

Constrained Spectral Clustering under a Local Proximity Structure Assumption



Qianjun Xu & Marie desJardins, University of Maryland, Baltimore County
Kiri Wagstaff, Jet Propulsion Laboratory

Problem

Clustering data sets that have *local proximity structure*

Technique:

- Spectral clustering
- Constraint incorporation
 - Must-link* constraints
Two points belonging to the same cluster
 - Cannot-link* constraints
Two points belonging to different clusters

Local proximity structure:

Locally, the points belong to the same cluster as their closest neighbors
Globally, the sub-clusters belonging to the same clusters

Example:

XOR data set

Spectral clustering

- Start from n by n similarity matrix A

$$A_{ij} = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

$$D_{ii} = \sum_j A_{ij}$$
- Normalize similarity matrix
 - (a) $P = D^{-1}A$
 - (b) $M = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$
 - (c) $N = (A + d_{\max}I - D) / d_{\max}$
- Compute the eigenvectors of the similarity matrix
- Clustering in the space spanned by the largest k eigenvectors

Motivation

Minimize cut value : $cut = \sum_{k=1}^K \sum_{i \in C_k} \sum_{j \in C_k} A_{ij}$

Different criteria

to balance the cut value and the cluster size

- Ratio cut
- Normalized cut

Understanding the eigenvector

Similar points will have similar values in the eigenvector
If k clusters are well separated, the eigenvector will be *piecewise constant, i.e.*, the points in the same cluster will have the same values while points in different clusters will have different values
If k clusters are not well separated, the eigenvectors will be approximately piecewise constant

The algorithms

- CSC** : (1) uses normalization (a)
 - For must-link constraint, set $A(i,j)=1$
 - For cannot-link constraint, set $A(i,j)=0$
 - Compute the P matrix and its eigenvectors
 - Clustering in the space spanned by P's eigenvectors
- KKM [1]** : similar to CSC, but uses normalization (c)
- CCL [2]** : Constrained Complete Linkage method
 - Impose must-link constraint, set distance(i, j) = 0
 - Propagate must-link constraint by running all-pairs-shortest-path and get new distance metric
 - Impose cannot-link constraint, set distance(i, j)=infinity
 - Run complete-linkage program

Results

Results on Soybean_small data set, 47 instances, 35 attributes, and 4 clusters.

Results on Iris data set, 150 instances, 4 attributes, 3 clusters.

Left: The 2-dimensional plot of Soybean_small data set, which is generated using the multi-dimensional scaling method. There is a small group of square points that is separated from the other square points by elements of the triangle cluster. The overlap between these two clusters will cause problem when the constraints are propagated.

Solution

- Actively select constraints that represent each sub-cluster
- Incorporate constraints
- Results: improved agreement with labels

The 2nd largest eigenvector derived by P matrix, m=9.

The 2nd largest eigenvector derived by P matrix with must-link constraint (1, 21), m=9.

The 2nd largest eigenvector derived by P matrix with 2 must-link constraints (1, 21) and (11, 31), m=9.

Parameter issues

The 2nd largest eigenvector derived by P with the same 2 must-link constraints, m=3.

The 2nd largest eigenvector derived by P with the same 2 must-link constraints, m=14.

The conditions that guarantee a piecewise constant eigenvector:

$$\forall S, \forall i \in S, \exists K_{i,S}, \forall S' [S' \neq S \rightarrow \sum_{j \in S'} P_{ij} = K_{i,S}]$$

This condition implies that for each point i, the sum of its similarity value with intra-cluster points and inter-cluster points must be a constant.

In other words, for each pair of clusters S and S', there exists a constant K which is the sum of the similarity value between any point in S with any point in S', and vice versa.

Parameter selection

$$A_{ij} = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- Select a sigma parameter such that the inter-cluster similarities are approximately 0
- Estimate the local neighborhood
 - For each point, compute the distances to all other points
 - Sort this distance array
 - Find m such that distances $d_{i,m}$ and $d_{i,m+1}$ have largest gap
 - Find the smallest m among all the points
 - Set $\exp(d_{i,m} / 2\sigma^2) = \epsilon$
- The first m items in the sorted distance array correspond to the largest sub-cluster that we guarantee to walk over in our random walk model
- Each point will have similarity value ϵ for its m closest neighbors, and near-0 similarity values for points farther than its mth closest neighbor
- m = 9 for the XOR data set

Example: one item (left, with black circle around it) and its plot of sorted distance array (right). The red circle indicates the m-neighborhood of this item.

Conclusions

- The eigenvectors of the P matrix will help us find each sub-cluster
- Incorporating these constraints will result in piecewise constant eigenvectors that can yield the correct partition for data sets that obey local proximity structure
- The sigma parameter will influence the result; we provide a parameter selection heuristic to solve this problem

References

- [1] Kamvar, S.D.; Klein, D.; and Manning, C.D. 2003. Spectral learning. In *Proceedings of IJCAI-03*, 561-566.
- [2] Klein, D.; Kamvar, S.D.; and Manning, C.D. 2002. From instance-level constraints to space-level constraints. In *Proceedings of ICML-02*, 307-314.

Funding source

This work is supported by NSF IIS-0325329.