Transductive Classification via Dual Regularization

Quanquan Gu Jie Zhou

State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
gqq03@mails.tsinghua.edu.cn, jzhou@tsinghua.edu.cn

Abstract. Semi-supervised learning has witnessed increasing interest in the past decade. One common assumption behind semi-supervised learning is that the data labels should be sufficiently smooth with respect to the intrinsic data manifold. Recent research has shown that the features also lie on a manifold. Moreover, there is a duality between data points and features, that is, data points can be classified based on their distribution on features, while features can be classified based on their distribution on the data points. However, existing semi-supervised learning methods neglect these points. Based on the above observations, in this paper, we present a dual regularization, which consists of two graph regularizers and a co-clustering type regularizer. In detail, the two graph regularizers consider the geometric structure of the data points and the features respectively, while the co-clustering type regularizer takes into account the duality between data points and features. Furthermore, we propose a novel transductive classification framework based on dual regularization, which can be solved by alternating minimization algorithm and its convergence is theoretically guaranteed. Experiments on benchmark semi-supervised learning data sets demonstrate that the proposed methods outperform many state of the art transductive classification methods.

1 Introduction

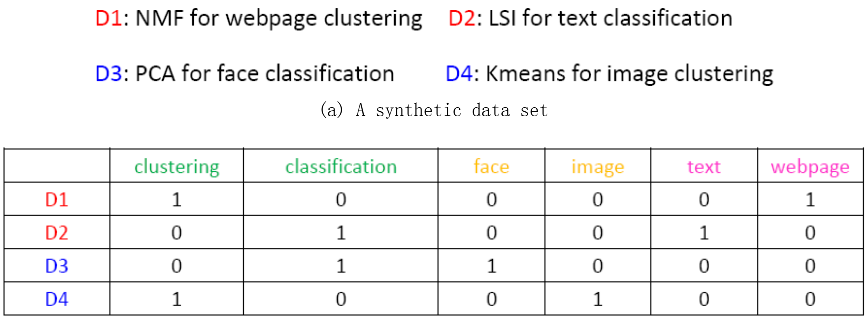
In many practical machine learning problems, the acquisition of sufficient labeled data is often expensive and/or time consuming. On the contrary, in many cases, large number of unlabeled data are far easier to obtain. Consequently, semi-supervised learning [1] [2], which aims to learn from both labeled and unlabeled data points, has received significant attention in the past decade. In general, semi-supervised learning can be categorized into two classes: (1) Transductive learning [3] [4] [5] [6] [7]: to estimate the labels of the given unlabeled data; and (2) Inductive learning [8]: to induce a decision function which has a low error rate on the whole sample space. In our study, we focus on transductive classification.

Many transductive classification methods have been proposed up to now [1] [2], among which graph based method [3] [4] [5] [6] [7] is one of the most popular approaches. One common assumption behind graph based transductive classification is that the data labels should be sufficiently smooth with respect to the intrinsic data manifold, i.e. *Cluster Assumption* [1] [3] [4]. This assumption can be achieved by graph regularization [9]. In detail, it models the whole data set as an undirected weighted graph, whose vertices correspond to the data points, and edges reflect the affinity between pairwise data points. Some of the vertices on the graph are labeled, while the remainder are unlabeled, and the goal of graph based transductive classification is to predict the labels of those unlabeled data points such that the predicted labels are sufficiently smooth with respect to the data graph. Other assumptions include *Local Learning Assumption* [6] [10], which says the label of each point can be predicted by the points in its neighborhood, and *local linear embedding assumption* [11] [5], which says if a data point can be reconstructed from its neighbors, then its label can be reconstructed from the labels of its neighbors by the same reconstruction coefficients.

The motivation of our work is twofold. First, recent research has shown that not only the data points are sampled from some low dimensional manifold embedded in the high dimensional ambient space [11] [12], namely data manifold, but also the features lie on a manifold [13] [14], namely feature manifold. Second, there is a duality between data points and features, i.e. data points can be classified based on their distribution on features while features can be classified based on their distribution on the data points. This is originally proposed in co-clustering literature [15] [16] [17], which suggests that clustering of features of a data matrix can lead to the improvement of data clustering. To demonstrate the usefulness of the duality between data points and features, we give an illustrative example in Fig. 1. As far as we know, existing semi-supervised learning methods fail to consider these points mentioned above together, which may further improve the performance of semi-supervised learning.

In this paper, we present a dual regularization. It consists of two graph regularizers and a co-clustering type regularizer. The graph regularizers explore the geometric structure of the data points and features respectively, while the co-clustering type regularizer utilizes the duality between the data points and features. Furthermore, we propose a novel framework for transductive classification based on dual regularization, which can be solved by alternating minimization algorithm and its convergence is theoretically guaranteed. Encouraging experimental results on benchmark semi-supervised learning data sets illustrate that the proposed methods outperform many existing transductive classification algorithms.

It is worth noting that in [18] [19], the authors proposed a co-clustering type regularization for transductive learning. However, in their method, besides partial supervision in the data points, partial supervision in the form of feature labels is also assumed, which is hardly obtained in many applications. Similar idea has also been applied for clustering [20]. Our method is different from theirs, since we do not require the supervision on the feature side. Furthermore, graph regularizers are adopted in our dual regularization, which considers the geometric structure of data points and features.



D1: NMF for webpage clustering D2: LSI for text classification

D3: PCA for face classification D4: Kmeans for image clustering

(a) A synthetic data set

	clustering	classification	face	image	text	webpage
D1	1	0	0	0	0	1
D2	0	1	0	0	1	0
D3	0	1	1	0	0	0
D4	1	0	0	1	0	0

(b) The document word matrix

Fig. 1. An illustrative example. Suppose we have 4 documents, i.e. D1, D2, D3 and D4. The true label of D1 and D2 is ”Data mining”, while the true label of D3 and D4 is ”Computer vision”. And suppose there are 6 words in the vocabulary, i.e. ”clustering, classification, face, image, text, webpage”. We assume D1 and D3 are labeled, while D2 and D4 are unlabeled and need to be predicted. If we only rely on the knowledge of the document side, then D2 is closer to D3 than D1, and D4 is closer to D1 than D3. So we will wrongly classify D2 as ”Computer vision” and D4 as ”data mining”. However, if we has the clusters information of the words, we may obtain 3 clusters, i.e. ”clustering, classification”, ”face, image” and ”text, webpage”, based on which D2 is closer to D1 than D3 while D4 is closer to D3 than D1. As a result, we will correctly classify D2 as ”Data mining”. Similar analysis can be conducted on the word side.

The remainder of this paper is organized as follows. In Section 2, we will briefly review graph-based transductive learning. In Section 3, we first present dual regularization, followed which we propose a novel transductive classification method based on dual regularization. The experiments on benchmark semi-supervised learning data sets are demonstrated in Section 4. Finally, we draw conclusions and point out the future work in Section 5.

2 A Brief Review of Graph Based Transductive Classification

Before we go any further, let’s first briefly review the general framework of graph based transductive classification [3] [4] [5] [6] [7], since it is the foundation of this paper.

In the setting of transductive classification, we are given a data set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, and a label set $\mathcal{L} = \{1, 2, \ldots, c\}$, the first l points $\mathbf{x}_i, 1 \leq i \leq l$ are labeled as $y_i \in \mathcal{L}$ and the remaining points $\mathbf{x}_u, l+1 \leq u \leq n$ are unlabeled. Each $\mathbf{x}_i$ is drawn from a fixed but usually unknown distribution $p(\mathbf{x})$. Typical graph based transductive classification seeks c optimal classification functions $f^j, 1 \leq j \leq c$, by minimizing the following criterion

$$J = \sum_{j=1}^{c} \sum_{i=1}^{l} L(y_i, f^j(\mathbf{x}_i)) + \lambda \sum_{j=1}^{c} ||f^j||_I^2, \qquad (1)$$

where $L(,)$ is some loss function, e.g. hinge loss or square loss, $||f^j||_I$ measures the smoothness of f^j with respect to the intrinsic data manifold, $\lambda > 0$ is the regularization parameter, which controls the balance between the loss and label smoothness.

Specifically, if we choose $L(,)$ as square loss, and $||f^j||_I$ as graph regularization, then Eq.(1) can be formulated as follows

$$\min_{\mathbf{F}} \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{C}(\mathbf{F}-\mathbf{Y})) + \lambda \mathrm{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}), \tag{2}$$

where $\mathbf{F} = (\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^c) \in \mathbb{R}^{n\times c}$ with $\mathbf{f}^j = [f^j(\mathbf{x}_1), f^j(\mathbf{x}_2), \ldots, f^j(\mathbf{x}_n)]^T$ is the class assignment matrix, $\mathbf{Y} \in \mathbb{R}^{n\times c}$ is the label matrix with $Y_{ij} = 1$ if $\mathbf{x}_i$ is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise, $\mathbf{L} \in \mathbb{R}^{n\times n}$ is called graph Laplacian [21], and $\mathbf{C} \in \mathbb{R}^{n\times n}$ is a diagonal matrix, with its ith diagonal element $C_{ii} = C_l > 0$ for $1 \le i \le l$, and $C_{ii} = C_u \ge 0$ for $l+1 \le i \le n$, where C_l and C_u are two parameters.

It is easy to show that the solution of Eq.(2) is

$$\mathbf{F} = (\mathbf{C} + \lambda\mathbf{L})^{-1}\mathbf{C}\mathbf{Y} \tag{3}$$

And the predicted label of $\mathbf{x}_i, l+1 \le i \le n$ is determined by

$$y_i = \arg\max_{1\le j\le c} \mathbf{F}_{ij}, l+1 \le i \le n \tag{4}$$

Most of the existing graph based transductive classification methods can be unified in Eq(2), and only differ in the setting of graph Laplacian $\mathbf{L}$ and/or the diagonal matrix $\mathbf{C}$. For example, [3] chose $\mathbf{L}$ as *combinational graph Laplacian* and $C_l = \infty, C_u = 0$. [4] set $\mathbf{L}$ as *normalized graph Laplacian* and $C_l = C_u = 1$. Both of the graph Laplacians mentioned above correspond to *Cluster Assumption* [1], while [6] selected $\mathbf{L}$ as *local learning graph Laplacian* which is based on *Local Learning Assumption*, and $C_l = 1, C_u = 0$. And [5] selected $\mathbf{L}$ as *local linear embedding graph Laplacian* and $C_l = C_u = 1$.

For convenience, we present in Table 1 the notation used in the rest of this paper.

Table 1. Notation used in this paper.

Notation	Description	Notation	Description
n	number of data points	$\mathbf{F}$	class assignment matrix of size $n \times c$
d	number of features	$\mathbf{f}_{i\cdot}$	ith row of $\mathbf{F}$
c	number of data classes	$\mathbf{f}_{\cdot i}$	ith column of $\mathbf{F}$
m	number offeature clusters	$\mathbf{G}$	feature partition matrix of size $d \times m$
$\mathcal{X}$	data set	$\mathbf{g}_{i\cdot}$	ith row of $\mathbf{G}$
$\mathbf{X}$	data matrix of size $d \times n$	$\mathbf{g}_{\cdot i}$	ith column of $\mathbf{G}$
$\mathbf{x}_{i\cdot}$	ith row of $\mathbf{X}$	$\mathbf{L}_F$	data graph Laplacian of size $n \times n$
$\mathbf{x}_{\cdot i}$	ith column of $\mathbf{X}$	$\mathbf{L}_G$	feature graph Laplacian of size $d \times d$

3 The Proposed Method

In this section, we will first present dual regularization. Then we will propose a transductive classification framework based on dual regularization, followed with the optimization algorithm as well as the proof of its convergence.

3.1 Dual Regularization

As we have mentioned above, we aim to explore the geometric structure on both the data point side and the feature side. To achieve this objective, we turn to graph regularization [9]. In detail, we construct two graphs i.e. data graph and feature graph, to explore the geometric structure of data manifold and feature manifold. In the following, we will introduce the construction of data graph and feature graph respectively. We will adopt *Cluster Assumption* [1] [3] [4] as a running example.

Data Graph We construct a data graph $\mathcal{G}_F$ whose vertices correspond to $\{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}\}$. According to *Cluster Assumption*, if data points $\mathbf{x}_{\cdot i}$ and $\mathbf{x}_{\cdot j}$ are close to each other, then their class labels $\mathbf{f}_{i\cdot}$ and $\mathbf{f}_{j\cdot}$ should be close as well. This is formulated as follows,

$$\frac{1}{2}\sum_{ij}||\frac{\mathbf{f}_{i\cdot}}{\sqrt{D_{ii}^F}} - \frac{\mathbf{f}_{j\cdot}}{\sqrt{D_{jj}^F}}||^2 W_{ij}^F \tag{5}$$

where W_{ij}^F is the affinity matrix of data graph measuring how close $\mathbf{f}_{i\cdot}$ and $\mathbf{f}_{j\cdot}$ will be, $D_{ii}^F = \sum_j W_{ij}^F$ is the diagonal degree matrix of data graph.

We define the data affinity matrix $\mathbf{W}^F$ as follows,

$$W_{ij}^F = \begin{cases} \exp\frac{-d(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j})^2}{\sigma^2}, \text{ if } \mathbf{x}_{\cdot j} \in \mathcal{N}(\mathbf{x}_{\cdot i}) \text{ or } \mathbf{x}_{\cdot i} \in \mathcal{N}(\mathbf{x}_{\cdot j}) \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{6}$$

where $\mathcal{N}(\mathbf{x}_{\cdot i})$ denotes the k-nearest neighbor of $\mathbf{x}_{\cdot i}$. $\exp\frac{-d(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j})^2}{\sigma^2}$ is Gaussian similarity where $\sigma > 0$ is the width. $d(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j})$ denotes the distance between $\mathbf{x}_{\cdot i}$ and $\mathbf{x}_{\cdot j}$.

Eq.(5) can be further rewritten as

$$\begin{aligned} &\frac{1}{2}\sum_{i,j}||\frac{\mathbf{f}_{i\cdot}}{\sqrt{D_{ii}^F}} - \frac{\mathbf{f}_{j\cdot}}{\sqrt{D_{jj}^F}}||^2 W_{ij}^F \\ &= \mathrm{tr}(\mathbf{F}^T(\mathbf{I} - (\mathbf{D}^F)^{-\frac{1}{2}}\mathbf{W}^F(\mathbf{D}^F)^{-\frac{1}{2}})\mathbf{F}) \\ &= \mathrm{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) \end{aligned} \tag{7}$$

where $\mathbf{F} \in \mathbb{R}^{n\times c}$ is the class assignment matrix of data points, and $\mathbf{L}_F = \mathbf{I} - (\mathbf{D}^F)^{-\frac{1}{2}}\mathbf{W}^F(\mathbf{D}^F)^{-\frac{1}{2}}$ is called *normalized graph Laplacian* (NLap) of the data graph $\mathcal{G}_F$. Eq.(7) reflects the label smoothness of the data points. The smoother the data labels are with respect to the underlying data manifold, the smaller the value of Eq.(7) will be.

Feature Graph Similar with the construction of the data graph $\mathcal{G}_F$, we construct a feature graph $\mathcal{G}_G$ whose vertices correspond to $\{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{d\cdot}\}$. According to *Cluster Assumption* again, if features $\mathbf{x}_{i\cdot}$ and $\mathbf{x}_{j\cdot}$ are near, then their cluster labels $\mathbf{g}_{i\cdot}$ and $\mathbf{g}_{j\cdot}$ should be near as well. This is formulated as follows

$$\frac{1}{2}\sum_{ij} ||\frac{\mathbf{g}_{i\cdot}}{\sqrt{D^G_{ii}}} - \frac{\mathbf{g}_{j\cdot}}{\sqrt{D^G_{jj}}}||^2 W^G_{ij} \tag{8}$$

where W^G_{ij} is the affinity matrix of feature graph measuring how close $\mathbf{g}_{i\cdot}$ and $\mathbf{g}_{j\cdot}$ will be, $D^G_{ii} = \sum_j W^G_{ij}$ is the diagonal degree matrix of feature graph.

Again we define the feature affinity matrix $\mathbf{W}^G$ as follows,

$$W^G_{ij} = \begin{cases} \exp\frac{-d(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot})^2}{\sigma^2}, \text{ if } \mathbf{x}_{j\cdot} \in \mathcal{N}(\mathbf{x}_{i\cdot}) \text{ or } \mathbf{x}_{i\cdot} \in \mathcal{N}(\mathbf{x}_{j\cdot}) \\ 0, \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{9}$$

where $\mathcal{N}(\mathbf{x}_{i\cdot})$ denotes the k-nearest neighbor of $\mathbf{x}_{i\cdot}$.

Eq.(8) can be further rewritten as

$$\begin{aligned} &\frac{1}{2}\sum_{i,j} ||\frac{\mathbf{g}_{i\cdot}}{\sqrt{D^G_{ii}}} - \frac{\mathbf{g}_{j\cdot}}{\sqrt{D^G_{jj}}}||^2 W^G_{ij} \\ &= \text{tr}(\mathbf{G}^T(\mathbf{I} - (\mathbf{D}^G)^{-\frac{1}{2}}\mathbf{W}^G(\mathbf{D}^G)^{-\frac{1}{2}})\mathbf{G}) \\ &= \text{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) \end{aligned} \tag{10}$$

where $\mathbf{G} \in \mathbb{R}^{d\times m}$ is the partition matrix of features, $\mathbf{L}_G = \mathbf{I} - (\mathbf{D}^G)^{-\frac{1}{2}}\mathbf{W}^G(\mathbf{D}^G)^{-\frac{1}{2}}$ is the *normalized graph Laplacian* (NLap) of the feature graph $\mathcal{G}_G$. Eq.(10) reflects the label smoothness of the features. The smoother the feature labels are with respect to the underlying feature manifold, the smaller the value of Eq.(10) will be.

Based on the two graph regularizers introduced above, we present a regularization as follows

$$\lambda\text{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) + \mu\text{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) + \eta||\mathbf{X} - \mathbf{G}\mathbf{S}\mathbf{F}^T||^2_F \tag{11}$$

where $||\cdot||_F$ is the Frobenius norm, $\lambda, \mu, \eta \geq 0$ are regularization parameters. It consists of three terms. The first term is graph regularizer defined in Eq.(7), which is also the same as the second term in Eq.(2). The second term is graph regularizer defined in Eq.(10). The third term is the most important one. It is a co-clustering [17] type regularizer. This term reflects the approximation error of matrix tri-factorization for co-clustering. The smaller it is, the better the approximation will be. It establishes a bridge between the data points in the first term and the features in the second term, through which the label information of data points and features can be transferred from one to another, which may benefit the classification of data points. Eq.(11) is called *Dual Regularization.*

By now, we have presented dual regularization. It is worth noting that although in the derivation, we adopt *Cluter Assumption*, other kinds of assumptions, e.g. *Local Learning Assumption* can also be used. In other words, besides the *normalized graph Laplacian* (NLap), other kinds of graph Laplacians can also be utilized in Eq.(11), e.g. *local learning graph Laplacian* (LLL). For the detail of LLL, please refer to [6] [10].

3.2 Transductive Classification via Dual Regularization

Based on the dual regularization in Eq.(11) presented above, we propose a novel transductive classification framework as follows,

$$\begin{aligned} J_{TCDR} &= \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{C}(\mathbf{F}-\mathbf{Y})) \\ &\quad + \lambda\mathrm{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) + \mu\mathrm{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) + \eta||\mathbf{X}-\mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2, \end{aligned} \tag{12}$$

where $||\cdot||_F$ is the Frobenius norm, $\mathbf{F}\in\mathbb{R}^{n\times c}$ is the class assignment matrix of the data points, $\mathbf{G}\in\mathbb{R}^{d\times m}$ is the partition matrix of the features, $\mathbf{L}_F\in\mathbb{R}^{n\times n}$ is graph Laplacian for data points, and $\mathbf{L}_G\in\mathbb{R}^{d\times d}$ is graph Laplacian for features. $\lambda,\mu,\eta\geq 0$ are regularization parameters. We call Eq.(12) *Transductive Classification via Dual Regularization* (TCDR). TCDR provides a unified framework for transductive classification. Different settings of the graph Laplacians $\mathbf{L}_F$ and $\mathbf{L}_G$, along with the diagonal matrix $\mathbf{C}$ lead to various instantiations of TCDR. When letting $\mu=\eta=0$ in Eq.(12), TCDR degenerates to traditional graph based transductive classification framework in Eq.(2). To this end, existing graph based transductive classification methods can be seen as the special case of TCDR.

By its definition, the elements in $\mathbf{F}$ and $\mathbf{G}$ can only take binary values, which makes the minimization in Eq.(12) very difficult, therefore we relax $\mathbf{F}$ and $\mathbf{G}$ into continuous nonnegative domain. Then the objective of TCDR in Eq.(12) turns out to be,

$$\begin{aligned} J_{TCDR} &= \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{C}(\mathbf{F}-\mathbf{Y})) \\ &\quad + \lambda\mathrm{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) + \mu\mathrm{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) + \eta||\mathbf{X}-\mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2, \\ &\quad \text{s.t. } \mathbf{F}\geq 0, \mathbf{G}\geq 0, \end{aligned} \tag{13}$$

To make the objective in Eq.(13) lower bounded, we use L_2 normalization on columns of $\mathbf{F}$ and $\mathbf{G}$ in the optimization, and compensate the norms of $\mathbf{F}$ and $\mathbf{G}$ to $\mathbf{S}$.

3.3 Optimization

As we see, minimizing Eq.(13) is with respect to $\mathbf{F},\mathbf{G}$ and $\mathbf{S}$. And we cannot give a closed-form solution. In the following, we will present an alternating scheme to optimize the objective. In other words, we will optimize the objective with respect to one variable when fixing the other variables. This procedure repeats until convergence.

Computation of S Optimizing Eq.(13) with respect to $\mathbf{S}$ is equivalent to optimizing

$$J_1 = ||\mathbf{X} - \mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2 \tag{14}$$

Setting $\frac{\partial J_1}{\partial \mathbf{S}} = 0$ leads to the following updating formula

$$\mathbf{S} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{X}\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1} \tag{15}$$

Computation of F Optimizing Eq.(13) with respect to $\mathbf{F}$ is equivalent to optimizing

$$\begin{aligned} J_2 \;=\; & \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{C}(\mathbf{F}-\mathbf{Y})) + \lambda\mathrm{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) + \eta||\mathbf{X} - \mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2, \\ & \text{s.t. } \mathbf{F} \geq 0, \end{aligned} \tag{16}$$

For the constraint $\mathbf{F} \geq 0$, we cannot get a closed-form solution of $\mathbf{F}$. In the following, we will present an iterative multiplicative updating solution. We introduce the Lagrangian multiplier $\boldsymbol{\alpha} \in \mathbb{R}^{n\times c}$, thus the Lagrangian function is

$$\begin{aligned} L(\mathbf{F}) = & \mathrm{tr}((\mathbf{F}-\mathbf{Y})^T\mathbf{C}(\mathbf{F}-\mathbf{Y})) \\ & + \lambda\mathrm{tr}(\mathbf{F}^T\mathbf{L}_F\mathbf{F}) + \eta||\mathbf{X} - \mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2 - \mathrm{tr}(\boldsymbol{\alpha}\mathbf{F}^T) \end{aligned} \tag{17}$$

Setting $\frac{\partial L(\mathbf{F})}{\partial \mathbf{F}} = 0$, we obtain

$$\boldsymbol{\alpha} = 2\mathbf{C}\mathbf{F} - 2\mathbf{C}\mathbf{Y} + 2\lambda\mathbf{L}_F\mathbf{F} - 2\eta\mathbf{A} + 2\eta\mathbf{F}\mathbf{B} \tag{18}$$

where $\mathbf{A} = \mathbf{X}^T\mathbf{G}\mathbf{S}$ and $\mathbf{B} = \mathbf{S}^T\mathbf{G}^T\mathbf{G}\mathbf{S}$.

Using the Karush-Kuhn-Tucker condition [22] $\boldsymbol{\alpha}_{ij}\mathbf{F}_{ij} = 0$, we get

$$[\mathbf{C}\mathbf{F} - \mathbf{C}\mathbf{Y} + \lambda\mathbf{L}_F\mathbf{F} - \eta\mathbf{A} + \eta\mathbf{F}\mathbf{B}]_{ij}\mathbf{F}_{ij} = 0 \tag{19}$$

Introduce $\mathbf{L}_F = \mathbf{L}_F^+ - \mathbf{L}_F^-$, $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$ and $\mathbf{B} = \mathbf{B}^+ - \mathbf{B}^-$ where $\mathbf{A}_{ij}^+ = (|\mathbf{A}_{ij}| + \mathbf{A}_{ij})/2$ and $\mathbf{A}_{ij}^- = (|\mathbf{A}_{ij}| - \mathbf{A}_{ij})/2$ [23], we obtain

$$[\mathbf{C}\mathbf{F} - \mathbf{C}\mathbf{Y} + \lambda\mathbf{L}_F^+\mathbf{F} - \lambda\mathbf{L}_F^-\mathbf{F} - \eta\mathbf{A}^+ + \eta\mathbf{A}^- + \eta\mathbf{F}\mathbf{B}^+ - \eta\mathbf{F}\mathbf{B}^-]_{ij}\mathbf{F}_{ij} = 0 \tag{20}$$

Eq.(20) leads to the following updating formula

$$\mathbf{F}_{ij} \leftarrow \mathbf{F}_{ij}\sqrt{\frac{[\mathbf{C}\mathbf{Y} + \lambda\mathbf{L}_F^-\mathbf{F} + \eta\mathbf{A}^+ + \eta\mathbf{F}\mathbf{B}^-]_{ij}}{[\mathbf{C}\mathbf{F} + \lambda\mathbf{L}_F^+\mathbf{F} + \eta\mathbf{A}^- + \eta\mathbf{F}\mathbf{B}^+]_{ij}}} \tag{21}$$

Computation of G Optimizing Eq.(13) with respect to $\mathbf{G}$ is equivalent to optimizing

$$\begin{aligned} J_3 \;=\; & \mu\mathrm{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) + \eta||\mathbf{X} - \mathbf{G}\mathbf{S}\mathbf{F}^T||_F^2 \\ & \text{s.t. } \mathbf{G} \geq 0, \end{aligned} \tag{22}$$

Since $\mathbf{G} \geq 0$, we introduce the Lagrangian multiplier $\boldsymbol{\beta} \in \mathbb{R}^{d\times m}$, thus the Lagrangian function is

$$L(\mathbf{G}) = \mu \mathrm{tr}(\mathbf{G}^T\mathbf{L}_G\mathbf{G}) + \eta||\mathbf{X} - \mathbf{GSF}^T||_F^2 - \mathrm{tr}(\boldsymbol{\beta}\mathbf{G}^T) \tag{23}$$

Setting $\frac{\partial L(\mathbf{G})}{\partial \mathbf{G}} = 0$, we obtain

$$\boldsymbol{\beta} = 2\mu\mathbf{L}_G\mathbf{G} - 2\eta\mathbf{P} + 2\eta\mathbf{GQ} \tag{24}$$

where $\mathbf{P} = \mathbf{XFS}^T$ and $\mathbf{Q} = \mathbf{SF}^T\mathbf{FS}^T$.

Using the Karush-Kuhn-Tucker complementarity condition [22] $\boldsymbol{\beta}_{ij}\mathbf{G}_{ij} = 0$, we get

$$[\mu\mathbf{L}_G\mathbf{G} - \eta\mathbf{P} + \eta\mathbf{GQ}]_{ij}\mathbf{G}_{ij} = 0. \tag{25}$$

Introduce $\mathbf{L}_G = \mathbf{L}_G^+ - \mathbf{L}_G^-$, $\mathbf{P} = \mathbf{P}^+ - \mathbf{P}^-$ and $\mathbf{Q} = \mathbf{Q}^+ - \mathbf{Q}^-$, we obtain

$$[\mu\mathbf{L}_G^+\mathbf{G} - \mu\mathbf{L}_G^-\mathbf{G} - \eta\mathbf{P}^+ + \eta\mathbf{P}^- + \eta\mathbf{GQ}^+ - \eta\mathbf{GQ}^-]_{ij}\mathbf{G}_{ij} = 0. \tag{26}$$

Eq.(26) leads to the following updating formula

$$\mathbf{G}_{ij} \leftarrow \mathbf{G}_{ij}\sqrt{\frac{[\mu\mathbf{L}_G^-\mathbf{G} + \eta\mathbf{P}^+ + \eta\mathbf{GQ}^-]_{ij}}{[\mu\mathbf{L}_G^+\mathbf{G} + \eta\mathbf{P}^- + \eta\mathbf{GQ}^+]_{ij}}} \tag{27}$$

3.4 Convergence Analysis

In this section, we will investigate the convergence of the updating formula in Eq.(21) and Eq.(27). We use the auxiliary function approach [24] to prove the convergence of the algorithm. Here we first introduce the definition of auxiliary function [24].

Definition 1. *[24] $Z(h, h')$ is an auxiliary function for $F(h)$ if the conditions*

$$Z(h, h') \geq F(h), Z(h, h) = F(h),$$

are satisfied.

Lemma 1. *[24] If Z is an auxiliary function for F, then F is non-increasing under the update*

$$h^{(t+1)} = \arg\min_h Z(h, h^{(t)})$$

Proof. $F(h^{(t+1)}) \leq Z(h^{(t+1)}, h^{(t)}) \leq Z(h^{(t)}, h^{(t)}) = F(h^{(t)})$

Lemma 2. *[23] For any nonnegative matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times k}$, $\mathbf{S} \in \mathbb{R}^{n\times k}$,$\mathbf{S}' \in \mathbb{R}^{n\times k}$, and $\mathbf{A}$, $\mathbf{B}$ are symmetric, then the following inequality holds*

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(\mathbf{AS'B})_{ip}\mathbf{S}_{ip}^2}{\mathbf{S}'_{ip}} \geq tr(\mathbf{S}^T\mathbf{ASB})$$

Theorem 1. *Let*

$$J(\mathbf{F}) = tr(\mathbf{F}^T\mathbf{C}\mathbf{F} - 2\mathbf{F}^T\mathbf{C}\mathbf{Y} + \lambda\mathbf{F}^T\mathbf{L}_F\mathbf{F} - 2\eta\mathbf{A}\mathbf{F}^T + \mathbf{F}\mathbf{B}\mathbf{F}^T) \tag{28}$$

Then the following function

$$\begin{aligned} Z(\mathbf{F},\mathbf{F}') = & \sum_{ij}\frac{(\mathbf{C}\mathbf{F}')_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}} - 2\sum_{ij}(\mathbf{C}\mathbf{Y})_{ij}\mathbf{F}'_{ij}(1+\log\frac{\mathbf{F}_{ij}}{\mathbf{F}'_{ij}}) \\ & + \lambda\sum_{ij}\frac{(\mathbf{L}_F^+\mathbf{F}')_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}} - \lambda\sum_{ijk}(\mathbf{L}_F^-)_{jk}\mathbf{F}'_{ji}\mathbf{F}'_{ki}(1+\log\frac{\mathbf{F}_{ij}\mathbf{F}_{ik}}{\mathbf{F}'_{ij}\mathbf{F}'_{ik}}) \\ & - 2\eta\sum_{ij}\mathbf{A}_{ij}^+\mathbf{F}'_{ij}(1+\log\frac{\mathbf{F}_{ij}}{\mathbf{F}'_{ij}}) + 2\eta\sum_{ij}\mathbf{A}_{ij}^-\frac{\mathbf{F}_{ij}^2+\mathbf{F}'^2_{ij}}{2\mathbf{F}'_{ij}} \\ & + \sum_{ij}\frac{(\mathbf{F}'\mathbf{B}^+)_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}} - \sum_{ijk}\mathbf{B}_{jk}^-\mathbf{F}'_{ij}\mathbf{F}'_{ik}(1+\log\frac{\mathbf{F}_{ij}\mathbf{F}_{ik}}{\mathbf{F}'_{ij}\mathbf{F}'_{ik}}) \end{aligned}$$

is an auxiliary function for $J(\mathbf{F})$. Furthermore, it is a convex function in $\mathbf{F}$ and its global minimum is

$$\mathbf{F}_{ij} = \mathbf{F}_{ij}\sqrt{\frac{[\mathbf{C}\mathbf{Y} + \lambda\mathbf{L}_F^-\mathbf{F} + \eta\mathbf{A}^+ + \eta\mathbf{F}\mathbf{B}^-]_{ij}}{[\mathbf{C}\mathbf{F} + \lambda\mathbf{L}_F^+\mathbf{F} + \eta\mathbf{A}^- + \eta\mathbf{F}\mathbf{B}^+]_{ij}}} \tag{29}$$

Proof. See Appendix.

Theorem 2. *Updating* $\mathbf{F}$ *using Eq.(21) will monotonically decrease the value of the objective in Eq.(13), hence it converges.*

Proof. By Lemma 1 and Theorem 1, we can get that $J(\mathbf{F}^0) = Z(\mathbf{F}^0,\mathbf{F}^0) \geq Z(\mathbf{F}^1,\mathbf{F}^0) \geq J(\mathbf{F}^1) \geq \ldots$ So $J(\mathbf{F})$ is monotonically decreasing. Since $J(\mathbf{F})$ is obviously bounded below, we prove this theorem.

Theorem 3. *Let*

$$J(\mathbf{G}) = tr(\mu\mathbf{G}^T\mathbf{L}_G\mathbf{G} - 2\eta\mathbf{G}^T\mathbf{P} + \mu\mathbf{G}\mathbf{Q}\mathbf{G}^T) \tag{30}$$

Then the following function

$$\begin{aligned} Z(\mathbf{G},\mathbf{G}') = & \sum_{ij}\frac{(\mathbf{L}_G^+\mathbf{G}')_{ij}\mathbf{G}_{ij}^2}{\mathbf{G}'_{ij}} - \sum_{ijk}(\mathbf{L}_G^-)_{jk}\mathbf{G}'_{ji}\mathbf{G}'_{ki}(1+\log\frac{\mathbf{G}_{ij}\mathbf{G}_{ik}}{\mathbf{G}'_{ij}\mathbf{G}'_{ik}}) \\ & - 2\eta\sum_{ij}\mathbf{P}_{ij}^+\mathbf{G}'_{ij}(1+\log\frac{\mathbf{G}_{ij}}{\mathbf{G}'_{ij}}) + 2\eta\sum_{ij}\mathbf{P}_{ij}^-\frac{\mathbf{G}_{ij}^2+\mathbf{G}'^2_{ij}}{2\mathbf{G}'_{ij}} \\ & + \mu\sum_{ij}\frac{(\mathbf{G}'\mathbf{Q}^+)_{ij}\mathbf{G}_{ij}^2}{\mathbf{G}'_{ij}} - \mu\sum_{ijk}\mathbf{Q}_{jk}^-\mathbf{G}'_{ij}\mathbf{G}'_{ik}(1+\log\frac{\mathbf{G}_{ij}\mathbf{G}_{ik}}{\mathbf{G}'_{ij}\mathbf{G}'_{ik}}) \end{aligned}$$

is an auxiliary function for $J(\mathbf{G})$. Furthermore, it is a convex function in $\mathbf{G}$ and its global minimum is

$$\mathbf{G}_{ij} = \mathbf{G}_{ij}\sqrt{\frac{[\mu\mathbf{L}_G^-\mathbf{G} + \eta\mathbf{P}^+ + \eta\mathbf{G}\mathbf{Q}^-]_{ij}}{[\mu\mathbf{L}_G^+\mathbf{G} + \eta\mathbf{P}^- + \eta\mathbf{G}\mathbf{Q}^+]_{ij}}} \tag{31}$$

Proof. For the limit of space, we omit it here.

Theorem 4. *Updating $\mathbf{G}$ using Eq.(27) will monotonically decrease the value of the objective in Eq.(13), hence it converges.*

Proof. By Lemma 1 and Theorem 3, we can get that $J(\mathbf{G}^0) = Z(\mathbf{G}^0, \mathbf{G}^0) \geq Z(\mathbf{G}^1, \mathbf{G}^0) \geq J(\mathbf{G}^1) \geq \ldots$ So $J(\mathbf{G})$ is monotonically decreasing. Since $J(\mathbf{G})$ is obviously bounded below, we prove this theorem.

4 Experiments

In this section, we evaluate the proposed methods on many benchmark semi-supervised learning data sets. Two instantiations of TCDR are evaluated: (1)*normalized graph Laplacian* (NLap) + TCDR, which chooses $\mathbf{L}_F$ and $\mathbf{L}_G$ as NLap. Note that it is just the method derived as a running example in Section 3; and (2) *local learning graph Laplacian* (LLL) + TCDR, which chooses $\mathbf{L}_F$ and $\mathbf{L}_G$ as LLL [6].

4.1 Data Sets

In our experiments, we use 9 benchmark semi-supervised learning data sets, which can be found in [1] [7].

g241c & g241n[1]**:** Each data set contains two classes with 350 points in each class, and the data sets are generated in a way of violating the cluster assumptions or misleading class structures.

USPS, COIL & Digit1: The first two data sets are generated from the famous USPS and COIL databases, such that the resultant image data did not appear to be manifold explicitly. The digit1 data set is generated by transforming the image of digit 1, and the image data appears a manifold structure strongly.

Cornell, Texas, Wisconsin & Washington: All these four data sets are selected from the famous WebKB database[2], and the web pages are classified into 5 ∼ 6 categories.

Table 1 summarizes the characteristics of the data sets mentioned above. For more details about these data sets, please refer to [1].

[1] http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html

[2] http://www.cs.cmu.edu/ webkb/

Table 2. Description of a subset of datasets

Datasets	#samples	#classes	#dimensions	Datasets	#samples	#classes	#dimensions
g241c	1500	2	241	cornell	826	6	4134
g241n	1500	2	241	texas	811	5	4029
USPS	1500	2	241	wisconsin	1210	6	4189
COIL	1500	6	241	washington	1165	6	4165
digit1	1500	2	241				

4.2 Methods & Parameter Settings

We compare our methods with some state of the art graph based transductive classification algorithms in the following.

Gaussian Field and Harmonic Function (GFHF) [3]: The width of the Gaussian similarity is set via the grid $\{2^{-3}\sigma_0^2, 2^{-2}\sigma_0^2, 2^{-1}\sigma_0^2, \sigma_0^2, 2\sigma_0^2, 2^2\sigma_0^2, 2^3\sigma_0^2\}$, where σ_0 is the mean distance between any two samples in the training set. And the size of $\mathcal{N}(\cdot)$ is searched by the grid $\{5, 10, 50, 80, n-1\}$

Learning with Local and Global Consistency (LLGC) [4]: The width of the Gaussian similarity and the size of $\mathcal{N}(\cdot)$ are also determined the same as that in GFHF, and the regularization parameter is set by searching the grid $\{0.1, 1, 10, 100\}$.

Transductive Classification with Local Learning Regularization (TCLLR) [6]: The neighborhood size for constructing local learning regularizer is searched by the grid $\{5, 10, 50, 80\}$, and the regularization parameter is set by searching the grid $\{0.1, 1, 10, 100\}$.

Transductive Classification via Dual Regularization (TCDR): In NLap+TCDR, we use *normalized graph Laplacian* on both the data point side and the feature side, and the width of the Gaussian similarity as well as the size of $\mathcal{N}(\cdot)$ are tuned the same as in GFHF. And we set $C_l = C_u = 1$. In LLL+TCDR, we use *local learning graph Laplacian* for both data points and features, and the neighborhood size is tuned the same as that in TLLR. And we set $C_l = 1, C_u = 0$. Besides, we set the number of feature clusters the same as the number of data classes for simplicity, i.e. $m = c$. And the regularization parameters, i.e. λ, μ, η are set by searching the grid $\{0.1, 1, 10, 100\}$.

For synthetic and image data sets, the distance between $\mathbf{x}_{\cdot i}$ and $\mathbf{x}_{\cdot j}$ (or $\mathbf{x}_{i\cdot}$ and $\mathbf{x}_{j\cdot}$) is computed as $d(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j}) = ||\mathbf{x}_{\cdot i} - \mathbf{x}_{\cdot j}||_2$ (or $d(\mathbf{x}_{i\cdot}, \mathbf{x}_{j\cdot}) = ||\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}||_2$).

For text data sets, we use TFIDF to weight the term-document matrix. The distance between two points $\mathbf{x}_{\cdot i}$ and $\mathbf{x}_{\cdot j}$ is defined as

$$d(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j}) = 1 - \frac{\langle \mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j} \rangle}{||\mathbf{x}_{\cdot i}|| \cdot ||\mathbf{x}_{\cdot j}||}. \tag{32}$$

And the distance between $\mathbf{x}_{i\cdot}$ and $\mathbf{x}_{j\cdot}$ can be computed analogously.

In order to compare these algorithms fairly, we randomly select $\{5\%, 10\%, \ldots, 45\%, 50\%\}$ data points as labeled samples, while the rest as unlabeled samples. Since the labeled set is randomly chosen, we repeat each experiment 20 times and calculate the average transductive classification accuracy. We run each algorithms under different parameter settings, and select the best average result to compare with each other.

4.3 Classification Results

The experimental results are shown in Fig. 2. In all figures, the x-axis represents the percentage of randomly labeled points, while the y-axis is the average transductive classification accuracy.

To illustrate the experimental results better, we also list the results of these algorithms on all the data sets with 10% labeled samples in Table 3.

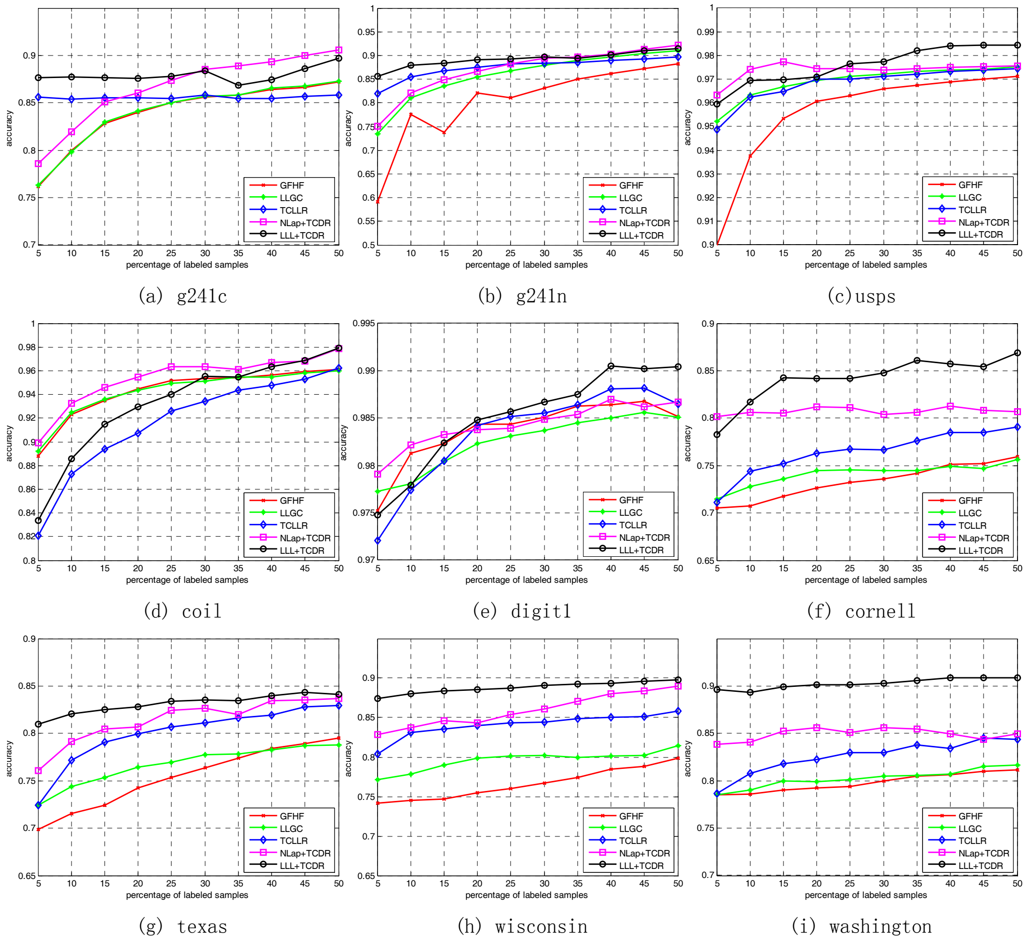
Table 3. Classification Accuracy with 10% labeled samples on the 9 data sets.

Data Sets	g241c	g241n	USPS	COIL	digit1	cornell	texas	wisconsin	washington
GFHF	0.7998	0.7763	0.9377	0.9230	0.9813	0.7075	0.7151	0.7451	0.7860
LLGC	0.7980	0.8101	0.9633	0.9250	0.9781	0.7276	0.7442	0.7787	0.7906
TCLLR	0.8541	0.8551	0.9623	0.8730	0.9774	0.7439	0.7719	0.8308	0.8078
NLap+TCDR	0.8195	0.8207	**0.9742**	**0.9323**	**0.9822**	0.8064	0.7914	0.8376	0.8407
LLL+TCDR	**0.8770**	**0.8799**	0.9694	0.8855	0.9779	**0.8173**	**0.8209**	**0.8807**	**0.8936**

It is obvious that TCDR outperforms other methods on all the data sets. In detail, we can see that NLap+TCDR outperforms LLGC consistently, while LLL+TCDR outperforms TCLLR consistently. This is due to that LLGC is the special case of NLap+TCDR and TCLLR is the special case of LLL+TCDR. And the consistent improvement indicates that clustering the features indeed benefits the classification of data points. In addition, it is worth noting that LLL+TCDR achieved higher classification accuracy than NLap+TCDR on text data sets. This indicates that *Local Learning Assumption* is more suitable for text classification than *Cluster Assumption*. The reason is probably that the data matrix of text data is very sparse, *normalized graph Laplacian* based on the distance defined in Eq.(32) may not be able to explore the geometric structure very well. In contrast, the *local learning graph Laplacian* can explore the geometric structure better in this case.

5 Conclusion & Future Work

In this paper, we present a dual regularization to explore the geometric structure in data manifold and feature manifold, along with the duality between data points and features. Furthermore, we propose a novel framework for transductive classification via dual regularization, which can be solved by alternating minimization algorithm and its convergence is theoretically guaranteed. Encouraging experimental results on benchmark semi-supervised learning data sets illustrate that the proposed methods outperform many existing approaches.



(a) g241c (b) g241n (c)usps

(d) coil (e) digit1 (f) cornell

(g) texas (h) wisconsin (i) washington

Fig. 2. Classification accuracy with respect to the proportion of labeled samples on the 9 data sets.

In the future work, we will devote to extending the dual regularization framework from transductive learning to inductive learning [8].

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.60721003, No.60673106 and No.60573062) and the Specialized Research Fund for the Doctoral Program of Higher Education. We thank the anonymous reviewers for their helpful comments.

References

1. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
2. Zhu, X.: Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison (2008)
3. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML. (2003) 912–919
4. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NIPS. (2003)
5. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. In: ICML. (2006) 985–992
6. Wu, M., Schölkopf, B.: Transductive classification via local learning regularization. In: AISTATS. (03 2007) 628–635
7. Wang, F., Li, T., Wang, G., Zhang, C.: Semi-supervised classification using local and global regularization. In: AAAI. (2008) 726–731
8. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research **7** (2006) 2399–2434
9. Smola, A.J., Kondor, R.I.: Kernels and regularization on graphs. In: COLT. (2003) 144–158
10. Gu, Q., Zhou, J.: Local learning regularized nonnegative matrix factorization. In: IJCAI. (2009)
11. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (December 2000) 2323–2326
12. Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15** (2003) 1373–1396
13. Sandler, T., Blitzer, J., Talukdar, P., Pereira, F.: Regularized learning with networks of features. In: NIPS. (2008)
14. Gu, Q., Zhou, J.: Co-clustering on manifolds. In: KDD. (2009)
15. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: KDD. (2001) 269–274
16. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: KDD. (2003) 89–98
17. Ding, C.H.Q., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: KDD. (2006) 126–135
18. Sindhwani, V., Hu, J., Mojsilovic, A.: Regularized co-clustering with dual supervision. In: NIPS. (2008) 556–562
19. Sindhwani, V., Melville, P.: Document-word co-regularization for semi-supervised sentiment analysis. In: ICDM. (2008) 1025–1030
20. Li, T., Ding, C.H.Q., Zhang, Y., Shao, B.: Knowledge transformation from word space to document space. In: SIGIR. (2008) 187–194
21. Chung, F.R.K.: Spectral Graph Theory. American Mathematical Society (February 1997)
22. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)
23. Ding, C.H., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence **99**(1) (2008)
24. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS. (2000) 556–562

Appendix: Proof of Theorem 1

Proof. We rewrite Eq.(28) as

$$L(\mathbf{F}) = \mathrm{tr}(\mathbf{F}^T\mathbf{C}\mathbf{F} - 2\mathbf{F}^T\mathbf{C}\mathbf{Y} + \lambda\mathbf{F}^T\mathbf{L}_F^+\mathbf{F} - \lambda\mathbf{F}^T\mathbf{L}_F^-\mathbf{F} \\ - 2\mathbf{F}^T\mathbf{A}^+ + 2\mathbf{F}^T\mathbf{A}^- + \mathbf{F}\mathbf{B}^+\mathbf{F}^T - \mathbf{F}\mathbf{B}^-\mathbf{F}^T) \tag{33}$$

By applying Lemma 2, we have

$$\mathbf{tr}(\mathbf{F}^T\mathbf{C}\mathbf{F}) \le \textstyle\sum_{ij} \frac{(\mathbf{C}\mathbf{F}')_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}}, \mathbf{tr}(\mathbf{F}^T\mathbf{L}_F^+\mathbf{F}) \le \sum_{ij} \frac{(\mathbf{L}_F^+\mathbf{F}')_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}} \ ,$$
$$\mathbf{tr}(\mathbf{F}\mathbf{B}^+\mathbf{F}^T) \le \textstyle\sum_{ij} \frac{(\mathbf{F}'\mathbf{B}^+)_{ij}\mathbf{F}_{ij}^2}{\mathbf{F}'_{ij}}$$

Moreover, by the inequality $a \le \frac{(a^2+b^2)}{2b}, \forall a, b > 0$, we have

$$\mathrm{tr}(\mathbf{F}^T\mathbf{A}^-) = \sum_{ij} \mathbf{A}_{ij}^-\mathbf{F}_{ij} \le \sum_{ij} \mathbf{A}_{ij}^- \frac{\mathbf{F}_{ij}^2 + \mathbf{F}'^2_{ij}}{2\mathbf{F}'_{ij}}$$

To obtain the lower bound for the remaining terms, we use the inequality that $z \ge 1 + \log z, \forall z > 0$, then

$$\mathrm{tr}(\mathbf{F}^T\mathbf{A}^+) \ge \sum_{ij} \mathbf{A}_{ij}^+\mathbf{F}'_{ij}(1 + \log \frac{\mathbf{F}_{ij}}{\mathbf{F}'_{ij}}) \ , \ \mathrm{tr}(\mathbf{F}^T\mathbf{L}_F^-\mathbf{F}) \ge \sum_{ijk} (\mathbf{L}_F^-)_{jk}\mathbf{F}'_{ji}\mathbf{F}'_{ki}(1 + \log \frac{\mathbf{F}_{ji}\mathbf{F}_{ki}}{\mathbf{F}'_{ji}\mathbf{F}'_{ki}})$$
$$\mathrm{tr}(\mathbf{F}^T\mathbf{C}\mathbf{Y}) \ge \sum_{ij} (\mathbf{C}\mathbf{Y})_{ij}\mathbf{F}'_{ij}(1 + \log \frac{\mathbf{F}_{ij}}{\mathbf{F}'_{ij}}) \ , \ \mathrm{tr}(\mathbf{F}\mathbf{B}^-\mathbf{F}^T) \ge \sum_{ijk} \mathbf{B}_{jk}^-\mathbf{F}'_{ij}\mathbf{F}'_{ik}(1 + \log \frac{\mathbf{F}_{ij}\mathbf{F}_{ik}}{\mathbf{F}'_{ij}\mathbf{F}'_{ik}})$$

By summing over all the bounds, we can get $\mathbf{Z}(\mathbf{F}, \mathbf{F}')$, which obviously satisfies (1) $\mathbf{Z}(\mathbf{F}, \mathbf{F}') \ge J_{TCDR}(\mathbf{F})$; (2)$\mathbf{Z}(\mathbf{F}, \mathbf{F}) = J_{TCDR}(\mathbf{F})$

To find the minimum of $\mathbf{Z}(\mathbf{F}, \mathbf{F}')$, we take

$$\frac{\partial Z(\mathbf{F}, \mathbf{F}')}{\partial \mathbf{F}_{ij}} = 2\frac{(\mathbf{C}\mathbf{F}')_{ij}\mathbf{F}_{ij}}{\mathbf{F}'_{ij}} - 2(\mathbf{C}\mathbf{Y})_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}} + 2\lambda\frac{(\mathbf{L}_F^+\mathbf{F}')_{ij}\mathbf{F}_{ij}}{\mathbf{F}'_{ij}} - 2\lambda(\mathbf{L}_F^-\mathbf{F}')_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}} \\ - 2\mathbf{A}_{ij}^+\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}} + 2\mathbf{A}_{ij}^-\frac{\mathbf{F}_{ij}}{\mathbf{F}'_{ij}} + 2\frac{(\mathbf{F}'\mathbf{B}^+)_{ij}\mathbf{F}_{ij}}{\mathbf{F}'_{ij}} - 2(\mathbf{F}'\mathbf{B}^-)_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}}$$

and the Hessian matrix of $Z(\mathbf{F}, \mathbf{F}')$

$$\frac{\partial^2 Z(\mathbf{F}, \mathbf{F}')}{\partial \mathbf{F}_{ij}\partial \mathbf{F}_{kl}} = \delta_{ik}\delta_{jl}(2\frac{(\mathbf{C}\mathbf{F}')_{ij}}{\mathbf{F}'_{ij}} + 2(\mathbf{C}\mathbf{Y})_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}^2} + 2\lambda\frac{(\mathbf{L}_F^+\mathbf{F}')_{ij}}{\mathbf{F}'_{ij}} + 2\lambda(\mathbf{L}_F^-\mathbf{F}')_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}^2} \\ + 2\mathbf{A}_{ij}^+\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}^2} + 2\frac{\mathbf{A}_{ij}^-}{\mathbf{F}'_{ij}} + 2\frac{(\mathbf{F}'\mathbf{B}^+)_{ij}}{\mathbf{F}'_{ij}} + 2(\mathbf{F}'\mathbf{B}^-)_{ij}\frac{\mathbf{F}'_{ij}}{\mathbf{F}_{ij}^2})$$

which is a diagonal matrix with positive diagonal elements. Thus $Z(\mathbf{F}, \mathbf{F}')$ is a convex function of $\mathbf{F}$. Therefore, we can obtain the global minimum of $Z(\mathbf{F}, \mathbf{F}')$ by setting $\frac{\partial Z(\mathbf{F}, \mathbf{F}')}{\partial \mathbf{F}_{ij}} = 0$ and solving for $\mathbf{F}$, from which we can get Eq.(29).