Adaptive Semi-supervised Learning with Discriminative Least Squares Regression

Minnan Luo[1], **Lingling Zhang**[1], **Feiping Nie**[2*], **Xiaojun Chang**[3], **Buyue Qian**[1], **Qinghua Zheng**[1]
[1]SPKLSTN Lab, Department of Computer Science, Xi'an Jiaotong University, Shaanxi, China.
[2]Center for OPTical Imagery Analysis and Learning, Northwestern Polytechnical University, China.
[3]School of Computer Science, Carnegie Mellon University, PA, USA.
{minnluo,qianbuyue,qhzheng}@xjtu.edu.cn, zhanglingling@stu.xjtu.edu.cn,
{feipingnie,cxj273}@gmail.com

Abstract

Semi-supervised learning plays a significant role in multi-class classification, where a small number of labeled data are more deterministic while substantial unlabeled data might cause large uncertainties and potential threats. In this paper, we distinguish the label fitting of labeled and unlabeled training data through a probabilistic vector with an adaptive parameter, which always ensures the significant importance of labeled data and characterizes the contribution of unlabeled instance according to its uncertainty. Instead of using traditional least squares regression (LSR) for classification, we develop a new discriminative LSR by equipping each label with an adjustment vector. This strategy avoids incorrect penalization on samples that are far away from the boundary and simultaneously facilitates multi-class classification by enlarging the geometrical distance of instances belonging to different classes. An efficient alternative algorithm is exploited to solve the proposed model with closed form solution for each updating rule. We also analyze the convergence and complexity of the proposed algorithm theoretically. Experimental results on several benchmark datasets demonstrate the effectiveness and superiority of the proposed model for multi-class classification tasks.

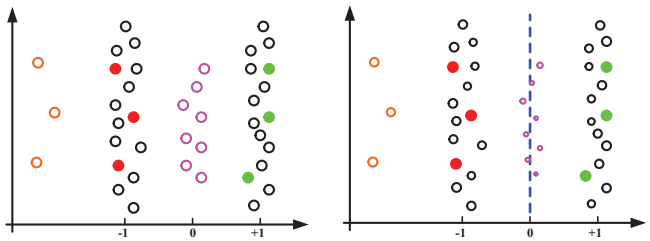
1 Introduction

Data are abundant due to the digital information explosion. However, a vast majority of large-scale data are usually collected without labels for classification. Labeling these data requires the efforts of human annotators who must often be quite skilled [Zhu *et al.*, 2003; Yang *et al.*, 2015]. Moreover, this process is difficult to scale and often error prone for its dramatically expensive and time-consuming cost [Subramanya and Talukdar, 2014; Chang *et al.*, 2014; Luo *et al.*, 2017]. As a result, it is crucial to develop semi-supervised learning (SSL) algorithms which potentially improve the generalization performance by employing a small amount of labeled data jointly with a large number of unlabeled data [Chapelle *et al.*, 2008; Zhu and Goldberg, 2009; Chang and Yang, 2016].

In the past decades, researchers have proposed a variety of semi-supervised learning algorithms. Most of them are started by representing the training data (labeled and unlabeled) as a graph and assume that data points are likely to have the same label if they are close to each other [Chapelle *et al.*, 2009]. For example, semi-supervised support vector machines (SVMs) with manifold regularization [Sindhwani and Keerthi, 2006; Belkin *et al.*, 2006; Liu *et al.*, 2011], graph kernels based SSL [Zhu *et al.*, 2004], semi-supervised discriminant analysis from graph perspective [Cai *et al.*, 2007a], SSL with local and global consistency [Zhou *et al.*, 2004], gaussian eld harmonic function [Zhu *et al.*, 2003] and so on. Since data points from different classes can be quite close in practice, the graph-based algorithms might fail to achieve desirable performance when two classes overlap significantly [Zhu and Goldberg, 2009]. Additionally, conventional semi-supervised learning algorithms treat the labeled and unlabeled instances equally in the training stage. In fact, the labeled data are usually more deterministic and should play a more significant role, while substantial unlabeled data might cause large uncertainties and potential threats. Thus, it is crucial to distinguish between different label types of training instances in the framework of SSL [Nie *et al.*, 2011b]. Wang *et al*. [Wang *et al.*, 2014] introduced an adaptive parameter to suppress the weights of unlabeled data points around the boundary. However, this method employs traditional least squares regression (LSR) loss function for classification, which usually penalizes the data points far away from the boundary even when they are classified correctly for sure [Bishop, 2006; Xiang *et al.*, 2012].

In this paper, we exploit an adaptive semi-supervised classification algorithm with a novel discriminative LSR. This idea is illustrated by a synthetic dataset (see Figure 1a) consisting of labeled data points (filled circles from two classes) and unlabeled ones (hollow circles). Specifically, the unlabeled data points with various uncertainties include the purple ones around the boundary, the brown ones far away from the boundary, and the black ones following the distribution of the labeled data. By assigning a probabilistic vector to the fitting of each sample, our method always ensure the significant importance of labeled data while characterizing the

*Corresponding author: Feiping Nie (feipingnie@gmail.com)



(a) Synthetic data (b) Classification result

Figure 1: (a) The synthetic dataset. (b) The contribution importance of each data point in the training stage is characterized by its size. The learned classifier (dotted blue line) by our model is not distorted by the unlabled data points far away from the boundary (the brown hollow circle) and the data points around the boundary (the purple hollow circle).

contribution of the unlabeled instance according to its uncertainty. See the classification performance of our model in Figure 1b. The size of labeled data remain unchanged, while the unlabeled data points with comparable deterministic affiliation become smaller. Specifically the unlabeled data points around the boundary that belong to multiple labels with similar probabilities turn to be much smaller. To overcome the shortcomings of the conventional LSR for classification, we equip each label with an adjustment vector. This strategy prevents the incorrect penalization of unlabeled samples far away from the boundary. As a result, our model performs robust to not only the unlabeled data with varying uncertainties but also the outliers far away from the boundary. In summary, we describe the contributions of this paper as follows:

1. We associate each label with an adjustment vector and develop a new discriminative LSR to prevent incorrect penalization on the data far away from the boundary.
2. Based on the proposed discriminative LSR, we concern the contribution difference of training data in semi-supervised learning and developed a unified label fitting for both labeled and unlabeled training data.
3. We exploit an efficient alternative algorithm to solve the proposed model with closed form solution for each updating rule, and analyze its convergence and computational complexity theoretically.

2 The Proposed Methodology

In this section, we first introduce a new discriminative least squares regression, and then propose an adaptive framework for semi-supervised classification.

2.1 Discriminative LSR

In this paper, we denote the training set by $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of the i-th instance. In the framework of semi-supervised classification, we suppose the first $n_l (n_l \leq n)$ instances, *i.e.*, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n_l}$, are labeled uniquely c different classes. The remaining $n_u = n - n_l$ instances, *i.e.*, $\mathbf{x}_{n_l+1}, \mathbf{x}_{n_l+2}, \cdots, \mathbf{x}_n$, are unlabeled. To identify the c classes uniquely, we associate the k-th class with a coding $\mathbf{t}_k = [-1, \cdots, -1, 1, -1, \cdots, -1]^\top \in \mathbb{R}^c$, where only the k-th element of $\mathbf{t}_k$ is equal to 1 and the other ones are -1.

Instead of using traditional regression loss which usually incorrectly penalizes the right classification, in this paper, we take $\mathbf{t}_k$ $(k = 1, 2, \cdots, c)$ as the regression target and associate each class with an adjustment variable $\mathbf{m}_k = [m_{k1}, m_{k2}, \cdots, m_{kc}]^\top \geq 0$ $(k = 1, 2, \cdots, c)$. Let $W \in \mathbb{R}^{d\times c}$ be a transformation matrix and $\mathbf{b} \in \mathbb{R}^c$ be a bias vector, the proposed discriminative regression loss (fitting loss) for the i-th instance which belongs to the k-th class is defined as

$$D_{ik}(W, \mathbf{b}, \mathbf{m}_k; \mathbf{x}_i) = \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{t}_k - \mathbf{t}_k \odot \mathbf{m}_k\|_2^2$$

where $\odot$ is a Hadamard product operator of vectors. In contrast to the traditional regression loss, the introduced term $\mathbf{t}_k \odot \mathbf{m}_k$ are capable of not only offsetting the incorrect penalization but also enlarging the margin between classes. For a better representation, we collect all of adjustment variables into matrix $M = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_c] \in \mathbb{R}^{c\times c}$. We take the instance $\mathbf{x}_i \in \mathbb{R}^3$ which might belong to one of $c = 3$ different classes for example. The discriminative regression loss of $\mathbf{x}_i$ belonging to the 2-th class is formulated as

$$\begin{aligned} D_{i2}(W, \mathbf{b}, \mathbf{m}_2; \mathbf{x}_i) &= \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{t}_2 - \mathbf{t}_2 \odot \mathbf{m}_2\|_2^2 \\ &= \left\| \begin{pmatrix} z_{i1} - (-1 - m_{21}) \\ z_{i2} - (1 + m_{22}) \\ z_{i3} - (-1 - m_{23}) \end{pmatrix} \right\|_2^2 \end{aligned}$$

where $\mathbf{z}_i = [z_{i1}, z_{i2}, z_{i3}]^\top$ is the predicted label vector for $\mathbf{x}_i$. If $\mathbf{x}_i$ belongs to the 2-th class with the predicted label vector $\mathbf{z}_i \gg \mathbf{t}_2$, the value of loss D_{i2} keeps small with an appropriate non-negative adjustment variable $\mathbf{m}_2$. This strategy effectively prevents the correct classification from being penalized. Moreover, the introduced nonnegative variable $\mathbf{m}_2$ intuitively enlarge the distance between different classes in the projected space as much as possible. Note that the proposed discriminative LSR is more efficient and usually have better generalization ability than the one in [Xiang *et al.*, 2012] which associates each instance with c non-negative variables.

2.2 Semi-supervised Classification with DLSR

Taking the advantages of discriminative LSR, we consider the contribution importance of varying training data in the framework of semi-supervised learning and propose a novel fitting loss for both labeled and unlabeled data in a unified formulation, *i.e.*,

$$\mathcal{L}(W, \mathbf{b}, M, Y) = \sum_{i=1}^{n}\sum_{k=1}^{c} y_{ik}^r D_{ik}(W, \mathbf{b}, \mathbf{m}_k; \mathbf{x}_i) \quad (1)$$

where $y_{ik} \in [0, 1]^c$ $(i = 1, 2\cdots, n; k = 1, 2, \cdots, c)$ refers to the probability of the i-th instance belonging to the k-th class. For a training instance $\mathbf{x}_i \in \mathcal{X}$ with probability vector $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{ic}] \in [0, 1]^c$, its contribution importance is defined as

$$\omega_i^r = \sum_{k=1}^{c} y_{ik}^r \quad (i = 1, 2, \cdots, n)$$

where $r \geq 1$ is an adaptive parameter. The advantages of this definition are embodied in the following three aspects. Firstly, it ensures the significance of labeled data since the contribution importance of any labeled instance is always 1 for any adaptive parameter $r \geq 1$. Secondly, it is capable of reflecting how much proportions of the unlabeled data play at the macro level because, on one hand, the unlabeled instance is considered to have the equal importance as labeled data when the adaptive parameter $r \to 1$; On the other hand, the unlabeled instance has no effect on the training stage as the adaptive parameter $r \to \infty$. Thirdly, the contribution importance with fixed $1 < r < \infty$ can distinguish between different types of unlabeled data at the micro level. This setting coincides with the fact that uncertain unlabeled instances should have less influence on the procedure of training. In summary, the contribution importance not only distinguish between labeled and unlabeled instances at the macro level but also differentiate the varieties of unlabeled data at the micro level.

Let $Y = [Y_l; Y_u] \in [0,1]^{n \times c}$ be the probability matrix for training data, where $Y_l = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{n_l}]^\top \in [0,1]^{n_l \times c}$ and $Y_u = [\mathbf{y}_{n_l+1}, \mathbf{y}_{n_l+2}, \cdots, \mathbf{y}_n]^\top \in [0,1]^{n_u \times c}$ refer to the probability matrices of labeled and unlabeled data, respectively. We employ the loss function (1) and formulate the idea of semi-supervised classification with discriminative LSR as the following optimization problem

$$\begin{aligned} \min_{\Omega} \quad & \mathcal{L}(W, \mathbf{b}, M, Y) + \lambda \|W\|_F^2 \\ \textit{s.t.} \quad & \sum_{k=1}^{c} y_{ik} = 1, y_{ik} \in [0,1], \mathbf{m}_k \geq 0 \ (\forall i, k); \end{aligned} \tag{2}$$

where $\Omega = \{W, \mathbf{b}, Y_u, M\}$ collects all of the optimization variables. $r \geq 1$ is the adaptive parameter which needs to be tuned; The regularization $\|W\|_F^2$ controls the complexity of model with a trade-off parameter λ. In contrast to the approach used in [Wang *et al.*, 2014], the proposed optimization problem (2) not only demonstrates excellent robustness to the unlabeled uncertainty instances that are near the classification boundary but also prevents the correct classification from being penalized.

3 Optimization Procedure

In this section, we exploit a simple and efficient alternative algorithm to solve the proposed optimization problem (2), and analyze its computational complexity and convergence theoretically.

3.1 Optimize Unlabeled Probability Matrix Y_u:

Consider the independence of training data, probability vector for each instance can be updated simultaneously by solving the following optimization problem

$$\min_{\mathbf{y}_i} \sum_{k=1}^{c} y_{ik}^r D_{ik} \quad \textit{s.t.} \quad \sum_{k=1}^{c} y_{ik} = 1; y_{ik} \in [0,1] \ (\forall i) \tag{3}$$

where $D_{ik} = D_{ik}(W, \mathbf{b}, \mathbf{m}_k; \mathbf{x}_i)$ is calculated with fixed parameters $W, \mathbf{b}$ and $\mathbf{m}_k$ $(k = 1, 2, \cdots, c)$. If the adaptive parameter r is set to 1, the optimization problem (3) has a trivial solution

$$y_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_j D_{ij}; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

for $k = 1, 2, \cdots, c$. In this case, the labeled data and unlabeled data are treated equally in the semi-supervised classification. For a more general case that $r > 1$, let the Lagrangian function of objective (3) be $\mathcal{L}_\alpha = \sum_{k=1}^{c} y_{ik}^r D_{ik} - \alpha(\sum_{k=1}^{c} y_{ik} - 1)$, where α is the Lagrangian multiplier. Setting the derivative of Lagrangian function $\mathcal{L}_\alpha$ in terms of y_{ik} to zero and combining with the constraint $\sum_{k=1}^{c} y_{ik} = 1$, we arrive at the following closed-form solution for optimization problem (3)

$$y_{ik} = \left[\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{ij}} \right)^{\frac{1}{r-1}} \right]^{-1} \tag{5}$$

Given this closed-form solution, the optimal probability vector for each unlabeled instance is updated efficiently.

3.2 Optimize W and $\mathbf{b}$:

To update W and $\mathbf{b}$ with fixed Y and M, we frist introduce the following Theorem 1.

Theorem 1. *Given $W \in \mathbb{R}^{d \times c}$, $\mathbf{b} \in \mathbb{R}^c$, $Y \in \mathbb{R}^{n \times c}$ and $M \in \mathbb{R}^{c \times c}$, let $T = [\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_c] \in \mathbb{R}^{c \times c}$ and $H = T + T \odot M$. The loss function $\mathcal{L}(W, \mathbf{b}, M, Y)$ can be reformulated as a compact matrix representation*

$$\begin{aligned} \mathcal{L}(W, \mathbf{b}, M, Y) = & Tr\left[(W^\top X + \mathbf{b}\mathbf{1}_n^\top) \Omega (W^\top X + \mathbf{b}\mathbf{1}_n^\top)^\top \right] \\ & - 2Tr\left[F H^\top (W^\top X + \mathbf{b}\mathbf{1}_n^\top) \right] + Tr\left[H G H^\top \right] \end{aligned}$$

where $\Omega = diag(\omega_1^r, \omega_2^r, \cdots, \omega_n^r) \in \mathbb{R}^{n \times n}$ is a diagonal matrix with its (i,i)-th diagonal element being the contribution importance of the i-th instance; $F = Y^r$ is computed in element wise; $G = diag(g_{11}, g_{22}, \cdots, g_{cc}) \in \mathbb{R}^{c \times c}$ is also a diagonal matrices with its (k,k)-th diagonal element being $g_{kk} = \sum_{i=1}^{n} y_{ik}^r$ $(k = 1, 2, \cdots, c)$.

Proof. The result is derived with some matrix theories. □

Based on Theorem 1, we update W and $\mathbf{b}$ according to the following Theorem 2.

Theorem 2. *Given fixed parameters Y and M, the solutions of optimization problem (2) with respect to variables W and $\mathbf{b}$ are derived as the following closed forms*

$$W = (X \Omega P X^\top + \lambda I_d)^{-1} X P^\top F H^\top \tag{6}$$

$$\mathbf{b} = (H F^\top - W^\top X \Omega) \mathbf{1}_n / \mathbf{1}_n^\top \Omega \mathbf{1}_n \tag{7}$$

where $P = I_n - \mathbf{1}_n \mathbf{1}_n^\top \Omega / \mathbf{1}_n^\top \Omega \mathbf{1}_n \in \mathbb{R}^{n \times n}$; I_d and I_n are identity matrices of size $d \times d$ and $n \times n$, respectively.

Proof. According to Theorem 1, the optimal W and $\mathbf{b}$ can be achieved through minimizing the objective function $f(W, \mathbf{b}) = Tr\left[(W^\top X + \mathbf{b}\mathbf{1}_n^\top) \Omega (W^\top X + \mathbf{b}\mathbf{1}_n^\top)^\top \right] -$

$2Tr\left[FH^\top(W^\top X+\mathbf{b}\mathbf{1}_n^\top)\right]+\lambda\|W\|_F^2$. By some matrix theories, we set the derivative of $f(W,\mathbf{b})$ with respect to variable $\mathbf{b}$ to zero, *i.e.*,

$$\frac{\partial f(W,\mathbf{b})}{\partial \mathbf{b}}=2[(W^\top X+\mathbf{b}\mathbf{1}_n^\top)\Omega-HF^\top]\mathbf{1}_n=0. \quad (8)$$

Thus, we have $\mathbf{b}=(HF^\top-W^\top X\Omega)\mathbf{1}_n/\mathbf{1}_n^\top\Omega\mathbf{1}_n$. Moreover, setting the derivative of $f(W,\mathbf{b})$ with respect to variable W to zero, *i.e.*,

$$\frac{\partial f(W,\mathbf{b})}{\partial W}=2X\Omega(X^\top W+\mathbf{1}_n\mathbf{b}^\top)-2XFH^\top+2\lambda W=0,$$

and substituting the derivation of $\mathbf{b}$ into the equation above, we arrive at

$$X\Omega PX^\top W+\lambda W=XP^\top FH^\top$$

where $P=I_n-\mathbf{1}_n\mathbf{1}_n^\top\Omega/\mathbf{1}_n^\top\Omega\mathbf{1}_n$. As a result, the optimal W is obtained in a closed form $W=(X\Omega PX^\top+\lambda I_d)^{-1}XP^\top FH^\top$. The proof is completed. □

3.3 Optimize Adjustment Variable M:

Given $W,\mathbf{b}$ and Y, the optimal M can be obtained through solving problem $\min_{M\geq 0}\ \mathcal{L}(W,\mathbf{b},M,Y)$. To this end, we introduce the following Theorem 3.

Theorem 3. *Given $W\in\mathbb{R}^{d\times c}$, $\mathbf{b}\in\mathbb{R}^c$ and $Y\in\mathbb{R}^{n\times c}$, let $F=Y^r\in\mathbb{R}^{n\times c}$ be calculated in element wise; $G=diag(g_{11},g_{22},\cdots,g_{cc})\in\mathbb{R}^{c\times c}$ be a diagonal matrix with its (k,k)-th element $g_{kk}=\sum_{i=1}^n y_{ik}^r$ $(k=1,2,\cdots,c)$. The solution of optimization problem (2) with respect to M is derived as a closed form*

$$M=T\odot\left[(W^\top X+\mathbf{b}\mathbf{1}_n^\top)FG^{-1}-T\right]. \quad (9)$$

Proof. According to the compact matrix representation of loss function $\mathcal{L}(W,\mathbf{b},M,Y)$ in Theorem 1, we denote $H=T+T\odot M$ with $T=[\mathbf{t}_1,\mathbf{t}_2,\cdots,\mathbf{t}_c]\in\mathbb{R}^{c\times c}$, and compute the derivative of $\mathcal{L}(W,\mathbf{b},M,Y)$ with respect to M, *i.e.*,

$$\begin{aligned}\frac{\partial\mathcal{L}(W,\mathbf{b},M,Y)}{\partial M}&=\frac{\partial H}{\partial M}\times\frac{\partial\mathcal{L}(W,\mathbf{b},M,Y)}{\partial H}\\&=T\times\left[2HG-2(W^\top X+\mathbf{b}\mathbf{1}_n^\top)F\right].\end{aligned}$$

Recall that matrix T is invertible, we set the derivative of $\mathcal{L}(W,\mathbf{b},M,Y)$ to 0, and thus obtain

$$H=T+T\odot M=(W^\top X+\mathbf{b})FG^{-1}. \quad (10)$$

Since the (k,k)-th element of T is 1 and the others are -1 for $k=1,2,\cdots,c$, $T\odot T=\mathbf{1}_{c\times c}$ holds, where $\mathbf{1}_{c\times c}$ refers to a matrix whose elements are all 1. Thus we have $M=T\odot\left[(W^\top X+\mathbf{b}\mathbf{1}_n^\top)FG^{-1}-T\right]$. The proof is completed. □

In summary, we describe the alternative algorithm for optimization problem (2) in Algorithm 1. Because of the closed-form solutions to each updating rule, Algorithm 1 is very simple to implement and efficient to run on large scale datasets. Specifically, the overall computational cost of Algorithm 1 is less than $\mathcal{O}(Td^2c)+\mathcal{O}(nd^2)+\mathcal{O}(ndc)$. For real world application, since the number of classes c is always much smaller than the dimensionality of data d and the number of data points n, the upper bound computational cost reduces to $\mathcal{O}(Td^2)+\mathcal{O}(nd^2)$. We also analyze the convergence of the proposed algorithm in the following Theorem 4.

Algorithm 1 Alternative optimization for problem (2)

Input: $X=[X_l;X_u]\in\mathbb{R}^{d\times n}$ with label matrix $F_l\in\mathbb{R}^{n_l\times c}$; λ; $r\geq 1$.
Initialization: $W^{(0)}\in\mathbb{R}^{d\times c}$, $\mathbf{b}^{(0)}\in\mathbb{R}^c$, $M^{(0)}\in\mathbb{R}^{c\times c}$ and $t=0$.
1: **while** not converge **do**
2: Update Y_u^{t+1} given $W^{(t)},\mathbf{b}^{(t)}$ and $M^{(t)}$ by Eq. (4) if $r=1$, by Eq. (5) otherwise;
3: Update $W^{(t+1)}$ and $\mathbf{b}^{t+1}$ given Y^{t+1} and M^t by Eq. (6) and Eq. (7);
4: Update $M^{(t+1)}$ given $W^{(t+1)},\mathbf{b}^{(t+1)}$ and $Y^{(t+1)}$ by Eq. (9);
5: $t=t+1$;
6: **end while**

Theorem 4. *The iterative updating rules in Algorithm 1 monotonically decrease the objective function value of the optimization problem (2) in each iteration until convergence.*

Proof. Let the value of objective function (2) be $\mathcal{F}(W^{(t)},\mathbf{b}^{(t)},M^{(t)},Y^{(t)})=\mathcal{L}(W^{(t)},\mathbf{b}^{(t)},M^{(t)},Y^{(t)})+\lambda\|W^{(t)}\|_F^2$ at the t-th iteration. In alternative optimization Algorithm 1, since we update variable Y with closed-form solution by $Y^{(t+1)}=\arg\min_Y\mathcal{L}(W^{(t)},\mathbf{b}^{(t)},M^{(t)},Y)$, followed by updating variables W and $\mathbf{b}$ according to $(W^{(t+1)},\mathbf{b}^{(t+1)})=\arg\min_{W,b}\mathcal{F}(W,\mathbf{b},M^{(t)},Y^{(t+1)})$, the inequality $\mathcal{F}(W^{(t+1)},\mathbf{b}^{(t+1)},M^{(t)},Y^{(t+1)})\leq\mathcal{F}(W^{(t)},\mathbf{b}^{(t)},M^{(t)},Y^{(t)})$ holds evidently. Furthermore, the adjustment variable M is updated through $M^{(t+1)}=\arg\min_M\mathcal{L}(W^{(t+1)},\mathbf{b}^{(t+1)},M,Y^{(t+1)})$, we have $\mathcal{F}(W^{(t+1)},\mathbf{b}^{(t+1)},M^{(t+1)},Y^{(t+1)})\leq\mathcal{F}(W^{(t)},\mathbf{b}^{(t)},M^{(t)},Y^{(t)})$. Recall objective function (2) is bounded below, thus its value decrease monotonically in each iteration until Algorithm 1 convergence. □

4 Experiment

In this section, several benchmark datasets of varying image types are used to validate the effectiveness and superiority of the proposed model, including the ORL database of faces [Cai *et al.*, 2007b], the extended Yale B database (YaleB) of face [Georghiades *et al.*, 2001], the face database CMU-PIE [Sim *et al.*, 2002] and the palm print database (PALM) [Yan *et al.*, 2007]. We download all of the datasets from different websites. The pixel value of the image is used as its feature representation. Images of the faces are resized to 32×32 while the palm print images in PALM are resized to 16×16.

We compare the proposed model with some state-of-the-art classification methods, including supervised learning methods: SVM [Fan *et al.*, 2005] and classification based on discriminative LSR (SDLSR) with respect to each data points [Xiang *et al.*, 2012], transductive learning methods: Semi-supervised classification with Gaussian Fields and Harmonic Functions (GFHF) [Zhu *et al.*, 2003] and Semi-Supervised Learning with ℓ_1-Norm Graph (ℓ_1-SEMI) [Nie *et al.*, 2011a], two inductive learning: semi-supervised learning with Flexi-

Table 1: Performance comparison on Accuracy STD with 10% training data are labeled.

	ORL		YaleB		CMU-PIE		PALM	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
SVM	0.513±0.027	0.501±0.014	0.187±0.015	0.187±0.021	0.276±0.012	0.279±0.015	0.801±0.012	0.811±0.013
SDLSR	0.660±0.028	0.658±0.023	0.937±0.007	0.940±0.010	0.895±0.013	0.900±0.012	0.925±0.008	0.931±0.012
GFHF	0.614±0.023	NA	0.582±0.022	NA	0.654±0.008	NA	0.914±0.012	NA
ℓ_1-SEMI	0.576±0.016	NA	0.387±0.026	NA	0.553±0.031	NA	0.891±0.013	NA
FME	0.665±0.018	0.672±0.014	0.928±0.008	0.933±0.012	0.881±0.015	0.884±0.015	0.942±0.008	0.945±0.012
ASL	0.674±0.020	0.666±0.013	0.964±0.004	0.962±0.004	0.943±0.009	0.946±0.008	0.963±0.010	0.964±0.015
Our Model	**0.718±0.012**	**0.725±0.012**	**0.996±0.001**	**0.995±0.002**	**0.954±0.005**	**0.957±0.006**	**0.982±0.005**	**0.986±0.007**

Table 2: Performance comparison on Accuracy STD with 20% training data are labeled.

	ORL		YaleB		CMU-PIE		PALM	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
SVM	0.521±0.014	0.514±0.026	0.315±0.030	0.319±0.033	0.378±0.012	0.383±0.011	0.917±0.007	0.921±0.009
SDLSR	0.662±0.018	0.662±0.017	0.982±0.002	0.981±0.003	0.955±0.013	0.957±0.009	0.986±0.005	0.988±0.003
GFHF	0.627±0.023	NA	0.735±0.020	NA	0.800±0.008	NA	0.981±0.004	NA
ℓ_1-SEMI	0.584±0.024	NA	0.576±0.016	NA	0.768±0.031	NA	0.976±0.005	NA
FME	0.672±0.018	0.689±0.023	0.990±0.003	0.989±0.003	0.952±0.015	0.953±0.008	0.989±0.004	0.991±0.003
ASL	0.691±0.027	0.689±0.013	0.996±0.002	0.995±0.002	0.969±0.009	0.968±0.004	0.992±0.005	0.993±0.005
Our Model	**0.732±0.012**	**0.736±0.021**	**0.997±0.001**	**0.997±0.002**	**0.973±0.005**	**0.974±0.005**	**0.994±0.005**	**0.996±0.004**

ble Manifold Embedding (FME) [Nie *et al.*, 2010] and Adaptive Semi-Supervised Learning (ASL) [Wang *et al.*, 2014].

For each dataset, two thirds of samples are randomly selected as the training data, while the remaining ones are served as the testing data. At the semi-supervised training stage, we randomly choose only 10% or 20% samples with labels. For the regularization parameter used in SDLSR, ℓ_1-SEMI, FME , ASL and our model, we tune them in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ and report the best results. Following [Nie *et al.*, 2010], we use the Gaussian function to compute the Laplacian matrix for all graph-based methods. The adaptive parameter used in ASL and our model is tuned from 1 to 2 with a step-size of 0.1.

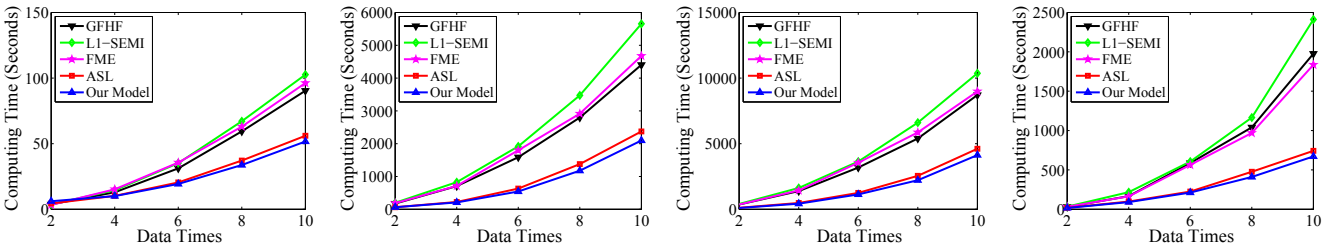
4.1 Experimental Results Comparison

For a fair comparison, we randomly split each dataset into training and testing dataset 50 times and report the average accuracy with standard deviation (STD) for the unlabeled data in training dataset and testing data. The experimental results on four datasets are collected in Table 1 and Table 2 with 10% and 20% labeled samples, where "*NA*" indicates that transductive semi-supervised classification approaches are not able to predict labels for samples out of training stage. From the experimental comparison, we have the following observations. (1) The performance of each algorithm becomes better as the increase of labeled data in the training stage. This phenomenon looks more obvious in supervised and transductive semi-supervised classification than inductive semi-supervised learning algorithms. (2) Thanks to the information proved by unlabeled data, the performance of semi-supervised classification algorithms exceeded SVM. However, both running in the supervised scenario, SDLSR achieves much better results than SVM. In fact, the performance of SDLSR is even over transductive semi-supervised classification approaches since it concerns the discriminative variables of each label for each samples. (3) Since our model takes the contribution importances of different training data into consideration, together with the introduced adjustment variables for each label, it consistently outperforms other methods over all datasets. ASL also considers the contribution importances but achieves the second best performance because it employs traditional LSR for classification.

We also conduct some numeric experiments to verity the time superiority of the proposed algorithm. With 20% labeled training data, we copy each datasets 2-times, 4-times, 6-times, 8-times and 10-times respectively and report the time performance over each extended datasets in Figure 2. Note that because ℓ_1-SEMI, ASL and our model are solved through alternative optimization algorithm, a equal convergence residual is shared for a fair comparison. We observe from the experimental result that: (1) The time cost of the graph-based methods, including GFHF, ℓ_1-SEMI and FME grow more rapidly than the proposed model and ASL with the increase of training data; (2) Our model consumes less time than other semi-supervised methods, especially when the number of data increases larger. This result gears to the computational complexity mentioned above and demonstrates the efficiency of the proposed model for large scale data applications.

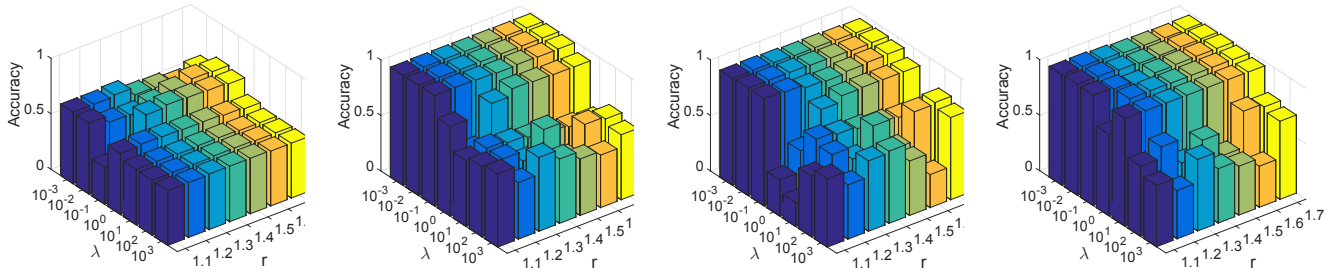
4.2 Sensitivity and Convergence

To illustrate the influence of parameters on the performance of semi-supervised classification, we report in Figure 3 the performance of the proposed model with varying values of parameters $\lambda \in [10^{-3}, 10^{3}]$ with step-size of 10 and $r \in [1.1, 1.7]$ with step-size of 0.1. We observe from the experi-



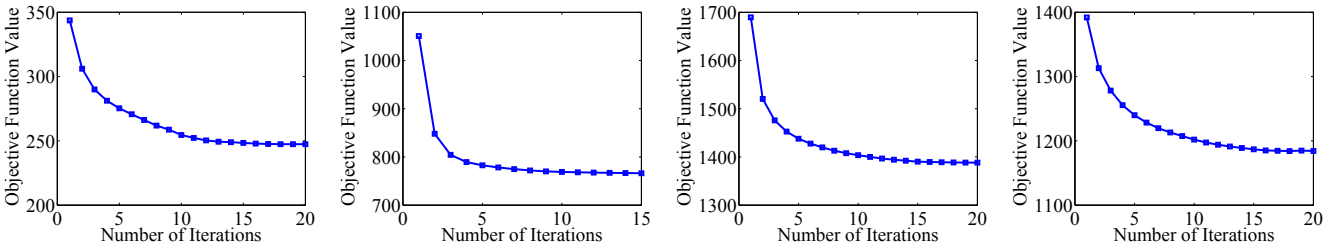
(a) ORL (b) YaleB (c) CMU-PIE (d) PALM

Figure 2: Time performance analysis of the competitors.



(a) ORL (b) YaleB (c) CMU-PIE (d) PALM

Figure 3: Sensitivity analysis on parameters λ and r.



(a) ORL (b) YaleB (c) CMU-PIE (d) PALM

Figure 4: Convergence curves of the objective values.

mental results that the performance varies with different settings on various datasets. As a result, how to identify the optimal values of the parameters is data dependent. Generally, the results could be satisfactory and relatively stable when λ and r fall in the range of $[10^{-2}, 10^{-1}]$ and $[1.5, 1.7]$, respectively. To illustrate the efficiency of our algorithm, we plot the corresponding convergence curves of the objective function (2) for two datasets in Figure 4, where the parameters λ and r are set as 0.1 and 1.4, respectively; 20% labeled data are involved in the training stage. It is evident that our algorithm converges within 20 iterations over all datasets, validating the efficiency and quickly converges of this algorithm.

5 Conclusion

In this paper, we develop a new discriminative LSR where each label is associated with an adjustment variable to eliminate the penalty of data that are classified correctly. Based on this discriminative LSR, we propose a novel semi-supervised classification which concerns the contribution importance of labeled and unlabeled training data simultaneously. An efficient alternative algorithm is exploited to solve the proposed challenging problems with theoretical analyses on its convergence and computational complexity. Experimental results, including the performance of accuracy and time cost, demonstrate the superiority and efficiency of the proposed model.

Acknowledgments

This work was funded by the National Science Foundation of China (No.61502377), the National Key Research and Development Program of China (No.2016YFB1000903), China Postdoctoral Science Foundation (No. 2015M582662), Ministry of Education Innovation Research Team (No.IRT13035) Project of China Knowledge Centre for Engineering Science and Technology and the Project of No.41418070102.

References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[Bishop, 2006] Christopher M Bishop. Pattern recognition. *Mach. Learn.*, 128:1–58, 2006.

[Cai *et al.*, 2007a] Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised discriminant analysis. In *ICCV*, 2007.

[Cai *et al.*, 2007b] Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, and Thomas Huang. Learning a spatially smooth subspace for face recognition. In *CVPR*, 2007.

[Chang and Yang, 2016] Xiaojun Chang and Yi Yang. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE Trans. Neural Netw. Learn. Syst.*, 2016.

[Chang *et al.*, 2014] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[Chapelle *et al.*, 2008] Olivier Chapelle, Vikas Sindhwani, and Sathiya S Keerthi. Optimization techniques for semi-supervised support vector machines. *J. Mach. Learn. Res.*, 9:203–233, 2008.

[Chapelle *et al.*, 2009] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *IEEE Trans. Neural Netw.*, 20(3):542–542, 2009.

[Fan *et al.*, 2005] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, 6:1889–1918, 2005.

[Georghiades *et al.*, 2001] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[Liu *et al.*, 2011] Xiaobai Liu, Xiaotong Yuan, Shuicheng Yan, and Hai Jin. Multi-class semi-supervised svms with positiveness exclusive regularization. In *ICCV*, 2011.

[Luo *et al.*, 2017] Minnan Luo, Xiaojun Chang, Liqiang Nie, Yi Yang, Alexander G Hauptmann, and Qinghua Zheng. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE Trans. Cybern.*, 2017.

[Nie *et al.*, 2010] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans. Image Process.*, 19(7):1921–1932, 2010.

[Nie *et al.*, 2011a] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Unsupervised and semi-supervised learning via ℓ_1-norm graph. In *ICCV*, 2011.

[Nie *et al.*, 2011b] Feiping Nie, Dong Xu, Xuelong Li, and Shiming Xiang. Semisupervised dimensionality reduction and classification through virtual label regression. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 41(3):675–685, 2011.

[Sim *et al.*, 2002] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *FG*, 2002.

[Sindhwani and Keerthi, 2006] Vikas Sindhwani and S Sathiya Keerthi. Large scale semi-supervised linear svms. In *ACM SIGIR*, 2006.

[Subramanya and Talukdar, 2014] Amarnag Subramanya and Partha Pratim Talukdar. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(4):1–125, 2014.

[Wang *et al.*, 2014] De Wang, Feiping Nie, and Heng Huang. Large-scale adaptive semi-supervised learning via unified inductive and transductive model. In *SIGKDD*, 2014.

[Xiang *et al.*, 2012] Shiming Xiang, Feiping Nie, Gaofeng Meng, Chunhong Pan, and Changshui Zhang. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(11):1738–1754, 2012.

[Yan *et al.*, 2007] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, 2007.

[Yang *et al.*, 2015] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vision*, 113(2):113–127, 2015.

[Zhou *et al.*, 2004] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *NIPS*, 2004.

[Zhu and Goldberg, 2009] Xiaojin Zhu and Andrew B Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.

[Zhu *et al.*, 2004] Xiaojin Zhu, Jaz Kandola, Zoubin Ghahramani, and John D Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *NIPS*, 2004.