SEMI-SUPERVISED CLUSTERING BASED ON SIGNED TOTAL VARIATION

Peter Berger, Thomas Dittrich, and Gerald Matz

Institute of Telecommunications, TU Wien (Vienna, Austria)
Email: firstname.lastname@nt.tuwien.ac.at

ABSTRACT

We consider the problem of semi-supervised clustering on signed graphs that model similarity and dissimilarity relations between nodes. We introduce a signed version of total variation and use it to formulate a convex optimization problem for the cluster labels. This optimization problem includes a 1-norm regularization to cover cases where only few cluster labels are known. We propose an ADMM-based algorithm to solve the optimization problem. The complexity of this algorithm scales linearly with the number of edges of the graph. Our scheme is suitable for distributed implementation and can therefore efficiently handle large-dimensional applications. Numerical experiments confirm that our clustering scheme is superior to existing methods.

1. INTRODUCTION

We address the problem of graph-based semi-supervised clustering, i.e., splitting a dataset characterized by a graph into disjoint classes ("clusters") under the assumption that the cluster affiliation is known for certain data points. Classical semi-supervised learning algorithms (e.g., [1–4]) use unsigned graphs in which edges connect similar data points. However, there are numerous problems, where dissimilarity information is available and helpful. Examples include constrained image segmentation [5] or the prediction of political positions [6]. In the latter example, in a social network similarity links can be extracted from follower/friendship relations and dissimilarity edges can be derived from blocking or quoting behavior [6]. Dissimilarity information can be modeled using signed graphs [7] and has led to clustering algorithms based on signed Laplacians [6, 7]. By contrast, we use total variation (TV) instead of the Laplacian quadratic form, which has the advantage of favoring locally constant functions and being connected to a minimum cut. The following list summarizes our contributions:

- We introduce a signed total variation measure that incorporates dissimilarity information in signed graphs;
- the signed total variation is shown to be a norm whenever the underlying graph is unbalanced;
- we formulate semi-supervised clustering with dissimilarity as a constrained minimization of signed total variation;
- we introduce a suitable ℓ_1 regularization to ensure reliable clustering even when only few labels are known;
- a low-complexity ADMM-based algorithm is developed for constrained signed total variation minimization;
- we illustrate accurate clustering performance of our scheme via numerical experiments with synthetic data.

Funding by WWTF Grant ICT15-119.

2. BACKGROUND

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ with vertex set $\mathcal{V} = \{1, \ldots, N\}$, edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and edge weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$. We first consider unsigned graphs with non-negative weights $W_{ij} \geq 0$ that quantify the similarity between nodes i and j. Our goal is to identify clusters of similar nodes when the cluster labels of some nodes are known. We restrict ourselves to two clusters, leaving the case of multiple clusters to future work. We denote the two clusters by $\mathcal{V}^+ \subset \mathcal{V}$ and $\mathcal{V}^- \subset \mathcal{V}$. The clusters must be nontrivial and every node has to belong to exactly one cluster; thus, $\mathcal{V}^+ \cup \mathcal{V}^- = \mathcal{V}$ and $\mathcal{V}^+ \cap \mathcal{V}^- = \emptyset$ (equivalently, $\mathcal{V}^- = \mathcal{V} \backslash \mathcal{V}^+$). The clusters can equivalently be described by a label vector $\mathbf{x} \in \mathbb{R}^N$ with $x_i = 1$ for $i \in \mathcal{V}^+$ and $x_i = -1$ for $i \in \mathcal{V}^-$. We assume that for a given set $\mathcal{L} \subset \mathcal{V}$ the cluster labels x_i, $i \in \mathcal{L}$, are known. We define $\mathcal{L}^+ = \{i \in \mathcal{L} : x_i = 1\}$ and $\mathcal{L}^- = \{i \in \mathcal{L} : x_i = -1\}$ such that $\mathcal{L}^+ \cup \mathcal{L}^- = \mathcal{L}$. With the min-cut approach from [2], the clusters $\mathcal{V}^+$ and $\mathcal{V}^- = \mathcal{V} \backslash \mathcal{V}^+$ are obtained via

$$\min_{\mathcal{V}^+} \gamma(\mathcal{V}^+, \mathcal{V} \backslash \mathcal{V}^+) \qquad \text{s.t.} \quad \mathcal{L}^+ \subseteq \mathcal{V}^+, \ \mathcal{L}^- \subseteq \mathcal{V} \backslash \mathcal{V}^+, \quad (1)$$

where the graph cut is defined as

$$\gamma(\mathcal{V}^+, \mathcal{V}^-) \triangleq \sum_{i \in \mathcal{V}^-} \sum_{j \in \mathcal{V}^+} W_{ij}.$$

The min-cut problem is related to the TV minimization

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

with the constraint set

$$\mathcal{Q} \triangleq \{\mathbf{x} \in \mathbb{R}^N : \ x_i = -1 \text{ for } i \in \mathcal{L}^-, \ x_i = 1 \text{ for } i \in \mathcal{L}^+\}.$$

Indeed, we have the following important equivalence (stated without proof due to lack of space, see also [1]).

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

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

A problem with (1) is its potential tendency to declare the known label sets as clusters (i.e., $\mathcal{V}^+ = \mathcal{L}^+$, $\mathcal{V}^- = \mathcal{V}\backslash\mathcal{L}^+$ or $\mathcal{V}^- = \mathcal{L}^-$, $\mathcal{V}^+ = \mathcal{V}\backslash\mathcal{L}^-$). This problem can be mitigated by favoring clusters $\mathcal{V}^-$ and $\mathcal{V}^+$ that have similar size. For example, clusters of similar size can be achieved through the use of the balanced cut (Cheeger cut) [8] or the ratio cut in spectral clustering [9].

The Laplacian form on the graph $\mathcal{G}$ is given by $\mathbf{x}^T\mathbf{L}\mathbf{x} = \frac{1}{2}\sum_i\sum_j (x_i - x_j)^2\, W_{ij}$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix formed with the diagonal degree matrix $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_N\}$, $d_i = \sum_{j=1} W_{ij}$. Numerous clustering and learning schemes use the Laplacian form instead of the TV [1, 3, 9–11]. One of the main reasons for this is that the latter is not differentiable. However, there has been enormous recent progress in developing efficient algorithms for non-differentiable optimization problems, cf. [12–15].

3. SIGNED TV MINIMIZATION

A core idea in [6, 7] is to model dissimilarity between two nodes i and j via a graph edge with negative weight $W_{ij} < 0$. The resulting signed graph can be characterized by its signed Laplacian $\overline{\mathbf{L}} = \overline{\mathbf{D}} - \mathbf{W}$ with $\overline{\mathbf{D}} = \mathrm{diag}\{\bar{d}_1, \ldots, \bar{d}_N\}$, $\bar{d}_i = \sum_{j=1} |W_{ij}|$. The induced Laplacian form reads

$$\mathbf{x}^T\overline{\mathbf{L}}\mathbf{x} = \frac{1}{2}\sum_i\sum_j (x_i - S_{ij}x_j)^2\, |W_{ij}|,$$

where we used the shorthand notation $S_{ij} = \mathrm{sign}(W_{ij})$. Note that for edges connected via negative edge weights, $(x_i - S_{ij}x_j)^2|W_{ij}| = (x_i + x_j)^2|W_{ij}|$ will be small if $x_i \approx -x_j$. This motivates us to introduce the new concept of the signed TV, given by

$$\|\mathbf{x}\|_{\mathrm{TV}} \triangleq \sum_i\sum_j |x_i - S_{ij}x_j|\, |W_{ij}|. \tag{3}$$

Note that for unsigned graphs, $\|\mathbf{x}\|_{\mathrm{TV}}$ simplifies to ordinary TV. Furthermore, the signed TV is a convex function. This is a consequence of the following result.

Theorem 2. *The signed total variation* $\|\mathbf{x}\|_{\mathrm{TV}}$ *in* (3) *is a semi-norm; for unbalanced graphs it is even a norm.*

Proof. The fact that the signed TV is a semi-norm follows from the easily verified properties $\|\mathbf{x}\|_{\mathrm{TV}} \geq 0$, $\|c\mathbf{x}\|_{\mathrm{TV}} = |c|\, \|\mathbf{x}\|_{\mathrm{TV}}$, and $\|\mathbf{x} + \mathbf{y}\|_{\mathrm{TV}} \leq \|\mathbf{x}\|_{\mathrm{TV}} + \|\mathbf{y}\|_{\mathrm{TV}}$.

A connected graph is called unbalanced if it contains a cycle with an odd number of edges with negative weight; otherwise the graph is balanced. In [7, Thm. 4.2] it was shown that a graph is balanced if and only if there exists an $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{x}\overline{\mathbf{L}}\mathbf{x} = 0$. Furthermore, we have that $\mathbf{x}\overline{\mathbf{L}}\mathbf{x} = 0$ if and only if $\|\mathbf{x}\|_{\mathrm{TV}} = 0$. Hence, for unbalanced graphs $\mathbf{x}\overline{\mathbf{L}}\mathbf{x} = \|\mathbf{x}\|_{\mathrm{TV}} = 0$ implies $\mathbf{x} = \mathbf{0}$, thus confirming that for such graphs $\|\mathbf{x}\|_{\mathrm{TV}}$ is a norm. □

Minimizing the signed TV $\|\mathbf{x}\|_{\mathrm{TV}}$ subject to the constraint $\mathbf{x} \in \mathcal{Q}$ suffers from two difficulties, specifically when only few labels are known. First, TV minimization tends to declare (one of) the label sets $\mathcal{L}^+$, $\mathcal{L}^-$ as clusters. Second, for unlabeled nodes the TV terms $|x_i - S_{ij}x_j|\, |W_{ij}|$ can be made zero by setting $x_i = x_j = 0$, no matter whether the edge weight has positive or negative sign; in this case, deciding on the cluster label would reduce to a coin flip.

Both of these difficulties can be resolved by introducing an ℓ_1 regularization term that exploits the fact that for any given node, the majority of neighbor nodes connected via positive edge weights will have the same cluster label. The unlabeled similar neighbors of a node $i \in \mathcal{V}$ are defined by

$$\mathcal{N}(i) = \{j \in \mathcal{V}\backslash\mathcal{L} : W_{ij} > 0\}.$$

Furthermore, for any set $\mathcal{A} \subset \mathcal{V}$ we define $\mathcal{N}(\mathcal{A}) = \bigcup_{i\in\mathcal{A}} \mathcal{N}(i)$. We next define the sets of unlabeled nodes that are in the similarity neighborhood of either $\mathcal{L}^-$ or $\mathcal{L}^+$ but not both, i.e.,

$$\mathcal{N}^- = \mathcal{N}(\mathcal{L}^-)\backslash\mathcal{N}(\mathcal{L}^+), \qquad \mathcal{N}^+ = \mathcal{N}(\mathcal{L}^+)\backslash\mathcal{N}(\mathcal{L}^-). \tag{4}$$

It is reasonable to impose that the number of nodes in $\mathcal{N}^-$ with positive label and the the number of nodes in $\mathcal{N}^+$ with negative label should be small. Since these cardinality constraints are difficult for optimization, we use a convex ℓ_1 relaxation, leading to the convex signed TV clustering problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_{\mathrm{TV}} + \lambda^- \sum_{i\in\mathcal{N}^-} |1 + x_i| + \lambda^+ \sum_{i\in\mathcal{N}^+} |1 - x_i|, \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{Q}. \end{aligned} \tag{5}$$

Note that our ℓ_1 relaxation is simpler and more elegant than the TV renormalization in [8]. The fewer positive or negative cluster labels are known, the more we can enforce label similarity by increasing the respective regularization parameter λ^- and λ^+.

4. ADMM FOR TV CLUSTERING

We next show how to solve (5) in an efficient manner via augmented ADMM [13]. Potential alternative non-smooth optimization algorithms for TV minimization are coordinate descent [15] and (preconditioned) primal-dual methods (e.g., [12, 14]). However, augmented ADMM combines the advantages of being amenable to distributed implementation, having a well-tested stopping criterion, and featuring a varying penalty strategy [16]. We next develop the update steps of augmented ADMM for (5). To this end, we note that the

signed TV can be written as $\|\mathbf{x}\|_{\mathrm{TV}} = \|\nabla_{\mathcal{G}}\mathbf{x}\|_1$ where $\nabla_{\mathcal{G}} : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ is the signed gradient operator defined by

$$(\nabla_{\mathcal{G}}\mathbf{x})_{ij} = (S_{ij}x_j - x_i)|W_{ij}|. \tag{6}$$

Furthermore, we use the signed divergence operator $\mathrm{div}_{\mathcal{G}} : \mathbb{R}^{N\times N} \to \mathbb{R}^N$ that can be derived (cf. [17]) as the negative adjoint of the signed gradient operator ($\mathrm{div}_{\mathcal{G}} = -\nabla_{\mathcal{G}}^*$):

$$(\mathrm{div}_{\mathcal{G}}\,\mathbf{Z})_i \triangleq \sum_{j\in\mathcal{V}} |W_{ij}|Z_{ij} - |W_{ji}|Z_{ji}S_{ji}. \tag{7}$$

Using the characteristic function

$$\chi_{\mathcal{Q}}(\mathbf{x}) \triangleq \begin{cases} 0, & \mathbf{x}\in\mathcal{Q}, \\ \infty, & \text{else}, \end{cases}$$

we can rewrite (5) in the form

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{Z}} \|\mathbf{Z}\|_1 + \lambda^- \sum_{i\in\mathcal{N}^-} |1+x_i| + \lambda^+ \sum_{i\in\mathcal{N}^+} |1-x_i| + \chi_{\mathcal{Q}}(\mathbf{x}), \\ &\text{s.t. } \nabla_{\mathcal{G}}\mathbf{x} = \mathbf{Z}. \end{aligned} \tag{8}$$

This is exactly the form suitable for augmented ADMM (cf. [13, eq. (3)]). The detailed derivation of the ADMM updates for (8) are omitted due to lack of space (see, e.g., [12]). The resulting scheme is summarized in Algorithm 1 (with $\sigma_\tau(x) = \mathrm{sign}(x)\max(0, |x| - \tau)$ denoting the soft-thresholding function). According to [13, Theorem 1], Algorithm 1 is guaranteed to converge to an optimal point of (5) for any penalty parameter $\rho > 0$. The cluster labels are then obtained by taking the signs of a minimizing vector $\mathbf{x}$.

Stopping Criterion and Varying Penalty Strategy. We use the stopping criterion from [16] which is based on the primal and dual residuals [13, 16]

$$\begin{aligned} \mathbf{r}^{(k+1)} &= \frac{1}{\rho}(\mathbf{Z}^{(k+1)} - \mathbf{Z}^{(k)}), \\ \mathbf{R}^{(k+1)} &= -\rho\,\mathrm{div}_{\mathcal{G}}\,\nabla_{\mathcal{G}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \\ &\quad - \mathrm{div}_{\mathcal{G}}(2\mathbf{Z}^{(k)} - \mathbf{Z}^{(k-1)} - \mathbf{Z}^{(k+1)}). \end{aligned} \tag{9}$$

Algorithm 1 will be stopped whenever the following two conditions are both satisfied:

$$\begin{aligned} \|\mathbf{r}^{(k)}\|_2 &\le \epsilon_{\mathrm{abs}}\sqrt{|\mathcal{E}|} + \epsilon_{\mathrm{rel}}\,\|\nabla_{\mathcal{G}}\mathbf{x}^{(k)}\|_{\mathrm{F}}, \\ \|\mathbf{R}^{(k)}\|_{\mathrm{F}} &\le \epsilon_{\mathrm{abs}}\sqrt{N} + \epsilon_{\mathrm{rel}}\,\|\,\mathrm{div}_{\mathcal{G}}\,\mathbf{Z}^{(k)}\|_2. \end{aligned} \tag{10}$$

Here ϵ_{abs} and ϵ_{rel} are the absolute and relative tolerance, respectively [16]. The choice of the penalty parameter ρ can heavily influence the convergence speed of the Algorithm. We used the varying penalty strategy described in [16], which automatically adjusts the parameter ρ to keep the primal and dual residual norms roughly at the same size.

Complexity. The signed graph gradient (6) (step 10 of Algorithm 1) and the signed divergence (7) (step 7) can be determined using the local neighborhood of a node (where $W_{ij} \neq 0$); their overall computation thus requires a number of operations that scales linearly with the number of edges of the graph. Furthermore, the matrices $\mathbf{Z}^{(k)}$ and $\tilde{\mathbf{Z}}^{(k)}$ are sparse, i.e., their elements in row i and column j is nonzero only if $W_{ij} \neq 0$. Since steps 8 and 9 of Algorithm 1 amount to per-node processing, the total number of operations in one iteration of Algorithm 1 scales linearly with the number of edges of the graph. The same observation imply that the algorithm can be implemented in a distributed manner (cf. [12]). Therefore, our signed TV clustering method is able to handle large-dimensional datasets in an efficient manner.

Algorithm 1 signed TV clustering

Input: $\mathbf{W}$, $\mathcal{L}^-$, $\mathcal{L}^+$, λ^-, λ^+

Initialization

1: $k = 0$, $\mathbf{Z}^{(0)} = \mathbf{Z}^{(-1)} = \mathbf{0}$, $\rho = 5$

2: $v_i = 2\sum_{j\in\mathcal{V}}(W_{ij}^2 + W_{ji}^2)$,

3: $c_i^+ = \lambda^+/(\rho v_i)$, $c_i^- = \lambda^-/(\rho v_i)$,

4: $x_i^{(0)} = \begin{cases} 1, & i\in\mathcal{L}^+ \\ -1, & i\in\mathcal{L}^- \\ 0, & \text{else} \end{cases}$

5: determine $\mathcal{N}^+$ and $\mathcal{N}^-$ via (4)

Iterations

6: **repeat**

7: $\quad \mathbf{z}^{(k)} = \mathrm{div}_{\mathcal{G}}(2\mathbf{Z}^{(k)} - \mathbf{Z}^{(k-1)})$

8: $\quad \tilde{x}_i^{(k)} = \begin{cases} x_i^{(k)}, & i\in\mathcal{L} \\ x_i^{(k)} + \frac{1}{\rho v_i} z_i^{(k)}, & \text{else} \end{cases}$

9: $\quad x_i^{(k+1)} = \begin{cases} 1 + \sigma_{c_i^+}(\tilde{x}_i^{(k)} - 1), & i\in\mathcal{N}^+ \\ -1 + \sigma_{c_i^-}(\tilde{x}_i^{(k)} + 1), & i\in\mathcal{N}^- \\ \tilde{x}_i^{(k)}, & \text{else} \end{cases}$

10: $\quad \tilde{\mathbf{Z}}^{(k)} = \mathbf{Z}^{(k)} + \rho\nabla_{\mathcal{G}}\mathbf{x}^{(k+1)}$

11: $\quad Z_{ij}^{(k+1)} = \min\{1, \max\{-1, \tilde{Z}_{ij}^{(k)}\}\}$

12: $\quad k = k + 1$

13: **until** stopping criterion is satisfied

Output: $\hat{\mathbf{x}} = \mathbf{x}^{(k)}$

Relaxation Parameter Adaptation. It remains to choose λ^- and λ^+ appropriately. Recall that the regularization terms with λ^- and λ^+ were introduced in order to assign a cluster label of $x_i = 1$ ($x_i = -1$) to the majority of nodes in $\mathcal{N}^+$ ($\mathcal{N}^-$). Furthermore, the cluster labels within $\mathcal{N}^-$ and $\mathcal{N}^+$ should be close to 1 in magnitude. These observations inspire a tuning of the relaxation parameters according to Algorithm 2, which uses prescribed parameter sets $\boldsymbol{\lambda}^+$ and $\boldsymbol{\lambda}^-$ whose elements are assumed to be sorted in increasing order.

	Algorithm 2			LapSVMd			LapRLSd		
	$M=2$	$M=5$	$M=10$	$M=2$	$M=5$	$M=10$	$M=2$	$M=5$	$M=10$
$L=0$	7.3 ± 12.5	4.0 ± 9.0	1.8 ± 5.1	13.7 ± 9.2	12.6 ± 8.6	6.1 ± 6.2	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	10.1 ± 9.1	9.7 ± 8.1	4.8 ± 5.4	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	7.7 ± 9.3	7.2 ± 8.0	3.4 ± 4.0	5.0 ± 5.4	3.6 ± 3.5	2.5 ± 2.1

Table 1: Error rates in percent (mean and standard deviation) achieved by our scheme and two other methods for various numbers of dissimilarity edges (L) and known labels (M).

5. NUMERICAL EXPERIMENTS

We created a dataset $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ consisting of $N = 500$ random vectors from the two moon model (cf. [18])

$$\mathbf{u}_i = \begin{pmatrix} x_i/2 \\ 0 \end{pmatrix} + \begin{pmatrix} \cos(\varphi_i) \\ x_i \sin(\varphi_i) \end{pmatrix} + \mathbf{n}_i$$

where $x_i \in \{-1, 1\}$ are randomly drawn cluster labels, $\varphi_i \sim \mathcal{U}(0, \pi)$ is a random angle, and $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is Gaussian noise with $\sigma = 0.3$. A graph was created using the k-nearest-neighbor method [19] with $k = 10$ and the edge weights chosen as $W_{ij} = \exp\left(-\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 / \kappa^2\right)$ with $\kappa^2 = 0.72$. We then added dissimilarity edges with weight $W_{ij} = -5$ between L randomly chosen pairs of nodes from different clusters and picked a set $\mathcal{L} = \mathcal{L}^- \cup \mathcal{L}^+$ of $M = |\mathcal{L}|$ known cluster labels (chosen randomly while ensuring at least one known label from each cluster).

We then clustered the graph using our scheme (Algorithm 2) and the algorithms from [6], i.e., Laplacian Regularized Least Squares with dissimilarity (LapRLSd) and Laplacian Support Vector Machines with dissimilarity (LapSVMd). We chose the algorithms from [6] since they incorporate both, labeled data and dissimilarity information. The relaxation vectors in Algorithm 2 were chosen as $\boldsymbol{\lambda}^- = \boldsymbol{\lambda}/|\mathcal{N}^-|$, $\boldsymbol{\lambda}^+ = \boldsymbol{\lambda}/|\mathcal{N}^+|$, with

$$\boldsymbol{\lambda} = [0, 1, 2, 3, 4, 5, 7, 10, 20, 50, 100, 500].$$

Furthermore, we used $x_{\min} = 0.9$. The relative and absolute tolerances were set to $\epsilon_{\text{rel}} = 10^{-3}$ and $\epsilon_{\text{abs}} = 10^{-7}$ (cf. (10)). For LapRLSd and LapSVMd we used the Gaussian (RBF) kernel $k(\mathbf{u}_i, \mathbf{u}_j) = \exp\left(-\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 / \kappa^2\right)$ and we used the optimal regularization parameters γ_A and γ_I, determined by searching the set 10^l, $l = -3, \ldots, 4$.

The clustering performance is quantified by an error rate defined as the percentage of mislabeled nodes among the set of nodes without prior known label, i.e.,

$$r(\hat{\mathbf{x}}) = \frac{|\{i \in \mathcal{V} \backslash \mathcal{L} : \hat{x}_i \neq x_i\}|}{|\mathcal{V} \backslash \mathcal{L}|}. \tag{11}$$

The resulting error rates (mean and standard deviation) obtained over 500 Monte-Carlo runs and different values of L and M are shown in Table 1. It is seen that Algorithm 2 clearly outperforms LapSVMd and LapRLSd for all parameter configurations. Our scheme achieves 3 to 4 times smaller error rates, particularly when there are fewer a priori known cluster labels and more dissimilarity edges. As expected, the performance of all schemes improves with increasing M and increasing L. However, only our method achieves an error rate below 1% for the least difficult setup with $L = M = 10$.

Algorithm 2 signed TV clustering with parameter tuning

Input: $\mathbf{W}$, $\mathcal{L}^-$, $\mathcal{L}^+$, $x_{\min}$, $\boldsymbol{\lambda}^+$, $\boldsymbol{\lambda}^-$

Initialization: $m = 0$, $n = 0$

repeat
  $\lambda^- = \boldsymbol{\lambda}^-(m),\ \lambda^+ = \boldsymbol{\lambda}^+(n)$
  $\hat{\mathbf{x}} \leftarrow$ **Algorithm 1**$(\mathbf{W}, \mathcal{L}^-, \mathcal{L}^+, \lambda^-, \lambda^+)$
  $\mathcal{M}^- = \{i \in \mathcal{N}^- \colon \hat{x}_i < 0\}$
  $\mathcal{M}^+ = \{i \in \mathcal{N}^+ \colon \hat{x}_i > 0\}$
  $\hat{x}^- = \min_{i \in \mathcal{M}^-} |\hat{x}_i|$
  $\hat{x}^+ = \min_{i \in \mathcal{M}^+} |\hat{x}_i|$
  $a = 0$
  if $\mathcal{M}^- = \emptyset$ **or** $\hat{x}^- < x_{\min}$ **then**
    $m \leftarrow m + 1,\ a = 1$
  end if
  if $\mathcal{M}^+ = \emptyset$ **or** $\hat{x}^+ < x_{\min}$ **then**
    $n \leftarrow n + 1, a = 1$
  end if
until $a = 0$

Output: $\hat{\mathbf{x}}$

6. CONCLUSIONS

In this paper, we considered the problem of semi-supervised clustering on signed graphs that incorporate information about the similarity and dissimilarity of data points. We introduced the concept of signed total variation and used it to formulate a suitably regularized convex cluster label optimization problem. A fast ADMM-based algorithm was proposed to solve this problem. Our numerical experiments demonstrated that our approach is superior to existing schemes and accurately recovers the cluster labels even in difficult conditions (few known labels and dissimilarity edges).

REFERENCES

[1] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. Int. Conf. Machine Learning*, pages 912–919, Washington, DC, USA, Aug. 2003.

[2] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. Int. Conf. Machine Learning*, pages 19–26, San Francisco, CA, USA, July 2001.

[3] W. Liu, J. Wang, and S. Chang. Robust and scalable graph-based semisupervised learning. *Proc. IEEE*, 100(9):2624–2638, Sept. 2012.

[4] K. Avrachenkov, P. Chebotarev, and A. Mishenin. Semi-supervised learning with regularized laplacian. *Optimization Methods and Software*, 32(2):222–236, 2017.

[5] X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Min. Knowl. Discov.*, 28(1):1–30, Jan. 2014.

[6] A. B. Goldberg, X. Zhu, and S. Wright. Dissimilarity in graph-based semi-supervised classification. In *Proc. Int. Conf. on Artificial Intelligence and Statistics*, pages 155–162, San Juan (Puerto Rico), Mar. 2007.

[7] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. W. De Luca, and S. Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proc. SIAM Int. Conf. Data Mining*, pages 559–570, Columbus (OH), May 2010.

[8] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Multiclass total variation clustering. In *Advances in Neural Information Processing Systems 26*, pages 1421–1429, Lake Tahoe (NV), Dec. 2013.

[9] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[10] X. Wang and I. Davidson. Flexible constrained spectral clustering. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 563–572, Washington, DC, USA, July 2010.

[11] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proc. Int. Conf. Machine Learning*, pages 824–831, Bonn, Germany, Aug. 2005.

[12] P. Berger, G. Hannak, and G. Matz. Graph signal recovery via primal-dual algorithms for total variation minimization. *IEEE J. Sel. Topics in Signal Processing*, 11(6):842–855, Sept. 2017.

[13] Y. Zhu. An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem. *J. Comput. Graph. Statist.*, 26(1):195–204, Feb. 2017.

[14] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *Proc. Int. Conf. Computer Vision*, pages 1762–1769, Barcelona (Spain), Nov. 2011.

[15] P. Berger, G. Hannak, and G. Matz. Coordinate descent accelerations for signal recovery on scale-free graphs based on total variation minimization. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pages 1739–1743, Kos (Greece), Aug. 2017.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.

[17] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Model. Simul.*, 7(3):1005–1028, Nov. 2008.

[18] T. Bühler and M. Hein. Spectral clustering based on the graph p-Laplacian. In *Proc. Int. Conf. Machine Learning*, pages 81–88, Montreal, Quebec, Canada, June 2009.

[19] K. Ozaki, M. Shimbo, M. Komachi, and Y. Matsumoto. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In *Proc. Conf. Computational Natural Language Learning*, pages 154–162, Portland, Oregon, USA, June 2011.