Efficient and Effective Optimal Transport-Based Biclustering

Chakib Fettal [1,2] Lazhar Labiod [1] Mohamed Nadif [1]

[1]Centre Borelli, Université Paris Cité

[2]Informatique CDC

Overview

Objective: Proposing a more efficient and effective optimal transport-based biclustering approach.

Approach:

- $\rightarrow$ Proposing two formulations of the block seriation problem through the lens of optimal transport. One that is hard (BCOT) and one that is fuzzy (BCOT_λ).
- $\rightarrow$ Solving the resulting problems via block coordinate descent which boils down to solving intermediate OT problems.

Outcome: Experimental results seem to support the superiority of our approach w.r.t other OT biclustering algorithms.

Optimal Transport

Given an a cost matrix $\mathbf{M}$, a source distribution $\mathbf{w}$, and a target distribution $\mathbf{v}$.

Discrete optimal transport:

$$\mathsf{OT}(\mathbf{M}, \mathbf{w}, \mathbf{v}) \triangleq \min_{\mathbf{Z} \in \Pi(\mathbf{w}, \mathbf{v})} \langle \mathbf{M}, \mathbf{Z} \rangle \tag{1}$$

Entropic discrete optimal transport:

$$\mathsf{OT}_\lambda(\mathbf{M}, \mathbf{w}, \mathbf{v}) \triangleq \min_{\mathbf{Z} \in \Pi(\mathbf{w}, \mathbf{v})} \langle \mathbf{M}, \mathbf{Z} \rangle - \lambda H(\mathbf{Z}) . \tag{2}$$

Biclustering

Bipartite Graphs:

$$\mathbf{A} = \begin{pmatrix} \mathbf{0}_{n\times n} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0}_{d\times d} \end{pmatrix}. \tag{3}$$

The matrix B is called *biadjacency matrix*.

Block Seriation:

$$\max_{\substack{\mathbf{Z}\in\Gamma(n,k)\\ \mathbf{W}\in\Gamma(d,k)}} \sum_{i,j,h} b_{ij} z_{ih} w_{jh}. \tag{4}$$

Biclustering using Optimal Transport

Let $L(\mathbf{B})$ be the anti-biadjacency of a bipartite graph such that $L(\mathbf{B})_{ij}$ discrepancy measure between node i and j.

We define the OT biclustering problem as

$$\mathrm{BCOT}(\mathbf{w}, \mathbf{v}, \mathbf{r}, \mathbf{c}) \triangleq \min_{\substack{\mathbf{Z} \in \Pi(\mathbf{w}, \mathbf{r}) \\ \mathbf{W} \in \Pi(\mathbf{v}, \mathbf{c})}} \left\langle L(\mathbf{B}), \mathbf{Z}\mathbf{W}^\top \right\rangle \tag{5}$$

$\rightarrow$ The transport plans $\mathbf{Z}$ and $\mathbf{W}$ resemble hard partition matrices.

Fuzzy Biclustering using Regularized OT

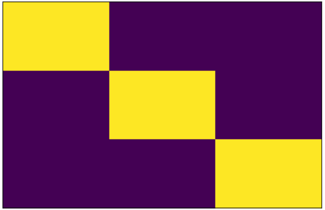
The previous formulation can be sped up.

We use entropic regularization to propose a fuzzy and faster to solve problem

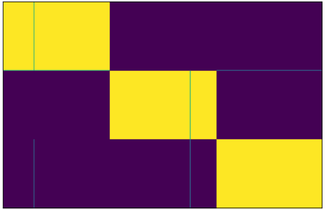
$$\mathrm{BCOT}_\lambda(\mathbf{w}, \mathbf{v}, \mathbf{r}, \mathbf{c}) \triangleq \min_{\substack{\mathbf{Z} \in \Pi(\mathbf{w}, \mathbf{r}) \\ \mathbf{W} \in \Pi(\mathbf{v}, \mathbf{c})}} \left\langle L(\mathbf{B}), \mathbf{Z}\mathbf{W}^\top \right\rangle - \lambda_{\mathbf{Z}} H\left(\mathbf{Z}\right) - \lambda_{\mathbf{W}} H\left(\mathbf{W}\right) \tag{6}$$

$\rightarrow$ The transport plans $\mathbf{Z}$ and $\mathbf{W}$ resemble fuzzy partition matrices.

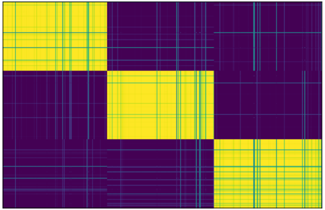
Biclustering Structure of Different Models



(a) Block Seriation.



(b) BCOT.



(c) BCOT_λ.

Figure: Biclusters formed using three different methods on the Pubmed dataset:
- Classical block seriation results in a hard biclustering.
- BCOT results in an almost hard.
- BCOT_λ results in a soft biclustering.

Optimization

Algorithm 1: BCOT

Input : $\mathbf{B}$ bi-adjacency matrix,
$\mathbf{w}$ and $\mathbf{v}$ row and column weights,
$\mathbf{r}$ and $\mathbf{c}$ row and column exemplar distributions.

Output: π^r, π^c row and column partitions

$\mathbf{W} \leftarrow \mathbf{W}_{init}$;

while *not converged* **do**

$\quad \mathbf{Z} \leftarrow \arg \mathsf{OT} \left(L(\mathbf{B})\mathbf{W}, \mathbf{w}, \mathbf{r}\right)$;

$\quad \mathbf{W} \leftarrow \arg \mathsf{OT} \left(L(\mathbf{B})^\top \mathbf{Z}, \mathbf{v}, \mathbf{c}\right)$;

end

Generate π^r, π^c from $\mathbf{Z}$ and $\mathbf{W}$;

Examples of Special Cases of our Model

- Modularity Maximization in Bipartite Graphs.
  $\rightarrow$ BCOT with $L(\mathbf{B}) = -(\mathbf{B} - \frac{1}{b_{\cdot\cdot}}\mathbf{B}\mathbf{1}\mathbf{1}^\top\mathbf{B})$.

- Modularity-Based Sparse Soft Graph Clustering.
  $\rightarrow$ BCOT$_\lambda$ with $L(\mathbf{B}) = -(\mathbf{B} - \frac{1}{b_{\cdot\cdot}}\mathbf{B}\mathbf{1}\mathbf{1}^\top\mathbf{B})$.

- Directional Co-clustering with a Conscience.
  $\rightarrow$ BCOT with $L(\mathbf{B}) = -\mathbf{B}$ and cluster size binding.

- Bipartite Correlation Clustering.
  $\rightarrow$ BCOT with

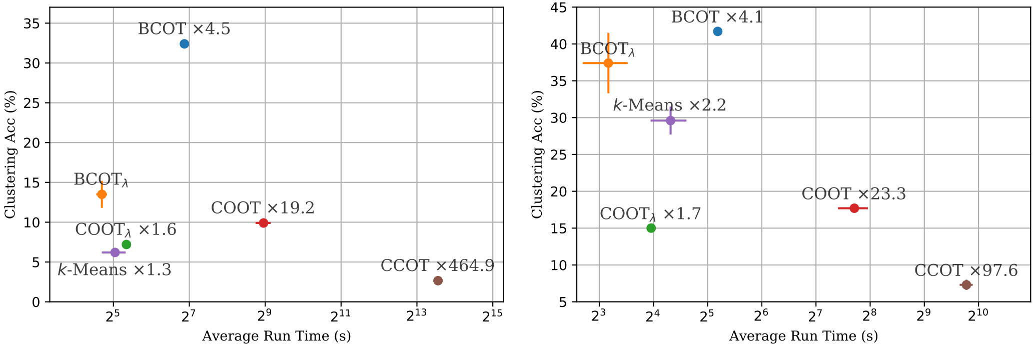
$$L(\mathbf{B})_{ij} = \begin{cases} -1 & \text{if } b_{ij} > 0 \\ +1 & \text{otherwise} \end{cases} \tag{7}$$

Experiments on Synthetic Data

Table: Biclustering performance on synthetic datasets.

Method	**A**	**B**	**C**	**D**
k-means	**100.0±0.0**	95.0±5.0	95.3±4.0	96.6±4.7
CCOT	54.4±3.5	70.0±0.0	29.7±0.4	55.7±1.8
CCOT-GW	99.1±0.0	83.5±0.0	83.4±0.0	75.3±0.0
COOT	99.8±0.0	78.8±2.0	99.8±0.0	93.7±1.2
COOT_λ	39.9±2.4	84.9±4.6	28.2±0.0	60.7±0.0
BCOT	99.8±0.0	80.4±2.2	99.6±0.1	91.3±0.7
BCOT_λ	**100.0±0.0**	99.1±0.4	**100.0±0.0**	**100.0±0.0**
BCOT (gnd $\mathbf{r}$, $\mathbf{c}$)	same $\mathbf{r}$, $\mathbf{c}$	99.9±0.0	same $\mathbf{r}$, $\mathbf{c}$	95.5±2.3
BCOT_λ (gnd $\mathbf{r}$, $\mathbf{c}$)	same $\mathbf{r}$, $\mathbf{c}$	**100.0±0.0**	same $\mathbf{r}$, $\mathbf{c}$	99.2±0.9

Experiments on Document-Term Matrices



(a) 20 Newsgroups. (b) Ohscal.

Figure: Accuracy against training time on NG20 and Ohscal.

Thank you.