Fair Correlation Clustering

Maciej Nemś

18th November 2021

Are algorithms fair?

How to determine if algorithm is fair?

Even if algorithm does not discriminate based on trait A, it can discriminate based on trait B, which is strongly correlated with trait A.

Let us set a hard constraint based on trait A, so that algorithm is fair.

Solution by Fair Correlation Clustering.

Correlation Clustering

Input

Let $G = (V, E)$ be full undirected graph with n vertices. Let $\sigma : E \mapsto \mathbb{R}$ be function assigning labels to edges. Label $\sigma(e)$ can be positive (meaning that connected vertices are similar) or negative (meaning that connected vertices are different).

Clustering

It is a division $\mathcal{C} = \{C_1, C_2, \ldots\}$ of set V into disjoint sets.

$E^+ = \{e \in E : \sigma(e) > 0\}$
$E^- = E \setminus E^+$
$E(C) = E \cap C^2$
$intra(\mathcal{C}) = \bigcup_{C \in \mathcal{C}} E(C)$
$inter(\mathcal{C}) = E \setminus intra(\mathcal{C})$

Cost function

$COST(G, \mathcal{C}) = \sum_{e \in intra(\mathcal{C}) \cap E^-} |\sigma(e)| + \sum_{e \in inter(\mathcal{C}) \cap E^+} |\sigma(e)|$

Fairness

Each vertex $v \in V$ has color $c(v)$.

Proportional Fairness

In each cluster number of vertices of each color must be proportional to the number of this color in whole graph.

α-fraction Fairness

$\alpha \in (0, 1)$. In each cluster number of vertices of each color must be at least α of vertices in this cluster.

Constrained correlation clustering

We have $G = (V, E)$. Let $\mathcal{F}$, be set of subsets of V. $\mathcal{F}$, is a set of allowed clusters with property that: $\forall_{F_1, F_2 \in \mathcal{F}} F_1 \cup F_2 \in \mathcal{F}$.

This constrain is satisfied if all subsets fulfill fairness condition.

Solution of this problem is division to clusters $\mathcal{C}$ with minimal $COST(G, \mathcal{C})$ such that $\forall_{C \in \mathcal{C}} C \in \mathcal{F}$.

Fairlet Decomposition

We call that a division of V for constrained correlation clustering into $\mathcal{P} = \{P_1, P_2, \ldots\}$. We call P_i a fairlet. Every $P_i \in \mathcal{F}$.

$FCOST^{in}$

$FCOST^{in}(P_i) = |E^- \cap intra(P_i)|$ - number of negative edges inside P_i
$FCOST^{in}(\mathcal{P}) = \sum_i FCOST^{in}(P_i)$

$FCOST^{out}$

$FCOST^{out}(P_i, P_j) = min(|E^-(P_i, P_j)|, |E^+(P_i, P_j)|)$
$FCOST^{out}(\mathcal{P}) = \sum_{i<j} FCOST^{out}(P_i, P_j)$

Fairlet Decomposition cost function

$FCOST(\mathcal{P}) = FCOST^{in}(\mathcal{P}) + FCOST^{out}(\mathcal{P})$

Reduction of Constrained Correlation Clustering

We have graph G and fairlet decomposition $\mathcal{P}$. Let $G^{\mathcal{P}}$ be a full graph with vertices: $\{p_1, \ldots, p_{|\mathcal{P}|}\}$, where p_i is $P_i \in \mathcal{P}$. Let $\sigma(p_i, p_j)$ be $max(|E^-(P_i, P_j)|, |E^+(P_i, P_j)|)$ multiplied by more frequent sign.

If number of edges $+, -$ is equal, we choose arbitrarily.

We get a reduced instance of Correlation Clustering (which is unconstrained)

Algorithm 1: Constrained Correlation Clustering

Result: Fair Clustering on G

1 Compute approximation of $\mathcal{P}$ Fairlet Decomposition;

2 Create graph $G^{\mathcal{P}}$: (p_i, p_j) has more frequent sign $E(P_i, P_j)$ and weight $max(|E^{-}(P_i, P_j)|, |E^{+}(P_i, P_j)|)$;

3 Compute approximation of $\mathcal{C}$ Correlation Clustering on $G^{\mathcal{P}}$;

4 **return** $\{\bigcup_{p_j \in C_i} P_j \ : \ C_i \in \mathcal{C}\}$

Lemma 3.1

Having $G, \mathcal{P}, \mathcal{C}$ there exists clustering $\mathcal{C}'$ on $G^{\mathcal{P}}$ such that:

$$COST(G^{\mathcal{P}}, \mathcal{C}') \leq COST(G, \mathcal{C}) + FCOST^{out}(\mathcal{P})$$

Lemma 3.2

Let $\mathcal{C}$ be clustering $G^{\mathcal{P}}$ and $\mathcal{C}'$ be clustering computed in step 4 of algorithm. Then:

$$COST(G, \mathcal{C}') \leq COST(G^{\mathcal{P}}, \mathcal{C}) + FCOST(\mathcal{P})$$

Lemma 3.3

For any constrained correlation clustering $\mathcal{C}$ of graph G there exists fairlet decomposition $\mathcal{P}$ fulfilling:

$$FCOST(\mathcal{P}) \leq COST(G, \mathcal{C})$$

Lemma 3.4

Assuming there exists A_α, α-approximation of searching for minimal fairlet decomposition and A_β, β-approximation for unconstrained correlation clustering, Algorithm 1 is $(\beta(1+\alpha)+\alpha)$-approximation for constrained correlation clustering.

By Lemma 3.3 we know there exists fairlet decomposition G with cost at most $COST(G, OPT)$.

A_α returns $\mathcal{P}$: $FCOST(\mathcal{P}) \leq \alpha COST(G, OPT)$.

By Lemma 3.1 we know that $G^{\mathcal{P}}$ has solution with cost at most $(1+\alpha)COST(G, OPT)$.

A_β returns $\mathcal{C}$ with cost at most $\beta(1+\alpha)COST(G, OPT)$.

By Lemma 3.2 we know that cost of clustering returned by Algorithm 1, is at most $(\beta(1+\alpha)+\alpha)COST(G, OPT)$.

We also know it is a constrained clustering solution, because sum of fairlets is in constrained clustering.

There exists β-approximation for correlation clustering

Authors show $\min(\log n, 2\rho r^2)$-approximation, where:

n - number of vertices

$r = \frac{\max_{P \in \mathcal{P}} |P|}{\min_{P \in \mathcal{P}} |P|}$

ρ is approximation factor of unweighted correlation clustering

Fairlet Decomposition Approximation

Reduction to fair clustering (without correlation), of problem from another paper. More precisely reduction to optimization of k-median. For fairlet division $\mathcal{P} = \{P_1, P_2, \ldots\}$ and metric space (M, d) we have:

$$MCOST(P_i) = \min_{u \in M} \sum_{v \in P_i} d(u, v)$$

$$MCOST(P) = \sum_{P_i \in \mathcal{P}} MCOST(P_i)$$

We want to minimize $MCOST(\mathcal{P})$

Reduction to fair clustering

We want to define metric space (M, d) for which *MCOST* approximates *FCOST*.

Let us define $\phi : V \to [0,1]^n$:

$$\phi(u)_v = \begin{cases} 1 & \text{if } u = v \text{ or } (u,v) \in E^+ \\ 0 & \text{if } (u,v) \in E^- \end{cases}$$

Let $M = [0,1]^n$, a $d(u,v) = |\phi(u) - \phi(v)|$

Reduction to fair clustering

Lemma 4.1

For any fairlet decomposition $\mathcal{P}$ we have:

$$MCOST(\mathcal{P}) \leq 2FCOST(\mathcal{P})$$

Lemma 4.2

For any fairlet decomposition $\mathcal{P}$ and $f = \max_{P \in \mathcal{P}} |P|$ we have:

$$FCOST(\mathcal{P}) \leq 2f \cdot MCOST(\mathcal{P})$$

With assumption there exists γ-approximation for fairlet decomposition with optimizing *MCOST*, which generates fairlets of size at most f we get $(4f\gamma)$-approximation for finding minimal *FCOST*.

Examples of approximating MCOST

Algorithms for $\alpha = 1/2$ and $\alpha = 1/C$

ON BLACKBOARD

Tested datasets

Amazon

Vertices represent products on page Amazon.

Colors are categories of products.

All products watched together have edge $+1$. All products not watched together have edge -1.

Authors take 1000 products equally divided in two categories.

reuters, victorian

Sets of English texts written by 16 different authors.

Vertices represent texts, colors represent authors.

For each pair of texts we create semantic embeding.

For top $\theta \in \{0.25, 0.5, 0.75\}$ edges embeding is computed by scalar product with $+1$, and for the rest with -1

Tested algorithms

Algorithm 1

For $\alpha = 1/2$ and $\alpha = 1/C$. To solve Correlation Clustering uses algorithm Local

Local

Solves Correlation Clustering based on local search.

Pivot

Randomised algorithm solving correlation clustering.

Single

Whole graph is a single cluster.

Rand

Random division into fairlets.

Measurements

ERROR

Ratio of incorrect edges (intra with -1 or inter with $+1$) to all edges.

Imbalance

Ratio of vertices breaking fairness rule to all vertices.

Results and experiments

Dataset	Unfair Alg. Error		Unfair Alg. Imbalance		Fair Alg. Error		
	Local	Pivot	Local	Pivot	Match + Local	Single	Rand
amazon	0.010	0.011	0.40	0.39	0.064	0.786	0.215
reuters, $\theta = 0.25$	0.096	0.161	0.64	0.59	0.230	0.754	0.255
reuters, $\theta = 0.50$	0.181	0.231	0.50	0.40	0.350	0.504	0.502
reuters, $\theta = 0.75$	0.188	0.241	0.15	0.25	0.199	0.252	0.746
victorian, $\theta = 0.25$	0.109	0.158	0.53	0.46	0.212	0.753	0.251
victorian, $\theta = 0.50$	0.183	0.268	0.31	0.23	0.348	0.502	0.499
victorian, $\theta = 0.75$	0.203	0.280	0.12	0.12	0.237	0.251	0.747
mean over datasets	0.139	0.193	0.38	0.35	0.234	0.459	0.543

Table 1: Error and Imbalance in $C = 2$ color case for various datasets and different threshold θ for the quantile used for positive edges. Notice how our algorithm Match + Local has cost comparable to Pivot and not much higher than Local while reducing the imbalance from the up 65% of the unfair algorithms to 0.

Results and experiments

Algorithm	ERROR	IMBALANCE for 1/2	IMBALANCE for equality
LOCAL	0.249	0.011	0.218
PIVOT	0.345	0.008	0.191
MATCH + LOCAL	0.255	0	0.180
REP. MATCH + LOCAL	0.321	0	0
SINGLE	0.5	0	0
RAND	0.5	0	0

Table 2: Experimental results for victorian, $\theta = 0.50$, using $C = 8$ colors.

Results and experiments

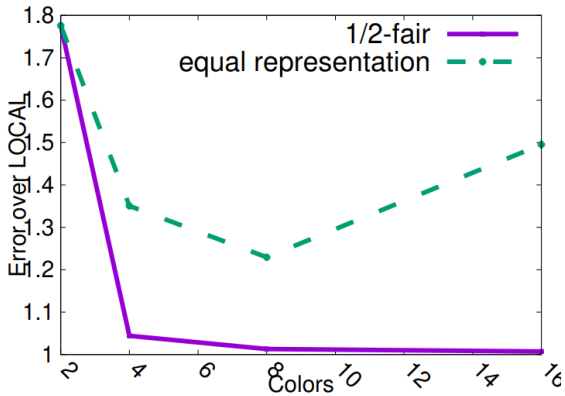

Figure 1: ERROR of our algorithms over that of the unfair LOCAL algorithms for $\alpha = 1/2$ and $\alpha = 1/C$, on a series of graphs from victorian, $\theta = 0.50$, and using $C = 2$ to $C = 16$ colors.

Bibliography I


S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian.
Fair correlation clustering.
CoRR, abs/2002.02274, 2020.


F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii.
Fair clustering through fairlets, 2018.