ICIC Express Letters
Part B: Applications
Volume **7**, Number **6**, June **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1193–1199**

HUMAN BEHAVIORAL RECOGNITION BASED SEMI-SUPERVISED LEARNING

Zhen Liu, Jun'an Yang, Hui Liu and Wei Wang

Department of Communication
Electronic Engineering Institution
No. 460, Huangshan Road, Hefei 230037, P. R. China
ahulz@163.com

Received November 2015; accepted February 2016

Abstract. *As an important semi-supervised learning, Laplacian support vector machine utilizes the unlabeled data for learning by adding the manifold regularizer into the objective function. However, the data adjacent graph in the manifold regularizer was not good at data structure representation as the label information was neglected. Moreover, the heat kernel parameter is usually empirical fixed and neglected the local distribution information, which might also degrade the learning performance. Inspired by human behavioral learning theory, a novel semi-supervised learning with local behavioral similarity was proposed to solve those problems. In detail, a new data adjacent graph considering label information was constructed by introducing behavioral similarity based edge weight. Besides, a local distribution parameter considering the underlying probability distribution in the neighborhood was applied. Extensive experiments on public datasets show the good performance and validity of the new algorithm.*
Keywords: Semi-supervised learning, Support vector machine, Manifold learning, Behavioral learning

1. **Introduction.** In recent years, semi-supervised learning [1] has attracted a significant amount of attention as it can take advantage of both labeled and unlabeled samples for learning. Many semi-supervised learning algorithms have been proposed during the past decade, such as Co-training [2,3], Tri-training [4], linear neighborhood propagation (LNP) [5], transductive support vector machine (TSVM) [6], Laplacian support vector machine (LapSVM) [7]. Among these methods, LapSVM focuses on the regularization in the reproducing kernel Hilbert space and only needs to solve one small SVM with the labeled data. It encodes both the labeled and unlabeled data by a data adjacent graph, where each instance is represented as a vertex and two vertices are connected by an edge weight if they have large similarity. However, the data adjacent graph heavily depends on the distance metric. In real application, there are always existing data regions with overlapping class and imbalance distribution. They may cause unreasonable representation of data structure and destroy label smoothness. Generally, LapSVM utilizes heat kernel function to compute the edge weights. The performance of the heat kernel weight highly depends on the parameter selection and how to exactly fix the parameter in different applications may be troublesome. Additionally, the kernel function only focuses on the samples themselves but neglects the underlying probability distribution in the local neighborhood.

In 2013, Bryan et al. [8] proposed the concept of human semi-supervised learning, and stated in detail how human behavioral cognition guides and improves the semi-supervised learning. More and more research fruits hold the viewpoint that human behavioral learning can effectively improve the performance of machine learning [8,9]. Inspired by these booming trends, we propose a novel semi-supervised learning approach called Local Behavioral Similarity based LapSVM (LBS-LapSVM) to overcome the problems in LapSVM.

The proposed algorithm adds the label information into the data adjacent graph by behavioral similarity based edge weight. Thus, the intra-class similarity is definitely larger than inter-class similarity, which is a superior property for classification. Besides, the local distribution parameter is also applied to modify the traditional heat kernel, which can not only reflect the underlying distribution in local neighborhood but also overcome the problem of heat kernel parameter selection. Extensive experimental results demonstrate that the proposed method can more effectively and stably enhance the learning performance.

The rest of the paper is organized as follows. The principle of LapSVM is reviewed in Section 2. In Section 3, the proposed LBS-LapSVM is presented. In Section 4, extensive experiments are performed. The conclusion is given in Section 5.

2. **Principle of LapSVM.** Formally, let us represent $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ as the labeled training set and $\{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$ as the unlabeled training set, where $\boldsymbol{x}_i$ and y_i represent the feature vector and the corresponding label respectively. The objective function of LapSVM is [7]

$$f^* =_f^{\arg\min} (1/l)\Sigma_{i=1}^{l} V(\boldsymbol{x}_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{1}$$

where $V(\boldsymbol{x}_i, y_i, f)$ is a hinge cost function of the committed errors on the labeled data, $\|f\|_K^2$ penalizes f in the reproducing kernel Hilbert space to keep the smoothness of the solution. The manifold regularizer, $\|f\|_I^2 = (1/2(l+u)^2)\,\Sigma_{i,j=1}^{l+u}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2 W_{ij}$, penalizes f along a low dimensional manifold, which can be rewritten as $(1/(l+u)^2)\,\boldsymbol{f}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{f}$. $\boldsymbol{L}$ is called graph Laplacian given by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where diagonal matrix $\boldsymbol{D}$ is given by $D_{ii} = \Sigma_{j=1}^{l+u} W_{ij}$, $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{l+u})]^{\mathrm{T}}$. γ_A and γ_I are the regularization parameters. W_{ij} is the edge weight between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the data adjacent graph and is computed by heat kernel function [10]

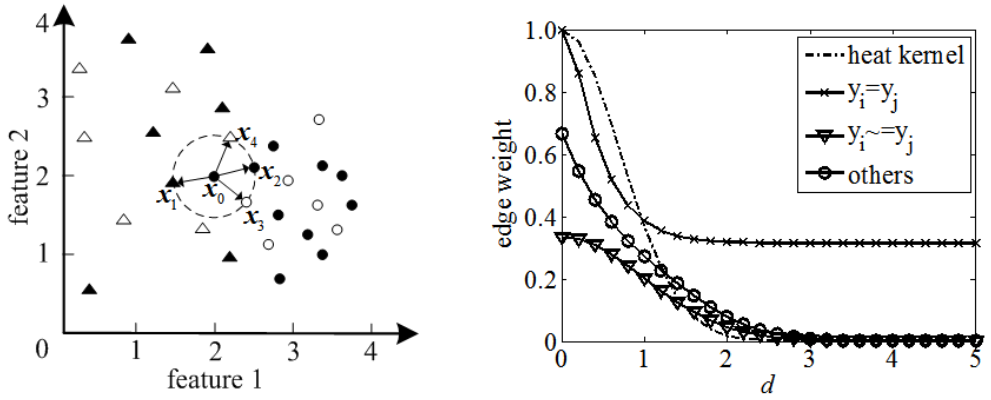
$$W_{ij} = \exp\left(-d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)/4t\right) \tag{2}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a general distance metric, and t is the heat kernel parameter. All edge weights between different samples in the whole dataset form the edge weight matrix $\boldsymbol{W}$.

The minimizer of optimization problem in Equation (1) admits a form of $f^*(\boldsymbol{x}) = \sum_{i=1}^{l+u} \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x}_i)$, where K is a kernel function. By introducing the Lagrange multiplier, $\boldsymbol{\alpha}^* = [\alpha_1^*, \alpha_2^*, \ldots, \alpha_{l+u}^*]$ can be computed by solving a quadratic programming problem like traditional SVM [7]. And there is an available toolbox called 'ManifoldLearn' [11] for implementing LapSVM.

3. **Local Behavioral Similarity for Semi-Supervised Learning.** In many real-world situations, humans are exposed to a combination of labeled data and far more unlabeled data when they need to make a classification decision. Understanding how humans combine information from labeled and unlabeled data to draw inferences can have significant social impact on the research of semi-supervised learning. Inspired by above analysis, we applied human behavioral learning strategy to semi-supervised learning.

3.1. **Behavioral similarity based edge weight for constructing new data adjacent graph.** In the manifold regularizer of LapSVM, the data adjacent graph is constructed by the edge weight matrix $\boldsymbol{W}$ without concerning the label information. As shown in Figure 1(a), the edge weights between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$ are the same according to Equation (2) as $\boldsymbol{x}_0$ has the same distance to them. However, it cannot give the accurate neighborhood structure information and does not conform to the characteristic of human behavioral recognition. When learning things, humans instinctively gather the things with the same label together even their features are not very similar and separate the things with different labels even they look similar. In Figure 1(a), human would give higher similarity between $\boldsymbol{x}_0$ and $\boldsymbol{x}_2$ when they know the label of $\boldsymbol{x}_0$ is same to $\boldsymbol{x}_2$'s but different from $\boldsymbol{x}_1$'s.



(a) (b)

FIGURE 1. Construction of data adjacent graph and behavioral similarity based edge weight

Based on the characteristic of human behavioral learning, the label information is added into the data adjacent graph. Define the behavioral similarity based edge weight as

$$W_{ij}^{BS} = \begin{cases} 1/\left(3 \times \sqrt{10/9 - W_{ij}}\right) & y_i = y_j \\ 1/\left(3 \times \sqrt{1/W_{ij}}\right) & y_i \neq y_j \\ 2/\left(3 \times \left(\sqrt{1 - W_{ij}} + \sqrt{1/W_{ij}}\right)\right) & \text{others} \end{cases} \tag{3}$$

where W_{ij} is the heat kernel computed by Equation (2). As shown in Figure 1(b), the calculation of W_{ij}^{BS} can be categorized on three settings: 1) if the labels of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are known and the same, W_{ij}^{BS} has a lager value and approaches a positive as the distance increases; 2) if the labels are known but different, W_{ij}^{BS} has a smaller value and approaches 0 as the distance increases; 3) if the labels are not all known, W_{ij}^{BS} has a value in the middle of the above two cases. Thus, the behavioral similarity based edge weight is divided into three regions based on the label information, which is a very good property for classification.

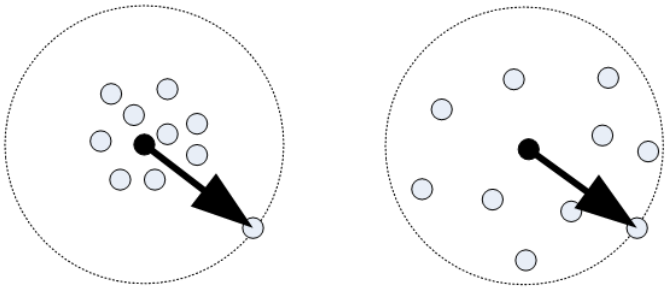
3.2. **Modified heat kernel with local distribution parameter.** In Equation (3), we still have to compute the heat kernel W_{ij} with Equation (2). The heat kernel parameter t varies in different applications and will badly hurt the learning performance if falsely set. What is more, the heat kernel weight only focuses on samples but ignore their neighborhoods. It is not the case in human behavioral paradigm that considers the underlying probability distribution in the neighborhood. Here, we define the local view distance from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$ as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j)/\rho_i \tag{4}$$

where $\rho_i = (1/N_k)\Sigma_{k=1}^{N_k} d(\boldsymbol{x}_i, \boldsymbol{x}_k)$ is called the local distribution parameter of $\boldsymbol{x}_i$, $\boldsymbol{x}_k$ is the k-th neighbor of $\boldsymbol{x}_i$, N_k is number of neighbors. The sensitivity of local view distance to $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is in inverse proportion to the density of the neighborhood distribution of central point. As shown in Figure 2, it can be explained that if the neighborhood of $\boldsymbol{x}_i$ is densely distributed, the points far away from $\boldsymbol{x}_i$ are almost impossibly similar to $\boldsymbol{x}_i$, vice versa.

Likewise, the local view distance from $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$ is $d(\boldsymbol{x}_j, \boldsymbol{x}_i)/\rho_j$. The square distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ may be generalized as $d(\boldsymbol{x}_i, \boldsymbol{x}_j)d(\boldsymbol{x}_j, \boldsymbol{x}_i)/\rho_i\rho_j = d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)/\rho_i\rho_j$. Eventually, the heat kernel function in Equation (2) can be modified as

$$W_{ij} = \exp\left(-d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)/\rho_i\rho_j\right) \tag{5}$$



(a) Dense distribution (b) Sparse distribution

FIGURE 2. Local view distance

The local distribution parameter could add the neighborhood information into the heat kernel. In addition, the modified heat kernel function with Equation (5) avoids the problem of heat kernel parameter setting in Equation (2).

3.3. **Overview of LBS-LapSVM.** In conclusion, the proposed LBS-LapSVM reconstructs the data adjacent graph. Firstly, the behavioral similarity based edge weight adds the information of the data's conditional distribution into the adjacent graph by utilizing the label message. Then, the information of marginal distribution of data' neighborhood is fed back by introducing the local distribution parameter and local view distance. LBS-LapSVM also avoids the problem of heat kernel parameter setting in LapSVM. The main procedure of LBS-LapSVM is summarized in Table 1.

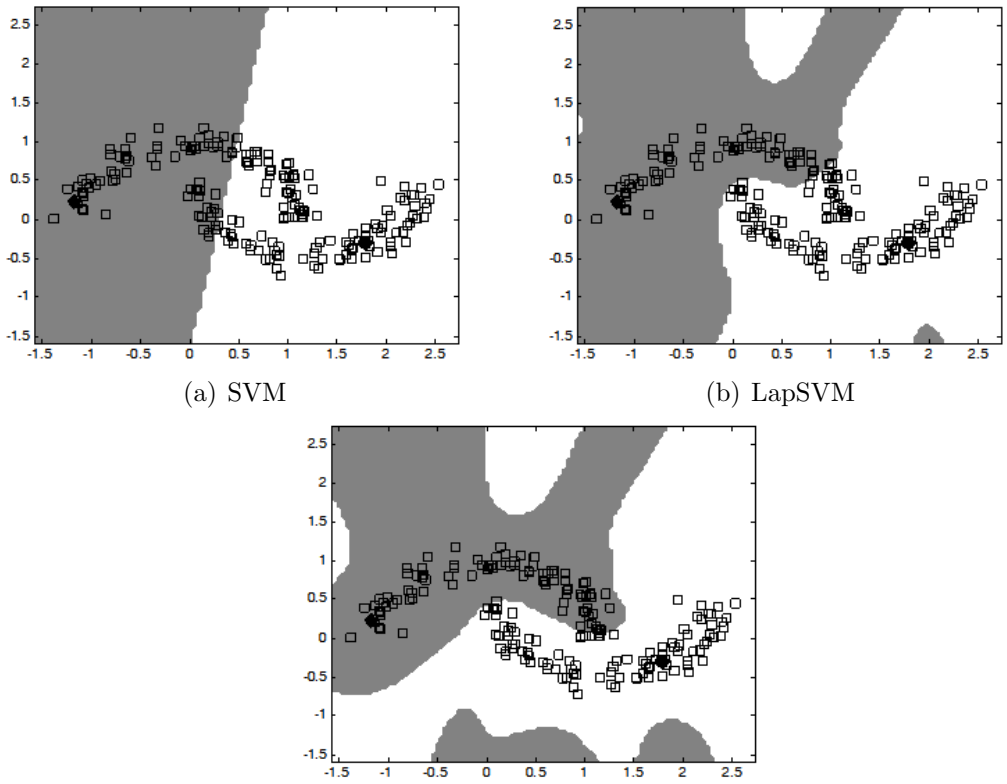
TABLE 1. The main procedure of LBS-LapSVM

Input: Labeled training set $\{(\boldsymbol{x_i}, y_i)\}_{i=l}^{l}$, unlabeled training set $\{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$.
1: Construct data adjacent graph of the whole $(l+u)$ samples using graph kernel: compute W_{ij} with Equation (5) and then compute local behavioral similarity based edge weight W_{ij}^{BS} with Equation (3).
2: Compute the new graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}^{BS}$, where $D_{ii} = \Sigma_{j=1}^{l+u} W_{ij}^{BS}$.
3: Choose a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then compute the Gram kernel matrix $\boldsymbol{K}$, where $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
4: Compute $\boldsymbol{\alpha}^*$ by solving a quadratic programming problem like traditional SVM for soft margin loss.
Output: The decision function $f^*(\boldsymbol{x}) = \Sigma_{i=1}^{l+u} \alpha_i^* K(\boldsymbol{x}, \boldsymbol{x_i})$.

4. **Experiments and Discussions.** Extensive experiments are performed on public data sets to demonstrate the validation of the proposed algorithm. The radial basis function (RBF) is chosen as the default kernel function. The parameters of all methods are set the default values as in the toolbox 'ManifoldLearn' [11] except our newly added parameter N_k. The value of local searching range N_k in LBS-LapSVM is empirically fixed as 8.

4.1. **Experiments on two moons data set.** The two moons data set contains 200 samples belonging to two non-linearly separable classes with only 1 labeled example for each class. The best decision surfaces are shown in Figure 3.

In Figure 3, SVM fails to find the optimal solution as it can only use the few labeled samples for learning. LapSVM decision surface seems to be acceptable, but it is also helpless to those most complex areas. LBS-LapSVM can effectively discover local intrinsic shape and cause the decision surface to appropriately adjust according to the geometry of the two classes. So the decision surface of LBS-LapSVM is intuitively most satisfying.



(a) SVM (b) LapSVM

(c) LBS-LapSVM

FIGURE 3. The best decision surface of two moons data set

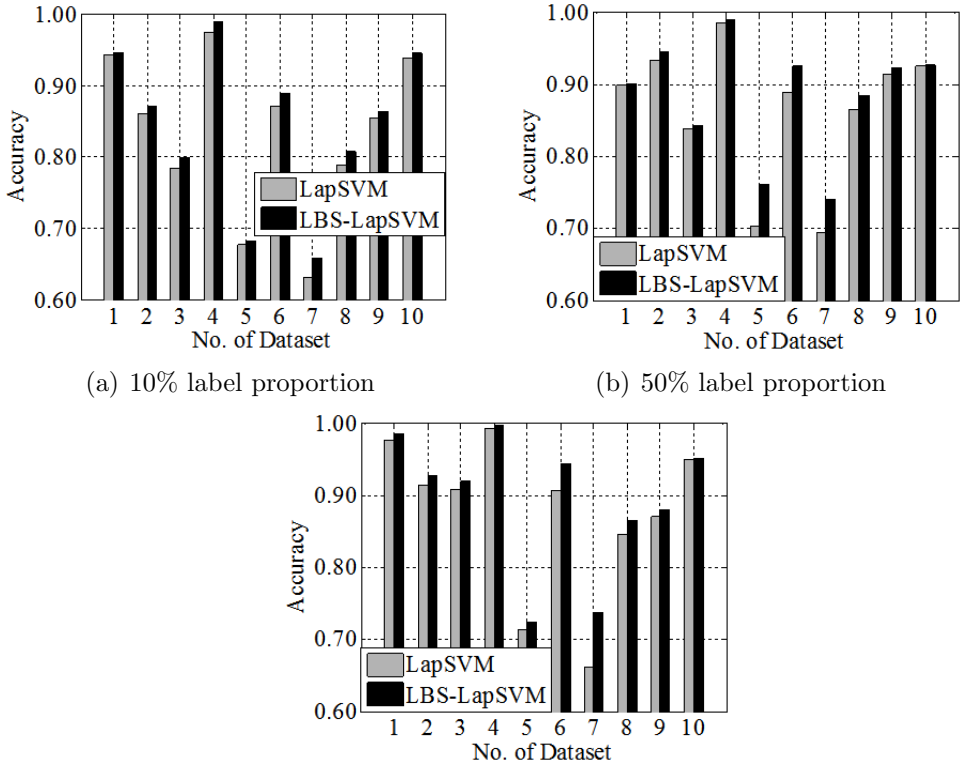
4.2. **Experiments on UCI data set.** In this section, we evaluate the proposed algorithm on 10 UCI data sets shown in Table 2. For easy description, each data set is numbered.

TABLE 2. Experimental UCI data sets

No.	Datasets	Attribute	Class	Samples
1	*SatImage*	36	6	6435
2	*Segment*	19	7	2310
3	*Ionosphere*	34	2	351
4	*Optdigits*	64	10	5620
5	*Diabetes*	8	2	768
6	*Glass*	9	7	214
7	*Haberman*	3	2	306
8	*Sonar*	60	2	208
9	*Vehicle*	18	4	846
10	*WaveForm*	40	3	5000

As experiments are designed for two-class problems, the multiclass data sets are converted into two-class data sets by randomly choosing two-classes. Then, 25% data are kept aside as test set, while the remaining 75% data are training set. The training set is partitioned into original labeled and unlabeled set with a certain proportion, e.g., 10%, 50% and 90%. We use classification accuracy as the evaluation measure. The experiments are repeated for 20 times and the average accuracies are shown in Figure 4. As SVM has very poor performance, we only give the results of LapSVM and LBS-LapSVM here.

It can be seen that LBS-LapSVM has higher accuracies than LapSVM in most cases with different datasets and different label proportions. It owes the better data structure



(a) 10% label proportion (b) 50% label proportion

(c) 90% label proportion

FIGURE 4. Average classification accuracy on UCI data sets with different label proportions

exploration by the new proposed data adjacent graph. In addition, the accuracies of both two methods improve along with the increase of label proportion as a whole. All the above experimental results demonstrate that LBS-LapSVM can effectively improve the learning performance.

4.3. **Analysis for the parameter.** The number of neighbors, N_k, is crucial to the property of local distribution parameter. For further studying LBS-LapSVM, the influence of parameter N_k on learning performance is considered. For diversity, six UCI data sets are chosen which are described as the No.3, No.5, No.6, No.7, No.8 and No.9 in Table 2. The label proportion is set as 50% and N_k is tuned in the range [2 4 6 8 10 12 14].

As shown in Figure 5, the learning performance will be badly hurt if N_k is too small or too large. If N_k is set too small, the local scope cannot cover all the affinitive examples

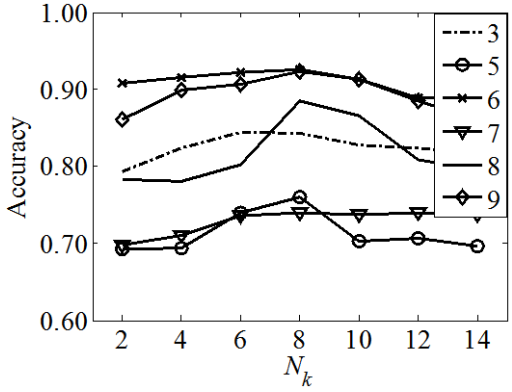


FIGURE 5. The impact of N_k on the algorithm performance

and the whole information of neighborhood structure cannot be fed back into the data adjacent graph. On the contrary, if N_k is fixed beyond normal scope, the information feedback may suffer interfere from false distribution of irrelevant data. To sum up, fixing the value of N_k at [6, 10] is recommended.

5. **Conclusions.** In this paper, a semi-supervised learning algorithm called LBS-LapSVM is proposed. Inspired by the human behavioral learning theory, the label information is added into the data adjacent graph. The information of marginal distribution of data's neighborhood is also fed back into the graph by introducing the local distribution parameter. Thus, the algorithm can do better than LapSVM in data structure exploration for learning and avoids the parameter setting problem of heat kernel. Validation of the proposed method was performed with extensive experiments. Results demonstrate that the proposed method can more effectively and stably enhance the learning performance. Furthermore, the setting of parameter N_k is analyzed based on experimental results and theory analysis. In the future, we will investigate a smart strategy to select the most informative unlabeled samples for more efficient learning and the reduction of the computational cost.

Acknowledgment. This work is supported by Key Laboratory of Electronic Restriction of Anhui Province, National High-tech R&D Program (863 Program), and Anhui Provincial Natural Science Foundation (NO. 1308085QF99, NO. 1408085MKL46).

REFERENCES

[1] J. Zhu, Semi-supervised learning literature survey, *Computer Science*, vol.37, no.1, pp.63-77, 2008.
[2] V. Ramanathan and H. Wechsler, PhishGILLNET-phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training, *Eurasip Journal on Information Security*, vol.2012, no.1, pp.1-22, 2012.
[3] J. X. Huang, J. Miao and B. He, High performance query expansion using adaptive co-training, *Information Processing and Management*, vol.49, no.2, pp.441-453, 2013.
[4] Z. H. Zhou and M. Li, Tri-training: Exploiting unlabeled data using three classifiers, *IEEE Trans. Knowledge and Data Engineering*, vol.17, no.11, pp.1529-1541, 2005.
[5] F. Wang and C. Zhang, Label propagation through linear neighborhoods, *IEEE Trans. Knowledge and Data Engineering*, vol.20, no.1, pp.55-67, 2008.
[6] A. Singla, S. Patra and L. Bruzzone, A novel classification technique based on progressive transductive SVM learning, *Pattern Recognition Letters*, vol.42, no.6, pp.101-106, 2014.
[7] M. Belkin, P. Niyogi and V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *The Journal of Machine Learning Research*, vol.7, no.3, pp.2399-2434, 2006.
[8] B. R. Gibson, T. T. Rogers and X. Zhu, Human semi-supervised learning, *Topics in Cognitive Science*, vol.5, no.1, pp.132-172, 2013.
[9] C. W. Kalish, T. T. Rogers and J. Lang, Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, vol.120, no.1, pp.106-118, 2011.
[10] V. Sindhwani, P. Niyogi and M. Belkin, Beyond the point cloud: From transductive to semi-supervised learning, *Proc. of the 22nd International Conf. Machine Learning*, Bonn, Germany, pp.824-831, 2005.
[11] V. Sindhwani, *ManifoldLearn*, http://manifold.cs.uchicago.edu/manifold_regularization/software.html.