Semi-supervised Learning Using Local Regularizer and Unit Circle Class Label Representation

Jia Lv[1,2]
1. School of Mathematical Sciences, Inner Mongolia University, Hohhot 010021 China;
2. College of Computer and Information Sciences, Chongqing Normal University, Chongqing 400047 China
Email: lvjia@cqnu.edu.cn

Abstract—**Semi-supervised learning, which aims to learn from partially labeled data and mostly unlabeled data, has been attracted more and more attention in machine learning and pattern recognition. A novel semi-supervised classification approach is proposed, which can not only handle semi-supervised binary classification problem but also deal with semi-supervised multi-class classification problem. The approach is based on local regularizer and unit circle class label representation. The former is minimized so as to cause the class labels to have the desired properties. The latter utilizes two-dimensional vector evenly distributed in circumference of unit circle to represent class label, so multi-class classification can be performed only once. Comparative classification experiments on some benchmark datasets validate the effectiveness of the presented approach.**

Index Terms—**Semi-supervised learning, classification, local regularizer, unit circle, class label representation**

I. INTRODUCTION

Most machine learning problems can be regarded as function estimation from the instances or the past experience by optimizing some objective criteria [1-4]. Labeling often requires expensive human labor and much time, while semi-supervised learning need not to label a lot of data, which is based upon both a small amount of labeled data and a large number of unlabeled data. Semi-supervised learning has become an increasingly attractive methodology, and it has been proven to be effective in the fields of pattern recognition and machine learning [5-9].

In general, the task of semi-supervised learning can be categorized into two classes: one is transductive learning and the other is inductive learning. The former takes advantage of a set of data points that contain both labeled and unlabeled data we want to classify to help train the classifier, then the unlabeled data are classified. In contrast, the latter exploits both labelled and unlabeled data to train the classifier which will be used to classify future unlabeled data, thus an independent unlabeled test dataset is used to evaluate the performance of the trained classifier. Now transductive learning algorithms are abundant, including Transductive Support Vector Machine (*TSVM*) [10], the Laplacian Regularization method (*Lap_Reg*) [11], the Normalized Laplacian Regularization method (*NLap_Reg*) [12]and Local Learning Regularization method (*LL_Reg*) [13], etc., which are applied to binary classification. Multi-class classification problem tends to be transformed into multiple binary classification problems and the algorithms mentioned above will be employed multiple times [13, 14]. A novel label representation method is presented in this paper, and the objective of multi-class classification problem is fulfilled directly via one multi-class classification model with local regularizer. Comparative classification experiments on six benchmark datasets illustrate the feasibility and effectiveness of this approach.

This paper is organized as follows: Section II gives semi-supervised binary classification problem, Section III describes the basic idea of local learning and presents local learning regularizer, Section IV presents unit circle class label presentation and gives semi-supervised multi-class classification approach on the basis of local learning and unit circle class label representation, Section V describes and analyzes the experimental results, then Section VI concludes the results.

II. SEMI-SUPERVISED BINARY CLASSIFICATION

Given a set of n data points

$$T=\{(x_1,y_1),\ldots,(x_l,y_l)\}\cup\{x_{l+1},\cdots,x_n\}, \qquad (1)$$

where $x_i \in R^d$ is a *d*-dimensional input vector for the *i*-th data points, the first l data points are labeled and $y_i \in \{-1,+1\}$ denotes the class label for $x_i, i=1,\ldots,l$, the remaining $n-l$ data points are unlabeled. The goal is to label the given unlabeled instances $x_{l+1},\ldots,x_n$ on all data instances.

Many semi-supervised binary classification can be roughly addressed as the following optimization problem [8, 9, 11-14].

Manuscript received March 1, 2011; revised June 1, 2011; accepted July 1, 2011.

doi:10.4304/jsw.7.5.951-958

$$\min_F (F^T RF + (F-Y)^T D(F-Y)),$$
(2)

where $F = (f_1, f_2, \ldots, f_n)^T \in R^n$ is the vector of real valued solution, R is the regularization matrix, $Y = (y_1, y_2, \ldots, y_n)^T$ is the class label vector of data points, $y_i \in \{-1,+1\}$, $i = 1,\ldots,l$, $y_i = 0$ is initial class label corresponding to $x_i, i = l+1,\ldots,n$, $D \in R^{n\times n}$ is a diagonal matrix and its *i*-th diagonal element D_{ii} is as follows: $D_{ii} = D_l > 0, i = 1,\ldots,l,$ $D_{ii} = D_u \geq 0, i = l+1,\ldots,n.$The first term $F^T RF$ in the objective function is called regularizer that specifies the desirable properties of f_i , and the second term $(F-Y)^T D(F-Y)$ is just similar to the square loss function that restricts that f_i should be close to $y_i, i = 1,\ldots,n$.

The problem (2) is an unconstrained optimization problem, so its solution can be obtained by

$$F = (R+D)^{-1} DY. \tag{3}$$

Class labels of the unlabeled data $x_{l+1},\ldots,x_n$ are computed as follows:

$$y_i = sign(f_i), i = l+1,...,n, \tag{4}$$

where sign(.) is the sign function and its definition is as follows:

$$\mathrm{si}\,gn(s) = \begin{cases} 1, & s \geq 0 \\ -1, & s < 0 \end{cases}. \tag{5}$$

III. Semi-supervised Binary Classification

A. Local Learning

It is difficult to handle the situation shown in Fig. 1 for global learning strategy [15] to the entire data points. Fig. 1 is a piece of an imaginary feature space. Gray and black circles are instances of two classes. Thin lines are the actual decision boundary between these two classes. Both crosses 1 and 2 represent rejected patterns. Cross 1 cannot be determined with enough confidence, while Cross 2 is actually an outlier which cannot be determined lower confidence.

Instead of a global learning model, local learning strategy is presented to construct models based on the local neighborhood information of each instances. Recent research has demonstrated that such a local learning strategy is superior to global learning, especially on datasets that are not evenly distributed [16, 17]. However, there are a small amount of labeled data and lots of unlabeled data in semi-supervised classification [18-25]. If local learning method is used directly, it is possible that the method cannot be satisfied because of a few or even no labeled data in the local region. This case has been reasonablely resolved by Local Learning Regularization method (*LL_Reg*) [13].

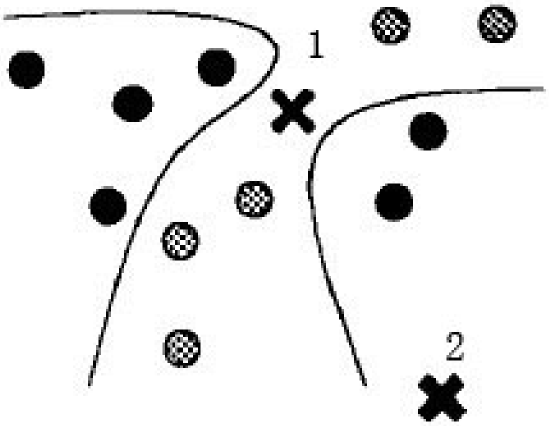

Figure 1. Two special instances in an imaginary pattern space

B. Local Regularizer

It is obvious that the nature of local learning is that a given test instance is well estimated only by its adjacent data points. Using the notation in the problem (2), for any $x_i, i = 1,\ldots,n$, the value of f_i can be well evaluated based on the data points located in the vicinity of x_i . The value of f_i should be similar to the output of the model that is trained locally with the data points $\{(x_j, f_i)\}(x_j \in N_i)$, where N_i denotes the set of neighboring data points of x_i . Linear model is employed in local learning model as follows:

$$g_i(x) = w_i^T (x - x_i) + b_i, \forall x \in N_i \subset R^d, \tag{6}$$

where $w_i \in R^d, b_i \in R, i = 1,\ldots,n$. To find the linear model, the optimization problem based on this model is required to solve as follows:

$$\min_{w_i\in R^d, b_i \in R} \lambda \|w_i\|^2 + \sum_{x_j \in N_i} (w_i^T (x_j - x_i) + b_i - f_j)^2, \tag{7}$$

where $\lambda > 0$ is a parameter that controls the trade-off between the first term and the second term in (7).

So the first term in the problem (2) is naturally replaced with $\sum_{i=1}^n (f_i - g_i(x_i))^2$ and it can be written in a compact form as:

$$\|F - G\|^2, G = (g_1(x_1),\ldots,g_n(x_n))^T. \tag{8}$$

Theorem 1. For each $x_i, i = 1,\ldots,n$, n_i is the number of instances in N_i , $X_i = (x_j - x_i)_{d\times n_i} \in R^{d\times n_i}$,

$x_j \in N_i$, $w_i \in R^d$, $b_i \in R$, $e = (1,\ldots,1)^T \in R^{n_i}$, I is the unit matrix of order n_i, $F_i = (f_j)^T_{x_j \in N_i} \in R^{n_i}$, then the linear model

$$g_i(x_i) = b_i = \alpha_i^T F_i, \qquad (9)$$

where

$$\alpha_i^T = \frac{e^T - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1}}{n_i - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}. \qquad (10)$$

and α_i can further be simplified [14].

Theorem 2. For each $x_i, i = 1,\ldots,n$, n_i is the number of instances in N_i, $X_i = (x_j - x_i)_{d\times n_i} \in R^{d\times n_i}$, $x_j \in N_i$, $w_i \in R^d$, $b_i \in R$, $e = (1,\ldots,1)^T \in R^{n_i}$, I is the unit matrix of order n_i, $F_i = (f_j)^T_{x_j \in N_i} \in R^{n_i}$, then the linear model

$$g_i(x_i) = \alpha_i^T F_i, \qquad (11)$$

where

$$\alpha_i^T = \frac{e^T (\lambda I + X_i^T X_i)^{-1}}{e^T (\lambda I + X_i^T X_i)^{-1} e}.$$

(12)

In the above equation, α_i is only related to X_i and is irrelevant to f_j. Thus α_i can be extended to the following matrix

$$A = (a_{ij}) \in R^{n\times n}, \qquad (13)$$

where if $x_j \in N_i$, then $a_{ij} = \alpha_{ij}$, otherwise $a_{ij} = 0$. Accordingly, equation (11) can be written in a compact form as:

$$G = AF. \qquad (14)$$

and $\|F - G\|^2$ in (8) is equal to

$$\|F - G\|^2 = F^T (I - A)^T (I - A) F, \qquad (15)$$

where I is the unit matrix of order n, therefore R of the first term in the problem (2) is just the following local learning regularizer:

$$R = (I - A)^T (I - A). \qquad (16)$$

IV. Semi-supervised Multi-class Classification

A. Semi-supervised Multi-class Classification Problem

Given a set of *n* data points,

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \cup \{x_{l+1}, \cdots, x_n\}, \qquad (17)$$

where $x_i \in R^d$ is a *d*-dimensional input vector for the data points, the first l data points are labeled and $y_i \in \{1,\ldots,c\}$ denotes the class label for $x_i, i = 1,\ldots,l$, the remaining $n - l$ data points are unlabeled. The goal is to predict the class labels $y_{l+1},\ldots,y_n$ of the given data points $x_{l+1},\ldots,x_n$.

B. Unit Circle Class Label Presentation Method

In semi-supervised binary classification, the initial class label y_i of the unlabeled data $x_i, i = l+1,\ldots,n$ is set to 0. The aim is that y_i is unbiased for either of the classes so to keep a balance between -1 and +1 of the class label. In essence, the corresponding class label will be assigned to x_i if f_i is nearly close to -1 or +1 by sign(.) in (5), so equation (5) can be changed into:

$$y_i = \arg\min_j |f_i - j|_{j=\{-1,+1\}} \quad i = l+1,\ldots,n. \qquad (18)$$

Similarly, for semi-supervised multi-class classification problem with c class labels, there is no need to solve multiple binary classification problems if the initial class labels of the unlabeled data can be biased for neither of c labels. Therefore unit circle in a two-dimensional plane is introduced in class label presentation, where the initial class labels of the unlabeled data are all fixed in circle center and c class labels are evenly distributed in circumference, and two-dimensional vector $C_j \in R^2$, $j = 1,\ldots,c$ is expressed as:

$$C_j = (\cos(\frac{2\pi}{c} \times j), \sin(\frac{2\pi}{c} \times j))^T, j = 1,\ldots,c. \qquad (19)$$

Actually, if $c = 2, C_1 = (-1,0)^T, C_1 = (1,0)^T$, semi-supervised multi-class classification is degenerated to binary classification. The optimal solution of multi-class classification is $F = (f_1, f_2, \ldots, f_n)^T \in R^{n\times 2}$, the class labels of the unlabeled data are determined by

$$y_i = \arg\min_j \|f_i - C_j\|_{j=1,\ldots,c} \quad i = l+1,\ldots,n. \qquad (20)$$

C. Local Regularizer-based Semi-supervised Multi-class Classification

In correspondence with the optimization problem (2), the class label vector in multi-class classification problem becomes the $n\times 2$ matrix M, $M = (m_1,\ldots,m_n)^T$, $m_i = C_{y_i}$, $y_i \in \{1,\ldots,c\}$, $i = 1,\ldots,l$, $m_i = (0,0)^T$, $i = l+1,\ldots,n$, so the problem (2) becomes

$$\min_F (F^T R F + (F - M)^T D (F - M)). \qquad (21)$$

Its optimal solution is analytically expressed as

$$F = (R + D)^{-1} DM, \qquad (22)$$

where $D \in R^{n\times n}$ is a diagonal matrix, its *i*-th diagonal element is $D_{ii} = D_l > 0,\ i = 1,\ldots,l,\ D_{ii} = D_u \ge 0,$ $i = l+1,\ldots,n.$

Similar to local learning regularization method, the corresponding results can be obtained, i.e.,

$$R = (I - A)^T (I - A), \tag{23}$$

where I is the unit matrix of $n\times n$. To summarize, the entire algorithm framework can be described as follows.

Step 1. Given the training dataset

$$T = \{(x_1, y_1),\ldots,(x_l, y_l)\} \cup \{x_{l+1},\cdots,x_n\}, \tag{24}$$

where $x_i \in R^d, i = 1,\ldots,n, y_i \in \{1,\ldots,c\}, i = 1,\ldots,l.$ Choose proper parameters $D_l > 0, D_u \ge 0, \lambda > 0,$ and the number of the nearest neighboring data points $K > 0$.

Step 2. Compute C_j

$$C_j = (\cos(\frac{2\pi}{c}\times j), \sin(\frac{2\pi}{c}\times j))^T, j = 1,\ldots,c. \tag{25}$$

Construct the $n\times 2$ matrix M , $M = (m_1,\ldots,m_n)^T$, where $m_i = C_{y_i}, y_i \in \{1,\ldots,c\}$, $y_i \in \{1,\ldots,c\}$, $i = 1,\ldots,l$, $m_i = (0,0)^T$, $i = l+1,\ldots,n$.

Step 3. Compute the K nearest neighboring data p $x_{j_1}, x_{j_2} \ldots, x_{j_K}, x_i \in R^d,\ i = 1,\ldots,n,\ j_k \in \{1,\ldots,n\},$ $k \in \{1,\ldots,K\}, X_i = (x_{j_1} - x_i, x_{j_2} - x_i,\ldots,x_{j_K} - x_i) \in R^{d\times K}$.

Step 4. Compute $\alpha_i = (\alpha_{ij_1},\ldots,\alpha_{ij_K})^T$

$$\alpha_i^T = \frac{e^T(\lambda I + X_i^T X_i)^{-1}}{e^T(\lambda I + X_i^T X_i)^{-1} e}, \tag{26}$$

where *e* is a vector of ones, that is, $e = (1,\ldots,1)^T \in R^K$, I is the $K\times K$ unit matrix.

Step 5. Extend α_i to the matrix

$$A = (a_{ij}) \in R^{n\times n}, \tag{27}$$

where if x_j belongs to the *K* nearest data points of x_i, then a_{ij} is equal to α_{ij}, otherwise a_{ij} is equal to 0.

Step 6. Calculate $F = (f_1, f_2,\ldots, f_n)^T \in R^{n\times 2}$

$$F = (R + D)^{-1} DM, \tag{28}$$

where

$$R = (I - A)^T (I - A), \tag{29}$$

I is the $n\times n$ unit matrix, $D \in R^{n\times n}$ is a diagonal matrix, its *i*-th diagonal element $D_{ii} = D_l > 0, i = 1,\ldots,l,$ $D_{ii} = D_u \ge 0, i = l+1,\ldots,n.$

Step 7. Determine the class labels $y_{l+1},\ldots,y_n$ of the unlabeled instances $x_{l+1},\ldots,x_n$ by

$$y_i = \arg\min_j \left\| f_i - C_j \right\|_{j=1,\ldots,c} \quad i = l+1,\ldots,n. \tag{30}$$

V. Experiments

A. The Datasets

Six benchmark datasets from UCI Machine Learning Repository are used in our experiments, and descriptions of the datasets are given in Table I. The original data sizes of shuttle are 43500 and the number of its classes is 7, and 1500 data points are randomly sampled from its first 5 classes in order to decrease computational complexity.

B. Experimental Methods

Leave-One-Out (LOO) classification error is computed for the sake of parameter selection. The LOO procedure of semi-supervised classification problem has *l* iterations. In the *i*-th iteration, the class label y_i of x_i, *i*=1,…,*l* is set to 0, and the algorithm is performed on the new dataset, then it is inferred whether x_i can correctly be classified in the *i*-th iteration.

Reference [13] presents a new LOO computation method for semi-supervised binary classification to avoid *l* iterations, seen as Theorem 3.

Theorem 3. Suppose $F = (f_1, f_2,\ldots, f_n)^T \in R^n$ is the solution of the quadratic programming problem (2), then $f_i^{(i)}$ can be computed as

$$f_i^{(i)} = \frac{f_i - D_l y_i (R + D)_{ii}^{-1}}{1 - (D_l - D_u)(R + D)_{ii}^{-1}} \quad 1 \le i \le l. \tag{31}$$

In addition to this algorithm, *Lap_Reg* and *NLap_Reg* are also implemented in our algorithm in order to compare the experimental results between global learning regularizer and local learning regularizer. *R* is replaced in

TABLE I.
DESCRIPTION OF THE DATASETS

No.	Datasets	Classes	Sizes	Features
1	DIGIT1	2	1500	241
2	IRIS	3	150	4
3	WINE	3	178	13
4	SHUTTLE	5	1500	9
5	OPTIGITS	10	1797	64
6	PENDIGITS	10	3498	16

the optimization solution (28) with $D-W$ and $I-D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, termed as *Lap_Reg*-based semi-supervised multi-class classification algorithm (*MCLap_Reg*) and *NLap_Reg*-based semi-supervised multi-class classification algorithm (*MCNLap_Reg*), respectively. $W=[w_{ij}]\in R^{n\times n}$ is a symmetric matrix and its elements w_{ij} representing the similarity between the data instances x_i and x_j can be computed as:

$$w_{ij}=\exp(-\frac{1}{\delta}\left\|x_i-x_j\right\|^2), \tag{32}$$

where if x_i is one of the K nearest neighboring points of x_j, otherwise w_{ij} is equal to 0, where δ is a positive hyper-parameter. D is a diagonal matrix whose diagonal element D_{ii} is the sum of the *i*-th row of W, i.e.,

$$D_{ii}=\sum_{j=1}^{n}w_{ij}, i=1,\ldots,n. \tag{33}$$

For three algorithms, their optimal parameters are set by LOO error from some grids introduced in the following:

10% of each dataset is randomly chosen as labeled data points for six datasets;

$K\in\{5,10,20,50,100\}$;

$D_l\in\{0.1,1,10,100\}$;

$D_u=10^{-6}$;

$\lambda\in\{0.1,1,10,100\}$;

$\delta\in\{\frac{1}{2^6},\frac{1}{2^5},\frac{1}{2^4},\frac{1}{2^3},\frac{1}{2^2},\frac{1}{2^1},1,2^1,2^2,2^3,2^4,2^5,2^6\}$.

C. Experimental Results

The mean classification error rates and standard deviations and of 20 independent runs are presented in Table II, and the best results are highlighted in bold face. Moreover, the computational complexities of three algorithms are listed in Table III. In addition to that, to illustrate how the average test error rates change with the number of randomly labeled data points, we also provide the results of three algorithms on two small training datasets IRIS and WINE with the labeled data sizes $l\in\{3,n\times10\%,n\times20\%,n\times30\%,n\times40\%,n\times50\%, n\times60\%,n\times70\%,n\times80\%,n\times90\%\}$, n is the number of dataset, seen as Fig. 2 and Fig. 3.

TABLE III.
THE AVERAGE TEST ERROR RATES (%) AND STANDARD DEVIATIONS (%) COMPARISON

No.	*MCLap_Reg*	*MCNLap_Reg*	*MCLL_Reg*
1	3.2444±1.1647	3.0926±0.9625	**2.3889±0.7261**
2	7.7037±5.6434	6.5185±4.0401	**4.0333±2.3243**
3	27.2981±3.4522	29.0683±5.0080	**11.7391±4.2314**
4	**3.0148±1.5529**	17.6222±5.4867	**3.3741±0.8945**
5	9.9011±2.9526	8.4425±1.9571	**7.1910±1.9195**
6	10.9130±0.9368	9.6046±1.8640	**4.7682±1.0608**

TABLE II.
AVERAGE COMPUTATIONAL COMPLEXITY (*S*) COMPARISON

No.	*MCLap_Reg*	*MCNLap_Reg*	*MCLL_Reg*
1	75.4007	84.0290	99.2299
2	0.1667	0.1846	0.3346
3	0.2482	0.2695	0.6170
4	67.3312	78.3869	71.1965
5	110.9601	127.4904	118.2410
6	797.1355	935.7579	839.2214

As can be seen from Table II, we can observe that *MCLL_Reg* is remarkably superior to *MCLap_Reg* and *MCNLap_Reg*, while *MCLap_Reg* and *MCNLap_Reg* are equivalent to each other except for the fourth dataset. Table III indicates that *MCLL_Reg* is a little more time consuming when the number of dataset as well as its dimensions is not a lot larger, in contrast with other two algorithms, while the computational time of *MCLL_Reg* is between that of *MCLap_Reg* and that of *MCNLap_Reg* when dataset sizes and dimensions are larger than them.

Fig. 2 and Fig. 3 illustrate that the number of randomly

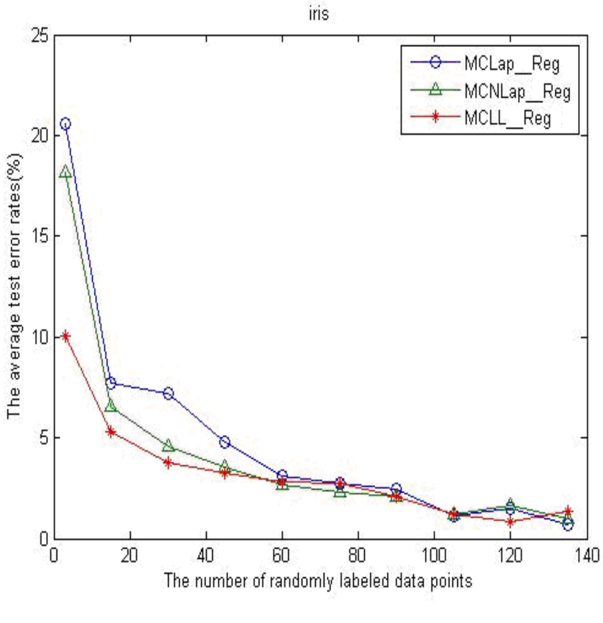

Figure 2. The average test error rates with respect to the number of randomly labeled points on IRIS dataset

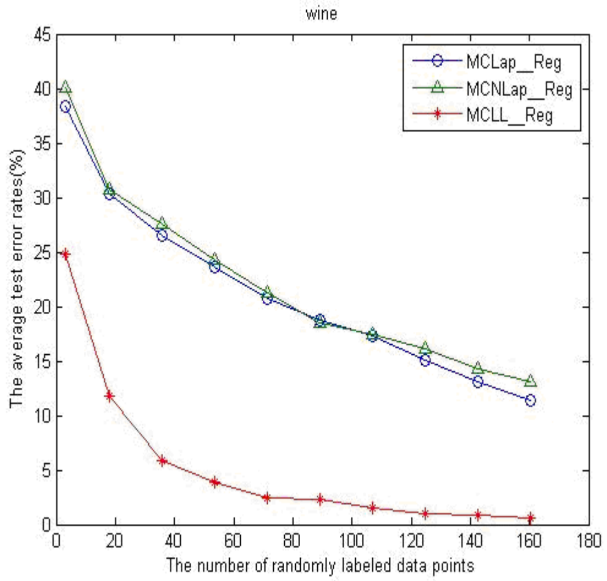

Figure 3. The average test error rates with respect to the number of randomly labeled points on WINE dataset

labeled points is larger, the average test error rate is lower on all algorithms, and furthermore, *MCLL_Reg* achieves a lower error rate.

VI. CONCLUSIONS

In this paper, a novel class label presentation method for semi-supervised classification problem is proposed on the basis of unit circle, which can impartially deal with each unlabeled data for fear of being misguiding by the labeled data during learning. Additionally, the class label of each data can be well estimated by its neighbors by means of local learning. So semi-supervised multi-class classification algorithm based on local regularizer and unit circle class label representation is developed. The experimental results show the superiority of our new algorithm.

In the further work, binary series class label representation method [11-14] will be introduced to our algorithm. Then we will compare the classification performance of binary series class label representation method with that of unit circle class label representation method.

APPENDIX A PROOF OF THEOREM 1

Let U be a d-dimensional unit matrix and

$$L=\lambda\|w_i\|^2+\sum_{x_j\in N_i}(w_i^T(x_j-x_i)+b_i-f_j)^2,$$

$$\begin{aligned}\nabla_{w_i}L=&2\lambda w_i+\\&\sum_{x_j\in N_i}(2w_i^T(x_j-x_i)(x_j-x_i)+2(x_j-x_i)b_i-2(x_j-x_i)f_j)\\&=2\lambda w_i+2X_iX_i^Tw_i+2b_iX_ie-2X_iF_i\\&=0,\end{aligned}$$

$$w_i=(\lambda U+X_iX_i^T)^{-1}X_iF_i-(\lambda U+X_iX_i^T)^{-1}b_iX_ie,$$

$$\begin{aligned}\nabla_{b_i}L&=\sum_{x_j\in N_i}(2b_i+2w_i^T(x_j-x_i)-2f_j)\\&=2n_ib_i+2w_i^TX_ie-2e^TF_i\\&=0,\end{aligned}$$

then

$$\begin{aligned}n_ib_i&=e^TF_i-w_i^TX_ie\\&=e^TF_i-((\lambda U+X_iX_i^T)^{-1}X_iF_i\\&\quad-(\lambda U+X_iX_i^T)^{-1}b_iX_ie)^TX_ie\\&=e^TF_i-F_i^TX_i^T((\lambda U+X_iX_i^T)^{-1})^TX_ie\\&\quad+e^TX_i^Tb_i((\lambda U+X_iX_i^T)^{-1})^TX_ie\\&=e^TF_i-F_i^TX_i^T((\lambda U+X_iX_i^T)^T)^{-1}X_ie\\&\quad+e^TX_i^Tb_i((\lambda U+X_iX_i^T)^T)^{-1}X_ie\\&=e^TF_i-F_i^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_ie\\&\quad+e^TX_i^Tb_i(\lambda U+X_iX_i^T)^{-1}X_ie\\&=e^TF_i-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_iF_i\\&\quad+e^TX_i^Tb_i(\lambda U+X_iX_i^T)^{-1}X_ie\\&=(e^T-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_i)F_i\\&\quad+e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_ieb_i,\end{aligned}$$

then

$$\begin{aligned}&n_ib_i-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_ieb_i\\&\quad=(e^T-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_i)F_i,\end{aligned}$$

and we can derive that

$$b_i=\frac{(e^T-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_i)F_i}{n_i-e^TX_i^T(\lambda U+X_iX_i^T)^{-1}X_ie}.$$

and also

$$\begin{aligned}&X_i^T(\lambda U+X_iX_i^T)^{-1}X_i=X_i^T(\lambda U+X_iX_i^T)^{-1}X_iI\\&=X_i^T(\lambda U+X_iX_i^T)^{-1}X_i(\lambda I+X_i^TX_i)(\lambda I+X_i^TX_i)^{-1}\\&=X_i^T(\lambda U+X_iX_i^T)^{-1}(X_i\lambda I+X_iX_i^TX_i)(\lambda I+X_i^TX_i)^{-1}\\&=X_i^T(\lambda U+X_iX_i^T)^{-1}(\lambda U+X_iX_i^T)X_i(\lambda I+X_i^TX_i)^{-1}\\&=X_i^TX_i(\lambda I+X_i^TX_i)^{-1},\end{aligned}$$

then

$$b_i = \frac{(e^T - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} F_i}{n_i - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}.$$

therefore

$$g_i(x_i) = b_i = \alpha_i^T F_i.$$

and

$$\alpha_i^T = \frac{e^T - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1}}{n_i - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e^T}. \quad \square$$

Appendix B Proof of Theorem 2

Noting that $n_i = e^T e$, where $e = (1,\ldots,1)^T \in R^{n_i}$, then it follows that

$$\alpha_i = \frac{e - X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}{n_i - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{(\lambda I + X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e - X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}{n_i - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{(\lambda I + X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e - X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}{e^T e - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{(\lambda I + X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e - X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}{e^T (\lambda I + X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e - e^T X_i^T X_i (\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{(\lambda I + X_i^T X_i - X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e}{e^T (\lambda I + X_i^T X_i - X_i^T X_i)(\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{\lambda I (\lambda I + X_i^T X_i)^{-1} e}{e^T \lambda I (\lambda I + X_i^T X_i)^{-1} e}$$

$$= \frac{(\lambda I + X_i^T X_i)^{-1} e}{e^T (\lambda I + X_i^T X_i)^{-1} e}. \quad \square$$

Acknowledgment

The author wishes especially to thank Naiyang Deng, Xinmin Yang, Yingjie Tian. This work was supported by the State Key Program of National Natural Science foundation of China (Grant No. 10831009) and National Natural Science foundation of China (Grant No.10971223, No.11071252, No. 70921061).

References

[1] N. Y. Deng, Y. J. Tian, C. H. Zhang, *Support Vector Machines: Theory, Algorithms, and Extensions.* CRC Press, 2012, in press.

[2] F. Wang, "A general learning framework using local and global regularization," *Pattern Recognition*, vol. 43, pp. 3120-3129, 2010.

[3] J. Lv. "Research on K-means clustering algorithm based on dynamic tunneling system," *Journal of Chongqing Normal University (Natural Science Edition)*, vol. 26, no. 1, pp. 73-77, 2009.

[4] J. Lv. "Realization of text classification system based on improved classification model," *Journal of Chongqing Normal University (Natural Science Edition)*, vol. 26, no. 2, pp.68-73, 2009.

[5] O. Chapelle, B. Scholkopf, A. Zien. *Semi-supervised Learning*, MIT Press, 2006.

[6] X. J. Zhu, *Semi-supervised Learning Literature Survey*, Computer Sciences TR 1530, Univ. of Wisconsin, 2008.

[7] G. Blanchard, G. Lee, C. Scott, "Semi-Supervised Novelty Detection," *Journal of Machine Learning Research*, vol. 11, pp. 2973-3009, 2010.

[8] H. Chen, L. Q. Li, J.T. Peng, "Semi-supervised learning based on high density region estimation," *Neural Networks*, vol. 23, pp. 812-818, 2010.

[9] M. Belkin, P. Niyogi, V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399-2434, 2006.

[10] T. Joachims, "Transductive learning via spectral graph partitioning," *Proc. 20th International Conference on Machine Learning*, pp. 290-297, 2003.

[11] X. Zhu, Z. Ghaharmani, J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," *Proc. 20th International Conference on Machine Learning*, pp. 912-919, 2003.

[12] D. Y. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Scholkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems* 16, MIT press, PP. 321-328, 2003.

[13] M. R. Wu, B. Scholkopf, "Transductive classification via local learning regularization," *Proc. of the 11th Int'1 Conf. on Artificial Intelligence and Statistics*, pp. 624-631, 2007.

[14] S.M. Xiang, F. P. Nie, C. S. Zhang, "Semi-supervised classification via local spline regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 11, pp. 2039-2053, 2011.

[15] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, Berlin, 1995.

[16] L. Bottou, V. Vapnik, "Local learning algorithms," *Neural Computation*, vol. 4, pp. 888-900, 1992.

[17] V. Vapnik, L. Bottou, "Local algorithms for pattern recognition and dependencies estimation," *Neural Computation*, vol. 5, pp. 893-909, 1993.

[18] A. Torralba, K. P. Murphy, W. T. Freeman, "Sharing visual features for multi-class and multi-view object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 854-869, 2007.

[19] M. Belkin, I. Matveeva, P. Niyogi, "Tikhonov regularization and semi-supervised learning on large graphs," IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1, pp. 1000-1003, 2004.

[20] H. B. Cheng, P. N, Tan, R. Jin, "Efficient algorithm for localized support vector machine," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 4, pp. 537-549, 2010.

[21] R. T. Peres, C.E. Pedreira, "A new local-global approach for classification," *Neural Networks*, vol. 23, pp. 887-891, 2010.

[22] F. Wang, C, S. Zhang, "Label propagation through linear neighborhoods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 1, pp. 55-67, 2008.

[23] F. Wang, T. Li, G. Wang, C. S. Zhang, "Semi-supervised classification using local and global regularization," *Proc. 23rd AAAI Conf. Artificial Intelligence*, PPP. 726-731, 2008.

[24] Q. X. Gao, H. Xu, Y. Y. Li, D, Y. Xie, "Two-dimensional supervised local similarity and diversity projection," *Pattern Recognition*, vol. 43, pp. 3359-3363, 2010.

[25] S. L. Sun, "Local within-class accuracies for weighting individual outputs in multiple classifier systems," *Pattern Recognition Letters*, vol. 31, pp. 119-124, 2010.

Jia Lv born in Dali county of Yunnan province in P. R. China, on May 30th, 1978. In 2000 she received the B.S. degree in computer science from Chongqing Normal University, Chongqing, China. In 2006 she received the M.S. degree in computer science and technology from Chongqing University, Chongqing, China. Now, she is pursuing for Ph.D. degree in applied mathematics in Inner Mongolia University, Hohhot, Inner Mongolia, China.

In 2000 she joined College of Computer and Information Sciences, Chongqing Normal University, Chongqing, where she is currently an ASSOCIATE PROFESSOR. From 2010 to 2011, she was a domestic visiting scholar with College of Sciences, Chinese Agricultural University, supported by a project of Chinese Ministry of Education for young key teachers of institutions of higher education. She has published more than 30 papers. Previous research interests included clustering, Web mining. Current research interests include machine learning, pattern recognition, support vector machine, and optimization method and its application.