2016 IEEE 16th International Conference on Data Mining

Adaptive Neighborhood Propagation by Joint L2,1-norm Regularized Sparse Coding for Representation and Classification

Lei Jia [#], Zhao Zhang [#], Lei Wang [#], Weiming Jiang [#], and Mingbo Zhao [$]

[#] School of Computer Science and Technology & Joint International Research Laboratory of Machine Learning and Neuromorphic Computing, Soochow University, Suzhou 215006, China

[$] Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

E-mails: {20155227021, 20145227001, 20164227009}@stu.suda.edu.cn, cszzhang@gmail.com

***Abstract*— We propose a new transductive label propagation method, termed *Adaptive Neighborhood Propagation* (Adaptive-NP) by joint L2,1-norm regularized sparse coding, for semi-supervised classification. To make the predicted soft labels more accurate for predicting the labels of samples and to avoid the tricky process of choosing the optimal neighborhood size or kernel width for graph construction, Adaptive-NP seamlessly integrates sparse coding and neighborhood propagation into a unified framework. That is, the sparse reconstruction error and classification error are combined for joint minimization, which clearly differs from traditional methods that explicitly separate graph construction and label propagation into independent steps, which may result in inaccurate predictions. Note that our Adaptive-NP alternately optimize the sparse codes and soft labels matrices, where the sparse codes are used as adaptive weights for neighborhood propagation at each iteration, so the tricky process of determining neighborhood size or kernel width is avoided. Besides, for enhancing sparse coding, we use the L2,1-norm constraint on the sparse coding coefficients and the reconstruction error at the same time for delivering more accurate and robust representations. Extensive simulations show that our model can deliver state-of-the-art performances on several public datasets for classification.**

Index Terms*— *Transductive learning; linear neighborhood propagation; L2,1-norm regularized sparse coding; classification

I. INTRODUCTION

Semi-supervised learning (SSL) [1] has been an important topic in the area of data mining [2][3]. Different from the unsupervised/supervised methods, SSL methods can apply small number of labeled data and large amount of unlabeled data for learning, which well suits the characteristics of real application data. That is, massive vision data or non-vision data generated in real world and virtual networks are high-dimensional and unlabeled, but labeled data is often costly to obtain and the labeling process is time consuming [2].

In the past years, various graph-based (G-SSL) methods [4][6-8][10][32-34] are presented due to elegant formulation and successful applications. Two basic assumptions of G-SSL are the cluster and manifold assumptions [12]. Label propagation (LP), as a typical G-SSL algorithm for label prediction, has been arousing considerable attention in more recent years due to its effectiveness and fast speed. LP is a process that propagates supervised class prior information of labeled data to the unlabeled data based on the relationships between both labeled and unlabeled data [5][8][10][30].

Typical transductive label propagation algorithms consist of the *SSL using Gaussian Fields and Harmonic Functions* (GFHF) [10], *Learning with Local and Global Consistency* (LLGC) [21], *Linear Neighborhood Propagation* (LNP) [5], *Special Label Propagation* (SLP) [8] and *Sparsity Induced Similarity Measure for Label Propagation* (SIS-LP) [22]. Note that these transductive methods aim to estimate the soft labels of unlabeled data by receiving information partially from the initial label and partially from its neighborhoods [5][8][10], where the latter part from neighborhoods is mainly decided by a weighted neighborhood graph. The neighbor of each sample is usually determined using K-neighborhood or ε-neighborhood. For assigning weights, two popular ways [13], i.e., Gaussian function and *Locally Linear Embedding* (LLE)-style reconstruction weights [18], are widely used in GFHF, LLGC, SLP and LNP. Note that the above weighting methods may suffer from two shortcomings. First, Gaussian function has to choose an optimal kernel width, which is not easy in reality; Second, both methods have to determine the neighborhood size and more importantly the neighbor size is always fixed as the same value for each point artificially, which is not reasonable in fact, since it does not consider the distributions of real data. To obtain adaptive edge weights for measuring pairwise similarities more accurately, several recent LP methods apply a sparse coding based weighting method, i.e., using the sparse codes as adaptive weights, such as SIS-LP [22] and *Label Propagation through Sparse Neighborhood* (LPSN) [27], etc. For obtaining adaptive weights, the aforementioned methods apply a L1-norm based sparse coding process that reconstructs each sample using a linear combination of compact samples so that more accurate neighborhood structures can be discovered to enhance the label predictions. It must be noted that although different weighting methods can be employed to encode the manifold smoothness degree, virtually all existing LP models suffer from one common drawback, that is, an independent graph weighting process is performed before the label estimation. Thus, the encoded weights separately cannot be ensured to be optimal for subsequent label prediction, which may result in decreased performance. It is also worth noting that the existing adaptive edge weights based methods, e.g., SIS-LP

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.26

and LPSN, apply the Frobenius norm that is very sensitive to noise and outliers in data [14-16] to encode the sparse reconstruction error, which may also cause inaccurate sparse representations for adaptive graph weighting.

In this paper, we propose a unified framework termed *Adaptive Neighborhood Propagation* (Adaptive-NP) method that seamlessly integrates sparse coding with neighborhood propagation for transductive label prediction. Different from existing LP methods that separate the graph construction and label propagation explicitly into two independent steps, our Adaptive-NP combines the sparse reconstruction error with the classification error for simultaneous minimization, which ensures the learnt sparse representations would be optimal as adaptive graph weights for more accurate label predictions. Note that our Adaptive-NP is solved in an alternate manner, i.e., optimizing the sparse representations and soft labels, where the sparse codes are used as the adaptive weights for neighborhood propagation at each iteration. Based on the unified framework, Adaptive-NP avoids the tricky process of determining neighborhood size or kernel width, and more importantly the learnt sparse representations are optimal for adaptive weight construction. In addition, to improve the robustness of the sparse coding process, we impose the L2,1-norm constraint [14-16] on sparse reconstruction error and also on the coding coefficients at the same time so that more accurate and robust representations can be delivered for encoding the neighborhoods to measure similarities, since L2,1-norm can explicitly ensure the representation matrix and the reconstruction error are sparse in rows, which can ensure the sparse properties of the representations and can also potentially reduce the reconstruction error to deliver enhanced performances. The convergence behavior is also theoretically analyzed, showing that the objective function of our Adaptive-NP is monotonically decreasing in iterations. Besides, the connection with other related LP criteria is also discussed and several existing LP methods can be regarded as special cases of our proposed formulation.

The paper is outlined as follows. Section II reviews the related models briefly. Section III shows the formulation of Adaptive-NP mathematically. The convergence analysis is also shown. Section IV describes the settings and evaluation results. Finally, the paper is concluded in Section V.

TABLE I. IMPORTANT NOTATIONS USED IN THE PAPER

Notation	Description
n	Dimensionality of samples
N	Number of all the samples
c	Total number of sample labels
$X=[X_L,X_U]\in\mathbb{R}^{n\times N}$	Whole set of samples
$X_L=[x_1,x_2,...,x_l]\in\mathbb{R}^{n\times l}$	Original labeled set
$X_U=[x_{l+1},x_{l+2},...,x_{l+u}]\in\mathbb{R}^{n\times u}$	Original unlabeled set
$Y=[y_1,y_2,...,y_{l+u}]\in\mathbb{R}^{c\times N}$	Initial class label matrix
$W=[w_1,w_2,...,w_N]\in\mathbb{R}^{N\times N}$	Graph weight matrix
$F=[f_1,f_2,...,f_{l+u}]\in\mathbb{R}^{c\times N}$	Predicted soft label matrix

II. NOTATIONS AND RELATED WORK

A. Linear Neighborhood Propagation (LNP)

We briefly review the LNP method [5] that is closely related to our model. Given a set of samples $X=[x_1,x_2,...,x_N]\in\mathbb{R}^{n\times N}$ and a class label set $L=\{1,2,...,c\}$, where n is the original dimensionality of each sample x_i, N is the number of samples. For SSL, l points x_i are considered as labeled, and the rest u samples are unlabeled, where $N=l+u$. Note that important notations used in this paper are shown in Table I.

Graph construction [4] [6-8] is a core of LNP, and it uses the reconstruction weights to encode the similarities between samples, i.e., LNP assumes the neighborhood of each point are linear, so each point can be reconstructed using a linear combination of its neighbors [5]. The minimization problem for obtaining the reconstruction weights x_i is defined as

$$\varepsilon=\left\|x_i-\sum_{i_j:x_{i_j}\in\mathbb{N}(x_i)}w_{\ddot{i}i_j}x_{i_j}\right\|^2, s.t.\sum_{i_j\in\mathbb{N}(x_i)}w_{\ddot{i}i_j}=1, w_{\ddot{i}i_j}\geq 0, \quad (1)$$

where $\mathbb{N}(x_i)$ is the K-neighbor set of sample x_i, x_{i_j} is the j-th neighbor of x_i and $w_{\ddot{i}i_j}$ denotes the contribution of each x_{i_j} for reconstructing x_i. After the reconstruction weights of all points are computed, a sparse weight W can be obtained by

$$W(i,j)=\left[w_{ij}\right]\in\mathbb{R}^{N\times N}, \quad (2)$$

In each propagation step, by letting each object absorb a fraction of label information from its neighbors and retain some label information of its initial set, the predicted labels $F=[f_1,f_2,...,f_{l+u}]\in\mathbb{R}^{c\times N}$ of LNP can be obtained by

$$F^{\mathrm{T}}=(1-\alpha)\left(I^N-\alpha W\right)^{-1}Y^{\mathrm{T}}, F\in\mathbb{R}^{c\times N}, \quad (3)$$

where $Y=[y_1,y_2,\ldots,y_{l+u}]\in\mathbb{R}^{c\times N}$ is the initial label matrix of all samples and I^N is an identify matrix in $\mathbb{R}^N$. Note that $y_{i,j}=1$ if x_j is labeled as $i(1\leq i\leq c)$ and else $y_{i,j}=0$. $0<\alpha<1$ is a control parameter. Finally, the label of each object can be assigned as $\arg\max_{i\leq c}F_{i,j}$, that is, the largest entry in each soft label vector f_i determines the final hard label of each point x_i. According to [5][17], the objective function of LNP can be formulated as

$$\min_F\ tr\left(F\left(I^N-W\right)F^{\mathrm{T}}\right)+\mu\sum_{i=1}^{l+u}\left\|f_i-y_i\right\|_2^2, \quad (4)$$

where $\|\bullet\|_2$ is L2-norm of a vector. By applying the matrix expression, the above objective function can be rewritten as

$$\begin{aligned}\min_F\ & tr\left(F\left(I^N-W\right)F^{\mathrm{T}}\right)+\mu\left\|F-Y\right\|_F^2\\ &=tr\left(F\left(I^N-W\right)F^{\mathrm{T}}\right)+tr\left(\mu(F-Y)(F-Y)^{\mathrm{T}}\right)\end{aligned}. \quad (5)$$

where $\|\bullet\|_F$ is Frobenius norm, the first term is the manifold smoothness term the second one is the label fitness term.

B. Sparsity Induced Similarity Measure for LP

We show another one more related approach called SIS-LP that also use the sparse coding to define the adaptive weight for LP. The main idea of SIS-LP consists of two steps. First,

SIS-LP seeks the *Sparsity Induced Similarity Measure* (SIS), and then propagates label information from labeled points to whole sets by using the SIS weights. Specifically, SIS-LP can be regarded for solving the following two sub-problems:

$$\begin{cases} \underset{S}{Min} \ \|S\|_{\ell_0}, \ S.t. \ f_k = G_k S \\ G_m = \left(D_{mm} - \left(D^{-1}S\right)_{mm}\right)^{-1} S_{mn} G_n \end{cases}, \quad (6)$$

where $\|S\|_{\ell_0}$ is the L0-norm of S, and $S = \left(s_1,...,s_{k-1},s_{k+1},s_N\right)^{\mathrm{T}}$ is the coding coefficient matrix after sparse decomposition. Let $F = \{f_1, f_2,..., f_N\}$ denotes all label vectors of sample set X. $G_k = \left(f_1,...,f_{k-1},f_{k+1},...,f_N\right)$ denotes the rest of label vectors in F, and D denotes a $N \times N$ diagonal matrix with $d_{ii} = \sum_j S_{i,j}$. In summary, we obtain S by solving the linear programming problem from the first problem in Eq. (6). Then, we can propagate labels from the labels G_n of labeled data to labels G_m of unlabeled data by S via the second formula in Eq. (6).

It is worth noting that virtually all existing LP methods, including LNP and SIS-LP, etc, perform weight construction and label prediction using two separable steps. Thus, learnt sparse representations or coding coefficients or affinity values cannot be ensured as optimal for subsequent label prediction. To this end, we will propose a unified adaptive LP framework that seamlessly integrates sparse coding with LP so that the label prediction results are more accurate.

III. Adaptive Neighborhood Propagation by Joint $L_{2,1}$-norm Regularized Sparse Coding

A. Proposed Formulation

We propose Adaptive-NP to boost the performance of LNP for predicting the unknown labels of unlabeled samples more accurately. The improvements over LNP are twofold. First, to overcome the drawback of existing methods that separate the graph construction and label propagation, especially for the sparse coding based adaptive LP methods, our Adaptive-NP proposes to combine the sparse reconstruction error with classification error for simultaneous minimization so that the learnt sparse representations can be ensured to be optimal for measuring the manifold smoothness for more accurate label predictions. Second, so as to reduce the sparse reconstruction error in sparse coding, we regularize the robust L2,1-norm [14][17][28] instead of Frobenius norm on the reconstruction error and also on the coding coefficients at the same time to compute the new adaptive weights so that more accurate and robust representations can be obtained [25-26]. Thus, the unified adaptive LP framework of Adaptive-NP can be formulated involving two variables F and S as:

$$\underset{F,S}{Min} \ \hat{J}(F,S) = \sum_{i=1}^{N}\|f_i - Fs_i\|_2^2 + \sum_{i=1}^{N} u_i \|f_i - y_i\|_2 + \alpha\left[\sum_{i=1}^{N}\|x_i - Xs_i\|_2 + \beta\sum_{i=1}^{N}\|s_i\|_2\right], \quad (7)$$

where $F = \left[f_1, \cdots f_N\right] \in \mathbb{R}^{c\times N}$ is a predicted soft label matrix, $Y \in \mathbb{R}^{c\times N}$ denotes the initial label matrix, S denotes a sparse representation matrix or sparse coding coefficients matrix, α and β are trade-off parameters, $\sum_{i=1}^{N}\|f_i - Fs_i\|_2^2$ is the manifold smoothness term, $\sum_{i=1}^{N} u_i \|f_i - y_i\|_2$ is a label fitness term, $\sum_{i=1}^{N}\|x_i - Xs_i\|_2 + \beta\sum_{i=1}^{N}\|s_i\|_2$ is a L2,1-norm based sparse coding term for computing the coding coefficients as the adaptive weights for label prediction, and μ_i is adjustable parameters for both labeled and unlabeled data, i.e., $\mu_i = 10^{10}$ for labeled data and else $\mu_i = 0$. It is clear that simultaneous minimization of $\sum_{i=1}^{N}\|x_i - Xs_i\|_2$ and $\sum_{i=1}^{N}\|f_i - Fs_i\|_2^2$ ensures that the learnt adaptive weights are optimal for measuring neighborhoods over both training samples and predicted soft labels. In addition, L2,1-norm can ensure the sparse property of the representations and can also potentially improve the manifold smoothness to deliver enhanced performances. It is worth noting that the unified framework of our Adaptive-NP can be performed alternately between the following steps:

(1) Adaptive weight construction by L2,1-norm based sparse coding: We fix the soft label matrix F and focus on learning the coding coefficients S to update the adaptive graph weights. The reduced problem can be obtained as

$$\underset{S}{Min} \ \hat{J}(S) = \sum_{i=1}^{N}\|f_i - Fs_i\|_2^2 + \alpha\left[\sum_{i=1}^{N}\|x_i - Xs_i\|_2 + \beta\sum_{i=1}^{N}\|s_i\|_2\right], \quad (8)$$

where L2,1-norm is imposed on the coding coefficients and the sparse reconstruction term $\|X - XS\|_{2,1}$. It is clear that the above problem forces the sparse representation matrix S to simultaneously minimize the reconstruction errors over both soft labels and samples, which can intuitively lead to the enhanced label prediction performance. After the adaptive weights are updated, we can learn the predicted soft labels by optimizing the following adaptive LP process.

(2) Adaptive label propagation and prediction: The sparse coding coefficient matrix S is fixed and we focus on propagating labels of labeled data to the remaining unlabeled data. We have the following LP formulation:

$$\underset{F}{Min} \ \hat{J}(F) = \sum_{i=1}^{N}\|f_i - Fs_i\|_2^2 + \sum_{i=1}^{N} u_i \|f_i - y_i\|_2, \quad (9)$$

which is similar to the objective function of LNP in form except that the weights are different. Based on the proposed unified framework, we can obtain an adaptive weight matrix S and a discriminative soft label matrix F jointly. In addition, we use an efficient method to solve the L2,1-norm based problems, which can potentially make the training phase efficient and can also reduce the computation cost, compared with solving the L0-norm or L1-norm based problems by *Orthogonal Matching Pursuit* (OMP) [31] or other iterative non-convex optimization.

B. Optimization

In this section, we show the optimization procedures for the objective function of our presented Adaptive-NP method in

Eq. (7). Before describing of optimization of our problem, let us introduce the Lr,p-norm first. For a matrix $A \in \mathbb{R}^{n\times N}$, its Lr,p-norm is defined as follows [14]:

$$\|A\|_{r,p} = \left(\sum_i^n \left(\sum_{j=1}^N |A_{ij}|^r \right)^{p/r} \right)^{1/p}. \quad (10)$$

Clearly, when $p = r = 2$, it identifies the commonly used Frobenius norm or L2-norm; when $p = 1, r = 2$, it becomes the commonly used robust L2,1-norm.

Next, we describe the optimization. It is clear that the formulation of our Adaptive-NP involves two main variables (S, F) to optimize. More importantly, the two variables depend on each other, thus the optimization problem of our Adaptive-NP cannot be solved directly. By following the common procedures, we optimize Adaptive-NP by using an alternate strategy, i.e., updating one of the variables each time by fixing the others. We first express the objective function of our Adaptive-NP in Eq. (7) using matrix form as

$$\begin{aligned} \underset{F,S,V}{Min}\ \hat{J}(F,S,V) = &\|F - FS\|_F^2 + tr\left((F-Y)UV(F-Y)^{\mathrm{T}}\right) \\ &+ \alpha\left(\|X - XS\|_{2,1} + \beta\|S\|_{2,1}\right) \end{aligned}, \quad (11)$$

where U is a diagonal matrix with μ_i as its elements, and V is also a diagonal matrix whose element is represented as $V_{ii} = 1/\left(2\left\|\hat{f}^i - \hat{y}^i\right\|_2\right)$ by the definition of L2,1-norm [14-16], i.e. $\sum_{i=1}^N u_i \|f_i - y_i\|_2$ can be treated as a weighted L2,1-norm, where $F = F^{\mathrm{T}} = \left[\hat{f}^1; \hat{f}^2; ...; \hat{f}^N\right]$, $\hat{Y} = Y^{\mathrm{T}} = \left[\hat{y}^1; \hat{y}^2; ...; \hat{y}^N\right]$. Next, we detail the optimization process.

(1) Given the soft label matrix *F*, update the sparse codes *S*. We first show how to solve S, that is to solve the problem in Eq. (8) with the soft label matrix F and V fixed. Therefore, the above problem can be simplified as

$$\underset{S}{Min}\ \hat{J}(S) = \|F - FS\|_F^2 + \alpha\left(\|X - XS\|_{2,1} + \beta\|S\|_{2,1}\right), \quad (12)$$

which is also the matrix form of the problem in Eq.(8). Note that the L2,1-norm based term $\|X - XS\|_{2,1} + \beta\|S\|_{2,1}$ for adaptive weights construction by sparse coding is generally convex [14], but the derivative does not exist when $m^i = 0$ or $s^i = 0$ where $i = 1,2,...,N$, $M = X - XS = \left[\hat{m}^1; \hat{m}^2; ...; \hat{m}^N\right]$. Thus, when $\hat{m}^i \neq 0$ and $\hat{s}^i \neq 0$, $i = 1,2,...,N$, the above formulation in Eq.(12) can be expressed involving several variables as

$$\begin{aligned} \underset{S,G,Q}{Min}\ \wp(S,G,Q) = &tr\left((F-FS)(F-FS)^{\mathrm{T}}\right) \\ &+ \alpha\left(tr\left((X-XS)G(X-XS)^{\mathrm{T}}\right) + \beta tr\left(SQS^{\mathrm{T}}\right)\right) \end{aligned}, \quad (13)$$

since $\|F - FS\|_F^2 = tr\left((F-FS)(F-FS)^{\mathrm{T}}\right)$, $\|S\|_{2,1} = tr\left(2S^{\mathrm{T}}QS\right)$, where $G \in \mathbb{R}^{n\times n}$ and $Q \in \mathbb{R}^{N\times N}$ are diagonal matrices whose diagonal elements are defined as

$$G_{ii} = 1/\left[2\left\|\hat{m}_t^i\right\|_2\right],\ Q_{ii} = 1/\left[2\left\|\hat{s}_t^i\right\|_2\right]. \quad (14)$$

Note that we initialize S with *LLE*-reconstruction weight matrix $W = \left[\tilde{w}_{ij}\right] \in \mathbb{R}^{N\times N}$ [18]. The weights for reconstructing each x_i are obtained for each data by the following problem:

$$\underset{\tilde{w}_i}{Min} \sum_{i=1}^n \left\| x_i - \sum_{j=1}^K \tilde{w}_{i,j} x_j \right\|_2^2,\ Subj \sum_{j=1}^K \tilde{w}_{i,j} = 1, \tilde{w}_{i,j} \geq 0, \quad (15)$$

where K is the number of nearest neighbors of each x_i. Note that when S is known, we can compute G and Q according to the definitions of G_{ii} and Q_{ii} in Eq. (14). After G and Q are updated, we are ready to update S. By taking the derivate of $\wp(S,G,Q)$ with respect to S, we can easily have

$$\begin{aligned} \partial\wp(S,G,Q)/\partial S = &\left(2F^{\mathrm{T}}FS + 2\alpha X^{\mathrm{T}}GXS + 2\partial\beta QS\right) \\ &-\left(2\partial X^{\mathrm{T}}GX + 2F^{\mathrm{T}}F\right) \end{aligned}. \quad (16)$$

By further setting $\partial\wp(S,G,Q)/\partial S = 0$, we can update S_{t+1} at the *(t+1)*-th iteration by

$$S_{t+1} = \left(F_t^{\mathrm{T}}F_t + \alpha X^{\mathrm{T}}G_tX + \alpha\beta Q_t\right)^{-1}\left(\alpha X^{\mathrm{T}}G_tX + F_t^{\mathrm{T}}F_t\right). \quad (17)$$

It should be noted that minimizing $tr\left(S^{\mathrm{T}}QS\right) = \|S\|_{2,1}/2$ will add explicit sparsity constraint on S [14-17].

(2) Given the adaptive weights using the sparse codes *S*, update the label matrix *F*. We show the optimization for delivering the soft labels for label assignment. Specifically, we update the soft label matrix F and diagonal matrix V by

$$\underset{F,V}{Min}\ \wp(F,V) = tr\left((F-FS)(F-FS)^{\mathrm{T}}\right) + tr\left((F-Y)UV(F-Y)^{\mathrm{T}}\right). \quad (18)$$

By taking the derivate of $\wp(F,V)$ with respect to the variable F, the solution can be formulated as

$$\partial\wp(F,V)/\partial F = 2F(I-S)(I-S)^{\mathrm{T}} + 2FUV - 2YUV, \quad (19)$$

and by setting the above derivate to zero, we can update F_{t+1} at the *(t+1)*-th iteration by

$$F_{t+1} = (YUV_t)\left((I-S_t)(I-S_t)^{\mathrm{T}} + UV_t\right)^{-1}. \quad (20)$$

When the soft label matrix F_{t+1} is updated, we can compute the diagonal entries for V as

$$V_{ii} = 1/\left[2\left\|\hat{f}_{t+1}^i - \hat{y}^i\right\|_2\right]. \quad (21)$$

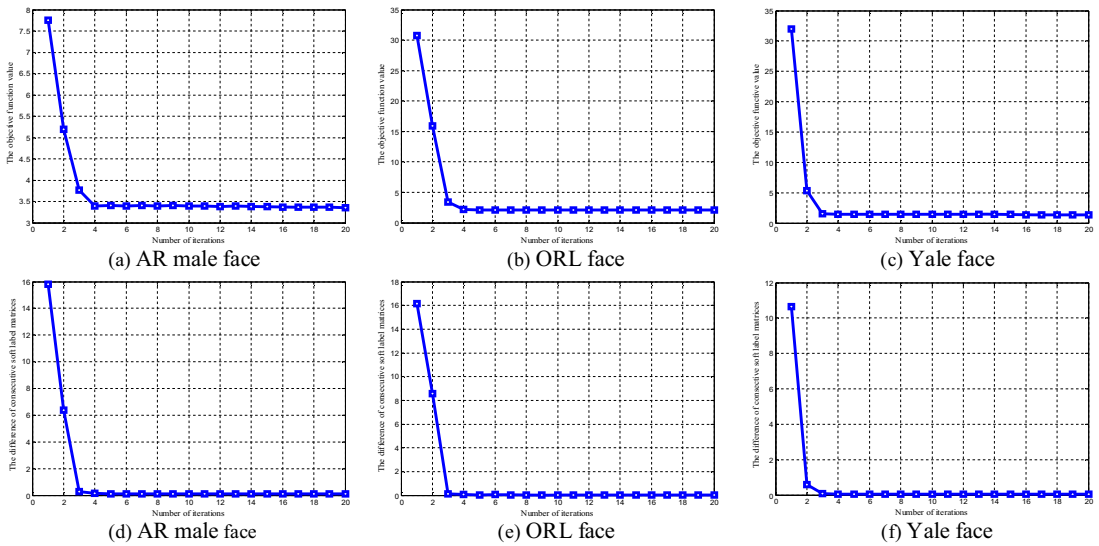
After convergence of algorithm, we can get the optimal adaptive weights S and label matrix F^*, where the position corresponding to the biggest element in the label vector f_i determines the class assignment of each x_i. That is, the hard label of each test data x_i can be assigned as $\arg\max_{i\leq c}(f_i)_i$, where $(f_i)_i$ is the i-th entry of estimated soft label vector f_i. For complete presentation of the approach, we summarize the procedures of Adaptive-NP in Algorithm 1.

C. Convergence Analysis

Adaptive-NP is solved in an alternative manner, so we want to show its convergence behavior. First, a lemma [29] that can assist the proof is provided.

Lemma 1. For any nonzero vectors a and b, the following inequality holds:

$$\|a\|_2 - \|a\|_2^2/\left[2\|b\|_2\right] \leq \|b\|_2 - \|b\|_2^2/\left[2\|b\|_2\right]. \quad (22)$$



(a) AR male face (b) ORL face (c) Yale face

(d) AR male face (e) ORL face (f) Yale face

Figure 1. Convergence behavior of Adaptive-NP, where the first row shows the results of objective function and the second row indicates the divergence between consecutive soft label matrices F.

Then, the convergence behavior of our Adaptive-NP is summarized in the following theorem.

Proposition 1. The objective function value of the problem of our Adaptive-NP during the iterations is non-increasing using the presented optimization procedures.

Proof: When we fix V as V_t in the t-th iteration to compute S_{t+1} and F_{t+1} the following inequality holds:

$$J(S_{t+1},V_t,F_{t+1}) \le J(S_t,V_t,F_t). \tag{23}$$

Because we have $\|M\|_{2,1}=\sum_i^N \left\|\widehat{m}_i\right\|_2$ and $\|S\|_{2,1}=\sum_i^N \left\|\widehat{s}_i\right\|_2$, the above inequality indicates:

$$\begin{aligned}
&\|F-FS\|_F^2+tr\left((F_{t+1}-Y)UV_t(F_{t+1}-Y)^{\mathrm{T}}\right)+\alpha\|M_{t+1}\|_{2,1}\\
&+\alpha\sum_{i=1}^{N}\left(\frac{\left\|\widehat{m}_{t+1}^i\right\|_2^2}{2\left\|\widehat{m}_t^i\right\|_2}-\left\|\widehat{m}_{t+1}^i\right\|_2\right)+\alpha\beta\|S_{t+1}\|_{2,1}+\alpha\beta\sum_{i=1}^{N}\left(\frac{\left\|\widehat{s}_{t+1}^i\right\|_2^2}{2\left\|\widehat{s}_t^i\right\|_2}-\left\|\widehat{s}_{t+1}^i\right\|_2\right)\\
&\le\|F-FS\|_F^2+tr\left((F_t-Y)UV_{t-1}(F_t-Y)^{\mathrm{T}}\right)+\alpha\|M_t\|_{2,1}\\
&+\alpha\sum_{i=1}^{N}\left(\frac{\left\|\widehat{m}_t^i\right\|_2^2}{2\left\|\widehat{m}_t^i\right\|_2}-\left\|\widehat{m}_t^i\right\|_2\right)+\alpha\beta\|S_t\|_{2,1}+\alpha\beta\sum_{i=1}^{N}\left(\frac{\left\|\widehat{s}_t^i\right\|_2^2}{2\left\|\widehat{s}_t^i\right\|_2}-\left\|\widehat{s}_t^i\right\|_2\right)
\end{aligned}. \tag{24}$$

Recalling the inequality in Lemma 1, we can obtain that

$$\begin{aligned}
&\frac{\left\|\widehat{m}_{t+1}^i\right\|_2^2}{2\left\|\widehat{m}_t^i\right\|_2}-\left\|\widehat{m}_{t+1}^i\right\|_2\ge\frac{\left\|\widehat{m}_t^i\right\|_2^2}{2\left\|\widehat{m}_t^i\right\|_2}-\left\|\widehat{m}_t^i\right\|_2\\
&\frac{\left\|\widehat{s}_{t+1}^i\right\|_2^2}{2\left\|\widehat{s}_t^i\right\|_2}-\left\|\widehat{s}_{t+1}^i\right\|_2\ge\frac{\left\|\widehat{s}_t^i\right\|_2^2}{2\left\|\widehat{s}_t^i\right\|_2}-\left\|\widehat{s}_t^i\right\|_2
\end{aligned}. \tag{25}$$

By combining Eq. (24) with Eq. (25), we can achieve the following result:

$$\begin{aligned}
&\|F-FS\|_F^2+tr\left((F_{t+1}-Y)UV_t(F_{t+1}-Y)^{\mathrm{T}}\right)+\alpha\|M_{t+1}\|_{2,1}+\alpha\beta\|S_{t+1}\|_{2,1}\\
&\le\|F-FS\|_F^2+tr\left((F_t-Y)UV_{t-1}(F_t-Y)^{\mathrm{T}}\right)+\alpha\|M_t\|_{2,1}+\alpha\beta\|S_t\|_{2,1}
\end{aligned}, \tag{26}$$

which inequality indicates that the objective function in Eq. (9) will monotonically decrease in the iterations. In addition, since the objective function has lower bounds, such as zero, the above iteration will converge. One problem should be pointed out here, that is, the above Proposition only indicates that the objective function is non-increasing. But we still do not know whether F converges, where F is the major variable to pursue. Thus, we also would like to measure the variance between two sequential Fs by the following metric:

$$Error(t)=\sum_{i=1}^{N}\left\|f_{t+1}^i\right\|_2-\left\|f_t^i\right\|_2\Big|, \tag{27}$$

which can guarantee that the final results will not be changed drastically. Note that we provide some convergence analysis results for illustration, where we show two groups of results. The first group is about the objective function and the other is the divergence between two consecutive soft label matrices F using the evaluation metric in Eq. (27). The convergence analysis results averaged over 20 times iterations are shown in Figure 1. In this study, three face image databases, i.e., Yale face, ORL face and AR male face sets are used. The detailed introduction about these face datasets are shown in Table II. As can be seen from Figure 1, the objective function values of Adaptive-NP are non-increasing during the iterations and they converge to a fixed value. Besides, the divergence between consecutive soft label matrices also converges to zero, which means that the final results will not be changed drastically. It is worth noting that the convergence speed is relatively fast, and the number of iterations is usually less than 10.

D. *Approach for Including Outside Samples*

We discuss the approach for our Adaptive-NP to include the out-of-sample samples using the similar method as LNP.

Given a new test data x_{new}, we first search its *K*-neighbors from labeled training samples, and then seek the coefficient vector $w(x_{new}, x)$ that measures the contribution of its neighbors for reconstructing x_{new}. Similar to LNP for inclusion, we use the same smoothness criterion for x_{new} by solving a problem:

$$\vartheta(f(x_{new})) = \sum_{i:x_i \in X_L, x_i \in \mathbb{N}(x_{new})} w(x_{new}, x_i)(f(x_{new}) - f_i)^2 . \quad (28)$$

Since $\vartheta(f)$ is convex in $f(x_{new})$, it is minimized when:

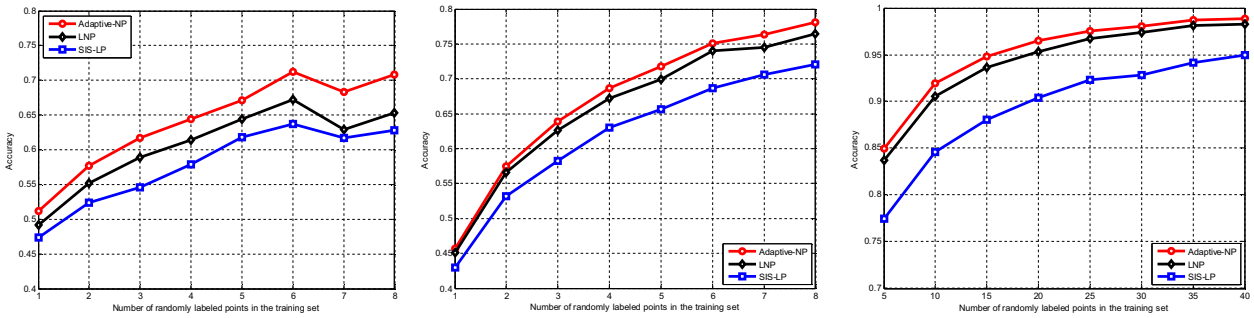
$$f(x_{new}) = \sum_{i:x_i \in X_L, x_i \in \mathbb{N}(x_{new})} w(x_{new}, x_i) f_i , \quad (29)$$

where $\mathbb{N}(x_{new})$ is the *K*-neighborhood of x_{new}. Finally, the label of x_{new} can be optimally reconstructed from the labeled samples in the training set, that is:

$$f(x_{new}) = \min_{f(x_{new})} \left\| f(x_{new}) - \sum_{i:x_i \in X_L, x_i \in \mathbb{N}(x_{new})} w(x_{new}, x_i) f_i \right\|^2 , \quad (30)$$

from which we can obtain the soft label vector $f(x_{new})$ of x_{new} by minimizing Eq. (30), where the position corresponding to the biggest element in the label vector $f(x_{new})$ determines the class assignment of x_{new}.

To illustrate the effectiveness of using our Adaptive-NP for including the out-of-sample points, we prepare a simulation to verify this. We describe the comparison results with those methods e.g., LNP and SIS-LP, which are closely related to our Adaptive-NP. Note that we adopt the same inductive method to extend SIS-LP to out-of-sample data as LNP does. In this study, three real-world image datasets, i.e., COIL20 object database, Yale face database and Georgia Tech face database are involved. Note that COIL20 object database has 72 subjects with 1440 records, and Georgia Tech face database has 15 subjects with 750 face images as a whole. The number of *K* used in nearest neighbor search here is set to 7 for each method. We show the results in Figure 2, where the number of unlabeled training set is also fixed in each simulation, the number of labeled training set is varied, and the rest is considered as outside points. Thus, we can observe the change trends. The horizontal axis is the number of labeled training set from each class. It is clear that the prediction performance can be effectively improved with the increasing number of labeled training data from each class. It can also be noted that our Adaptive-NP delivering consistent better results than the other two methods in investigated cases.



(a) Yale face (b) Georgia Tech face (c) COIL20 object

Figure 2. Out-of-Sample data inclusion performance of each method on (a) Yale face; (b) Georgia Tech face; (c) COIL20 object.

E. *Connection to Sparse Coding Based Adaptive LP Methods*

We mainly show the differences and connections with SIS-LP [22] and *LP through Sparse and its applications* (LPSN) [27]. The main difference between our proposed Adaptive-NP and existing sparse coding based adaptive LP methods, e.g., SIS-LP and LPSN, is that our Adaptive-NP seamlessly integrates sparse coding and neighborhood propagation into a unified framework, but existing SIS-LP and LPSN explicitly separate the processes of sparse coding and label propagation into two independent steps. We illustrate the differences by using the following working principle diagrams in Figure.3.

SIS weight is a sparse decomposition coefficient matrix, which used to measure the similarities among data points. The main idea of SIS-LP is that coefficients in such a sparse decomposition reflect the point's neighborhood structure thus providing better similarity measures among the decomposed data point and the rest of the samples.

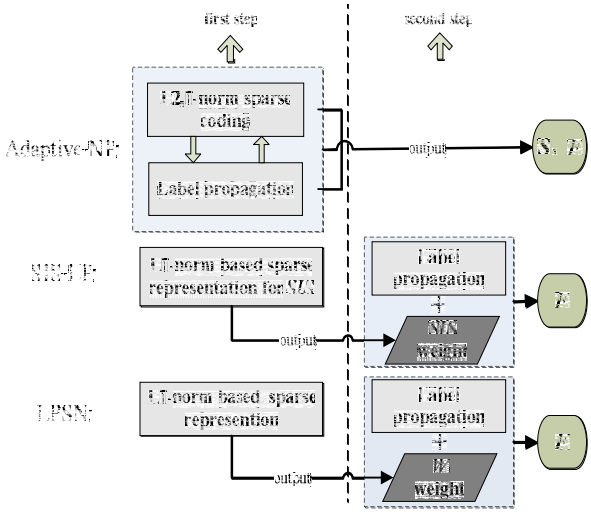


Figure 3. Comparison of the working principles of our Adaptive-NP and the existing SIS-LP and LPSN methods.

TABLE II. DESCRIPTION OF THE FACE RECOGNITION DATABASES

Dataset Name	#Classes (c)	#Dim (n)	#Points
Yale face	15	1024	165
Indian male face	39	1024	435
ORL face	40	1024	400
AR male face	50	1024	1300
AR female face	50	1024	1300
MIT face	10	1024	3240
YaleB face	38	1024	2414

TABLE III. DESCRIPTION OF THE THREE UCI DATABASES

Dataset Name	#Classes (c)	#Dim (n)	#Points
Statlog	7	19	2310
Balance Scale	3	4	625
SPECT Heart	2	22	267

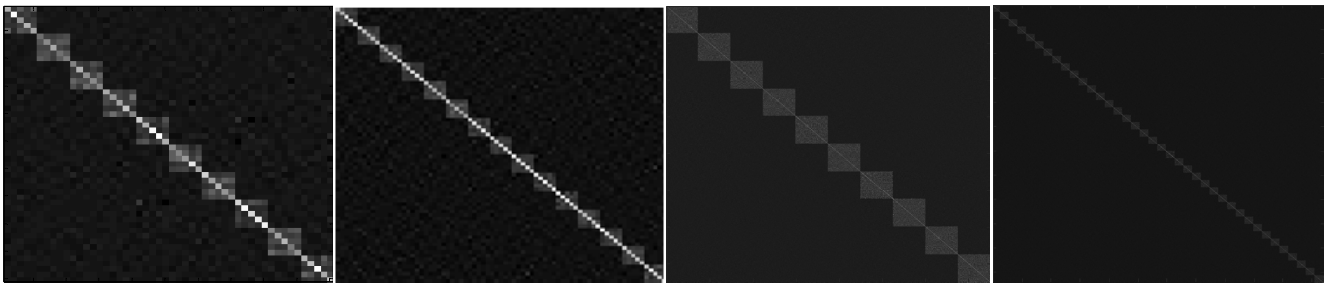
In Figure 3, W weight in LSPN denotes the reconstruction coefficient matrix, where W_{ij} is the contribution of $x_j \in N(x_i)$ to x_i, and $N(x_i)$ represents the neighborhood of x_i. LPSN algorithm assumes that, for the *i*-th sample x_i, the soft label of x_i can be linearly reconstructed by its sparse neighborhoods. From Figure 3, it is clear that: (1) our proposed Adaptive-NP only needs one step, which ensures the reconstruction error and classification error can be jointly minimized. Besides, the optimization of L2,1-norm based problems is more efficient than those of L0- and L1-norm minimization used in SIS-LP and LPSN, which will reduce the computational cost and is time-saving during the training phase; (2) SIS-LP and LPSN explicitly involves two separable steps, that is, graph weight construction plus label propagation, which cannot ensure the learnt sparse representations are optimal for classification.

IV. EXPERIMENTS

We perform experiments to illustrate the effectiveness of our Adaptive-NP in several aspects in this section, i.e., we mainly evaluate our algorithm by visual observation of constructed weights, by quantitative evaluation of data classification and by investigating the effects of model parameters. Note that the classification performance of our Adaptive-NP is compared with several related label propagation models, including SLP [8], LNP [5], LLGC [21], LapLDA [9], GFHF [10], *Prior Class Dissimilarity based LNP* (CD-LNP) [7], SIS-LP [22] and SparseNP [17]. For fair comparison, all the simulations are repeated 15 times and the averaged results are reported. For transductive classification, we randomly split each given dataset into a labeled set and an unlabeled set. Besides, we use the grid search approach to select the parameters. All the experiments are carried out on a PC with Intel (R) Core (TM) i5-4590 @ 3.30Hz 8.00 GB.

A. Visualization of Adaptive Weights by Sparse Coding

For the purpose of classification, the weight matrix should have powerful discriminating capability for different classes. In ideal conditions, the weight matrix should also be enforced to be block-diagonal so that each sample can be reconstructed by the samples of the same class as much as possible, which can potentially improve the classification performance. Thus, in this study, we mainly illustrate the adaptive weights in our method for visual evaluation. Four public face sets including Yale, MIT, AR male and YaleB (The detailed descriptions of these datasets will be shown in Table II) are used as examples. For each database, we select 5 face images from each class to form the labeled training set and treat the rest as unlabeled. Figure 4 illustrates the constructed adaptive reconstruction weight matrices over each database, which the elements of the weight matrices reflect the similarity information between any pair of samples. The larger the reconstruction weight values, the lighter the corresponding pixels in the illustration of the weight matrices, where each column of the results corresponds to the adaptive reconstruction weight vector of each sample and the larger weight values mean the bigger contribution of reconstructing given sample. The reconstruction contribution degree measures the similarity of the sample pair, i.e., the more similar the sample pair is, the larger the contribution degree for reconstructing the paired sample is. From Figure 4, we find that our Adaptive-NP can ensure the block-diagonal structures of the coding coefficients well, that is, less inter-connections that may decrease performance are included. Note that the similar sample pairs are more likely to come from the same subject/class, which is the major reason why the computed reconstruction weight matrices have block-diagonal structures. It should be noted that constructed reconstruction weights are adaptive in our model, because no neighborhood size is pre-defined, which is actually one contribution of our manuscript. Moreover, less unfavorable mixed signs are included in the visualization results. In addition, because we regularize the L2,1-norm on the weight matrix for iterative optimization, the adaptive graph weights would be potentially discriminated and robust to noise.

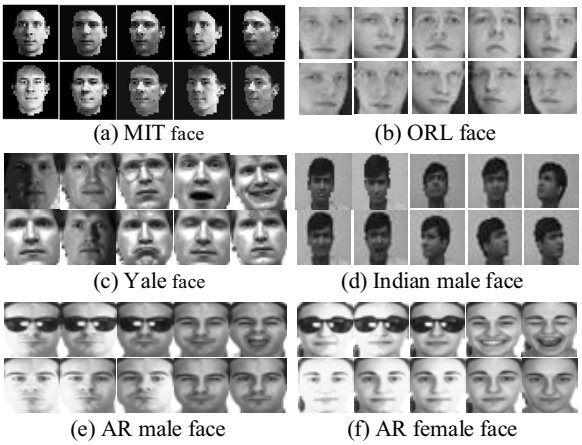


(a) AR male face (b) Yale face (c) MIT face (d) YaleB face

Figure 4. Illustration of weight matrix S constructed using the proposed algorithm on the face database.

B. Face Recognition

We examine our Adaptive-NP method for face representation and recognition. The face recognition performance is mainly compared with those of SLP, LNP, LLGC, LapLDA, GFHF, CD-LNP, SIS-LP and SparseNP. In this experiment, six real face image databases, i.e., Yale, Indian male face, ORL, AR male face, AR female face and MIT, we tested. The detailed descriptions about the databases are summarized in Table II, including dataset name, class number, data dimension and the number of images. As is common practice, all the images are resized into 32×32 pixels, so each image corresponds to a data point in a 1024-dimensional space. Meanwhile, we have shown some typical image examples of those face databases in Figure 5. Table IV shows the statistics of each method, where we report the mean accuracy (%), standard deviation (%) and run time (s) for each algorithm. Under each setting, we vary the number of labeled object face images from 1 to 8 with interval 1 and took the average. We observe from the results in Table IV that our Adaptive-NP obtains higher accuracies compared with the other recent LP methods in most cases, which can be attributed to the more reasonable and general formulation for seamlessly integrating the sparse coding and neighborhood propagation into a unified framework so that the sparse reconstruction error and the classification error can be minimized simultaneously. Note that SIS-LP method works well in some cases and obtains the better results than other remaining methods on the AR face datasets. SparseNP also performs in some cases and delivers higher accuracies than others on Yale, Indian male face and ORL face databases. The results of GFHF, LLGC, SLP and LNP are comparative with each other in most cases.



(a) MIT face (b) ORL face (c) Yale face (d) Indian male face (e) AR male face (f) AR female face

Figure 5. Image examples of face databases.

TABLE IV. PERFORMANCE COMPARISON OF EACH ALGORITHM UNDER DIFFERENT SETTINGS BASED ON SIX FACE DATABASES.

Setting / Method	Yale	AR-female	AR-male	Indian-male	MIT	ORL
	Mean/Std/Time	***Mean/Std /Time***	***Mean/Std/Time***	***Mean/Std/Time***	***Mean/Std/Time***	***Mean/Std/Time***
SparseNP	63.82/6.88/0.030	30.36/8.82/2.104	32.85/9.26/2.141	57.54/8.71/0.154	88.43/5.97/	89.52/6.25/0.114
SLP	63.04/7.34/0.026	29.46/8.65/1.081	31.99/9.32/1.102	56.38/9.62/0.109	39.86//19.1/	88.88/6.18/0.091
LNP	62.16/7.45/0.026	29.76/7.87/0.899	31.65/8.57/0.934	55.91/8.73/0.099	88.43/5.97/	87.78/6.79/0.083
LLGC	51.24/7.18/0.025	25.91/6.47/1.035	27.88/6.80/1.058	46.90/5.73/0.101	67.91/13.5/	78.63/6.05/0.087
LapLDA	60.33/6.58/0.374	37.42/8.15/0.984	41.19/8.90/1.018	54.52/7.07/0.501	67.49/13.3/	85.51/4.54/0.478
GFHF	62.72/7.50/0.025	29.57/8.42/0.869	32.13/8.95/0.901	56.53/8.74/0.099	89.17/4.86/	88.67/6.45/0.080
CD-LNP	57.64/7.65/0.027	22.52/4.44/1.059	23.64/4.70/1.082	51.82/7.90/0.115	94.94/3.98/	82.20/5.89/0.098
SIS-LP	61.78/15.7/1.854	62.37/19.3/102.9	65.30/18.1/104.1	51.56/17.6/8.221	51.72/21.2/	82.58/14.9/7.171
Adaptive-NP	**73.28**/10.5/0.038	**85.61**/11.8/2.301	**86.20**/11.2/2.369	**59.93**/12.7/0.246	**97.74**/2.99/	**92.13**/6.00/0.190

TABLE V. PERFORMANCE COMPARISON OF EACH ALGORITHM UNDER DIFFERENT SETTINGS BASED ON THREE UCI DATABASES

Setting / Method	SPECT-Heart	Balance-scale	Statlog
	Mean/Std/Time	***Mean/Std/Time***	***Mean/Std/Time***
SparseNP	61.36/1.57/0.060	71.22/3.20/0.401	75.68/4.24/16.34
SLP	63.71/2.14/0.030	70.86/4.21/0.166	72.83/16.4/4.232
LNP	62.87/2.53/0.027	69.57/3.34/0.140	73.03/5.53/3.403
LLGC	61.91/2.25/0.030	69.69/3.74/0.171	70.63/8.63/4.110
LapLDA	55.19/14.9/0.015	72.55/4.40/0.048	76.53/2.56/0.419
GFHF	63.70/2.10/0.026	69.71/4.00/0.141	71.75/4.08/3.406
CD-LNP	64.42/3.31/0.032	65.07/3.03/0.173	73.78/4.84/3.861
SIS-LP	65.42/6.62/1.142	56.63/7.02/4.808	66.21/3.92/164.3
Ours	**82.51**/3.74/0.031	**86.85**/5.24/0.203	**80.15**/2.94/6.703

C. Classification on UCI Datasets

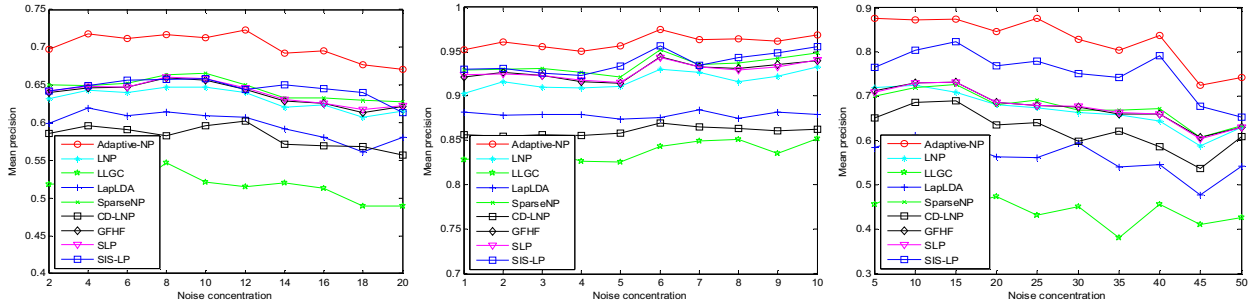
This study tests the recognition power of our Adaptive-NP on three UCI datasets. We evaluate the performances of our model and compare with the others using three UCI datasets, including a Statlog database, a Balance Scale database and a SPECT Heart database. Table III describes the information about these three datasets. Table V summarizes the averaged accuracy (%), standard deviation (%) and running time (s) under various numbers of labeled training set, from which we can clearly see that better results have been delivered by our Adaptive-NP due to the formulation of seamlessly integrates the sparse coding and neighborhood propagation into a unified model, which clearly differs from traditional label propagation methods that explicitly separate graph construction and label propagation into two independent steps. For enhancing sparse coding, we regularize the L2,1-norm constraint on the coding coefficients and the reconstruction error at the same time for delivering the more accurate and robust representations.

D. Robustness Analysis

Note that our proposed Adaptive-NP by joint L2,1-norm regularized sparse coding for adaptive weight construction and robust adaptive label prediction, so we would like to evaluate our algorithm for dealing with the cases with noise corruptions, along with illustrating the comparison results.

In this case study, we examine each method to recognize the face images under corrupted image pixels with varying noise concentration values. Three real-world face image databases, i.e., ORL face database, Indian male face database and Yale face database, are employed. For each database, random Gaussian noise is manually added to given training

data by $Data = Data + \sqrt{Var} \times randn(size(Data))$, where var is noise concentration. We prepare the following experimental settings for evaluations. That is, we aim at fixing the number of training face images from each database as 9, 7 and 5 respectively, and add random noise to the training sets. The evaluation results on the noisy cases are shown in Figure 6 (a-c), where the horizontal axis denotes the noise concentration and the vertical axis is the mean accuracy averaged over 20 times random splits of training/test samples. We have the following observations. The overall performance of each algorithm is decreased when the level of noise is increased from low to high level, but it should be noted that the result of our method go down slower than other methods. In other words, our method can obtain a relatively promising and stable superiority performance, that is, our proposed method is more robust to noise when the noise level is increased.



(a) Indian male face (b) ORL face (c) Yale face

Figure 6. Classification results of each method under different random noise concentrations on three face datasets.

E. Parameter Sensitive Analysis

We discuss the effects of different parameter selections on the classification performance of our Adaptive-NP method in this section. We illustrate the results by respectively fixing one of the two parameters (α, β) and use the approach of grid search to explore the effects. For each pair of parameters, we average the results based on 15 random splits with varied parameter α and β from $\{10^{-9}, 10^{-7}, ...10^{-1}, 10^{1}, ..., 10^{7}, 10^{9}\}$.

Due to the page limitation, we only use the Indian females face dataset (totally 22 subjects with 242 faces as a whole) as an example and we choose 9 samples of each class as training set. The parameter selection results are illustrated in Figure 7, where the vertical axis is the mean accuracy. We can observe from the results that our Adaptive-NP performs well in a wide range of parameters in each group, that is, our method is not very sensitive to the model parameters. Note that the above parameter settings are used in the face recognition simulations. It is also worth noting that similar findings for the parameter selections can be found from other datasets in most cases, but the results will not be shown due to page limitation.

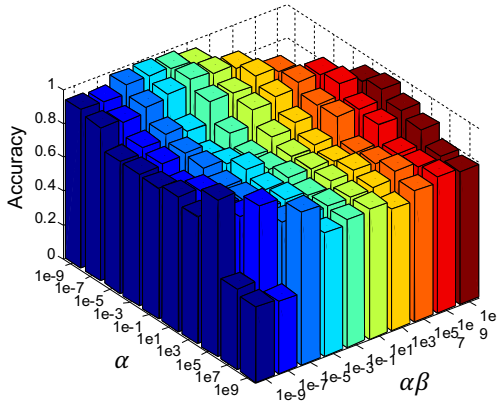


Figure 7. Parameter sensitivity analysis on the Indian face dataset.

V. Concluding Remarks

We have discussed the adaptive transductive label propagation via joint L2,1-norm regularized sparse coding for predicting the labels of samples. Different from virtually all the existing label propagation methods that pre-computes the edge weights separately before the label prediction, our algorithm explicitly incorporates the adaptive edge weights construction by L2,1-norm based sparse coding into the neighborhood propagation framework to form a unified model that can ensure the sparse reconstruction and label reconstruction errors can be jointly minimized so that the prediction results will be more accurate. Besides, the regularized L2,1-norm can make the process of sparse coding robust to noise and can also potentially reduce the sparse reconstruction error. The convergence behavior of our method is also theoretically and experimentally studied.

We have evaluated our algorithm for visual observation of weights and the quantitative comparison of data classification with several related label propagation models. Visualization of graph weights show that the true subspaces can be accurately discovered in most cases, which would potentially improve the label prediction power. Classification on real image databases and standard UCI datasets also demonstrates that remarkable results can be delivered by our technique, compared with eight related methods. In addition, we will investigate to extend this model to out-of-sample scenario for handling new data. Also, extending our method to other related application areas, e.g., image segmentation and retrieval, is also worth studying.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (61402310,61672365, 61672364 and 61373093), Major Program of Natural Science Foundation of Jiangsu Higher Education Institutions of China (15KJA520002), Special Funding of China Postdoctoral Science Foundation (2016T90494), Postdoctoral Science

Foundation of China (2015M580462), Postdoctoral Science Foundation of Jiangsu Province of China (1501091B), Natural Science Foundation of China Jiangsu Province (BK20140008, BK20141195) and the Graduate Student Innovation Project of Jiangsu Province of China (SJZZ16_0236, SJZZ15_0154). Dr. Zhao Zhang is the corresponding author of this paper.

REFERENCES

[1] M. Haji, and H. A. Toliyat, "Pattern recognition," In: *Proceeding of Electric Machines and Drives Conference*, pp. 899-904, 2001.

[2] J. Han, M. Kamber. (2000). Data Mining Concepts Models Methods & Algorithms. (2nded.) [Online]. Available from http://grail.csuohio.edu.

[3] A. Blum and S. Chawla, "Learning from Labeled and Unlabeled Data using Graph Mincuts," In: *Proceeding of 18th International Conference on Machine Learning*, San Francisco, USA, pp. 19-26, 2001.

[4] F. Wang and C. S. Zhang, "Label propagation through linear neighborhoods," *IEEE Trans. on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 985-992, Jan 2006.

[5] Z. Zhang, M. B. Zhao and T. W. S. Chow, "Graph based Constrained Semi-Supervised Learning Framework via Label Propagation over Adaptive Neighborhood," *IEEE Trans. on Knowledge and Data Engineering*, vol. 27, no. 9, pp. 2362-2376, 2015.

[6] C. Zhang, S. Wang and D. Li, "Prior class dissimilarity based linear neighborhood propagation," *Knowledge-Based Systems*, vol. 83, pp. 58-65, 2015.

[7] F. Nie, S. Xiang and Y. Liu, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing & Applications*, vol.19, no. 4, pp. 549-555, 2010.

[8] H. Tang, T. Fang and P. F. Shi, "Laplacian linear discriminant analysis," *Pattern Recognition*, vol. 39, no. 1, pp. 136-139, 2006.

[9] X. Zhu, Z. Ghahramani and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," In: *Proceeding of the International Conference on Machine Learning*, pp. 912-919, 2003.

[10] F. Zang and J. S. Zhang, "Label propagation through sparse neighborhood and its applications," *Neurocomputing*, vol. 97, no. 1, pp. 267-277, 2012.

[11] M. Belkin and P. Niyogi, "Semi-Supervised Learning on Riemannian Manifolds," *Machine Learning*, vol. 56, no. 1, pp. 209-239, 2004.

[12] Z. Zhang, W. M. Jiang, F. Z. Li, L. Zhang, M. B. Zhao, and L. Jia "Projective Label Propagation by Label Embedding," In: *Proceeding of the International Conference on Computer Analysis of Images and Patterns*, vol. 9257, pp. 470-481, Sept 2015.

[13] C. P. Hou, F. P. Nie, X. L. Li, D. Y. Yi and Y Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," *IEEE Trans. on Cybernetics*, vol. 44, no. 6, pp. 793-804, 2014.

[14] F. Nie, H. Huang, X. Cai and C. Ding, "Efficient and robust feature selection via joint L2,1-norms minimization," In: *Proceedings of Neural Information Processing Systems*, British Columbia, Canada, pp. 1813-1821, 2010.

[15] S. Z. Yang, C. P. Hou, F. P. Nie and Y. Wu, "Unsupervised maximum margin feature selection via L2,1-norm minimization," *Neural Computing & Applications*, vol. 21, no. 7, pp. 1791-1799, 2012.

[16] Z. Zhang, L. Zhang, M. B. Zhao, W. M. Jiang, Y. C Liang and F. Z. Li, "Semi-Supervised Image Classification by Nonnegative Sparse Neighborhood Propagation," In: *Proceeding of the ACM International Conference on Multimedia Retrieval*, vol. 4, no. 6, pp. 13-14, 2015.

[17] J. Wang, "Locally Linear Embedding," *Springer Berlin Heidelberg*, vol. 8215, no. 2, pp. 203-220, 2012.

[18] L. Wang, Y. Zhang and J. Feng, "On the Euclidean Distance of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.27, no.8, pp.1334-1342, 2005.

[19] A. Martinez and R. Benavente, "The AR Face Database," CVC Technical Report, 1998.

[20] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Scholkopf, "Learning with Local and Global Consistency," In: *Neural Information Processing Systems*, vol. 16, no. 4, pp. 321-328, March 2004.

[21] H. Cheng and Z. Liu, "Sparsity induced similarity measure for label propagation," In: *Proceeding of IEEE International Conference on Computer Vision*, vol. 30, pp.317-324, 2009.

[22] Y. N. Hai, S. Y. Yuan and H. Ran, "Label Propagation Algorithm Based on Non-negative Sparse Representation," In: *proceeding of the 2010 international conference on Life system modeling and simulation and intelligent computing*, vol. 6330, no. 2, pp. 348-357, 2010.

[23] Z. Zhang, M. Zhao and T. W. S. Chow, "Label propagation and soft-similarity measure for graph based Constrained Semi-Supervised Learning," In: *Proceeding of International Joint Conference on Neural Networks*, Beijing, China, pp. 2927-2934, July 2014.

[24] J. Wright, A. Ganesh and Z. Zhou, "Robust face recognition via sparse representation," In: *Proceeding of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1-2, Sept 2008.

[25] H. Wang, F. Nie and W. Cai, "Semi-supervised Robust Dictionary Learning via Efficient l-Norms Minimization," In: *Proceeding of the IEEE International Conference on Computer Vision*, 2013.

[26] F. Zang, J.Zhang, "Label propagation through sparse neighborhood and its applications," *Neurocomputing*, vol. 97, no. 1, pp. 267-277, 2012.

[27] Y. Yang, H. T. She and Z. Ma, "L2,1-Norm Regularized Discriminative Feature Selection for Unsupervised," In: *International Joint Conference on Artificial Intelligence*, pp. 1589-1594, 2011.

[28] F. Nie, H. Huang and X. Cai, "Efficient and Robust Feature Selection via Joint L2,1-Norms Minimization," In: *Proceeding of Neural Information Processing Systems*, British Columbia, Canada, pp. 1813-1821, Jan 2010.

[29] Z. Zhang, Y. Zhang, F. Z. Li, M. B. Zhao, L. Zhang and S.C. Yan, "Discriminative Sparse Flexible Manifold Embedding with Novel Graph for Robust Visual Representation and Label Propagation." *Pattern Recognition*, vol.61, pp.492-510, Jan 2017.

[30] Y. Pati, R. Rezaiifar and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," In: *Proceeding of the Annual Asilomar Conference on Signals,* Pacific Grove, vol. 1, pp. 40-44. 1993.

[31] L. Gao, J. Song, and F. Nie, "Optimal graph learning with partial tags and multiple features for image and video annotation," In: *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4371-4379, June 2015.

[32] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," In: *Proceeding of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 977-986, 2014.

[33] S. Xiang, F. Nie, and C. Zhang. "Semi-Supervised Classification via Local Spline Regression," *IEEE Transactions on Software Engineering*, vol. 32, no.11, pp. 2039-2053, 2010.