https://doi.org/10.1016/j.asoc.2017.06.017

Semi-Supervised Learning using Hidden Feature Augmentation

Wenlong Hang, Kup-Sze Choi, Shitong Wang, Pengjiang Qian

***Abstract*—Semi-supervised learning methods are conventionally conducted by simultaneously utilizing abundant unlabeled samples and a few labeled samples given. However, the unlabeled samples are usually adopted with assumptions, e.g. cluster and manifold assumptions, which degrade the performance when the assumptions become invalid. The reliable hidden features embedded in both the labeled and the unlabeled samples can potentially be used to tackle this issue. In this regard, we investigate the feature augmentation technique to improve the robustness of semi-supervised learning in this paper. By introducing an orthonormal projection matrix, we first transform both the unlabeled and labeled samples into a shared hidden subspace to determine the connections between the samples, and utilize the hidden features, the raw features, and zero vectors determined to develop a novel feature augmentation strategy. Finally, a hidden feature transformation model is proposed to compute the desired projection matrix by applying the maximum joint probability distribution principle in the augmented feature space. The effectiveness of the proposed method is evaluated in terms of the hinge and square loss functions respectively, based on two types of semi-supervised classification formulations developed using only the labeled samples with their original features and hidden features. The experimental results have demonstrated the effectiveness of the proposed feature augmentation technique for semi-supervised learning.**

***Index Terms*—Semi-supervised learning, Cluster assumption, Manifold assumption, Hidden features, Joint probability distribution**

I. Introduction

Semi-supervised learning finds applications in various domains, such as machine learning, pattern recognition, image processing, computer vision and bioinformatics. The performance of semi-supervised learning is usually dependent on the availability of abundant labeled samples, without which promising model learning performance is difficult to achieve. In many real-world applications, collecting full classes of labeled samples is labor-intensive or even impracticable, whereas acquiring a large amount of unlabeled data is relatively feasible. Hence, semi-supervised learning using unlabeled data has received considerable attention [1, 2, 5, 8]. In this approach, the intrinsic structure of the data is critical to the performance. For unlabeled data, it is necessary make assumptions on the intrinsic data structure but the validity can adversely affect the learning performance, as demonstrated in many empirical studies [30]. Specifically, the wide variation in modalities and distributions in different datasets, or even the variation of data distribution among different clusters within a dataset, makes it impractical to accurately model every dataset based on a few common and straightforward assumptions. For example, in manifold learning [3], Laplacian matrix is enlisted to depict the manifold structure existing in a dataset. Nevertheless, different structures could be disclosed with different choices of k-nearest-neighbors. Therefore, it is of significance to investigate more reliable and robust strategies for using unlabeled samples in semi-supervised learning. This is the motivation underlying the research in this paper.

It has been illustrated empirically that for a classifier, the negative influence incurred by feature errors is far less than that by incorrect labels [4]. That is, under a certain assumption, e.g. manifold preservation, the process of automated labeling may result in wrongly tagged labels which can propagate and seriously affect the performance of the classifier. Ideally, a robust semi-supervised learning method should take advantages of both labeled and unlabeled samples while avoiding the negative effects due to incorrect labels. However, achieving these goals simultaneously is difficult. In practice, classification approaches safeguarding invalid assumptions often produce little or even no improvement in classification performance. The need to make assumptions in existing semi-supervised learning algorithms is indeed a major hurdle against effective leveraging of the unlabeled samples.

Since the adverse effect due to attribution error is essentially less than that caused by class label error [4], it is more beneficial to make use of the hidden features based on the complete unlabeled samples information, rather than relying on brute-force assumptions in the model training process. In fact, hidden features also play an important role in the human cognition. The theory of adaptive control of thought (ACT) developed by John Anderson is a well-known theory in the community of cognitive psychology [21]. According to ACT, declarative knowledge and procedural knowledge are assumed to be two critical atomic components of human cognition. Explicit declarative knowledge is not always accessible. It is more common that based on the existing declarative knowledge, along with some unconscious data inference (hidden features) and retrieval of information, procedural

This work was supported in part by the Research Grants Council of the Hong Kong SAR (PolyU5134/12E), the National Natural Science Foundation of China under Grants 61272210, the Natural Science Foundation of Jiangsu Province under Grant BK20130155, BK20130161, the Jiangsu 333 expert engineering grant (BRA2011142), and the 2013 Postgraduate Student's Creative Research Fund of Jiangsu Province.

W.L. Hang is with the School of Digital Media, Jiangnan University, Wuxi 214122, China, and also with the School of Nursing, Hong Kong Polytechnic University, Hong Kong (e-mail: hwl881018@163.com)

K.-S. Choi is with the Centre for Smart Health, School of Nursing, Hong Kong Polytechnic University, Hong Kong (e-mail: kschoi@ieee.org)

P.J. Qian and S.T. Wang (corresponding author) are with the School of Digital Media, Jiangnan University, Wuxi 214122, China (wxwangst@aliyun.com; qianpjiang@126.com)

knowledge gradually can generates new declarative knowledge. In other words, human can unconsciously abstract hidden features from the procedural knowledge to infer and create new declarative knowledge. Enlightened by the learning approach of human cognition, we adopt an analogy to propose the research in this paper.

In this paper, the problem of semi-supervised learning is investigated from the perspective of hidden feature augmentation. We firstly introduce an orthonormal projection matrix to transform both the labeled and unlabeled samples into a common subspace such that the connection of objects (to be classified) belonging to the same class can be maximized. According to the principle of maximum joint probability distribution between the labeled and unlabeled samples in the augmented feature space, we then propose the Hidden Feature Transformation (HFT) model to obtain the hidden feature projection matrix shared by them. Finally, semi-supervised classification formulations are developed respectively based on two typical loss functions. The merits of the proposed approach are highlighted as follows.

(1) With the hidden feature augmentation strategy, a new mechanism to effectively leverage unlabeled samples in semi-supervised learning is first proposed. Instead of brute-force assumptions, hidden features embedded in both the labeled and unlabeled samples are mined and utilized. The proposed approach avoids the propagation of labeling errors commonly existing in semi-supervised learning, which enables practical and reliable utilization of then unlabeled data by establishing connections between the labeled and unlabeled data.

(2) The novel HFT model is proposed based on the maximum joint probability principle to extract hidden features in the samples. It guarantees the labeled and unlabeled data in the augmented feature space have similar data distributions. In addition, the desired hidden space projection matrix can be obtained analytically.

(3) In the two proposed semi-supervised classification formulations, the augmented labeled instances and their original and hidden features are only involved, which make it easy to solve the classification problems.

(4) Besides classification, the proposed hidden space augmentation based semi-supervised learning mechanism can also be conveniently used in other applications, such as semi-supervised clustering, regression, and fuzzy inference.

The remainder of this paper is organized as follows. A brief review of the related work is provided in Section 2. The proposed semi-supervised learning framework and the feature augmentation modality, as well as the estimation of the hidden space projection matrix and the semi-supervised classification formulations are presented in Section 3. Extensive experimental studies on the proposed method and the associated analyses are discussed in Section 4. Conclusions and the possible avenues for future research are given in the last section.

II. Related Works

Semi-supervised classification methods attempt to improve the performance of classifier training by exploiting both the relatively ample yet unlabeled data (with potential to disclose the intrinsic data pattern), and a small amount of labeled samples available. Numerous semi-supervised classification approaches have been developed in the past decades [5-7]. Most of them rely on making two major types of assumptions on the unlabeled data, namely, the cluster assumption and the manifold assumption [1, 5, 8]. The cluster assumption presumes that similar instances have the same class labels and the classification boundaries should pass through low density regions [9-12]. The manifold assumption supposes that the data are distributed in some low dimensional manifolds and similar instances have the same data feature. For example, data distribution and geometric structure are generally depicted by the Laplacian graph [3, 13-15].

However, empirical studies have shown that in some cases, utilization of unlabeled data indeed decreases the learning performance since the assumptions made cannot be met [30, 31]. It is in fact difficult to estimate the potentially useful but unknown data pattern of the unlabeled data. Different assumptions can lead to different outcomes, or even deteriorates classification effectiveness if the assumptions made turn out to be inappropriate. This is evident from the example demonstrated in Fig. 1, where two categories of objects – round-shaped objects and quadrilateral objects – are presented in Fig. 1(a); and the classification results achieved with different assumptions on a specific dataset are shown in Fig. 1(b). On the one hand, because of the rough shape similarity between the rectangles and the ellipses and that between the circles and the squares, making the cluster assumption may result in the classification hyperplane shown with the blue dotted line in Fig. 1(b). On the other hand, based on the manifold assumption, another classification hyperplane, shown with the red dotted line in the figure, may be obtained according to the different data densities of the different classes. The figure illustrates intuitively the effect of assumptions on semi-supervised classification using unlabeled data.

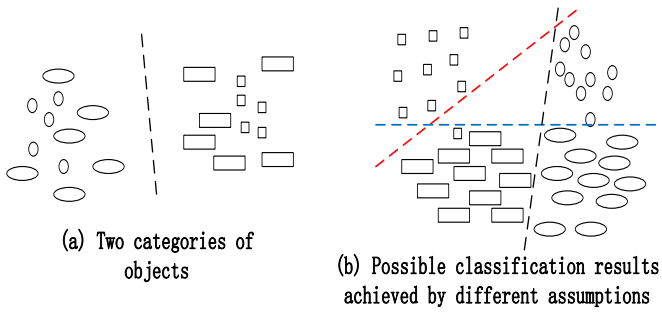


Fig. 1. Effect of assumptions on semi-supervised learning.

In some assumption-based methods, the pseudo-label generation trick is typically applied to part of the unlabeled data to expand the labeled dataset [1, 5, 8]. However, if the pseudo-labels generated based on the assumptions are erroneous, information due to the wrong labels can propagate iteratively and eventually degrade the classification performance significantly. A data editing technique has been

proposed as a counter measure [27]. However, this technique relies on the neighboring information and only works well in some dense-data scenarios.

Other semi-supervised learning methods concerning data leveraging as well as negative influence avoidance have also been investigated. For example, Wang et al. developed a safety-aware semi-supervised learning (SA-SSCCM) method [18] which is a compromise between the modified cluster assumption and least-square support vector machine (LS-SVM) [20] in order to tackle the sole dependence of the cluster assumption. The safe semi-supervised SVMs (S4VMs) was also proposed to only exploit the candidate low-density separators to assist model training, under the assumption that the ground-truth label could be attained by one of the low-density separators obtained [19]. This method attempts to achieve better performance based on the extrinsic data density property, which, to some extent, is equivalent to the assumption of density-based spatial cluster [32].

In light of the fact that inappropriate assumptions not only mismatch the intrinsic data pattern but also degrade the performance of model training, there is a need to develop more reliable data-leveraging mechanisms, rather than resorting to brute-force assumptions.

III. Semi-Supervised Learning based on Feature Augmentation

Instead of making assumptions on the data pattern, we focus on semi-supervised classification with the feature augmentation strategy in this paper. Before introducing the details of this approach, the notations used in the paper are first described.

A. Mathematical Notations

The mathematical notations and definitions used in this paper are introduced as follows. $\mathbf{0}_n$, $\mathbf{1}_n \in \mathbb{R}^n$ denotes the $n\times 1$ column vectors of all zeros and all ones respectively. For simplicity, $\mathbf{0}$ and $\mathbf{1}$ are used instead of $\mathbf{0}_n$ and $\mathbf{1}_n$ when the dimension is explicit. Suppose a d dimensional dataset S with N samples, $S=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\}$, is given, the N_l labeled samples are denoted by $L=\left\{\left(\mathbf{x}_1^L,y_1\right),\left(\mathbf{x}_2^L,y_2\right),...,\left(\mathbf{x}_{N_l}^L,y_{N_l}\right)\right\}$ with labels $y_i\in\{-1,+1\}$, and the remaining $N_u=N-N_l$ unlabeled samples are denoted as $U=\left\{\left(\mathbf{x}_{N_l+1}^U,y_{l+1}\right),\left(\mathbf{x}_{N_l+2}^U,y_{N_l+2}\right),...,\left(\mathbf{x}_N^U,y_N\right)\right\}$. The classification problem can be formulated as an optimization problem which aims at finding the decision function f by minimizing the structural risk $\Phi(f)+C\sum_{i=1}^{N}\mathrm{L}\left(f\left(\mathbf{x}_i\right),y_i\right)$, where $\Phi(f)$ is a regularization term and $\mathrm{L}(\cdot)$ is a certain convex non-negative loss function to assesses the effectiveness of the f . In our study, two classical loss functions are concerned, i.e., the hinge loss function and the least square loss function [28].

B. Feature Augmentation Based Semi-Supervised Framework

In the paper, we introduce a common orthonormal projection matrix $\mathbf{P}\in\mathbb{R}^{r\times d}$ to transform both the labeled and unlabeled samples into shared hidden subspace, in which r is the dimension of hidden feature space and the range of r will be discussed later. The shared hidden subspace is referred to as the *augmented future space*, which is composed of the generated hidden features, the original features, and the zero vectors. In the augmented future space, the HFT model is proposed based on the principle of maximum joint probability to calculate the desired projection matrix **P**. Eventually, augmented labeled samples, produced by using the ample unlabeled and the common projection matrix **P**, are exploited to improve the performance of the classifier. The framework of the semi-supervised learning approach is illustrated in Fig. 2, where the proposed assumption-free method is compared with the conventional assumption-based method.

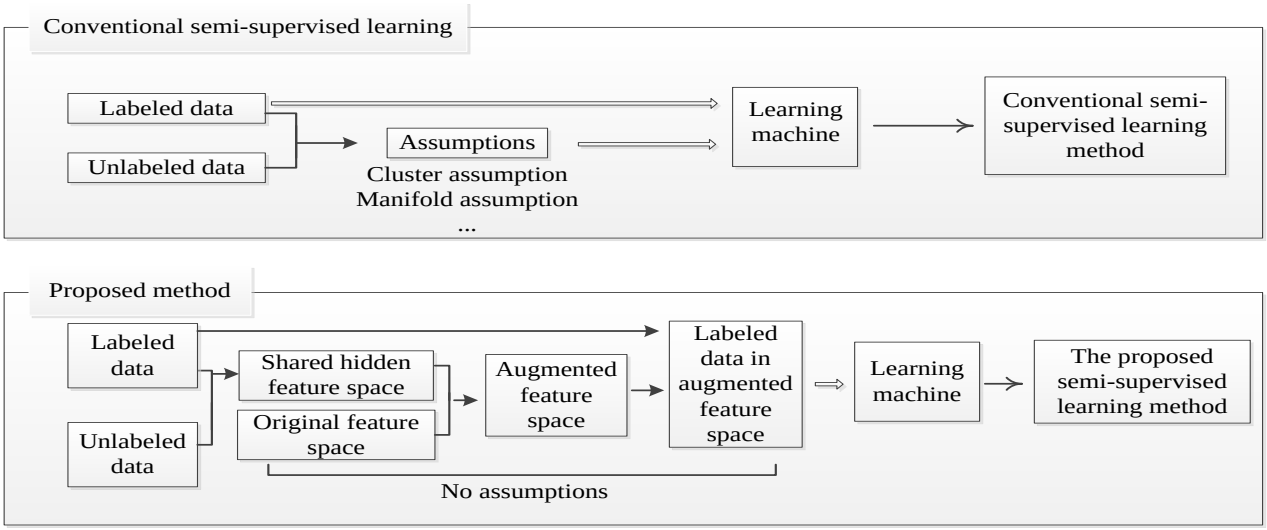


Fig. 2. The proposed assumption-free method versus conventional supervised-learning method.

The proposed method figures out the shared hidden space via the feature augmentation mechanism, where the hidden features, generated by using the projection matrix P for each

original sample in the labeled dataset or unlabeled dataset, are regarded as a type of implicit knowledge. Based on the supplementary knowledge, i.e., the augmented features, and the original features, classification of the samples can be performed more effectively.

C. Feature Augmentation

A simple domain-adaption-based feature augmentation method has been proposed by Daumé recently [22], where the original feature space $\mathbb{R}^d$ is projected into the augmented feature space $\mathbb{R}^{3d}$ by merely replicating the original features and zeroes. Specifically, for any data point $\mathbf{x} \in \mathbb{R}^d$ from the source or target domain (where useful knowledge can be exploited from the source domain to help learning in the target domain), define the feature mapping functions $\Phi^S(\mathbf{x}) = \langle \mathbf{x}, \mathbf{0}, \mathbf{x} \rangle$ and $\Phi^T(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}, \mathbf{0} \rangle$ for the source domain and the target domain respectively. Inspired by this idea, while developing a different way of feature augmentation, we apply the hidden features into the feature space in order to establish connections between the labeled and unlabeled samples. In this paper, the primitive feature augmentation structure is slightly modified by replacing the original components with the hidden features for both the labeled and the unlabeled sets. That is

$$\mathbf{X}_i^L = \left[\mathbf{P}\mathbf{x}_i^L, \mathbf{x}_i^L, \mathbf{0}\right]^T,\ \mathbf{X}_j^U = \left[\mathbf{P}\mathbf{x}_j^U, \mathbf{0}, \mathbf{x}_j^U\right]^T, \tag{1}$$

Note that it is not meaningful to directly use the method proposed by Daumé [22] for semi-supervised learning tasks simply by padding zeros to equalize the dimensions of the data from the labeled dataset and the unlabeled dataset, which would result in the absence of correspondences between the two datasets. With the projection matrix **P** in our method, the hidden features obtained can be used to facilitate the discovery of the connection of the samples belonging to the same classes in the entire dataset. In addition, the new feature augmentation method proposed only makes use of the hidden features to establish the connection between individual samples. This precludes the need of making assumptions and prevents the propagation of labeling noise in the model training procedure.

D. Hidden Feature Extraction in Augmented Feature Space

In this section, the HFT model is first presented, followed by the projection matrix **P** which is obtained based on the principle of maximum joint probability distribution in the augmented feature space and the HFT model.

1) The HFT model

Using the novel feature augmentation mechanism presented in (1), all samples, either from the labeled dataset or the unlabeled dataset, are mapped into the augmented feature space established. For the labeled dataset L and the unlabeled dataset U in the augmented feature space, with the Gaussian distribution function $G(\mathbf{X}, \mathbf{X}_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\|\mathbf{X} - \mathbf{X}_i\|^2 / 2\sigma^2\right)$, the Parzen-window based density functions [23] can be respectively expressed as

$$G\left(\mathbf{X}, \mathbf{X}_i^L, \sigma^2\right) = \frac{1}{N_l \cdot \sqrt{2\pi}\sigma} \sum_{i=1}^{N_l} e^{-\frac{\left\|\mathbf{X} - \mathbf{X}_i^L\right\|^2}{2\sigma^2}} \text{ and} \tag{2}$$

$$G\left(\mathbf{X}, \mathbf{X}_j^U, \sigma^2\right) = \frac{1}{N_u \cdot \sqrt{2\pi}\sigma} \sum_{j=1}^{N_u} e^{-\frac{\left\|\mathbf{X} - \mathbf{X}_j^U\right\|^2}{2\sigma^2}}, \tag{3}$$

where σ denotes the Gaussian kernel bandwidth. Furthermore, the Gaussian probability density of $P_L(\mathbf{X})$ and $P_U(\mathbf{X})$ can be separately expressed as

$$P_L(\mathbf{X}) = \frac{1}{N_l} \sum_{i=1}^{N_l} G\left(\mathbf{X}, \mathbf{X}_i^L, \sigma^2\right) \text{ and} \tag{4}$$

$$P_U(\mathbf{X}) = \frac{1}{N_u} \sum_{j=1}^{N_u} G\left(\mathbf{X}, \mathbf{X}_j^U, \sigma^2\right). \tag{5}$$

To measure the difference between these two density distributions, the integrated squared error J_0 between $P_L(\mathbf{X})$ and $P_U(\mathbf{X})$ is calculated, i.e.,

$$J_0 = \int \left(P_L(\mathbf{X}) - P_U(\mathbf{X})\right)^2 d\mathbf{X}. \tag{6}$$

By minimizing J_0, the labeled dataset and unlabeled dataset are expected to have the maximal commonality in the projective hidden feature space. It can be shown that by expressing J_0 as $J_0 = \int P_L^2(\mathbf{X}) d\mathbf{X} - 2\int P_L(\mathbf{X}) P_U(\mathbf{X}) d\mathbf{X} + \int P_U^2(\mathbf{X}) d\mathbf{X}$ and based on the results in [24, 34], the relationship $\int G(\mathbf{X}, \mathbf{X}_i, \sigma_1) G(\mathbf{X}, \mathbf{X}_j, \sigma_2) d\mathbf{X} = G(\mathbf{X}_i - \mathbf{X}_j, \sigma_1 + \sigma_2)$ holds. Hence, the following equations can be obtained.

$$\begin{aligned} \int P_L^2(\mathbf{X}) d\mathbf{X} &= \frac{1}{N_l^2} \sum_{i=1}^{N_l} \sum_{j=1}^{N_l} G\left(\mathbf{X}_i^L, \mathbf{X}_j^L, 2\sigma^2\right) \\ &= \frac{1}{N_l} \sum_{i=1}^{N_l} \left[\frac{1}{N_l} \sum_{j=1}^{N_l} G\left(\mathbf{X}_i^L, \mathbf{X}_j^L, 2\sigma^2\right) \right], \end{aligned} \tag{6.1}$$

$$\begin{aligned} \int P_U^2(\mathbf{X}) d\mathbf{X} &= \frac{1}{N_u^2} \sum_{i=1}^{N_u} \sum_{j=1}^{N_u} G\left(\mathbf{X}_i^U, \mathbf{X}_j^U, 2\sigma^2\right) \\ &= \frac{1}{N_u} \sum_{i=1}^{N_u} \left[\frac{1}{N_u} \sum_{j=1}^{N_u} G\left(\mathbf{X}_i^U, \mathbf{X}_j^U, 2\sigma^2\right) \right], \end{aligned} \tag{6.2}$$

$$\int P_L(\mathbf{X}) P_U(\mathbf{X}) d\mathbf{X} = \frac{1}{N_l N_u} \sum_{i=1}^{N_l} \sum_{j=1}^{N_u} G\left(\mathbf{X}_i^L, \mathbf{X}_j^U, 2\sigma^2\right). \tag{6.3}$$

Here, the expression $\frac{1}{N_l} \sum_{j=1}^{N_l} G\left(\mathbf{X}_i^L, \mathbf{X}_j^L, 2\sigma^2\right)$ can be regarded as another estimate of $P_L\left(\mathbf{X}_i^L\right)$ with bandwidth $\sqrt{2}\sigma$, and thus $\int P_L^2(\mathbf{X}) d\mathbf{X}$ can be approximated by $\frac{1}{N_l} \sum_{i=1}^{N_l} P_L\left(\mathbf{X}_i^L\right)$ and further

reduced to $\frac{1}{N_l}$. Similarly, $\int P_U^2(\mathbf{X})d\mathbf{X}$ can be approximated by $\frac{1}{N_u}$. Eventually, J_0 can be approximated as

$$J_0 \approx \frac{1}{N_l}+\frac{1}{N_u}-\frac{2}{N_l N_u}\sum_{i=1}^{N_l}\sum_{j=1}^{N_u} G\left(\mathbf{X}_i^L,\mathbf{X}_j^U,2\sigma^2\right), \tag{7}$$

and the minimization of J_0 is then equivalent to the maximization of J_0' in the following expression

$$J_0' = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u} G\left(\mathbf{X}_i^L,\mathbf{X}_j^U,2\sigma^2\right) = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u} \frac{1}{\sqrt{4\pi}\sigma} e^{-\frac{\left\|\mathbf{X}_i^L-\mathbf{X}_j^U\right\|^2}{4\sigma^2}}. \tag{8}$$

By using the Taylor series expansion

$$e^x = 1+\frac{x^1}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots=\sum_{n=0}^{\infty}\frac{x^n}{n!}, \tag{9}$$

with the exponential function e^x at 0, we arrive at the following approximation

$$e^{-x} \approx 1-x. \tag{10}$$

Therefore, in order to maximize the joint probability distribution of the labeled dataset L and the unlabeled dataset U with the new representations in the augmented feature space, we can instead minimize the following objective

$$J = J_0' \approx \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\left(\mathbf{X}_i^L-\mathbf{X}_j^U\right)^2, \tag{11}$$

subject to the orthogonal condition of the projection matrix $\mathbf{P}$, i.e. $\mathbf{P}\mathbf{P}^T=\mathbf{I}_{d\times d}$. Finally, the following optimization objective function is obtained.

$$\begin{aligned} \min\quad & J(\mathbf{P}) = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\left(\mathbf{X}_i^L-\mathbf{X}_j^U\right)^2 \\ \text{s.t.}\quad & \mathbf{P}\mathbf{P}^T=\mathbf{I}_{d\times d}. \end{aligned} \tag{12}$$

After mathematical derivations, this optimization problem can be reformulated as:

$$\min\quad J(\mathbf{P}) = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}\left(\mathbf{x}_i^L\right)^T\mathbf{P}^T\mathbf{P}\mathbf{x}_i^L+\left(\mathbf{x}_i^L\right)^T\mathbf{x}_i^L+\\ \left(\mathbf{x}_j^U\right)^T\mathbf{P}^T\mathbf{P}\mathbf{x}_j^U+\left(\mathbf{x}_j^U\right)^T\mathbf{x}_j^U\\ -2\left(\mathbf{x}_i^L\right)^T\mathbf{P}^T\mathbf{P}\mathbf{x}_j^U\end{pmatrix} \tag{13}$$

$$\text{s.t.}\quad \mathbf{P}\mathbf{P}^T=\mathbf{I}_{d\times d}.$$

By taking the derivative of the above objective function with respect to $\mathbf{P}$, we have

$$\frac{\partial J}{\partial \mathbf{P}} = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}2\mathbf{P}\mathbf{x}_i^L\left(\mathbf{x}_i^L\right)^T+2\mathbf{P}\mathbf{x}_j^U\left(\mathbf{x}_j^U\right)^T-\\ 2\mathbf{P}\left(\mathbf{x}_i^L\left(\mathbf{x}_j^U\right)^T+\mathbf{x}_j^U\left(\mathbf{x}_i^L\right)^T\right)\end{pmatrix}. \tag{14}$$

Then the projection matrix $\mathbf{P}$ can be estimated using the following gradient descent procedure [25-27, 33]

$$\mathbf{P}\leftarrow\mathbf{P}-\eta\frac{\partial J}{\partial \mathbf{P}}\left(\mathbf{I}_{d\times d}-\mathbf{P}\mathbf{P}^T\right)=\mathbf{P}-\eta\nabla\mathbf{P}. \tag{15}$$

That is,

$$\begin{aligned}\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)} - \\ \eta\left(\sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}2\mathbf{P}^{(t)}\mathbf{x}_i^L\left(\mathbf{x}_i^L\right)^T+2\mathbf{P}^{(t)}\mathbf{x}_j^U\left(\mathbf{x}_j^U\right)^T-\\ 2\mathbf{P}^{(t)}\left(\mathbf{x}_i^L\left(\mathbf{x}_j^U\right)^T+\mathbf{x}_j^U\left(\mathbf{x}_i^L\right)^T\right)\end{pmatrix}\right)\\ \times\left(\mathbf{I}_{d\times d}-\left(\mathbf{P}^{(t)}\right)^T\mathbf{P}^{(t)}\right)^T\end{aligned} \tag{16}$$

The step size η in (16) can either be manually set or analytically obtained using the techniques described below. After putting (15) into the objective function in (13), the following continuous function $g(\eta)$ and its derivative with respect to η can be obtained.

$$g(\eta) = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}\left(\mathbf{x}_i^L\right)^T(\mathbf{P}-\eta\nabla\mathbf{P})^T(\mathbf{P}-\eta\nabla\mathbf{P})\mathbf{x}_i^L+\left(\mathbf{x}_i^L\right)^T\mathbf{x}_i^L+\\ \left(\mathbf{x}_j^U\right)^T(\mathbf{P}-\eta\nabla\mathbf{P})^T(\mathbf{P}-\eta\nabla\mathbf{P})\mathbf{x}_j^U+\left(\mathbf{x}_j^U\right)^T\mathbf{x}_j^U-\\ 2\left(\mathbf{x}_i^L\right)^T(\mathbf{P}-\eta\nabla\mathbf{P})^T(\mathbf{P}-\eta\nabla\mathbf{P})\mathbf{x}_j^U\end{pmatrix}. \tag{17}$$

$$\frac{\partial g}{\partial \eta} = \sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}2\eta\left(\nabla\mathbf{P}\mathbf{x}_i^L\right)^T\nabla\mathbf{P}\mathbf{x}_i^L-\left(\mathbf{P}\mathbf{x}_i^L\right)^T\nabla\mathbf{P}\mathbf{x}_i^L-\\ \left(\nabla\mathbf{P}\mathbf{x}_i^L\right)^T\mathbf{P}\mathbf{x}_i^L+2\eta\left(\nabla\mathbf{P}\mathbf{x}_j^U\right)^T\nabla\mathbf{P}\mathbf{x}_j^U-\\ \left(\mathbf{P}\mathbf{x}_j^U\right)^T\nabla\mathbf{P}\mathbf{x}_j^U-\left(\nabla\mathbf{P}\mathbf{x}_j^U\right)^T\mathbf{P}\mathbf{x}_j^U-\\ 2\begin{pmatrix}2\eta\left(\nabla\mathbf{P}\mathbf{x}_i^L\right)^T\nabla\mathbf{P}\mathbf{x}_j^U-\left(\mathbf{P}\mathbf{x}_i^L\right)^T\nabla\mathbf{P}\mathbf{x}_j^U\\ -\left(\nabla\mathbf{P}\mathbf{x}_i^L\right)^T\mathbf{P}\mathbf{x}_j^U\end{pmatrix}\end{pmatrix}. \tag{18}$$

Set $\frac{\partial g}{\partial \eta}=0$, the step size η is given by

$$\eta = \frac{\sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}\left(\mathbf{x}_i^L\right)^T\left(\mathbf{P}^T\nabla\mathbf{P}+\nabla\mathbf{P}^T\mathbf{P}\right)\mathbf{x}_i^L+\\ \left(\mathbf{x}_j^U\right)^T\left(\mathbf{P}^T\nabla\mathbf{P}+\nabla\mathbf{P}^T\mathbf{P}\right)\mathbf{x}_j^U-\\ 2\left(\mathbf{x}_i^L\right)^T\left(\mathbf{P}^T\nabla\mathbf{P}+\nabla\mathbf{P}^T\mathbf{P}\right)\mathbf{x}_j^U\end{pmatrix}}{\sum_{i=1}^{N_l}\sum_{j=1}^{N_u}\begin{pmatrix}2\left(\mathbf{x}_i^L\right)^T\nabla\mathbf{P}^T\nabla\mathbf{P}\mathbf{x}_i^L+2\left(\mathbf{x}_j^U\right)^T\nabla\mathbf{P}^T\nabla\mathbf{P}\mathbf{x}_j^U\\ -4\left(\mathbf{x}_i^L\right)^T\nabla\mathbf{P}^T\nabla\mathbf{P}\mathbf{x}_j^U\end{pmatrix}} \tag{19}$$

In summary, the procedure to obtain the projection matrix $\mathbf{P}$ is presented in Algorithm 1.

Algorithm 1: Hidden Feature Transformation

Input: labeled dataset L, unlabeled dataset U

Initialize: $t=0$, $\forall$ row in the orthonormal matrix $\mathbf{P}^{(0)}$, obtain the gradient $\partial J/\partial \mathbf{P}$ according to (14).

repeat

$t=t+1$.

Compute the increment $\nabla \mathbf{P}$ according to (15).

Use (17) to update step size η.

Obtain the tth projection matrix $\mathbf{P}^{(t)}=\mathbf{P}^{(t-1)}-\eta\nabla\mathbf{P}$ by (16).

Until

$\left\|J(t)-J(t-1)\right\|\le\delta$ or $t\ge t_{\max}$.

Output: $\mathbf{P}$

2) Hidden Feature Formulation

The hidden features of the given dataset, i.e., $\mathbf{Px}^L$ and $\mathbf{Px}^U$ of the labeled and unlabeled samples respectively, can be readily obtained with equation (1) and Algorithm 1. However, it is necessary to determine the number of hidden features r in the HFT model. Assume the samples in L and U are independent identity distribution with their labels randomly discarded during the sample generation process, it is sufficient to set r to any number greater than the total number of classes of the dataset. On the other hand, since the hidden features are combined with the origin features to determine the unlabeled samples, it is necessary to restrict r below a reasonable threshold even if the dimension of the original features d is relatively high to refrain from the difficulty in handling high-dimensional data.

Regarding the time complexity of the HFT procedure, as the solution of $\mathbf{P}$ is obtained using the gradient descent strategy and the step length is determined analytically, the computational complexity of $\mathbf{P}$ is $O(trd^2)$, where t is the iteration number allowed to compute $\mathbf{P}$.

E. Formative Semi-supervised Learning

With the projection matrix $\mathbf{P}$ and the hidden features, the labeled and unlabeled samples can be readily expressed in the new augmented feature space. A notable feature of the proposed feature-augmentation-based semi-supervised classification mechanism is that only the labeled samples, together with the original and the corresponding hidden features, are involved in the classifier training. The unlabeled samples are no longer needed.

In theory, the presented hidden space augmentation strategy is applicable for most semi-supervised learning models, e.g. classification, clustering, and regression, as it can be conveniently incorporated into the corresponding frameworks. In this study, we focus on semi-supervised SVM classification formulations in terms of two typical loss functions – hinge loss function and least square loss function – to verify the effectiveness of the proposed approach.

With the hinge loss function, the semi-supervised SVM formulation can be represented as

$$\begin{aligned}&\min\ \frac{1}{2}\left(\|\mathbf{w}\|^2+\|\mathbf{v}\|^2\right)+C\sum_{i=1}^{N}\xi_i\\&\text{s.t.}\quad y_i\left(\left(\mathbf{w}+\mathbf{P}^T\mathbf{v}\right)^T\mathbf{x}_i+b\right)\ge 1-\xi_i,\xi_i\ge 0.\end{aligned}\tag{20}$$

Using a mapping function $\phi(\bullet)$ to map all samples into the high dimensional Hilbert space, the dual problem of (20) is given by

$$\begin{aligned}&\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i y_i\begin{pmatrix}\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)+\\ \phi(\mathbf{Px}_i)^T\phi(\mathbf{Px}_j)\end{pmatrix}y_j\alpha_j+\sum_{i=1}^{N}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0,\ 0\le\alpha_i\le C.i=1,2,...,N\end{aligned}\tag{21}$$

Likewise, for the least square loss function, the dual problem is given by

$$\begin{aligned}&\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i y_i\begin{pmatrix}\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)+\\ \phi(\mathbf{Px}_i)^T\phi(\mathbf{Px}_j)+\dfrac{\delta_{ij}}{C}\end{pmatrix}y_j\alpha_j+\sum_{i=1}^{N}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0.i=1,2,...,N\end{aligned}\tag{22}$$

The optimization problems in (21) and (22) can both be solved by transforming them into standard quadratic programming problems.

Besides, suppose $\mathbf{K}(\cdot)$ is the kernel function involved, the decision function of the above semi-supervised classification problems can be presented as $f(\mathbf{x})=\sum_{i=1}^{N}\alpha_i\left(\mathbf{K}(\mathbf{x},\mathbf{x}_i)+\mathbf{K}(\mathbf{Px},\mathbf{Px}_i)\right)$ according to the Representer theorem [3].

IV. Experiments

In this section, we conducted extensive experiments on benchmarking datasets to evaluate the effectiveness of the proposed hidden feature augmentation-based semi-supervised learning method. We firstly introduce the datasets used in the paper.

A. Data Preparation

The proposed approach is evaluated using 15 benchmarking datasets, which are available from the UCI repository (http://www.ics.uci.edu/~mlearn/MLRepository.html), the KEEL-dataset repository (http://sci2s.ugr.es/keel/datasets), and the Max Planck Institute for Biological Cybernetics (http://www.kyb.tuebingen.mpg.de/ssl-book). Table 1 summarizes the characteristics of the datasets.

TABLE 1. ATTRIBUTES OF THE BENCHMARKING DATASETS.

ID	Data	#Instances	#Features
1	Sonar	208	60
2	Australian	690	14
3	Ionosphere	351	34
4	Breast	699	10
5	Monk2	601	6

6	German_org	1000	24
7	Vehicle	846	18
8	Wine	174	13
9	Diabetes	768	8
10	Heart	270	13
11	WDBC	569	14
12	Clean1	476	166
13	Spectfheart	267	44
14	Bci	400	117
15	Class1	163	9

B. Experimental Setup

A series of experiments are conducted to compare the performance of the proposed approach with that of several off-the-shelf semi-supervised classification methods. The 8 methods involved in the experiments are listed below.

1) Baseline Methods

(1) Support Vector Machines (SVM) [29].
(2) Least Square Support Vector Machines (LS-SVM) [20].
(3) Laplacian Regularized Least Squares (LapRLS) [3].
(4) Laplacian Support Vector Machines (LapSVM) [3].
(5) Safe Semi-Supervised Support Vector Machines (S4VM) [19].
(6) Safety-Aware Semi-Supervised Classification Method Based on Class Memberships (SA-SSCCM) [18].
(7) The proposed Semi-Supervised Learning Method Based on the Augmented Hidden Features (SSLAHF) using the hinge loss function (SSLAHF_h).
(8) The proposed SSLAHF method using the square loss function (SSLAHF_s).

2) Implementation Details

To satisfy the requirement of the transductive learning method S4VM and thus ensuring the comparison is made on equal ground, all the methods are implemented by following the procedures in transductive learning, i.e., learning is performed on both the labeled and unlabeled data, whereas performance is predicted on given unlabeled data. For each dataset, the experiments are configured in two ways: (1) 10 instances are randomly selected and labeled; and (2) 10% instances are randomly selected and serve as labeled instances. The remaining data are used as unlabeled instances. In order to evaluate the effectiveness of the proposed SSLAHF in the control of the influence of label error, 10% and 20% of the labeled data are selected and tagged with wrong labels purposely to generate label noise for the second configuration.

The parameters of the methods are set as follows. Following the setting in [18], both linear and radial basis function (RBF) kernels are used for all the datasets. The width of the RBF kernel σ is set to be the average distance between all the instances. The regularization parameter C_1 is set by searching the value from the set $\{1e\text{-}5,1e\text{-}4,1e\text{-}3,1e\text{-}2,1e\text{-}1,1e0,1e1,1e2,1e3,1e4,1e5\}$ through the leave-one-out strategy. The regularization parameter C_2 of S4VM is used to weight the loss of unlabeled instance, which is fixed to 0.1 [19]. The number of times of sampling in each trial is set to 100. For SA-SSCCM, λ_1 and λ_2 are set to 100 and 0.1 respectively, and ε is set to 10^{-3}. The parameter λ is selected using the leave-one-out strategy from the set $\{0.05,0.1,0.15,...,0.5\}$. For LapSVM and LapRLS, the number of nearest neighbors for the Laplacian is selected from $\{5,10,15,20\}$. The extrinsic regularization parameter γ_A and the intrinsic regularization parameter γ_I are set to 1e-6 and 1e-2 respectively. The dimension of the common hidden space r is selected from $r \in \{2,3,4,5,6\}$ when the dimension of original features d is less than 60; and from $r \in \{8,9,10,11,12\}$ otherwise. All the experiments are conducted using MATLAB on a computer with an Intel Core i3-3240 3.4GHz CPU and 4GB RAM.

C. Performance Comparison

The average accuracy and standard deviation (SD) of the 8 methods with experiments conducted using 10 labeled instances is shown in Tables 2 and 3 respectively, where linear kernel is adopted for Table 2 and RBF kernel for Table 3. Each row in the tables gives the average and SD of classification accuracy of the individual methods for a dataset. The values in boldface represent the best performance obtained. The last row in the tables shows the average accuracy of each individual method over all the 15 datasets. In the same way, the average performance and SD with 10% instances are labeled is shown in Table 4 (linear kernel) and Table 5 (RBF kernel) respectively.

From the results in the four tables above, the following observations can be made.

TABLE 2. COMPARISON OF ACCURACY (MEAN ± SD) WITH 10 LABELED INSTANCES USING LINEAR KERNEL.

Linear / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.5823 ± 0.0059	0.5919 ± 0.0051	0.6081 ± 0.0055	0.6167 ± 0.0063	0.6020 ± 0.0061	0.6077 ± 0.0088	0.6212 ± 0.0047	**0.6283 ± 0.0079**
Australian	0.7248 ± 0.0177	0.6994 ± 0.0163	0.7375 ± 0.0066	0.6500 ± 0.0163	**0.7475 ± 0.0142**	0.6896 ± 0.0157	0.7224 ± 0.0176	0.7042 ± 0.0166
Ionosphere	0.6372 ± 0.0080	0.7032 ± 0.0045	0.7114 ± 0.0072	0.7196 ± 0.0043	0.6950 ± 0.0076	0.6902 ± 0.0034	0.7264 ± 0.0058	**0.7305 ± 0.0055**
Breast	0.8322 ± 0.0873	0.8319 ± 0.0872	0.7932 ± 0.0063	0.8186 ± 0.0049	0.8412 ± 0.0078	0.8350 ± 0.0064	**0.9087 ± 0.0049**	0.9041 ± 0.0016
Monk2	0.5347 ± 0.0035	0.5866 ± 0.0143	0.5345 ± 0.0035	0.5514 ± 0.0034	0.5549 ± 0.0053	**0.6253 ± 0.0104**	0.5899 ± 0.0078	0.5931 ± 0.0019

German_org	0.6787± 0.0014	0.6330± 0.0048	0.6146± 0.0045	0.5978± 0.0027	0.6193± 0.0039	0.6469± 0.0158	**0.6889± 0.0010**	0.6429± 0.0060
Vehicle	0.6271± 0.0007	0.6254± 0.0005	0.6645± 0.0060	0.6317± 0.0036	0.6707± 0.0053	0.6488± 0.0039	**0.7268± 0.0007**	0.7251± 0.0005
Wine	0.8816± 0.0033	0.8823± 0.0031	0.8816± 0.0033	0.7804± 0.0069	0.8829± 0.0035	0.8404± 0.0083	**0.8968± 0.0014**	0.8856± 0.0080
Diabetes	0.6390± 0.0022	0.6313± 0.0057	0.6298± 0.0060	0.5701± 0.0041	0.6363± 0.0018	0.6276± 0.0092	0.6533± 0.0033	**0.6537± 0.0030**
Heart	0.6127± 0.0022	0.6146± 0.0019	0.6154± 0.0019	0.6269± 0.0062	0.6146± 0.0028	0.6073± 0.0028	0.6127± 0.0017	**0.6619± 0.0078**
WDBC	**0.8524± 0.0088**	0.8503± 0.0084	0.8394± 0.0065	0.7372± 0.0151	0.8517± 0.0054	0.8270± 0.0058	0.8376± 0.0094	0.8363± 0.0090
Clean1	0.5794± 0.0032	0.5803± 0.0026	0.5796± 0.0031	0.5803± 0.0027	**0.5839± 0.0028**	0.5088± 0.0028	0.5794± 0.0033	0.5755± 0.0027
Spectfheart	0.7709± 0.0056	0.7700± 0.0060	0.7190± 0.0056	0.7206± 0.0075	0.7291± 0.0038	**0.7947± 0.0001**	0.7709± 0.0056	0.7714± 0.0058
Bci	0.5318± 0.0006	0.5318± 0.0007	0.5318± 0.0006	0.5328± 0.0005	0.5315± 0.0006	0.5105± 0.0003	**0.5356± 0.0001**	0.5319± 0.0007
Class1	0.6259± 0.0074	0.6196± 0.0060	0.6322± 0.0057	0.5902± 0.0014	0.6203± 0.0062	0.5944± 0.0021	0.6266± 0.0055	**0.6343± 0.0013**
Average Accuracy	0.6740	0.6768	0.6729	0.6483	0.6787	0.6702	0.6998	0.6986

TABLE 3. COMPARISON OF ACCURACY (MEAN±SD) WITH 10 LABELED INSTANCES USING RBF KERNEL.

RBF / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.6071± 0.0054	0.6072± 0.0056	0.6146± 0.0052	0.6167± 0.0053	0.6000± 0.0058	0.6116± 0.0047	0.6136± 0.0056	**0.6197± 0.0060**
Australian	0.7345± 0.0163	0.7281± 0.0156	0.7463± 0.0069	**0.7510± 0.0066**	0.7212± 0.0139	0.7287± 0.0159	0.7385± 0.0130	0.7330± 0.0129
Ionosphere	0.6727± 0.0040	0.7220± 0.0058	0.7135± 0.0052	0.7264± 0.0048	0.7170± 0.0058	0.6440± 0.0001	0.7132± 0.0055	**0.7279± 0.0044**
Breast	0.6192± 0.0030	0.6289± 0.0031	0.5691± 0.0042	0.5557± 0.0054	0.6060± 0.0102	0.6550± 0.0000	**0.7354± 0.0134**	0.7041± 0.0110
Monk2	0.5570± 0.0035	0.5562± 0.0036	0.5570± 0.0035	0.5455± 0.0050	0.5805± 0.0064	**0.6570± 0.0000**	0.5555± 0.0034	0.5514± 0.0046
German_org	0.6706± 0.0023	0.6727± 0.0024	0.6082± 0.0030	0.6099± 0.0032	0.6706± 0.0033	0.6733± 0.0024	0.6678± 0.0033	**0.6801± 0.0040**
Vehicle	0.7011± 0.0012	0.6769± 0.0032	0.6778± 0.0030	0.6602± 0.0041	**0.7408± 0.0000**	0.6890± 0.0091	0.7010± 0.0013	0.6849± 0.0034
Wine	0.8582± 0.0105	0.8582± 0.0105	0.8582± 0.0105	0.8620± 0.0085	0.7823± 0.0004	0.7791± 0.0365	0.8608± 0.0096	**0.8671± 0.0084**
Diabetes	0.6376± 0.0037	**0.6472± 0.0035**	0.6161± 0.0029	0.6061± 0.0030	0.6088± 0.0078	0.5681± 0.0183	0.6372± 0.0041	0.6442± 0.0035
Heart	0.5827± 0.0020	0.5823± 0.0022	0.5827± 0.0020	0.5527± 0.0037	0.5046± 0.0024	**0.5838± 0.0017**	0.5835± 0.0022	0.5554± 0.0033
WDBC	0.8372± 0.0088	0.8492± 0.0087	0.8401± 0.0075	0.8372± 0.0076	**0.8556± 0.0060**	0.8490± 0.0088	0.8254± 0.0085	0.8317± 0.0097
Clean1	0.5813± 0.0023	0.5811± 0.0034	0.5811± 0.0023	0.5732± 0.0022	0.5504± 0.0049	0.5818± 0.0033	**0.5872± 0.0022**	0.5768± 0.0033
Spectfheart	0.7789± 0.0024	0.7789± 0.0024	0.7279± 0.0060	0.7267± 0.0053	0.7397± 0.0258	0.7822± 0.0014	0.7806± 0.0020	**0.7834± 0.0019**
Bci	0.5310± 0.0006	0.5231± 0.0009	0.5323± 0.0005	0.5285± 0.0010	0.5154± 0.0001	0.5228± 0.0009	**0.5372± 0.0008**	0.5285± 0.0008
Class1	0.6039± 0.0103	0.6137± 0.0090	0.6007± 0.0089	0.6007± 0.0119	**0.6288± 0.0086**	0.5431± 0.0047	0.6170± 0.0105	0.6183± 0.0085
Average Accuracy	0.6649	0.6684	0.6550	0.6502	0.6548	0.6579	0.6769	0.6738

TABLE 4. COMPARISON OF ACCURACY (MEAN±SD) WITH 10% INSTANCES LABELED USING LINEAR KERNEL.

Linear / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.6420± 0.0033	0.6457± 0.0026	0.6654± 0.0019	0.6186± 0.0031	0.6527± 0.0037	0.6202± 0.0061	**0.6686± 0.0016**	0.6574± 0.0017
Australian	0.8444± 0.0001	0.8472± 0.0002	0.7971± 0.0010	0.8304± 0.0003	0.8496± 0.0001	0.8045± 0.0026	**0.8562± 0.0001**	0.8551± 0.0001

Ionosphere	0.7446± 0.0068	0.7965± 0.0020	0.7797± 0.0016	0.6921± 0.0031	0.7842± 0.0026	0.7383± 0.0093	0.7978± 0.0017	**0.7997± 0.0015**
Breast	0.9554± 0.0004	0.9483± 0.0002	0.9278± 0.0076	0.9141± 0.0013	0.9440± 0.0005	0.9433± 0.0044	**0.9597± 0.0006**	0.9494± 0.0002
Monk2	0.5447± 0.0011	0.6148± 0.0085	0.5560± 0.0010	0.5528± 0.0024	0.5569± 0.0032	0.6260± 0.0104	0.6129± 0.0050	**0.6219± 0.0047**
German_org	0.6919± 0.0005	0.7131± 0.0002	0.6449± 0.0010	0.7132± 0.0003	0.7063± 0.0004	0.6949± 0.0006	0.6921± 0.0006	**0.7290± 0.0002**
Vehicle	0.7486± 0.0000	0.7491± 0.0005	0.7223± 0.0023	0.7663± 0.0002	0.7459± 0.0007	0.7484± 0.0000	0.7486± 0.0000	**0.7631± 0.0002**
Wine	0.9130± 0.0002	0.9168± 0.0004	0.9112± 0.0001	0.8453± 0.0051	0.9019± 0.0006	0.8268± 0.0090	0.9112± 0.0001	**0.9335± 0.0009**
Diabetes	0.7200± 0.0008	**0.7467± 0.0002**	0.6601± 0.0017	0.6540± 0.0012	0.7460± 0.0003	0.6291± 0.0031	0.7314± 0.0004	0.7457± 0.0002
Heart	**0.7683± 0.0039**	0.7621± 0.0022	0.7449± 0.0039	0.7765± 0.0021	0.7658± 0.0020	0.7033± 0.0043	0.7605± 0.0020	**0.7683± 0.0022**
WDBC	0.9383± 0.0006	**0.9396± 0.0002**	0.9162± 0.0006	0.9250± 0.0003	0.9230± 0.0003	0.9191± 0.0004	0.9227± 0.0003	0.9287± 0.0002
Clean1	0.6492± 0.0011	**0.6528± 0.0009**	0.6506± 0.0011	0.6470± 0.0020	**0.6528± 0.0012**	0.6388± 0.0032	0.6434± 0.0013	0.6474± 0.0015
Spectfheart	0.7983± 0.0002	0.7959± 0.0001	0.7415± 0.0018	0.6921± 0.0039	0.7448± 0.0015	0.7959± 0.0001	**0.7985± 0.0002**	0.7959± 0.0001
Bci	0.5956± 0.0049	0.5953± 0.0050	0.5956± 0.0049	0.5992± 0.0045	0.5969± 0.0051	0.5942± 0.0052	**0.6048± 0.0039**	0.5958± 0.0044
Class1	0.6606± 0.0021	0.6496± 0.0019	0.6401± 0.0043	0.6343± 0.0015	0.6438± 0.0061	0.6248± 0.0099	**0.6708± 0.0022**	0.6321± 0.0009
Average Accuracy	0.7477	0.7582	0.7302	0.7241	0.7476	0.7272	0.7586	0.7609

TABLE 5. COMPARISON OF ACCURACY (MEAN±SD) WITH 10% INSTANCES LABELED USING RBF KERNEL.

RBF / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.6701± 0.0015	0.6679± 0.0020	0.6684± 0.0016	0.6674± 0.0029	0.6529± 0.0029	0.6733± 0.0018	**0.6829± 0.0018**	0.6797± 0.0022
Australian	0.8533± 0.0001	0.8195± 0.0097	0.7770± 0.0008	0.7225± 0.0041	0.8417± 0.0001	0.8541± 0.0001	0.8557± 0.0001	**0.8576± 0.0000**
Ionosphere	0.8278± 0.0033	0.8434± 0.0021	0.8263± 0.0016	0.7892± 0.0016	0.8101± 0.0071	0.8436± 0.0064	0.8437± 0.0025	**0.8440± 0.0026**
Breast	0.6477± 0.0007	0.6566± 0.0000	0.6267± 0.0019	0.6234± 0.0018	0.6485± 0.0000	0.6566± 0.0000	**0.7017± 0.0083**	0.6566± 0.0000
Monk2	0.6294± 0.0021	0.6468± 0.0009	0.6451± 0.0007	0.6231± 0.0009	0.6418± 0.0007	**0.6580± 0.0004**	0.6344± 0.0010	0.6348± 0.0006
German_org	0.6754± 0.0012	0.6951± 0.0006	0.6652± 0.0014	0.6807± 0.0020	0.6787± 0.0006	0.7013± 0.0000	0.6783± 0.0009	**0.7037± 0.0003**
Vehicle	0.7094± 0.0043	0.7461± 0.0003	0.7377± 0.0003	0.7404± 0.0012	0.7415± 0.0000	0.7450± 0.0008	0.7301± 0.0006	**0.7516± 0.0004**
Wine	0.8969± 0.0013	0.8988± 0.0012	0.8969± 0.0013	0.8876± 0.0008	0.7056± 0.0011	0.7422± 0.0110	0.8950± 0.0017	**0.9186± 0.0004**
Diabetes	0.7107± 0.0006	0.7163± 0.0003	0.6495± 0.0008	0.6500± 0.0005	0.6457± 0.0000	0.6375± 0.0017	0.7161± 0.0003	**0.7233± 0.0003**
Heart	0.6414± 0.0059	0.6243± 0.0058	0.6193± 0.0049	**0.6432± 0.0047**	0.5243± 0.0021	0.6226± 0.0054	0.6160± 0.0053	0.6416± 0.0059
WDBC	0.9342± 0.0002	0.9396± 0.0003	0.9211± 0.0004	0.8824± 0.0007	0.9301± 0.0003	**0.9402± 0.0003**	0.9303± 0.0002	0.9348± 0.0003
Clean1	**0.7324± 0.0012**	0.7284± 0.0008	0.7235± 0.0007	0.7091± 0.0009	0.5065± 0.0007	0.7289± 0.0008	0.7310± 0.0013	0.7263± 0.0011
Spectfheart	**0.7959± 0.0013**	**0.7959± 0.0001**	0.7452± 0.0013	0.7266± 0.0016	0.7905± 0.0001	0.7905± 0.0001	**0.7959± 0.0001**	**0.7959± 0.0001**
Bci	0.5867± 0.0030	0.5764± 0.0023	0.5986± 0.0037	0.6011± 0.0038	0.5106± 0.0012	0.5753± 0.0023	**0.6083± 0.0029**	0.5953± 0.0029
Class1	0.6429± 0.0037	0.6299± 0.0051	0.6510± 0.0063	0.6490± 0.0056	0.6578± 0.0084	0.5510± 0.0068	0.6531± 0.0056	**0.6605± 0.0059**
Average Accuracy	0.7303	0.7323	0.7168	0.7037	0.6858	0.7147	0.7382	0.7416

(1) It can be seen from the results obtained using different benchmarking datasets that not all assumptions can be precisely met for the unlabeled samples, which adversely affect the training process and lead to unsatisfactory classification performance. Obviously, the classification performance of SSLAHF_h and SSLAHF_s, as shown in Tables 2 to 5, are better than that of the other six algorithms, demonstrating the potential of the proposed feature augmentation method. Instead of making assumptions, the SSLAHF successfully leverages the hidden features in handling the benchmarking datasets to establish correspondence between the labeled and unlabeled samples, thereby obtaining improved classification performance.

(2) The overall performance of semi-supervised SVM and LS-SVM are comparable, regardless of whether 10 samples or 10% samples are labeled. In most cases, the performance of SSLAHF is better than SVM and LS-SVM which rely on cluster assumptions or manifold assumptions. Overall, the performance of all the algorithms on 10% labeled samples are better than that on 10 labeled samples due to the increasing number of real labeled samples. However, for those based on assumptions lose their advantages when the assumptions cannot capture the real attributions of the data. This is exactly the reason why a secure semi-supervised learning method is needed, and developed here using the actual relationship of hidden features between the labeled and unlabeled samples rather than relying on assumption.

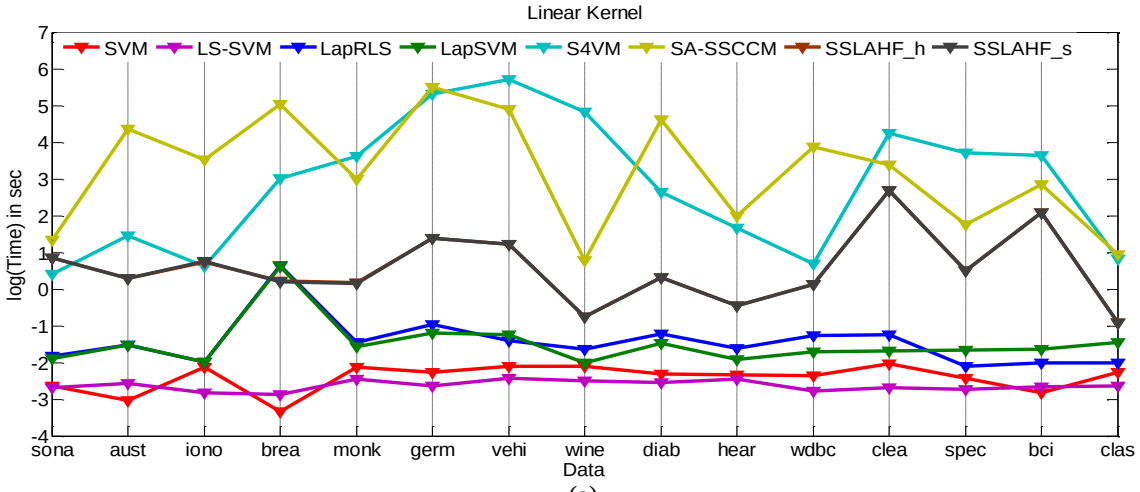
(3) Refer to the classification accuracy obtained using 10 labeled samples in Table 2 and Table 3, SSLAHF significantly outperforms all the other methods in 10 of the 15 datasets for linear kernel, and 8 of the 15 for RBF kernel. More importantly, unlike the assumption-based algorithms, out of the 15 datasets, performance degradation is only observed in 2 and 4 datasets respectively for SSLAHF_h and SSLAHF_s when linear kernel is adopted, whereas performance degradation occurs in 6 datasets for S4VM and LapRLS for 6 of the 15 datasets, 9 for SA-SSCCM, and 8 for LapSVM. Similar situation is also observed when RBF kernel is used. On the other hand, the classification performance evaluated using 10% labeled samples (Tables 4 and 5) is similar to that evaluated using 10 labeled samples given in (Tables 2 and 3). This is apparently because the assumptions made fail in the same way for both cases.

(4) Although some of the assumption-based semi-supervised learning methods significantly outperform the traditional classification SVM methods for most datasets, when the average classification accuracy over all the 15 datasets is concerned, their performance is indeed worse than that of the traditional methods. This is because when the assumptions made do not agree with the real data distribution, the performance on the unlabeled samples will degrade markedly. For the proposed SSLAHF, while the performance does not outperform SVM for some datasets, the average accuracy is better than the other methods as it is more secure to leverage the hidden features which effectively avoid the propagation of erroneous information created by the inappropriate assumptions.

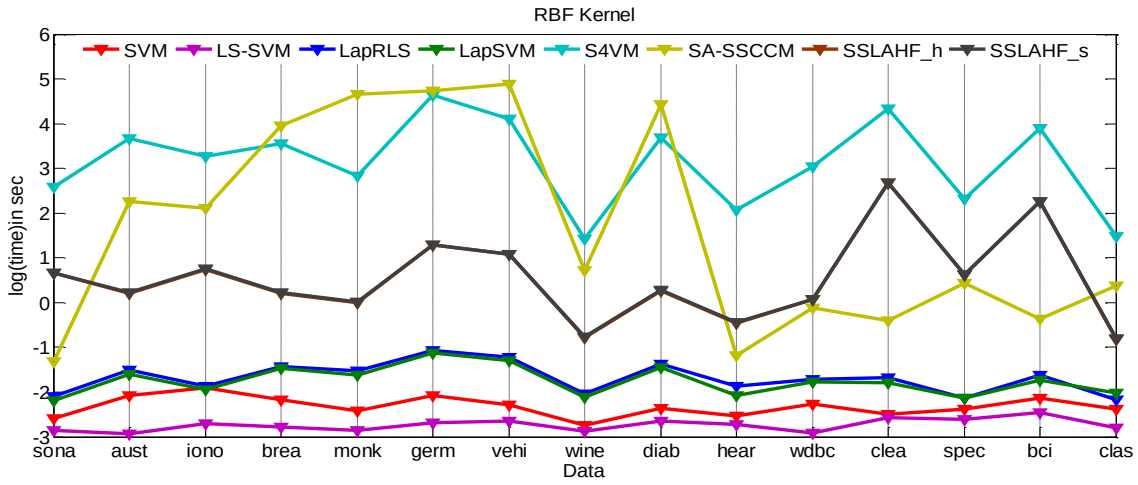
Besides, the average running time, i.e. the total training and testing time, of the 8 methods on the 15 datasets using 10 labeled samples is shown in Fig. 3, with Fig. 3(a) showing the timing performance when linear kernel is adopted and Fig. 3(b) for RBF kernel. Obviously, the traditional semi-supervised learning methods have the shortest running time. The running time of the proposed SSLAHF is a little longer than that of traditional methods and the algorithms based on manifold assumptions, but much shorter than that of S4VM and SA-SSCCM.

D. Effect of Wrongly Tagged Data

Most existing semi-supervised classification methods are implemented by including unlabeled samples into the training sets through an automated labeling process, based on manifold or cluster preserving. If some of the originally unlabeled samples are incorrectly labeled, errors would prorogated by the automated labeling process and could seriously affect model training. The proposed SSLAHF can effectively alleviate the



(a)



(b)

Fig. 3. The total running time (in seconds) of eight algorithms on fifteen datasets with linear kernel and RBF.

negative effect of labeling error because it only makes use of the feature information of the labeled and unlabeled samples. To demonstrate the advantage, experiment is conducted with 10% labeled samples, together with 10% and 20% wrongly tagged data respectively. RBF kernel is used in the experiment. The results are shown in Table 6 and Table 7, where 10% and 20% of the samples are wrongly tagged respectively. The classification accuracy and SD for each dataset using the 8 methods are given in each row. The values in boldface represent the best results for a given dataset. The last row in Table 6 and Table 7 show the average accuracy of each individual method over all the 15 datasets.

Several interesting observations can be obtained from the results in Table 6 and Table 7.

(1) The performance of both SSLAHF_h and SSLAHF_s only degrades slightly in 1 of the 15 datasets when 10% of the samples are wrongly tagged, whereas the performance of SSLAHF_h degrades slightly in 1 dataset when the proportion of wrongly tagged data increased to 20%. Besides, SSLAHF_s always outperforms to the inductive SVM methods. The promising result of SSLAHF is attributed to the use of the hidden features between the labeled and unlabeled samples, without the need of making assumptions that may be inappropriate.

(2) The average classification accuracy of SSLAHF is better than that of the other semi-supervised learning methods, whether the proportion of wrongly labeled samples is 10% or 20%. The result demonstrates that SSLAHF is able to establish, to some extent, connections among samples belonging to the same classes in the labeled and unlabeled datasets. It also exposes the major limitation of assumption-based algorithms that erroneous information is prorogated and amplified when

TABLE 6. CLASSIFICATION ACCURACY (MEAN ± SD) AT 10% INSTANCES WRONGLY LABELED.

10% / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.6364±0.0014	0.6406±0.0013	0.6492±0.0031	0.6080±0.0057	0.6021±0.0030	0.6503±0.0026	0.6481±0.0021	**0.6556±0.0016**
Australian	0.8238±0.0024	0.8074±0.0085	0.7481±0.0050	0.7153±0.0013	0.8386±0.0002	0.8385±0.0002	**0.8457±0.0001**	0.8332±0.0004
Ionosphere	0.8016±0.0037	0.7918±0.0021	0.6693±0.0032	0.5883±0.0052	0.6592±0.0139	0.6463±0.0085	**0.8149±0.0022**	0.7959±0.0026
Breast	0.6445±0.0013	0.6566±0.0001	0.6067±0.0084	0.5089±0.0014	0.6467±0.0001	0.6434±0.0001	**0.6952±0.0014**	0.6566±0.0001
Monk2	0.5440±0.0028	0.5434±0.0028	0.5524±0.0012	0.5503±0.0030	0.5574±0.0032	**0.6588±0.0014**	0.5550±0.0030	0.5584±0.0019
German_org	0.6643±0.0017	0.6860±0.0006	0.5863±0.0073	0.6756±0.0014	0.6726±0.0003	0.6846±0.0006	0.6708±0.0012	**0.7004±0.0004**
Vehicle	0.7005±0.0010	0.7188±0.0006	0.7100±0.0048	**0.7365±0.0007**	0.7287±0.0003	0.7320±0.0004	0.7205±0.0005	0.7344±0.0008
Wine	0.8112±0.0031	0.8174±0.0031	0.7752±0.0055	0.7354±0.0077	0.6739±0.0013	0.6901±0.0044	0.8534±0.0028	**0.8596±0.0023**
Diabetes	0.7046±0.0012	0.7082±0.0010	0.5974±0.0091	0.6444±0.0017	0.6395±0.0016	0.6719±0.0078	0.6997±0.0010	**0.7097±0.0010**
Heart	0.5901±0.0087	0.5897±0.0075	0.6171±0.0078	**0.6407±0.0027**	0.5321±0.0029	0.6148±0.0036	0.5979±0.0046	0.6206±0.0041
WDBC	0.7938±0.0023	0.8115±0.0017	0.7879±0.0032	0.8578±0.0009	0.8473±0.0010	0.8645±0.0009	**0.8670±0.0006**	0.8619±0.0010
Clean1	0.6473±0.0017	0.6534±0.0038	0.6464±0.0023	0.6275±0.0025	0.5406±0.0019	0.6534±0.0040	0.6974±0.0005	**0.7023±0.0008**

Spectfheart	0.7660± 0.0016	0.7705± 0.0014	0.6266± 0.0047	0.5647± 0.0050	0.7892± 0.0001	0.7739± 0.0008	**0.7954± 0.0001**	**0.7954± 0.0001**
Bci	0.5661± 0.0023	0.5594± 0.0024	0.5731± 0.0052	0.5633± 0.0033	0.5100± 0.0003	0.5592± 0.0024	**0.5750± 0.0032**	**0.5750± 0.0035**
Class1	0.6143± 0.0036	0.6204± 0.0066	0.5844± 0.0047	0.6156± 0.0022	0.6361± 0.0101	0.5282± 0.0031	0.6463± 0.0074	**0.6503± 0.0072**
Average Accuracy	0.6872	0.6917	0.6487	0.6422	0.6583	0.6807	0.7122	0.7140

TABLE 7. CLASSIFICATION ACCURACY (MEAN ± SD) AT 20% INSTANCES WRONGLY LABELED.

20% / Dataset	SVM	LS-SVM	LapRLS	LapSVM	S4VM	SA-SSCCM	SSLAHF_h	SSLAHF_s
Sonar	0.6316± 00030	0.6332± 0.0028	0.6374± 0.0050	0.5947± 0.0023	0.5679± 0.0023	0.6289± 0.0035	0.6385± 0.0035	**0.6422± 0.0042**
Australian	0.8110± 0.0053	0.7596± 0.0143	0.7060± 0.0092	0.7659± 0.0021	0.8211± 0.0008	0.8225± 0.0006	**0.8425± 0.0003**	0.8247± 0.0006
Ionosphere	0.7538± 0.0029	0.7633± 0.0019	0.6247± 0.0033	0.5532± 0.0016	0.6013± 0.0176	0.6389± 0.0029	**0.7649± 0.0029**	0.7647± 0.0022
Breast	0.6197± 0.0059	0.6248± 0.0059	0.4994± 0.0217	0.5003± 0.0005	0.5890± 0.0126	0.5434± 0.0046	**0.6677± 0.0167**	0.6566± 0.0115
Monk2	0.5414± 0.0045	0.5419± 0.0044	0.5505± 0.0046	0.5243± 0.0035	0.5507± 0.0041	**0.6588± 0.0014**	0.5429± 0.0034	0.5414± 0.0030
German_org	0.6584± 0.0012	0.6798± 0.0006	0.5624± 0.0073	0.6599± 0.0011	0.6539± 0.0004	0.6781± 0.0006	0.6590± 0.0021	**0.6941± 0.0004**
Vehicle	0.6825± 0.0053	0.6752± 0.0007	0.5913± 0.0205	0.7110± 0.0009	0.6904± 0.0011	**0.7316± 0.0004**	0.6869± 0.0041	0.7029± 0.0017
Wine	0.7981± 0.0021	0.7938± 0.0035	0.7447± 0.0085	0.7012± 0.0115	0.6547± 0.0017	0.6720± 0.0355	0.8099± 0.0048	**0.8261± 0.0033**
Diabetes	0.6793± 0.0012	0.6945± 0.0015	0.5509± 0.0117	0.6350± 0.0024	0.6295± 0.0004	0.6551± 0.0000	0.6647± 0.0021	**0.6958± 0.0014**
Heart	0.5593± 0.0063	0.5613± 0.0050	0.6113± 0.0032	**0.6360± 0.0045**	0.5214± 0.0024	0.5942± 0.0044	0.5914± 0.0034	0.6033± 0.0061
WDBC	0.7242± 0.0014	0.7404± 0.0020	0.7422± 0.0033	0.7852± 0.0024	0.7916± 0.0023	0.8086± 0.0024	**0.8117± 0.0013**	0.8010± 0.0022
Clean1	0.6238± 0.0032	0.6490± 0.0046	0.6096± 0.0025	0.5888± 0.0024	0.5378± 0.0016	0.6487± 0.0050	0.6683± 0.0012	**0.6767± 0.0014**
Spectfheart	0.7564± 0.0022	0.7672± 0.0014	0.5963± 0.0040	0.5510± 0.0025	0.7851± 0.0002	0.7676± 0.0012	**0.7903± 0.0001**	**0.7903± 0.0001**
Bci	0.5428± 0.0017	0.5439± 0.0013	0.5536± 0.0026	0.5564± 0.0023	0.4942± 0.0008	0.5406± 0.0012	**0.5622± 0.0025**	0.5503± 0.0014
Class1	0.5966± 00128	0.6088± 0.0078	0.5527± 0.0043	0.5715± 0.0061	0.5762± 0.0053	0.5154± 0.0081	0.6020± 0.0132	**0.6102± 0.0099**
Average Accuracy	0.6653	0.6691	0.6088	0.6223	0.6310	0.6603	0.6869	0.6920

the labeling errors due to invalid assumptions exists.

V. CONCLUSIONS

Existing semi-supervised learning methods requires making assumptions on the intrinsic pattern of the samples. Inappropriate assumptions can lead to wrongly labeled data and the ensuing propagation of the erroneous information, which eventual degrades the model performance. In this regard, a novel semi-supervised learning method, SSLAHF, exploiting hidden features is proposed in this paper. Without the need of assumption making, the method effectively reduces the risk of wrongly labeled samples. Extensive experiments have been conducted with benchmarking datasets to demonstrate the effectiveness of SSLAHF in classification problems. Experiments are also conducted to demonstrate that the performance of SSLAHF is not affected by the presence of wrongly tagged samples.

Although the proposed SSLAHF shows encouraging classification performance in semi-supervised learning problems, there are still issues that deserve further investigation. For example, in SSLAHF, it is necessary to identify the new feature representation underlying the labeled and unlabeled samples before the projection matrix can be obtained. The accuracy of the projection matrix is affected by the relevance of the feature identified. Besides, the computation time of SSLAHF is relatively long. Research will be conducted to reduce the computation complexity.

REFERENCES

[1] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-Supervised Learning," *Cambridge, MA, USA: MIT Press*, 2006.

[2] L. Chen, I. W. Tsang, D. Xu. "Laplacian embedded regression for scalable manifold regularization," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 23, no. 6, pp. 902–915, 2012.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,"*J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 2399–2434, 2006.

[4] X. Q. Zhu and X.D. Wu, "Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts," *Artificial Intelligence Review*, vol. 22, no.3, pp. 177–210, 2004.

[5] X. Zhu, "Semi-supervised learning literature survey," *Ph.D. dissertation, Dept. Comput. Sci., Wisconsin-Madison, Univ., Madison, WI, USA*, Jul. 2008.

[6] M. Loog and A. C. Jensen, "Semi-Supervised Nearest Mean Classification Through a Constrained Log-Likelihood," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 26, no. 5, pp. 995–1006, 2015.

[7] Z. H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowl. Inf. Syst.*, vol. 24, no. 3, pp. 415–439, 2010.

[8] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "Semi-boost: Boosting for semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2000–2014, Nov. 2009.

[9] Y. F. Li, J. Kwok, and Z. H. Zhou, "Semi-supervised learning using label mean," *In Proc. 26th Int. Conf. Mach. Learn.*, pp. 633–640, 2009.

[10] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Large scale transductive SVMs," *J. Mach. Learn. Res.*, vol. 7, pp. 1687–1712, Jan. 2006.

[11] G. Fung and O. L. Mangasarian, "Semi-supervised support vector machine for unlabeled data classification," *Optim. Methods Softw.*, vol. 15, no. 1, pp. 99–105, 2001.

[12] T. Joachims, "Transductive inference for text classification using support vector machines," *In Proc. 16th Int. Conf. Mach. Learn.*, 1999, pp. 200–209.

[13] Y. Bengio, O. B. Alleau, and N. L. Roux, "Label propagation and quadratic criterion, in Semi-Supervised Learning," *Cambridge, MA, USA: MIT Press*, 2006, pp. 193–216.

[14] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *Dept. Lang.Technol., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CALD-02-107*, 2002.

[15] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," *in Proc. 18th Int. Conf. Mach. Learn.*, pp. 19–26, 2001.

[16] T. Yang and C. E. Priebe, "The effect of model misspecification on semi-supervised classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2093–2103, Oct. 2011.

[17] Y. Wang, S. Chen, and Z. H. Zhou, "New semi-supervised classification method based on modified cluster assumption," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 23, no. 5, pp. 689–702, May 2012.

[18] Y. Y. Wang and S. C. Chen, "Safety-aware semi-supervised classification," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 24, no. 11, pp. 1763–1772, November 2013.

[19] Y. Li, Z. Zhou, "Towards making unlabeled data never hurt," *In Proceedings of the 28th International Conference on Machine Learning, Omnipress*, pp. 1081–1088, 2011.

[20] J. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vanderwalle, "Least Squares Support Vector Machines," *World Scientific*, 2002.

[21] J. R. Anderson, "Cognitive Psychology and Its Applications (seventh edition)," *New York: Freeman*, 2010.

[22] H. Daumé, III, "Frustratingly easy domain adaptation," *in Proc. ACL*, 2007.

[23] E. Parzen, "On Estimation of a Probability Density Function and Mode, "*Annals of Math. Statistics*, vol. 33, pp. 1065-1076, Sept. 1962.

[24] Z. H. Deng, Fu-Lai Chung, S. T. Wang, "FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation," *Pattern Recognition*, vo. 41, no. 4, pp.1363-1372, 2008.

[25] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthonormality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[26] Y.Z. Jiang, Fu-Lai Chung, S.T. Wang, et al. "Collaborative Fuzzy Clustering From Multiple Weighted Views," *IEEE Transactions on Cyb.*, vol.45, no.4, pp.688-701, 2015.

[27] N. Del Buono and T. Politi, "A continuous technique for the weighted low-rank approximation problem," *In Proc. Int. Conf. Comput. Sci. Appl., Assisi, Italy*, pp. 988–997, 2004.

[28] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, 2004.

[29] V. Vapnik, "The Nature of Statistical Learning," *Springer-Verlag*, 1995.

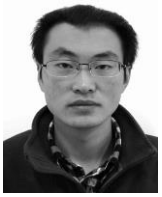
[30] K. Chen and S. Wang. "Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.33, no.1, pp.129–143, 2011.

[31] F. G. Cozman, I. Cohen, and M. C. Cirelo. "Semi-supervised learning of mixture models," *In Proceedings of the 20th International Conference on Machine Learning*, pp. 99–106, 2003.

[32] Ester M, Kriegel H P, Sander J, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*. vol.96, pp. 226-231, 1996.

[33] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Jul. 2010.

[34] S.T. Wang, J. Wang, Fu-Lai Chung, "Kernel density estimation, kernel methods, and fast learning in large data sets," *IEEE Transactions on Cyb.*, vol.44, no.1, pp. 1-20, 2014.



Wenlong Hang is a Ph.D. candidate at the School of Digital Media, Jiangnan University. Wuxi, China.

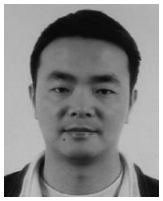
His research interests include pattern recognition, data mining and their applications.

Kup-Sze Choi received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong.

He is an Associate Professor with the School of Nursing, Hong Kong Polytechnic University, Hong Kong, and the Director of the Centre for Smart Health. His research interests include computational intelligence, data mining, computer graphics and virtual reality, and their applications in biomedicine and health care.

Shitong Wang received the M.S. degree in computer science from Nanjing University of Aeronautics and Astronautics, China, in 1987. He visited London University and Bristol University in U.K., Hiroshima International University and Osaka Prefecture University in Japan, Hong Kong University of Science and Technology, Hong Kong Polytechnic University, as a Research Scientist, for over six years.

Currently, he is a Full Professor of the School of Digital Media, Jiangnan University, China. His research interests include artificial intelligence, neuro-fuzzy systems, pattern recognition, and image processing. He has published about 100 papers in international/national journals and has authored seven books.



Pengjiang Qian received the B.S. degree in computer science and technology from Jiangnan University, Wuxi, China, in 2000, the M.S. degree in software engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in information technology and engineering from Jiangnan University, in 2011.

He is an Associate Professor with the School of Digital Media, Jiangnan University. He is currently with the Case Western Reserve University, Cleveland, OH, USA, as a Visiting Scholar and doing research in medical image processing. He has published nearly 30 papers in international/national journals and conferences. His current research interests include data mining, pattern recognition, bioinformatics, and their applications, such as medical image processing.