Pattern Recognition Letters 37 (2014) 192–200

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Laplacian minimax probability machine

K. Yoshiyama *, A. Sakurai [1]

Graduate School of Science for Open and Environmental System, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawaken 223-8522, Japan

ARTICLE INFO

Article history:
Available online 12 January 2013

Keywords:
Semi-supervised learning
Manifold regularization
Minimax probability machine
Laplacian SVM
Laplacian RLS

ABSTRACT

In this paper, we propose a Laplacian minimax probability machine, which is a semi-supervised version of minimax probability machine based on the manifold regularization framework. We also show that the proposed method can be kernelized on the basis of a theorem similar to the representer theorem for non-linear cases. Experiments confirm that the proposed methods achieve competitive results, as compared to existing graph-based learning methods such as the Laplacian support vector machine and the Laplacian regularized least square, for publicly available datasets from the UCI machine learning repository.

1. Introduction

The objective of semi-supervised learning is to utilize many unlabeled samples coupled with a few labeled samples to improve the generalization performance of a learned model. Recently, many semi-supervised learning methods have been proposed from different viewpoints, such as density-based, cluster-based or graph-based (e.g., Belkin et al., 2006; Bousquet et al., 2004; Chapelle and Zien, 2005;Sindhwani et al., 2005; Zhu et al., 2003), and correspondingly, by formulating different forms of loss function and/or regularization terms based on original optimization problems.

In most of these proposals, existing learning methods have been extended for use in semi-supervised settings. We follow these approaches to extend the minimax probability machine (MPM) to a semi-supervised framework (Yoshiyama and Sakurai, 2012), and we propose a Laplacian minimax probability machine (Lap-MPM) by adopting graph-based regularization as explained in (Belkin et al., 2006), which leads to explicit modification of the MPM bias with the square root of a graph-based regularization term. In addition, we show that the proposed Lap-MPM can be suitably kernelized for non-linear cases.

Our experiments show that the proposed methods achieve competitive results, as compared to existing graph-based semi-supervised methods, i.e., the Laplacian regularized least-square (Lap-RLS) and the Laplacian support vector machine (Lap-SVM), using 20 benchmark datasets from the UCI machine learning repository.

The remainder of this paper is organized as follows. In Section 2, we review related studies, and we present MPM, Lap-RLS, and Lap-SVM, which constitute the basis of the proposed method. In Section 3, we describe an approach for extending MPM to a semi-supervised framework, and we show that it can be suitably kernelized for non-linear cases. In Section 4, we discuss the computational complexity of our proposed algorithm to solve Lap-MPM with the algorithms to solve Lap-SVM. In Section 5, we present the empirical results of our experiments, and we compare the proposed method with existing semi-supervised methods, i.e., Lap-SVM and Lap-RLS. In Section 6, we discuss relevant issues and state our conclusion.

2. Related work

In Section 2.1, we summarize MPM (Lanckriet et al., 2002), in Section 2.2, we review the regularization framework (Belkin et al., 2005) and mention other graph-based methods.

2.1. Minimax probability machine

As in (Lanckriet et al., 2002), consider the set of hyperplanes $\mathcal{H}(\mathbf{a},b)=\{\boldsymbol{a}^T\boldsymbol{z}-b=0|\boldsymbol{a},\boldsymbol{z}\in\mathbb{R}^d,b\in\mathbb{R}\}$ that hopefully separate two classes $\mathcal{X}$ and $\mathcal{Y}$ with maximum lowerbound of worst-case correct classification probability. MPM maximizes α, a lower bound of membership probability to each class with respect to all distributions having the prescribed means and covariance matrices. This is expressed as

$$\max_{\alpha,\boldsymbol{a}\neq 0,b}\ \alpha\quad \text{s.t.}\quad \inf_{\boldsymbol{x}\sim(\bar{\boldsymbol{x}},\Sigma_{\boldsymbol{x}})} \boldsymbol{Pr}\{\boldsymbol{a}^T\boldsymbol{x}\geqslant b\}\geqslant\alpha,\quad \inf_{\boldsymbol{y}\sim(\bar{\boldsymbol{y}},\Sigma_{\boldsymbol{y}})} \boldsymbol{Pr}\{\boldsymbol{a}^T\boldsymbol{y}\leqslant b\}\geqslant\alpha, \tag{1}$$

where $(\bar{\boldsymbol{x}},\Sigma_{\boldsymbol{x}})$ refers to the class of distributions having prescribed mean $\bar{\boldsymbol{x}}$ and covariance $\Sigma_{\boldsymbol{x}}$, but are otherwise arbitrary; likewise for $\boldsymbol{y}$. By exploiting the Marshall–Olkin Theorem (Bertsimas and

* Corresponding author. Tel.: +81 45 566 1659; fax: +81 45 566 1617.
E-mail addresses: k_yoshiyama@ae.keio.ac.jp (K. Yoshiyama), sakurai@ae.keio.ac.jp (A. Sakurai).
[1] Principal corresponding author.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.01.004

Popescu, 2005; Lanckriet et al., 2002), the optimization problem (1) can be rewritten as

$$\max_{\alpha,\boldsymbol{a}\neq 0,b} \alpha \quad \text{s.t.} \quad \boldsymbol{a}^T\overline{\boldsymbol{y}}+\kappa(\alpha)\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{y}}\boldsymbol{a}} \leqslant b \leqslant \boldsymbol{a}^T\overline{\boldsymbol{x}}-\kappa(\alpha)\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}}\boldsymbol{a}}, \tag{2}$$

where $\kappa(\alpha)=\sqrt{\frac{\alpha}{1-\alpha}}$. Since maximizing α is equivalent to maximizing $\kappa(\alpha)$, we can maximize κ without considering α. Furthermore, the upper and lower bound of b in (2) are monotonically and unboundedly decreasing and increasing function of κ respectively. Thus, we can eliminate b at the optimum, which converts (2) into the following optimization problem:

$$\max_{\kappa,\ \boldsymbol{a}\neq 0} \kappa \quad \text{s.t.} \quad \boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}}) \geqslant \kappa\left(\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}}\boldsymbol{a}}+\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{y}}\boldsymbol{a}}\right). \tag{3}$$

If $\overline{\boldsymbol{x}}=\overline{\boldsymbol{y}}$, $\kappa=0$. In this case, MPM does not have a meaningful solution. Assuming $\overline{\boldsymbol{x}}\neq\overline{\boldsymbol{y}}$, we get $\boldsymbol{a}\neq 0$ and $\left(\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{x}}\boldsymbol{a}}+\sqrt{\boldsymbol{a}^T\Sigma_{\boldsymbol{y}}\boldsymbol{a}}\right)\neq 0$. Further, the right-hand side of the inequality constraint in (3) implies that $\boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}})\geqslant 0$, and if an $\boldsymbol{a}$ satisfies the inequality in (3), so does $c\boldsymbol{a}$ with $c\geqslant 0$. Therefore, we set $\boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}})=1$ without loss of generality. Finally, the problem (3) reduces to the following optimization problem with respect to $\boldsymbol{a}$:

$$\kappa_*^{-1}=\min_{\boldsymbol{a}}\left(\|\Sigma_{\boldsymbol{x}}^{1/2}\boldsymbol{a}\|_2+\|\Sigma_{\boldsymbol{y}}^{1/2}\boldsymbol{a}\|_2\right) \quad \text{s.t.} \quad \boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}})=1. \tag{4}$$

If the problem (4) is feasible and convex, and its objective is bounded below, there exists an optimal $\boldsymbol{a}_*$. In addition, the optimal b can be computed as $b_*=\boldsymbol{a}_*^T\overline{\boldsymbol{x}}-\kappa_*\sqrt{\boldsymbol{a}_*^T\Sigma_{\boldsymbol{x}}\boldsymbol{a}_*}=\boldsymbol{a}_*^T\overline{\boldsymbol{y}}+\kappa_*\sqrt{\boldsymbol{a}_*^T\Sigma_{\boldsymbol{y}}\boldsymbol{a}_*}$.

2.2. Manifold regularization

To utilize unlabeled samples, a manifold regularization framework (Belkin et al., 2006) was proposed to introduce a regularization that exploits the geometry of marginal distribution.

Suppose that there is a probability distribution P on $X\times\mathbb{R}$, which generates samples. Labeled examples are $(\boldsymbol{x},y)$ pairs generated according to P, and unlabeled examples are $\boldsymbol{x}\in X$ generated according to the marginal distribution P_X of P. In the manifold regularization framework, a roughness penalty on possible solutions f is imposed by adding a penalty term $\|f\|_I^2$ to an objective, where the norm is defined on a manifold $\mathcal{M}$, the support of P_X.

Since P_X is unknown in most applications, an approximation based on the labeled and unlabeled samples shall be considered. In (Belkin et al., 2006), $\int_{\mathcal{M}}\langle\nabla_{\mathcal{M}}f,\nabla_{\mathcal{M}}f\rangle$ is used as $\|f\|_I^2$, and it is approximated on the basis of labeled and unlabeled samples using the graph Laplacian associated with the samples. Here, the graph is an approximation of the manifold $\mathcal{M}$, where a node $\boldsymbol{x}$ in the graph is a point in $\mathcal{M}$ and the weight w_{ij} on an edge connecting two nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is the adjacency of the nodes. If we choose exponential weights, e.g., $w_{ij}=\exp[-\sigma_s\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2]$, when the number of points approaches infinity, after appropriate scaling, the graph Laplacian converges to the true Laplace–Beltrami operator on the manifold (Theorem 3 in (Belkin and Niyogi, 2005)). Therefore, we consider $\frac{1}{(\ell+u)^2}\sum_{i,j=1}^{\ell+u}w_{ij}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2$ instead of $\|f\|_I^2$.

Suppose that we are given a set of labeled samples $\{\boldsymbol{x}_i,y_i\}_{i=1}^{\ell}$ and a set of unlabeled samples $\{\boldsymbol{x}_j\}_{j=\ell+1}^{\ell+u}$, and $\|f\|^2$ is an appropriate smoothness measure on f in the function space of possible solutions; then, the optimization problem with the manifold regularization is

$$\begin{aligned}&\arg\min_f \frac{1}{\ell}\sum_{i=1}^{\ell}V(\boldsymbol{x}_i,y_i,f)+\gamma_A\|f\|^2+\frac{\gamma_I}{(\ell+u)^2}\sum_{i,j=1}^{\ell+u}w_{ij}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2\\ &=\arg\min_f \frac{1}{\ell}\sum_{i=1}^{\ell}V(\boldsymbol{x}_i,y_i,f)+\gamma_A\|f\|^2+\frac{\gamma_I}{(\ell+u)^2}\boldsymbol{f}^T L\boldsymbol{f},\end{aligned} \tag{5}$$

where $\boldsymbol{f}=[f(\boldsymbol{x}_1),\ldots,f(\boldsymbol{x}_{\ell+u})]^T$, $V(\boldsymbol{x}_i,y_i,f)$ is some loss function, γ_A and γ_I control the complexity of f in the function space and in the intrinsic geometry of P_X respectively, and $L=D-W$ is the graph Laplacian. Here, W is the edge weights matrix of the data adjacency graph having elements w_{ij}. $D\in\mathbb{R}^{(\ell+u)\times(\ell+u)}$ is a diagonal matrix whose elements $\forall i\ D_{ii}=\sum_{j=1}^{\ell+u}w_{ij}$ and otherwise 0.

In this manifold regularization framework, by choosing squared loss $(y_i-f(\boldsymbol{x}_i))^2$ and hinge loss $\max[0,1-y_if(\boldsymbol{x}_i)]$ as the loss function for RLS and SVM respectively, RLS and SVM are extended to semi-supervised versions, Lap-RLS and Lap-SVM (Belkin et al., 2006).

Other related works based on the manifold regularization framework are presented in (Goldberg et al., 2007,). They incorporated dissimilarity into their objective function. In (Chapell et al., 2008), the graph Laplacian was combined with a semi-supervised SVM. Further, a smoothness measure analogous to the manifold regularization was used in (Li et al., 2008).

3. Laplacian minimax probability machine

In this section, we show that MPM can be extended to a manifold-regularized version, and we propose an algorithm, block coordinate descent, to solve the Lap-MPM optimization problem. Furthermore, we can obtain a kernelized version of Lap-MPM, called Lap-KMPM, on the basis of a theorem similar to Corollary 5 in (Lanckriet et al., 2002).

3.1. Linear case

Here, our objective is to construct linear Lap-MPM to exploit the unlabeled samples. In order to incorporate the manifold regularization framework, we introduce a manifold regularization term to the objective in the optimization problem (4), as in the case of Belkin et al. (2006).

Let $\{\boldsymbol{x}_i\}_{i=1}^{N_x}$, $\{\boldsymbol{y}_i\}_{i=1}^{N_y}$, $f\in\mathcal{H}(\boldsymbol{a},b)$ be as in Section 2.1, and $\{\boldsymbol{z}_i\}_{i=1}^{N_z}$ denote unlabeled samples. Then, the optimization problem (4) becomes

$$\begin{aligned}\kappa_*^{-1}=\min_{\boldsymbol{a}}&\left(\|\Sigma_{\boldsymbol{x}}^{1/2}\boldsymbol{a}\|_2+\|\Sigma_{\boldsymbol{y}}^{1/2}\boldsymbol{a}\|_2+\frac{\gamma_I}{(\ell+u)^2}\sum_{i,j=1}^{\ell+u}w_{ij}(f(\boldsymbol{t}_i)-f(\boldsymbol{t}_j))^2\right)\\ &\text{s.t.}\quad \boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}})=1,\end{aligned} \tag{6}$$

where $\boldsymbol{t}\in\{\boldsymbol{x}_i\}_{i=1}^{N_x}\cup\{\boldsymbol{y}_i\}_{i=1}^{N_y}\cup\{\boldsymbol{z}_i\}_{i=1}^{N_z}$. Since $(f(\boldsymbol{t}_i)-f(\boldsymbol{t}_j))$ in the problem (6) is equal to $(\boldsymbol{a}^T\boldsymbol{t}_i-\boldsymbol{a}^T\boldsymbol{t}_j)$, the optimization problem (6) can be rewritten as

$$\begin{aligned}\kappa_*^{-1}=\min_{\boldsymbol{a}}&(\|\Sigma_{\boldsymbol{x}}^{1/2}\boldsymbol{a}\|_2+\|\Sigma_{\boldsymbol{y}}^{1/2}\boldsymbol{a}\|_2+\frac{\gamma_I}{(\ell+u)^2}\boldsymbol{a}^TZLZ^T\boldsymbol{a})\\ &\text{s.t.}\quad \boldsymbol{a}^T(\overline{\boldsymbol{x}}-\overline{\boldsymbol{y}})=1,\end{aligned} \tag{7}$$

where $Z\in\mathbb{R}^{d\times n}$ is a matrix composed of all labeled and unlabeled samples, $n=N_x+N_y+N_z$, and L is the graph Laplacian given by $L=D-W$. Note that the elements of Z are ordered by samples belonging to the class $\mathcal{X}$, $\mathcal{Y}$, and unlabeled samples, and the elements of W are constructed in the same order.

Although the introduction of the manifold regularization term is straightforward, it is clear that the first and second terms appearing in the objective of (7) and the third term differ in scale and/or dimension. Therefore, we introduce the square root of the manifold regularization term as our regularization term, where the normalizing factor $\frac{1}{(\ell+u)^2}$ and regularization parameter γ_I are coerced into one parameter λ. Note that even if we introduce the square root term instead of the original form, the representer theorem still holds, as will be shown in the following section, which is