TV-SVM: Total Variation Support Vector Machine for Semi-Supervised Data Classification

Xavier Bresson∗ Ruiliang Zhang∗

Abstract

We introduce semi-supervised data classification algorithms based on total variation (TV), Reproducing Kernel Hilbert Space (RKHS), support vector machine (SVM), Cheeger cut, labeled and unlabeled data points. We design binary and multi-class semi-supervised classification algorithms. We compare the TV-based classification algorithms with the related Laplacian-based algorithms, and show that TV classification perform significantly better when the number of labeled data is small.

1 Introduction

1.1 Notation

Let \( \{x_i, y_i\}_{1 \leq i \leq N} \) denote \( N \) data points, where \( x_i \in \mathbb{R}^d \) is its attributes with dimension \( d \), while \( y_i \in \{+1, -1\} \) (binary classification) or \( y_i \in \{1, \ldots, c\} \) (multi-class classification). The total number of data points is \( N \) including \( n \) labeled data and \( N - n \) unlabeled data. \( K: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \text{Sym}(\mathbb{R}) \) is an operator-valued, positive definite kernel. Finally, we use the abbreviation \( f_i = f(x_i) \).

2 Binary (two-class) data classification

2.1 Regularized Least Square (RLS)

The standard RLS problem for binary classification is as follows [8]. Find a function \( f: \mathbb{R}^d \rightarrow \mathbb{R} \) such that

\[
\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in n} (y_i - f_i)^2 + \frac{\lambda}{2} \| f \|^2_{H_K},
\]

(1)

where \( \eta, \lambda > 0 \). Representer theorem states the existence of a minimizing function \( f^*(x) = \sum_{j \in n} K(x, x_j)\alpha_j^* \) (or \( f(x) = Kx\alpha \) with matrix representation) and the norm of \( f \) in the RKHS is \( \| f \|^2_{H_K} = \alpha^T K \alpha \). Problem (1) is equivalent to

\[
\min_{\alpha \in \mathbb{R}^n} \frac{\eta}{2} \| y - K\alpha \|^2_2 + \frac{\lambda}{2} \alpha^T K \alpha
\]

(2)

Taking the derivative w.r.t. \( \alpha \) provides the minimizer:

\[
\alpha^* = (\eta K + \lambda I_n)^{-1}(\eta y)
\]

(3)

Finally, unseen data points are classified as follows:

\[
x \in C_1 \text{ if } f^*(x) \geq 0
\]

(4)

\[
x \in C_2 \text{ if } f^*(x) < 0
\]

(5)

*Department of Computer Science, City University of Hong Kong, Hong Kong.
The TV-based RLS problem for binary semi-supervised classification is as follows [6]:

\[
\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in \mathcal{N}} (y_i - f_i)^2 + \frac{\lambda}{2} \|f\|^2_{H_K} + \frac{\gamma}{2} \sum_{i,j \in \mathcal{N}} w_{i,j} |f_i - f_j|, \quad (6)
\]

where \(\|Df\|^2 = \sum_{i,j \in \mathcal{N}} w_{i,j} |f_i - f_j|^2 = f^T L f\) is the Dirichlet energy and \(L = D - W\) is the graph Laplacian. Observe that training data points are composed of \(n\) labeled points and \(N - n\) unlabeled points. Let us consider matrix \(J = diag(1, \ldots, 1, 0, \ldots, 0)\) with the first \(n\) diagonal entries as 1 and the rest 0 and \(y = [y_1, \ldots, y_n, 0, \ldots, 0]\) with \(N - n\) entries as 0. This allows to write \(\sum_{i \in \mathcal{N}} (y_i - f_i)^2 = \|y - Jf\|^2\).

\[\min_{\alpha \in \mathbb{R}^N} \frac{\eta}{2} \|y - J \alpha\|^2 + \frac{\lambda}{2} \alpha^T K \alpha + \frac{\gamma}{2} (K \alpha)^T L (K \alpha). \quad (7)\]

Taking the derivative w.r.t. \(\alpha\) provides the minimizer:

\[\alpha^* = (\eta J K + \lambda I_N + \gamma L K)^{-1}(\eta y) \quad (8)\]

Finally, unseen data points are classified as follows:

\[\begin{align*}
x & \in C_1 \text{ if } f^*(x) \geq 0 \\
x & \in C_2 \text{ if } f^*(x) < 0
\end{align*}\]

\[\begin{align*}
x & \in C_1 \text{ if } f^*(x) \geq 0 \\
x & \in C_2 \text{ if } f^*(x) < 0
\end{align*}\]

The Laplacian-based RLS problem for binary semi-supervised classification is as follows [1]:

\[
\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in \mathcal{N}} (y_i - f_i)^2 + \frac{\lambda}{2} \|f\|^2_{H_K} + \frac{\gamma}{2} \sum_{i,j \in \mathcal{N}} w_{i,j} |f_i - f_j|, \quad (11)
\]

where \(\|Df\| = \sum_{i,j \in \mathcal{N}} w_{i,j} |f_i - f_j|\) is the graph TV of function \(f\). Unlike previous optimization problems, minimizing (11) needs advanced optimization techniques as TV term is non-differentiable. However, recent advances in \(\ell^1\) optimization provide efficient tools to deal with problem (11). In this work, we propose a splitting step coupled with an augmented Lagrangian method. Although one splitting variable is enough for minimizing (11), experimental observations suggest more accurate results using two splitting variables \(g, h\). The proposed iterative optimization algorithm is as follows:

\[
(f^{n+1}, h^{n+1}, g^{n+1}) = \min_{f \in H_K, h, g} \frac{\eta}{2} \|y - J h\|^2 + \frac{\lambda}{2} \|f\|^2_{H_K} + \gamma \|Dg\| + \frac{r_1}{2} \|f - g\|^2 + \frac{r_2}{2} \|h - g\|^2 \quad (12)
\]

\[\begin{align*}
\lambda_1^{n+1} &= \lambda_1^n + r_1 (f^{n+1} - g^{n+1}) \\
\lambda_2^{n+1} &= \lambda_2^n + r_2 (h^{n+1} - g^{n+1}) \quad (13)\]

The sub-minimization problem w.r.t. \(f\) is:

\[
\min_{f \in H_K} \frac{\lambda}{2} \|f\|^2_{H_K} + \frac{r_1}{2} \|f - (g - \frac{\lambda_1}{r_1})\|^2 \quad (15)
\]

which solution is given by \(f^{n+1} = K \alpha^{n+1}\), with

\[\alpha^{n+1} = (\lambda I_N + r_1 K)^{-1} (r_1 g^n - \lambda_1^n) \quad (16)\]
We summarize the iterative algorithm:

The Cheeger-based RLS problem for binary semi-supervised classification is as follows:

\[ \min_{\eta} \frac{\eta}{2} \| y - Jh \|_2^2 + \frac{r_1}{2} \| h - (g - \frac{\lambda_2}{r_2}) \|_2^2 \] (17)

which solution is given by

\[ h^{n+1} = (\eta I + r_2 I_N)^{-1}(\eta y + r_2 g^n - \lambda_2^o) \] (18)

The sub-minimization problem w.r.t. \( h \) is:

\[ \min_{g} \gamma \| Dg \| + \frac{r_1}{2} \| g - (f + \frac{\lambda_1}{r_1}) \|_2^2 + \frac{r_2}{2} \| g - (h + \frac{\lambda_2}{r_2}) \|_2^2 \] (19)

which can be written as

\[ \min_{g} \gamma \| Dg \| + \frac{r_1 + r_2}{2} \| g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2} \|_2^2 \] (20)

with \( z_1 = f + \frac{\lambda_1}{r_1} \) and \( z_2 = h + \frac{\lambda_2}{r_2} \). Different techniques can be applied to solve the TV ROF problem [9]. We use the primal-dual method [3] which is guaranteed to converge in \( O(\frac{1}{k}) \), \( k \) being the iteration number. Finally, we project each function \( f, h, g \) on the unit ball (i.e. \( f^{n+1} \leftarrow N, \| f^{n+1} \|_2 \) and constraint them to be zero-mean (i.e. \( f^{n+1} \leftarrow f^{n+1} - \text{mean}(f^{n+1}) \)).

We summarize the iterative algorithm:

\[ \alpha^{n+1} = (\lambda I_N + r_1 K)^{-1}(r_1 g^n - \lambda_1^o) \] (21)

\[ f^{n+1} = K\alpha^{n+1} \] (22)

\[ h^{n+1} = (\eta I + r_2 I_N)^{-1}(\eta y + r_2 g^n - \lambda_2^o) \] (23)

\[ \bar{g}^{n+1} = \arg\min_{g} \gamma \| Dg \| + \frac{r_1 + r_2}{2} \| g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2} \|_2^2 \] (24)

with \( z_1 = f + \frac{\lambda_1}{r_1} \), \( z_2 = h + \frac{\lambda_2}{r_2} \)

\[ \tilde{g}^{n+1} = N, \frac{\bar{g}^{n+1}}{\| \bar{g}^{n+1} \|_2} \] (25)

\[ g^{n+1} = \tilde{g}^{n+1} - \text{mean}(\tilde{g}^{n+1}) \] (26)

2.4 Cheeger-based RLS

The Cheeger-based RLS problem for binary semi-supervised classification is as follows:

\[ \min_{f \in H_K} \sum_{i,j \in N} w_{i,j} | f_i - f_j | \quad \text{s.t.} \quad f_i = y_i, \forall i \in n \] (28)

Based on [2], the following algorithm is proposed:

\[ g^{n+1} = f^n + c \cdot \text{sign}(f^n) \] (29)

\[ e^{n+1} = \text{RLS}(g^{n+1}) \] (30)

\[ h^{n+1} = \arg\min_{h} TV(h) + \frac{E^n}{2c} \| h - e^{n+1} \|_2^2 \] (31)

\[ t^{n+1} = h^{n+1} - \text{median}(h^{n+1}) \] (32)

\[ s^{n+1} = \begin{cases} y_i & \forall i \in n \\ f^{n+1}(i) & \forall i \notin n \end{cases} \] (33)

\[ f^{n+1} = N, \frac{s^{n+1}}{\| s^{n+1} \|_2} \] (34)

where RLS(\( g \)) is as follow

\[ \min_{e \in \mathcal{H}_K} \frac{\lambda}{2} \| e \|_{\mathcal{H}_K}^2 + \frac{r}{2} \| e - g \|_2^2, \] (35)
which solution is given by $e^{n+1} = K \alpha^*$, with

$$\alpha^* = (\lambda I + rK)^{-1} rg. \tag{36}$$

### 2.5 Support Vector Machine (SVM)

The standard SVM method for binary classification is as follows [5]. Find a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$\min_{f \in H_K, b \in \mathbb{R}} \frac{\lambda}{2} \|f\|_{H_K}^2, \quad \text{s.t. } y_i(f_i + b) \geq 1, i = 1, \ldots, n. \tag{37}$$

where $\lambda > 0$. To deal with non-separable case, the above problem can be rewritten with a slack variable $\xi$:

$$\min_{f \in H_K, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \sum_{i=1}^n \xi_i, \quad \text{s.t. } y_i(f_i + b) \geq 1 - \xi_i, i = 1, \ldots, n, \quad \xi_i \geq 0, i = 1, \ldots, n \tag{38}$$

Representer theorem states the existence of a minimizing function $f^*(x) = \sum_{j \in \mathbb{R}} K(x, x_j) \alpha_j^*$ and $\|f\|_{H_K}^2 = \alpha^T K \alpha$. Problem (38) is equivalent to

$$\min_{\alpha, \xi, b \in \mathbb{R}} \frac{\lambda}{2} \alpha^T K \alpha + \mu \sum_{i=1}^n \xi_i, \quad \text{s.t. } y_i\left(\sum_{j=1}^n K(x_i, x_j) \alpha_j + b\right) \geq 1 - \xi_i, i = 1, \ldots, n, \quad \xi_i \geq 0, i = 1, \ldots, n \tag{39}$$

By using the Lagrangian multiplier technique, problem (39) can be reformulated as:

$$\min_{\alpha, \xi, \beta, \beta_{\xi} \in \mathbb{R}^n, b \in \mathbb{R}} \frac{\lambda}{2} \alpha^T K \alpha + \mu \sum_{i=1}^n \xi_i, \quad \text{s.t. } y_i\left(\sum_{j=1}^n K(x_i, x_j) \alpha_j + b\right) \geq 1 - \xi_i, i = 1, \ldots, n, \quad \beta_i, \beta_{\xi} \geq 0, i = 1, \ldots, n \tag{40}$$

where $\beta, \beta_{\xi}$ are Lagrangian multipliers, $1$ is a vector whose elements are all ones, and $Y = \text{diag}(y_1, \ldots, y_n)$. Let us consider the Lagrangian optimality conditions. Taking the derivative w.r.t. $b$ and setting to 0 gives

$$\beta^T Y 1 = 0 \Rightarrow \beta^T y = 0. \tag{41}$$

Taking the derivative w.r.t. $\xi$ and setting to 0 gives

$$\mu 1 - \beta - \beta_{\xi} = 0 \Rightarrow 0 \leq \beta_i \leq \mu, i = 1, \ldots, n. \tag{42}$$

Taking the derivative w.r.t. $\alpha$ and setting to 0 gives

$$\alpha = \frac{Y \beta}{\lambda} \tag{43}$$

By substituting (43) back into (40), we reach the following dual optimization problem:

$$\max_{\beta \in \mathbb{R}^n} \beta^T 1 - \frac{1}{2} \beta^T Q \beta, \quad \text{s.t. } \beta^T y = 0, \quad 0 \leq \beta_i \leq \mu, i = 1, \ldots, n \tag{44}$$
where \( Q = Y(\frac{\lambda}{\gamma})Y \). The above problem can be solved using several efficient SVM solvers s.a. libSVM [4]. Once the optimal \( \beta^* \) is obtained, it is straightforward to get the optimal \( \alpha^* \):

\[
\alpha^* = \frac{Y \beta^*}{\lambda} \tag{45}
\]

and

\[
f^*(x) = \sum_{i=1}^{n} \alpha_i^* K(x, x_i). \tag{46}
\]

Finally, unseen data points are classified as follows:

\[
x \in C_1 \text{ if } f^*(x) \geq 0 \tag{47}
\]
\[
x \in C_2 \text{ if } f^*(x) < 0 \tag{48}
\]

### 2.6 Laplacian-based SVM

The Laplacian-based SVM problem with slack variable for binary semi-supervised classification is as follows [1]:

\[
\begin{align*}
\min_{f \in H_K, \xi \in \mathbb{R}^N, b \in \mathbb{R}} & \quad \frac{\lambda}{2} \|f\|^2_{H_K} + \mu \sum_{i \in N} \xi_i + \frac{\gamma}{2} \sum_{i, j \in N} w_{i,j} |f_i - f_j|^2, \\
\text{s.t.} & \quad y_i (f_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
& \quad \xi_i \geq 0, i = 1, \ldots, N
\end{align*} \tag{49}
\]

By using Lagrangian multipliers technique, problem (49) becomes:

\[
\begin{align*}
\min_{\alpha, \xi, \beta, \beta_\xi} & \quad \frac{\lambda}{2} \alpha^T K \alpha + \mu \xi^T \mathbf{1} + \frac{\gamma}{2} \alpha^T K \mathbf{L} K \alpha + \beta^T (\mathbf{1} - \xi - Y(K \alpha + b \mathbf{1})) - \beta_\xi^T \xi, \\
\text{s.t.} & \quad \beta_i, \beta_{\xi i} \geq 0, i = 1, \ldots, N
\end{align*} \tag{50}
\]

Applying the same steps as (41), (42) and (43), we get

\[
\max_{\beta \in \mathbb{R}^N} \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta, \tag{51}
\]

where

\[
Q = Y(\lambda I + \gamma L)^{-1}KY \tag{52}
\]

Optimal \( \alpha^* \) is obtained by solving the following linear system:

\[
\alpha^* = (\lambda I + \gamma L)^{-1} Y \beta^* \tag{53}
\]

and

\[
f^*(x) = \sum_{i=1}^{N} \alpha_i^* K(x, x_i). \tag{54}
\]

Finally, unseen data points are classified as follows:

\[
x \in C_1 \text{ if } f^*(x) \geq 0 \tag{55}
\]
\[
x \in C_2 \text{ if } f^*(x) < 0 \tag{56}
\]
2.7 Total Variation-based SVM

The TV-based SVM for binary semi-supervised classification is as follows:

\[
\min_{f \in H_K, \xi \in \mathbb{R}^N, \lambda \in \mathbb{R}} \frac{\lambda}{2} \|f\|^2_{H_K} + \mu \sum_{i \in N} \xi_i + \gamma \sum_{i,j \in N} w_{i,j} |f_i - f_j|, \\
\text{s.t. } y_i(f_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
\xi_i \geq 0, i = 1, \ldots, N
\]

where \( \|Df\| = \sum_{i,j \in N} w_{i,j} |f_i - f_j| \) is the graph TV of function \( f \). Like for TV-based RLS, we use a splitting step coupled with an augmented Lagrangian method. The proposed iterative optimization algorithm is as follows:

\[
(f^{n+1}, h^{n+1}, g^{n+1}) = \min_{f \in H_K, h, g, \xi, \beta} \frac{\lambda}{2} \|f\|^2_{H_K} + \mu \sum_{i=1}^{N} \xi_i + \gamma \|Dg\| + \\
\text{s.t. } y_i(h_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
\xi_i \geq 0, i = 1, \ldots, N
\]

Let \( \lambda_1^{n+1} = \lambda_1^n + r_1 (f^{n+1} - g^{n+1}) \)

\[
\lambda_2^{n+1} = \lambda_2^n + r_2 (h^{n+1} - g^{n+1})
\]

The sub-minimization problem w.r.t. \( f \) is:

\[
\min_{f \in H_K} \frac{\lambda}{2} \|f\|^2_{H_K} + \frac{r_1}{2} \|f - (g - \frac{\lambda_1^n}{r_1})\|^2
\]

which solution is given by \( f^{n+1} = K\alpha^{n+1} \), with

\[
\alpha^{n+1} = (\lambda N + r_1 K)^{-1} (r_1 g^n - \lambda_2^n)
\]

The sub-minimization problem w.r.t. \( h \) is:

\[
\min_{h, \xi, b} \mu \sum_{i=1}^{N} \xi_i + \frac{r_2}{2} \|h - c\|^2 \\
\text{s.t. } y_i(h_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
\xi_i \geq 0, i = 1, \ldots, N
\]

where \( e = g - \frac{\lambda_2^n}{r_2} \). Problem (65) is equivalent to

\[
\min_{h, \xi, b} \mu \xi^T \mathbf{1} + \frac{r_2}{2} \|h - e\|^2 + \beta^T (1 - \xi - Y(h + b)) - \beta^T \xi \\
\text{s.t. } \beta, \beta_\xi \geq 0, i = 1, \ldots, N
\]

Applying the same steps as (61), (62) and (63) did, we get

\[
\max_{\beta \in \mathbb{R}^N} \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \beta^T \mathbf{P}, \\
\text{s.t. } \beta^T \mathbf{y} = 0, \\
0 \leq \beta_i \leq \mu, i = 1, \ldots, N
\]

where \( Q = YY^T \) and \( P = Yc \). Problem (64) can be solved by gradient descent method, and the solution \( \beta^* \) can be used to obtain the optimal \( h^{n+1} \):

\[
h^{n+1} = \frac{1}{r_2} Y \beta^* + e
\]
The sub-minimization problem w.r.t. $g$ is:

$$
\min_g \gamma \|Dg\| + \frac{r_1}{2} \|g - (f + \frac{\lambda_1}{r_1})\|^2 + \frac{r_2}{2} \|g - (h + \frac{\lambda_2}{r_2})\|^2
$$

which can be written as

$$
\min_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|^2
$$

with $z_1 = f + \frac{\lambda_1}{r_1}$ and $z_2 = h + \frac{\lambda_2}{r_2}$.

We summarize the iterative algorithm:

$$
\begin{align*}
\alpha^{n+1} &= (\lambda N + r_1 K)^{-1}(r_1 g^n - \lambda^n_1) \\
f^{n+1} &= K \alpha^{n+1} \\
\beta^* &= \max_{\beta \in \mathbb{R}^N} \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \beta^T P, \text{ s.t. } \beta^T y = 0, 0 \leq \beta_i \leq C, i = 1, \ldots, N \\
h^{n+1} &= \frac{1}{r_2} Y \beta^* + e \\
\bar{g}^{n+1} &= \arg\min_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|^2 \\
&\text{with } z_1 = f + \frac{\lambda_1}{r_1}, z_2 = h + \frac{\lambda_2}{r_2} \\
\tilde{g}^{n+1} &= \frac{N \cdot \bar{g}^{n+1}}{\|\bar{g}^{n+1}\|_2} \\
g^{n+1} &= \tilde{g}^{n+1} - \text{mean}(\tilde{g}^{n+1})
\end{align*}
$$

### 2.8 Cheeger-based SVM

The Cheeger-based SVM problem for binary semi-supervised classification is as follows:

$$
\min_{f \in H_K} \frac{\sum_{i,j \in N} w_{i,j}|f_i - f_j|}{\sum_{i \in N}|f_i - \text{median}(f)|} \text{ s.t. } f_i = y_i, \forall i \in n \\
\text{s.t. } y_i(f_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
\xi_i \geq 0, i = 1, \ldots, N
$$

Based on [2], the following algorithm is proposed:

$$
\begin{align*}
g^{n+1} &= f^n + c . \text{sign}(f^n) \\
e^{n+1} &= \text{SVM}(g^{n+1}) \\
h^{n+1} &= \arg\min_h TV(h) + \frac{c}{2} \|h - e^{n+1}\|^2 \\
t^{n+1} &= h^{n+1} - \text{median}(h^{n+1}) \\
s^{n+1} &= \left\{ \begin{array}{ll} \ell(i) & \forall i \in n \\
\ell^{n+1}(i) & \forall i \notin n \end{array} \right. \\
f^{n+1} &= N \cdot \frac{s^{n+1}}{\|s^{n+1}\|_2}
\end{align*}
$$

where $\text{SVM}(g)$ is as follow:

$$
\begin{align*}
&\min_{e, \xi, b} \frac{\lambda}{2} \|e\|^2_{H_K} + \mu \sum_{i \in N} \xi_i + \frac{r}{2} \|e - g\|^2 \\
&\text{s.t. } y_i(e_i + b) \geq 1 - \xi_i, i = 1, \ldots, N \\
&\xi_i \geq 0, i = 1, \ldots, N
\end{align*}
$$
Problem (88) is equivalent to

$$\min_{e, \lambda, b, \beta, \beta_\xi} \frac{\lambda}{2}||e||_H^2 + \mu \xi^T 1 + \frac{r}{2}||e - g||_2^2 + \beta^T (1 - \xi - Y(e + b)) - \beta_\xi^T \xi$$

s.t. $\beta_\xi, \beta_\xi i \geq 0, i = 1, \ldots, N$ (89)

Applying the same steps as (41), (42) and (43), we get

$$\max_{\beta \in \mathbb{R}^N} \beta^T 1 - \frac{1}{2} \beta^T Q \beta - \frac{1}{2} P \beta,$$

s.t. $\beta^T y = 0$, $0 \leq \beta_i \leq \mu, i = 1, \ldots, N$ (90)

where $Q = YGY$, $P = r g^T (G + G^T) Y$ and $G = (\lambda I + r K)^{-1} K$. The above problem can be solved by gradient descent method, and the solution $\beta^*$ can be used to obtain the optimal $\alpha^*$:

$$\alpha^* = (\lambda I + r K)^{-1} (Y \beta^* + rg)$$

and

$$e_{n+1}^n = K \alpha^*$$

2.9 Experimental results

| # labels per class | 1     | 5     | 10    | 50    |
|--------------------|-------|-------|-------|-------|
| Lap-RLS            | 18.09 | 10.48 | 7.77  | 4.14  |
| Lap-SVM            | 13.79 | 9.84  | 7.61  | 4.77  |
| TV-RLS             | 3.18  | 3.16  | 3.13  | 3.16  |
| TV-SVM             | 3.18  | 3.13  | 3.13  | 3.08  |
| Cheeger-RLS        | 4.06  | 3.74  | 4.03  | 2.84  |
| Cheeger-SVM        | 3.87  | 3.74  | 4.00  | 2.73  |

Table 1: Binary semi-supervised classification algorithms tested on the sets of 4’s and 9’s from USPS dataset. Error is averaged over 10 runs with randomly selected labels.

3 Multi-class data classification

3.1 Laplacian-based RLS

The Laplacian-based RLS problem for multi-class semi-supervised classification is as follows:

$$\min_{f = (f_1, \ldots, f_c) \in H_K} \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in \mathcal{N}} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f_k\|^2_{H_K} + \frac{\gamma}{2} \sum_{k=1}^c \sum_{i,j \in \mathcal{N}} w_{i,j} |f_i^k - f_j^k|^2,$$

s.t. $\sum_{k=1}^c f_i^k = 1, f_i^k \geq 0, \forall i \in \mathcal{N}$ (93)

where the last constraint being the simplex constraint. Problem (93) is equivalent to

$$\min_{f = (f_1, \ldots, f_c) \in H_K} \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in \mathcal{N}} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f_k\|^2_{H_K} + \frac{\gamma}{2} \sum_{k=1}^c \|D f_k\|^2,$$

s.t. $f_k^k = g_k^k, \sum_{k=1}^c g_i^k = 1, g_i^k \geq 0, \forall i \in \mathcal{N}$ (94)
Finally, unseen data points are classified as follows:

The simplex projection is done by Michelot’s method \[7\].

The TV-based RLS problem for multi-class semi-supervised classification is as follows:

\[ 3.2 \text{ Total Variation-based RLS} \]

Finally, unseen data points are classified as follows:

\[ x \in C_k \text{ if } f^*_k(x) = \max_{j} \{f^*_j(x)\}_{1 \leq j \leq c} \]  

\[ (\alpha^k)^{n+1} = \arg\min_{\alpha \in R^N} \frac{\eta}{2} \|y^k - J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 L(K \alpha_k) + \frac{r}{2} \|K \alpha_k - (g^k - \frac{\lambda^k}{r})\|^2 \]  

\[ (f^k)^{n+1} = K(\alpha^k)^{n+1} \]  

\[ (g^k)^{n+1} = \Pi_{\sum_{g=1}^n (f^k + \frac{\lambda^k}{r})} \]  

This leads to the proposed iterative algorithm:

The simplex projection is done by Michelot’s method \[7\].

Finally, unseen data points are classified as follows:

\[ x \in C_k \text{ if } f^*_k(x) = \max_{j} \{f^*_j(x)\}_{1 \leq j \leq c} \]  

\[ (\alpha^k)^{n+1} = \arg\min_{\alpha \in R^N} \frac{\eta}{2} \|y^k - J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 L(K \alpha_k) + \frac{r}{2} \|K \alpha_k - (g^k - \frac{\lambda^k}{r})\|^2 \]  

\[ (f^k)^{n+1} = K(\alpha^k)^{n+1} \]  

\[ (g^k)^{n+1} = \Pi_{\sum_{g=1}^n (f^k + \frac{\lambda^k}{r})} \]  

This leads to the proposed iterative algorithm:

\[ (\alpha^k)^{n+1} = \arg\min_{\alpha \in R^N} \frac{\eta}{2} \|y^k - J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 L(K \alpha_k) + \frac{r}{2} \|K \alpha_k - (g^k - \frac{\lambda^k}{r})\|^2 \]  

\[ (f^k)^{n+1} = K(\alpha^k)^{n+1} \]  

\[ (g^k)^{n+1} = \Pi_{\sum_{g=1}^n (f^k + \frac{\lambda^k}{r})} \]  

\[ (g^k)^{n+1} = \frac{N}{\|g^k\|^{n+1}} \]  

Finally, unseen data points are classified as follows:

\[ x \in C_k \text{ if } f^*_k(x) = \max_{j} \{f^*_j(x)\}_{1 \leq j \leq c} \]  

\[ (\alpha^k)^{n+1} = \arg\min_{\alpha \in R^N} \frac{\eta}{2} \|y^k - J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 + \frac{\lambda}{2} \|J^k K \alpha_k\|^2 L(K \alpha_k) + \frac{r}{2} \|K \alpha_k - (g^k - \frac{\lambda^k}{r})\|^2 \]  

\[ (f^k)^{n+1} = K(\alpha^k)^{n+1} \]  

\[ (g^k)^{n+1} = \Pi_{\sum_{g=1}^n (f^k + \frac{\lambda^k}{r})} \]  

\[ (g^k)^{n+1} = \frac{N}{\|g^k\|^{n+1}} \]  

Finally, unseen data points are classified as follows:

\[ x \in C_k \text{ if } f^*_k(x) = \max_{j} \{f^*_j(x)\}_{1 \leq j \leq c} \]
3.3 Cheeger-based RLS

The Cheeger-based RLS problem for multi-class semi-supervised classification is as follows:

\[
\min_{f=(f^1, \ldots, f^c) \in H_K} \sum_{k=1}^{c} \sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k| \quad \text{s.t.} \quad f_i^k = l_i^k, \; \forall i \in n
\]  

\[ (108) \]

The following algorithm is proposed:

\[
\begin{align*}
(g_k^{n+1}) &= (f_k^n) + c \cdot \text{sign}((f_k^n)) \\
(e_k^{n+1}) &= \text{RLS}(g_k^{n+1}) \\
(h_k^{n+1}) &= \arg\min_h \; TV(h_k) + \frac{E_n}{2c} \|h_k - (e_k^{n+1})\|_2^2 \\
(f_k^{n+1}) &= (h_k^{n+1}) - \text{median}((h_k^{n+1})) \\
(s_k^{n+1}) &= \left\{ \begin{array}{ll} g_k^{n+1} & \forall i \in n \\ (f_k^{n+1})(i) & \forall i \notin n \end{array} \right. \\
(t_k^{n+1}) &= \Pi_{s_k=1}(s_k) \\
(f_k^{n+1}) &= N \frac{(s_k^{n+1})}{\|s_k^{n+1}\|_2} \\
\end{align*}
\]

\[ (109) \]

\[ (110) \]

\[ (111) \]

\[ (112) \]

\[ (113) \]

\[ (114) \]

\[ (115) \]

\[ (116) \]

where RLS(\(g\)) is exact the same as \(\hat{g}\).

Finally, unseen data points are classified as follows:

\[ x \in C_k \text{ if } f_i^k(x) = \max_j \{f_j^k(x)\}_{1 \leq j \leq c} \]  

\[ (117) \]

3.4 Laplacian-based SVM

The Laplacian-based SVM for multi-class semi-supervised classification is as follows:

\[
\min_{f=(f^1, \ldots, f^c) \in H_K, \beta \in \mathbb{R}^c, \xi \in \mathbb{R}^N \times c} \frac{\lambda}{2} \sum_{k=1}^{c} \|f_k^\prime\|_{H_K}^2 + \mu \sum_{k=1}^{c} \sum_{i \in N} \xi_i^k + \frac{\gamma}{2} \sum_{k=1}^{c} \sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k|^2 + \frac{\gamma}{2} \sum_{k=1}^{c} \|Df_k^\prime\|_2^2, \\
\text{s.t.} \quad y_i^k \left( f_i^k + b_k \right) \geq 1 - \xi_i^k, \xi_i^k \geq 0, \; i \in N, \; k \in c \\
\sum_{k=1}^{c} f_i^k = 1, \; f_k(i) \geq 0, \; \forall i \in N
\]

Problem \([HS]\) is equivalent to

\[
\min_{f=(f^1, \ldots, f^c) \in H_K, \beta \in \mathbb{R}^c, \xi \in \mathbb{R}^N \times c} \frac{\lambda}{2} \sum_{k=1}^{c} \|f_k^\prime\|_{H_K}^2 + \mu \sum_{k=1}^{c} \sum_{i \in N} \xi_i^k + \frac{\gamma}{2} \sum_{k=1}^{c} \|Df_k^\prime\|_2^2, \\
\text{s.t.} \quad y_i^k \left( f_i^k + b_k \right) \geq 1 - \xi_i^k, \xi_i^k \geq 0, \; i \in N, \; k \in c \\
f_k = \hat{g}_k, \; \sum_{k=1}^{c} g_i^k = 1, \; g_i^k \geq 0, \; \forall i \in N
\]

Notes that, each \(f_k^\prime\) can be solved independently by using the same procedure as below (superscript \(k\) is ignored for convenience):

\[
\min_{f \in H_K, \beta \in \mathbb{R}, \xi \in \mathbb{R}^N} \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \xi^T 1 + \frac{\gamma}{2} f^T Lf + \frac{\gamma}{2} \|f - c\|_2^2, \\
\text{s.t.} \quad y_i(f_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \; i \in N
\]

\[ (118) \]
where \( e = g - \frac{1}{l} \), and \( l \) is the Lagrangian multiplier. Problem (118) is equivalent to

\[
\begin{align*}
\min_{\beta, \gamma} & \quad \alpha^T K \alpha + \mu \xi^T 1 + \frac{\gamma}{2} \alpha^T KLK \alpha + \frac{r}{2} \|K \alpha - e\|^2 \\
& + \beta^T (1 - \xi - Y(K \alpha + b)) - \beta^T \xi \\
\text{s.t.} & \quad \beta, \beta \geq 0, i \in N
\end{align*}
\] (119)

Applying the same steps as (41), (42) and (43), we get

\[
\begin{align*}
\max_{\beta \in \mathbb{R}^N} & \quad \beta^T 1 - \frac{1}{2} \beta^T Q \beta - \frac{1}{2} P \beta, \\
\text{s.t.} & \quad \beta^T y = 0, \\
& \quad 0 \leq \beta_i \leq \mu, i = 1, \ldots, N
\end{align*}
\] (120)

where \( Q = YGY, P = re^T (G + G^T)Y \) and \( G = (\lambda I + \gammaLK + rK)^{-1}K \). The above problem can be solved by gradient descent method, and the solution \( \beta^* \) can be used to obtain the optimal \( \alpha^* \), which is:

\[
\alpha^* = (\lambda I + \gammaLK + rK)^{-1} (Y \beta^* + re)
\] (121)

and

\[
f = Ka^*
\] (122)

This leads to the following iterative algorithm:

\[
\begin{align*}
(f^k)^{n+1} &= \text{computed by using (120), (121) and (122)} \\
(g^k)^{n+1} &= \Pi_{\sum g^k = 1}(f^k)^{n+1} + \frac{b^k}{r})
\end{align*}
\] (123)

The simplex projection is done by Michelot’s method [7].

Finally, unseen data points are classified as follows:

\[
x \in C_k \text{ if } f_k^*(x) = \max_j \{f_j^*(x)\}_{1 \leq j \leq c}
\] (125)

### 3.5 Total Variation-based SVM

The TV-based SVM for multi-class semi-supervised classification is as follows:

\[
\begin{align*}
\min_{f \in (f^1, \ldots, f^c)} & \quad \lambda \sum_{k=1}^c \sum_{i=1}^n \|f_i^k - f_i^k (x)\|^2, \\
\text{s.t.} & \quad y_i^k (f_i^k + b^k) \geq 1 - \xi_i, \xi_i \geq 0, i \in N, k \in c
\end{align*}
\] (126)

Problem (126) is equivalent to

\[
\begin{align*}
\min_{f \in (f^1, \ldots, f^c)} & \quad \lambda \sum_{k=1}^c \sum_{i=1}^n \|f_i^k - f_i^k (x)\|^2, \\
\text{s.t.} & \quad y_i^k (f_i^k + b^k) \geq 1 - \xi_i, \xi_i \geq 0, i \in N, k \in c
\end{align*}
\] (127)

Notes that, each \( f^k \) can be solved independently:

\[
\begin{align*}
\min_{f \in H_K, b \in \mathbb{R}^N} & \quad \lambda \sum_{k=1}^c \|f\|^2, \\
\text{s.t.} & \quad y_i (f_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in N
\end{align*}
\] (128)
where $e = g - \frac{1}{r}$, and $l$ is the Lagrangian multiplier. Problem (128) is equivalent to

$$
\min_{b \in R, \alpha, \xi, \beta, \beta_{x} \in R^{N}} \frac{1}{2} \alpha^{T} K \alpha + \mu \xi^{T} 1 + \frac{r}{2} \|K \alpha - e\|^{2}_{2} + \beta^{T} (1 - \xi - \gamma (\alpha + b)) - \beta_{x}^{T} \xi
$$

(129)

s.t. $\beta, \beta_{x} \geq 0, i \in N$

Applying the same steps as (41), (42) and (43), we get

$$
\max_{\beta \in R^{N}} \beta^{T} 1 - \frac{1}{2} \beta^{T} Q \beta - \frac{1}{2} P \beta,
$$

(130)

s.t. $\beta^{T} y = 0$

$$
0 \leq \beta_{i} \leq \mu, i = 1, \ldots, N
$$

where $Q = Y G Y$, $P = r e^{T} (G + G^{T}) Y$ and $G = (\lambda I + r K)^{-1} K$. The above problem can be solved by gradient descent method, and the solution $\beta^{*}$ can be used to obtain the optimal $\alpha^{*}$, which is:

$$
\alpha^{*} = (\lambda I + r K)^{-1} (Y \beta^{*} + r e)
$$

(131)

and

$$
f = K \alpha^{*}
$$

(132)

This leads to the proposed iterative algorithm:

$$
(f^{k})^{n+1} = \text{computed by using (130), (131) and (132)}
$$

(133)

$$
(g^{k})^{n+1} = \arg \min_{g} \gamma \|D g^{k}\| + \frac{r}{2} \|g^{k} - (f^{k} + \frac{l}{r})\|^{2}_{2}
$$

(134)

$$
\hat{g}^{k+1} = \Pi_{-1} (g^{k})^{n+1}
$$

(135)

$$
(g^{k})^{n+1} = N \cdot \frac{(g_{k})^{n+1}}{\| (g^{k})^{n+1} \|_{2}}
$$

(136)

Finally, unseen data points are classified as follows:

$$
x \in C_{k} \text{ if } f^{k}_{x}(x) = \max_{j} \{f^k_j(x)\}_{1 \leq j \leq c}
$$

(137)

3.6 Cheeger-based SVM

The Cheeger-based SVM with slack variable problem for multi-class classification is as follows:

$$
\min_{f = (f^{1}, \ldots, f^{c}) \in R^{K}} \sum_{i=1}^{c} \sum_{j \in N} w_{ij} | f_{i}^{j} - f_{j}^{j} | \text{ s.t. } f_{i}^{j} = y_{i}^{j}, \forall i \in n
$$

(138)

The following algorithm is proposed:

$$
(g^{k})^{n+1} = (f^{k})^{n} + c \cdot \text{sign}((f^{k})^{n})
$$

(139)

$$
(e^{k})^{n+1} = \text{SVM}( (g^{k})^{n+1} )
$$

(140)

$$
(h^{k})^{n+1} = \arg \min_{h_{k}} TV (h^{k}) + \frac{E_{n}}{2c} \| h^{k} - (e^{k})^{n+1} \|^{2}_{2}
$$

(141)

$$
(l^{k})^{n+1} = (h^{k})^{n+1} - \text{median}( (h^{k})^{n+1} )
$$

(142)

$$
(s^{k})^{n+1} = \begin{cases} y_{i}^{k}(i) & \text{if } i \in n \\ (t^{k})^{n+1}(i) & \text{if } i \notin n \end{cases}
$$

(143)

$$
(\hat{g}^{k})^{n+1} = \Pi_{-1} (s^{k})^{n+1}
$$

(144)

$$
(f^{k})^{n+1} = N \cdot \frac{(\hat{g}^{k})^{n+1}}{\| (\hat{g}^{k})^{n+1} \|_{2}}
$$

(145)

where SVM() is as (83).

Finally, unseen data points are classified as follows:

$$
x \in C_{k} \text{ if } f^{k}_{x}(x) = \max_{j} \{f^k_j(x)\}_{1 \leq j \leq c}
$$

(146)
3.7 Experimental results

| # labels per class | 1   | 5   | 10  | 50  |
|--------------------|-----|-----|-----|-----|
| Lap-RLS            | 20.06 | 6.64 | 4.03 | 3.3 |
| Lap-SVM            | 49.95 | 14.21 | 6.27 | 2.82 |
| TV-RLS             | 2.0  | 2.06 | 1.91 | 1.98 |
| TV-SVM             | 1.75 | 1.82 | 1.77 | 1.85 |
| Cheeger-RLS        | 3.35 | 1.95 | 1.85 | 1.87 |
| Cheeger-SVM        | 2.94 | 2.08 | 1.72 | 1.74 |

Table 2: Multi-class semi-supervised classification algorithms tested on four classes (0’s, 1’s, 4’s and 9’s) from USPS dataset. Error is averaged over 10 runs with randomly selected labels.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] X. Bresson, T. Laurent, D. Uminsky, and J.H. von Brecht. Convergence and Energy Landscape for Cheeger Cut Clustering. *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[3] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[4] C.C. Chang and C.J. Lin. LIBSVM: a Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 27(3), 2011.

[5] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[6] S.H. Kang, B. Shafei, and G. Steidl. Supervised and Transductive Multi-Class Segmentation Using p-Laplacians and RKHS methods. 2012.

[7] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of n. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

[8] T. Poggio and S. Smale. The Mathematics of Learning: Dealing with Data. *Notices of the AMS*, 50(5):537–544, 2003.

[9] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D*, 60(1-4):259 – 268, 1992.