SEMI-SUPERVISED MULTICLASS CLUSTERING BASED ON SIGNED TOTAL VARIATION

Peter Berger, Thomas Dittrich, Gabor Hannak, 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 for multiple (more than two) classes. The proposed clustering algorithm uses the (dis)similarity of given data to learn the unknown cluster labels. We quantify label (dis)similarity in terms of the new concept of signed total variation (TV). The clustering task is formulated as a convex optimization problem with an ℓ_1-norm regularization term that helps when only few labels are known. We solve the optimization problem by developing an ADMM-based algorithm whose per-iteration complexity scales linearly with the number of edges and the number of clusters. Our algorithm admits a distributed implementation and can therefore efficiently handle large-dimensional problems. Numerical experiments demonstrate the superiority of our scheme.

1. INTRODUCTION

We address the problem of graph-based semi-supervised clustering, i.e., grouping the nodes of a graph under the assumption that the cluster affiliation is known for certain data points. Classical semi-supervised learning algorithms (e.g., [1–4]) group the nodes based on similarity relations between nodes. However, there are numerous problems where some nodes are known to have different class labels. Incorporating such dissimilarity information extends the range of applications and can significantly improve the clustering accuracy. Noticeable applications include constrained image segmentation [5] or the prediction of political positions [6]. Signed graphs can model (dis)similarity information [7] and motivated clustering algorithms based on signed Laplacians [6, 7]. For unsigned graphs it is well known that clustering based on TV outperforms Laplacian clustering [8–10]. In [11] we found that the same is true for two clusters in signed graphs.

Contributions. We generalize the signed TV based semi-supervised clustering approach from [11] to more than two classes. Furthermore, we introduce an ℓ_1 regularization that improves clustering performance in cases where only few cluster labels are known. We show that this regularization can be directly incorporated into the TV term via a weight adjustment of the edges incident to the sampled nodes. We develop a low-complexity distributed ADMM-based algorithm for signed total variation minimization that solves the regularized clustering problem. Our numerical experiments demonstrate accurate clustering performance of our scheme.

Funding by WWTF Grant ICT15-119.

2. BACKGROUND

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ with vertex set $\mathcal{V} = \{1, \dots, N\}$, edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and edge weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$. We admit for signed graphs, i.e., graphs with possibly negative edge weights. In *similarity-based/unsigned clustering* the node set $\mathcal{V}$ is partitioned into K clusters $\mathcal{V}_1, \dots, \mathcal{V}_K$ ($\bigcup_{k=1}^{K} \mathcal{V}_k = \mathcal{V}$, $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$) based on the idea that nodes are more similar within a cluster than across clusters. In *signed clustering*, dissimilarity relations between nodes are taken into account besides similarity relations. Dissimilarity of two nodes indicates that they likely belong to different clusters.

Unsigned clustering. The amount of similarity between two nodes i and j is captured by the non-negative weight $W_{ij} \geq 0$. The goal is to determine the unknown clusters $\mathcal{V}_1, \dots, \mathcal{V}_K$ when the graph topology is given in terms of the non-negative weight matrix $\mathbf{W}$ and the number of clusters K is known. Most clustering algorithms determine the clusters by approximate minimization of the graph cut, i.e.,

$$\min_{\mathcal{V}_1,\dots,\mathcal{V}_K} \sum_{k=1}^{K} \left(\sum_{i \in \mathcal{V}_k} \sum_{j \in \mathcal{V} \setminus \mathcal{V}_k} W_{ij} \right).$$

Side constraints on the size of the clusters are imposed to avoid trivial solutions. A frequently used relaxation is to replace the graph cut with the Laplacian quadratic form $\frac{1}{2} \sum_i \sum_j (x_i - x_j)^2 W_{ij}$ [1, 3, 12–14]. However, a tighter relaxation is obtained by using the TV $\sum_i \sum_j |x_i - x_j| W_{ij}$ [8, 9]. In fact, for semi-supervised clustering (that is, given the labels of some sampled nodes) one can even show the equivalence between TV based clustering and minimum cuts for the case of two clusters [1, 11, 15].

Signed clustering. We model dissimilarity of two nodes i and j by a negative edge weight $W_{ij} < 0$, with the magnitude $|W_{ij}|$ describing the amount of dissimilarity. We next review the basic idea of clustering on signed graphs (i.e., with dissimilarity) for $K = 2$ clusters. The clusters can be conveniently described by a label vector $\mathbf{x} \in \{-1, 1\}^N$ with $x_i = 1$ for $i \in \mathcal{V}_1$ and $x_i = -1$ for $i \in \mathcal{V}_2$. In [6, 7], dissimilarity is incorporated by using the signed Laplacian

$\overline{\mathbf{L}} = \overline{\mathbf{D}} - \mathbf{W}$ with $\overline{\mathbf{D}} = \mathrm{diag}\{\bar{d}_1, \ldots, \bar{d}_N\}$, $\bar{d}_i = \sum_j |W_{ij}|$. 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 $S_{ij} = \mathrm{sign}(W_{ij})$ serves as a binary indicator for dis/similarity. This motivated us to introduce the following signed total variation [11]:

$$\|\mathbf{x}\|_{\mathrm{TV}} \triangleq \sum_i \sum_j |x_i - S_{ij} x_j| \, |W_{ij}| \, . \tag{1}$$

For edges (i, j) with negative weights $|x_i - S_{ij} x_j| \, |W_{ij}| = |x_i + x_j| \, |W_{ij}|$ is small whenever $x_i \approx -x_j$.

3. SEMI-SUPERVISED MULTICLASS TV CLUSTERING

3.1. Basic Method

We next generalize the TV based bi-partition clustering scheme from [11] to $K \geq 2$ clusters. We assume that the number of clusters K is known. The sets $\mathcal{L}_k \subset \mathcal{V}_k$, $k = 1, \ldots, K$ denote groups of nodes that are known a priori to belong to cluster $\mathcal{V}_k$, and $\mathcal{L} = \bigcup_{k=1}^K \mathcal{L}_k$ is the set of all nodes with known cluster labels. We represent the cluster affiliation using the binary indicator matrix $\mathbf{X} \in \{-1, 1\}^{N \times K}$ with $X_{ik} = 1$ if node i belongs to cluster k and $X_{ik} = -1$ otherwise. Hence, we have $\sum_k X_{ik} = -K + 2$. The symmetry of the indicator values $\{-1, 1\}$ about zero facilitates the use of dissimilarity information. Let $\mathbf{x}_i$ denote the rows of $\mathbf{X}$, i.e., $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$. We propose the following novel definition for the signed TV of a multi-cluster indicator matrix:

$$\|\mathbf{X}\|_{\mathrm{TV}} = \sum_{(i,j) \in \mathcal{E}_{\mathrm{sim}}} \|\mathbf{x}_i - \mathbf{x}_j\|_1 |W_{ij}| + \sum_{(i,j) \in \mathcal{E}_{\mathrm{dis}}} \|\mathbf{x}_i + \mathbf{x}_j\|_+ |W_{ij}|$$

with

$$\|\mathbf{x}\|_+ = \sum_k (x_k)_+ = \sum_k \max\{0, x_k\} \, ,$$

and

$$\mathcal{E}_{\mathrm{sim}} = \{(i, j) : W_{ij} > 0\}, \qquad \mathcal{E}_{\mathrm{dis}} = \{(i, j) : W_{ij} < 0\},$$

for the set of the similarity and dissimilarity edges. Observe that as opposed to the 1-norm, with the $+$-seminorm we have $\|\mathbf{x}_i + \mathbf{x}_j\|_+ = 0$ when dissimilar nodes are in the different clusters (for $X_{ik} \in \{-1, 1\}$). The metric $\|\mathbf{X}\|_{\mathrm{TV}}$ is consistent with the unsigned TV for multiple clusters used in [8].

The task of semi-supervised clustering using similarity and dissimilarity edges and signed TV amounts to an expensive combinatorial optimization problem. We thus propose the following relaxation that replaces the condition $X_{ik} \in \{-1, 1\}$ by $X_{ik} \in [-1, 1]$:

$$\min_{\mathbf{X}} \ \|\mathbf{X}\|_{\mathrm{TV}} \qquad \text{s.t.} \quad \mathbf{X} \in \mathcal{Q}.$$

The constraint set reads

$$\begin{aligned} \mathcal{Q} = \Big\{ & \mathbf{X} \in [-1, 1]^{N \times K} : \\ & X_{ik} = 1 \text{ for } i \in \mathcal{L}_k, \\ & X_{ik} = -1 \text{ for } i \in \mathcal{L} \backslash \mathcal{L}_k, \\ & \sum_k X_{ik} = -K + 2 \ \text{ for } i = 1, \ldots, N \Big\}. \end{aligned} \tag{2}$$

After finding a minimizer, node i is attributed to the cluster for which X_{ik} is maximal. For unsigned similarity graphs, a similar approach (with additional terms favoring clusters of similar size) achieves good clustering results [8].

3.2. Regularization

There are two other important issues which need to be specifically addressed. First, the label sets may be separated out as clusters when only relatively few labels are known. Second, the TV tends to assign zero values since both $(X_{ik} + X_{jk})_+$ and $|X_{ik} - X_{jk}|$ can be minimized by setting $X_{ik} = X_{jk} = 0$, thereby being penalized neither by similarity nor by dissimilarity edges. Both problems only occur when the number of known labels is rather small. We resolve them by introducing an ℓ_1 norm regularization. Let us define the unlabeled similar neighbors of a node $i \in \mathcal{V}$ by

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

Furthermore, for any $\mathcal{A} \subset \mathcal{V}$ we define $\mathcal{N}(\mathcal{A}) = \bigcup_{i \in \mathcal{A}} \mathcal{N}(i)$. The similarity neighborhood of $\mathcal{L}_k$ is denoted by

$$\mathcal{N}_k = \left\{ n \in \mathcal{N}(\mathcal{L}) : \sum_{k \in \mathcal{L}_k} W_{kn} > \sum_{l \in \mathcal{L}_l} W_{ln} \text{ for all } l \neq k \right\}. \tag{3}$$

We now exploit the fact that typically the majority of the nodes in $\mathcal{N}_k$ will also belong to cluster k (i.e., $X_{jk} = 1$ for $j \in \mathcal{N}_k$ and $X_{jk} = -1$ for $j \in \mathcal{N}_l$, $l \neq k$). Since incorporating the cardinalities of these sets is difficult for optimization, we use a convex ℓ_1 relaxation, leading to the following optimization problem for multiclass total variation clustering:

$$\min_{\mathbf{X} \in \mathcal{Q}} \|\mathbf{X}\|_{\mathrm{TV}} + R(\mathbf{X}), \tag{4}$$

$$R(\mathbf{X}) = \sum_{k=1}^K \Big(\lambda_k \sum_{j \in \mathcal{N}_k} |1 - X_{jk}| + \sum_{l \neq k} \lambda_l \sum_{j \in \mathcal{N}_l} |1 + X_{jk}| \Big).$$

Here, $\mathcal{Q}$ is as in (2) and $\lambda_1, \ldots, \lambda_K \geq 0$ are regularization parameters which can be tuned automatically, see Section 4. Similar regularization terms (termed "region force") were recently used in [9, 16].

4. ADMM FOR MULTICLASS TV CLUSTERING

Next, we show how to solve (4) in an efficient manner via the augmented ADMM [17]. While the closely related (preconditioned) primal-dual method (e.g., [18, 19]) appears to be an

Algorithm 1 Multiclass signed TV clustering

Input: $\mathbf{W}, \mathcal{L}_1, \ldots, \mathcal{L}_K, \lambda_1, \ldots, \lambda_K$

Initialization

1: determine $\widetilde{\mathbf{W}}$ via (5).

2: $m = 0, \quad \mathbf{Z}_k^{(0)} = \mathbf{Z}_k^{(-1)} = \mathbf{0}, \quad \rho = 0.1$

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

4: determine $\mathcal{N}_k$ via (3)

5: $X_{nk}^{(0)} = \begin{cases} 1, & n \in \mathcal{L}_k \cup \mathcal{N}_k \\ -1, & n \in \bigcup\limits_{l\neq k}(\mathcal{L}_l \cup \mathcal{N}_l) \\ 0, & \text{else} \end{cases}$

Iterations

6: **repeat**

7: $\quad Y_{nk}^{(m)} = \mathrm{div}_{\widetilde{\mathbf{W}}}\left(2\mathbf{Z}_k^{(m)} - \mathbf{Z}_k^{(m-1)}\right)_n$

8: $\quad \tilde{X}_{nk}^{(m)} = \begin{cases} X_{nk}^{(m)}, & n \in \mathcal{L} \\ X_{nk}^{(m)} + \frac{1}{\rho v_n} Y_{nk}^{(m)}, & \text{else} \end{cases}$

9: $\quad \mathbf{y}_n = \pi_{\mathcal{M}}(\tilde{\mathbf{X}}^T \mathbf{e}_n)$

10: $\quad \mathbf{x}_k^{(m+1)} = (y_{1k}, \ldots, y_{Nk})^T$

11: $\quad \tilde{\mathbf{Z}}_k^{(m+1)} = \mathbf{Z}_k^{(m)} + \rho\, \nabla_{\widetilde{\mathbf{W}}}\, \mathbf{x}_k^{(m+1)}$

12: $\quad Z_{ij}^{k,(m+1)} =$
$\begin{cases} \max\left\{-1, \min\left\{1, \tilde{Z}_{ij}^{k,(m+1)}\right\}\right\}, & W_{ij} > 0, \\ \max\left\{0, \min\left\{1, \tilde{Z}_{ij}^{k,(m+1)}\right\}\right\}, & W_{ij} < 0 \end{cases}$

13: $\quad m = m + 1$

14: **until** stopping criterion is satisfied

Output: $\hat{\mathbf{X}} = \mathbf{X}^{(m)}$

appealing alternative, we prefer ADMM due to its well-tested stopping criterion and varying penalty strategy [20].

As a first step, that also helps to understand the regularization term $R(\mathbf{X})$, we incorporate $R(\mathbf{X})$ into the TV total variation objective. This is accomplished by modifying the weight matrix $\mathbf{W}$. For each cluster $k = 1, \ldots, K$ and each node node $j \in \mathcal{N}_k$ we pick an anchor node $l_{kj} \in \mathcal{L}_k$ with $W_{l_{kj}j} \geq 0$ and define a new weight matrix $\widetilde{\mathbf{W}} \in \mathbb{R}^{N\times N}$ via

$$\widetilde{W}_{ij} = \begin{cases} W_{ij} + \lambda_k, & j \in \mathcal{N}_k \text{ and } i = l_{kj}, \\ W_{ij}, & \text{else.} \end{cases} \tag{5}$$

Note that since $W_{l_{kj}j} \geq 0$, this modification leaves the signs of the edge weights unchanged. Since $X_{nk} = 1$ for $n \in \mathcal{L}_k$ and $X_{nk} = -1$ for $n \in \mathcal{L}\backslash\mathcal{L}_k$ we can rewrite (4) as

$$\min_{\mathbf{X}\in\mathcal{Q}} \sum_{(i,j)\in\mathcal{E}_{\mathrm{sim}}} \|\mathbf{x}_i - \mathbf{x}_j\|_1 |\widetilde{W}_{ij}| + \sum_{(i,j)\in\mathcal{E}_{\mathrm{dis}}} \|\mathbf{x}_i + \mathbf{x}_j\|_+ |\widetilde{W}_{ij}|, \tag{6}$$

which is the signed total variation of $\mathbf{X}$ with respect to the modified weight matrix $\widetilde{W}$. This shows that the regularization terms have the effect of increasing the edge weights from the sampled nodes to their similarity neighborhood, thereby even more forcing these nodes to end up in the same cluster.

To derive the update steps of the augmented ADMM applied to (4) (respectively (6)) we require the signed gradient operator $\nabla_{\mathbf{W}} : \mathbb{R}^N \to \mathbb{R}^{N\times N}$,

$$(\nabla_{\mathbf{W}}\mathbf{x})_{ij} = (x_i - S_{ij}x_j)|W_{ij}|. \tag{7}$$

Furthermore, we use the signed divergence operator $\mathrm{div}_{\mathbf{W}} : \mathbb{R}^{N\times N} \to \mathbb{R}^N$ that can be derived (cf. [21]) as the negative adjoint of the signed gradient operator ($\mathrm{div}_{\mathbf{W}} = -\nabla_{\mathbf{W}}^*$):

$$(\mathrm{div}_{\mathbf{W}}\,\mathbf{Z})_i \triangleq \sum_{j\in\mathcal{V}} |W_{ji}| Z_{ji} S_{ji} - |W_{ij}| Z_{ij}. \tag{8}$$

The update steps of the augmented ADMM applied to problem (4) are summarized in Algorithm 1. It should be understood implicitly that statements made for a generic cluster index k are to be performed for all indices $k = 1, \ldots, K$. Step 9 of Algorithm 1 projects the rows of $\tilde{\mathbf{X}}$ onto the set

$$\mathcal{M} = \left\{\mathbf{x} : x_k \geq -1,\ \sum_{k=1}^{K} x_k = 2 - K\right\}. \tag{9}$$

Observe that $\pi_{\mathcal{M}}(\tilde{\mathbf{x}}) = \pi_{\mathcal{M}+1}(\tilde{\mathbf{x}} + \mathbf{1}) - \mathbf{1}$ where $\mathcal{M} + 1 = \{\mathbf{x} : 0 \leq x_k \text{ and } \sum_k x_k = 2\}$. The projection onto $\mathcal{M} + 1$ can be solved efficiently in $\mathcal{O}(K \log K)$ operations [22, 23]. For Algorithm 1 we use the stopping criterion and the varying penalty strategy from [20] which are both based on the primal and dual residuals (see also [11, 17]).

Complexity. The signed graph gradient (7) (step 11 of Algorithm 1) and the signed divergence (8) (step 7) can be calculated using the local neighborhood of a node (where $\widetilde{W}_{ij} \neq 0$). Thus their overall computation requires a number of operations that scales linearly with the number of edges of the graph. Furthermore, the matrices $\mathbf{Z}_k^{(m)}$ and $\tilde{\mathbf{Z}}_k^{(m)}$ are sparse with non-zero elements at the row and column indices (i, j) with $\widetilde{W}_{ij} \neq 0$. Since the steps 7, 11 and 12 have to be performed for each cluster index $k = 1, \ldots, K$, their calculation requires $\mathcal{O}(K|\mathcal{E}|)$ operations. The calculation of $\tilde{\mathbf{X}}$ in step 8 only requires $\mathcal{O}(KN)$ operations and is therefore governed by the cost of steps 7, 11 and 12. The projection of one vector onto $\mathcal{M}$ (respectively $\mathcal{M} + 1$) requires $\mathcal{O}(K \log K)$ operations [22]. Therefore the overall computational cost of one iteration of Algorithm 1 is $\mathcal{O}(K|\mathcal{E}|) + \mathcal{O}(NK \log K)$. We underline that because all steps of Algorithm 1 can be calculated using only the local neighborhood of a node, the algorithm can be implemented in a distributed manner (cf. [19]). Therefore, our signed TV clustering method scales well and is perfectly suited for handling large-dimensional datasets. The overall per-node complexity of a fully distributed implementation is $\mathcal{O}(Kd_{\max}) + \mathcal{O}(K \log K)$ (with $d_{\max}$ being the maximum number of neighbors across the N nodes).

Algorithm 2 Signed TV clustering with parameter tuning

Input: $\mathbf{W}, \mathcal{L}_1, \ldots, \mathcal{L}_K, x_{\min}$

Initialization: $\lambda(m) = \begin{cases} 0, & \text{for } m = 1, \\ 2^m, & \text{for } m > 1 \end{cases}$

$m_1 = m_2 = \cdots = m_K = 1$

repeat
  $\lambda_k = \lambda(m_k)$
  $\hat{\mathbf{X}} \leftarrow$ **Algorithm 1**$(\mathbf{W}, \mathcal{L}_1, \ldots, \mathcal{L}_K, \lambda_1, \ldots, \lambda_K)$
  $\mathcal{M}^k = \{n \in \mathcal{N}_k : \hat{X}_{nk} > 0\}$
  $\hat{x}^k = \min_{n \in \mathcal{M}^k} \hat{X}_{nk}$
  if $\mathcal{M}^k = \emptyset$ **or** $\hat{x}^k < x_{\min}$ **then**
    $m_k \leftarrow m_k + 1, \ a = 1$
  end if
until $a = 0$

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

Relaxation parameter adaptation. It remains to choose the regularization parameters $\lambda_1, \ldots, \lambda_K$ appropriately. Recall that the regularization was introduced in order to assign $X_{nk} = 1$ to the majority of nodes $n \in \mathcal{N}_k$ and $X_{nk} = -1$ to the majority of nodes $n \in \bigcup_{l \neq k} \mathcal{N}_l$. In addition, the cluster labels within the similarity neighborhoods should be close to 1 in magnitude. These observations led us to tuning the relaxation parameters according to Algorithm 2.

5. NUMERICAL EXPERIMENTS

We consider a dataset consisting of K noisy spirals: we created $N = 500$ random vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ according to (cf. [24])

$$\mathbf{u}_n = \begin{pmatrix} r_n(\cos(\varphi_n + (f_n - 1)\frac{2\pi}{K})) \\ r_n(\sin(\varphi_n + (f_n - 1)\frac{2\pi}{K})) \end{pmatrix} + \boldsymbol{\varepsilon}_n$$

where $f_n \in \{1, \ldots, K\}$ are randomly drawn cluster labels, $\varphi_n \sim \mathcal{U}(0, 4\pi/K)$ is a random angle, $r_n = (\frac{K^2 \varphi_n}{2\pi} + 2)/2$ is the radius and $\boldsymbol{\varepsilon}_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is a Gaussian jitter with $\sigma^2 = 0.045$. A similarity graph was created using the k-nearest-neighbor method [25] with $k = 10$ and edge weights $W_{ij} = \exp\left(-\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 / \kappa^2\right)$ with $\kappa^2 = 0.72$. For each pair of clusters we then added P dissimilarity edges with weight $W_{ij} = -10$ between randomly chosen pairs of nodes, which results in a total number of $L = PK(K-1)/2$ dissimilarity edges. We picked a set $\bigcup_{k=1}^K \mathcal{L}_k$ of $M = |\mathcal{L}|$ known cluster labels (chosen uniformly at random while ensuring at least one known label from each cluster). We then clustered the graph using our scheme (Algorithm 2) and the unsigned multiclass TV algorithm (UMTV) from [8]. The parameter $x_{\min}$ of Algorithm 2 was set to $x_{\min} = 0.9$.

The clustering performance is quantified by the percentage of mislabeled nodes among the set of nodes without prior known label. Table 1 depicts the error rates (mean and standard deviation) obtained over 500 Monte-Carlo runs and different values of P, M and K. For the case where the number of known labels is small, Algorithm 2 clearly outperforms UMTV. With increasing number of samples the performance difference diminishes and both algorithms deliver accurate clustering results. We also observe that the inclusion of only a single dissimilarity edge between each pair of clusters already significantly improves the clustering accuracy. The more dissimilarity edges there are, the more pronounced is the performance advantage of our method. We emphasize that UMTV optimizes for clusters of similar size; thus, we expect an even more pronounced gain of Algorithm 2 versus UMTV when cluster sizes vary. For a comparison between the dissimilarity-based method in [6] and Algorithm 2 we refer the interested reader to [11].

6. CONCLUSION

In this paper we incorporated dissimilarity into multiclass TV clustering. We introduced ℓ_1 regularization terms for the case of few cluster labels. We showed that these terms have the effect of putting more emphasis on the given node labels by increasing the weights of edges originating from the sampled nodes. We derived an ADMM based algorithm which can be implemented in a distributed environment. Our numerical experiments demonstrated that our approach outperforms unsigned TV clustering, and hence, including dissimilarity information into multiclass TV clustering significantly improves the accuracy of the results.

	UMTV; $K=3$	Algorithm 2; $K=3$			UMTV; $K=5$	Algorithm 2; $K=5$		
	$P=0$	$P=0$	$P=1$	$P=2$	$P=0$	$P=0$	$P=1$	$P=2$
$M=K$	27.0 ± 16.6	9.4 ± 12.4	5.2 ± 9.4	3.7 ± 9.1	4.0 ± 6.1	3.2 ± 5.9	1.8 ± 4.2	1.4 ± 3.8
$M=2K$	10.5 ± 13.3	5.0 ± 8.6	3.2 ± 6.9	1.8 ± 5.1	1.7 ± 3.9	2.0 ± 4.5	1.1 ± 3.3	0.7 ± 2.3
$M=5K$	1.1 ± 4.0	1.5 ± 4.5	0.9 ± 3.1	0.6 ± 2.0	0.3 ± 1.1	0.3 ± 1.3	0.3 ± 1.5	0.2 ± 1.0

Table 1: Error rates in percent (mean and standard deviation) achieved by our scheme and UMTV from [8] for $K = 3$ and $K = 5$ clusters and various numbers of dissimilarity edges P and known labels M.

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] K. Yin, Tai XC., and S. J. Osher. An effective region force for some variational models for learning and clustering. *Technical report, UCLA*, 2016.

[10] S. S. Rangapuram, P. K. Mudrakarta, and M. Hein. Tight continuous relaxation of the balanced k-cut problem. *CoRR*, abs/1505.06478, 2015.

[11] P. Berger, T. Dittrich, and G. Matz. Semi-supervised clustering based on signed total variation. In *Proc. Global Conf. on Signal and Information Processing*, Anaheim, California, USA, Nov. 2018.

[12] 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.

[13] 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.

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

[15] E. Merkurjev, E. Bae, A. L. Bertozzi, and X.-C. Tai. Global binary optimization on graphs for classification of high-dimensional data. *Journal of Mathematical Imaging and Vision*, 52(3):414–435, July 2015.

[16] K. Yin and XC. Tai. An effective region force for some variational models for learning and clustering. *Journal of Scientific Computing*, 74(1):175–196, Jan. 2018.

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

[18] 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.

[19] 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.

[20] 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.

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

[22] W. Wang and M. Á. Carreira-Perpiñán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *CoRR*, abs/1309.1541, 2013.

[23] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50(1):195–200, July 1986.

[24] H. Chang and D.-Y. Yeung. Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191 – 203, 2008.

[25] 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.