Fair Constrained Spectral clustering

Neha Afreen

IIT Guwahati
Department of Computer Science and engineering

Under the guidance of
Prof. V. Vijaya Saradhi

18th May 2023

Content

Spectral Clustering

Data set X is converted into a graph or similarity matrix W. Let us denote D as degree matrix of W and L as the laplacian matrix

$$L = D - W \tag{1}$$

Spectral clustering find an embedding space in which the data points are easily separable by k means algorithm. This embedding space is given by the k smallest eigen vectors of L.

Spectral Clustering

Let us define an encoding H as follows

$$H_{il} = \begin{cases} \frac{1}{\sqrt{|C_l|}}, & i \in C_l \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Spectral Clustering

From "A tutorial on spectral clustering" [1]

Spectral clustering problem

$\Downarrow$

finding minimum cost cut

$\Downarrow$

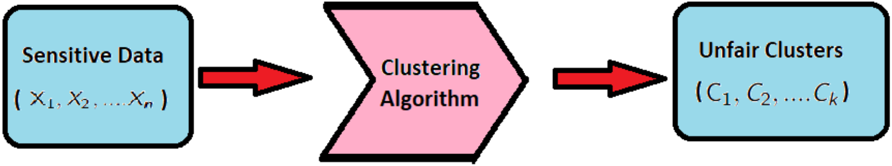
Objective Function of Spectral clustering

finding H that minimize $Tr(H^T LH)$

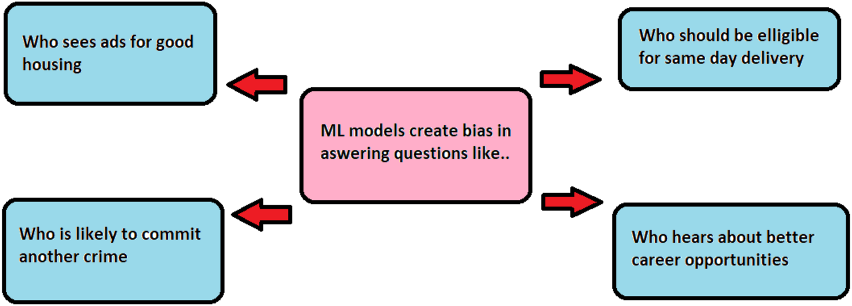
$\Downarrow$

K eigen vectors of L

Why to incorporate fairness

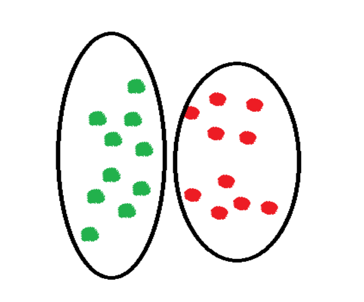


Why to incorporate fairness

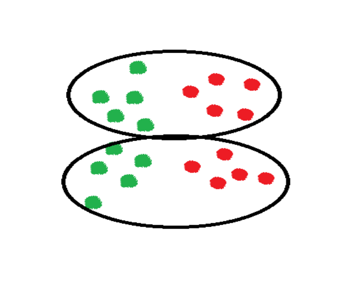


Fairness in clustering

Unfair clustering



Fair clustering



Fairness notion

Balance (A fairness notion): According to Chierichetti et al. (2017) [2], fairness notion is ensured if every cluster contains nearly the same number of data from each group V_s. Balance of any cluster C_l is defined as

$$Balance = \min_{i \in k}\left(\min_{s \in [h]} \frac{|C_i \cap V_s|}{|C_i|}\right) \tag{3}$$

Incorporating fairness in Spectral clustering

Let $f^{(s)}$ defines the group membership of V_s such that

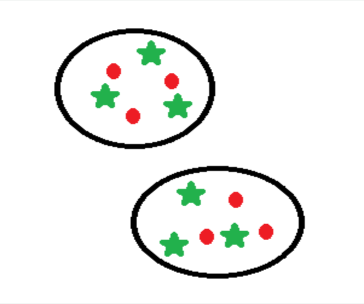
$$f_s^{(i)} = \begin{cases} 1, & \text{if } v_i \in V_s \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

then the bellow equation guarantees fairness for each cluster, l

$$\forall s \in [h-1] : \sum_{i=1}^{n} (f_i^s - \frac{V_s}{n}) H_{il} = 0 \tag{5}$$

Fairness in clustering

Example



Explanation

Defining for star data point and above cluster

$$0 = 3 * (1 - \frac{6}{12}) * (\frac{1}{\sqrt{6}}) + 3 * (0 - \frac{6}{12}) * (\frac{1}{\sqrt{6}}) + 3 * (1 - \frac{6}{12}) * 0 + 3 * (0 - \frac{6}{12}) * 0$$

Incorporating fairness in Spectral clustering

Objective function for Fair sepctral clustering

$$\underbrace{\min_{H} Tr(H^T L H)}_{\text{SC}} \quad \text{subject to} \quad \underbrace{F^T H = 0_{(h-1)xk}}_{\text{FSC}} \tag{6}$$

We can substitute H $= ZY$, where Z $\in$ R$^{n*(n-h+1)}$ and forms the orthonormal subspace of F^T and Y $\in$R$^{(n-h+1)*k}$

$$\min_{Y} Tr(Y^T Z^T L Z Y) \quad \text{subject to} \quad Y^T Y = I_k \tag{7}$$

Embedding space is given by the k smallest eigenvectors of $Z^T L Z$

Increasing accuracy of spectral clustering

In constrained spectral clustering we impose some information as constraints to the data in the form of **Must-link** and **Cannot-link** to enhance accuracy.

If M is the set of must-link constraints then

$$M = (x_i, x_j) \quad | \quad x_i \text{ and } x_j \text{ are similar} \tag{8}$$

If C is the set of cannot-link constraints then

$$C = (x_i, x_j) \quad | \quad x_i \text{ and } x_j \text{ are dissimilar} \tag{9}$$

Increasing accuracy of spectral clustering

The constraint matrix Q is defined as

$$q_{ij} = \begin{cases} -1, & \text{if } (x_i, x_j) \in C \\ +1, & \text{if } (x_i, x_j) \in M \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Increasing accuracy of spectral clustering

Constrained spectral embedding for K-way data clustering[3]

$$J_{cSC} = \gamma J_{SC} + (1-\gamma) J_{CM} \tag{11}$$

$$J_{cSC} = \gamma Tr(H^T L H) + (1-\gamma) Tr(H^T L_Q H) \tag{12}$$

where $L_Q = D_Q - Q$

$$J_{cSC} = H^T(\gamma L + (1-\gamma) L_Q) H = H^T L_{cSC} H \tag{13}$$

where

$$L_{cSC} = D_{cSC} - W_{cSC} \tag{14}$$

$$D_{cSC} = \gamma D + (1-\gamma) D_q \tag{15}$$

$$W_{cSC} = \gamma W + (1-\gamma) Q \tag{16}$$

Increasing accuracy of spectral clustering

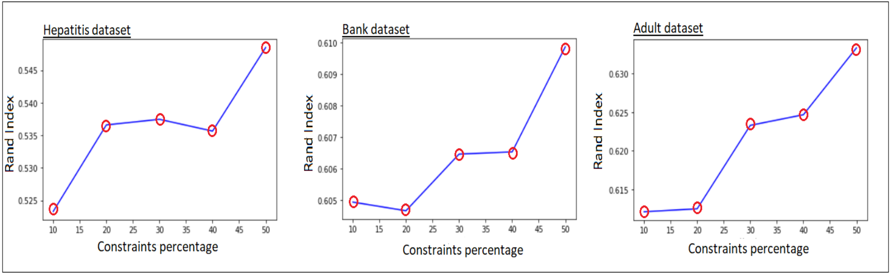
Objective function of constrained spectral clustering

$$\min H^T L_{cSC} H \tag{17}$$

The linearly separable embedding space is given by the eigenmap of L_{cSC}

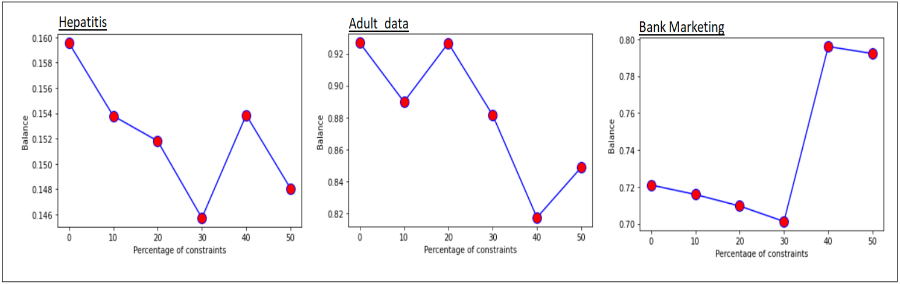
Increasing accuracy of spectral clustering

Results for constrained



Fairness in constrained spectral clustering

Fairness in constrained spectral clustering



Incorporating fairness in constraint set

Definition of fairness in constraints set

In order to improve fairness, the fraction of (v_i, v_j) where i≠j data pairs in must-link constraints set should be close to the fraction of (v_l, v_l) data pairs in cannot-link constraints set if V contains h groups V_s and $1 =< i, j, l <= h$.

Defining few matrices

Let the data $V = v_1, v_2, ..., v_n$ has [h] groups. The constraint matrix Q is defined as

$$q_{ij} = \begin{cases} -1, & \text{if } (x_i, x_j) \in C \\ +1, & \text{if } (x_i, x_j) \in M \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Let us defined two separate matrix Q_M and Q_C

$$Q_M = \begin{cases} +1, & \text{if } (x_i, x_j) \in M \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

and

$$Q_C = \begin{cases} -1, & \text{if } (x_i, x_j) \in C \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Defining few matrices

$group(v_i)$ indicates the group to which the data point v_i belongs to.

$$e = [1, 1, ..., 1]_{1*n} \tag{21}$$

Let Q_M^S can be given as

$$Q_M^S(i,j) = \begin{cases} 1, & \text{if } group(v_i) \neq group(v_j), (v_i, v_j) \in M \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Let Q_C^S can be given as

$$Q_C^S(i,j) = \begin{cases} -1, & \text{if } group(v_i) = group(v_j), (v_i, v_j) \in C \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Defining few matrices

$$\text{Total number of must-link constraints} = |e^T Q_M e| \tag{24}$$

$$\text{Total number of cannot-link constraints} = |e^T Q_C e| \tag{25}$$

Proportion of (v_i, v_j) where i≠j data pairs in must-link constraints set

$$= \frac{e^T(Q_M \odot Q_M^S)e}{|e^T Q_M e|} \tag{26}$$

Proportion of (v_l, v_l) data pairs in cannot-link constraints set

$$= \frac{e^T(Q_C \odot Q_C^S)e}{|e^T Q_C e|} \tag{27}$$

Defining few matrices

Difference between the fractions

$$= \left|\frac{e^T(Q_M \odot Q_M^S)e}{|e^T Q_M e|} - \frac{e^T(Q_C \odot Q_C^S)e}{|e^T Q_C e|}\right| \tag{28}$$

We can define a matrix Q_e such that

$$Q_e = \left|\frac{Q_M \odot Q_M^S}{|e^T Q_M e|} - \frac{Q_C \odot Q_C^S}{|e^T Q_C e|}\right| \tag{29}$$

Spectral clustering with fair constraints

$$\min_{H} \gamma_1 Tr(H^T L H) + \gamma_2 Tr(H^T L_Q H) + \gamma_3 Tr(H^T L_{Qe} H)$$

subject to $\gamma_1 + \gamma_2 + \gamma_3 = 1,$ and $H^T H = I_k$

Taking $\gamma_1 = \gamma_2 = \gamma_3 = \frac{1}{3}$ Then the above objective function will become

$$\frac{1}{3} \min_{H} Tr(H^T (L + L_Q + L_{Qe}) H) \text{ subject to } H^T H = I_k \tag{30}$$

$$\frac{1}{3} \min_{Y} Tr(H^T L_r H) \text{ subject to } H^T H = I_k \tag{31}$$

Where $L_r = L + L_Q + L_{Qe}$

SCFC

- **Input:** 1. Data set $V = x_1, x_2, ..., x_n$
  2. Group membership vectors f_s as given by 2.3
  3. Constraints in the form of ML and CL.
- Make similarity graph
- Find laplacian matrix, L_r
- Compute the k smallest eigenvectors of L_r.(columns of Y)
- Apply k-means clustering to the rows of Y
- **Output:**
  Final Clusters

Fair spectral clustering with fair constraints

$$\min_{H} \gamma_1 Tr(H^T L H) + \gamma_2 Tr(H^T L_Q H) + \gamma_3 Tr(H^T L_{Qe} H)$$

$$\text{subject to} \quad \gamma_1 + \gamma_2 + \gamma_3 = 1,\ H^T H = I_k \quad \text{and} \quad F^T H = 0_{(h-1)xk}$$

Substituting H $= ZY$

$$\frac{1}{3} \min_{Y} Tr(Y^T Z^T (L + L_Q + L_{Qe}) ZY) \ \text{subject to} \ \ Y^T Y = I_k$$

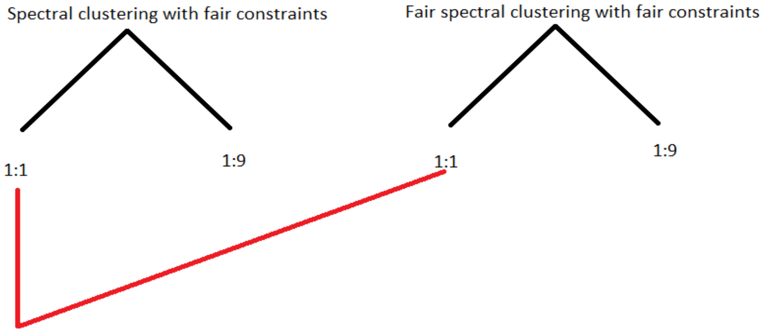
$$\frac{1}{3} \min_{Y} Tr(Y^T Z^T L_p ZY) \ \text{subject to} \ \ Y^T Y = I_k$$

Where $L_P = L + L_Q + L_{Qe}$

FSCFC

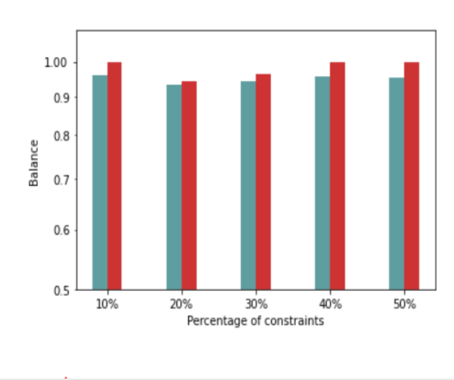
- **Input:** 1. Data set $V = x_1, x_2, ..., x_n$
  2. Group membership vectors f_s as given by 2.3
  3. Constraints in the form of ML and CL.
- Make similarity graph
- Find laplacian matrix, L_p
- Find $F \in R^{nx(h-1)}$ having column vectors $f^{(s)} - \frac{|V_s|}{n}.1_n$ $s \in [h-1]$
- Compute Z whose columns form the orthonormal basis of F^T.
- Compute the k smallest eigenvectors of $Z^T L_P Z$.(columns of Y)
- Apply k-means clustering to the rows of H = ZY
- **Output:**
  Final Clusters

Experimental setup

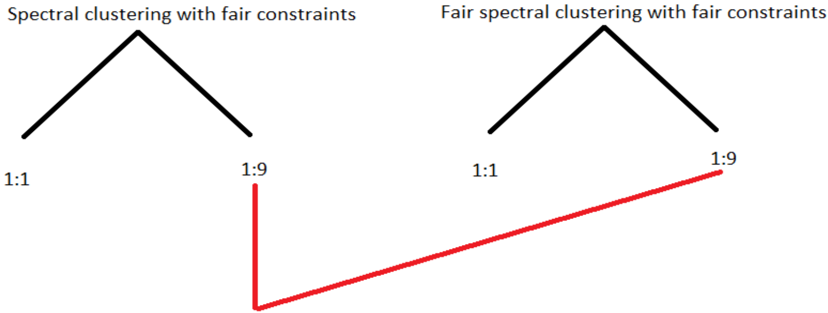


Results

Adult data set

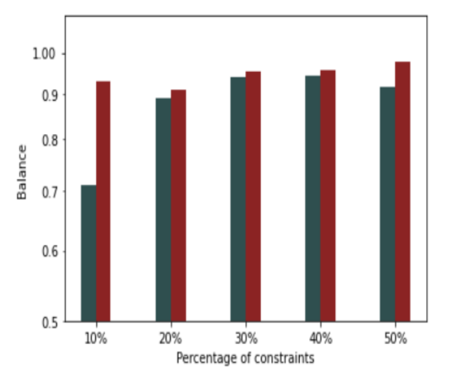


Experimental setup

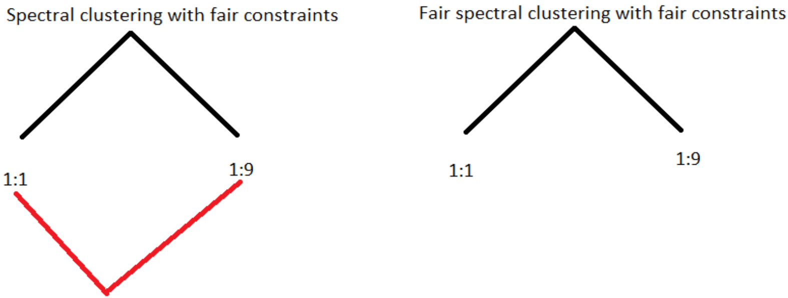


Results

Adult data set

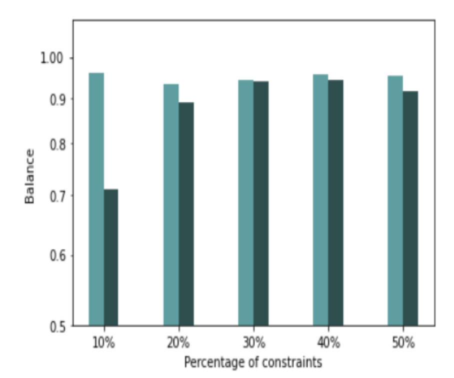


Experimental setup

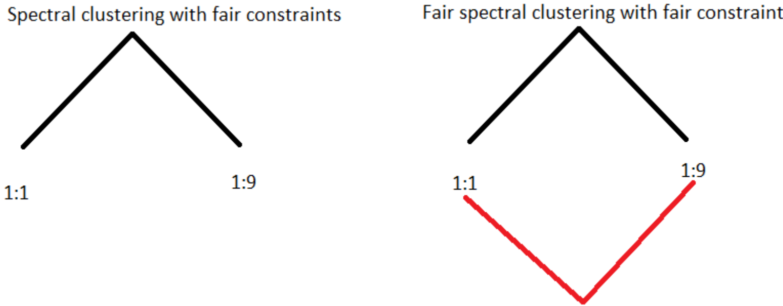


Results

Adult data set

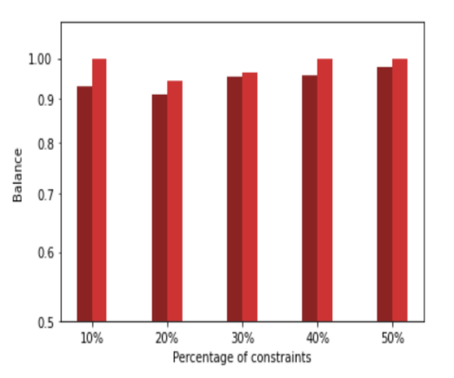


Experimental setup



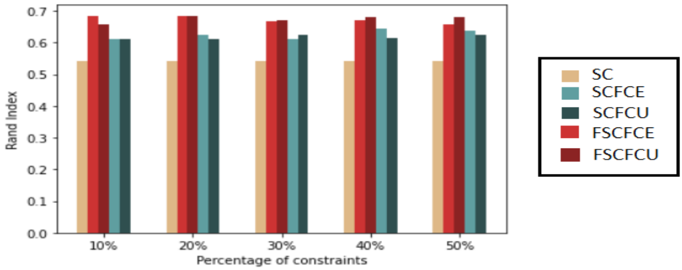
Results

Adult data set



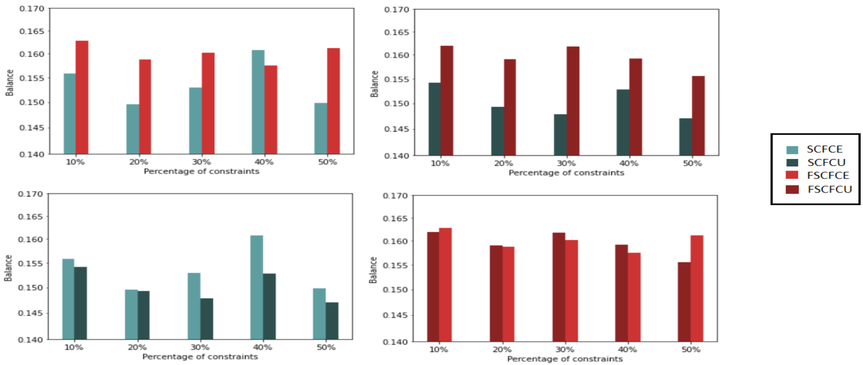
Results

Adult data set



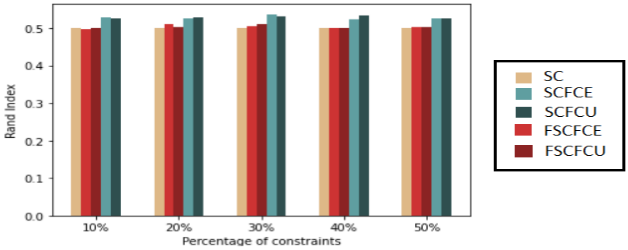
Results

Hepatitis set



Results

Hepatitis data set



References

- Ulrike von Luxburg", "A tutorial on spectral clustering".
- "Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii,S. Fair clustering through fairlets"2017.
- "G. Wacquet, É. Poisson Caillault 1, D. Hamad 1, P.-A. Hébert 1", "Constrained spectral embedding for K-way data clustering"
- "Guarantees for Spectral Clustering with Fairness Constraints"

Thank You!!