Discriminative Semi-Supervised Feature Selection via Manifold Regularization

Zenglin Xu[†] **Rong Jin**[‡] **Michael R. Lyu**[†] **Irwin King**[†]
[†] Computer Science & Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{zlxu, lyu, king}@cse.cuhk.edu.hk

[‡]Computer Science & Engineering
Michigan State University
East Lansing, MI, 48824
rongjin@cse.msu.edu

Abstract

We consider the problem of semi-supervised feature selection, where we are given a small amount of labeled examples and a large amount of unlabeled examples. Since a small number of labeled samples are usually insufficient for identifying the relevant features, the critical problem arising from semi-supervised feature selection is how to take advantage of the information underneath the unlabeled data. To address this problem, we propose a novel discriminative semi-supervised feature selection method based on the idea of manifold regularization. The proposed method selects features through maximizing the classification margin between different classes and simultaneously exploiting the geometry of the probability distribution that generates both labeled and unlabeled data. We formulate the proposed feature selection method into a convex-concave optimization problem, where the saddle point corresponds to the optimal solution. To find the optimal solution, the level method, a fairly recent optimization method, is employed. We also present a theoretic proof of the convergence rate for the application of the level method to our problem. Empirical evaluation on several benchmark data sets demonstrates the effectiveness of the proposed semi-supervised feature selection method.

1 Introduction

Feature selection has been playing an important role in both research and application communities of machine learning [Guyon and Elisseeff, 2003]. It has been employed in a variety of real-world applications, such as natural language processing, image processing and bioinformatics, where high dimensionality of data is usually observed. It is also used in distributed communication systems and sensor networks, where each mobile equipment or sensor has very limited computational power. Overall, feature selection is a very important method that is often applied to reduce the computational cost or to save storage space, for problems with either high dimensionality or limited resources.

Feature selection can be conducted in a supervised or unsupervised manner, in terms of whether the label information is utilized to guide the selection of relevant features. Generally, supervised feature selection methods require a large amount of labeled training data. It however could fail to identify the relevant features that are discriminative to different classes, provided the number of labeled samples is small. On the other hand, while unsupervised feature selection methods could work well with unlabeled training data, they ignore the label information and therefore are often unable to identify the discriminative features. Given the high cost in manually labeling data, and at the same time abundant unlabeled data are often easily accessible, it is desirable to develop feature selection methods that are capable of exploiting both labeled and unlabeled data. This motivates us introduce semi-supervised learning into the feature selection process. Particularly, the method of semi-supervised SVM with manifold regularization has demonstrated good performance [Belkin *et al.*, 2006]. In this work, we try to employ the idea of manifold regularization to semi-supervised feature selection.

Semi-supervised feature selection studies how to better identify the relevant features that are discriminative to different classes by effectively exploring the information underlying the huge amount of unlabeled data. In [Zhao and Liu, 2007], a filter-based semi-supervised feature selection method was proposed, which rank features via some information measure. As argued in [Guyon and Elisseeff, 2003], the filter-based feature selection could discard important features that are less informative by themselves but are informative when combined with other features. Moreover, it can also ignore the underlying learning algorithm that is used to train classifiers from labeled data. Therefore, it is hard to find features that are particularly useful to a given learning algorithm. To avoid these disadvantages, we propose a novel semi-supervised feature selection method based on the idea of manifold regularization.

In the proposed method, an optimal subset of features is identified by maximizing a performance measure that combines classification margin with manifold regularization. The manifold regularization in the proposed feature selection method assures that the decision function is smooth on the manifold constructed by the selected features of the unlabeled data. This therefore better exploits the underlying structural information of the unlabeled data. Moreover, we success-

fully formulate the presented semi-supervised feature selection method into a concave-convex problem, where the saddle point corresponds to the optimal solution. We then derive an extended level method [Xu *et al.*, 2009], a fairly recent optimization method, to find the optimal solution of the concave-convex problem. The proof of the convergence rate is also presented in this work. Finally, experiments on several benchmark data sets indicate the promising results of the proposed method in comparison with the state-of-the-art approaches for feature selection.

The rest of this paper is organized as follows. In Section 2, we review previous work on feature selection. In Section 3, we derive the discriminative semi-supervised feature selection model. We then successfully employ the level method to solve the optimization problem for semi-supervised feature selection. Section 4 presents the experimental evaluation on the proposed semi-supervised feature selection method, followed by the conclusion in Section 5.

2 Related work

The goal of feature selection is to choose a subset of features that maximizes a generalized performance criterion. A typical performance criterion is the maximum margin criterion [Guyon and Elisseeff, 2003], which naturally leads to SVM. An example based on the maximum margin criterion is SVM-Recursive Feature Elimination (SVM-RFE) [Guyon *et al.*, 2002] where features with smallest weights were removed iteratively. In [Fung and Mangasarian, 2000], L_1-norm of weights in SVM was suggested to replace L_2-norm for feature selection when learning an SVM model. In addition, several studies such as [Weston *et al.*, 2003] explored L_0-norm when computing the weights of features. Compared with supervised feature selection, unsupervised feature selection is more challenging in that there is no categorical information available [Dy and Brodley, 2004].

Extended from supervised feature selection and unsupervised feature selection, semi-supervised feature selection works on both the labeled data and the unlabeled data. Traditional semi-supervised feature selection algorithms are almost either filter-based methods including spectral analysis [Zhao and Liu, 2007] , or search based methods including forward search [Ren *et al.*, 2008]. These methods usually neglect the interaction among features and the interaction between the feature selection heuristics and the corresponding classifier. Instead, our proposed semi-supervised feature selection method works in an embedded way: the feature selection process is integrated to the semi-supervised classifier by taking advantage of manifold regularization. This therefore takes good care of the correlation among features and the integration between the features and the semi-supervised classifiers.

3 Semi-supervised Feature Selection Model

Before presenting the semi-supervised feature selection model, we firstly introduce the notations that will be used throughout this paper. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{R}^{n \times d}$ denote the entire data set, which consists of n data points in d-dimensional space. The data set includes both the labeled examples and the unlabeled ones. We assume that the first l examples within $\mathbf{X}$ are labeled by $\mathbf{y} = (y_1, y_2, \ldots, y_l)$ where $y_i \in \{-1, +1\}$ represents the binary class label assigned to $\mathbf{x}_i$. For convenience, we also denote the collection of labeled examples by $\mathbf{X}_\ell = (\mathbf{x}_1, \ldots, \mathbf{x}_l)$, and the unlabeled examples by $\mathbf{X}_u$, such that $\mathbf{X} = (\mathbf{X}_\ell, \mathbf{X}_u)$. We then introduce the indicator variable $\mathbf{p}$, where $\mathbf{p} = (p_1, \ldots, p_d)^\top$ and $p_i \in \{0, 1\},\ i = 1, \ldots, d$, to represent which features are selected. We further introduce a diagonal matrix $\mathrm{D}(\mathbf{p}) = \mathrm{diag}(p_1, \ldots, p_d)$. Then the input data are now represented as $\mathbf{X}\mathrm{D}(\mathbf{p})$. In order to indicate that m features are selected, we will have $\mathbf{p}^\top \mathbf{e} = m$.

It is important to note that determining the number of selected features is a model selection problem, which is beyond the scope of this study. In this work, we assume that the number of selected features, i.e., m, has been decided by an external oracle. It should also be noted that the number of required features usually is dependent on the objective of the task, and there is no single number of features that are optimal for all tasks.

3.1 Semi-supervised SVM Based on Manifold Regularization

Following the framework of manifold regularization [Belkin *et al.*, 2006], a semi-supervised SVM can be obtained by penalizing a regularization term defined as:

$$\|\mathbf{f}\|_I^2 \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{n} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij} = \mathbf{f}^\top \mathcal{L} \mathbf{f},$$

where W_{ij} are the edge weights defined on a pair of nodes $(\mathbf{x}_i, \mathbf{x}_j)$ of the adjacency graph. $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]$ denotes the decision function values over all data examples. The graph Laplacian $\mathcal{L}$ is defined as $\mathcal{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix and $D_{ii} = \sum_{j=1}^{n} W_{ij}$. According to [Belkin *et al.*, 2006], $\|\mathbf{f}\|_I^2$ indeed reflects the smoothness of the decision function with respect to the marginal distribution of $\mathbf{X}$.

Considering a linear SVM where the decision function can be represented as $f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i - b$, the manifold regularization term $\|\mathbf{f}\|_I^2$ is equal to $\mathbf{w}^\top \mathbf{X}^\top \mathcal{L} \mathbf{X} \mathbf{w}$. Note that the bias term b has no effect on calculating the regularization term. Then, the semi-supervised SVM can be represented as follows:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{l} \xi_i + \frac{\rho}{2} \mathbf{w}^\top \mathbf{X}^\top \mathcal{L} \mathbf{X} \mathbf{w} \qquad (1) \\ \text{s. t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \xi_i,\ i = 1, \ldots, l, \\ & \xi_i \geq 0,\ i = 1, \ldots, l, \end{aligned}$$

where ξ denotes the margin error and ρ is a trade-off parameter between the two regularization terms of $\mathbf{w}$ satisfying $\rho \geq 0$.

In order to efficiently solve the optimization problem (1), we calculate its dual. We therefore introduce the following lemma:

Lemma 1. *The dual problem of (1) can be written as:*

$$\max_{\alpha \in \mathcal{Q}} \quad \alpha^\top \mathbf{e} - \frac{1}{2}(\alpha \circ \mathbf{y})^\top \mathbf{X}_\ell (\mathbf{I} + \rho \mathbf{X}^\top \mathcal{L} \mathbf{X})^{-1} \mathbf{X}_\ell^\top (\alpha \circ \mathbf{y})$$

where $\mathcal{Q} = \{\alpha \in [0, C]^l | \alpha^\top \mathbf{y} = 0\}$, $\mathbf{I} \in \mathbb{R}^{n \times n}$ *is the identity matrix, and* $\circ$ *is an operator of element-wise product.*

3.2 Semi-supervised Feature Selection Model Based on Manifold Regularization

We have Proposition 1 to describe the optimization problem with respect to the feature indicator and the decision function.

Proposition 1. *The optimal feature subset for the problem (1) can be obtained by solving the following combinatorial problem:*

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\mathbf{p}\in\mathcal{P}} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{l}\xi_i \\ & + \frac{\rho}{2}\mathbf{w}^\top \mathrm{D}(\mathbf{p})\mathbf{X}^\top \mathcal{L}\mathbf{X}\mathrm{D}(\mathbf{p})\mathbf{w} \\ \text{s. t.} \quad & y_i(\mathbf{w}^\top \mathrm{D}(\mathbf{p})\mathbf{x}_i - b) \geq 1 - \xi_i,\ i = 1, \ldots, l, \\ & \xi_i \geq 0,\ i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $\mathcal{P} = \{\mathbf{p} \in \{0, 1\}^d | \mathbf{p}^\top \mathbf{e} = m\}$.

To simplify the presentation, we introduce a matrix $\mathbf{Z}$ as follows

$$\mathbf{Z} = \mathbf{X}^\top \mathcal{L}\mathbf{X} \tag{3}$$

For the convenience of discussion, we assume matrix $\mathbf{Z}$ is non-singular, although the derivation below can be easily extended to the singular case by simply replacing matrix inverse with matrix pseudo inverse.

The theorem below shows that (2) can be reformulated into a min-max optimization, which is the key for speeding up the computation.

Theorem 1. *The problem in (2) is equivalent to the following min-max optimization problem*

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\alpha\in\mathcal{Q}} \phi(\mathbf{p}, \alpha) \tag{4}$$

where $\phi(\mathbf{p}, \alpha)$ *is defined as*

$$\begin{aligned} \phi(\mathbf{p}, \alpha) &= \alpha^\top \mathbf{e} \\ &- \frac{1}{2\rho}(\alpha \circ \mathbf{y})^\top \mathbf{X}_\ell \left(\mathbf{Z}^{-1} - [\mathbf{Z} + \rho\mathbf{Z}\mathrm{D}(\mathbf{p})\mathbf{Z}]^{-1}\right) \mathbf{X}_\ell^\top (\alpha \circ \mathbf{y}). \end{aligned}$$

When ρ *is very small (i.e.,* $\rho \ll 1$*),* $\phi(\mathbf{p}, \alpha)$ *is approximated as*

$$\phi(\mathbf{p}, \alpha) = \alpha^\top \mathbf{e} - \frac{1}{2}(\alpha \circ \mathbf{y})^\top \mathbf{X}_\ell \mathrm{D}(\mathbf{p})\mathbf{X}_\ell^\top (\alpha \circ \mathbf{y}).$$

The proof will appear in the long version of this paper. As indicated by the above theorem, when ρ is small, the manifold regularization term can be ignored.

One of the major challenges in solving the optimization problem in (2), or the equivalence in (4) arises from the constraint that $\{p_i\}_{i=1}^d$ have to be binary variables. To avoid the combinatorial nature of the problem, we relax the binary variable $p_i \in \{0, 1\}$ to a continuous one, i.e., $p_i \in [0, 1]$, and convert the discrete optimization problem in (4) into the following continuous optimization problem

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\alpha\in\mathcal{Q}} \phi(\mathbf{p}, \alpha) \tag{5}$$

where domain $\mathcal{P}$ is changed to

$$\mathcal{P} = \{\mathbf{p} \in [0, 1]^d | \mathbf{p}^\top \mathbf{e} = m\}.$$

Theorem 2. *The problem in (5) is indeed a convex-concave optimization problem, and therefore its optimal solution is the saddle point of* $\phi(\mathbf{p}, \alpha)$.

The proof of this theorem will appear in the long version of this paper. As indicated by the above theorem, the problem in (5) is essentially a convex problem and therefore its global optimal solution can be found via standard techniques.

Although (5) is a convex-concave optimization problem with a guarantee to find the global optimal solution, solving it efficiently is very challenging. To reduce the computational complexity, in the following proposition, we consider a variant of min-max optimization problem for (5).

Proposition 2. *(5) is equivalent to the following min-max optimization problem*

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\alpha\in\mathcal{Q}} h(\mathbf{p}, \alpha) \tag{6}$$

where

$$h(\mathbf{p}, \alpha) = \alpha^\top \mathbf{e} - \frac{1}{2}(\alpha \circ \mathbf{y})^\top \mathbf{X}_\ell \Gamma \mathbf{X}_\ell^\top (\alpha \circ \mathbf{y}) \tag{7}$$

and Γ *is defined as*

$$\Gamma = \mathrm{D}(\mathbf{p})\left(\mathbf{I} + \rho\mathbf{Z}\right)^{-1}\mathrm{D}(\mathbf{p}). \tag{8}$$

We then proceed to simplify Γ in $h(\mathbf{p}, \alpha)$. The proposition provides a simple upper bound for Γ.

Proposition 3. *We introduce the matrix* $\mathbf{A}$ *as*

$$\mathbf{A} = (1 - \tau)^2 \mathrm{D}(\mathbf{p}) + \frac{\tau^2}{\rho}\mathbf{Z}^{-1} \tag{9}$$

where τ *is a parameter. We have* $\mathbf{A} \succeq \Gamma$ *for any* $\tau \in [0, 1]$.

Using the result in Proposition 3, we replace Γ with $\mathbf{A}$, which results in the following optimization problem

$$\min_{\mathbf{p}\in\mathcal{P}} \max_{\alpha\in\mathcal{Q},\tau\in[0,1]} \alpha^\top \mathbf{e} - \frac{1}{2}(\alpha \circ \mathbf{y})^\top \mathbf{X}_\ell \mathbf{A} \mathbf{X}_\ell^\top (\alpha \circ \mathbf{y}) \tag{10}$$

Because $\mathbf{A}$ is linear in $\mathbf{p}$, (10) is substantially simpler to solve than (6). The following lemma reveals the relationship between (6) and (10).

Proposition 4. *The optimal value of (10) provides a lower bound for the optimal value of (6)*

It is interesting to examine (10) with a fixed τ. When $\tau = 0$, the problem in (10) is reduced to a supervised feature selection algorithm. Now we can use (10) to approximate (6).

3.3 Optimization Method

Before introducing the optimization method to solve the optimization problem, we first discuss the relationship between the model of semi-supervised feature selection and multiple kernel learning [Lanckriet *et al.*, 2004]. Note that for a linear kernel, the kernel matrix $\mathbf{K}$ can be written as:

$$\mathbf{K} = \mathbf{X}_\ell \mathbf{X}_\ell^\top = \sum_{i=1}^{d} \mathbf{v}_i \mathbf{v}_i^\top = \sum_{i=1}^{d} \mathbf{K}_i,$$

where $\mathbf{v}_i$ is the ith feature of $\mathbf{X}_\ell$. The term $\mathbf{K}_i = \mathbf{v}_i \mathbf{v}_i^\top$ can then be regarded as a base kernel which is calculated on

a single feature. Therefore, the term $(1-\tau)^2\mathbf{X}_\ell \mathrm{D}(\mathbf{p})\mathbf{X}_\ell^\top$ can be written as $(1-\tau)^2\sum_{i=1}^d p_i\mathbf{K}_i$. We further define $\mathbf{H} = \mathbf{X}_\ell^\top(\mathbf{X}^\top\mathcal{L}\mathbf{X})^{-1}\mathbf{X}_\ell$ which can be seen as a kernel matrix defined on the entire data set. The overall optimization problem can be formulated as, by switching $\mathbf{p}$ and τ, $\max_{0\le\tau\le 1}\psi(\tau)$, where $\psi(\tau)$ is defined as

$$\min_{\mathbf{p}\in\mathcal{P}}\max_{\alpha\in\mathcal{Q}} \quad \alpha^\top\mathbf{e} - \tfrac{1}{2}(\alpha\circ\mathbf{y})^\top\mathbf{M}(\alpha\circ\mathbf{y}) \tag{11}$$

where

$$\mathbf{M} = (1-\tau)^2\sum_{i=1}^d p_i\mathbf{K}_i + \frac{\tau^2}{\rho}\mathbf{H}.$$

Therefore, the optimization problem (11) is related to a kernel learning problem. According to [Lanckriet *et al.*, 2004], the dual problem of $\psi(\tau)$ can be formulated as an semi-definite programming (SDP) problem. However, the SDP problem involves high computational and storage complexity.

Indeed, (11) can be regarded as a concave-convex problem, since (11) is concave in α and convex in $\mathbf{p}$. The saddle point of (11) corresponds to the optimal solution. According to the literatures of multiple kernel learning and convex optimization, we can formulate an alternating procedure to solve the concave-convex problem: in each iteration, the solution of α and that of $\mathbf{p}$ are alternatively optimized. Several optimization methods, such as the cutting plane method [Sonnenburg *et al.*, 2006], the subgradient descent method [Rakotomamonjy *et al.*, 2008], and the level method [Xu *et al.*, 2009], could be employed. Among them, the level method has shown its significant improvements over the other two methods on the convergence speed [Xu *et al.*, 2009]. Although there are some simplification methods of SDP could lead to a QP problem [Cortes *et al.*, 2008] due to the constraints on $\mathbf{p}$, they may not apply to our case. In the following, we discuss how to derive an extended level method to solve the concave-convex optimization problem related to semi-supervised feature selection.

To facilitate the description, we denote the objective function of (11) as follows:

$$\varphi(\mathbf{p},\alpha) \;=\; \alpha^\top\mathbf{e} - \frac{1}{2}(\alpha\circ\mathbf{y})^\top\mathbf{M}(\alpha\circ\mathbf{y}). \tag{12}$$

For the optimal solution $(\mathbf{p}^*,\alpha^*)$, we have

$$\begin{aligned}\varphi(\mathbf{p},\alpha^*) &= \max_{\alpha\in\mathcal{Q}}\varphi(\mathbf{p},\alpha)\\ &\ge \varphi(\mathbf{p}^*,\alpha^*) \ge \varphi(\mathbf{p}^*,\alpha) = \min_{\mathbf{p}\in\mathcal{P}}\varphi(\mathbf{p},\alpha).\end{aligned}$$

The level method iteratively updates both the lower and the upper bounds for $\varphi(\mathbf{p},\alpha)$ in order to find the saddle point.

To obtain the bounds, we first construct the cutting plane model. Let $\{\mathbf{p}^j\}_{j=1}^i$ denote the solutions for $\mathbf{p}$ obtained in the last i iterations. Let $\alpha^j = \arg\max_{\alpha\in\mathcal{Q}}\varphi(\mathbf{p}^j,\alpha)$ denote the optimal solution that maximizes $\varphi(\mathbf{p}^j,\alpha)$. We calculate the gradient of $\varphi(\mathbf{p},\alpha)$ over $\mathbf{p}$ as the follows: $\nabla_{\mathbf{p}}\varphi(\mathbf{p},\alpha) = -\frac{1}{2}[(\alpha\circ\mathbf{y})^\top\mathbf{K}_1(\alpha\circ\mathbf{y}),\ldots,(\alpha\circ\mathbf{y})^\top\mathbf{K}_d(\alpha\circ\mathbf{y})]^\top$. We construct a cutting plane model $g^i(\mathbf{p})$ as follows:

$$g^i(\mathbf{p}) = \max_{1\le j\le i}\varphi(\mathbf{p}^j,\alpha^j) + (\mathbf{p}-\mathbf{p}^j)^\top\nabla_{\mathbf{p}}(\mathbf{p}^j,\alpha^j) \tag{13}$$

Next, we construct both the lower and the upper bounds for the optimal value $\varphi(\mathbf{p}^*,\alpha^*)$. We define two quantities $\underline{\varphi}^i$ and $\overline{\varphi}^i$ as follows:

$$\underline{\varphi}^i \;=\; \min_{\mathbf{p}\in\mathcal{P}} g^i(\mathbf{p}) \tag{14}$$

$$\overline{\varphi}^i \;=\; \min_{1\le j\le i}\varphi(\mathbf{p}^j,\alpha^j). \tag{15}$$

It can be shown that $\{\underline{\varphi}^j\}_{j=1}^i$ and $\{\overline{\varphi}^j\}_{j=1}^i$ provide a series of increasingly tight bounds for $\varphi(\mathbf{p}^*,\alpha^*)$ according to [Xu *et al.*, 2009].

We furthermore define the gap Δ^i as

$$\Delta^i = \overline{\varphi}^i - \underline{\varphi}^i. \tag{16}$$

In the third step, we define the current level as $\ell^i = \lambda\overline{\varphi}^i + (1-\lambda)\underline{\varphi}^i$. We then construct the level set $\mathcal{L}^i$ using the estimated bounds $\overline{f}^i$ and $\underline{\varphi}^i$ as follows:

$$\mathcal{L}^i = \{\mathbf{p}\in\mathcal{P} : g^i(\mathbf{p}) \le \ell^i\}, \tag{17}$$

where $\lambda\in(0,1)$ is a predefined constant. The new solution, denoted by $\mathbf{p}^{i+1}$, is computed as the projection of $\mathbf{p}^i$ onto the level set $\mathcal{L}^i$, which is equivalent to solving the following optimization problem:

$$\min_{\mathbf{p}\in\mathcal{L}^i}\|\mathbf{p}-\mathbf{p}^i\|_2^2 \tag{18}$$

The projection can often be solved efficiently, since only very few linear constraints of $\mathcal{L}$ are active. This sparse nature usually leads to significant speedup.

We summarize the steps of the level method for semi-supervised feature selection in Algorithm 1.

Algorithm 1 Level method for semi-supervised feature selection

1: Initialize $\mathbf{p}^0 = \frac{m}{d}\mathbf{e}$ and $i=0$
2: **repeat**
3: Obtain α^i by solving SVM with $\mathbf{M} = (1-\tau)^2\mathbf{X}_\ell\mathrm{D}(\mathbf{p}^i)\mathbf{X}_\ell^\top + \frac{\tau^2}{\rho}\mathbf{H}$
4: Construct the cutting plane model $g^i(\mathbf{p})$ in (13)
5: Calculate the lower bound $\underline{\varphi}^i$ and the upper bound $\overline{\varphi}^i$ in (15), and the gap Δ^i in (16)
6: Obtain $\mathbf{p}^{i+1}$ via the projection step (18)
7: **until** $\Delta^i \le \varepsilon$

Finally, we show the convergence behavior of the level method for semi-supervised feature selection in Theorem 3.

Theorem 3. *To obtain a solution* $\mathbf{p}$ *that satisfies the stopping criterion, i.e.,* $|\max_{\alpha\in\mathcal{Q}}\varphi(\mathbf{p},\alpha) - \varphi(\mathbf{p}^*,\alpha^*)| \le \varepsilon$, *the maximum number of iterations* N *that the level method requires is bounded by* $N \le \frac{2c(\lambda)L^2}{\varepsilon^2}$, *where* $c(\lambda) = \frac{1}{(1-\lambda)^2\lambda(2-\lambda)}$, $L = \frac{1}{2}\sqrt{d}C^2\max_{1\le i\le d}|\mathbf{v}_i|^2$.

4 Experiments

We denote by **FS-Manifold** the proposed discriminative feature selection method based on manifold regularization. We

compare our algorithm with the following state-of-the-art approaches for feature selection: **Fisher** [Bishop, 1995], L_0**-SVM** [Weston *et al.*, 2003] and L_1**-SVM** [Fung and Mangasarian, 2000]. The description of the selected comparison methods is as follows:

- **Fisher** [Bishop, 1995] calculates a Fisher/Correlation score for each feature,
- L_0**-SVM** [Weston *et al.*, 2003] approximates the L_0-norm by minimizing a logarithm function,
- L_1**-SVM** [Fung and Mangasarian, 2000] replaces L_2-norm of the weights $\mathbf{w}$ with L_1-norm in SVM and leads to a sparse solution.

For all the comparison methods, features with the largest scores are selected. SVM is used as the evaluation classifier since it is usually regarded as the state-of-the-art classification method.

It is important to note that we also compare the above methods with the semi-supervised feature selection method proposed in [Zhao and Liu, 2007], which selects features according to the spectral and the normalized mutual information. However, due to the weak interaction among features and the class labels, it is instable in the scenario of small training samples and it usually performs significantly worse than $L0-SVM$. Therefore, we do not include its results in this work.

4.1 Experimental Settings

We adopt two types of data sets: digit characters and text documents. For the data sets of digits, we select three tasks from the USPS data set, i.e., *4 vs 7*, *2 vs 3*, and *3 vs 8*, to make the learning tasks more challenging. For each task, we randomly select 400 digit images to form the data set. For the data sets of text documents, three subsets, i.e., *auto vs motor*, *baseball vs hockey*, and *gun vs mideast*, of text documents are selected from the 20-Newsgroups repository. For convenience, we denote them by DS1, DS2, and DS3, respectively.

For both types of data sets, the training examples are randomly selected such that each category has the same number of examples. The remaining examples are then used as the test data. The test data are also used as the unlabeled data for the semi-supervised feature selection algorithm. As the USPS data sets are used to examine how the property of features changes with the number of labeled examples, we vary the number of training examples within the set of $\{6, 10, 20, 30, 40\}$. For each setting of the training samples, the number of selected features is set to 10 and 20, respectively. This is because a small number of features (pixels) are enough to identify the digits. For the text data sets, we fix the number of training document to be 50, since the scales of the text data sets are significantly larger than those of the USPS data sets. For each text data set, we consider two settings that the number of required features is equal to 50 and 100, respectively. It is interesting to note that the features (words) in the text data sets are very sparse and therefore more features are needed to represent the documents. In all cases, every experiment is repeated with 30 random trials.

We select parameters C trade-off parameter τ by a 5-fold cross validation. The parameter ρ is fixed to 10, since the trade-off is naturally cared by the parameter τ. We adopt the Cosine similarity measure and the binary weights to construct the graph. The number of neighbors is set to 20 for all cases. In addition, we set the parameter λ in the level method to 0.9 since a larger λ value accelerates the convergence of the algorithm.

4.2 Experimental Results

We plot the results on the USPS data sets in Figures 1 and 2, when the number of required features is set to 10 and 20, respectively. Firstly, it can be observed that the maximum margin based methods (SVM-based methods) usually perform better in identifying the discriminative features comparing with the non-SVM based method, **Fisher**. For example, for the task of *4 vs 7*, when the number of training samples is 30 and the number of required features is 10, the improvement of **FS-Manifold** over **Fisher** is over 3%. This indicates the advantage of embedding the feature selection process to the classifier. Secondly, compared with the supervised feature selection methods, **FS-Manifold** achieves promising test accuracy. In a number of cases, **FS-Manifold** outperforms the supervised feature selection methods. This is because the information supplied by the manifold structure of the unlabeled data helps to identify the global smooth features where data lie in.

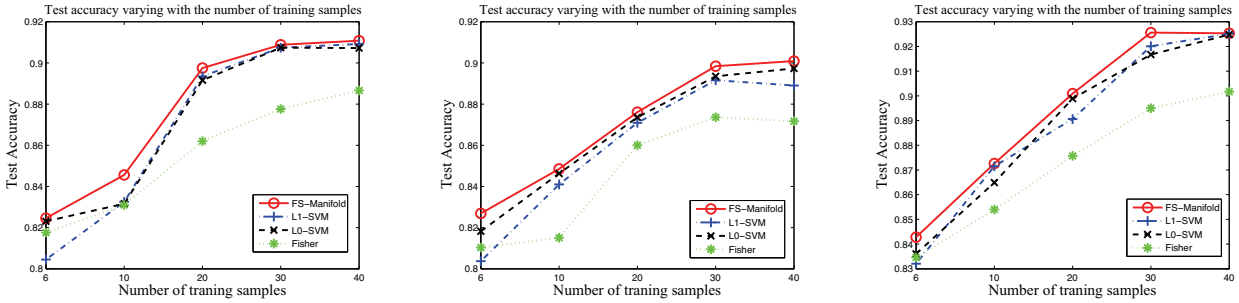
We then report the averaged prediction accuracy and the standard deviation on the text data sets in Table 1. We can observe that the proposed semi-supervised feature selection method performs better than other methods in almost all cases. For example, in the *gun vs mideast* data set, the improvement of **FS-Manifold** over **Fisher** is nearly 4% when the number of selected features is equal to 50. Furthermore, it is important to note that, for each data set, **FS-Manifold** achieves smaller deviation values than other feature selection methods. This phenomenon, which may be due to the global smoothness induced by the manifold regularization, suggests that **FS-Manifold** is more robust in selecting features.

5 Conclusion

We have presented a discriminative semi-supervised feature selection method via manifold regularization. The proposed method selects features through maximizing the margin between different classes and at the same time exploiting the geometry of the probability distribution that generates the data. We successfully formulate the resulting semi-supervised feature selection method as a concave-convex optimization problem, where the saddle point corresponds to the optimal solution. We then derive an extended level method to find the optimal solution of the concave-convex problem. Empirical evaluation with several benchmark data sets demonstrates the effectiveness of our proposed feature selection method over the state-of-the-art feature selection methods.

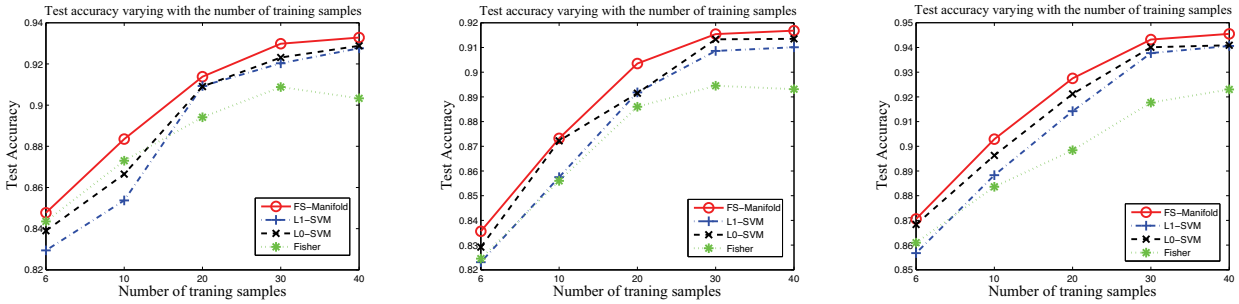
Acknowledgement

The work was supported by the National Science Foundation (IIS-0643494), National Institute of Health (1R01GM079688-01) and Research Grants Council of Hong Kong (CUHK4158/08E and CUHK4128/08E).



(a) *2 vs 3* (b) *3 vs 8* (c) *4 vs 7*

Figure 1: The comparison among different feature selection algorithms when the number of selected features is equal to 10.



(a) *2 vs 3* (b) *3 vs 8* (c) *4 vs 7*

Figure 2: The comparison among different feature selection algorithms when the number of selected features is equal to 20.

Table 1: The classification accuracy (%) on text data sets. The best result, and those not significantly worse than it (t-test with 95% confidence level), are highlighted.

Data	#F	FS-Manifold	L_1-SVM	L_0-SVM	Fisher
DS1	50	**82.9**±2.4	82.2±2.9	82.3±2.9	82.3±3.5
	100	**83.5**±2.2	82.9±2.6	**83.2**±2.6	**83.4**±2.6
DS2	50	**89.7**±3.9	88.7±8.6	89.1±4.9	**89.8**±6.9
	100	**91.1**±3.4	**90.9**±5.8	90.3±3.7	90.3±5.6
DS3	50	**84.2**±4.3	82.0±4.4	82.9±4.3	81.3±4.7
	100	**85.8**±3.9	84.1±4.2	85.2±4.4	84.3±4.1

References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Bishop, 1995] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, London, 1995.

[Cortes *et al.*, 2008] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Learning sequence kernels. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 2–8, 2008.

[Dy and Brodley, 2004] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[Fung and Mangasarian, 2000] Glenn Fung and Olvi L. Mangasarian. Data selection for support vector machine classifiers. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 64–70, New York, NY, USA, 2000. ACM.

[Guyon and Elisseeff, 2003] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[Guyon *et al.*, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[Lanckriet *et al.*, 2004] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[Rakotomamonjy *et al.*, 2008] Alain Rakotomamonjy, Francis R. Bach, Stéphane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[Ren *et al.*, 2008] Jiangtao Ren, Zhengyuan Qiu, Wei Fan, Hong Cheng, and Philip S. Yu. Forward semi-supervised feature selection. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '08)*, pages 970–976, 2008.

[Sonnenburg *et al.*, 2006] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[Weston *et al.*, 2003] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

[Xu *et al.*, 2009] Zenglin Xu, Rong Jin, Irwin King, and Michael Lyu. An extended level method for efficient multiple kernel learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1825–1832. 2009.

[Zhao and Liu, 2007] Zheng Zhao and Huan Liu. Semi-supervised feature selection via spectral analysis. In *SDM*, pages 641–646, 2007.