Nonparametric Maximum Margin Similarity for Semi-Supervised Learning

Yingzhen Yang, Xinqi Chu, Zhangyang Wang, Thomas S. Huang. Beckman Institute for Advanced Science and Technology, UIUC.

Introduction

1. Nonparametric Label Propagation (LP) has been proven to be effective for semi-supervised learning problems, and it predicts the labels for unlabeled data by a harmonic solution of an energy minimization problem which encourages local smoothness of the labels in accordance with the similarity graph.
2. On the other hand, the success of LP algorithms highly depends on the underlying similarity graph. Most similarity graphs for LP are constructed empirically and the objective function over the similarity graphs is defined as sum of the product of pairwise similarity and the squared label difference.
3. **We relate LP to a novel nonparametric maximum margin similarity framework with the concept of similarity margin, and present a new semi-supervised learning algorithm called Maximum Margin Similarity Graph (MMSG).** The conventional LP algorithm can be interpreted as a special case of our MMSG algorithm when the separation parameter is sufficiently large.
4. By the sample-based similarity margin rather than the expectation based margin, our framework leads to an tractable optimization problem which is solved by the projected subgradient method.

Formulation

- Definitions:
  The similarity function over $\mathbb{R}^d \times \mathbb{R}^d$ is defined as any bounded pairwise function $S\colon \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$. The labeled data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are drawn i.i.d. from some distribution P on $\mathbb{R}^d \times \{-1, 1\}$. We consider binary classification in this paper, and the multi-class case can be handled in the one-vs-all manner. The similarity margin of the datum $\mathbf{x} \in \mathbb{R}^d$ is defined as the difference of sum of $\mathbf{x}$'s similarity to the data with the same label as $\mathbf{x}$, and the sum of $\mathbf{x}$'s similarity to the data with different label:

$$\gamma_{\mathbf{x}} = \frac{1}{n}\Big(\sum_{j\colon y_j = y(\mathbf{x})} S(\mathbf{x}, \mathbf{x}_j) - \sum_{j\colon y_j \neq y(\mathbf{x})} S(\mathbf{x}, \mathbf{x}_j)\Big) \qquad (1)$$

- Theoretical Guarantee:
  Intuitively, the similarity margin for each datum should be large so as to separate different classes.

Formulation of Maximum Margin Sim

To facilitate optimization algorithms, the small similarity margin is penalized by hinge loss. For the separation parameter $\gamma > 0$, the hinge loss of the similarity margins for the data $\mathcal{D}$ is defined as

$$H_{\gamma,\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - \frac{\gamma_i}{\gamma}\} \qquad (2)$$

where $\gamma_i = \gamma_{\mathbf{x}_i}$ is the similarity margin of $\mathbf{x}_i$. Theorem 1 shows that with a high probability, there exists a linear classifier in the transformed space with hinge loss bounded by $H_{\gamma,\mathcal{D}}$.

Theorem 1 *Given the data* $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, *define the mapping* $F_{\mathcal{D}}\colon \mathbb{R}^d \to \mathbb{R}^n$ *as* $F_{\mathcal{D}}(\mathbf{x}) = \frac{1}{\sqrt{n}}\big(S(\mathbf{x}, \mathbf{x}_1), S(\mathbf{x}, \mathbf{x}_2), \ldots, S(\mathbf{x}, \mathbf{x}_n)\big)$. *For* $\delta_1, \delta_2, \delta_3 > 0$, *with probability at least* $1 - \delta_1 - \delta_2 - \delta_3$ *over the data* $\mathcal{D}$, *there exists a linear classifier in the transformed space induced by* $F_{\mathcal{D}}$ *such that this classifier has hinge loss at most* $H_0 = H_{\gamma,\mathcal{D}} + \frac{\sqrt{\frac{2}{n}\log\frac{n}{\delta_1}}}{\gamma} + \sqrt{\frac{2}{n\gamma^2}\log\frac{1}{\delta_2}} + \delta_3(1 + \frac{1}{\gamma})$ *with respect to the margin* γ. *Namely, there exists a vector* $\boldsymbol{\beta} \in \mathbb{R}^n$ *such that* $\mathbb{E}_{(\mathbf{x},y)\sim P}\Big[\max\{0, 1 - \frac{y\langle\boldsymbol{\beta}, F_{\mathcal{D}}(\mathbf{x})\rangle}{\gamma}\}\Big] \le H_0$.

- Maximum Margin Similarity Graph for Semi-Supervised Learning:
  We propose Maximum Margin Similarity Graph for semi-supervised learning, and MMSG minimizes the hinge loss of the similarity margins by projected subgradient method. The data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^d$ are comprised of labeled and unlabeled set, and the first l points have labels $y_i \in \{-1, 1\}$ for $1 \le i \le l$. In the following text, ℓ_i is the label of $\mathbf{x}_i$ for $i \in \{1, \ldots, n\}$, and $\ell_i = y_i$ for $i \in \{1, \ldots, l\}$.

- The optimization problem of MMSG is presented below, where the discreteness condition is relaxed so that $\boldsymbol{\ell}$ takes real values for unlabeled data:

$$\begin{aligned} \min_{\boldsymbol{\ell}} \quad & H_{\boldsymbol{\ell}} = \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - \frac{\gamma_i}{\gamma}\} \\ s.t. \quad & \ell_i \in [-1, 1],\ i \in \{1 + 1, \ldots, n\} \\ & \ell_i = y_i,\ i \in \{1, \ldots, l\} \end{aligned}$$

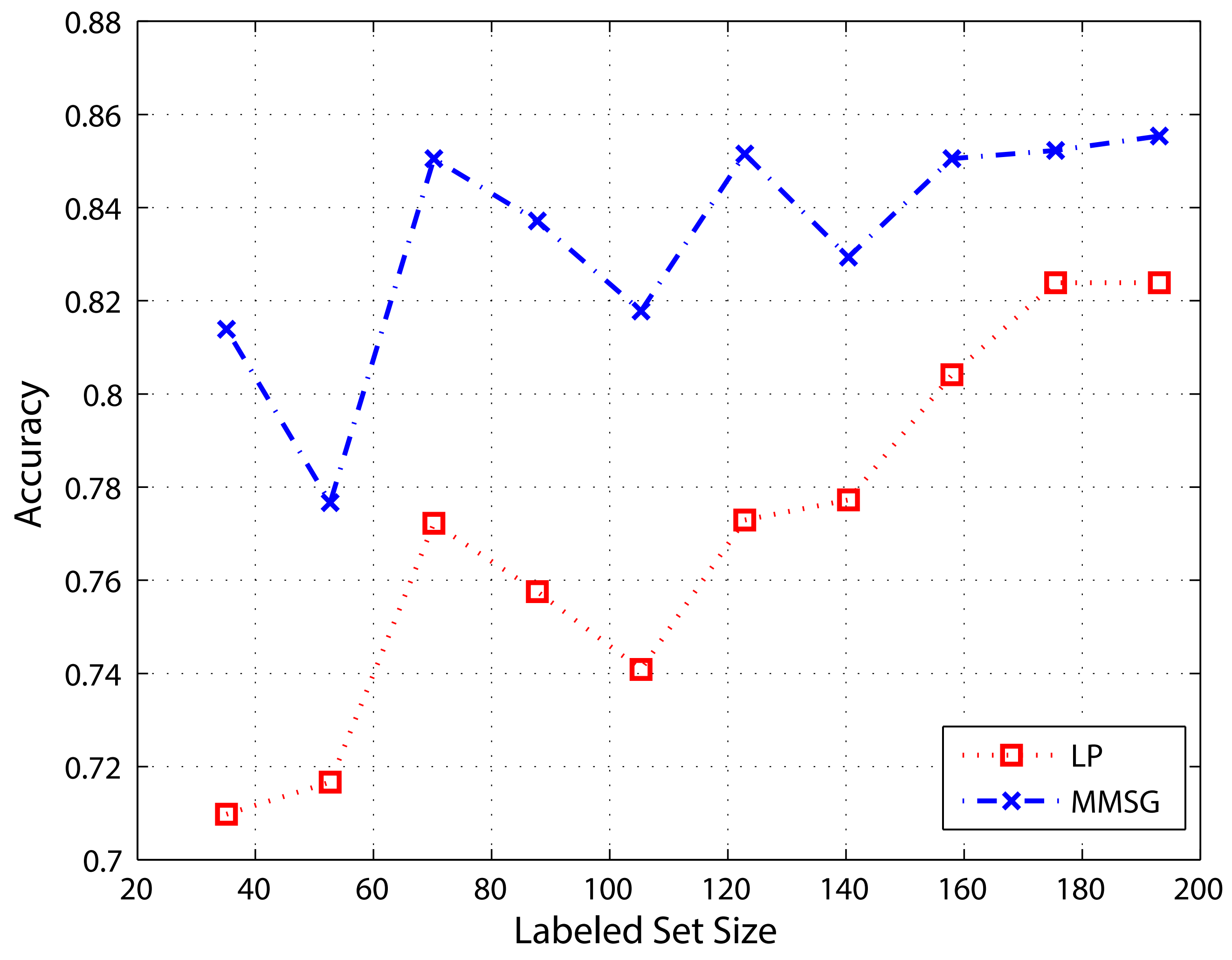
Formulation Continued

γ_i is rewritten as $\gamma_i = \frac{1}{n}\sum\limits_{j=1}^{n} S(\mathbf{x}_i, \mathbf{x}_j)\big(1 - \frac{1}{2}(\ell_i - \ell_j)^2\big)$ and it is convex function of $\boldsymbol{\ell}$. Note that when the separation parameter is sufficiently large ($\gamma \ge \max_i\{\gamma_i\}$), $H_{\boldsymbol{\ell}} = 1 - \frac{\sum\limits_{i=1}^{n}\gamma_i}{n\gamma}$, and the optimization problem is reduced to that of Label Propagation. In the iteration $k \ge 0$ of the projected subgradient method, we set $\boldsymbol{\ell}^{(k+1)} = P_C(\boldsymbol{\ell}^{(k)} - \eta_k \mathbf{g}^{(k)})$ where η_k is the learning rate, $\mathbf{g}^{(k)} \in \mathbb{R}^n$ is the subgradient of $H_{\boldsymbol{\ell}}$ at $\boldsymbol{\ell}^{(k)}$:

$$\mathbf{g}_t^{(k)} = \frac{1}{n^2\gamma}\Big(\big(\sum_{j=1}^{n} S(\mathbf{x}_t, \mathbf{x}_j)(\ell_t^{(k)} - \ell_j^{(k)})\big)\mathbb{1}_{\gamma_t < \gamma} + \\ \sum_{i \neq t} S(\mathbf{x}_i, \mathbf{x}_t)(\ell_t^{(k)} - \ell_i^{(k)})\mathbb{1}_{\gamma_i < \gamma}\Big),\ 1 \le t \le n$$

Experimental Results

we conduct experiments on the UCI Ionosphere data set, and the accuracy of LP and MMSG with respect to different labeled set size are shown in the figure below. The bounded similarity function $S(u, v) = \exp\big(-\frac{1}{0.03}(1 - \frac{u^T v}{\|u\|_2\|v\|_2})\big)$, is used throughout our experiments. For each labeled set size, we perform 5 trials randomly. We set the the separation parameter γ as the average of $\{\gamma_i\}$ using the labels produced by Label Propagation. It is observed that the accuracy curves of MMSG and LP share the same tendency with respect to the size of labeled set, and MMSG always achieves better accuracy by minimizing the hinge loss of the similarity margins


Accuracy
Labeled Set Size
LP
MMSG