2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Semi-Supervised Local-Learning-based Feature Selection

Jim Jing-Yan Wang, Jin Yao and Yijun Sun

Abstract— **Local-learning-based feature selection has been successfully applied to high-dimensional data analysis. It utilizes class labels to define a margin for each data sample and selects the most discriminative features by maximizing the margins with regard to a feature weight vector. However, it requires that all data samples are labeled, which makes it unsuitable for semi-supervised learning where only a handful of training samples are labeled while most are unlabeled. To address this issue, we herein propose a new semi-supervised local-learning-based feature selection method. The basic idea is to learn the class labels of unlabeled samples in a new feature subspace induced by the learned feature weights, and then use the learned class labels to define the margins for feature weight learning. By constructing and optimizing a unified objective function, the feature weights and class labels are learned simultaneously in an iterative algorithm. The experiments performed on some benchmark data sets show the advantage of the proposed algorithm over stat-of-the-art semi-supervised feature selection methods.**

I. Introduction

HUGH-throughput technologies now routinely produce large data sets characterized by unprecedented numbers of features. Accordingly, feature selection has become increasingly important in a wide range of scientific disciplines. One example where feature selection plays a critical role, is the use of oligonucleotide microarray for the identification of cancer-associated gene expression profiles of diagnostic or prognostic value [1], [2], [3]. Typically, the number of samples is less than a few hundreds, while the number of features associated with the raw data is in the order of thousands or even tens of thousands. Amongst this enormous number of genes, only a small fraction is likely to be relevant for cancerous tumor growth and/or spread. The abundance of irrelevant features poses serious problems for existing machine learning algorithms, and represents one of the most recalcitrant problems for their applications in oncology and other scientific disciplines dealing with copious features.

Numerous feature selection algorithms have been developed in the past twenty years. One famous approach is the multi-level model proposed in [4]. This method can automatically find a sub-space feature set that is close to the ground-truth feature set, thus make a huge contribution in the NMF-based feature selection field. One of the most effective methods for high-dimensional data analysis is the recently proposed local-learning-based method [5]. The key idea is to decompose an arbitrarily complex nonlinear problem into a set of locally linear ones through local learning, and then learn feature relevance globally within the large margin framework. Specifically, for each sample, it first defines the nearest neighbor from the same class as its nearest hit (NH) and the nearest neighbor from a different class as its nearest miss (NM). Then a margin is defined as the difference between the distances of the sample to NM and NH in a weighted feature space, and a feature weight vector is learned by maximizing the margins with regard to the feature weight vector. This method is very simple and computationally efficient. It has been demonstrated that it can achieve close-to-optimal solution even in the presence of one million irrelevant features. The method has been successfully applied to various applications [6], [7], [8], [9].

One major issue associated with the above method is that it requires that the class labels of all training samples are available. For many real-world applications, due to the high cost of obtaining sample labels, only a handful of training samples are labeled while most are unlabeled. In the machine learning literature, this is called semi-supervised learning [10], [11], [12], [13]. In this paper, we develop a new feature selection method where we extend the concept of local learning to semi-supervised learning settings. The basic idea is to learn the class labels of unlabeled samples in a new feature subspace induced by the learned feature weights, and then use the learned class labels to define the margins for feature weight learning. By constructing and optimizing a unified objective function, the feature weights and class labels are learned simultaneously in an iterative algorithm. The experiments are performed on some benchmark data sets that demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. In Section II, we present a brief review of some state-of-the-art methods for semi-supervised feature selection that we will compare with in the numeric experiment. In Section III, we give a detailed description of the proposed method. In Section IV, we evaluate our method on two real-world data sets, and the paper is concluded in Section V.

II. Literature Review

A number of methods have been recently proposed to learn useful features from both labeled and unlabeled samples. Zhao et al. [14] proposed a locality sensitive semi-supervised feature selection method by using labeled samples

Jim Jing-Yan Wang is with University at Buffalo, The State University of New York, Buffalo, NY 14203, USA, and Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, 215006, China. Jin Yao and Yijun Sun are with University at Buffalo, The State University of New York, Buffalo, NY 14203, USA. E-mail: jimjywang@gmail.com, jinyao@buffalo.edu, yijunsun@buffalo.edu.
This work is supported by Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, China (Grant No. KJS1324), and US National Science Foundation (Grant No. DBI-1322212).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

to maximize margins between data samples from different classes, while using unlabeled samples to discover the geometrical structure of the data space. Xu et al. [15] proposed the manifold regularization-based discriminative method by maximizing classification margins between different classes, and simultaneously exploiting the geometry of the probability distribution that generates both labeled and unlabeled samples. Zhao and Lu [16] presented the spectral analysis-based algorithm by exploiting both labeled and unlabeled samples through a regularization framework to address the "small labeled-sample" problem.

III. Proposed Method

A. Problem Formulation

Suppose that we have a training data set $\mathbf{X} = [\mathbf{x}_i, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i$ is the i-th sample. Without loss of generality, we assume that only the first l data samples are labeled, while the rest $n-l$ samples are unlabeled. The known labels are organized in a l-dimensional binary vector $\widehat{\mathbf{y}}_l = [\widehat{y}_1, \cdots, \widehat{y}_l] \in \{+1, -1\}^l$.

The problem of feature selection is to map the feature vector of a sample $\mathbf{x}$ to a new feature space by weighting the features using a nonnegative feature weight vector $\mathbf{w} = [w_1, \cdots, w_d]^\top \in \mathbb{R}^d$, $\mathbf{x} \mapsto \mathbf{w} \circ \mathbf{x}$, where $\circ$ is the element-wise product sign, and w_c is the weight for the c-th feature. In order to encourage the sparsity of the feature weights, we impose the constrain $\sum_{c=1}^d w_c = \alpha$ to the learning of $\mathbf{w}$. Moreover, we try to learn the class labels for the samples, which are organized in a class label vector $\mathbf{y} = [y_1, \cdots, y_n] \in \mathbb{R}^n$, where $y_i \in \mathbb{R}$ is the class label of the i-th sample. Note that instead of learning a binary class label vector, we relax it to the learning of a real-valued vector. We impose the constrain $y_i = \widehat{y}_i, \forall i = 1, \cdots, l$, to respect the known labels. We consider the following two problems to formulate the objective function to learn the feature weight vector $\mathbf{w}$ and the class label vector $\mathbf{y}$ simultaneously.

- **Large Margin Regularization**: As with our previous work [5], we use the margin of each sample in the weighted feature space induced by $\mathbf{w}$ to regularize the learning of $\mathbf{w}$. Given a data sample $\mathbf{x}_i$, we find its two types of nearest neighbors, NH and NM, in the feature space weighted by $\mathbf{w}$. The margin of $\mathbf{x}_i$ is then defined as

$$\rho_i(\mathbf{w}) = d(\mathbf{x}_i, \mathrm{NM}(\mathbf{x}_i)|\mathbf{w}) - d(\mathbf{x}_i, \mathrm{NH}(\mathbf{x}_i)|\mathbf{w}) \quad (1)$$

  where $d(\mathbf{x}_i, \mathrm{NM}(\mathbf{x}_i)|\mathbf{w})$ is the distance between $\mathbf{x}_i$ and $\mathrm{NM}(\mathbf{x}_i)$. The feature weight vector $\mathbf{w}$ is learned by maximizing the margins of the training samples with regard to $\mathbf{w}$. In the semi-supervised learning setting, since $\mathbf{w}$ and $y_i|_{i=l+1}^n$ are both unknown, it is impossible to find the exact NH and HM during the learning procedure. Instead, we first estimate the probability of a sample $\mathbf{x}_j$ being the NH or NM of $\mathbf{x}_i$ from the previously learned $\mathbf{w}$ and $\mathbf{y}$ respectively. We denote $p_{ij}^{\mathbf{w},\mathbf{y}}$ as the probability of $\mathbf{x}_j$ being the NH of $\mathbf{x}_i$ given $\mathbf{w}$ and $\mathbf{y}$, while $q_{ij}^{\mathbf{w},\mathbf{y}}$ as the probability of $\mathbf{x}_j$ being the NH of $\mathbf{x}_i$ given $\mathbf{w}$ and $\mathbf{y}$. For any pair of $(\mathbf{x}_i, \mathbf{x}_j)$, if y_i and y_j have the same sign (different signs), the probability $p_{ij}^{\mathbf{w},\mathbf{y}}$ ($q_{ij}^{\mathbf{w},\mathbf{y}}$) is estimated via the standard kernel density estimation:

$$\begin{aligned} p_{ij}^{\mathbf{w},\mathbf{y}} &= \frac{I(y_i y_j > 0)\kappa(\mathbf{x}_i, \mathbf{x}_j|\mathbf{w})}{\sum_{k=1}^n I(y_i y_k > 0)\kappa(\mathbf{x}_i, \mathbf{x}_k|\mathbf{w})} \\ q_{ij}^{\mathbf{w},\mathbf{y}} &= \frac{I(y_i y_j \le 0)\kappa(\mathbf{x}_i, \mathbf{x}_j|\mathbf{w})}{\sum_{k=1}^n I(y_i y_k \le 0)\kappa(\mathbf{x}_i, \mathbf{x}_k|\mathbf{w})} \end{aligned} \quad (2)$$

  where $I(x) = 1$ if x is true, and 0 otherwise, and $\kappa(\mathbf{x}_i, \mathbf{x}_j|\mathbf{w}) = \exp\left(-\frac{\|\mathbf{w}\circ\mathbf{x}_i - \mathbf{w}\circ\mathbf{x}_j\|_2^2}{2\sigma^2}\right)$ is a Gaussian kernel function. For the purpose of this paper, we use the squared Euclidean distance to measure the distance between $\mathbf{x}_i$ and NH or NM. Then, the margin of $\mathbf{x}_i$ can be computed as

$$\begin{aligned} \rho_i(\mathbf{w}, \mathbf{y}) =& \sum_{j=1}^n \|(\mathbf{w}\circ\mathbf{x}_i) - (\mathbf{w}\circ\mathbf{x}_j)\|_2^2 q_{ij}^{\mathbf{w},\mathbf{y}} \\ & - \sum_{j=1}^n \|(\mathbf{w}\circ\mathbf{x}_i) - (\mathbf{w}\circ\mathbf{x}_j)\|_2^2 p_{ij}^{\mathbf{w},\mathbf{y}} \\ =& \sum_{j=1}^n \|(\mathbf{w}\circ\mathbf{x}_i) - (\mathbf{w}\circ\mathbf{x}_j)\|_2^2 \left(q_{ij}^{\mathbf{w},\mathbf{y}} - p_{ij}^{\mathbf{w},\mathbf{y}}\right) . \end{aligned} \quad (3)$$

  Thus, the large margin regularization problem for feature selection can be formulated as the following minimization problem

$$\begin{aligned} \min_{\mathbf{w},\mathbf{y}} \; & -\sum_{i=1}^n \left(\sum_{j=1}^n \|(\mathbf{w}\circ\mathbf{x}_i) - (\mathbf{w}\circ\mathbf{x}_j)\|_2^2 \left(q_{ij}^{\mathbf{w},\mathbf{y}} - p_{ij}^{\mathbf{w},\mathbf{y}}\right) \right) \\ \text{s.t.} \; & \sum_{c=1}^d w_c = \alpha, w_c|_{c=1}^d \ge 0 . \end{aligned} \quad (4)$$

- **Manifold Regularization**: Besides the margin regularization, we also use the manifold structure to regularize the learning process by adopting the scheme of Local Linear Embedding (LLE) [17]. Given a data sample $\mathbf{x}_i$, we first find its nearest neighbors $\mathcal{N}_i^{\mathbf{w}}$ in the weighted feature space. We assume that $\mathbf{x}_i$ can be reconstructed by the linear combination of the nearest neighbors as

$$(\mathbf{w}\circ\mathbf{x}_i) \approx \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij}(\mathbf{w}\circ\mathbf{x}_j) \quad (5)$$

  where $A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}$ is the reconstruction coefficients, with constrains $\sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} = 1$ and $A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}} \ge 0$. Then the manifold regularization problem in the weighted feature space is formulated by minimizing the reconstruction error with regard to both $\mathbf{w}$ and $A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}$ for all the

training samples,

$$\begin{aligned}\min_{\mathbf{w},A}\ & \sum_{i=1}^{n}\left\|(\mathbf{w}\circ\mathbf{x}_i)-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}(\mathbf{w}\circ\mathbf{x}_j)\right\|_2^2\\ s.t.\ & \sum_{c=1}^{d}w_l=\alpha, w_c|_{c=1}^{d}\geq 0,\\ & \sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}=1, A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}\geq 0, A_{ij}|_{j\notin\mathcal{N}_i^{\mathbf{w}}}=0,\\ & \forall i=1,\cdots,n\end{aligned} \tag{6}$$

where $A=[A_{ij}]\in\mathbb{R}_+^{n\times n}$. Similar to LLE, we also regularize the learning of label learning using the same reconstruction formulation as,

$$\begin{aligned}\min_{\mathbf{y},A}\ & \sum_{i=1}^{n}\left\|y_i-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}y_j\right\|_2^2\\ s.t.\ & y_i=\widehat{y}_i,\forall i=1,\cdots,l\\ & \sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}=1, A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}\geq 0, A_{ij}|_{j\notin\mathcal{N}_i^{\mathbf{w}}}=0,\\ & \forall i=1,\cdots,n\end{aligned} \tag{7}$$

By considering the problems in (2), (6) and (7) simultaneously, we obtain the following formulation for the learning problem,

$$\begin{aligned}\min_{\mathbf{w},\mathbf{y},A}\ & \left\{-\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\|(\mathbf{w}\circ\mathbf{x}_i)-(\mathbf{w}\circ\mathbf{x}_j)\|_2^2\left(q_{ij}^{\mathbf{w},\mathbf{y}}-p_{ij}^{\mathbf{w},\mathbf{y}}\right)\right)\right.\\ & +\beta\sum_{i=1}^{n}\left\|(\mathbf{w}\circ\mathbf{x}_i)-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}(\mathbf{w}\circ\mathbf{x}_j)\right\|_2^2\\ & \left.+\gamma\sum_{i=1}^{n}\left\|y_i-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}y_j\right\|_2^2\right\}\\ s.t.\ & \sum_{c=1}^{d}w_l=\alpha, w_c|_{c=1}^{d}\geq 0,\\ & y_i=\widehat{y}_i,\forall i=1,\cdots,l,\\ & \sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}=1, A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}\geq 0, A_{ij}|_{j\notin\mathcal{N}_i^{\mathbf{w}}}=0,\\ & \forall i=1,\cdots,n\end{aligned} \tag{8}$$

where β and γ are the tradeoff parameters.

B. Optimization

We use an alternative optimization strategy to solve (8). The three variables are optimized in turn in an iterative algorithm.

1) Optimizing $\mathbf{w}$: When $\mathbf{w}$ is optimized, we fix both $\mathbf{y}$ and A, remove a term irrelevant to $\mathbf{w}$, and the following optimization problem is obtained,

$$\begin{aligned}\min_{\mathbf{w},\mathbf{y},A}\ & \left\{f(\mathbf{w})=-\sum_{i=1}^{n}\sum_{j=1}^{n}\|(\mathbf{w}\circ\mathbf{x}_i)-(\mathbf{w}\circ\mathbf{x}_j)\|_2^2\left(q_{ij}^{\mathbf{w},\mathbf{y}}-p_{ij}^{\mathbf{w},\mathbf{y}}\right)\right.\\ & +\beta\sum_{i=1}^{n}\left\|(\mathbf{w}\circ\mathbf{x}_i)-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}(\mathbf{w}\circ\mathbf{x}_j)\right\|_2^2\\ & \left.=\sum_{c=1}^{d}v_c\times w_c^2\right\}\\ s.t.\ & \sum_{c=1}^{d}w_l=\alpha, w_c|_{c=1}^{d}\geq 0.\end{aligned} \tag{9}$$

where

$$\begin{aligned}\mathbf{v}=\sum_{i=1}^{n}\Bigg(&-\sum_{j=1}^{n}|\mathbf{x}_i-\mathbf{x}_j|^2\left(q_{ij}^{\mathbf{w},\mathbf{y}}-p_{ij}^{\mathbf{w},\mathbf{y}}\right)\\ &+\beta\left|\mathbf{x}_i-\sum_{j\in\mathcal{N}_i^{\mathbf{w}}}A_{ij}\circ\mathbf{x}_j\right|^2\Bigg),\end{aligned} \tag{10}$$

$|\mathbf{x}|^2$ is the element-wise squared vector of vector $\mathbf{x}$, and v_c is the c-th element of $\mathbf{v}$. The Lagrange function of this problem is

$$\mathcal{L}=\sum_{c=1}^{d}v_cw_c^2-\sum_{c=1}^{d}\delta_cw_c+\epsilon\left(\sum_{c=1}^{d}w_c-\alpha\right) \tag{11}$$

where $\delta_c\geq 0$ is the Lagrange multiplier for the c-th constrain $w_c\geq 0$, while $\epsilon\geq 0$ is the Lagrange multiplier for the constrain $\sum_{c=1}^{d}w_l=\alpha$, By setting the derivative of $\mathcal{L}$ with regard to w_c to zero, we have

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial w_c}=0&\Rightarrow v_cw_c+\delta_c+\epsilon=0\\ &\Rightarrow w_c(v_cw_c+\delta_c+\epsilon)=0\end{aligned} \tag{12}$$

Using the (Karush-Kuhn-Tucker) KKT condition $w_c\delta_c=0$ [18], [19], [20], we have

$$w_c(v_cw_c+\epsilon)=0 \tag{13}$$

We can obtain the solution for w_c for the following two cases:

- Case I, If $v_c=0$: $w_c=0$;
- Case II, If $v_c\neq 0$: $w_c=-\frac{\epsilon}{v_c}$. By substituting it to the constrain $\sum_{c=1}^{d}w_c=\alpha$, we have

$$\sum_{c=1}^{d}w_c=\alpha\Rightarrow-\sum_{c:v_c\neq 0}\frac{\epsilon}{v_c}=\alpha\Rightarrow\epsilon=-\frac{\alpha}{\sum_{c:v_c\neq 0}\frac{1}{v_c}} \tag{14}$$

Thus the solution for w_c is

$$w_c = \alpha \frac{\frac{1}{v_c}}{\sum_{c':v_{c'}\neq 0} \frac{1}{v_{c'}}} \tag{15}$$

2) Optimizing $\mathbf{y}$*:* To optimize the class label vector $\mathbf{y}$, we fix $\mathbf{w}$ and A, remove irrelevant terms, and obtain the following optimization problem with regard to $\mathbf{y}$,

$$\begin{aligned} \min_{\mathbf{y}} \ & \gamma \sum_{i=1}^{n} \left\| y_i - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} y_j \right\|_2^2 \\ & = \gamma \left\| \mathbf{y}(I-A)^\top \right\|_2^2 \\ s.t. \ & y_i = \widehat{y}_i, \forall i = 1, \cdots, l \end{aligned} \tag{16}$$

where I is a $n \times n$ identity matrix. Since it is constrained that for any labeled samples $\mathbf{x}_i|_{i=1}^{l}$, its learned label y_i should be identify to its known label $\widehat{y}_i$, we only need to solve the class labels for the unlabeled samples $\mathbf{x}_i|_{i=l+1}^{n}$. To this end, we separate the class label vector to two sub-vectors $\mathbf{y}_l$ and $\mathbf{y}_u$, $\mathbf{y} = [\mathbf{y}_l \ \mathbf{y}_u]$, where $\mathbf{y}_l = [y_1, \cdots, y_l]$ contains the labels of the labeled samples, while $\mathbf{y}_u = [y_{l+1}, \cdots, y_n]$ contains the labels of the unlabeled samples. Moreover, we also denote matrix $Q = (I-A)^\top$ and separate it in a similar way, $Q = \begin{bmatrix} Q_l \\ Q_u \end{bmatrix}$ where Q_l contains the first l rows of Q and Q_u contains the remaining rows. By substituting $\mathbf{y}_l = \widehat{\mathbf{y}}_l$, we rewrite the problem in (16) as

$$\min_{\mathbf{y}_u} \left\{ g(\mathbf{y}_u) = \gamma \left\| [\widehat{\mathbf{y}}_l \ \mathbf{y}_u] \begin{bmatrix} Q_l \\ Q_u \end{bmatrix} \right\|_2^2 = \gamma \left\| \widehat{\mathbf{y}}_l Q_l + \mathbf{y}_u Q_u \right\|_2^2 \right\} \tag{17}$$

Its solution can be obtained by setting the derivative of g with regard to $\mathbf{y}_u$ to zero,

$$\begin{aligned} & \frac{\partial g(\mathbf{y}_u)}{\partial \mathbf{y}_u} = 2\gamma \left(\widehat{\mathbf{y}}_l Q_l + \mathbf{y}_u Q_u \right) Q_u^\top = 0 \\ & \Rightarrow \mathbf{y}_u = -\widehat{\mathbf{y}}_l Q_l Q_u^\top \left(Q_u Q_u^\top \right)^{-1} \end{aligned} \tag{18}$$

3) Optimizing A*:* To solve the reconstruction coefficients in A, we fix $\mathbf{w}$ and $\mathbf{y}$, remove the term irrelevant to A, and obtain the following optimization problem,

$$\begin{aligned} \min_{A} \ & \beta \sum_i \left\| (\mathbf{w} \circ \mathbf{x}_i) - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} (\mathbf{w} \circ \mathbf{x}_j) \right\|_2^2 \\ & + \gamma \sum_i \left\| y_i - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} y_j \right\|_2^2 \\ s.t. \ & \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} = 1, A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}} \geq 0, A_{ij}|_{j\notin\mathcal{N}_i^{\mathbf{w}}} = 0, \\ & \forall i = 1, \cdots, n \end{aligned} \tag{19}$$

It can be observed that the reconstruction coefficients for each data sample are independent, thus we can solve the coefficients for every data sample individually. For the i-th data sample, the problem is

$$\begin{aligned} \min_{A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}} \ & \left\{ g(A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}}) = \beta \left\| \mathbf{w} \circ \mathbf{x}_i - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} (\mathbf{w} \circ \mathbf{x}_j) \right\|_2^2 \right. \\ & + \gamma \left\| y_i - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} y_j \right\|_2^2 \\ & = \left\| \boldsymbol{\xi}_i - \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} \boldsymbol{\xi}_j \right\|_2^2 = \left\| \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_j) \right\|_2^2 \\ & \left. = \sum_{j,k\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} \left((\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)^\top (\boldsymbol{\xi}_i - \boldsymbol{\xi}_k) \right) A_{ik} \right\} \\ s.t. \ & \sum_{j\in\mathcal{N}_i^{\mathbf{w}}} A_{ij} = 1, A_{ij}|_{j\in\mathcal{N}_i^{\mathbf{w}}} \geq 0 \end{aligned} \tag{20}$$

where $\boldsymbol{\xi}_i = \begin{bmatrix} \sqrt{\beta}(\mathbf{w} \circ \mathbf{x}_i) \\ \sqrt{\gamma} y_i \end{bmatrix}$. Apparently, this problem can be solved as a Quadratic Programming (QP) problem [21], [22], [23].

The pseudo-code of the proposed algorithm is summarized in Algorithm 1.

Algorithm 1 Learning algorithm of the semi-supervised local-learning-based feature selection.

Input: Training data damples $\mathbf{x}_1, \cdots, \mathbf{x}_n$;
Input: Class labels for the first l samples in $\widehat{\mathbf{y}}_l$;
Input: Parameters α, β and γ.
Initialize the feature weight vector $\mathbf{w}^0$;
Initialize the neighborhood reconstruction coefficient matrix A^0;
Initialize the class label vector $\mathbf{y}_u^0$ for the unlabeled samples;
for $t = 1, \cdots, T$ **do**
  Update vector $\mathbf{v}^t$ as in (10) by fixing $\mathbf{w}^{t-1}$ and $\mathbf{y}^{t-1}$;
  Update feature weight vector $\mathbf{w}^t$ from vector $\mathbf{v}^t$;
  Update class label vector $\mathbf{y}_u^t$ of unlabeled samples as in (18) by fixing A^{t-1};
  Update neighborhood reconstruction coefficient matrix A^t as in (20) by fixing $\mathbf{w}^t$ and $\mathbf{y}^t$;
end for
Output: Feature weight vector $\mathbf{w}^T$.

IV. Experiments

We evaluate the performance of the proposed algorithm on two real world data sets.

A. Experiment I: Face Recognition

In the first experiment, we use the face image data set downloaded from the CMU Pose, Illumination, and Expression (PIE) database [24]. Since our algorithm currently can

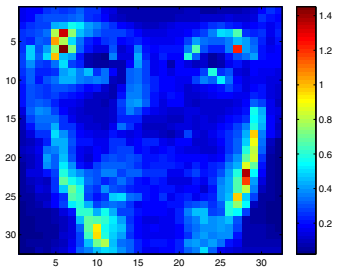


Fig. 1. Feature weights learned from the face image data set.

only deal with binary classification problems, we randomly select the images of two persons from the database. For each person, there are 170 images with various combinations of pose, illumination, and expression, and each images is of 32×32 pixels. We randomly select 50 samples for each person as labeled samples, while leaving the remaining 120 samples as unlabeled.

Fig. 1 presents the feature weights learned from the face image data set. It is interesting to see that the learned feature weights when mapped back to a two-dimensional image also has a face pattern. Some regions important for face recognizing, such as eyes and contour of a face, are highlighted, while the regions that may be irrelevant to face recognition, such as the lower left and right corners and cheeks, are assigned with small weights.

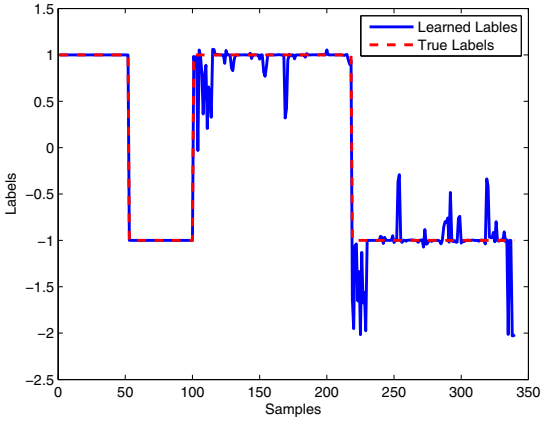


Fig. 2. Predicted class labels of the face image data set.

Our algorithm is able to learn feature weights and class labels simultaneously. Fig. 2 reports the predicted labels of unable samples as well as their true labels. We can see that the predicted labels are consistent with the true class labels. All positive samples obtain class label larger than zero, while all negative samples obtain class label smaller than zero, indicating that all of the unlabeled samples are predicted correctly.

We also show the reconstruction coefficient matrix in Fig. 3. It can be seen that by learning it in a weighted feature space, we can learn a discriminative reconstruction

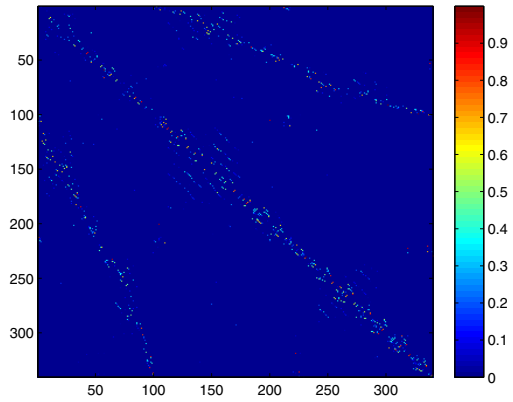


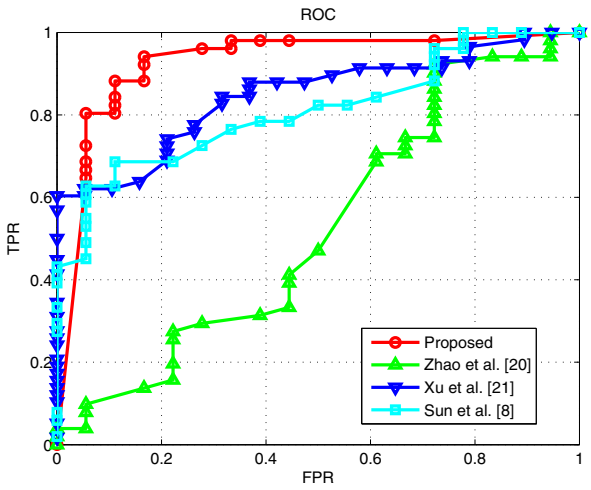
Fig. 3. Reconstruction coefficient matrix learned from the face image data set.

coefficient matrix to represent the manifold structure. It is not only smooth, but also discriminative. A sample is only constructed by its neighboring samples from the same class. This is an evidence that the local-learning-base method can sense the class conditional neighbors effectively.

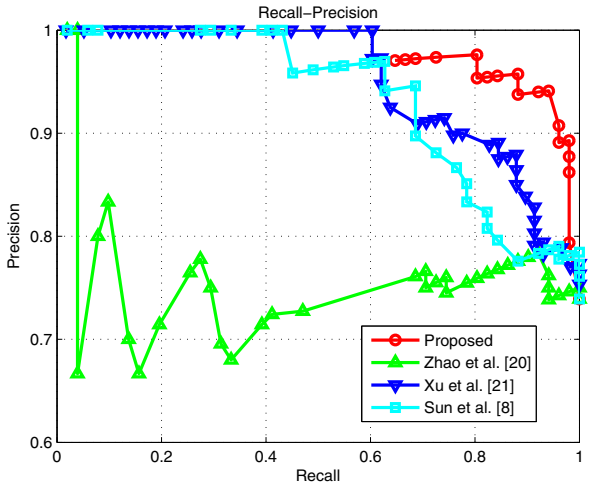
B. Experiment II: Diffuse large B-cell lymphoma outcome prediction

In the second experiment, we compare the proposed method against several semi-supervised feature selection method. The Diffuse Large B-Cell Lymphoma (DLBCL) gene expression data set [25] is used in this experiment. This data set contains the gene expression data of 69 samples, and for each sample, there are 5,469 features corresponding to 5,469 genes. In this data same, there are 48 positive samples and 21 negative samples. To conduct the experiment, we split the entire data in to ten folds randomly, and each fold is used as labeled samples in turn, while the remaining nine folds as unlabeled samples. The target is to predict the label of unlabeled samples. To measure the prediction performance of the prediction, we employ the Receiver Operating Characteristic (ROC) [26], [27] and recall-precision curves [28], [29] as performance metrics.

We compare our method to two state-of-the-art semi-supervised proposed by Zhao et al. [14] and Xu et al. [15]. Moreover, we also include the supervised local-learning-based feature selection method proposed by Sun et al. [5]. To adapt this algorithm to the partially labeled data set, we only use the labeled samples. The ROC and recall-precision curves are given in Fig. 4. From this figure, we can see that Zhao et al. [14]'s method is inferior to all other methods. This is not surprising because it relies on a precise local structure defined in the original feature space. However, in this data set, most of the original features are noisy, which leads to a unreliable estimate of local structure. Xu et al. [15]'s method is better than Zhao et al. [14]'s method. Although it also relies on the local structure defined in the original feature space, it learn the feature weights and the classifier jointly.



(a) ROC



(b) Recall-precision curve

Fig. 4. ROC and recall-precision curves on the DLBCL data set.

Benefiting from the discriminative ability of the classifier, it can also learn a discriminative feature weight vector. It is obvious that the proposed method outperform the other methods in most cases. Differently from these two methods which explore the local data structure via the original feature space, the proposed method learn the data structure and the class labels in the weighted feature space. This is the main reason that it beats the other methods. It is also interesting to note that Sun et al. [5] also obtain comparable performed as Xu et al. [15]'s method, although it only used the labeled samples. This is because that it uses the margin defined in the weighted feature space to refine the feature weights. However, it still cannot beat the proposed method, due to the fact that it cannot explore the data structure by only using the labeled samples.

V. Conclusion

In this paper, we extend the local-learning-based feature selection to semi-supervised learning problem. We proposed to learn the feature weight, the labels of unlabeled samples and the data structure jointly. By doing this, we can learn the margin for each data sample, and thus the local-learning can be performed in the partially labeled data set. The encouraging performances demonstrate the advantage of the proposed method over other semi-supervised feature selection method.

References

[1] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[2] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[3] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[4] Q. Sun, P. Wu, Y. Wu, M. Guo, and J. Lu, "Unsupervised multi-level non-negative matrix factorization model: Binary data case." *Journal of Information Security*, vol. 3, no. 4, 2012.

[5] Y. Sun, S. Todorovic, and S. Goodison, "Local-Learning-Based Feature Selection for High-Dimensional Data Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1610–1626, 2010.

[6] N. Koutsouleris, S. Borgwardt, E. M. Meisenzahl, R. Bottlender, H.-J. Möller, and A. Riecher-Rössler, "Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the fepsy study," *Schizophrenia bulletin*, vol. 38, no. 6, pp. 1234–1246, 2012.

[7] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[8] L. A. Cooper, J. Kong, F. Wang, T. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, "Morphological signatures and genomic correlates in glioblastoma," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1624–1627.

[9] I. Takeuchi and M. Sugiyama, "Target neighbor consistent feature weighting for nearest neighbor classification," in *Advances in Neural Information Processing Systems*, 2011, pp. 576–584.

[10] M. Bilenko, S. Basu, and R. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 2004, pp. 81–88.

[11] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine Learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[12] L. Käll, J. Canterbury, J. Weston, W. Noble, and M. MacCoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets," *Nature Methods*, vol. 4, no. 11, pp. 923–925, 2007.

[13] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings, Twentieth International Conference on Machine Learning*, vol. 2, 2003, pp. 912–919.

[14] J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, no. 10-12, pp. 1842–1849, 2008.

[15] Z. Xu, I. King, M.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[16] Z. Zhao and H. Lu, "Semi-supervised feature selection via spectral analysis," in *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007, pp. 641–646.

[17] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[18] L. Qi and H. Jiang, "Semismooth Karush-Kuhn-Tucker equations and convergence analysis of Newton and quasi-Newton methods for solving these equations," *Mathematics of Operations Research*, vol. 22, no. 2, pp. 301–325, 1997.
[19] H.-C. Wu, "The Karush-Kuhn-Tucker optimality conditions in an optimization problem with interval-valued objective function," *European Journal of Operational Research*, vol. 176, no. 1, pp. 46–59, 2007.
[20] ——, "The Karush-Kuhn-Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions," *European Journal of Operational Research*, vol. 196, no. 1, pp. 49–60, 2009.
[21] M. Forti and A. Tesi, "New conditions for global stability of neural networks with application to linear and quadratic programming problems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 42, no. 7, pp. 354–366, 1995.
[22] L. dos Santos Coelho and V. Mariani, "Combining of chaotic differential evolution and quadratic programming for economic dispatch optimization with valve-point effect," *IEEE Transactions on Power Systems*, vol. 21, no. 2, pp. 989–996, 2006.
[23] C.-H. Guo, Y.-Q. Bai, and J.-B. Jian, "An improved sequential quadratic programming algorithm for solving general nonlinear programming problems," *Journal of Mathematical Analysis and Applications*, vol. 409, no. 2, pp. 777–789, 2014.
[24] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) database," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 53 – 8.
[25] M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, and T. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.
[26] E. DeLong, D. DeLong, and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.
[27] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[28] V. V. Raghavan, G. S. Jung, and P. Bollmann, "Critical investigation of recall and precision as measures of retrieval system performance," *ACM transactions on office information systems*, vol. 7, no. 3, pp. 205–229, 1989.
[29] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," vol. 2006, 2006, pp. 233–240.