Semi-Supervised Classifications via Elastic and Robust Embedding

Yun Liu,[1] **Yiming Guo,**[2] **Hua Wang,**[3] **Feiping Nie,**[4,1*] **Heng Huang**[1*]
[1]Department of Computer Science and Engineering, University of Texas at Arlington, Texas, USA
[2]Computer Science Department, Illinois Institute of Technology, Chicago, Illinois, USA
[3]Division of Computer Science, Colorado School of Mines, Colorado, USA
[4]School of Computer Science, OPTIMAL, Northwestern Polytechnical University, Xian 710072, Shaanxi, P. R. China
yunliu09@gmail.com, yguo46@hawk.iit.edu, huawang@mines.edu, feipingnie@gmail.com, heng@uta.edu

Abstract

Transductive semi-supervised learning can only predict labels for unlabeled data appearing in training data, and can not predict labels for testing data never appearing in training set. To handle this out-of-sample problem, many inductive methods make a constraint such that the predicted label matrix should be exactly equal to a linear model. In practice, this constraint might be too rigid to capture the manifold structure of data. In this paper, we relax this rigid constraint and propose to use an elastic constraint on the predicted label matrix such that the manifold structure can be better explored. Moreover, since unlabeled data are often very abundant in practice and usually there are some outliers, we use a non-squared loss instead of the traditional squared loss to learn a robust model. The derived problem, although is convex, has so many non-smooth terms, which make it very challenging to solve. In the paper, we propose an efficient optimization algorithm to solve a more general problem, based on which we find the optimal solution to the derived problem.

Introduction

In most machine learning and data mining applications, such as image annotations and categorizations, we often have a small set of labeled data together with a large collection of unlabeled data. Due to the small size of labeled data, the traditional supervised classification methods cannot be applied. The semi-supervised method involving both labeled and unlabeled data to train classification model is more realistic to solve the problems. The semi-supervised learning can be viewed as label propagation from labeled data to unlabeled data. In its simplest form, the label propagation is like a random walk on a similarity graph (Szummer and Jaakkola 2002; Nie et al. 2010b). Using the diffusion kernel, the semi-supervised learning is like a diffusive process of the labeled information (Kondor and Lafferty 2002). The harmonic function approach (Zhu, Ghahramani, and Lafferty 2003) emphasizes the harmonic nature of the diffusive function. The consistency labeling approach (Zhou et al. 2004) focuses on the iterative spread of labels.

*Corresponding authors. This work was partially supported by the following grants: NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, NIH R01 AG049371.

The success of semi-supervised learning is based on how much information unlabeled data carry about the distribution of labels in the pattern space (Nie et al. 2011b). In general, researchers hypothesize a low-dimensional manifold structure along which labels can be assumed to vary smoothly (Belkin, Niyogi, and Sindhwani 2006; Sindhwani et al. 2005). Because the linear models are strongly preferred due to their ease of and empirical performance, the linear manifold regularization was introduced to semi-supervised learning (Belkin, Niyogi, and Sindhwani 2006; Sindhwani et al. 2005). However, the linear embedding often restricts the identification of manifold structure and reduces the classification performance of semi-supervised learning.

To solve this problem, we propose a new elastic constraint on the predicted label matrix such that a better manifold structure can be identified. Meanwhile, we consider that the outliers often exist in the unlabeled data and confuse the learning model. In order to reduce the outliers effect, we replace the traditional squared loss function by a non-squared one which is robust to the outliers. Although our new objective is a convex function, it includes many non-smooth terms which make the optimization problem very challenging. We provide a new efficient optimization algorithm to solve a more general problem and also solve the proposed objective. Extensive experiments have been performed on both single-label image categorizations and multi-label image annotations to evaluate the new method. In all empirical results, our approach outperforms other related methods.

Related Work Revisited

Suppose we have n training data points $\{x_1, x_2, ..., x_n\}$, denote the data matrix by $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d\times n}$, where d is the dimensionality. For a graph based method, a graph on the data is constructed based on the training data points. The similarity matrix $A \in \mathbb{R}^{n\times n}$ of the graph is calculated to encode the similarities between data pairs. In semi-supervised learning, only a few data points of the training data are labeled, and the remaining data points are unlabeled. Without loss of generality, suppose the first l data points $x_1, ..., x_l$ are labeled as $y_1, ..., y_l$, respectively. For every i, $y_i \in \{1, 2, ..., c\}$, where c is the number of classes. The task of semi-supervised learning is to predict the labels for the unlabeled data points. Denote the label matrix of the first l data points by $Y_l \in \mathbb{R}^{l\times c}$, where the (i, j)-th element of Y_l is

1 if $y_i = j$ and 0 otherwise. Denote the predicted label matrix by $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix}$, where $F_l \in \mathbb{R}^{l\times c}$ and $F_u \in \mathbb{R}^{(n-l)\times c}$. Label propagation (Zhu, Ghahramani, and Lafferty 2003) is a popular method for semi-supervised learning, which is to solve the following problem:

$$\min_{F,F_l=Y_l} \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 \tag{1}$$

Eq.(1) can be rewritten as: $\min_{F,F_l=Y_l} tr(F^T L F)$, where L is the Laplacian matrix defined as $L = D - A$, the D is a diagonal matrix with the i-th diagonal element $D_{ii} = \sum_i A_{ij}$. With writing L as blockwise matrix, Eq.(1) is rewritten as

$$\min_{F_u} tr\left(\begin{bmatrix} Y_l \\ F_u \end{bmatrix}^T \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \begin{bmatrix} Y_l \\ F_u \end{bmatrix}\right) \tag{2}$$

By setting the derivative of Eq.(2) w.r.t. F_u to zero, we get the optimal solution F_u as: $F_u = -L_{uu}^{-1} L_{ul} Y_l$. The problem in Eq.(1) can be generalized as

$$\min_{F} \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 + \alpha\left\|F_l - Y_l\right\|_F^2 \tag{3}$$

When $\alpha \to \infty$, Eq.(3) is reduced to Eq.(1).

Label propagation can only be used in the transductive setting, that is, the method can only predict the labels for the seen training data points, and can not predict the labels for new data points unseen in the training procedure. To handle this out-of-sample problem, a manifold regularization method was proposed recently (Belkin, Niyogi, and Sindhwani 2006; Sindhwani et al. 2005). The basic idea of this inductive method is to learn a linear model by constraining $X^T W + 1b^T = F$, where 1 is a vector with all the elements as one. Thus the problem in Eq.(3) becomes

$$\min_{F,X^T W+1b^T=F} \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 + \alpha\left\|F_l - Y_l\right\|_F^2 \tag{4}$$

By adding a regularization term $\|W\|_F^2$ of the linear model to avoid overfitting, problem (4) can be rewritten as

$$\min_{W,b} tr(W^T X L X^T W) + \alpha\left\|X_l^T W + 1b^T - Y_l\right\|_F^2 + \beta\|W\|_F^2 \tag{5}$$

where $X_l = [x_1, x_2, ..., x_l] \in \mathbb{R}^{d\times l}$.

Elastic and Robust Embedding for Semi-Supervised Classification

As can be seen in Eq.(4), the manifold regularization method made a constraint that the predicted label matrix F must be exactly equal to the linear model $X^T W + 1b^T$. In practice, this constraint is too rigid to capture the manifold structure of data for label propagation (Nie et al. 2010c; 2013). In this paper, we propose to use an elastic constraint on the label propagation such that the manifold structure can be better explored and a linear model is also learned for induction. Specifically, instead of using the rigid constraint $X^T W + 1b^T = F$, we use an elastic constraint $\left\|X^T W + 1b^T - F\right\|_F^2 \le \delta$ on Eq.(3), which results in the following problem:

$$\min_{F,\|X^T W+1b^T-F\|_F^2\le\delta} \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 + \alpha\left\|F_l - Y_l\right\|_F^2 \tag{6}$$

Obviously, when $\delta \to 0$, Eq.(6) is reduced to Eq.(4).

In practice, labeled data are usually few and we can make a reasonable assumption that all the labels of the labeled data are correct. As in the label propagation method (Zhu, Ghahramani, and Lafferty 2003), we can set the parameter α in Eq.(6) to infinite to fully make use of the label information. Then the problem in Eq.(6) becomes

$$\min_{F,F_l=Y_l,\|X^T W+1b^T-F\|_F^2\le\delta} \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 \tag{7}$$

By adding a regularization term $\|W\|_F^2$ of the linear model to avoid overfitting, this problem can be equivalently rewritten as follows:

$$\begin{aligned}\min_{F,F_l=Y_l,W,b} &\tfrac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2^2 \\ &+\alpha\sum_{i=1}^{n}\left\|W^T x_i + b - f_i\right\|_2^2 + \beta\|W\|_F^2\end{aligned} \tag{8}$$

In practice, unlabeled data are often very abundant, and usually there are some outliers in the unlabeled data. It is known that the traditional squared loss is sensitive to outliers, we propose to use the loss without the square and thus the learned model is robust to outliers (Nie et al. 2010a; 2011a). The proposed problem is as follows:

$$\begin{aligned}\min_{F,F_l=Y_l,W,b} &\tfrac{1}{2}\sum_{i,j=1}^{n} A_{ij}\left\|f_i - f_j\right\|_2 \\ &+\alpha\sum_{i=1}^{n}\left\|W^T x_i + b - f_i\right\|_2 + \beta\|W\|_F^2\end{aligned} \tag{9}$$

As in Eq.(8), the problem in Eq.(9) is still convex. However, solving the problem (9) is very challenging since there are more than n^2 non-smooth terms of ℓ_2 norm (without square). Traditional sparsity induced norms minimization methods are difficult to solve this kind of problem with so many sparsity induced norms.

Optimization Algorithm

Optimization Algorithm for A General Problem

First, let us consider a more general problem as follows:

$$\min_x f(x) + \sum_i \left\|g_i(x)\right\|_2, \tag{10}$$

where $g_i(x)$ is a vector output function. It can be seen that the problem (9) is a special case of the problem (10).

Eq. (10) is non-smooth, we turn to solve the following smooth problem:

$$\min_x f(x) + \sum_i \sqrt{g_i^T(x) g_i(x) + \delta} \tag{11}$$

When $\delta \to 0$, Eq.(11) is reduced to Eq.(10).

By setting the derivative of Eq.(11) w.r.t. x to zero, we have

$$f'(x) + \sum_i \frac{g_i(x)}{\sqrt{g_i^T(x)g_i(x)+\delta}} = 0 \tag{12}$$

Denote

$$s_i = \frac{1}{2\sqrt{g_i^T(x)g_i(x)+\delta}} \tag{13}$$

Then Eq.(12) is rewritten as

$$f'(x) + \sum_i 2s_i g_i(x) = 0 \tag{14}$$

Note that s_i is dependent on x, this equation is difficult to solve. However, if s_i is given for every i, then solving Eq.(14) is equivalent to solving the following problem:

$$\min f(x) + \sum_i s_i g_i^T(x)g_i(x) \tag{15}$$

Based on the above analysis, we propose an iterative algorithm to find the solution of Eq.(12), and thus the optimal solution of problem (11). We will give a theoretical analysis to prove the convergence of the proposed algorithm. The detailed algorithm is described in Algorithm 1. In the algorithm, We first guess a solution x, then we calculate s_i based on the current solution x and update the current solution x by the optimal solution of problem (15) based on the calculated s_i, this procedure is iteratively performed until converges.

Algorithm 1: The algorithm to solve the problem (11).

Initialize x ;
while *not converge* **do**
  1. For each i, calculate s_i according to Eq.(13). ;
  2. Update x by solving the problem (15) ;
end
Output: x.

Convergence Analysis of Algorithm 1

To prove the convergence of the Algorithm 1, we need the following lemma:

Lemma 1 *For any vectors $\tilde{x}, x$ with the same size, the following inequality holds:*

$$\sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx}{2\sqrt{x^Tx+\delta}}.$$

Proof: We begin with an obvious inequality $-\left(\sqrt{\tilde{x}^T\tilde{x}+\delta} - \sqrt{x^Tx+\delta}\right)^2 \le 0$, we have

$$\begin{aligned}
&-\left(\sqrt{\tilde{x}^T\tilde{x}+\delta} - \sqrt{x^Tx+\delta}\right)^2 \le 0 \\
&\Rightarrow 2\sqrt{\tilde{x}^T\tilde{x}+\delta}\sqrt{x^Tx+\delta} - (\tilde{x}^T\tilde{x}+\delta) \le x^Tx+\delta \\
&\Rightarrow \sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}+\delta}{2\sqrt{x^Tx+\delta}} \le \frac{\sqrt{x^Tx+\delta}}{2} \\
&\Rightarrow \sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}+\delta}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx+\delta}{2\sqrt{x^Tx+\delta}} \\
&\Rightarrow \sqrt{\tilde{x}^T\tilde{x}+\delta} - \frac{\tilde{x}^T\tilde{x}}{2\sqrt{x^Tx+\delta}} \le \sqrt{x^Tx+\delta} - \frac{x^Tx}{2\sqrt{x^Tx+\delta}}
\end{aligned}$$

which completes the proof. □

As a result, we have the following theorem:

Theorem 1 *The Algorithm 1 will monotonically decrease the objective of the problem (11) in each iteration.*

Proof: In step 2 of Algorithm 1, suppose the updated x is $\tilde{x}$. According to step 2, we know that

$$f(\tilde{x}) + \sum_i s_i g_i^T(\tilde{x})g_i(\tilde{x}) \le f(x) + \sum_i s_i g_i^T(x)g_i(x) \tag{16}$$

Note that $s_i = \frac{1}{2\sqrt{g_i^T(x)g_i(x)+\delta}}$, so we have

$$f(\tilde{x}) + \sum_i \frac{g_i^T(\tilde{x})g_i(\tilde{x})}{2\sqrt{g_i^T(x)g_i(x)+\delta}} \le f(x) + \sum_i \frac{g_i^T(x)g_i(x)}{2\sqrt{g_i^T(x)g_i(x)+\delta}}$$

According to Lemma 1, we have

$$\begin{aligned}
&\sum_i \sqrt{g_i^T(\tilde{x})g_i(\tilde{x})+\delta} - \sum_i \frac{g_i^T(\tilde{x})g_i(\tilde{x})}{2\sqrt{g_i^T(x)g_i(x)+\delta}} \\
&\le \sum_i \sqrt{g_i^T(x)g_i(x)+\delta} - \sum_i \frac{g_i^T(x)g_i(x)}{2\sqrt{g_i^T(x)g_i(x)+\delta}}
\end{aligned} \tag{17}$$

Summing the above two equations on two sides, we arrive at

$$f(\tilde{x}) + \sum_i \sqrt{g_i^T(\tilde{x})g_i(\tilde{x})+\delta} \le f(x) + \sum_i \sqrt{g_i^T(x)g_i(x)+\delta}$$

The above equalities hold when and only when the algorithm converges. Thus the Algorithm 1 will monotonically decrease the objective of the problem (11) in each iteration until the algorithm converges. □

In the convergence, the equality in Eq. (12) holds, thus the KKT condition (Boyd and Vandenberghe 2004) of problem (11) is satisfied. Therefore, the Algorithm 1 will converge to a stationary point, which is usually an optimum solution to the problem (11).

Optimization Algorithm to Problem (9)

In this subsection, we describe how to solve the problem (9) based on the Algorithm 1. According to the step 2 in Algorithm 1, i.e. Eq.(15), the key step of solving problem (9) is to solve the following problem:

$$\begin{aligned}
&\min_{F,F_l=Y_l,W,b} \tfrac{1}{2}\sum_{i,j=1}^n A_{ij}\hat{S}_{ij}\left\|f_i - f_j\right\|_2^2 \\
&+\alpha\sum_{i=1}^n S_{ii}\left\|W^Tx_i + b - f_i\right\|_2^2 + \beta\left\|W\right\|_F^2
\end{aligned} \tag{18}$$

where $\hat{S}_{ij} = \frac{1}{2\sqrt{\|f_i-f_j\|_2^2+\delta}}$ and $S_{ii} = \frac{1}{2\sqrt{\|W^Tx_i+b-f_i\|_2^2+\delta}}$ are calculated by the current solution of W,b and F, and δ is a small enough (approaching to zero) positive constant.

Denote $\hat{L} = \hat{D} - \hat{A}$, where $(\hat{A})_{ij} = A_{ij}\hat{S}_{ij}$, $\hat{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j(\hat{A})_{ij}$, and denote S as a diagonal matrix, where the i-th diagonal element is S_{ii}. Problem (18) can be rewritten as

$$\begin{aligned}
&\min_{F,F_l=Y_l,W,b} tr(F^T\hat{L}F) + \beta\left\|W\right\|_F^2 \\
&+\alpha tr(X^TW + 1b^T - F)^TS(X^TW + 1b^T - F)
\end{aligned} \tag{19}$$

By setting the derivative of Eq.(19) w.r.t. b to zero, we have

$$b = \frac{1}{1^T S 1}(F^T S 1 - W^T X S 1) \tag{20}$$

Substituting Eq.(20) into the problem (19), we have

$$\begin{aligned} &\min_{F, F_l = Y_l, W} tr(F^T \hat{L} F) + \beta \|W\|_F^2 \\ &+\alpha tr(PX^T W - PF)^T S (PX^T W - PF) \end{aligned} \tag{21}$$

where $P = I - \frac{1}{1^T S 1} 1 1^T S$. By setting the derivative of Eq.(21) w.r.t. W to zero, we have

$$W = \alpha(\alpha X H X^T + \beta I)^{-1} X H F \tag{22}$$

where $H = P^T S P = S - \frac{1}{1^T S 1} S 1 1^T S$.

Then according to Eq.(21) and Eq.(22), we have

$$\begin{aligned} &\min_{F, F_l = Y_l, W} tr(F^T \hat{L} F) + \alpha tr(F^T H F) + \beta tr(W^T W) \\ &+\alpha tr(W^T X H X^T W) - 2\alpha tr(F^T H X^T W) \\ &\Leftrightarrow \min_{F, F_l = Y_l, W} tr(F^T \hat{L} F) + \alpha tr(F^T H F) \\ &+tr(W^T(\alpha X H X^T W + \beta I) W) - 2\alpha tr(F^T H X^T W) \\ &\Leftrightarrow \min_{F, F_l = Y_l} tr(F^T \hat{L} F) + \alpha tr(F^T H F) \\ &-\alpha^2 tr(F^T H X^T (\alpha X H X^T + \beta I)^{-1} X H F) \end{aligned} \tag{23}$$

Define a matrix M as following

$$M = \hat{L} + \alpha H - \alpha^2 H X^T (\alpha X H X^T + \beta I)^{-1} X H, \tag{24}$$

and write M as a blockwise matrix, then Eq.(23) can be rewritten as

$$\min_{F_u} tr\left(\begin{bmatrix} Y_l \\ F_u \end{bmatrix}^T \begin{bmatrix} M_{ll} & M_{lu} \\ M_{ul} & M_{uu} \end{bmatrix} \begin{bmatrix} Y_l \\ F_u \end{bmatrix} \right) \tag{25}$$

By setting the derivative of Eq.(25) w.r.t. F_u to zero, we get the optimal solution F_u as follows:

$$F_u = -M_{uu}^{-1} M_{ul} Y_l \tag{26}$$

Based on the above analysis, we propose an iterative algorithm to solve the problem (9). The detailed algorithm is described in Algorithm 2. Since problem (9) is convex, the algorithm will converge to the optimal solution according to Theorem 1.

Algorithm 2: The algorithm to solve the problem (9).

Input: The training data $X \in \mathbb{R}^{d \times n}$ and the label matrix $Y_l \in \mathbb{R}^{l \times c}$ for the first l data points, the similarity matrix $A \in \mathbb{R}^{n \times n}$.

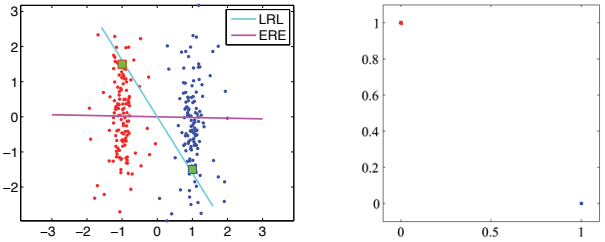
Initialize $W \in \mathbb{R}^{d \times c}$, $b \in \mathbb{R}^{c \times 1}$, $F \in \mathbb{R}^{n \times c}$;

while *not converge* **do**

1. Calculate $\hat{L} = \hat{D} - \hat{A}$, where $(\hat{A})_{ij} = \frac{A_{ij}}{2\sqrt{\|f_i - f_j\|_2^2 + \delta}}$, $\hat{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j (\hat{A})_{ij}$;
2. Calculate the diagonal matrix S, where the i-th diagonal element $S_{ii} = \frac{1}{2\sqrt{\|W^T x_i + b - f_i\|_2^2 + \delta}}$;
3. Calculate $H = S - \frac{1}{1^T S 1} S 1 1^T S$ and calculate M according to Eq.(24) ;
4. Update F_u by Eq.(26), $F = \begin{bmatrix} Y_l \\ F_u \end{bmatrix}$;
5. Update W by Eq.(22) ;
6. Update b by Eq.(20) ;

end

Output: $W \in \mathbb{R}^{d \times c}$, $b \in \mathbb{R}^{c \times 1}$, $F \in \mathbb{R}^{n \times c}$.



(a) Projection W (b) Embedding of F

Figure 1: Projection and Embedding results on a toy dataset.

Empirical Studies

A Toy Example

We generate a dataset randomly distributed on two different Gaussians with some noises (see Fig.1). The proposed Elastic and Robust Embedding (ERE) method and its most related methods, i.e. the Linearly Regularized Laplacian (LRL) method as defined in Eq. (5) (Belkin, Niyogi, and Sindhwani 2006; Sindhwani et al. 2005) are evaluated on this dataset. The results are shown in Fig.1. From Fig.1(a) we can see that our method successfully finds the optimal projection W while the LRL method is failed in this case. Fig.1(b) is the embedding result of the label matrix F learned by our method, which is very close to the ideal class indicator matrix. The results indicate that the proposed ERE method can effectively explore the distribution of unlabeled data to find the optimal projection W and the learned label matrix F is much more suitable for classification.

Implementation Details of Our Method

The proposed method has two parameters α and β in Eq. (9). Although it is tedious to seek an optimal combination of them, we can demonstrate that the performance of the proposed method is not sensitive to these parameters when they are sampled in certain value ranges. Upon some preliminary tests, we bound the parameters in the ranges of $1 \leq \alpha \leq 10$ and $0.01 \leq \beta \leq 0.1$.

Besides the feature input in vector form, our method also requires an input graph in the form of pairwise similarity to explore the manifold structures of the input data. Following (He et al. 2005), we construct the nearest-neighbor graph for an input data set, where the neighborhood size for graph construction is set as optimal by searching the grid of $\{1, 2, \ldots, 10\}$.

Classification Rules by Our Method

Given the output decision matrix F from Algorithm 2 for both labeled and unlabeled data points, their relevances to the classes of interest are ranked, upon which we can assign labels to the unlabeled data points.

For single-label image classification tasks, in which each image belongs one and only one class, we classify an unlabeled image $x_i\,(l+1 \leq i \leq n)$ by: $l\,(x_i) = \arg\max_k F_{ik}$.

For multi-label image classification tasks, in which one image could belong to more than one class, we need a threshold to make label prediction (Wang, Huang, and Ding 2009). For every class we learn a threshold from the labeled data as follows: $\tau_k = \sum_i^l Y_{l(ik)} F_{ik} / \sum_i^l Y_{l(ik)}$, which is the average decision score of all the labeled data points belonging to the k-th class. Then we determine the class membership for an unlabeled image $x_i\,(l+1 \leq i \leq n)$ by the following rule: assign x_i to the k-th class if $F_{ik} \geq b_k$, and not otherwise.

For the classification of an out-of-sample data point $x \in \mathbb{R}^d$, we first compute its decision vector by $f = W^T x + b$, where W is the output projection matrix and b is output bias vector from Algorithm 2. Then we predict labels for x by applying the same classification rules as above.

Improved Single-Label Classifications

We experiment with two single-label image data sets (Caltech-101 and MSRC-v1) for four image classification tasks (following (Dueck and Frey 2007; Lee and Grauman 2009)), which are broadly used in computer vision studies.

Experimental setups. We compare the proposed ERE method against its most related methods, *i.e.*, the LRL method as defined in Eq. (5). We also compare our method to the following widely used semi-supervised method including Transductive SVM (TSVM) (Joachims 1999) method, Gaussian Field and Harmonic function (GFHF) (Zhu, Ghahramani, and Lafferty 2003) method, Green's Function (GF) (Ding et al. 2007) method and Transductive Classification via Dual Regularization (TCDR) (Gu and Zhou 2009) method. As a baseline, we also report the classification performances of SVM on the same data sets, though it is a supervised classification method. We implement SVM and TSVM methods using the SVM^{light} software package[1]. Following (Joachims 1999), we fix the regularization parameter $C = 1$ and use the Gaussian kernel (*i.e.*, $\mathcal{K}(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right)$) where γ is fined tuned in the range of $\left\{10^{-5}, \ldots, 10^{-1}, 1, 10^1, \ldots, 10^5\right\}$. For SVM and TSVM methods, we employ one-vs-other strategy to deal with multi-class data sets. We implement the other compared methods and set the parameters to be optimal following their original works.

Experimental results. For each of the four classification tasks from the two image data sets, we randomly select 20% images as labeled data and the rest as unlabeled data, on which we perform all the compared methods. A 5-fold cross-validation is conducted on the labeled data to fine tune the parameters of the compared methods. We repeat each test case for 20 times and report the average performance. Because we experiment with single-label data, the macro average classification accuracies over all the classes of the compared methods are reported in Table 1, from which we have a number of interesting observations as following.

[1]http://svmlight.joachims.org/

First, the proposed method is consistently better than the other compared methods with a significant margin, which clearly demonstrate its effectiveness on single-label image data. Second, our new method obviously outperforms its non-robust and rigid counterpart, *i.e.*, the LRL method, which is consistent with our previous theoretical analyses: our method does better in the robustness against noises and outlier samples that are inevitable in large-scale data and its elasticity between the predicted and ground truth labels for the labeled data. These important results concretely confirm the correctness and advantage of our learning objective in Eq. (9) over that of the LRL method in Eq. (5). Third, the four methods, including SVM, TSVM, GFHF and GF methods, which take as input only one single input data format of either feature description in vector form or pairwise similarities between data points in graph form, generally do not perform well, which can be explained as follows. Although both the feature vectors and the similarity graph are describing a same set of objects, they may emphasize different aspects of the data and could reinforce the discriminability of each other (Wang, Huang, and Ding 2010a).

Improved Multi-Label Classifications

Now we evaluate the proposed method in multi-label image classification tasks. Multi-label classification can be seen as a generalization of traditional single-label classification, which is more challenging yet more close to real-world computer vision applications. We experiment with the following two broadly used multi-label image data sets: Mediamill (Snoek et al. 2006) and MSRC-v2.

Experimental setups. As in previous subsection, we still compare the proposed approach against its two most related methods, *i.e.*, TCDR and LRL methods. In addition, we also compare our method to three recent multi-label image classification methods including Multi-Label Green's Function (MLGF) (Wang, Huang, and Ding 2009) method that uses graph input, Multi-Label Least Square (MLLS) (Ji et al. 2010) method that uses only feature vector input, and multi-label feature transform (MLFT) (Wang, Huang, and Ding 2010b) method that integrates the inputs in the both formats.

Experimental results. We conduct standard 5-fold cross-validation and report the average results over the five trial on the two data sets by the compared methods in Table 2 and Table 3. From the results, we can see that our method still performs the best on the both test data sets. Besides, the TCDR and LRL methods are worse than the other three compared methods, as well as ours. This is because these two methods are designed for single-label classification, while the rest three compared methods are designed for multi-label data sets to leverage the cross-label correlations. Although our method is not particularly designed for multi-label classification, as in (Chang et al. 2014), it is naturally capable to deal with multi-label data.

Inductive Classification on Out-of-Sample Data

Although semi-supervised learning methods (Joachims 1999) have been successfully used in many computer vision applications solve the lack labeled data by utilizing a large

Table 1: Comparison of the average macro classification accuracy (mean $\pm$ standard deviation).

Methods	Caltech-101 (7 classes)	Caltech-101 (20 classes)	Caltech-101 (all)	MSRC-v1
SVM	0.75 ± 0.15	0.50 ± 0.17	0.37 ± 0.12	0.76 ± 0.20
TSVM	0.81 ± 0.04	0.56 ± 0.04	0.41 ± 0.03	0.78 ± 0.06
GFHF	0.51 ± 0.09	0.38 ± 0.07	0.21 ± 0.04	0.56 ± 0.09
GF	0.63 ± 0.06	0.42 ± 0.03	0.26 ± 0.02	0.60 ± 0.06
TCDR	0.81 ± 0.09	0.62 ± 0.04	0.51 ± 0.07	0.82 ± 0.07
LRL	0.77 ± 0.09	0.49 ± 0.12	0.40 ± 0.05	0.74 ± 0.06
ERE	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.69 \pm 0.03}$	$\mathbf{0.62 \pm 0.05}$	$\mathbf{0.88 \pm 0.09}$

Table 2: Multi-label classification on mediamill data.

Methods	Macro average		Micro average	
	Precision	F1	Precision	F1
MLGF	0.204	0.206	0.201	0.304
MLFT	0.397	0.425	0.386	0.570
MLLS	0.385	0.410	0.352	0.560
TCDR	0.362	0.403	0.326	0.533
LRL	0.364	0.397	0.341	0.519
ERE	**0.411**	**0.421**	**0.412**	**0.581**

Table 3: Multi-label classification on MSRC-v2 data.

Methods	Macro average		Micro average	
	Precision	F1	Precision	F1
MLGF	0.216	0.279	0.237	0.287
MLFT	0.256	0.304	0.259	0.312
MLLS	0.255	0.291	0.256	0.301
TCDR	0.242	0.278	0.241	0.299
LRL	0.201	0.252	0.226	0.264
ERE	**0.291**	**0.363**	**0.315**	**0.340**

amount of cheap unlabeled data, another important practical issue is how to effectively deal with out-of-sample images that are not available at the training phase. Most transductive learning methods, such as the GFHF and GF methods, are not able to generalize beyond the (labeled and unlabeled) training data. However, in many real world applications, test images are only available at the testing phase but not at the training phase. As a result, the inductive capability to classify unseen images at the training phase is desired for practical use. Motivated by the prior studies (Belkin, Niyogi, and Sindhwani 2006; Sindhwani et al. 2005), as an important advantage, our model gracefully solved this problem by replacing the soft label indications by a linear projection on the input data, as in Eq. (9). Therefore, in this subsection, we evaluate the inductive classification capability of the proposed method.

Experimental setups. In this experimental study, we use the Caltech-101 and the MSRC-v2 data sets. The former is a single-label data set, while the latter is a multi-label one. For each data set, we randomly split it into two halves, and use one half as training data and the other one as testing data. For the training data, we randomly label 40% of the images and leave the rest 60% of the images as unlabeled. Our task is to learn a model from the training data to classify the images in the testing data.

Table 4: Multi-label classification on MSRC-v2 data.

Methods	Caltech-101	MSCR-v2			
	Accuracy	Macro average		Micro average	
		Precision	F1	Precision	F1
PCA	0.327	0.174	0.250	0.219	0.281
LDA	0.536	0.230	0.325	0.262	0.298
LPP	0.315	0.147	0.240	0.215	0.256
DLE	0.553	0.242	0.332	0.281	0.304
LRL	0.452	0.201	0.252	0.226	0.264
ERE	**0.631**	**0.287**	**0.352**	**0.309**	**0.338**

Besides comparing our method to the LRL method, we also compare it to the following standard and widely used projection methods in statistical learning: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) (He et al. 2005) and Discriminant Laplacian Embedding (DLE) (Wang, Huang, and Ding 2010a) method. PCA and LDA methods take feature vectors as input, while LPP method takes graphs as input. DLE method aims at integrating the both forms of inputs to seek a better low-dimensional discriminative subspace.

Experimental results. Each of the test cases for the compared methods on the two data sets is repeated for 20 times and the average results are reported in Table 4. From the results we can see that, PCA and LPP methods do not lead to satisfactory classification performance because they are unsupervised methods. LDA and DLE methods are generally better, because the former is a supervised method and the latter is a semi-supervised method. However, none of these compared methods could outperform the proposed method, which demonstrate the inductive classification capability of our new method that is able to generalize to the out-of-sample images beyond the training data and adds to its practical value.

Conclusions

In this paper, we propose an elastic and robust embedding method for semi-supervised classification. To handle the out-of-sample problem in transductive semi-supervised learning, many inductive methods make a rigid constraint on the predicted label matrix to learn a linear model. In practice, this constraint might be too rigid. We relax this rigid constraint and use an elastic constraint such that the man-

ifold structure can be better explored. Moreover, we use a non-squared loss instead of the traditional squared loss to improve the robustness to outliers that often lie in the abundant unlabeled data. We proposed an efficient optimization algorithm to solve the new challenging objective. Experimental results on toy example, single-label and multi-label image categorizations clearly show the effectiveness of the proposed method.

References

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge University Press.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 1171–1177.

Ding, C.; Simon, H.; Jin, R.; and Li, T. 2007. A learning framework using Green's function and kernel regularization with application to recommender system. In *SIGKDD*.

Dueck, D., and Frey, B. 2007. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*.

Gu, Q., and Zhou, J. 2009. Transductive classification via dual regularization. In *ECML/PKDD*.

He, X.; Yan, S.; Hu, Y.; Niyogi, P.; and Zhang, H. 2005. Face recognition using laplacianfaces. *IEEE TPAMI* 27(3):328–340.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2010. A shared-subspace learning framework for multi-label classification. *TKDD*.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *ICML*, 200–209.

Kondor, R., and Lafferty, J. 2002. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, 315–322.

Lee, Y., and Grauman, K. 2009. Foreground focus: Unsupervised learning from partially matching images. *IJCV* 85(2):143–166.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010a. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*.

Nie, F.; Xiang, S.; Liu, Y.; and Zhang, C. 2010b. A general graph-based semi-supervised learning with novel class discovery. *Neural Computing Applications* 19(4):549–555.

Nie, F.; Xu, D.; Tsang, I. W.-H.; and Zhang, C. 2010c. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing* 19(7):1921–1932.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. 2011a. Unsupervised and semi-supervised learning via l1-norm graph. In *IEEE International Conference on Computer Vision (ICCV)*, 2268–2273.

Nie, F.; Xu, D.; Li, X.; and Xiang, S. 2011b. Semisupervised dimensionality reduction and classification through virtual label regression. *IEEE Trans. on Systems, Man, and Cybernetics, Part B* 41(3):675–685.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. H. Q. 2013. Adaptive loss minimization for semi-supervised elastic embedding. In *IJCAI*, 1565–1571.

Sindhwani, V.; Niyogi, P.; Belkin, M.; and Keerthi, S. 2005. Linear manifold regularization for large scale semi-supervised learning. In *ICML Workshop*.

Snoek, C. G. M.; Worring, M.; van Gemert, J. C.; Geusebroek, J.-M.; and Smeulders, A. W. M. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*.

Szummer, M., and Jaakkola, T. 2002. Partially labeled classification with Markov random walks. In *NIPS*, 945.

Wang, H.; Huang, H.; and Ding, C. 2009. Image annotation using multi-label correlated green's function. In *ICCV*.

Wang, H.; Huang, H.; and Ding, C. 2010a. Discriminant laplacian embedding. In *AAAI*.

Wang, H.; Huang, H.; and Ding, C. 2010b. Multi-label Feature Transform for Image Classifications. *ECCV*.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Scholkopf, B. 2004. Learning with local and global consistency. In *NIPS*, 321.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*.