's more, the old sparse learning optimization algorithms cannot be applied to solve non-smooth ratio minimization problem. Therefore, the authors are also eager to give a new optimization algorithm to solve this difficult problem.

Related Theories and Definitions

Similarity Matrix: W is the similarity matrix, indicating the distance or similarity defined between i-th element and j-th element in (i,j) position. The most popular way is $W_{ij} = S_{ij} = exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$

Degree Matrix: It is a diagonal matrix whose values in main diagonal are the degrees of every nodes. In a weighted graph, the degree of a node is defined as $D_{ii} = \sum_{j=1}^{n} w_{ij}$.

Normalized Cut: $Ncut = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)},$, where $cut(A,B) = \sum_{i \in A, j \in B} W_{ij}$ and $assoc(A,V) = \sum_{i \in A, j \in V} W_{ij}$. When $r = -\frac{d1}{d2}$, the objective is to minimize the formula following: $\min_{1^T D y = 0} \frac{\frac{1}{2}\sum_{i,j} W_{ij}(y_i - y_j)^2}{\sum_i D_{ii} y_i^2}$

Ratio Cut: $Rcut = \frac{cut(A,B)}{|A|} + \frac{cut(A,B)}{|B|},$, where $|A|$ represents the number of points in A. Similarly, when $r = -\frac{d1}{d2}$, the objective is to minimize the formula following:

$$\min_{1^T y = 0} \frac{\frac{1}{2}\sum_{i,j} W_{ij}(y_i - y_j)^2}{\sum_i y_i^2}$$.

Relaxation: Using its necessary condition rather than sufficient condition, the tide constraint original objective gets relaxation.

Authors' New Idea about Relaxation

Firstly, the authors proof the equation $\frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i-y_j|}{\sum_i |D_{ii}y_i|} = \frac{1}{2}Ncut$, which is not hard to proof. Then the new objective function can be: $\min_{y=[1,\ldots,1,-\frac{d_1}{d_2},\ldots,-\frac{d_1}{d_2}]^T} \frac{\frac{1}{2}\sum_{i,j} W_{ij}|y_i-y_j|}{\sum_i |D_{ii}y_i|}$. In this way, the goal is to minimize a L1-norm function, which will cause more sparse solutions and discrete values thus more close to the ideal solution.

Authors' New Algorithm

The algorithm is designed to solve the objective function with non-smooth terms mentioned in this paper.

1. Initial y such that $1^T Dy = 0$
2. While not converge , do

   Calculate the objective $\lambda = \frac{\frac{1}{2}\sum_{i\in A, j\in B} W_{ij}|y_i - y_j|}{\sum_{i\in A}|D_{ii}y_i|}$;

   Calculate vector b, where $b_i = sign(D_{ii}y_i)$;

   Compute the matrix S, where $S_{ij} = \frac{1}{2|y_i - y_j|}$

   Compute $W = W \bullet S$

   Get $L = D - W$

   Update y by $y = \arg\min_{1^T Dy=0}\left(y^T Ly - \lambda b^T y\right)$

Summary

This paper introduces a new idea about L1-norm relaxation on the basis of the old methods, spectral relaxation, in graph clustering. The L1-norm relaxation will result in sparse solutions and discrete solutions thus more close to the ideal solution. However, since the L1-norm term is in the objective function, those traditional sparse learning algorithms cannot be applied. Therefore, the authors propose a new optimization algorithm to solve this problem with non-smooth term in objective function.

My idea

To better understand this paper, I read some theories in spectral clustering which helps me to know some definitions and theories in this paper that authors did not mention. Comparing with traditional relaxation, I think the creative idea of this paper is turning the squared objective to L1-norm objective, and more importantly, the authors figure out a clever way to solve the non-smooth objective function.

Paper two

Multi-View K-Means Clustering on Big Data

Motivation

Nowadays, there are many multi-view clustering methods. But since they are all graph based approaches, they are computationally expensive thus cannot be applied to large scale datasets. Therefore, the authors want to propose a new method to handle multi-view clustering on big data on the basis of classic K-Means algorithm, which has low computational cost and can be easily applied in parallel.

Related Definitions

$X^{(v)} \in R^{d_v \times n}$ (input data matrix): the input data in v-th view has n elements with d_v dimensions.

$F^{(v)} \in R^{d_v \times K}$ (centroid matrix): there are K centers with d_v dimensions in v-th view.

$G^{(v)} \in R^{n \times K}$ (clustering indicator matrix): 0 or 1 in the matrix, where $x_{ij}^{(v)} = 1$ represents that the i-th element in v-th view belongs to j-th centers.

$\alpha \in R^{M}$ (weight factor): $\alpha^{(v)}$ is the weight for the v-th view.

Main Algorithm

Input: $\{X^{(1)},...,X^{(M)}\}$ where $X^{(v)} \in R^{d_v \times n}$, number of clusters K, parameter γ

Output: clustering indicator matrix G where $G \in R^{n \times K}$, $\{F^{(1)},...,F^{(M)}\}$ where $F^{(v)} \in R^{d_v \times K}$, $\{\alpha^{(1)},...,\alpha^{(M)}\}$

1. Initialization:

   Initialize $G \in R^{n \times K}$ randomly

   Initialize the diagonal matrix $D^{(v)} = I_n$

   Initialize the $\alpha^{(v)} = \dfrac{1}{M}$
2. While not converge, do

Calculate the diagonal matrix $D^{(v)} = (\alpha^{(v)})^{\gamma} D^{(v)}$

Update the centroid matrix $F^{(v)} = X^{(v)} D^{(v)} G \left(G^T D^{(v)} G \right)^{-1}$

Update the cluster indicator matrix G for each data g one by one: $g = e_k$, where $e_k \in I_K = [e_1, \ldots, e_K]$. And which column value in I_K is decided by the following equation $k = \arg\min_j \sum_{v=1}^{M} d^{(v)} \left\| x^{(v)} - F^{(v)} e_j \right\|_2^2$

Update the diagonal matrix $D_{ii}^{(v)} = \frac{1}{2\left\| e^{(v)i} \right\|}$, where $e^{(v)i}$ is the i-th row from

$$E^{(v)} = X^{(v)T} - GF^{(v)T}$$

Update $\alpha^{(v)} = \frac{\left(\gamma H^{(v)}\right)^{\frac{1}{1-\gamma}}}{\sum_{v=1}^{M} \left(\gamma H^{(v)}\right)^{\frac{1}{1-\gamma}}}$

Contribution

This paper proposes an enhanced K-Means algorithm which is capable to handle multi-view clustering work in large scale data. It is not computationally expensive but it can be applied in parallel easily as well. It is also robust since it is not sensitive to outliers and is more stable when changing different initializations than other clustering methods.

My idea

Honestly speaking, unlike the previous paper, I did not understand the mathematical theories well since I cannot tell why to do that in this algorithm. But I am able to read the mathematical equations and expressions and know what to do in this algorithm. From the analysis in the paper and our knowledge about K-Means and spectral clustering, we can clearly understand that K-Means is less computationally expensive than graph based clustering, and K-Means can work on big datasets since it can divide into several data chunks and then update cluster centroid matrix after all data chunk are processed thus suitable to work in parallel. Graph based clustering method is more capable to handle multi-view clustering work before, but now this paper fixes this gap.