

Learning by Propagability

Bingbing Ni Shuicheng Yan Ashraf Kassim
 Loong Fah Cheong
 Department of Electrical and Computer Engineering
 National University of Singapore
 Singapore, 117576
 {g0501096, eleyans, eleashra, elec1f}@nus.edu.sg

Abstract

In this paper, we present a novel feature extraction framework, called learning by propagability. The whole learning process is driven by the philosophy that the data labels and optimal feature representation can constitute a harmonic system, namely, the data labels are invariant with respect to the propagation on the similarity-graph constructed by the optimal feature representation. Based on this philosophy, a unified formulation for learning by propagability is proposed for both supervised and semi-supervised configurations. Specifically, this formulation offers the semi-supervised learning two characteristics: 1) unlike conventional semi-supervised learning algorithms which mostly include at least two parameters, this formulation is parameter-free; and 2) the formulation unifies the label propagation and optimal representation pursuing, and thus the label propagation is enhanced by benefiting from the graph constructed with the derived optimal representation instead of the original representation. Extensive experiments on UCI toy data, handwritten digit recognition, and face recognition all validate the effectiveness of our proposed learning framework compared with the state-of-the-art methods for feature extraction and semi-supervised learning.

1 Introduction

Learning tasks (e.g., classification and regression) in computer vision research often encounter two interactional issues, namely high feature dimensionality and insufficient training data. Thus it is often necessary to transform the high dimensional data into a low dimensional feature space. Many algorithms have been proposed for this purpose [11], and among them, Principal Component Analysis (PCA) [7] and Linear Discriminant Analysis (LDA) [1] are the two most popular ones.

Without class information, PCA simply attempts to project the high dimensional data into a low dimension linear subspace according to the criterion that the preserved dimension captures the most of the variability of the original data. On the other hand, if the class label information is provided, LDA aims to maximizing the inter-class scatter and at the same time minimizing the intra-class scatter. Unlike PCA and LDA whose objective functions are not directly related with the final classification performances, the recently proposed algorithm, called Neighborhood Component Analysis (NCA) [5], was designed to directly optimize the expected leave-one-out classification error on the training data.

However, the second issue of insufficient training data often degenerates the performances of these algorithms. Despite of the insufficient labeled training samples, samples without labels are relatively much easier to obtain. Known as semi-supervised learning, a variety of algorithms have been proposed to improve the algorithmic learnability by using additional unlabeled training data. Most algorithms of this category are based on graph [2], e.g., LapSVM and LapRLS [3], and local and global consistency [12], etc. Although these algorithms have proved to be effective and achieved great success in many applications, much more research is still demanded for semi-supervised learning owing to the following observations: 1) almost all these algorithms are based on graphs defined on the original high dimensional feature space, which is unnecessary to be best for characterizing the similarities of the sample pairs, since unfavorable features and noise may exist within the original representation; 2) most previous semi-supervised algorithms were directly designed for two-class classification problems, and extra process is required for handling multi-class classification problems; 3) most semi-supervised learning algorithms result in optimization problems with extra parameters which lack elegant ways for automatic setting; and 4) the objective function does not directly target the algorithmic discriminating power, instead

is often the trade-off between the algorithmic discriminating power and the label smoothness.

Recently, there existed one attempt to tackle the second issue as above-mentioned, an algorithm, called Semi-supervised Discriminant Analysis (SDA) [4], was proposed to learn a linear subspace by taking the graph constraints into consideration. However, SDA is still based on the graph defined on the original high dimensional feature space, which also has extra parameters to determine.

In this paper, we present a unified parameter-free learning framework for both supervised and semi-supervised learning configurations, where the above four issues are resolved within in the scenario of semi-supervised configuration. The whole framework is based on the philosophy that the class labels and the optimal feature representation can construct a harmonic system, namely the class labels are invariant with respect to the propagation on the similarity-graph defined on the optimal representation. For supervised learning tasks, the pursuing of the optimal feature transformation matrix targets to build such a harmonic system, while for semi-supervised configurations, it achieves the propagation of the labels from the labeled data to the unlabeled data in the same time. Based on this unified framework, an iterative procedure is presented for seeking the solution.

Here we would like to highlight some aspects of our proposed unified learning framework of Learning by Propagability (FLP):

1. For label propagation, the graph is finally constructed on the derived optimal low dimensional feature space, which better characterizes the data similarity than the graph built directly on the original high dimensional feature space.
2. FLP is parameter-free and hence the results are more stable and not sensitive to any factors. While for almost all previous semi-supervised learning algorithms, there exist at least two parameters: one parameter is used to determine the neighbors for graph construction, and the other is used to balance the importance between the discriminating power and the geometric label smoothness.
3. Unlike most previous semi-supervised algorithms which work for two-class classification problems, FLP directly works on general multi-class classification problems.
4. The objective of FLP still characterizes the discriminating power, and hence it is naturally better than other semi-supervised learning algorithms which make trade-offs between discriminating power and geometric label smoothness.

This paper is organized as follows. Section 2 introduces the learning by Propagability framework followed by the iterative procedure for seeking the solution in Section 3. Section 4 presents the extensive experimental results. Section 5 concludes this paper followed by the discussion of future work.

2 Unified Formulation for Learning by Propagability

For notational consistency, in this paper, we use lower case alphabets to represent scalars, lower case bold alphabets to represent vectors, upper case alphabets to represent matrix and upper case bold alphabets to represent data sets.

For a classification problem, we assume that the training sample data (both labeled and unlabeled) are given as $\mathbf{X} = [\mathbf{X}^l, \mathbf{X}^u] = [\mathbf{x}_1, \dots, \mathbf{x}_{N_1}, \mathbf{x}_{N_1+1}, \dots, \mathbf{x}_N]$, where $\mathbf{x}_i \in \mathbb{R}^m$, N is the total number of training samples and N_1 is the number of labeled training samples. We use $\mathbf{\Pi} = [\mathbf{\Pi}^l, \mathbf{\Pi}^u] = [\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{N_1}, \tilde{\boldsymbol{\pi}}_{N_1+1}, \dots, \tilde{\boldsymbol{\pi}}_N]$ to represent the corresponding label/class probability distribution information, namely, for labeled data $\mathbf{X}^l = [\mathbf{x}_1, \dots, \mathbf{x}_{N_1}]$, each corresponding K dimensional vector $\boldsymbol{\pi}_i = [\pi_i^1, \dots, \pi_i^k, \dots, \pi_i^K]^T, i = 1, \dots, N_1$ has a 1-of- K representation in which a particular element π_i^k is equal to 1 and all other elements are equal to 0, indicating its class label; for unlabeled data $\mathbf{X}^u = [\mathbf{x}_{N_1+1}, \dots, \mathbf{x}_N]$, each corresponding K dimensional vector $\tilde{\boldsymbol{\pi}}_i = [\tilde{\pi}_i^1, \dots, \tilde{\pi}_i^k, \dots, \tilde{\pi}_i^K]^T, i = N_1 + 1, \dots, N$ represents its probability distribution of belonging to each class, with all its elements taking some real value in the interval $[0, 1]$, which are unknown and need to be updated by some learning algorithm. Therefore, for both labeled and unlabeled data, the sum-to-1 constraints are always satisfied, i.e., $\sum_{k=1}^K \pi_i^k = 1, \forall i = 1, \dots, N_1$ and $\sum_{k=1}^K \tilde{\pi}_i^k = 1, \forall i = N_1 + 1, \dots, N$. Note for notational convenience, we will ignore the notational difference between all the $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\pi}}$ in the rest of the paper, i.e. $\boldsymbol{\pi}_i = \tilde{\boldsymbol{\pi}}_i, i = N_1 + 1, \dots, N$. Since in practice the feature dimension m is often very large and unfavorable features may exist for certain task, it is usually necessary to transform the original high-dimensional data into a low-dimensional feature space for facilitating and enhancing subsequent process.

2.1 Motivations

Dimensionality reduction under supervised or semi-supervised learning configurations has attracted much attention in the past decade. Our work in this paper is motivated from the following three aspects.

2.1.1 Harmoniousness between Label and Representation:

For supervised learning problems, the harmoniousness between label and representation is easy to characterize, e.g., Linear Discriminant Analysis (LDA) [1] searches for a low-dimensional representation that minimizes the intra-class scatter and at the same time maximizes the inter-class scatter. However, for semi-supervised learning with unlabeled data, the criterion based on class labels is not applicable in modeling the harmoniousness between the class label and the representation any more. Instead, in the literature many semi-supervised learning algorithms without feature extraction, namely with only label propagation, are proposed for utilizing these unlabeled data. These algorithms essentially can be considered as special classification methods instead of learning processes. These semi-supervised learning algorithms are mostly based on graphs built from the original high-dimensional features, which may include unfavorable features and noises. Intuitively the semi-supervised learning may benefit from the feature extraction process which may remove these unfavorable features and noises.

In this work, we aim to conduct label propagation and feature extraction simultaneously. More specifically speaking, a low-dimensional feature space and the class labels for the unlabeled data are learned within a unified formulation, and the ultimate target is that the class labels are invariant after the propagation on the graph constructed on the derived optimal low-dimensional feature space.

2.1.2 Parameter-free Semi-supervised Learning:

Semi-supervised learning has proved to be effective in utilizing the unlabeled data for enhancing the algorithmic performance. However, a common issue suffered by most previous semi-supervised learning algorithms is that they have extra parameters within the formulation, and these parameters may greatly affect the final algorithmic performance. Moreover, the optimal parameters for different algorithms and different data sets may be different. It is desirable to develop a parameter-free semi-supervised learning algorithm for the purpose of robust systems.

In this work, the unlabeled data are used to propagate expected label information to the labeled data, such that the prediction confidence for the labeled data is maximized, and the derived formulation is parameter-free.

2.1.3 Consistency between Feature Extraction and Classification:

Dimensionality reduction or feature extraction has been widely studied in computer vision literature, but most algorithms are based on certain intuitive motivations, which

do not directly maximize the consequent classification accuracy. As a consequence, the feature extraction and final classification are separated as two independent steps, and thus the overall optimum cannot be guaranteed. All unsupervised learning algorithms belong to this category. Among the few exceptions, Neighborhood Component Analysis (NCA) [5] learns a Mahalanobis distance measure to be used in the KNN classification algorithm, and the derived transformation is optimal in the sense of soft k -nearest-neighbor classification. However, NCA is designed for supervised learning, and cannot be directly used for semi-supervised learning.

In this work, we aim to conduct feature extraction which is directly based on the classification criteria used in the classification stage and applicable for both supervised and semi-supervised configurations.

2.2 Problem Formulation

In this subsection, we introduce our solution to feature extraction based on graph propagability. The task of feature extraction is to find a matrix $A = [A_1, A_2, \dots, A_d] \in \mathbb{R}^{m \times d}$ to transform the original high-dimensional data \mathbf{x} into a low-dimensional form $\mathbf{y} \in \mathbb{R}^d$ (usually $d \ll m$) as

$$\mathbf{y} = A^T \mathbf{x}. \quad (1)$$

2.2.1 Similarity Graph:

Based on the low-dimensional representation transformed by matrix A , a similarity graph $\mathbf{G} = \{\mathbf{\Pi}, \mathbf{W}\}$ with vertices as label set $\mathbf{\Pi}$ and edge weights as matrix W can be defined as

$$W_{ij} = \exp(-\|A^T \mathbf{x}_i - A^T \mathbf{x}_j\|^2), \forall i \neq j. \quad (2)$$

2.2.2 Propagation on Graph:

The underlying philosophy of classification tasks is to predict the class label of a new datum based on the label observations of other data, namely the label of the new datum is propagated from other observed data. The weight W_{ij} measures the relative similarity between different sample pairs. Intuitively, for the datum \mathbf{x}_i , the larger the weight W_{ij} is, the more contribution the label of sample \mathbf{x}_j offers to the prediction of the label for sample \mathbf{x}_i . Since the sum of the predicted class probability vector of the sample \mathbf{x}_i should be 1, the propagation coefficients are normalized as

$$\begin{aligned} p_{ij} &= \frac{W_{ij}}{\sum_{j \neq i} W_{ij}} = \frac{\exp(-\|A\mathbf{x}_i - A\mathbf{x}_j\|^2)}{\sum_{j \neq i} \exp(-\|A\mathbf{x}_i - A\mathbf{x}_j\|^2)}, \\ p_{ii} &= 0. \end{aligned} \quad (3)$$

Then, we have $\sum_j p_{ij} = 1, \forall i$.

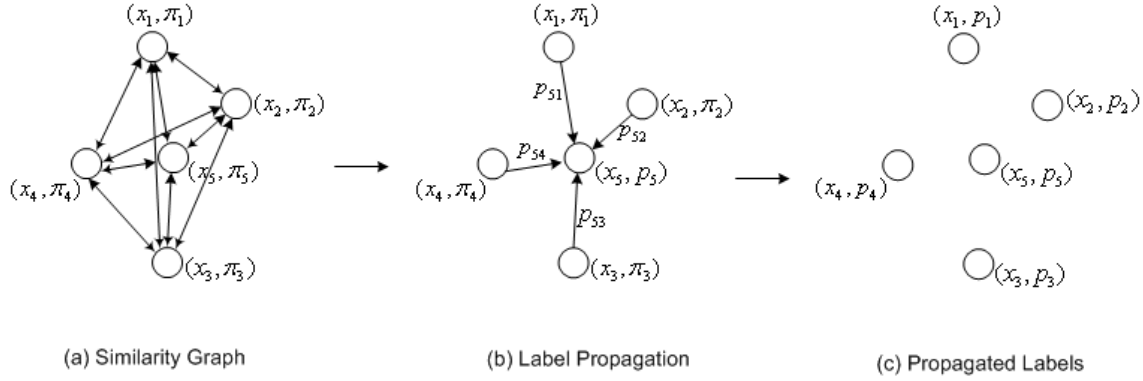


Figure 1. An illustration of the class label propagation on the normalized similarity graph. (a) shows the constructed similarity graph. In (b), the surrounding uncertain labels (π_1, \dots, π_4) are propagated to p_5 . After label propagation, the uncertain labels (π_1, \dots, π_4) are updated in (c) via graph harmoniousness.

Here we denote $\mathbf{p}_i = [p_i^1, \dots, p_i^k, \dots, p_i^K]^T, i = 1, \dots, N$ as the propagated class probability vector for sample \mathbf{x}_i , then we have

$$p_i^k = \sum_{j \neq i} p_{ij} \pi_j^k, k = 1, 2, \dots, K. \quad (4)$$

Note that the propagated label satisfies the conditions 1) $p_i^k \geq 0$, and 2) $\sum_{k=1}^K p_i^k = \sum_{k=1}^K \sum_{j \neq i} p_{ij} \pi_j^k = \sum_{j \neq i} p_{ij} \{ \sum_{k=1}^K \pi_j^k \} = 1$. Figure 1 gives an illustration on the class label propagation on the normalized similarity graph.

2.2.3 Objective Function:

Before formally describing the objective function for learning by propagability, we first define a concept called *graph harmoniousness*.

Definition For a graph $\mathbf{G} = \{\mathbf{\Pi}, \mathbf{W}\}$, the graph harmoniousness is defined as $H(\mathbf{G}) = \frac{1}{2} \sum_{i=1}^{N_1} \|\mathbf{p}_i - \boldsymbol{\pi}_i\|^2$, where \mathbf{p}_i is defined in Eqn. (4).

The graph harmoniousness characterizes the label propagability on the graph, which on the other hand measures the classification capability of the derived feature representation based on the soft nearest neighbor method. Note that the supervised configuration can be considered as a special case of semi-supervised configuration by setting $N_1 = N$, namely all the samples are with labels. For simplicity, we only present the formulation for semi-supervised learning afterwards.

For semi-supervised learning, we propose to simultaneously learn the transformation matrix A and the undetermined labels $\mathbf{\Pi}^u = [\boldsymbol{\pi}_{N_1+1}, \dots, \boldsymbol{\pi}_N]$ by seeking the best harmoniousness, namely,

$$\arg \min_{A, \mathbf{\Pi}^u} \{H(\mathbf{G}) = \frac{1}{2} \sum_{i=1}^{N_1} \|\mathbf{p}_i - \boldsymbol{\pi}_i\|^2\}, \quad s.t. \quad (5)$$

1. $\pi_i^k \geq 0, \forall k, i = N_1 + 1, \dots, N;$
2. $\sum_{k=1}^K \pi_i^k = 1, i = N_1 + 1, \dots, N.$

We shall note that 1) we do not include the terms on the unlabeled data in the objective function regarding that the number of the unlabeled data is much larger than the labeled data and therefore the term on the unlabeled data will dominant the objective function, which in turn will degrade the algorithmic performance. 2) We do not use the cross-entropy as in [5] since some π_j^k on the unlabeled data can be or nearly be zero, which makes the numerical computation unstable. We also would like to highlight some aspects of this formulation: 1) this formulation is for multi-class classification problems, not only for two-class problems, and there exist two constraints for the class labels; and 2) it integrates feature extraction and label propagation within a unified formulation.

2.3 Iterative Solution

The objective function $H(\mathbf{G})$ is nonlinear and the optimization problem is constrained, and hence there does not

exist a closed-form solution. Naturally, we present a procedure to optimize the transformation matrix A and the under-terminated labels Π^u iteratively.

2.3.1 Optimize A for given Π^u :

Assume that the values of Π^u are known, the optimization problem defined in (5) is casted as:

$$\arg \min_A \{H(A) = \frac{1}{2} \sum_{i=1}^{N_1} \|\pi_i - \mathbf{p}_i\|^2\} \quad (6)$$

It is an unconstrained optimization problem with non-convex objective function, any gradient descent based optimization method could converge to a local minimum. The gradient of the cost function is calculated as,

$$\frac{\partial H}{\partial A} = - \sum_k \sum_i (\pi_i^k - p_i^k) \frac{\partial p_i^k}{\partial A} \quad (7)$$

$$\begin{aligned} &= - \sum_k \sum_i (\pi_i^k - p_i^k) (-2A) \left\{ \left[\sum_{j \neq i} \pi_j^k p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right. \\ &\quad \left. - \left[\sum_{j \neq i} \pi_j^k p_{ij} \right] \left[\sum_{j \neq i} p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right\} \\ &= 2A \sum_k \sum_i (\pi_i^k - p_i^k) \left\{ \left[\sum_{j \neq i} \pi_j^k p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right. \\ &\quad \left. - \left[p_i^k \right] \left[\sum_{j \neq i} p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right\} \quad (8) \end{aligned}$$

$$\begin{aligned} &= 2A \sum_k \sum_i (\pi_i^k - p_i^k) \left\{ \left[\sum_{j \neq i} \pi_j^k p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right. \\ &\quad \left. - \left[\sum_{j \neq i} p_i^k p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right] \right\} \quad (9) \end{aligned}$$

$$= 2A \sum_{k,i,j} (\pi_i^k - p_i^k) (\pi_j^k - p_j^k) p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T, \quad (10)$$

where $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$.

2.3.2 Optimize Π^u for given A :

When the transformation matrix A is fixed, the optimization problem defined in (5) is a constrained optimization problem. To make it more computationally tractable, we convert this constrained optimization problem into an unconstrained one by setting

$$\pi_i^k = \frac{\exp(c_i^k)}{\sum_{m=1}^K \exp(c_i^m)}, i = 1, \dots, N, k = 1, \dots, K, \quad (11)$$

and then the constraints defined in (5) are naturally satisfied. Similarly, there does not exist closed-form solution, and we utilize the gradient descent method to obtain a local

optimum. The gradient with respect to the new variable c_i^k is

$$\begin{aligned} \frac{\partial H}{\partial c_i^k} &= \frac{\partial (\frac{1}{2} \sum_m \|\pi_i^m - p_i^m\|^2)}{\partial c_i^k} \\ &\quad + \frac{\partial (\frac{1}{2} \sum_{j \neq i} \sum_m \|\pi_j^m - p_j^m\|^2)}{\partial c_i^k} \quad (12) \end{aligned}$$

$$\begin{aligned} &= \sum_m (\pi_i^m - p_i^m) \frac{\partial \pi_i^m}{\partial c_i^k} \\ &\quad - \sum_{j \neq i} \sum_m (\pi_j^m - p_j^m) p_{ij} \frac{\partial \pi_i^m}{c_i^k}. \quad (13) \end{aligned}$$

The above two steps are iteratively conducted to obtain the stepwise result (A_t, Π_t^u) until satisfying the following stop criteria,

$$\begin{cases} \|A_t - A_{t+1}\| < \varepsilon md, \\ \|\Pi_t^u - \Pi_{t+1}^u\| < \varepsilon(N - N_1)K, \end{cases} \quad (14)$$

where ε is a manually defined threshold and empirically set to be 10^{-6} in this work.

3 Algorithmic Analysis

In this section, we discuss two related works, and then analyze the algorithmic convergency property.

3.1 Related Works

Zhu et al. [13] proposes an approach for semi-supervised learning based on a Gaussian random field model defined with respect to a weighted graph representing labeled and unlabeled data. Our work is different from this work in many aspects: 1) Zhu's work propagates the labels based on the graph defined on the original feature representation, while our work seeks the best feature representation and labels for unlabeled data simultaneously; 2) Zhu's work was designed for two-class classification problems, while our work is directly designed for multi-class classification problems; and 3) Zhu's work is applicable for semi-supervised learning, while our work is general and applicable for both semi-supervised and supervised learning configurations. In our experiment part, we shall further show the advantages of simultaneous feature extraction and label propagation.

To the best of our knowledge, Cai et al. [4] presented the first work for simultaneous feature extraction and semi-supervised learning (SDA). Our work has the advantages over SDA in the following aspects: 1) Similar to Zhu's work, Cai's work is also based on the graph defined on the original feature representation, and hence may suffer from the unfavorable features and noises; and 2) Cai's work is based on the LDA algorithm, the effectiveness of which is

limited by the strong assumption that the samples of each class follow some Gaussian distribution and the covariance matrices for different classes are the same; while our algorithm does not have such assumptions and hence is more general.

There also exists some recent extension on NCA by Salakhutdinov and Hinton [9], however, our work is substantially different from theirs, i.e., [9] mainly focuses on the nonlinear mapping function learning under the supervised configuration, and an extra regularization parameter is used for extension to semi-supervised learning. However, our proposed method could learn a linear transformation matrix as well as label propagation in a unified graph harmoniousness framework without any parameters.

3.2 Convergence Analysis

The optimization problem in Eqn. (5) is non-convex due to the non-convexity of the objective function, and hence we cannot guarantee that the solution will be globally optimal. Here, instead we prove that the iterative procedure will converge to a local optimum. Denote the objective function as $H(A, \mathbf{\Pi}^u) = \frac{1}{2} \sum_{i=1}^{N_1} \|\mathbf{p}_i - \boldsymbol{\pi}_i\|^2$, then we have

$$H(A_t, \mathbf{\Pi}_t^u) \geq H(A_{t+1}, \mathbf{\Pi}_t^u) \geq H(A_{t+1}, \mathbf{\Pi}_{t+1}^u). \quad (15)$$

Therefore, the objective function is non-increasing, and we have $H(A, \mathbf{\Pi}^u) \geq 0$, which means that the objective function has a lower-bound. Then we can conclude that the objective function will converge to a local optimum according to "Cauchy's criterion for convergence" [8].

4 Experiments

In this section, we systematically evaluate the effectiveness of our proposed framework of learning by propagability (FLP) under the semi-supervised configuration. First, we examine the algorithmic properties, including algorithmic convergence, advantage of the integration of label propagation and feature extraction, and parameter sensitivity. Then we give an extensive comparison of FLP with the state-of-the-art feature extraction and semi-supervised learning algorithms including Linear Discriminant Analysis (LDA) [1], Neighborhood Component Analysis (NCA) [5], and Semi-supervised Discriminant Analysis (SDA) [4], on three UCI toy data sets¹, the USPS handwritten digit dataset [6], and the CMU PIE [10] face dataset. Note that the supervised version of FLP is very similar to NCA although their objective functions are slightly different in this case, and our offline experiments show their performances are similar, therefore we only report the results of NCA in this work.

¹Available at <http://www.ics.uci.edu/mllearn/MLRepository.html>.

4.1 Experimental Configurations

The data sets used in our experiments are summarized as follows.

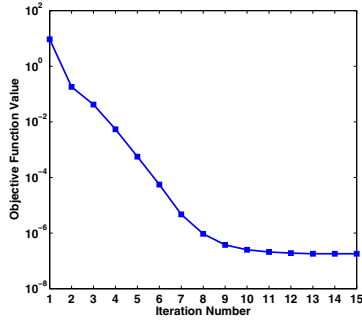
- UCI iris dataset: It includes 3 classes and 150 samples in total. The feature dimension is 4.
- UCI wine dataset. It includes 3 classes and 178 samples in total. The feature dimension is 13.
- UCI yeast dataset. It includes 8 classes and 1459 samples in total. The feature dimension is 8.
- USPS handwritten digit dataset: It includes 10 classes (e.g., 0 – 9 digit characters) and 11000 samples in total. The original data format is of 32×32 pixels. For the sake of computational efficiency as well as noise filtering, we first reduce its dimension to 128 by PCA.
- CMU PIE face dataset: The original database includes 68 individuals and totally 20000 face images with varying pose, illumination, and expression conditions. We use a subset from the pose with index 27, and it includes 43 images per person with different illumination conditions. The original image is in size of 64×64 pixels. We also first reduce the dimension to 128 by PCA.

If it is not otherwise specified, in each run, for the iris, wine, yeast, and handwritten digit, the data set is divided into three subsets. Two samples of each class are randomly selected as the labeled set, and other 20 samples of each class are randomly selected as the semi-set, and then all the remaining data are used as the testing set. For the face database, following the configuration in [4] for SDA evaluation, we use 1 sample each subject as labeled set, and hence only semi-supervised algorithms SDA and FLP can be used for the experiments. We use 20 images each subject as the semi-set for computational efficiency, and the rest images are used for testing.

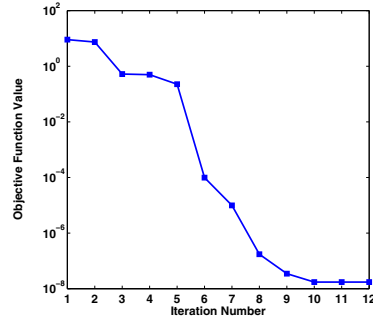
In the classification stage, the nearest neighborhood classifier is used to evaluate the recognition accuracies of LDA and SDA algorithms for both the semi-set and the testing set. NCA and our proposed FLP directly use their own probabilistic ways for final classification. The offline experiments show that after feature extraction, the performances of NCA (or FLP) with nearest neighbor and probabilistic way are similar, and hence in this work we do not report the results from nearest neighbor method for them separately.

4.2 Algorithmic Property Evaluation

In this subsection, we examine some properties of our proposed FLP and SDA: 1) convergence property, 2) advantage of integrating label propagation and feature extraction, and 3) parameter sensitivity.



(a) UCI Wine Dataset



(b) Handwritten Digit Dataset

Figure 2. Objective function value decreases along with the increase of the iteration number for the UCI wine dataset and the USPS handwritten digit dataset.

4.2.1 Algorithmic Convergency:

As proved in the previous section, the iterative procedure guarantees a local optimum solution for our objective function in Eqn. (5). In Figure 2 we show how the objective function value decreases with increasing iterations on the UCI wine dataset and USPS handwritten digit dataset. Our offline experiments show that generally FLP converges after about 6-10 iterations.

4.2.2 Advantages of Integrating Label Propagation and Feature Extraction:

Most previous semi-supervised learning algorithms only consider label propagation, and the graph is built directly based on the original feature representation. SDA is a semi-supervised algorithm for feature extraction, but it does not directly consider label propagation. In this subsection, we show the necessity to integrate the label propagation and feature extraction for better semi-supervised learning. More specifically, we compare our proposed semi-supervised FLP algorithm under the scenarios with feature extraction and without feature extraction (namely, directly set $A = I$). This means that the labels can only be propagated on the original similarity graph by simply optimizing Eqn. 5 with respect to Π^u . The comparison experiments are conducted on the UCI iris dataset and USPS handwritten digit dataset. Figure 3 shows the recognition accuracy of FLP with different reduced feature dimensions, and the results on both datasets indicate that the integration of label propagation and feature extraction can boost the performance compared with the counterpart without feature extraction.

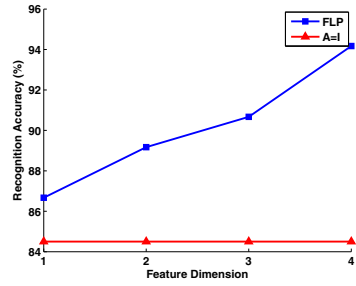
4.2.3 Sensitivity to Parameter for SDA:

Figure 4 shows the classification accuracies of SDA on USPS handwritten digit dataset and the CMU PIE face dataset with respect to different values of the parameter β , which balances the term on discriminating power and another term for label smoothness in SDA [4]. From the comparison, we can observe that the recognition accuracy can be significantly affected by the selection of different values of β , moreover, the trends of the recognition accuracy on different datasets with respect to β can be totally different. It indicates that SDA is sensitive to the selection of this parameter. As a parameter-free approach, our proposed FLP algorithm however has no such problems.

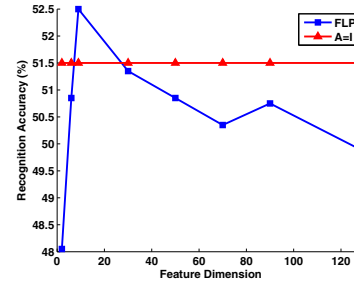
4.3 Classification Accuracy Evaluation

The detailed comparisons on recognition accuracies of different algorithms over different datasets are summarized in Table 1. For each dataset and each algorithm, we explore all possible dimensions for the transformation matrices, and then report the best result. For each configuration, we randomly generate 10 splits/runs, and the reported results are given in terms of means and standard deviations, which is of sufficient statistical importance. For SDA, different values for parameters as in Figure 4 are tested and the best result is reported. For FLP, we randomly initialize the transformation matrix A and unknown variables $c_i^k, i = N_1 + 1, \dots, N, k = 1, \dots, K$ for each run. The detailed recognition accuracies over different feature dimensions for all the evaluated algorithms are displayed in Figure 5, where the results are obtained from the UCI wine dataset and the CMU PIE face dataset.

From the results in Table 1 and Figure 5, we can have a set of interesting observations:

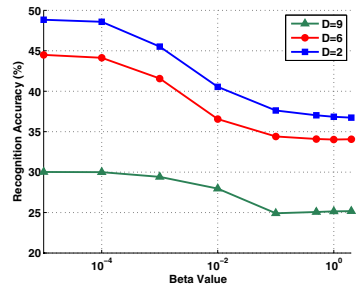


(a) UCI Iris Dataset

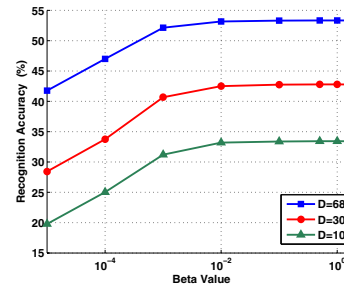


(b) Handwritten Digit Dataset

Figure 3. Recognition accuracy (on semi-set) comparison of FLP without feature extraction ($A = I$) and FLP with feature extraction of different dimensions on the UCI iris dataset and the USPS handwritten digit dataset. Note that the recognition accuracy of FLP without feature extraction is plotted as a line since it is a fixed value for a certain dataset. Note that for UCI iris dataset, feature extraction at very dimension can improve the performance, and for the USPS dataset, if the dimension is set around the number of classes (i.e., 10 in this example), the performance can also be improved.



(a) Handwritten Digit Dataset



(b) CMU PIE Face Dataset

Figure 4. Recognition accuracy (on testing set) over different β (shown as *beta* in the figure) for SDA algorithm on the USPS handwritten digit dataset and the CMU PIE face dataset. More details on the parameter β of SDA algorithm are referred to [4]. Note that three curves represent the recognition accuracies corresponding to three different reduced feature dimensions.

1. For the supervised algorithms, NCA algorithm shows to be much better than LDA in all the cases.
2. From the comparison of LDA and SDA, we can conclude that semi-supervised learning may bring extra advantages by utilizing the unlabeled data for model training. The classification accuracy is significantly improved by benefitting from these unlabeled data.
3. Our proposed semi-supervised learning algorithm FLP is the best among all the evaluated algorithms, i.e., FLP outperforms all other algorithms in terms of high average recognition accuracy and small deviations. Note that 1) we do not report the conventional *t-test* for the experimental results since the higher mean values and lower deviations achieved by FLP indicate its better performance measured by *t-test*; and 2) although

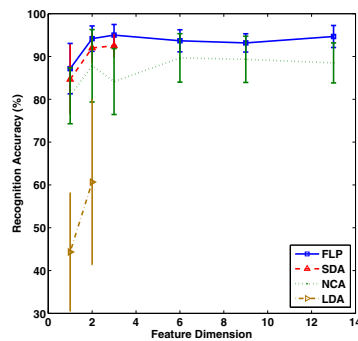
theoretically only local minimum solution is guaranteed for FLP, our statistics of the experimental results show that our algorithm practically provides good performance.

4. In this work, we do not compare our algorithm with other popular semi-supervised learning algorithms, e.g., those introduced in [13][12], since SDA has shown to be superior over these algorithms [4]. Thus in this work, we only compare our algorithm with the state-of-the-art semi-supervised algorithm, namely, SDA.

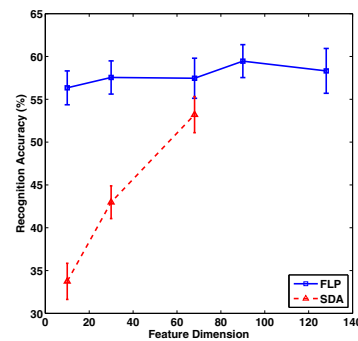
An additional experiment is performed to demonstrate the performances of our algorithm under different unlabeled:labeled partitions for the training data. For each

Table 1. Recognition accuracies (%) for different algorithms on different data sets. Note that the supervised learning methods, e.g., LDA and NCA, use labeled set for training only, while semi-supervised algorithms use both labeled set and semi-set for model learning. The results in the lines with *semi* mean the accuracies on the semi-set and those in the lines with *test* means the accuracies on the testing set. For the cells with two values, the left value is the average recognition accuracy of 10 runs, and the right one is the standard deviation.

| | | Iris | Wine | Yeast | USPS Digit | CMU PIE |
|-----|------|--------------------|--------------------|--------------------|--------------------|--------------------|
| LDA | semi | 65.49 ± 24.38 | 60.67 ± 19.31 | 15.69 ± 3.81 | 18.05 ± 3.81 | n/a |
| NCA | semi | 92.83 ± 5.61 | 87.83 ± 8.46 | 33.31 ± 6.46 | 40.21 ± 5.77 | n/a |
| SDA | semi | 88.83 ± 6.67 | 92.33 ± 3.16 | 38.44 ± 6.77 | 48.50 ± 6.44 | 53.21 ± 2.11 |
| FLP | semi | 94.50 ±4.65 | 95.00 ±2.48 | 40.88 ±7.53 | 52.50 ±6.39 | 59.46 ±1.92 |
| LDA | test | 66.91 ± 25.29 | 62.05 ± 16.27 | 19.00 ± 8.70 | 17.32 ± 2.23 | n/a |
| NCA | test | 92.98 ± 3.24 | 83.10 ± 9.70 | 32.76 ± 6.32 | 40.97 ± 4.57 | n/a |
| SDA | test | 89.41 ± 5.40 | 90.89 ± 5.39 | 37.00 ± 6.89 | 48.84 ± 4.05 | 53.33 ± 1.82 |
| FLP | test | 93.45 ±3.09 | 93.13 ±3.32 | 40.03 ±5.40 | 49.87 ±4.41 | 54.73 ±2.13 |



(a) UCI Wine Dataset



(b) CMU PIE Face Dataset

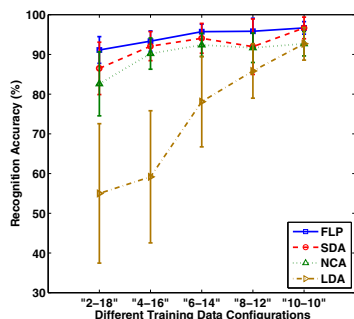
Figure 5. Recognition accuracies (on semi-set) over different feature dimensions for LDA, NCA, SDA, and FLP on the UCI wine dataset and the CMU PIE face dataset in terms of mean (data points) and standard deviation (vertical bars). Note that for LDA, the largest feature dimension can be only $K-1$ [1] and the largest feature number for SDA can be K [4], while NCA and FLP do not have this kind of constraints.

dataset, we use 20 samples for training (including both labeled and unlabeled data) and the rest for testing. We vary the ratio between the number of labeled and unlabeled training data (e.g., 2 – 18 means 2 labeled samples and 18 unlabeled samples from each class) and for each configuration, we randomly generate 10 splits. We report the classification results in terms of both means and standard deviations and compare the results from all the learning algorithms. Note we report the best result for SDA by varying its parameters as before. Fig. 6 shows the results on the UCI wine dataset and the USPS handwritten digit dataset (Note we do not show the results from all datasets due to the space limitation, however, the un-reported results are similar). We could observe that our proposed FLP algorithm consistently performs best under all different training data configurations,

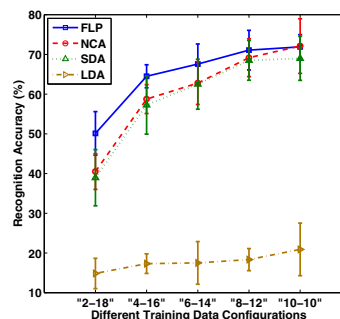
i.e., highest recognition rates with low deviations, which validate the robustness of FLP in terms of different training data configurations. Note that the result of NCA approaches FLP in the 10 – 10 setting, where the number of labeled training data is relatively high, and extra benefit from unlabeled data vanishes.

5 Conclusions and Future Work

In this paper, we have proposed a criteria to measure the optimality of a graph with class probability vectors as vertices and similarity graph constructed from the pursuing of the optimal feature representation. Then, a general learning philosophy of learning by propagability was presented for both supervised and semi-supervised config-



(a) UCI Wine Dataset



(b) Handwritten Digit Dataset

Figure 6. The recognition accuracies under different labeled:unlabeled training data partitions. Note that for each dataset, we use dimension d as $K-1$ for all algorithms (i.e., $d = 2$ for the UCI wine dataset and $d = 9$ for the USPS handwritten digit dataset). The mean values are represented by points and the standard deviations are represented by vertical bars.

urations. Simultaneous feature extraction and label propagation brought great improvement in classification accuracy compared with the state-of-the-art algorithm for semi-supervised learning tasks. We are planning to further exploit learning by propagability in three aspects: 1) to study learning by propagability under unsupervised learning configuration, namely, clustering with feature extraction; 2) to accelerate the computational speed by taking advantages of novel nonlinear optimization toolbox, since the current version of algorithm still suffers from the heavy computational cost; and 3) to study learning by propagability under the scenarios with uncertain labels.

Acknowledgment

This work was supported by MDA grant of R-705-000-018-279, Singapore.

References

- [1] P. Belhumeur, J. Hespanha, and D. Kiregeman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, 1997.
- [2] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. *COLT*, 2004.
- [3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 2006.
- [4] D. Cai, X. He, and J. Han. Semi-Supervised Discriminant Analysis. *IEEE International Conference on Computer Vision*, 2007.
- [5] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood component analysis. *Neural Information Processing Systems*, pp. 513–520, 2004.
- [6] J. Hull. A Database for Handwritten Text Recognition Research. *IEEE PAMI*, vol. 16, no. 5, pp. 550-554, 1994.
- [7] I. Jolliffe. Principal component analysis. *Springer-Verlag, New York*, 1986.
- [8] W. Rudin. Principles of Mathematical Analysis, 3rd Edition. *McGray-Hill*, 1976.
- [9] R. Salakhutdinov and G. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. *AI and Statistics*, 2007.
- [10] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *Proceeding of the European Conference on Computer Vision*, vol. 25, pp. 1615–1618, 2003.
- [11] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen. Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing. *IEEE Trans. Knowl. Data Eng*, vol. 18, no. 2, pp. 320-333, 2006.
- [12] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems* 16, 2003.
- [13] X. Zhu, Z. Ghahramani and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *International Conference on Machine Learning*, vol. 20, 2003.