Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Noise-Robust Semi-Supervised Learning by Large-Scale Sparse Coding

Zhiwu Lu[1] and **Xin Gao**[2] and **Liwei Wang**[3] and **Ji-Rong Wen**[1] and **Songfang Huang**[4]
[1]School of Information, Renmin University of China, Beijing 100872, China
[2]Computer, Electrical and Mathematical Sciences and Engineering Division,
King Abdullah University of Science and Technology (KAUST), Thuwal, Jeddah 23955, Saudi Arabia
[3]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, China
[4]IBM China Research Lab, Beijing, China
{zhiwu.lu, xin4gao, wangliwei.pku, jirong.wen}@gmail.com, huangsf@cn.ibm.com

Abstract

This paper presents a large-scale sparse coding algorithm to deal with the challenging problem of noise-robust semi-supervised learning over very large data with only few noisy initial labels. By giving an L_1-norm formulation of Laplacian regularization directly based upon the manifold structure of the data, we transform noise-robust semi-supervised learning into a generalized sparse coding problem so that noise reduction can be imposed upon the noisy initial labels. Furthermore, to keep the scalability of noise-robust semi-supervised learning over very large data, we make use of both nonlinear approximation and dimension reduction techniques to solve this generalized sparse coding problem in linear time and space complexity. Finally, we evaluate the proposed algorithm in the challenging task of large-scale semi-supervised image classification with only few noisy initial labels. The experimental results on several benchmark image datasets show the promising performance of the proposed algorithm.

Introduction

Semi-supervised learning has been widely applied to many challenging image analysis tasks (Lu, Zhao, and Cai 2006; Xu and Yan 2009; Lu and Ip 2010; Fergus, Weiss, and Torralba 2010; Tang et al. 2009; Liu et al. 2009) such as image representation, image classification, and image annotation. In these image analysis tasks, the manual labeling of training data is often tedious, subjective as well as expensive, while the access to unlabeled data is much easier. In the literature, through exploiting the large number of unlabeled data with reasonable assumptions, semi-supervised learning (Blum and Mitchell 1998; Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004) has been shown to reduce the need for expensive labeled data and achieve promising results in different challenging image analysis tasks.

Among various semi-supervised learning methods, one influential work is graph-based semi-supervised learning (Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004) which models the entire dataset as a graph. The basic idea behind this semi-supervised learning is label propagation over the graph with the cluster consistency (Zhou et al. 2004). Since the graph is at the heart of graph-based semi-supervised learning, graph construction has been studied extensively (Wang and Zhang 2008; Yan and Wang 2009; Cheng et al. 2010; Zhuang et al. 2012; Sun, Hussain, and Shawe-Taylor 2014). However, these graph construction methods are not developed directly for noise reduction, and the corresponding semi-supervised learning still suffers from significant performance degradation due to the inaccurate labeling of data points commonly encountered in different image analysis tasks. Moreover, the labeling of images may be contributed by the community (e.g. Flickr) and we can only obtain noisy labels for all the images.

In this paper, we focus on developing a novel noise-robust semi-supervised learning algorithm to deal with the challenging problem of semi-supervised learning with noisy initial labels. As summarized in (Wang and Zhang 2008), graph-based semi-supervised learning can be formulated as a quadratic optimization problem based on Laplacian regularization (Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004; Lu and Peng 2013a). Considering the success of L_1-norm optimization for noise reduction (Elad and Aharon 2006; Mairal, Elad, and Sapiro 2008; Wright et al. 2009), if we give a new L_1-norm formulation of Laplacian regularization, we can propose noise-robust L_1-norm semi-supervised learning. Fortunately, derived from the eigenvalue decomposition of the normalized Laplacian matrix $\mathcal{L}$, we can represent $\mathcal{L}$ in a symmetrical decomposition form and then define L_1-norm Laplacian regularization. Since all the eigenvectors of $\mathcal{L}$ are explored in this symmetrical decomposition, our L_1-norm Laplacian regularization is actually defined based upon the manifold structure of the data. By limiting the solution of semi-supervised learning to the space spanned by the eigenvectors of $\mathcal{L}$, our noise-robust semi-supervised learning is formulated as a generalized sparse coding problem.

To keep the scalability for very large data, we utilize both nonlinear approximation and dimension reduction techniques to solve this generalized sparse coding problem. More specifically, we first construct a large graph over the data only based upon a limited number of clustering centers (obtained by k-means clustering). Due to the special definition of this graph, we can find the eigenvectors of $\mathcal{L}$ (to be used in this generalized sparse coding problem) in linear time complexity. For the sake of more efficiency, we choose to only work with a small subset of eigenvec-

tors during solving the generalized sparse coding problem. By considering nonlinear approximation and dimension reduction together, we develop a large-scale sparse coding algorithm of linear time and space complexity. In this paper, the proposed algorithm is then applied to the challenging task of semi-supervised image classification over very large data with only few noisy initial labels. Although there exist other large-scale semi-supervised learning algorithms (Zhang, Kwok, and Parvin 2009; Liu, He, and Chang 2010; Fergus, Weiss, and Torralba 2010) in the literature, they are not originally developed to deal with noisy initial labels.

To emphasize the main contributions of this paper, we summarize the following distinct advantages of our new semi-supervised learning algorithm:

- This is he first work to deal with the challenging problem of semi-supervised learning over very large data with only few noisy initial labels, to our knowledge.
- Our semi-supervised learning algorithm developed based on large-scale sparse coding has been shown to achieve promising results in semi-supervised image classification over very large data with only few noisy initial labels.
- Our new L_1-norm Laplacian regularization can be similarly applied to many other problems in machine learning and patter recognition, considering the wide use of Laplacian regularization in the literature.

The remainder of this paper is organized as follows. In Section 2, we develop a large-scale sparse coding (LSSC) algorithm for semi-supervised learning over very large data with only few noisy initial labels. In Section 3, the proposed LSSC algorithm is applied to large-scale noise-robust image classification. Sections 4 and 5 present our experimental results and conclusions, respectively.

Noise-Robust Semi-Supervised Learning

In this section, we first give the problem formulation for noise-robust semi-supervised learning by defining new L_1-norm Laplacian regularization. Moreover, we extend our noise-robust semi-supervised learning to large-scale problems by exploiting nonlinear approximation and dimension reduction techniques. Finally, we develop a large-scale sparse coding algorithm to solve the challenging problem of noise-robust semi-supervised learning over very large data.

Problem Formulation

We first introduce semi-supervised learning problem as follows. Here, we only consider the two-class problem, while the multi-class problem will be discussed in Section 3. Given a dataset $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_l, \mathbf{x}_{l+1}, ..., \mathbf{x}_n\}$ and a label set $\{1, -1\}$, the first l data points $\mathbf{x}_i$ $(i \leq l)$ are labeled as $y_i \in \{1, -1\}$ and the remaining data points $\mathbf{x}_u$ $(l+1 \leq u \leq n)$ are unlabeled with $y_u = 0$. The goal of semi-supervised learning is to predict the labels of the unlabeled data points, i.e., to find a vector $\mathbf{f} = [f_1, ..., f_n]^T$ corresponding to a classification on $\mathcal{X}$ by labeling each $\mathbf{x}_i$ with a label $\mathrm{sign}(f_i)$, where $\mathrm{sign}(\cdot)$ denotes the sign function. Let $\mathbf{y} = [y_1, ..., y_n]^T$, and we can observe that $\mathbf{y}$ is consistent with the initial labels according to the decision rule.

We further model the whole dataset $\mathcal{X}$ as a graph $\mathcal{G} = \{\mathcal{V}, W\}$ with its vertex set $\mathcal{V} = \mathcal{X}$ and weight matrix $W = [w_{ij}]_{n\times n}$, where w_{ij} denotes the similarity between data points $\mathbf{x}_i$ and $\mathbf{x}_j$. It should be noted that the weight matrix W is usually assumed to be nonnegative and symmetric. For example, we can define the weight matrix W as

$$w_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/(2\sigma^2)), \tag{1}$$

where the variance σ is a free parameter that can be determined empirically. In fact, to eliminate the need to tune this parameter, we can adopt many graph construction methods (Wang and Zhang 2008; Yan and Wang 2009; Cheng et al. 2010; Zhuang et al. 2012; Sun, Hussain, and Shawe-Taylor 2014) that have been developed in the literature. Derived from the weight matrix W, the normalized Laplacian matrix $\mathcal{L}$ of the graph $\mathcal{G}$ can be computed by

$$\mathcal{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \tag{2}$$

where I is an $n \times n$ identity matrix, and D is an $n \times n$ diagonal matrix with its i-th diagonal element being equal to the sum of the i-th row of W (i.e. $\sum_j w_{ij}$).

Since Laplacian regularization plays an important role in graph-based semi-supervised learning, we then provide a symmetrical decomposition of the normalized Laplacian matrix $\mathcal{L}$. More concretely, as a nonnegative definite matrix, $\mathcal{L}$ can be decomposed into

$$\mathcal{L} = V \Sigma V^T, \tag{3}$$

where V is an $n \times n$ orthonormal matrix with each column being an eigenvector of $\mathcal{L}$, and Σ is an $n \times n$ diagonal matrix with its diagonal element Σ_{ii} being an eigenvalue of $\mathcal{L}$ (sorted as $0 \leq \Sigma_{11} \leq ... \leq \Sigma_{nn}$). Derived from the above eigenvalue decomposition, we can denote $\mathcal{L}$ in a symmetrical decomposition form:

$$\mathcal{L} = (\Sigma^{\frac{1}{2}} V^T)^T \Sigma^{\frac{1}{2}} V^T = B^T B, \tag{4}$$

where $B = \Sigma^{\frac{1}{2}} V^T$. Since B is computed with all the eigenvectors of $\mathcal{L}$, we can regard B as being explicitly defined based upon the manifold structure of the data.

We can directly utilize $B = \Sigma^{\frac{1}{2}} V^T$ to define a new L_1-norm smoothness measure, instead of the traditional smoothness measure used as Laplacian regularization for graph-based semi-supervised learning. In spectral graph theory, the smoothness of a vector $\mathbf{f} \in R^n$ is measured by $\Omega(\mathbf{f}) = \mathbf{f}^T \mathcal{L} \mathbf{f}$. Different from the traditional smoothness $\Omega(\mathbf{f})$, in this paper, the L_1-norm smoothness of a vector $\mathbf{f} \in R^n$ is measured by $\tilde{\Omega}(\mathbf{f}) = ||B\mathbf{f}||_1$, just as our recent work (Lu and Peng 2012; 2013b; Lu and Wang 2015). Considering the success of L_1-norm optimization for noise reduction (Elad and Aharon 2006; Mairal, Elad, and Sapiro 2008; Wright et al. 2009), we then define the objective function of our noise-robust semi-supervised learning as follows

$$\tilde{Q}(\mathbf{f}) = \frac{1}{2}||\mathbf{f} - \mathbf{y}||_2^2 + \lambda ||B\mathbf{f}||_1. \tag{5}$$

The first term of $\tilde{Q}(\mathbf{f})$ is the fitting constraint, while the second term is the L_1-norm smoothness constraint used as

Laplacian regularization. Here, the fitting constraint is not formulated as an L_1-norm term. The reason is that, otherwise, most elements of $\mathbf{f}$ would tend to zeros (i.e. sparsity) by solving $\min_{\mathbf{f}} ||\mathbf{f}-\mathbf{y}||_1 + \lambda||B\mathbf{f}||_1$ given that $\mathbf{y}$ has very few nonzero elements (i.e. very few initial labeled data are usually provided for semi-supervised learning). In other words, the labels of data points are almost not propagated across the entire dataset, which *actually conflicts* with the original goal of semi-supervised learning. Hence, the fitting constraint of $\tilde{Q}(\mathbf{f})$ remains as an L_2-norm term.

It is worth noting that our L_1-norm formulation of Laplacian regularization plays an important role in the explanation of noise-robust semi-supervised learning in the framework of sparse coding. Concretely, by limiting the solution $\mathbf{f}$ to the space spanned by the eigenvectors V (i.e. $\mathbf{f} = V\alpha$), we can readily formulate our noise-robust semi-supervised learning as a sparse reconstruction problem (see the next subsection). In fact, the sparsity can be considered to be induced into the compressed domain for our noise-robust semi-supervised learning, since sparse coding is regarded as learning compressible model (Zhang, Schneider, and Dubrawski 2010) in the literature. Furthermore, we can also readily apply dimension reduction to our noise-robust semi-supervised learning by working with only a small subset of eigenvectors (i.e. only partial columns of V are used), which is especially suitable for image analysis tasks where the datasets are commonly very large. Although there exist other L_1-norm generalizations of Laplacian regularization in (Chen et al. 2011; Petry, Flexeder, and Tutz 2011; Zhou and Scholköpf 2005; Zhou, Lu, and Peng 2013) which approximately take the form of $\sum_{i<j} w_{ij}|f_i - f_j|$, they are not defined based upon the eigenvectors of the Laplacian matrix and the strategy of dimension reduction is difficult to be used for $\mathbf{f}$. Hence, the sparse coding algorithms developed directly using these L_1-norm generalizations incur very large time cost.

Large-Scale Extension

To keep the scalability of our noise-robust semi-supervised learning over very large data, we can reduce the dimension of $\mathbf{f}$ dramatically by working only with a small subset of eigenvectors of $\mathcal{L}$, instead of $\mathbf{f} = V\alpha$. That is, similar to (Fergus, Weiss, and Torralba 2010; Chapelle, Scholköpf, and Zien 2006), we significantly reduce the dimension of $\mathbf{f}$ by requiring it to take the form of $\mathbf{f} = V_m\alpha$ where V_m is an $n \times m$ matrix whose columns are the m eigenvectors with smallest eigenvalues (i.e. the first m columns of V). In fact, such dimension reduction can ensure that $\mathbf{f}$ is as smooth as possible in terms of our L_1-norm smoothness. Hence, the objective function of our noise-robust semi-supervised learning can now be derived from Eq. (5) as follows:

$$\begin{aligned}\tilde{Q}(\alpha) &= \frac{1}{2}||(V_m\alpha)-\mathbf{y}||_2^2 + \lambda||(\Sigma^{\frac{1}{2}}V^T)(V_m\alpha)||_1 \\ &= \frac{1}{2}||V_m\alpha-\mathbf{y}||_2^2 + \lambda||\sum_{i=1}^{m}\Sigma^{\frac{1}{2}}(V^TV_{\cdot i})\alpha_i||_1 \\ &= \frac{1}{2}||V_m\alpha-\mathbf{y}||_2^2 + \lambda\sum_{i=1}^{m}\Sigma_{ii}^{\frac{1}{2}}|\alpha_i|. \end{aligned} \quad (6)$$

The first term of $\tilde{Q}(\alpha)$ denotes the reconstruction error, while the second term denotes the weighted L_1-norm sparsity regularization over the reconstruction coefficients. That is, our L_1-norm semi-supervised learning problem $\mathbf{f}^* = \arg\min_{\mathbf{f}} \tilde{Q}(\mathbf{f})$ has been successfully transformed into a generalized sparse coding problem $\alpha^* = \arg\min_{\alpha} \tilde{Q}(\alpha)$.

It should be noted that the formulation $\mathbf{f} = V_m\alpha$ used in Eq. (6) has two distinct advantages. Firstly, we can derive a sparse reconstruction problem from the original semi-supervised learning problem, and correspondingly we can explain our noise-robust semi-supervised learning in the framework of sparse coding. In fact, the second term of $\tilde{Q}(\alpha)$ corresponds to both Laplacian regularization and sparsity regularization. By unifying these two types of regularization, we thus obtain novel noise-robust semi-supervised learning. Secondly, since $\tilde{Q}(\alpha)$ is minimized with respect to $\alpha \in R^m$ ($m \ll n$), we can readily develop fast sparse coding algorithms for our noise-robust semi-supervised learning. That is, although many sparse coding algorithms scale polynomially with m, they have linear time complexity with respect to n. More importantly, we *have eliminated the need to compute the full matrix* B in Eq. (5), which is especially suitable for image analysis on large datasets. In fact, we only need to compute the m smallest eigenvectors of $\mathcal{L}$.

We further pay our main attention to finding the m smallest eigenvectors of $\mathcal{L}$. Fortunately, we can keep the scalability of this step by exploiting the following nonlinear approximation technique. Concretely, given k clustering centers $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k\}$ obtained by k-means clustering over the dataset $\mathcal{X}$, we find the approximation $\hat{\mathbf{x}}_i$ of any data point $\mathbf{x}_i$ by Nadaraya-Watson kernel regression (Härdle 1992):

$$\hat{\mathbf{x}}_i = \sum_{j=1}^{k} z_{ij}\mathbf{u}_j, \quad (7)$$

where $Z = [z_{ij}]_{n\times k}$ collects the regression coefficients. A natural assumption here is that z_{ij} should be larger if $\mathbf{x}_i$ is closer to $\mathbf{u}_j$. We can emphasize this assumption by setting $z_{ij} = 0$ as $\mathbf{u}_j$ is not among the $r(\leq k)$ nearest neighbors of $\mathbf{x}_i$. This restriction naturally leads to a sparse matrix Z. Let $\mathcal{N}_r(i)$ denote the indexes of r clustering centers that are nearest to $\mathbf{x}_i$. We compute z_{ij} ($j \in \mathcal{N}_r(i)$) as

$$z_{ij} = \frac{\exp(-||\mathbf{x}_i - \mathbf{u}_j||^2/(2\sigma^2))}{\sum_{j'\in\mathcal{N}_r(i)}\exp(-||\mathbf{x}_i - \mathbf{u}_{j'}||^2/(2\sigma^2))}, \quad (8)$$

where the parameter σ is defined the same as Eq. (1). The weight matrix $W \in R^{n\times n}$ of the graph $\mathcal{G}$ over the dataset $\mathcal{X}$ can now be computed as:

$$W = \hat{Z}\hat{Z}^T, \quad (9)$$

where $\hat{Z} = ZD_z^{-1/2}$ and D_z is a $k \times k$ diagonal matrix with its i-th diagonal element being the sum of the i-th column of Z. Since each row of Z sums up to 1, the degree matrix is I and the normalized Laplacian matrix $\mathcal{L}$ is $I - W$. This means that finding the m smallest eigenvectors of $\mathcal{L}$ is equivalent to finding the m largest eigenvectors of W.

Let the singular value decomposition (SVD) of $\hat{Z}$ be:

$$\hat{Z} = V_z\Sigma_z U_z^T, \quad (10)$$

Algorithm 1 Large-Scale Sparse Coding (LSSC)

Input: $\mathcal{X}$, $\mathbf{y}$, k, σ, r, m, and λ
Output: the predicted labels by $\mathrm{sign}(\mathbf{f}^*)$
Step 1. Find k clustering centers $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k\}$ by k-means clustering over the dataset $\mathcal{X}$.
Step 2. Compute the weight matrix W of the graph $\mathcal{G}$ over the dataset $\mathcal{X}$ according to Eq. (9).
Step 3. Find the m largest eigenvectors of W according to Eq. (12) and store them in V_m.
Step 4. Solve the problem $\alpha^* = \arg\min_\alpha \tilde{Q}(\alpha)$ using the modified FISTA.
Step 5. Compute $\mathbf{f}^* = V_m\alpha^*$.

where $\Sigma_z = \mathrm{diag}(\sigma_1, ..., \sigma_k)$ with σ_i being a singular value of $\hat{Z}$ (sorted as $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k \geq 0$), V_z is an $n \times k$ matrix with each column being a left singular vector of $\hat{Z}$, and U_z is a $k \times k$ matrix with each column being a right singular vector of $\hat{Z}$. It is easy to check that each column of V_z is an eigenvector of $W = \hat{Z}\hat{Z}^T$, and each column of U_z is an eigenvector of $\hat{Z}^T\hat{Z}$ (the eigenvalues are $\sigma_1^2, ..., \sigma_k^2$ in both cases). Since $\hat{Z}^T\hat{Z} \in R^{k\times k}$, we can compute U_z within $O(k^3)$ time. V_z can then be computed as:

$$V_z = \hat{Z}U_z\Sigma_z^{-1}. \tag{11}$$

Hence, to find the m ($m < k$) smallest eigenvectors of $\mathcal{L} = I - W$, we first find the m largest eigenvectors $U_m \in R^{k\times m}$ of $\hat{Z}^T\hat{Z}$ (the eigenvalues store in $\Sigma_m^2 = \mathrm{diag}(\sigma_1^2, ..., \sigma_m^2)$) and then compute the m largest eigenvectors V_m of W as:

$$V_m = \hat{Z}U_m\Sigma_m^{-1}. \tag{12}$$

which can then be used in Eq. (6). Since both finding V_m (including k-means) and solving $\min_\alpha \tilde{Q}(\alpha)$ have a *linear time and space complexity* with respect to n ($m, k, r \ll n$), our semi-supervised learning is scalable to very large data.

The Proposed Algorithm

In theory, any fast sparse coding algorithm can be adopted to solve the problem $\min_\alpha \tilde{Q}(\alpha)$. In this paper, we only consider the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck and Teboulle 2009), since its implementation mainly involves lightweight operations such as vector operations and matrix-vector multiplications. To adjust FISTA for our noise-robust semi-supervised learning, we only need to modify the soft-thresholding function as:

$$\mathrm{soft}(\alpha_i, \frac{\lambda\Sigma_{ii}^{\frac{1}{2}}}{||V_m||_s^2}) = \mathrm{sign}(\alpha_i)\max\{|\alpha_i| - \frac{\lambda\Sigma_{ii}^{\frac{1}{2}}}{||V_m||_s^2}, 0\}, \tag{13}$$

where $||V_m||_s$ represents the spectral norm of the matrix V_m. For large problems, it is often computationally expensive to directly compute the Lipschitz constant $||V_m||_s^2$. In practice, it can be efficiently estimated by a backtracking line-search strategy (Beck and Teboulle 2009). The complete large-scale sparse coding (LSSC) algorithm for our noise-robust semi-supervised learning is outlined in Algorithm 1.

Application to Image Classification

To show the advantage of the proposed LSSC algorithm, we apply it to semi-supervised image classification over very large data with only few noisy initial labels. It should be noted that semi-supervised image classification has been studied extensively (Xu and Yan 2009; Lu and Ip 2010; Fergus, Weiss, and Torralba 2010; Tang et al. 2009), which can be considered as the basis of many image analysis tasks such as image annotation and retrieval. In these tasks, the manual labeling of training data is often tedious and expensive, while the access to unlabeled data is much easier. The original motivation of semi-supervised image classification is just to reduce the need for expensive labeled data by exploiting the large number of unlabeled data. In other words, the original task of semi-supervised image classification is to learn with *both labeled and unlabeled data*.

In this paper, we consider a more challenging task, i.e., semi-supervised image classification with *both correctly and incorrectly labeled data*. In general, the occurrence of noisy initial labels may be due to the subjective manual labeling of training data. Fortunately, this challenging problem can be addressed to some extent by our LSSC algorithm. That is, due to the use of L_1-norm Laplacian regularization, our LSSC algorithm can effectively suppress the negative effect of noisy initial labels. Since we focus on verifying this noise-robustness advantage, we directly apply our LSSC algorithm to semi-supervised image classification over very large data with only few noisy initial labels. Hence, we only need to extend Algorithm 1 to multi-class problems commonly encountered in image classification.

We first introduce multi-class semi-supervised image classification problem similar to (Zhou et al. 2004). Given a dataset $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_l, \mathbf{x}_{l+1}, ..., \mathbf{x}_n\}$ and a label set $\{1, ..., C\}$ (C is number of classes), the first l images $\mathbf{x}_i$ ($i \leq l$) are labeled as: $y_{ij} = 1$ if $\mathbf{x}_i$ belongs to class j ($1 \leq j \leq C$) and $y_{ij} = 0$ otherwise, while the remaining images $\mathbf{x}_u$ ($l+1 \leq u \leq n$) are unlabeled with $y_{uj} = 0$. The goal of semi-supervised image classification is to predict the labels of the unlabeled images, i.e., to find a matrix $F = [f_{ij}]_{n\times C}$ corresponding to a classification on the dataset $\mathcal{X}$ by labeling each image $\mathbf{x}_i$ with a label $\arg\max_{1\leq j\leq C} f_{ij}$. Let $Y = [y_{ij}]_{n\times C}$, and we can readily observe that Y is exactly consistent with the initial labels according to the decision rule. When noisy initial labels are provided for semi-supervised image classification, some entries of Y may be inconsistent with the ground truth labels.

Based on the above preliminary notations, we further formulate our multi-class LSSC problem for semi-supervised image classification as:

$$\min_F \tilde{Q}(F) = \min_F \frac{1}{2}||F - Y||_{fro}^2 + \lambda||BF||_1, \tag{14}$$

where $||\cdot||_{fro}$ denotes the Frobenius norm of a matrix. The above multi-class LSSC problem can then be decomposed into the following C independent subproblems:

$$\min_{F_{\cdot j}} \tilde{Q}(F_{\cdot j}) = \min_{F_{\cdot j}} \frac{1}{2}||F_{\cdot j} - Y_{\cdot j}||_2^2 + \lambda||BF_{\cdot j}||_1, \tag{15}$$

where $F_{\cdot j}$ and $Y_{\cdot j}$ denote the j-th column of F and Y, respectively. Since each subproblem, $\min_{F_{\cdot j}} \tilde{Q}(F_{\cdot j})$ ($1 \leq$

Table 1: Details of the four large image datasets.

Datasets	#classes	#features	#images	#labeled
MNIST_HALF	10	784	35K	100
MNIST	10	784	70K	100
NUS_HALF	81	1500	135K	405
NUS_WIDE	81	1500	269K	405

$j \leq C$), can be regarded as two-class semi-supervised learning, we can readily solve it by Algorithm 1. Let $F^*_{\cdot j} = \arg\min_{F_{\cdot j}} \tilde{Q}(F_{\cdot j})$, and we can classify image $\mathbf{x}_i$ into the class that satisfies $\arg\max_{1 \leq j \leq C} f^*_{ij}$.

Experimental Results

In this section, we evaluate the proposed LSSC algorithm on four large image datasets listed in Table 1, where MNIST_HALF and NUS_HALF are derived from MNIST[1] and NUS_WIDE[2] (Chua et al. 2009), respectively. The proposed LSSC algorithm is compared with three state-of-the-art large-scale semi-supervised learning methods: PVM (Zhang, Kwok, and Parvin 2009), Eigenfunction (Fergus, Weiss, and Torralba 2010), and LGC-SSL (Liu, He, and Chang 2010). We also report the performance of one baseline method LIBLINEAR (i.e. large-scale SVM) (Fan et al. 2008). All these methods are implemented in MATLAB 7.12 and run on a 3.40 GHz, 32GB RAM Core 2 Duo PC.

Experimental Setup

In the experiments, we consider three noise levels (i.e. 0%, 15%, and 30%) for large-scale semi-supervised image classification. Here, the noise level denotes the percentage of inaccurately labeled images among a limited number of initial labeled images. In this paper, to generate an inaccurately labeled image, we first randomly select a labeled image and then change each of its initial labels to a random wrong label. It should be noted that a labeled image may have multiple labels (i.e. belong to multiple classes) for the two natural image datasets (i.e. NUS_HALF and NUS_WIDE).

The classification results on unlabeled images are averaged over 10 splits of each dataset just as in (Karlen et al. 2008) and then used for overall performance evaluation. In particular, we choose accuracy as the measure of classification results for the two handwritten digit datasets (i.e. MNIST_HALF and MNIST), while mean average precision (MAP) is used as the measure for the two natural images datasets given that we actually perform multi-label image classification over these two datasets.

We find that our LSSC algorithm is not sensitive to λ in our experiments, and thus fix this parameter at $\lambda = 0.01$ for all the four datasets. Meanwhile, we uniformly set $k = 5000$ by considering a tradeoff of running efficiency and effectiveness. We follow the strategy of parameter selection used in (Chapelle and Zien 2005) and select the rest parameters (i.e. σ, r, and m) for our LSSC algorithm by five-fold cross-validation over initial labeled images (to generate the labeled set and validation set, with the other images forming the unlabeled set). In our experiments, the fold generation process is repeated randomly two times for a total of 10 splits of each dataset. The optimal values of these parameters are chosen with respect to the validation results. For example, we set $\sigma = 0.4$, $r = 3$, and $m = 18$ for the two handwritten digit datasets. For fair comparison, the same strategy of parameter selection is adopted for all the other related methods.

[1]http://yann.lecun.com/exdb/mnist/

[2]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Classification Results

Although our original motivation is to apply our LSSC algorithm to large-scale noise-robust image classification, we first make comparison in the less challenging task of large-scale semi-supervised image classification with clean initial labeled images to verify its effectiveness in dealing with the scarcity of labeled images. The number of initial labeled images provided for each dataset is listed in Table 1, where the initial labeled images are indeed scarce for all the four datasets. The comparison results are reported in Tables 2 and 3 (with the noise level being 0%). In general, we can observe that our LSSC algorithm performs better than (or at least comparably to) the other four methods for large-scale image classification. That is, due to the use of L_1-norm Laplacian regularization for problem formulation, our LSSC algorithm can effectively suppress the negative effect of the complicated manifold structure hidden among all the images on large-scale image classification.

We make further comparison in the challenging task of large-scale noise-robust image classification with noisy initial labeled images. The comparison results are reported in Tables 2 and 3, where two noise levels (i.e. 15% and 30%) are considered for initial labeled images. The immediate observation is that our LSSC algorithm generally achieves obvious gains over the other three SSL methods, especially when more noisy labels are provided initially for each dataset. That is, our L_1-norm Laplacian regularization indeed can help to find a smooth and also sparse solution for large-scale semi-supervised learning and thus effectively suppress the negative effect of noisy initial labels. More importantly, although all the four SSL methods suffer from more performance degradation when the percentage of noisy initial labels grows, the performance of Eigenfunction and our LSSC is shown to degrade degrade most slowly in large-scale noise-robust image classification.

The comparison in terms of both accuracy and running time over the MNIST dataset is shown in Table 4. Here, only 30% noise level is considered. We find that the running time taken by our LSSC algorithm is comparable to that taken by PVM and LGC-SSL. Since our LSSC algorithm is shown to achieve significant gains over these two closely related methods, it is preferred by overall consideration in practice. Moreover, excluding the running time (i.e. 237.5 seconds) taken by k-means clustering to find clustering centers (e.g. random sampling can be used instead of k-means), our LSSC algorithm itself is considered to run very efficiently over such a large dataset.

As we have mentioned, the parameter k is uniformly set to $k = 5000$ for all the four datasets by considering a tradeoff of running efficiency and effectiveness of our LSSC algo-

Table 2: Comparison of different large-scale learning algorithms with varying noise levels on the two handwritten digit datasets. The classification results are measured by accuracy (%) along with standard deviation (in brackets). The first three methods make use of k-means clustering ($k = 5000$) to find clustering centers.

Datasets	Noise level	LSSC (ours)	PVM	LGC-SSL	Eigenfunction	LIBLINEAR
MNIST_HALF	0%	**89.8** (1.1)	80.3 (0.8)	88.9 (1.0)	73.8 (2.3)	74.3 (2.1)
	15%	**88.8** (1.7)	73.1 (1.0)	82.5 (0.6)	67.3 (1.4)	67.1 (1.1)
	30%	**86.1** (3.6)	63.1 (2.1)	72.1 (1.7)	60.4 (2.6)	57.8 (2.5)
MNIST	0%	**93.1** (0.7)	81.4 (0.9)	90.4 (0.7)	73.8 (1.6)	75.5 (1.3)
	15%	**91.1** (2.0)	73.1 (1.4)	83.5 (1.6)	68.6 (2.8)	67.3 (3.0)
	30%	**89.0** (3.6)	64.3 (2.1)	74.4 (2.8)	61.9 (4.0)	58.9 (3.2)

Table 3: Comparison of different large-scale learning algorithms with varying noise levels on the two natural image datasets. The classification results are measured by MAP (%) along with standard deviation (in brackets). The first three methods make use of k-means clustering ($k = 5000$) to find clustering centers.

Datasets	Noise level	LSSC (ours)	PVM	LGC-SSL	Eigenfunction	LIBLINEAR
NUS_HALF	0%	**19.1** (0.5)	17.8 (0.8)	18.9 (0.4)	12.1 (0.2)	18.0 (0.4)
	15%	**17.3** (0.5)	14.7 (0.7)	15.8 (0.7)	10.9 (0.2)	13.9 (0.7)
	30%	**15.4** (0.6)	12.4 (0.3)	13.3 (0.4)	9.9 (0.3)	10.8 (0.2)
NUS_WIDE	0%	18.5 (0.6)	18.2 (0.4)	**18.8** (0.4)	11.8 (0.3)	18.0 (0.3)
	15%	**16.8** (0.6)	14.2 (0.6)	15.6 (0.4)	10.3 (0.3)	14.0 (0.4)
	30%	**14.6** (0.4)	11.6 (0.4)	12.6 (0.2)	8.7 (0.4)	10.1 (0.3)

Table 4: The classification results in terms of both accuracy and running time (including k-means clustering if applicable) over the MNIST dataset. Here, only 30% noise level is considered and the running time taken by k-means clustering ($k = 5000$) is 237.5 seconds.

Methods	accuracy (%)	Running time (sec.)
LSSC (ours)	**89.0**	23.7+237.5
PVM	64.3	21.5+237.5
LGC-SSL	74.4	19.4+237.5
Eigenfunction	61.9	18.9
LIBLINEAR	58.9	2.7

Table 5: The classification accuracies (%) for different number of clustering centers (i.e. k) over the MNIST dataset. Here, only 30% noise level is considered.

k	1000	2000	3000	4000	5000	6000
LSSC (ours)	**84.2**	**86.6**	**87.0**	**88.3**	**89.0**	**89.0**
PVM	63.9	63.3	63.6	64.1	64.3	64.2
LGC-SSL	71.5	72.8	74.0	74.3	74.4	74.5

rithm. To make this clearer, the comparison among LSSC, PVM, and LGC-SSL with different number of clustering centers is shown in Table 5. Here, only 30% noise level over the MNIST dataset is considered. It can be clearly observed that all the three methods tend to achieve better results when more clustering centers are used for nonlinear approximation (see Eq. (7)). In particular, our LSSC algorithm is shown to approach the highest accuracy for $k = 5000$. Additionally, we can still find that our LSSC algorithm performs the best, when k changes from 1000 to 6000.

Conclusions

We have proposed a large-scale sparse coding algorithm to deal with the challenging problem of noise-robust semi-supervised learning over very large data with only few noisy initial labels. By giving an L_1-norm formulation of Laplacian regularization directly based upon the manifold structure of the data, we have transformed noise-robust semi-supervised learning into a generalized sparse coding problem so that noise reduction can be imposed upon the noisy initial labels. Furthermore, to keep the scalability of noise-robust semi-supervised learning over very large data, we have adopted both nonlinear approximation and dimension reduction techniques to solve this generalized sparse coding problem in linear time and space complexity. When applied to the challenging task of large-scale noise-robust image classification, our LSSC algorithm has been shown to achieve promising results on several benchmark datasets. In the future work, considering the wide use of Laplacian regularization in the literature, we will apply our new L_1-norm Laplacian regularization to other challenging problems in machine learning and pattern recognition.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China under Grants 61202231 and 61222307, National Key Basic Research Program (973 Program) of China under Grant 2014CB340403, Beijing Natural Science Foundation of China under Grant 4132037, Ph.D. Programs Foundation of Ministry of Education of China under Grant 20120001120130, the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China under Grant 14XNLF04, and the IBM Faculty Award.

References

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proc. Conference on Learning Theory (COLT)*.

Chapelle, O., and Zien, A. 2005. Semi-supervised classification by low density separation. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 57–64.

Chapelle, O.; Scholköpf, B.; and Zien, A., eds. 2006. *Semi-Supervised Learning*. MIT Press.

Chen, X.; Lin, Q.; Kim, S.; Carbonell, J. G.; and Xing, E. P. 2011. Smoothing proximal gradient method for general structured sparse learning. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, 105–114.

Cheng, B.; Yang, J.; Yan, S.; Fu, Y.; and Huang, T. 2010. Learning with ℓ^1-graph for image analysis. *IEEE Trans. Image Processing* 19(4):858–866.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. 2009. NUS-WIDE: A real-world web image database from National University of Singapore. In *Proc. CIVR*.

Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing* 15(12):3736–3745.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Fergus, R.; Weiss, Y.; and Torralba, A. 2010. Semi-supervised learning in gigantic image collections. In *Advances in Neural Information Processing Systems 22*, 522–530.

Härdle, W. 1992. *Applied Nonparametric Regression*. Cambridge University Press.

Karlen, M.; Weston, J.; Erkan, A.; and Collobert, R. 2008. Large scale manifold transduction. In *Proc. ICML*, 448–455.

Liu, J.; Li, M.; Liu, Q.; Lu, H.; and Ma, S. 2009. Image annotation via graph learning. *Pattern Recognition* 42(2):218–228.

Liu, W.; He, J.; and Chang, S.-F. 2010. Large graph construction for scalable semi-supervised learning. In *Proc. ICML*, 679–686.

Lu, Z., and Ip, H. 2010. Combining context, consistency, and diversity cues for interactive image categorization. *IEEE Trans. Multimedia* 12(3):194–203.

Lu, Z., and Peng, Y. 2012. Image annotation by semantic sparse recoding of visual content. In *Proc. ACM Multimedia*, 499–508.

Lu, Z., and Peng, Y. 2013a. Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications. *International Journal of Computer Vision* 103(3):306–325.

Lu, Z., and Peng, Y. 2013b. Latent semantic learning with structured sparse representation for human action recognition. *Pattern Recognition* 46(7):1799–1809.

Lu, Z., and Wang, L. 2015. Noise-robust semi-supervised learning via fast sparse coding. *Pattern Recognition* 48(2):605–612.

Lu, K.; Zhao, J.; and Cai, D. 2006. An algorithm for semi-supervised learning in image retrieval. *Pattern Recognition* 39(4):717–720.

Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *IEEE Trans. Image Processing* 17(1):53–69.

Petry, S.; Flexeder, C.; and Tutz, G. 2011. Pairwise fused Lasso. Technical Report 102, Department of Statistics, University of Munich.

Sun, S.; Hussain, Z.; and Shawe-Taylor, J. 2014. Manifold-preserving graph reduction for sparse semi-supervised learning. *Neurocomputing* 124:13–21.

Tang, J.; Yan, S.; Hong, R.; Qi, G.-J.; and Chua, T.-S. 2009. Inferring semantic concepts from community-contributed images and noisy tags. In *Proc. ACM Multimedia*, 223–232.

Wang, F., and Zhang, C. 2008. Label propagation through linear neighborhoods. *IEEE Trans. Knowledge and Data Engineering* 20(1):55–67.

Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 31(2):210–227.

Xu, D., and Yan, S. 2009. Semi-supervised bilinear subspace learning. *IEEE Trans. Image Processing* 18(7):1671–1676.

Yan, S., and Wang, H. 2009. Semi-supervised learning by sparse representation. In *Proc. SIAM Conference on Data Mining (SDM)*, 792–801.

Zhang, K.; Kwok, J. T.; and Parvin, B. 2009. Prototype vector machine for large scale semi-supervised learning. In *Proc. ICML*, 1233–1240.

Zhang, Y.; Schneider, J. G.; and Dubrawski, A. 2010. Learning compressible models. In *Proc. SIAM Conference on Data Mining (SDM)*, 872–881.

Zhou, D., and Scholköpf, B. 2005. Regularization on discrete spaces. In *Proc. 27th Annual meeting of the German Association for Pattern Recognition (DAGM)*, 361–368.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 321–328.

Zhou, G.; Lu, Z.; and Peng, Y. 2013. L1-graph construction using structured sparsity. *Neurocomputing* 120:441–452.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. ICML*, 912–919.

Zhuang, L.; Gao, H.; Lin, Z.; Ma, Y.; Zhang, X.; and Yu, N. 2012. Non-negative low rank and sparse graph for semi-supervised learning. In *Proc. CVPR*, 2328–2335.