Large-Scale Robust Semisupervised Classification

Lingling Zhang, Minnan Luo, Zhihui Li, Feiping Nie, Huaxiang Zhang, Jun Liu, and Qinghua Zheng

***Abstract*—Semisupervised learning aims to leverage both labeled and unlabeled data to improve performance, where most of them are graph-based methods. However, the graph-based semisupervised methods are not capable for large-scale data since the computational consumption on the construction of graph Laplacian matrix is huge. On the other hand, the substantial unlabeled data in training stage of semisupervised learning could cause large uncertainties and potential threats. Therefore, it is crucial to enhance the robustness of semisupervised classification. In this paper, a novel large-scale robust semisupervised learning method is proposed in the framework of capped $\ell_{2,p}$-norm. This strategy is superior not only in computational cost because it makes the graph Laplacian matrix unnecessary, but also in robustness to outliers since the capped $\ell_{2,p}$-norm used for loss measurement. An efficient optimization algorithm is exploited to solve the nonconvex and nonsmooth challenging problem. The complexity of the proposed algorithm is analyzed and discussed in theory detailedly. Finally, extensive experiments are conducted over six benchmark data sets to demonstrate the effectiveness and superiority of the proposed method.**

***Index Terms*—Classification, ridge regression, robustness, semisupervised learning.**

I. Introduction

MORE and more data emerges along with the development of Internet, where the labeled data is extremely deficient. For many practical applications, supervised learning requires some data annotation work beforehand, which is tedious and time-consuming [1]–[3]. The quality of data annotation also seriously affects the algorithm performance. Note that the acquisition of unlabeled data is relatively inexpensive. Therefore, how to effectively utilize the unlabeled data to mine the available information is crucial [4]–[6].

Semisupervised learning is a considerable learning paradigm, which learns from both labeled and unlabeled data to improve the performance [7]–[11]. Typically, semisupervised learning can be categorized into transductive learning and inductive learning according to the statistical learning theory. Transductive learning is to infer the correct labels of unlabeled samples by spreading the labels from labeled data to unlabeled data. The local and global consistency (LGC) [12] and Gaussian fields and harmonic functions (GFHF) [13] are typical transductive learning approaches. The key of LGC is to let every point iteratively propagate its label information to its neighbors until a stable global state is achieved. GFHF solves the same problem as LGC based on a Gaussian random field model. Zhou and Schölkopf [14] further investigated the transductive algorithm using random walks (RWs) and spectral graph theory. Note that LGC, GFHF, and RW all use the quadratic form of graph embedding. In [15], a novel transductive method unsupervised and semisupervised learning via ℓ_1-norm graph (L1-SEMI) is proposed to achieve robustness over the noisy samples via ℓ_1-norm of spectral embedding. The drawback of transductive learning is that it could not predict the labels of samples which are not involved in the procedure of training.

Instead, inductive learning can predict not only the labels of unlabeled data using for training but also the labels of testing samples by mapping sample representations to the corresponding labels [16], [17]. This is the reason why inductive learning is more attractive and practical than transductive learning. In the past decades, many inductive learning methods were developed for classification. For example, the methods Laplacian regularized least squares [18] and Laplacian support vector machines (SVMs) are rooted in a general framework for semisupervised learning called manifold regularization [19], [20]. Nie *et al.* [21] combined a flexible penalty term to cope better with the samples which reside on the nonlinear manifold, namely flexible manifold embedding (FME). In [22], a new semisupervised elastic embedding method (AEE) with a novel adaptive loss is proposed to achieve robustness to the outliers. Note that these methods are based on graph-based learning which provides an efficient approach for incorporating the relationship among labeled and unlabeled data [23], [24]. However, the construction of graph Laplacian matrix requires tremendous computational cost, particularly for large-scale data.

Manuscript received June 20, 2017; accepted December 29, 2017. Date of publication January 17, 2018; date of current version February 14, 2019. This work was supported in part by the "Fundamental Theory and Applications of Big Data With Knowledge Engineering" through the National Key Research and Development Program of China under Grant 2016YFB1000903, in part by the National Natural Science Foundation of China under Grant 61772322, in part by the National Science Foundation of China under Grant 61502377, Grant 61532004, Grant 61532015, Grant 61672418, and Grant 61672419, in part by the Project of China Knowledge Centre for Engineering Science and Technology, in part by the Ministry of Education Innovation Research Team under Grant IRT 17R86, and in part by the Innovative Research Group of the National Natural Science Foundation of China under Grant 61721002. This paper was recommended by Associate Editor Y. Jin. *(Corresponding authors: Feiping Nie; Huaxiang Zhang.)*

L. Zhang, M. Luo, J. Liu, and Q. Zheng are with the MOEKLINNS Laboratory, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zhanglingling@stu.xjtu.edu.cn; minnluo@mail.xjtu.edu.cn; liukeen@mail.xjtu.edu.cn; qhzheng@mail.xjtu.edu.cn).

Z. Li is with the Research and Development Department, Beijing Etrol Technologies Company, Ltd., Beijing 100095, China (e-mail: zhihuilics@gmail.com).

F. Nie is with the Department of Computer Science, Northwestern Polytechnical University, Xi'an 710049, China (e-mail: feipingnie@gmail.com).

H. Zhang is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China (e-mail: huaxzhang@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2789420

Although semisupervised learning could combine labeled and unlabeled data information to surpass the classification performance intuitively, substantial unlabeled data in training stage could cause large uncertainties and potential threats. Many outliers hiding in unlabeled data may distort the classifier learned only from labeled data. Therefore, guaranteeing the robustness to outliers is crucial for semisupervised classification performance. To address this issue, Wang *et al.* [25] performed the large-scale adaptive semisupervised learning (ASL) method based on ridge regression, where ridge regression [26] is generally used to solve the classification problem and achieve excellent performance in supervised learning. ASL adaptively suppresses the weights of boundary points to achieve better robustness to outliers. However, this method introduces an additional probability of label for each instance, which increase the model's computation.

Specially, Jiang *et al.* [27] proposed a novel dictionary learning method which uses capped ℓ_1-norm [28] as a robust loss function. This model achieves an excellent property of robustness to outliers and outperforms other dictionary learning methods. We propose a novel semisupervised classification method based on ridge regression in this paper. On the one hand, the method in this paper is superior in computational cost because it makes the similarity matrix unnecessary. On the other hand, inspired by the effectiveness of capped ℓ_1-norm, we leverage a novel loss function based on capped $\ell_{2,p}$-norm to enhance the model's robustness. Note that the capped $\ell_{2,p}$-norm is indeed an extensive version of capped ℓ_1-norm, i.e., the capped $\ell_{2,p}$-norm becomes traditional capped ℓ_1-norm if $p = 1$. Undoubtedly, the loss function based on capped $\ell_{2,p}$-norm could get better robustness than capped ℓ_1-norm by setting the appropriate value of p. Extensive numeric experiments demonstrate that our method significantly outperforms other semisupervised methods on six benchmark data sets, particularly in large-scale data applications. The contributions of this paper are summarized as follows.

1) A novel large-scale semisupervised classification method is presented on the basis of ridge regression, which does not require to construct the graph Laplacian matrix. Therefore, the proposed method is of significant in big data applications.
2) Instead of using the ℓ_2-norm or $\ell_{2,p}$-norm, we employ the capped $\ell_{2,p}$-norm-based loss function to achieve the model's robustness and accuracy for classification.
3) An efficient alternative algorithm is exploited to solve the proposed nonconvex and nonsmooth optimization problem. The computational complexity of the algorithm is analyzed detailedly in theory.

II. Robust Semisupervised Learning

In this section, we exploit a novel large-scale semisupervised learning method in the framework of ridge regression, which utilizes the capped $\ell_{2,p}$-norm to enhance the model's robustness to outliers. We denote the training set by $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where n is the total number of data points and $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^\top \in \mathbb{R}^d$ refers to a d-dimensional feature of the ith data point. In the framework of semisupervised

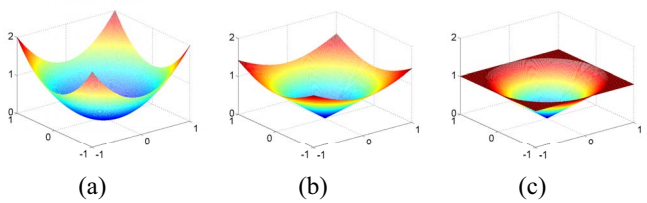


Fig. 1. Squared ℓ_2-norm loss versus $\ell_{2,p}$-norm loss versus capped $\ell_{2,p}$-norm loss. (a) Squared ℓ_2-norm loss. (b) $\ell_{2,p}$-norm loss ($p = 1$). (c) Capped $\ell_{2,p}$-norm loss ($p = 1,\ \varepsilon = 1$).

case, the feature matrix of training data points is represented by $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where the first m ($m < n$) data point $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ are labeled and the remain ones, i.e., $\mathbf{x}_{m+1}, \mathbf{x}_{m+2}, \ldots, \mathbf{x}_n$, are unlabeled data points whose labels are not given. Let $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ic}]^\top \in \mathbb{R}^c$ be the label vector of the ith labeled data point for $i = 1, 2, \ldots, m$, where c refers to the number of semantic categories and $y_{ij} = 1$ when $\mathbf{x}_i$ is in the jth class whereas $y_{ij} = 0$ otherwise. In addition, the label matrix

$$Y = \left[\underbrace{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m}_{Y_l \in \mathbb{R}^{m\times c}}, \underbrace{\mathbf{y}_{m+1}, \ldots, \mathbf{y}_n}_{Y_u \in \mathbb{R}^{(n-m)\times c}}\right]^\top \in \text{Ind} \tag{1}$$

means $Y \in \mathbb{R}^{n\times c}$ is an indicator matrix used for label assignment over the entire training data points. In particular, $Y_l \in$ Ind is given for m labeled data points; $Y_u \in$ Ind refers to the predicted label matrix for $(n - m)$ unlabeled data points.

Squared ℓ_2-norm $\|\cdot\|_2^2$ is usually leveraged in ridge regression to compute the loss over entire data points. It is simple but not robust to the large losses which are typically introduced by outliers, especially when the class number of the data points is large [22]. However, substantial unlabeled data during semisupervised learning stage could cause large uncertainties and potential threats. It leads the learned classifier maybe distorted far way from the ground truth if these outlier data points dominated the loss function [25]. Therefore, it is necessary to develop a method that is robust to the outliers.

Instead of using ℓ_2-norm-based loss measurement, $\ell_{2,p}$-norm $\|\cdot\|_2^p$ is typically exploited to achieve the joint sparsity by making p falls inside the range of $(0, 2]$. This strategy also enhances the model's robustness to outliers [29]–[31]. For a better representation, we illustrate the squared ℓ_2-norm and $\ell_{2,p}$-norm in Fig. 1(a) and (b). It is evident that $\ell_{2,p}$-norm is much more capable to tolerate the bias caused by outliers than traditional squared ℓ_2-norm. As a result, $\ell_{2,p}$-norm achieves more robustness by selecting $0 < p \leq 2$.

To further improve the robustness, we first define a novel norm of a matrix as follows, namely capped $\ell_{2,p}$-norm.

Definition 1: For any matrix $A = (a_{ij}) \in \mathbb{R}^{n\times m}$ and parameter $0 < p \leq 2$, its capped $\ell_{2,p}$-norm is defined as

$$g(A) = \sum_{i=1}^{n} \min\left(\|A^i\|_2^p, \varepsilon\right) \tag{2}$$

with thresholding parameter $\varepsilon \geq 0$.

As illustrated in Fig. 1(c), the loss value of capped $\ell_{2,p}$-norm, i.e., $\min(\|\mathbf{y}_i' - \mathbf{y}_i\|_2^p, \varepsilon)$ does not increase anymore when

$\|\mathbf{y}_i' - \mathbf{y}_i\|_2^p$ is large than ε. Therefore, the capped $\ell_{2,p}$-norm can further suppress the bias when the outliers are far away from the normal data distribution. Taking the robustness of capped $\ell_{2,p}$-norm into consideration, we propose a novel robust semisupervised classification model formalized as the following optimization problem:

$$\min_{W,\mathbf{b},Y_u\in\mathrm{Ind}} \sum_{i=1}^{n} \min\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\|_2^p, \varepsilon\Big) + \lambda\|W\|_F^2 \quad (3)$$

where $W = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c] \in \mathbb{R}^{d\times c}$ is the classifier matrix whose column $\mathbf{w}_i$ refers to the classifier with respect to the ith integrate. c-dimensional vector $\mathbf{b}$ refers to the bias term. Note that W and $\mathbf{b}$ correlate each data point $\mathbf{x}_i$ with its labels $\mathbf{y}_i$ for $i = 1, 2, \ldots, n$. Specifically, the first term $\sum_{i=1}^{n} \min(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\|_2^p, \varepsilon)$ in objective function (3) describes the loss value based on the capped $l_{2,p}$-norm. The second term $\lambda\|W\|_F^2$ is included to penalize nonzero weights with the cost proportional to a shrinkage coefficient λ. It is evident that the first term turns to $\ell_{2,p}$-norm-based loss measurement when the parameter ε is set as ∞. Indeed, the proposed problem (3) is a more general extension of ridge regression. It is equivalent to the traditional ridge regression when parameters p and ε are set as 2 and ∞, respectively.

III. Optimization Algorithm

Note that the proposed objective function is indeed the sum of a concave function and a convex function. This nonconvex and nonsmooth optimization problem (3) with respect to variables W, $\mathbf{b}$ and Y_u is challenging for optimization. In order to address this problem, we first introduce the following theorem on concave duality in [32].

Theorem 1 [32]: Let $g : \mathbb{R}^d \to \mathbb{R}$ be a nonconvex function and $h : \mathbb{R}^d \to \Omega \subset \mathbb{R}^d$ be a map with range Ω. If there exists a concave function $\bar{g}(\mathbf{u})$ defined on Ω such that $g(\mathbf{z}) = \bar{g}(h(\mathbf{z}))$ holds, the nonconvex function g can be rewritten according to concave duality [33] as

$$g(\mathbf{z}) = \inf_{\mathbf{v}\in\mathbb{R}^d}\Big\{\mathbf{v}^\top h(\mathbf{z}) - g^*(\mathbf{v})\Big\} \quad (4)$$

where $g^*(\mathbf{v})$ is the concave dual of $\bar{g}(\mathbf{u})$ defined as

$$g^*(\mathbf{v}) = \inf_{\mathbf{u}\in\Omega}\Big\{\mathbf{v}^\top \mathbf{u} - \bar{g}(\mathbf{u})\Big\}. \quad (5)$$

Specifically, the minimum of the right-hand side of (4), denoted by $\mathbf{v}^*$, is achieved at

$$\mathbf{v}^* = \frac{\partial \bar{g}(\mathbf{u})}{\partial \mathbf{u}}|_{\mathbf{u}=h(\mathbf{z})}. \quad (6)$$

According to Theorem 1, we define a concave function $\bar{g} : \mathbb{R} \to \mathbb{R}$ such that $\forall z > 0$

$$\bar{g}(z) = \min\Big(\sqrt{z^p}, \varepsilon\Big). \quad (7)$$

Let $h(\omega) = \omega^2$, then the capped l_2, l_p-norm loss measurement could be reformulated as

$$\min\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^p, \varepsilon\Big) = \bar{g}(h(\omega)) \quad (8)$$

where $\omega = \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2$. As a result, we obtain the following theorem.

Theorem 2: The objective function of optimization problem (3) can be rewritten as

$$\min_{W,\mathbf{b},Y_u,v_{ii}\geq 0(\forall i)} \sum_{i=1}^{n} L_i(v_{ii}, W, \mathbf{b}; \varepsilon) + \lambda\|W\|_F^2 \quad (9)$$

where

$$L_i(W, \mathbf{b}, Y_u, v_{ii}; \varepsilon) = \begin{cases} v_{ii}z - v_{ii}\Big(\frac{2}{p}v_{ii}\Big)^{\frac{2}{p-2}} + \Big(\frac{2}{p}v_{ii}\Big)^{\frac{p}{p-2}}, & z^{\frac{p}{2}} < \varepsilon \\ v_{ii}z - v_{ii}\varepsilon^{\frac{2}{p}} + \varepsilon, & z^{\frac{p}{2}} \geq \varepsilon \end{cases}$$

and $z = \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2$.

Proof: According to Theorem 1, the $\ell_{2,p}$-norm loss measurement in objective function (3) can be reformulated as

$$\begin{aligned}\min\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^p, \varepsilon\Big) &= \bar{g}\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2\Big) \\ &= \inf_{v_{ii}\geq 0} v_{ii}h(\omega) - g^*(v_{ii}) = \inf_{v_{ii}\geq 0} v_{ii}z - g^*(v_{ii}) \end{aligned} \quad (10)$$

where $z = h(\omega) = \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2$. The concave dual of $\bar{g}(z)$, denoted by $g^*(v_{ii})$ is defined as

$$\begin{aligned} g^*(v_{ii}) &= \inf_z \big[v_{ii}z - \bar{g}(z)\big] \\ &= \inf_z \begin{cases} v_{ii}z - \sqrt{z^p}, & z^{\frac{p}{2}} < \varepsilon \\ v_{ii}z - \varepsilon, & z^{\frac{p}{2}} \geq \varepsilon. \end{cases} \end{aligned} \quad (11)$$

By optimizing z for problem (11), it turns that

$$g^*(v_{ii}) = \begin{cases} v_{ii}\Big(\frac{2}{p}v_{ii}\Big)^{\frac{2}{p-2}} - \Big(\frac{2}{p}v_{ii}\Big)^{\frac{p}{p-2}}, & z^{\frac{p}{2}} < \varepsilon \\ v_{ii}\varepsilon^{\frac{2}{p}} - \varepsilon, & z^{\frac{p}{2}} \geq \varepsilon. \end{cases} \quad (12)$$

As a result, the $\ell_{2,p}$-norm loss measurement in objective function (3) could be expressed as

$$\min\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^p, \varepsilon\Big) = \inf_{v_{ii}\geq 0} L_i(W, \mathbf{b}, Y_u, v_{ii}; \varepsilon) \quad (13)$$

with

$$\begin{aligned} &L_i(W, \mathbf{b}, Y_u, v_{ii}; \varepsilon) \\ &\quad = \begin{cases} v_{ii}z - v_{ii}\Big(\frac{2}{p}v_{ii}\Big)^{\frac{2}{p-2}} + \Big(\frac{2}{p}v_{ii}\Big)^{\frac{p}{p-2}}, & z^{\frac{p}{2}} < \varepsilon \\ v_{ii}z - v_{ii}\varepsilon^{\frac{2}{p}} + \varepsilon, & z^{\frac{p}{2}} \geq \varepsilon. \end{cases} \end{aligned} \quad (14)$$

Therefore, the objective function (3) can be reformulated as

$$\begin{aligned} &\min_{W,\mathbf{b},Y_u} \sum_{i=1}^{n} \min\Big(\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^p, \varepsilon\Big) + \lambda\|W\|_F^2 \\ &\iff \min_{W,\mathbf{b},Y_u} \sum_{i=1}^{n} \inf_{v_{ii}\geq 0} L_i(W, \mathbf{b}, Y_u, v_{ii}; \varepsilon) + \lambda\|W\|_F^2 \\ &\iff \min_{W,\mathbf{b},Y_u,v_{ii}\geq 0,\forall i} \sum_{i=1}^{n} L_i(W, \mathbf{b}, Y_u, v_{ii}; \varepsilon) + \lambda\|W\|_F^2. \end{aligned} \quad (15)$$

The proof is completed. ■

Based on Theorem 2, the optimal classier related parameters W and $\mathbf{b}$, and the predicted label matrix for unlabeled

data points Y_u can be learned via addressing optimization problem (9) through an alternative optimization algorithm.

A. Optimizing Y_u

When the classifier related parameters W and $\mathbf{b}$ and weighting parameter v_{ii} are fixed for $i = 1, 2, \ldots, n$, the label matrix Y_u of unlabeled data points can be predicted via solving the following optimization problem:

$$Y_u = \arg\min_{Y_u \in \text{Ind}} \|X_u^\top W + \mathbf{1}\mathbf{b}^\top - Y_u\|_F^2 \tag{16}$$

where $\mathbf{1}$ denotes the vector whose elements are all one. Since the rows of indicator matrix Y_u are independent from each other, we update one row each time while keeping the other rows fixed. To be more specific, the updating of one row is conducted by finding the element being 1 that results in the minimum of the optimization problem

$$\min_{\mathbf{y}_i \in \text{Ind}} \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\|_2^2 \tag{17}$$

for $i = m+1, m+2, \ldots, n$. Let $\mathbf{y}_i' = W^\top \mathbf{x}_i + \mathbf{b}$, we have

$$\begin{aligned}\|\mathbf{y}_i' - \mathbf{y}_i\|_2^2 &= \sum_{j=1}^{c} \left(\mathbf{y}'_{ij} - \mathbf{y}_{ij}\right)^2 \\ &= \sum_{j=1}^{c} \mathbf{y}'^2_{ij} - 2\sum_{j=1}^{c} \mathbf{y}'_{ij}\mathbf{y}_{ij} + \sum_{j=1}^{c} \mathbf{y}_{ij}^2.\end{aligned}$$

Because only one element of $\mathbf{y}_i \in \text{Ind}$ is 1, we arrive at $\sum_{j=1}^{c} \mathbf{y}_{ij}^2 = 1$. Therefore the value of function (18) only lies on $\sum_{j=1}^{c} \mathbf{y}'_{ij}\mathbf{y}_{ij}$. It is evident that the optimization problem (17) arrives at its minimum value when $\mathbf{x}_i$ is in the k_0th class where $k_0 = \arg\max_j \ \mathbf{y}_{ij}$.

B. Optimizing v_{ii}

We formulate the gradient of function $\bar{g}(z)$ with respect to z as

$$\frac{\partial \bar{g}(z)}{\partial z} = \begin{cases} \frac{p}{2} z^{\frac{p}{2}-1}, & \text{if } 0 < z < \varepsilon^{\frac{2}{p}} \\ 0, & z^{\frac{p}{2}} \geq \varepsilon. \end{cases} \tag{18}$$

According to (6) in Theorem 1, when $z = h(\omega) = \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2$ with fixed values of W, $\mathbf{b}$, and Y_u, we have

$$\begin{aligned} v_{ii} &= \frac{\partial \bar{g}(z)}{\partial z}|_{z=\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2} \\ &= \begin{cases} \frac{p}{2}\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^{p-2}, & \text{if } 0 < \|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^p < \varepsilon \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{19}$$

for $i = 1, 2, \ldots, n$. Note that parameter v_{ii} $(i = 1, 2, \ldots, n)$ indeed provides a weight for the ith instance among both labeled and unlabeled data points. Specifically, if the loss brought by ith data point is larger than the predefined threshold, we remove this data point and avoid it.

C. Optimizing W and $\mathbf{b}$

When the variables v_{ii} $(i = 1, 2, \ldots, n)$ are fixed, the optimization problem (9) for variables W and $\mathbf{b}$ becomes

$$\{W, \mathbf{b}\} = \arg\min \sum_{i=1}^{n} v_{ii}\|W^\top \mathbf{x}_i + \mathbf{b} - \mathbf{y}^i\|_2^2 + \lambda\|W\|_F^2. \tag{20}$$

Let $V = \text{diag}(v_{11}, v_{22}, \ldots, v_{nn})$ be a $n \times n$ diagonal matrix, then the optimization problem (20) can be reformulated as

$$\begin{aligned}\phi(W, \mathbf{b}) = \text{Tr}\Big(\Big(X^\top W + \mathbf{1}\mathbf{b}^\top - Y\Big)^\top V\Big(X^\top W + \mathbf{1}\mathbf{b}^\top - Y\Big)\Big) \\ + \lambda\|W\|_F^2\end{aligned} \tag{21}$$

where $\text{Tr}(\cdot)$ stands for the trace of matrix. With the equivalent objective $\phi(W, \mathbf{b})$ for optimization problem (20), the optimal solution can be solved easily since ϕ is jointly convex with respect to W and $\mathbf{b}$. Setting the derivative of objective function ϕ with respect to $\mathbf{b}$ to 0, we have

$$\frac{\partial \phi(W, \mathbf{b})}{\partial \mathbf{b}} = 2\mathbf{1}^\top V\Big(X^\top W + \mathbf{1}\mathbf{b}^\top - Y\Big) = 0. \tag{22}$$

Then, we obtain

$$\mathbf{b} = \frac{1}{\mathbf{1}^\top V \mathbf{1}} Y^\top V\mathbf{1} - \frac{1}{\mathbf{1}^\top V\mathbf{1}} W^\top XV\mathbf{1}. \tag{23}$$

After that, we set the derivative of ϕ with respect to W to 0, we arrive at

$$\frac{\partial \phi(W, \mathbf{b})}{\partial W} = 2XV\Big(X^\top W + \mathbf{1}\mathbf{b}^\top - Y\Big) + 2\lambda W = 0. \tag{24}$$

By substituting (23) into (24), we get

$$XVHX^\top W + \lambda W = XVHY \tag{25}$$

where $H = I - [(\mathbf{1}\mathbf{1}^\top V)/(\mathbf{1}^\top V\mathbf{1})] \in \mathbb{R}^{n\times n}$ denotes a centering matrix. Let $A = XVHX^\top + \lambda I \in \mathbb{R}^{d\times d}$ and $B = XVHY \in \mathbb{R}^{d\times c}$, we obtain

$$W = A^{-1}B. \tag{26}$$

There, the optimization problem (20) for minimal estimation of W and $\mathbf{b}$ can be addressed via updating $\mathbf{b}$ and W alternatively according to (23) and (26).

In a summary, we describe the alternative optimization algorithm in Algorithm 1 to solve the proposed challenging semisupervised classification method. We specially initialize the classifier related parameters $W^0, \mathbf{b}^0$ with all of the labeled data by addressing the following optimization problem:

$$(W_0, \mathbf{b}_0) = \arg\min_{W,\mathbf{b}} \|X_l^\top W + \mathbf{1}\mathbf{b}^\top - Y_l\|_F^2 + \lambda\|W\|_F^2. \tag{27}$$

It is evident that the objective function (27) is jointly convex with respect to variables W and $\mathbf{b}$. We set the derivatives of the function (27) equal to zero with respect to $\mathbf{b}$ and obtain

$$\mathbf{b}_0 = \frac{1}{l}\Big(Y_l^\top \mathbf{1} - W^\top X_l\mathbf{1}\Big). \tag{28}$$

By substituting (28) into (27) and setting its derivative with respect to W equal to zero, we arrive at

$$W_0 = \Big(X_l HX^\top + \lambda I\Big)^{-1} XHY_l \tag{29}$$

Algorithm 1 Large-Scale Robust Semisupervised Classification

Input: Labeled data matrix X_l with label matrix Y_l, unlabeled data matrix X_u, number of classes c, regularization parameter λ and γ.

Output: Predicted label Y_u.

Initialize: $\{W_0, \mathbf{b}_0\} = \arg\min_{W,\mathbf{b}} \|X_l^\top W + \mathbf{1}\mathbf{b}^\top - Y_l\|_F^2 + \lambda\|W\|_F^2$ and set initial $t = 0$.

1: **while** not converge **do**

2: Update each row of Y_u by solving $\mathbf{y}_i^{t+1} = \arg\min_{\mathbf{y}\in Ind} \|W^\top \mathbf{x}_i + \mathbf{b}_t - \mathbf{y}\|_2^2$ $(i = l+1, l+2, \cdots, n)$ while fixing the remaining rows;

3: Update the diagonal matrix $V^{t+1} = diag(v_{11}^{t+1}, v_{22}^{t+1}, \cdots, v_{nn}^{t+1})$ with

$$v_{ii}^{t+1} = \begin{cases} \frac{p}{2}\|W_t^\top \mathbf{x}_i + \mathbf{b}_t - \mathbf{y}^i\|_2^{p-2}, & \|W_t^\top \mathbf{x}_i + \mathbf{b}_t - \mathbf{y}^i\|_2^p \le \varepsilon; \\ 0, & \text{otherwise.} \end{cases} \quad (i = 1, 2, \cdots, n)$$

4: Update $\{W_{t+1}, \mathbf{b}_{t+1}\} = \arg\min_{W,\mathbf{b}} Tr\big((X^\top W + \mathbf{1}\mathbf{b}^\top - Y^{t+1})^\top V^{t+1}(X^\top W + \mathbf{1}\mathbf{b}^\top - Y^{t+1})\big) + \lambda\|W\|_F^2$.

5: **end while**

where $H = I - (1/l)\mathbf{1}\mathbf{1}^\top$. As a result, closed-form initial solutions W and $\mathbf{b}$ are gained over the labeled data points. Because only a small number of data points are usually labeled in the framework of semisupervised learning, the procedure of initialization will not cost too much computation.

D. Discussion

We briefly discuss the computational complexity of the proposed algorithm. During the training, the major computational cost lies in the updating of classifier related parameters W, i.e., $W = A^{-1}B$ with the most computational cost $\mathcal{O}(d^3)$, where the $d \times d$ matrix C and $d \times c$ matrix B is calculated with computational cost $\mathcal{O}(nd^2)$ and $\mathcal{O}(ndc)$, respectively. Indeed, the $d \times d$ matrix inverse computation can be avoided by updating W via solving the following problem:

$$W_{t+1} = \arg\min_W A^\top \|W\|_F^2 - 2B^\top W. \tag{30}$$

In this case, W can be updated iteratively using gradient descent method as $W_{t+1} = W_t - \alpha(AW_t - B)$ with a computational cost of $\mathcal{O}(Td^2c)$, where T refers to the number of iterations. As a result, the total computational cost of the proposed algorithm is upper bounded by $\mathcal{O}(Td^2c) + \mathcal{O}(nd^2) + \mathcal{O}(ndc)$. Because in practice the number of categories c is always much smaller than the instance dimensionality d and the number of data points n, the total computational cost can also written as $\mathcal{O}(Td^2) + \mathcal{O}(nd^2)$. It indicates that the computational cost of our algorithm is linear with respect to the number of data points n, and therefore our algorithm is able to scale to large-scale data. The scalability maybe restricted when the number of features d is so large. In this case, we can decrease the dimension d using some dimension reduction methods.

Compared with graph-based methods, our proposed algorithm need not consume additional time on establishing the Laplacian matrix with computational cost $\mathcal{O}(n^2)$ or $\mathcal{O}(nk)$, where k is the sparsity of graph Laplacian matrix.

IV. Experiment

A. Experimental Setup

We perform the experiments over six data sets. The JAFFE [34] contains 213 facial expression images from ten classes. The USPS [35] includes 9298 handwritten digit images, which has ten categories representing digits from 0 to 9. The MSRA50 [36] contains 1799 face images from 12 different classes. The YaleB [37] consists of 2414 face images from 38 individuals captured under the various lighting conditions. The face database CMU-PIE [38] is collected from 68 individuals with varying expressions, which includes 3329 images. The PALM [39] contains 2000 hand images of 100 individuals. For all data sets, we directly use pixel values to represent the images. Thus the image feature dimensions are 576, 256, 1024, 1024, 1024, and 576, respectively in these six data sets.

We compare our method with the following six classification algorithms, where the adaptive boosting (AdaBoost) and SVM are the common valid supervised classifiers and the others are designed for semisupervised learning. Note that the semisupervised classifiers contain the transductive learning methods (GFHF and L1-SEMI) and the inductive learning methods (FME and ASL).

1) *AdaBoost [40]:* It improves the classification performance by constructing a "strong" classifier as the linear combination of "weak" classifiers.
2) *SVM [41]:* It is a supervised model for classification and regression, which constructs a hyperplane in a high-dimensional space to achieve a good separation of data samples.
3) *GFHF [13]:* It is a representative graph-based semisupervised learning method. The method formulates the learning problem in terms of Gaussian random field on Laplacian graph.
4) *L1-SEMI [15]:* Different from minimizing a ℓ_2-norm in traditional graph-based learning methods, L1-SEMI exploits sparsity by minimizing the ℓ_1-norm to improve robustness.
5) *FME [21]:* It is a unified manifold embedding framework for semisupervised learning. It combines a flexible penalty term to cope better with the samples which reside on the nonlinear manifold.
6) *ASL [25]:* It achieves robustness by suppressing the weights of boundary points adaptively. ASL is scalable to large-scale data because it need not establish the graph Laplacian matrix.

TABLE I
PERFORMANCE COMPARISON ON JAFFE (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.406±0.025	0.400±0.033	0.484±0.026	0.489±0.048	0.675±0.018	0.661±0.051	0.794±0.032	0.818±0.014
SVM	0.790±0.015	0.700±0.012	0.733±0.009	0.727±0.011	0.817±0.015	0.756±0.013	0.911±0.011	0.818±0.016
GFHF	0.829±0.012	NA	0.945±0.013	NA	**0.988±0.014**	NA	0.999±0.011	NA
L1-SEMI	0.885±0.013	NA	0.947±0.014	NA	0.972±0.022	NA	0.999±0.025	NA
FME	0.650±0.019	0.686±0.014	0.700±0.016	0.736±0.035	0.810±0.015	0.790±0.017	0.837±0.014	0.862±0.015
ASL	0.876±0.010	0.907±0.011	0.947±0.015	0.956±0.018	0.946±0.012	0.957±0.016	0.989±0.015	0.992±0.005
Our Model	**0.887±0.015**	**0.932±0.015**	**0.956±0.015**	**0.977±0.015**	0.978±0.015	**0.985±0.015**	**1.000±0.001**	**1.000±0.002**

TABLE II
PERFORMANCE COMPARISON ON USPS (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.554±0.014	0.534±0.025	0.615±0.035	0.588±0.034	0.647±0.023	0.634±0.013	0.687±0.012	0.672±0.021
SVM	0.673±0.013	0.663±0.005	0.750±0.025	0.741±0.008	0.823±0.021	0.810±0.018	0.892±0.013	0.881±0.019
GFHF	0.564±0.014	NA	0.671±0.018	NA	0.742±0.008	NA	0.837±0.025	NA
L1-SEMI	**0.796±0.011**	NA	0.839±0.019	NA	**0.897±0.052**	NA	**0.923±0.034**	NA
FME	0.743±0.018	0.732±0.013	0.823±0.012	0.837±0.010	0.864±0.045	0.870±0.012	0.884±0.033	0.879±0.013
ASL	0.769±0.015	0.752±0.011	0.791±0.011	0.803±0.025	0.832±0.015	0.856±0.009	0.868±0.009	0.875±0.021
Our Model	0.785±0.015	**0.762±0.015**	**0.843±0.015**	**0.852±0.015**	0.875±0.015	**0.877±0.015**	0.901±0.015	**0.890±0.015**

TABLE III
PERFORMANCE COMPARISON ON MSRA50 (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.334±0.016	0.296±0.025	0.386±0.032	0.365±0.024	0.532±0.026	0.495±0.011	0.840±0.014	0.808±0.028
SVM	0.524±0.023	0.532±0.012	0.583±0.021	0.560±0.018	0.646±0.013	0.690±0.013	0.871±0.011	0.863±0.012
GFHF	0.416±0.015	NA	0.506±0.013	NA	0.556±0.016	NA	0.901±0.027	NA
L1-SEMI	0.356±0.012	NA	0.474±0.015	NA	0.527±0.023	NA	0.956±0.012	NA
FME	0.653±0.011	0.583±0.017	0.842±0.019	0.744±0.013	0.865±0.022	0.803±0.014	0.912±0.016	0.923±0.016
ASL	0.853±0.017	0.801±0.011	0.885±0.010	0.875±0.015	0.910±0.014	0.900±0.011	0.956±0.017	0.988±0.003
Our Model	**0.886±0.013**	**0.836±0.015**	**0.902±0.009**	**0.892±0.012**	**0.941±0.012**	**0.953±0.010**	**1.000±0.001**	**0.992±0.006**

B. Performance Comparison

In the experiments, we randomly choose $2/3$ of data over each data set as the training samples. The remaining samples are served as the corresponding testing data. To study the performance of proposed method over different radios labeled data among the training samples, we assign the value of radio as 10%, 20%, 40%, and 80%, respectively.

For a fair comparison, we record the classification results of all methods using the best tuned. The parameters of our method, including the thresholding parameter ε, regularization parameter λ and p ($0 < p \leq 2$), are tuned in [0.6:0.2:1.4], $\{10^{-5}, 10^{-4}, \ldots, 10^{-1}\}$ and [0.1:0.1:2], respectively. The transductive semisupervised methods GFHF and L1-SEMI are parameter-free. The inductive method FME has two regularization parameters, which are tuned in $\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}$. For ASL, it has only one adaptive parameter, which is tuned from 1 to 2 with a stepsize of 0.1. Under the condition mentioned above, we repeat every method 30 times to compute the average classification accuracy and standard deviation (STD) to evaluate the performance of methods.

We report the experimental results over six data sets from Tables I–VI, where the column "unlabeled" shows the classification results over unlabeled samples among the sampled training data. Similarly, the column "testing" shows the results over testing samples. For transductive semisupervised methods, the learned classifiers cannot be used to predict the out-of-sample testing data. We use the symbol "NA" to represent the meaning. By comparing the performance of different methods, we have the following observations.

1) Our method consistently performs better over all data sets than other semisupervised and the supervised methods. The reason is that the capped $\ell_{2,p}$-norm used for

TABLE IV
PERFORMANCE COMPARISON ON YALEB (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.190±0.012	0.213±0.017	0.241±0.032	0.238±0.028	0.312±0.022	0.335±0.032	0.387±0.015	0.401±0.023
SVM	0.535±0.011	0.521±0.018	0.692±0.023	0.689±0.016	0.792±0.013	0.787±0.019	0.923±0.015	0.923±0.021
GFHF	0.315±0.013	NA	0.428±0.033	NA	0.564±0.011	NA	0.703±0.015	NA
L1-SEMI	0.670±0.013	NA	0.799±0.017	NA	0.900±0.014	NA	0.913±0.015	NA
FME	0.584±0.025	0.615±0.022	0.662±0.013	0.676±0.018	0.752±0.017	0.748±0.018	0.860±0.015	0.853±0.008
ASL	0.944±0.011	0.963±0.005	0.986±0.014	0.985±0.013	0.992±0.012	0.993±0.013	0.998±0.015	0.995±0.009
Our Model	**0.964±0.015**	**0.967±0.015**	**0.996±0.015**	**0.997±0.015**	**0.998±0.001**	**0.994±0.004**	**1.000±0.002**	**0.998±0.006**

TABLE V
PERFORMANCE COMPARISON ON CMU-PIE (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.169±0.024	0.187±0.023	0.256±0.018	0.273±0.023	0.342±0.032	0.331±0.031	0.402±0.016	0.432±0.021
SVM	0.462±0.034	0.435±0.025	0.646±0.019	0.610±0.018	0.779±0.011	0.763±0.021	0.796±0.015	0.784±0.011
GFHF	0.225±0.032	NA	0.388±0.013	NA	0.466±0.025	NA	0.554±0.012	NA
L1-SEMI	0.415±0.023	NA	0.601±0.011	NA	0.780±0.023	NA	0.915±0.013	NA
FME	0.442±0.013	0.434±0.013	0.608±0.010	0.578±0.014	0.760±0.013	0.777±0.018	0.900±0.018	0.908±0.016
ASL	0.871±0.011	0.866±0.014	0.927±0.013	0.937±0.015	0.964±0.014	0.954±0.014	0.970±0.014	0.960±0.014
Our Model	**0.891±0.010**	**0.900±0.011**	**0.945±0.010**	**0.943±0.016**	**0.968±0.015**	**0.964±0.019**	**0.971±0.015**	**0.968±0.015**

TABLE VI
PERFORMANCE COMPARISON ON PALM (ACCURACY ± STD) WITH RESPECT TO 10%, 20%, 40%, AND 80% LABELED TRAINING DATA

	10% Labeled		20% Labeled		40% Labeled		80% Labeled	
	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing	Unlabeled	Testing
Adaboost	0.218±0.014	0.234±0.019	0.327±0.026	0.336±0.026	0.450±0.024	0.441±0.032	0.493±0.014	0.526±0.015
SVM	0.704±0.014	0.720±0.012	0.775±0.010	0.820±0.011	0.858±0.013	0.870±0.015	0.957±0.019	0.958±0.016
GFHF	0.879±0.017	NA	0.925±0.013	NA	0.946±0.015	NA	0.986±0.002	NA
L1-SEMI	0.915±0.012	NA	0.920±0.011	NA	0.932±0.016	NA	0.993±0.001	NA
FME	0.562±0.011	0.545±0.016	0.744±0.014	0.708±0.018	0.880±0.012	0.840±0.013	0.976±0.013	0.975±0.014
ASL	0.755±0.013	0.762±0.014	0.832±0.015	0.850±0.019	0.923±0.010	0.907±0.018	0.983±0.010	0.986±0.007
Our Model	**0.947±0.015**	**0.926±0.012**	**0.958±0.017**	**0.958±0.014**	**0.976±0.019**	**0.963±0.015**	**0.987±0.008**	**0.993±0.005**

loss measurement is capable to effectively improve the robustness by lowering the influence of outliers.

2) In generally, semisupervised methods perform better than supervised methods, especially when the labeled samples among training points are fewer. This is because semisupervised methods can effectively utilize the unlabeled samples information during the training stage.
3) In most case, the inductive semisupervised methods outperform the transductive methods. In particular, the ASL and our model dramatically perform better than transductive approaches because they suppress the weights of boundary points to improve the robustness.
4) The supervised method SVM significantly performs better than AdaBoost, particularly over the data sets YaleB, CMU-PIE, and PALM. It demonstrates that the SVM has better stability than AdaBoost over the data sets with high-dimensional descriptors and large categories.

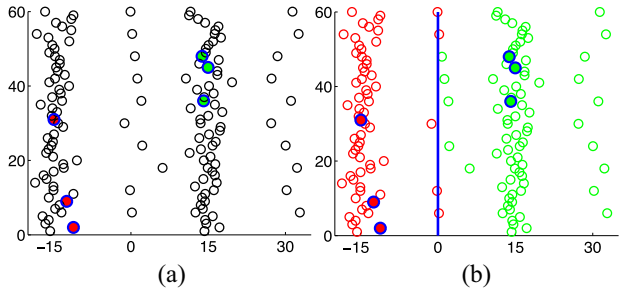


Fig. 2. Demonstration on synthetic data. (a) Original synthetic data. (b) Synthetic data after classification.

5) The performance of each method becomes better as the ratio of labeled training samples increasing. More labeled data are helpful to improve the performance of both supervised and semisupervised learning.

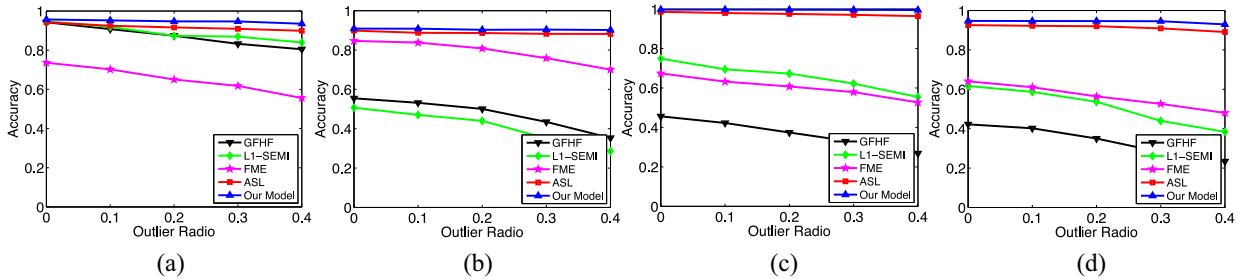


Fig. 3. Classification accuracies over 20% labeled training data for four face recognition data sets. (a) JAFFE. (b) MSRA50. (c) YaleB. (d) CMU-PIE.

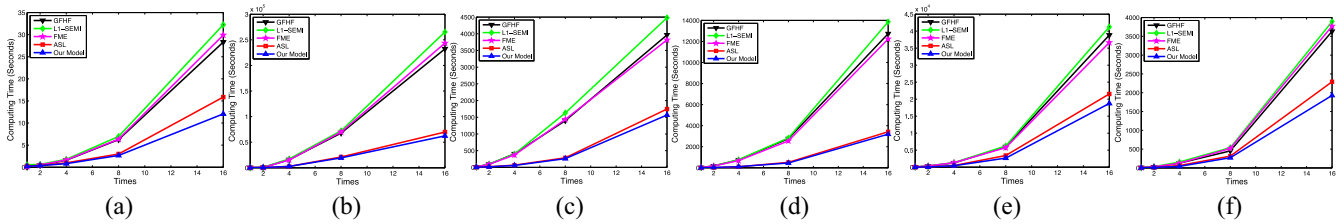


Fig. 4. Time performance analysis over 20% labeled training data for all data sets with different times. (a) JAFFE. (b) USPS. (c) MSRA50. (d) YaleB. (e) CMU-PIE. (f) PALM.

C. Robustness Analysis

To study the property of robustness to outliers, we carried out experiments over a synthetic data set [25] and four real face data sets with artificial block occlusion [27].

1) Demonstration on Synthetic Data With Boundary Outliers: In the experiment, original data is generated as Fig. 2(a). Specifically, we sample 60 data points for first class from a Gaussian distribution with a mean of -15, and an STD 4. Similarly, 60 data points for second class are sampled with a mean of 15, and an STD 4. We sample ten outliers around decision boundary with a mean of 0. In addition, we add ten outliers far away the decision boundary with a mean of 30. Three points for each class are sampled randomly as labeled data (i.e., the blue circle boxes in the left and right). Fig. 2(b) shows the classification result using our robust semisupervised model, where the blue line is the learned decision boundary. It is evident that our model learns the correct classifier in spite of the disturbance from some outliers far away from the normal data distribution.

2) Demonstration on Real Data With Block Occlusion: We provide some experimental results on classification task with synthetic occlusion over four face data sets including JAFFE, MSRA50, YaleB, and CMU-PIE. We set the labeled radio as 20% and then we add block occlusion into training set, where the side of block occlusion is $1/3$ original image size. We set the percentage of images with block occlusion (outlier radio) as 10%, 20%, 30%, and 40%, respectively. Fig. 3 shows the classification results of all semisupervised methods over the four face data sets with synthetic outliers. The observation indicates the following.

1) The classification performance of all methods decreases with the increase of outlier radio, which is caused by the existence of block occlusion.
2) The accuracy of our method and ASL drops slowly in contrast to GFHF, L1-SEMI, and FME, in particularly our method performs slightly better than ASL model over all data sets.

D. Scalability Analysis

To verity the time superiority of the proposed method, we design experiments over six benchmark data sets which are copied to 2, 4, 8, and 16 times of the original data sets, respectively. The experiment run on a computer with 16 $\times$ 2.8-GHz cores and 128-GB memory. Considering that some methods including L1-SEMI, ASL, and our method utilize the alternative optimization algorithms to solve the classification problems, the time consumptions of them are decided by the equal convergence residual for fair comparison.

Fig. 4 shows the time performance of all methods over 20% labeled training data for six extended data sets. We can conclude as follows.

1) The total computing time of all methods increases with the times of data sets boosting.
2) ASL and the proposed method consume less time than other graph-based semisupervised methods, especially when the number of data is larger, which is because our proposed method and ASL need not consume additional time on constructing the Laplacian matrix.
3) Compared to our method, ASL consumes more time because it introduces an additional probability distribution of label for each data point in the procedure of ASL's optimization. In summary, the proposed method achieves a better scalability on large-scale data applications.

E. Sensitivity Analysis

There are three parameters in the proposed method, such as the thresholding parameter ε, regularization parameter λ, and p $(0 < p \leq 2)$. In this section, we analysis the influence of varying parameters on the performance of classification.

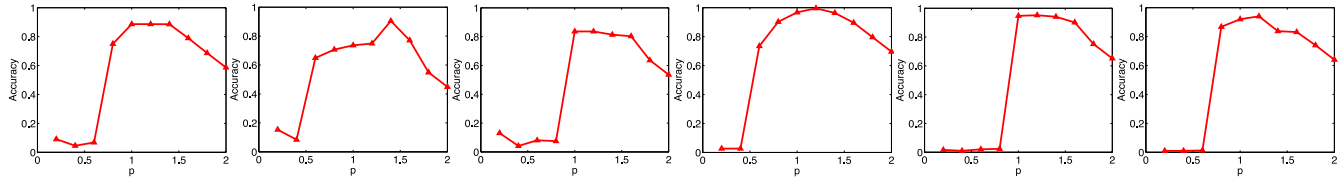


Fig. 5. Sensitivity analysis on parameter p with $\lambda = 0.1$ and $\varepsilon = 0.8$ over 20% labeled training data for all data sets. (a) JAFFE. (b) USPS. (c) MSRA50. (d) YaleB. (e) CMU-PIE. (f) PALM.

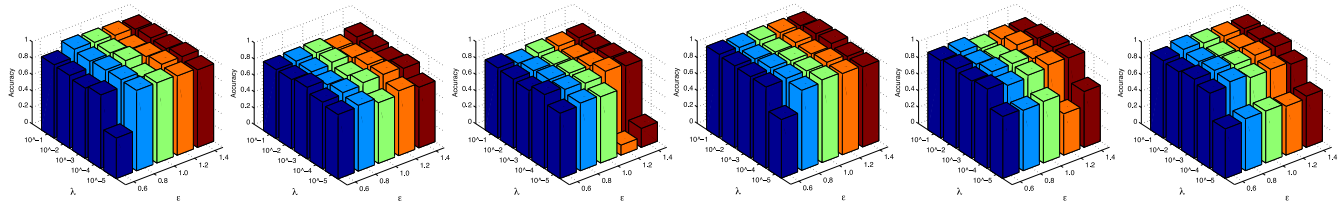


Fig. 6. Sensitivity analysis on regularization parameter λ and ε with $p = 1$ over 20% labeled training data for all data sets. (a) JAFFE. (b) USPS. (c) MSRA50. (d) YaleB. (e) CMU-PIE. (f) PALM.

In the condition of parameter $\varepsilon = 0.1$ and $\lambda = 0.8$, we study the influence of parameter p over 20% labeled training data. There, we plot the sensitivity performance curves as the increase of p in Fig. 5. It indicates that the classification accuracy increases with the increase of p. After reaching its maximum, the performance decreases gradually. The results demonstrate that a proper p could achieve the best robustness by reducing the effect of outliers through sparsity. In addition, the best parameter p is distinguishing over different data sets because of the diverse properties related to the data sets.

We also evaluate the impacts of parameters ε and λ on the performance by assigning ε and λ vary from 0.6 to 1.4 and 10^{-5} to 10^{-1}, respectively. With $p = 1$, Fig. 6 shows the sensitivity analysis result on regularization parameters λ and ε over 20% labeled training data. We achieve at the following conclusions.

1) Our method has strong robustness since the parameters λ and ε can achieve well classification performance on a large scale in all data sets.
2) The selection of best parameters is closely related to the property of data sets. In other words, the optimal parameters selected over different data sets are distinguishable.

According to the mentioned above, the classification performance of the proposed method is fluctuant with varying values of parameters p, ε, and λ. When the parameter p in interval $[1.2, 1.4]$, ε in $[0.8, 1.4]$ and λ in $[10^{-1}, 10^{-4}]$, the classification accuracy could be satisfactory and relatively stable.

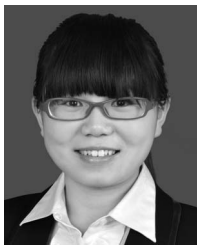
V. Conclusion

This paper proposes a novel method for semisupervised classification. Compared with other graph-based methods, our method need not construct the Laplacian matrix. As a result, our method shows superiority in computational cost for large-scale data applications. Moreover, based on the proposed $\ell_{2,p}$-norm, this method show better robustness to the outliers. An efficient alternative optimization algorithm is proposed to solve the challenging problem. We also analyze the computational complexity of the proposed algorithm in the theory.

References

[1] S. Yang, Z. Feng, Y. Ren, H. Liu, and L. Jiao, "Semi-supervised classification via kernel low-rank representation graph," *Knowl. Based Syst.*, vol. 69, pp. 150–158, Oct. 2014.
[2] M. Luo *et al.*, "Adaptive unsupervised feature selection with structure regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2650978.
[3] X. Chang, Z. Ma, M. Lin, Y. Yang, and A. G. Hauptmann, "Feature interaction augmented sparse learning for fast Kinect motion detection," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3911–3920, Aug. 2017.
[4] S. Yang *et al.*, "Semisupervised dual-geometric subspace projection for dimensionality reduction of hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3587–3593, Jun. 2014.
[5] M. Luo *et al.*, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.
[6] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, Aug. 2017.
[7] M. Zhao *et al.*, "Route selection for cabling considering cost minimization and earthquake survivability via a semi-supervised probabilistic model," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 502–511, Apr. 2017.
[8] Z. Zhang, M. Zhao, and T. W. S. Chow, "Graph based constrained semi-supervised learning framework via label propagation over adaptive neighborhood," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2362–2376, Sep. 2015.
[9] I. Triguero, S. García, and F. Herrera, "SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 622–634, Apr. 2015.
[10] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1180–1197, May 2017.
[11] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, Oct. 2017.
[12] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. NIPS*, Whistler, BC, Canada, 2004, pp. 321–328.

[13] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, vol. 3. Washington, DC, USA, 2003, pp. 912–919.
[14] D. Zhou and B. Schölkopf, "Learning from labeled and unlabeled data using random walks," in *Pattern Recognition*. Heidelberg, Germany: Springer, 2004, pp. 237–244.
[15] F. Nie, H. Wang, H. Huang, and C. Ding, "Unsupervised and semi-supervised learning via l1-norm graph," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 2268–2273.
[16] L. Zhuang *et al.*, "Constructing a nonnegative low-rank and sparse graph with data-adaptive features," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3717–3728, Nov. 2015.
[17] M. Luo *et al.*, "Adaptive semi-supervised learning with discriminative least squares regression," in *Proc. IJCAI*, Melbourne, VIC, Australia, 2017, pp. 2421–2427.
[18] V. Sindhwani, P. Niyogi, M. Belkin, and S. Keerthi, "Linear manifold regularization for large scale semi-supervised learning," in *Proc. 22nd ICML Workshop Learn. Partially Classified Training Data*, vol. 28. Bonn, Germany, 2005, pp. 80–83.
[19] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Jan. 2006.
[20] M. Hein, J.-Y. Audibert, and U. Von Luxburg, "From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians," in *Learning Theory*. Springer: Heidelberg, Germany, 2005, pp. 470–485.
[21] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
[22] F. Nie, H. Wang, H. Huang, and C. H. Ding, "Adaptive loss minimization for semi-supervised elastic embedding," in *Proc. IJCAI*, Beijing, China, 2013, pp. 1565–1571.
[23] X. Pei, Z. Lyu, C. Chen, and C. Chen, "Manifold adaptive label propagation for face clustering," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1681–1691, Aug. 2015.
[24] F. Dornaika and Y. El Traboulsi, "Learning flexible graph-based semi-supervised embedding," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 206–218, Jan. 2016.
[25] D. Wang, F. Nie, and H. Huang, "Large-scale adaptive semi-supervised learning via unified inductive and transductive model," in *Proc. SIGKDD*, New York, NY, USA, 2014, pp. 482–491.
[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.
[27] W. Jiang, F. Nie, and H. Huang, "Robust dictionary learning with capped l1-norm," in *Proc. IJCAI*, Buenos Aires, Argentina, 2015, pp. 3590–3596.
[28] T. Zhang, "Multi-stage convex relaxation for feature selection," *Bernoulli*, vol. 19, no. 5, pp. 2277–2293, 2013.
[29] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Proc. NIPS*, Vancouver, BC, Canada, 2010, pp. 1813–1821.
[30] H. Wang, F. Nie, and H. Huang, "Learning robust locality preserving projection via p-order minimization," in *Proc. AAAI*, Austin, TX, USA, 2015, pp. 3059–3065.
[31] J. Wen, Z. Lai, W. K. Wong, J. Cui, and M. Wan, "Optimal feature selection for robust classification via l2, 1-norms regularization," in *Proc. ICPR*, Stockholm, Sweden, 2014, pp. 517–521.
[32] T. Zhang, "Multi-stage convex relaxation for learning with sparse regularization," in *Proc. NIPS*, Vancouver, BC, Canada, 2009, pp. 1929–1936.
[33] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[34] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Nara, Japan, 1998, pp. 200–205.
[35] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.
[36] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[37] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[38] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 15th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, 2002, pp. 46–51.
[39] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[40] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Stat. Interface*, vol. 2, no. 3, pp. 349–360, 2009.
[41] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Multiple Classifier Systems*. Heidelberg, Germany: Springer, 2005, pp. 278–285.



Lingling Zhang received the B.S. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2015, where she is currently pursuing the Ph.D. degree.

Her current research interests include cross-media information mining, few-shot learning, and e-learning.

Minnan Luo received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2014.

She is currently an Assistant Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. She is also a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. Her current research interests include machine learning and optimization, video analysis, and cross-media retrieval.

Zhihui Li received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008.

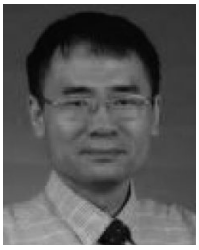
She has been a Data Analyst with Beijing Etrol Technologies Company, Ltd., Beijing. Her current research interests include artificial intelligence, machine learning, and computer vision.

Feiping Nie received the B.S. degree in computer science from the North China University of Water Conservancy and Electric Power, Zhengzhou, China, in 2000, the M.S. degree in computer science from Lanzhou University, Lanzhou, China, in 2003, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

His current research interests include machine learning, pattern recognition, data mining, and image processing.

Huaxiang Zhang received the Ph.D. degree from Shanghai Jiaotong University, Shanghai, China, in 2004.

He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China, where he was an Associate Professor with the Department of Computer Science, from 2004 to 2005. He has authored over 170 journal and conference papers and holds ten invention patents. His current research interests include machine learning, pattern recognition, evolutionary computation, and cross-media retrieval.



Jun Liu received the B.S., M.S., and Ph.D. degrees in computer science from Xi'an Jiaotong University, Xi'an, China, in 1995, 1998, and 2004, respectively.

He is currently a Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University. His current research interests include text mining, data mining, intelligent network learning environment, and multimedia e-learning.

Qinghua Zheng received the B.S. degree in computer software, the M.S. degree in computer organization and architecture, and the Ph.D. degree in system engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, 1993, and 1997, respectively.

He was a Post-Doctoral Researcher with Harvard University, Cambridge, MA, USA, in 2002. He is currently a Professor with Xi'an Jiaotong University. His current research interests include computer network security and intelligent e-learning theory and algorithm.