Semi-Supervised Clustering Based on Signed Total Variation

Peter Berger, Thomas Dittrich, and Gerald Matz

Institute of Telecommunications, Technische Universität Wien, Austria
<firstname>.<lastname>@nt.tuwien.ac.at

Introduction

Motivation

- Problem: semi-supervised clustering, i.e., splitting a dataset into disjoint classes under the assumption that the cluster affiliation is known for certain data points
- Assumption: nodes within a cluster are similar and nodes from different clusters are dissimilar
- Example social network: similarity links $\leftrightarrow$ follower/friends dissimilarity links $\leftrightarrow$ blocking or quoting behavior
- Question: how can dissimilarity information be incorporated into total variation based clustering

Contributions

- Introduce the signed total variation
- Formulate semi-supervised two-class clustering with dissimilarity based on the signed total variation
- Introduce a suitable ℓ_1 regularization to ensure reliable clustering even when only few labels are known
- Develope a low-complexity ADMM-based algorithm

Modeling of the data

- Data is represented by a graph $\mathcal{G}(\mathcal{V}, \mathbf{W})$ with node set $\mathcal{V} = \{1, \ldots, N\}$ and weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$
- $\mathcal{V}^+$ and $\mathcal{V}^- = \mathcal{V} \backslash \mathcal{V}^+$ denote the clusters
- Modeling of the clusters: label vector $\mathbf{x} \in \mathbb{R}^N$ with $x_i = 1$ for $i \in \mathcal{V}^+$ and $x_i = -1$ for $i \in \mathcal{V}^-$
- Denote sampled nodes by $\mathcal{L} \subset \mathcal{V}$, $\mathcal{L}^+ = \{i \in \mathcal{L} : x_i = 1\}$, $\mathcal{L}^- = \{i \in \mathcal{L} : x_i = -1\}$

Total variation based unsigned clustering

- Consider unsigned weight matrix $\mathbf{W}$, $W_{ij} \geq 0$
- A positive weight $W_{ij} > 0$ models similarity between i and j
- Min-cut approach determines $\mathcal{V}^+$ and $\mathcal{V}^- = \mathcal{V} \backslash \mathcal{V}^+$ via

$$\min_{\mathcal{V}^+} \sum_{j \in \mathcal{V}^+} \sum_{i \in \mathcal{V} \backslash \mathcal{V}^+} W_{ij} \quad \text{s.t.} \quad \mathcal{L}^+ \subseteq \mathcal{V}^+, \ \mathcal{L}^- \subseteq \mathcal{V} \backslash \mathcal{V}^+ \qquad (1)$$

- Constrained total variation minimization:

$$\min_{\mathbf{x}} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} |x_i - x_j| \, W_{ij} \quad \text{s.t.} \quad x_i = -1 \text{ for } i \in \mathcal{L}^-, \ x_i = 1 \text{ for } i \in \mathcal{L}^+ \qquad (2)$$

- If the min-cut problem (1) has a unique solution $\{\mathcal{V}^-, \mathcal{V}^+\}$, then (2) yields the equivalent solution

$$x_i = \begin{cases} -1, & i \in \mathcal{V}^-, \\ 1, & i \in \mathcal{V}^+ \end{cases}$$

Signed Clustering

Signed Laplacian

- Negative weight $W_{ij} < 0$ models dissimilarity between i and j
- Signed graph Laplacian: $\bar{\mathbf{L}} = \bar{\mathbf{D}} - \mathbf{W}$ with the signed degree matrix $\bar{\mathbf{D}} = \operatorname{diag}\{\bar{d}_1, \ldots, \bar{d}_N\}$, $\bar{d}_i = \sum_{j=1} |W_{ij}|$
- Induced Laplacian form:

$$\mathbf{x}^T \bar{\mathbf{L}} \mathbf{x} = \frac{1}{2} \sum_i \sum_j (x_i - \operatorname{sign}(W_{ij}) x_j)^2 |W_{ij}|$$

- For negative edge weights, $(x_i - \operatorname{sign}(W_{ij}) x_j)^2 |W_{ij}| = (x_i + x_j)^2 |W_{ij}|$ will be small if $x_i \approx -x_j$

Signed total variation

- This motivates the new concept of the signed total variation:

$$\|\mathbf{x}\|_{\text{TV}} \triangleq \sum_i \sum_j |x_i - \operatorname{sign}(W_{ij}) x_j| \, |W_{ij}|$$

- The signed total variation $\|\mathbf{x}\|_{\text{TV}}$ is a semi-norm and convex
- For unbalanced graphs (contains a cycle with an odd number of edges with negative weight) it is a norm

Regularization

- Problem 1: total variation minimization tends to declare (one of) the label sets $\mathcal{L}^+$, $\mathcal{L}^-$ as clusters
- Problem 2: the signed total variation tends to assign zero values since both $|x_i + x_j|$ and $|x_i - x_j|$ can be minimized by setting $x_i = x_j = 0$
- Regularized signed total variation clustering problem:

$$\min_{\mathbf{x}} \ \|\mathbf{x}\|_{\text{TV}} + \lambda^- \sum_{i \in \mathcal{N}^-} |1 + x_i| + \lambda^+ \sum_{i \in \mathcal{N}^+} |1 - x_i| \quad \text{s.t.} \quad x_i = -1 \text{ for } i \in \mathcal{L}^-, \ x_i = 1 \text{ for } i \in \mathcal{L}^+, \qquad (3)$$

where

$$\mathcal{N}(i) = \{j \in \mathcal{V} \backslash \mathcal{L} : W_{ij} > 0\},$$
$$\mathcal{N}(\mathcal{A}) = \bigcup_{i \in \mathcal{A}} \mathcal{N}(i) \text{ for } \mathcal{A} \subset \mathcal{V},$$
$$\mathcal{N}^- = \mathcal{N}(\mathcal{L}^-) \backslash \mathcal{N}(\mathcal{L}^+), \quad \mathcal{N}^+ = \mathcal{N}(\mathcal{L}^+) \backslash \mathcal{N}(\mathcal{L}^-)$$

- Regularization terms with λ^- and λ^+ are introduced to assign $x_i = 1$ $(x_i = -1)$ to the majority of nodes in $\mathcal{N}^+$ $(\mathcal{N}^-)$
- Regularization parameters can be tuned automatically, see Algorithm 1

Algorithm

- Propose augmented ADMM to solve (3)
- Resulting algorithm can be implemented in a distributed manner

Algorithm 1 Signed TV clustering with parameter tuning

```
Input: W, L⁻, L⁺, x_min (slightly smaller than 1)
  Initialize: λ⁻ = 0, λ⁺ = 0
  repeat
      calculate minimizer x of (3)
      M⁻ = {i ∈ N⁻ : x_i < 0}
      M⁺ = {i ∈ N⁺ : x_i > 0}
      x⁻ = min_{i∈M⁻} |x_i|
      x⁺ = min_{i∈M⁺} |x_i|
      a = 0
      if M⁻ = ∅ or x⁻ < x_min then
          increase λ⁻, a = 1
      end if
      if M⁺ = ∅ or x⁺ < x_min then
          increase λ⁺, a = 1
      end if
  until a = 0
Output: x
```

Simulations

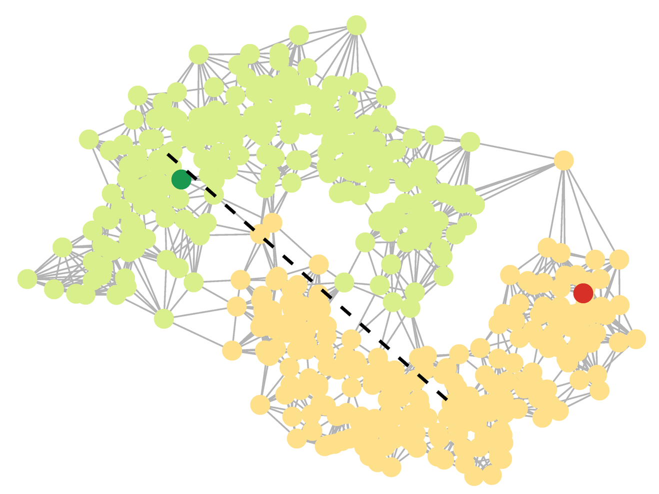
Setup

- Simulations on two-moon datasets with $N = 500$ nodes
- Coordinates of each node generated from a random angle on a center curve and Gaussian jitter (variance $\sigma^2 = 0.09$)
- Graph generated as kNN graph with $k = 10$ neighbors and Gaussian kernel for edge weights (parameter $\sigma_1 = 0.6$)
- M samples drawn randomly while ensuring at least one known label from each cluster
- L randomly chosen dissimilarity edges between pairs of nodes from different clusters

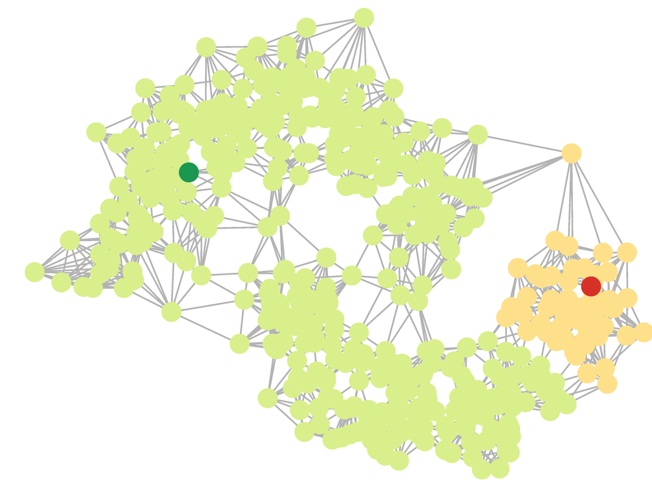
Illustrative example

- Different colors represent different clusters
- Sampled nodes represented by dark colors
- Dissimilarity edges represented by dashed lines

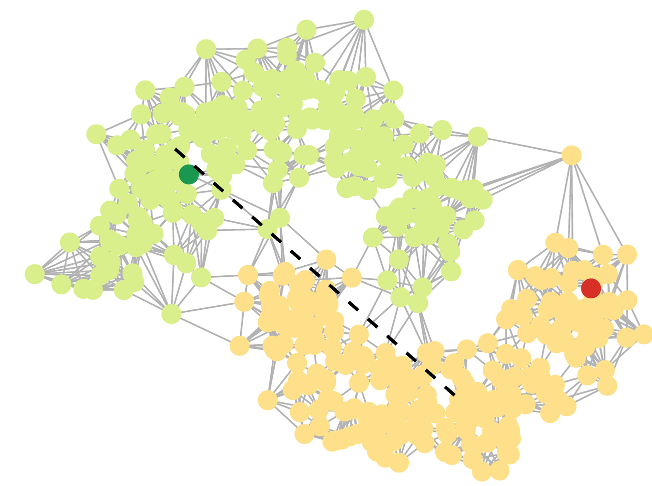
Ground truth



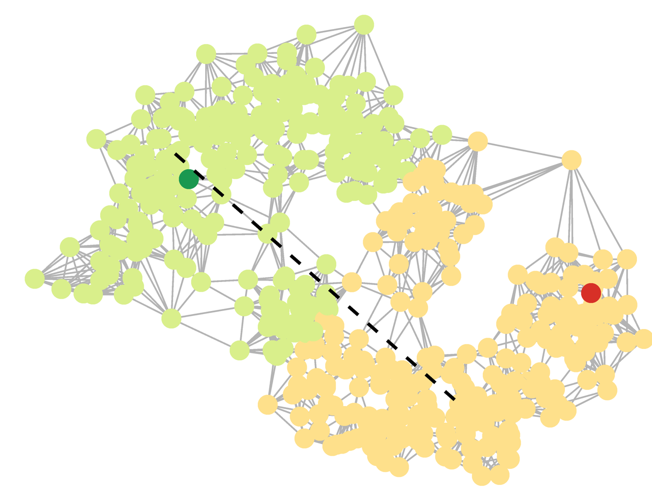
Unsigned total variation



Signed total variation



Laplacian regularized least squares with dissimilarity (parameters determined by grid search) [Goldberg et al., PMLR'07]



Monte Carlo simulations

Error rates in percent (mean and standard deviation)

	Algorithm 1			LapRLSd		
	$M=2$	$M=5$	$M=10$	$M=2$	$M=5$	$M=10$
$L=0$	7.3 ± 12.5	4.0 ± 9.0	1.8 ± 5.1	13.6 ± 9.2	12.8 ± 8.8	6.1 ± 6.2
$L=5$	3.0 ± 9.1	1.2 ± 3.5	1.0 ± 2.4	8.4 ± 7.2	5.8 ± 4.7	3.4 ± 3.2
$L=10$	1.4 ± 6.0	0.9 ± 1.9	0.7 ± 0.7	5.0 ± 5.4	3.6 ± 3.5	2.5 ± 2.1

Discussion

- Incorporation of dissimilarity substantially improves performance
- Total variation is directly connected to a minimum cut and therefore outperforms Laplacian based algorithms
- Most state of the art algorithms have free parameters
- Proposed algorithm has no free parameters

Funded by WWTF Grant ICT15-119