Pattern Recognition Letters 33 (2012) 580–588

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Logistic label propagation

Takumi Kobayashi *, Kenji Watanabe, Nobuyuki Otsu

National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba 305-8568, Japan

ARTICLE INFO

Article history:
Received 24 December 2010
Available online 24 December 2011
Communicated by L. Yin

Keywords:
Semi-supervised learning
Logistic function
Label propagation
Similarity
Gradient descent

ABSTRACT

In this paper, we propose a novel method for semi-supervised learning, called logistic label propagation (LLP). The proposed method employs the logistic function to classify input pattern vectors, similarly to logistic regression. To cope with unlabeled samples as well as labeled ones in the semi-supervised learning framework, the logistic functions are learnt by using similarities between samples in a manner similar to label propagation. In the proposed method, these two methods of logistic regression and label propagation are effectively incorporated in terms of posterior probabilities. LLP estimates the labels of input samples by using the learnt logistic function, whereas the method of label propagation has to optimize the whole labels whenever an input sample comes. In addition, we suggest the way to provide proper parameter setting and initialization, which frees the users from determining a parameter value in trial and error. In experiments on classification (estimating labels) in the semi-supervised learning framework, the proposed method exhibits favorable performances compared to the other methods.

1. Introduction

A fundamental procedure in pattern recognition is to classify pattern vectors. Various classification methods have been developed along with the advances in machine learning, such as SVM (Vapnik, 1998) and kernel-based methods (Schölkopf and Smola, 2001). Classifiers are generally learnt by using given training (labeled) samples. For example, in two-class (binary) classification, we have to prepare both positive and negative labeled samples. Although the performance of the classifier depends on the amount of such labeled samples, the task to label (annotate) samples by hand requires heavy human labor, making it difficult to prepare plenty of labeled samples in practical situations. In such cases, it is an effective approach to exploit both the labeled and unlabeled samples, which leads to semi-supervised and transductive learning, because we can easily collect unlabeled samples just by measuring the data without annotating it. The semi-supervised learning (Belkin and Niyogi, 2006; Cai et al., 2007) has attracted great deal of attentions over the last decade. We give brief reviews of the semi-supervised learning methods in Section 2.

The method of label propagation (LP) (Zhu et al., 2003) is frequently used in the framework of semi-supervised learning, such as for patch labeling (Bishop and Ulusoy, 2005), image matting (Levin et al., 2008; Grady et al., 2005), image annotation (Kang et al., 2006) and image classification (Cheng et al., 2009). The label propagation method estimates the label values based on the graph Laplacian (Belkin and Niyogi, 2003), i.e., similarities between samples, given a few labeled samples. The labels are retrieved as if the given labels propagate through the graph over the whole unlabeled samples. In this label propagation, the unlabeled samples are effectively incorporated by the graph Laplacian without imposing prior models on the sample distribution. A probabilistic interpretation can also be given to the method of label propagation (Grady et al., 2005), and from this probabilistic viewpoint, the estimated label values are regarded as posterior probabilities for the classes at each sample.

The method of logistic regression (LR) (Bishop, 2007) is one of the well-known supervised classification methods. The method produces promising performances, such as in biological (EEG) signals (Tomioka et al., 2007). In the logistic regression, the classifier based on the logistic function is learnt from labeled samples and estimates the class posterior probabilities of the input pattern vectors.

In this paper, we propose a novel method for semi-supervised learning, called logistic label propagation (LLP). The proposed method employs the logistic function to classify input pattern vectors as in logistic regression, and the classifier is optimized in the framework of semi-supervised learning as in label propagation using the graph Laplacian so as to cope with both labeled and unlabeled samples. The contributions of the proposed method are listed below:

- From the probabilistic viewpoint, both methods of label propagation and logistic regression are efficiently integrated.
- We utilize the graph Laplacian with the given labels in a discriminative manner as in label propagation, while the other semi-supervised methods such as Laplacian SVM (Belkin and

* Corresponding author. Tel.: +81 29 861 5491; fax: +81 29 861 3313.
E-mail addresses: takumi.kobayashi@aist.go.jp (T. Kobayashi), kenji.watana be@aist.go.jp (K. Watanabe), otsu.n@aist.go.jp (N. Otsu).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.12.005

Niyogi, 2006) and semi-supervised DA (Cai et al., 2007) introduce the graph Laplacian just for regularization.

- The proposed method provides the classifiers to estimate the labels of the newly input pattern vectors by using the learnt logistic functions at a quite low computational cost, unlike the label propagation method which requires exhaustive computation to deal with such newly input samples.
- The logistic classifiers are learnt (optimized) by using both the labeled and the unlabeled samples, which possibly avoids the over-fitting problem that has been addressed in the logistic regression.

The rest of this paper is organized as follows: the next section briefly reviews the related works, including the logistic regression and the label propagation. In Section 3, we describe the details of the proposed method, logistic label propagation (LLP), and the associated practical issues. Then, the experimental results on classification (label estimation) in the semi-supervised learning framework are shown in Section 4. Finally, Section 5 contains our concluding remarks.

The preliminary version of the proposed LLP has been published in (Watanabe et al., 2010). The method proposed in this paper is generalized from that in the points that we introduce the classification cost on the labeled samples together with a balancing parameter and propose some practically effective methods regarding to the balancing parameter and initialization.

2. Related works

In this section, we briefly review the methods that optimize the classifiers in the framework of semi-supervised learning.

Although the class-dependent structure is not directly exploited even by using unlabeled samples, it is possible to exploit the whole structure (manifold) of the pattern vectors through the unlabeled samples. The framework of semi-supervised learning has mainly been studied in the following three directions.

The first direction is based on transductive learning. The well-known method is transductive SVM (Vapnik, 1998; Joachims, 1999). Another is the semi-supervised logistic regression (Amini and Gallinari, 2002), which is closely related to the proposed method. These methods incorporate the unlabeled samples into supervised learning framework, iteratively relabeling (estimating) the labels of the unlabeled samples, and thus are susceptible to local minima. On the other hand, the proposed method differs from the method of semi-supervised logistic regression (Amini and Gallinari, 2002) in the following point: we define the objective cost function incorporating the unlabeled samples in a unified manner via graph Laplacian, and thereby the labels of the unlabeled samples are directly obtained without iteratively relabeling.

The second direction is to incorporate generative models, resulting in a hybrid model of the generative models and the discriminative models of supervised learning (Fujino et al., 2008; Lasserre et al., 2006). Ng and Jordan (2002) showed that the generative classifiers often work better than discriminative classifiers in the case of a few labeled samples, and thus the hybrid model would effectively perform on the semi-supervised problems. Practically speaking, however, it is difficult to appropriately determine the (parametric) generative model in advance.

The third direction is to use a graph-based representation, called graph Laplacian (Belkin and Niyogi, 2003), of the sample distribution. In the graph Laplacian, nodes stand for the samples and (weighted) edges between nodes represent the pairwise relationships between samples. Usually, predefined similarity measures between samples are assigned to the edge weights. Such defined graph Laplacian is used not only in the unsupervised learning to discover the (lower-dimensional) manifold of sample distribution (Belkin and Niyogi, 2003; Yan et al., 2007), but also in the semi-supervised learning. For example, the Laplacian support vector machine (LapSVM) introduces the unlabeled samples into the framework of SVM (Vapnik, 1998) and the method of semi-supervised discriminant analysis (SDA) (Cai et al., 2007; Zhang and Yeung, 2008) has also been proposed to incorporate the unlabeled samples into the well-known discriminant analysis. These methods define the energy cost function in the semi-supervised framework, consisting of the cost derived from discriminative learning and the energy over the graph Laplacian using the similarities. The latter graph-based energy plays a role of regularization in optimizing the classifiers. In the proposed method, we also follow this direction, but use the graph Laplacian not just as regularization but for discriminative learning in a manner similar to label propagation (see Section 3).

The similarity measures between samples are inherently required to construct the graph Laplacian. The performance of the semi-supervised classifier based on the graph Laplacian depends on what kind of similarity measure is used. There are a lot of works for measuring effective similarities: the most commonly used similarities are k-NN based similarity and Gaussian kernel similarity (Belkin and Niyogi, 2002), and the more sophisticated similarities are proposed by Cheng et al. (2009), Wang and Zhang (2007) and Wang et al. (2009) who assume linear relationship among the local neighborhoods of sample vectors. In this study, the design of the similarities is out of our focus and we suppose certain type of similarity is already given.

In the following subsections, we briefly describe the methods of logistic regression (Bishop, 2007) and label propagation (Zhu et al., 2003), since our proposed method is related to those methods.

2.1. Logistic regression

The method of logistic regression (LR) (Bishop, 2007) is applied to learn the classifier for pattern (feature) vectors in the probabilistic and supervised framework. Let $\boldsymbol{x}_i \in \mathcal{R}^d$ be the ith d-dimensional feature vector and $\boldsymbol{y}_i \in \{0,1\}^C$ be its binary label vector in which only the component y_{ic} associated with the assigned class ($c \in \{1,\ldots,C\}$) is 1 and the others are 0. In the logistic regression, the logistic function is employed to estimate the label vector $\boldsymbol{y}_i$ from the feature vectors $\boldsymbol{x}_i$:

$$\hat{y}_{ic} = \begin{cases} \frac{\exp(\boldsymbol{w}_c'\boldsymbol{x}_i)}{1+\sum_{q=1}^{C-1}\exp(\boldsymbol{w}_q'\boldsymbol{x}_i)} & (c \neq C) \\ \frac{1}{1+\sum_{q=1}^{C-1}\exp(\boldsymbol{w}_q'\boldsymbol{x}_i)} & (c = C) \end{cases}, \tag{1}$$

where $\boldsymbol{w}_c$ ($c = 1,\ldots,C-1$) are the coefficient vectors for the respective classes. The estimated label values $\hat{y}_{ic}$ are regarded as posterior probabilities over classes ($1 \sim C$). Thus, in the logistic regression, to optimize $\boldsymbol{w}_c$, we minimize the following cost function derived from the probabilistic perspective:

$$J(\boldsymbol{w}) = -\log\left(\prod_{c=1}^{C}\prod_{i=1}^{N}\hat{y}_{ic}^{y_{ic}}\right) = -\sum_{c=1}^{C}\sum_{i=1}^{N} y_{ic}\log(\hat{y}_{ic}), \tag{2}$$

where N denotes the number of samples. The minimization of this cost function means that log-likelihood is maximized across samples. The optimization is actually performed based on gradient descents (Bishop, 2007):

$$\frac{\partial J}{\partial \boldsymbol{w}_c} = \sum_{i=1}^{N}(\hat{y}_{ic} - y_{ic})\boldsymbol{x}_i. \tag{3}$$

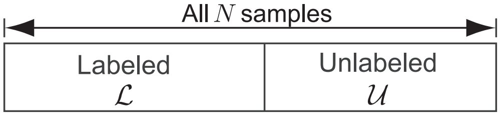


Fig. 1. All N samples are divided into labeled ($\mathcal{L}$) and unlabeled ($\mathcal{U}$) ones.

2.2. Label propagation

The method of label propagation (LP) (Zhu et al., 2003) integrates labeled and unlabeled samples for estimating binary class labels in the framework of semi-supervised learning. We consider a binary (two-class) label estimation problem and denote the class label value of the ith sample by $y_i \in [0,1]$ in which positive and negative labels are indicated by 1 and 0, respectively. Suppose, in all N samples, some of samples are labeled as positive ($y = 1$) or negative ($y = 0$), and the others are not assigned any labels. Let $\mathcal{L}$ be the index set of those labeled samples and $\mathcal{U}$ be that of the remaining unlabeled samples ($\mathcal{L} \cap \mathcal{U} = \phi, |\mathcal{L} \cup \mathcal{U}| = N$), as shown in Fig. 1. Given a symmetric similarity $s_{ij} = s_{ji}$ between the i, jth samples ($1 \leqslant i,j \leqslant N$), the method of LP estimates the labels $\boldsymbol{y}_{\mathcal{U}}$ of the unlabeled samples so as to minimize the following cost function with the given (fixed) $\boldsymbol{y}_{\mathcal{L}}$:

$$\begin{aligned} J(\boldsymbol{y}_{\mathcal{U}}) &= \frac{1}{2}\sum_{i,j=1}^{N} s_{ij}(y_i - y_j)^2, \text{ under } y_l \in \{0,1\} (\forall l \in \mathcal{L}) \text{ are given} \\ &= \frac{1}{2}\sum_{i,j \in \mathcal{U}} s_{ij}(y_i - y_j)^2 + \sum_{u \in \mathcal{U}, l \in \mathcal{L}} s_{ul}(y_u - y_l)^2 + const \\ &= \boldsymbol{y}'_{\mathcal{U}}(\boldsymbol{D}_{\mathcal{U}\mathcal{U}} - \boldsymbol{S}_{\mathcal{U}\mathcal{U}})\boldsymbol{y}_{\mathcal{U}} - 2\boldsymbol{y}'_{\mathcal{U}}\boldsymbol{S}_{\mathcal{U}\mathcal{L}}\boldsymbol{y}_{\mathcal{L}} + const, \end{aligned} \quad (4)$$

where $\boldsymbol{S}$ is a similarity matrix, $S_{ij} = s_{ij}$, and $\boldsymbol{D}$ is a diagonal matrix, $D_{ii} = \sum_{j=1}^{N} s_{ij}$. The optimum labels are obtained in a closed form:

$$\hat{\boldsymbol{y}}_{\mathcal{U}} = (\boldsymbol{D}_{\mathcal{U}\mathcal{U}} - \boldsymbol{S}_{\mathcal{U}\mathcal{U}})^{-1}\boldsymbol{S}_{\mathcal{U}\mathcal{L}}\boldsymbol{y}_{\mathcal{L}}. \quad (5)$$

LP estimates the class labels by solving the above analytic form (Eq. (5)), not by using any classifiers. Therefore, it has a difficulty in estimating the labels of newly input samples which are not taken into account in the training samples. In such cases, we have to reconstruct the whole similarity matrix including the newly input samples and solve Eq. (5) again at every time when new samples come, which results in significant computational cost even for the large amount of training samples.

A probabilistic interpretation is given to the LP from the viewpoint of random walks (Grady et al., 2005); namely, when the binary label values (0/1) are assigned to $\boldsymbol{y}_{\mathcal{L}}$, the estimated label values $\hat{\boldsymbol{y}}_{\mathcal{U}}$ are regarded as posterior probabilities for the positive class.

3. Logistic label propagation

We incorporate the logistic function in Eq. (1) into the framework of label propagation described in Section 2.2 in order to cope with the unlabeled samples in the semi-supervised manner. So, the proposed method is named *logistic label propagation (LLP)*.

3.1. Definition

We use the same notations as in Sections 2.1 and 2.2. In the proposed method, we minimize the following cost function using similarities s_{ij} between samples:

$$J(\boldsymbol{w}) = \sum_{c=1}^{C}\left\{\frac{1}{2}\sum_{i \in \mathcal{U}, j \in \mathcal{U}} s_{ij}(\hat{y}_{ic} - \hat{y}_{jc})^2 + \sum_{u \in \mathcal{U}, l \in \mathcal{L}} s_{ul}(\hat{y}_{uc} - y_{lc})^2 - \sum_{l \in \mathcal{L}} \eta_l y_{lc} \log(\hat{y}_{lc})\right\} \quad (6)$$

$$\text{s.t.} \ \forall l \in \mathcal{L}, \forall c \in \{1,\ldots,C\}, y_{lc} \in \{0,1\} \text{ is given}, \quad (7)$$

$$\forall i, \hat{y}_{ic} = \begin{cases} \dfrac{\exp(\boldsymbol{w}'_c \boldsymbol{x})}{1+\sum_{q=1}^{C-1}\exp(\boldsymbol{w}'_q \boldsymbol{x})} & (c \neq C) \\ \dfrac{1}{1+\sum_{q=1}^{C-1}\exp(\boldsymbol{w}'_q \boldsymbol{x})} & (c = C) \end{cases}, \quad (8)$$

where η_l is a balancing parameter at the lth labeled sample. We estimate the class label values of unlabeled samples $\hat{\boldsymbol{y}}_{\mathcal{U}}$ as well as those of labeled ones $\hat{\boldsymbol{y}}_{\mathcal{L}}$ by using logistic functions in Eq. (8). In Eq. (6), the first two terms indicate the cost derived from label propagation (Section 2.2) to cope with the unlabeled samples in a semi-supervised manner, and the last term[1] measures the classification errors across the labeled samples by negative log-likelihood as in logistic regression (Section 2.1). These terms are balanced in each labeled sample by using the parameters η_l. The way to determine the parameter values is described in Section 3.2.2. By minimizing the above-defined cost function, we can obtain the optimum logistic functions that favorably estimate the class label values (posterior probabilities) from the semi-supervised perspective.

We apply the gradient descent approach to minimize the cost function (Eq. (6)) in a manner similar to logistic regression in Eq. (3). The derivative of the cost function in Eq. (6) with respect to the coefficient vector $\boldsymbol{w}_c$ is written by

$$\frac{\partial J}{\partial \boldsymbol{w}_c} = 2\sum_{u \in \mathcal{U}}\left(\hat{y}_{uc}\epsilon_{uc} - \hat{y}_{uc}\sum_{q=1}^{C}\hat{y}_{uq}\epsilon_{uq}\right)\boldsymbol{x}_u + \sum_{l \in \mathcal{L}} \eta_l(\hat{y}_{lc} - y_{lc})\boldsymbol{x}_l, \quad (9)$$

$$\text{where } \forall i \in \mathcal{U}, \ \forall c, \epsilon_{ic} = \sum_{u \in \mathcal{U}}(D_{iu} - s_{iu})\hat{y}_{uc} - \sum_{l \in \mathcal{L}} s_{il} y_{lc}. \quad (10)$$

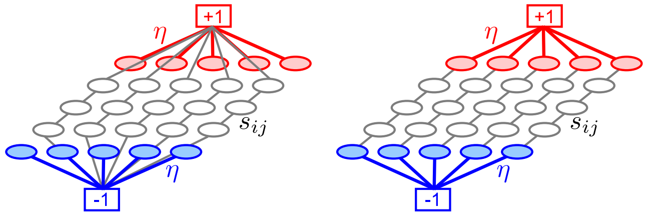
Details are described in Appendix A. By using the gradient, we can apply, e.g., quasi-newton method and conjugate gradient method (Nocedal and Wright, 1999). We present how to select the favorable initial points for the optimization in Section 3.2.1.

From the probabilistic viewpoint, the logistic function in Eq. (8) is suitable in the framework of label propagation, since the function approximates the posterior probability and outputs the value ranged from 0 to 1. Therefore, it can be said that the LLP estimates the posterior probability more accurately than label propagation which does not explicitly impose such probabilistic constraint ($0 \leqslant \hat{y} \leqslant 1$) on the estimated values. The method of locality preserving projection (He and Niyogi, 2004) also considers the graph Laplacian using similarities in the lower dimensional space into which sample vectors are mapped via linear (or kernel-based) projections. The locality preserving projection, however, is an unsupervised method using only 'unlabeled' samples without labeled ones, and the embedded space is unbounded, not providing probabilistic interpretations.

A graphical model of LLP is shown in Fig. 2(a) in the case of a binary classification. There are N nodes of samples and two special nodes standing for the binary classes (y). We introduce three types of edges: the first type is defined between unlabeled samples (the first term in Eq. (6)), the second type is between the unlabeled and the labeled samples (the second term in Eq. (6)), and the third type links the labeled samples and the class node, i.e., the given label (the third term in Eq. (6)). The first and the second types of edges are weighted by the similarity measures s_{ij}, while the third type edge is assigned with the parameter value η_l. It should be noted that the labeled samples are not directly connected to the unlabeled ones, but indirectly via the class nodes. Thus, the associated terms, i.e., the first and the second terms in Eq. (6), exhibit high discriminative power as in label propagation.

As a simple alternative model of LLP, it is conceivable to directly connect the labeled samples to the unlabeled ones, as shown in Fig. 2(b), and its cost function is then described by

[1] In (Watanabe et al., 2010), we did not take this term into account.



(a) Graphical model of LLP (b) Alternative model

Fig. 2. Graphical model used in LLP. The empty circle nodes indicate the unlabeled samples, while the filled nodes denote the labeled ones. Those nodes are connected by the edge (solid line) with the weight of s_{ij}. There are two special nodes standing for positive and negative classes. (a) In LLP, the labeled nodes are linked to the unlabeled ones not directly but via class nodes. (b) We can consider the alternative model of the LLP, in which the labeled and the unlabeled nodes are directly connected. This model, however, makes it difficult to appropriately determine the parameter value (η) corresponding to the edge weight between the labeled and the unlabeled samples.

$$\widetilde{J}(\boldsymbol{w}) = \sum_{c=1}^{C}\left\{\frac{1}{2}\sum_{i=1,j=1}^{N} s_{ij}(\hat{y}_{ic}-\hat{y}_{jc})^2 - \sum_{l\in\mathcal{L}}\eta_l y_{lc}\log(\hat{y}_{lc})\right\}. \tag{11}$$

This alternative model is, however, impractical, because it is difficult to appropriately determine the balancing parameter η. Actually, the first term in Eq. (11) has the trivial global optimum solution, namely all the components are uniform ($\hat{\boldsymbol{y}}_i = \frac{1}{C}\mathbf{1}$), while the second term enforces the solution to be the discriminative solution of logistic regression applied to the labeled samples. These two solutions are contradictory to each other and the parameters η_l cannot balance those properly. In practice, we can find that, by varying those parameter values, the obtained solution switches between those solutions.

Based on the above arguments, we can say that the model in the proposed LLP (Fig. 2(a)) is meaningful, combining two discriminative costs derived from label propagation and logistic regression, respectively. The parameter η can appropriately balance those two discriminative costs, as described in Section 3.2.2.

The other semi-supervised methods by Belkin and Niyogi (2006) and Cai et al. (2007) are also based on the graph-based formulations (graph Laplacian), while the transductive methods of Joachims (1999) and Amini and Gallinari (2002) are not. Those methods simply introduce the graph-based energy as regularization of the classifier weights; the graph-based energy is described by $\sum_{i=1,j=1}^{N} s_{ij}(\hat{y}_{ic}-\hat{y}_{jc})^2$. This graph-based regularization is similar to Eq. (11) and Fig. 2(b), but it is defined without considering the given labels unlike the proposed method.

3.2. Practical issues

In this subsection, we address two practical issues in the proposed method: one is how to select the initial points for the optimization, and the other is how to determine the balancing parameter value η_l.

3.2.1. Initialization

Since the cost function in Eq. (6) is not necessarily convex, the gradient-descent based optimization can be trapped by one of local minima. In this case, the initial points in the optimization affects the obtained solution and it is important to select 'good' initial points for $\boldsymbol{w}$ such that the obtained local minima is close to the global minimum. In this study, we determine the initial point $\boldsymbol{w}^{(0)}$ based on the solution by linear regression which is analytically obtained.

In the proposed method, the logistic functions in Eq. (8) are employed to approximate the label values in the label propagation through the cost function in Eq. (6). Thus, we determine the initial logistic functions that directly approximate the solution of label propagation. First, we obtain the solution $\hat{\boldsymbol{y}}_c^*$ of the label propagation by analytically solving Eq. (5). Note that the given labels are also utilized by setting $\hat{y}_{lc}^* = y_{lc},\ (l\in\mathcal{L})$. Next, the logistic models are fitted to the solution $\hat{\boldsymbol{y}}_c^*$ in a least square sense:

$$\hat{y}_{ic}^* \approx \frac{\exp(\boldsymbol{w}_c^{(0)\prime}\boldsymbol{x}_i)}{1+\sum\limits_{c=1}^{C-1}\exp(\boldsymbol{w}_c^{(0)\prime}\boldsymbol{x}_i)} \rightarrow \exp(\boldsymbol{w}_c^{(0)\prime}\boldsymbol{x}_i) - \hat{y}_{ic}^*\sum_{c=1}^{C-1}\exp(\boldsymbol{w}_c^{(0)\prime}\boldsymbol{x}_i) \approx \hat{y}_{ic}^* \tag{12}$$

$$\boldsymbol{W}^{(0)\prime}\boldsymbol{x}_i \approx \log\left(\underbrace{\begin{bmatrix} 1-\hat{y}_{i1}^* & -\hat{y}_{i1}^* & \cdots & -\hat{y}_{i1}^* \\ \vdots & \ldots & \vdots & \vdots \\ -\hat{y}_{iC-1}^* & -\hat{y}_{iC-1}^* & \cdots & 1-\hat{y}_{iC-1}^* \end{bmatrix}^{-1}\begin{bmatrix}\hat{y}_{i1}^* \\ \vdots \\ \hat{y}_{iC-1}^*\end{bmatrix}}_{\equiv\boldsymbol{\lambda}_i}\right) \tag{13}$$

$$\Rightarrow \boldsymbol{W}^{(0)\prime}\boldsymbol{X} \approx \underline{\log}([\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_n]) \tag{14}$$

$$\therefore\ \boldsymbol{W}^{(0)} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}'\boldsymbol{X} - \underline{\log}([\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_n])\|^2, \tag{15}$$

where $\underline{\log}(\boldsymbol{M})$ denotes the matrix consisting of truncated logarithm of the components;

$$\{\underline{\log}(\boldsymbol{M})\}_{ij} = \begin{cases}\log(M_{ij}) & (M_{ij} > \Delta), \\ \log(\Delta) & (\text{otherwise}),\end{cases} \tag{16}$$

where Δ is the small positive number (say, $\Delta = 2e^{-16}$). Eq. (15) corresponds to a linear least square problem, which can be analytically solved. The solution $\boldsymbol{W}^{(0)}$ is employed as the initial points in the optimization of $\boldsymbol{W}$.

3.2.2. Balancing parameter

We have no prior knowledge about the parameter η_l which balances the cost derived from label propagation and that from logistic regression in Eq. (6). Such parameter is generally determined based on the empirical classification performance, e.g., by cross validations. The empirical determination, however, requires additional exhaustive computation besides the optimization process and also somewhat larger amount of labeled samples to statistically estimate the classification performance. Therefore, we suggest the following way to settle the parameter value η_l.

The parameter $\eta_l (l\in\mathcal{L})$ are associated with the labeled samples ($\mathcal{L}$). By focusing on the last two terms in Eq. (6) connected to the labeled samples, the parameter η_l actually balances those terms:

$$\sum_{u\in\mathcal{U}} s_{lu}(y_{lc}-\hat{y}_{uc})^2 + \eta_l\{-y_{lc}\log(\hat{y}_{lc})\}. \tag{17}$$

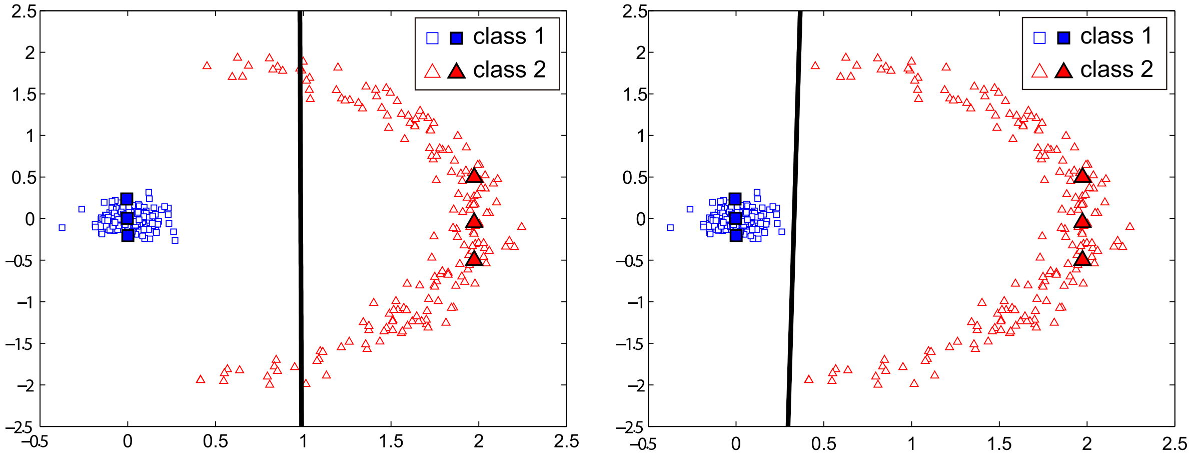
By comparing the coefficients with respect to y_{lc}, we can determine η_l so as to equally balance these two terms as follows,

$$\eta_l = \sum_{u\in\mathcal{U}} s_{lu}. \tag{18}$$

In the case that we do not have any unlabeled samples, $\mathcal{U} = \phi$, however, the above-defined η_l results in 0. To avoid such unfavorable situations, we simply introduce the lower bound of η_l by

$$\eta_l = \max\left(\sum_{u\in\mathcal{U}} s_{lu}, 1\right). \tag{19}$$

It should be noted that we use the similarity bounded by $0 \leqslant s_{ij} \leqslant 1$ and simply set the lower bound of η as 1. In the experiments (Section 4), the parameter η_l is determined by Eq. (19) on all datasets.



(a) Logistic regression (LR)

(b) Logistic label propagation (LLP)

Fig. 3. Experimental results on toy examples. There are two classes denoted by blue squares and red triangles. We labeled only three samples per class, denoted by filled markers, and regarded the others as unlabeled samples (unfilled markers). The obtained classifiers ($\boldsymbol{w}$) are indicated by black solid lines. (a) Logistic regression learnt from the labeled samples produces unfavorable classifier, (b) while logistic label propagation learnt from all samples provides the favorable classifier successfully classifying the samples.

3.3. Kernel logistic label propagation

At the last of this section, we mention the extension of LLP to the kernel-based method via kernel tricks (Schölkopf and Smola, 2001). The method of LLP defined in Section 3.1 uses the logistic function with linear coefficients $\boldsymbol{w}_c$ for the feature vector $\boldsymbol{x}$. By replacing it with the kernel-based logistic function as in kernel logistic regression (Zhu and Hastie, 2005), we can easily obtain the method of kernel logistic label propagation (KLLP) as follows:

$$\hat{y}_c = \frac{\exp\left(\sum_{i=1}^{N}\alpha_{ic}k(\boldsymbol{x}_i,\boldsymbol{x})\right)}{1+\sum_{q=1}^{C-1}\exp\left(\sum_{i=1}^{N}\alpha_{iq}k(\boldsymbol{x}_i,\boldsymbol{x})\right)} = \frac{\exp(\boldsymbol{\alpha}_c'\boldsymbol{k}(\boldsymbol{x}))}{1+\sum_{q=1}^{C-1}\exp(\boldsymbol{\alpha}_q'\boldsymbol{k}(\boldsymbol{x}))}, \qquad (20)$$

where $k(\boldsymbol{x},\boldsymbol{z})$ indicates the kernel function for the sample vectors of $\boldsymbol{x}$ and $\boldsymbol{z}$, $\boldsymbol{k}(\boldsymbol{x}) = [k(\boldsymbol{x}_1,\boldsymbol{x}),\ldots,k(\boldsymbol{x}_N,\boldsymbol{x})]' \in \mathcal{R}^N$, and $\boldsymbol{\alpha}_c = [\alpha_{1c},\ldots,\alpha_{Nc}]' \in \mathcal{R}^N$ denotes the coefficient vectors for the cth class. In KLLP, we minimize the objective cost function in Eq. (6) under Eq. (20) instead of Eq. (8), and then obtain the following derivative by simply replacing $\boldsymbol{w}_c$ and $\boldsymbol{x}_i$ with $\boldsymbol{\alpha}_c$ and $\boldsymbol{k}(\boldsymbol{x}_i)$ in Eq. (9), respectively:

$$\frac{\partial J}{\partial \boldsymbol{\alpha}_c} = 2\sum_{u\in\mathcal{U}}\left(\hat{y}_{uc}\epsilon_{uc} - \hat{y}_{uc}\sum_{q=1}^{C}\hat{y}_{uq}\epsilon_{uq}\right)\boldsymbol{k}(\boldsymbol{x}_u) + \sum_{l\in\mathcal{L}}\eta_l(\hat{y}_{lc}-y_{lc})\boldsymbol{k}(\boldsymbol{x}_l), \qquad (21)$$

$$\text{where } \forall i\in\mathcal{U},\ \forall c, \epsilon_{ic} = \sum_{u\in\mathcal{U}}(D_{iu}-s_{iu})\hat{y}_{uc} - \sum_{l\in\mathcal{L}}s_{il}y_{lc}. \qquad (22)$$

We can apply the same practical techniques described in Section 3.2 to the KLLP.

4. Experimental results

We conducted experiments on classification in the framework of semi-supervised learning. There are three experiments; a toy example, benchmark datasets from UCI repository, and ETH-80 dataset (Leibe and Schiele, 2003) for object recognition. As to the similarity, we employed the similarity proposed in (Wang and Zhang, 2007; Wang et al., 2009) on all datasets.

4.1. Toy example

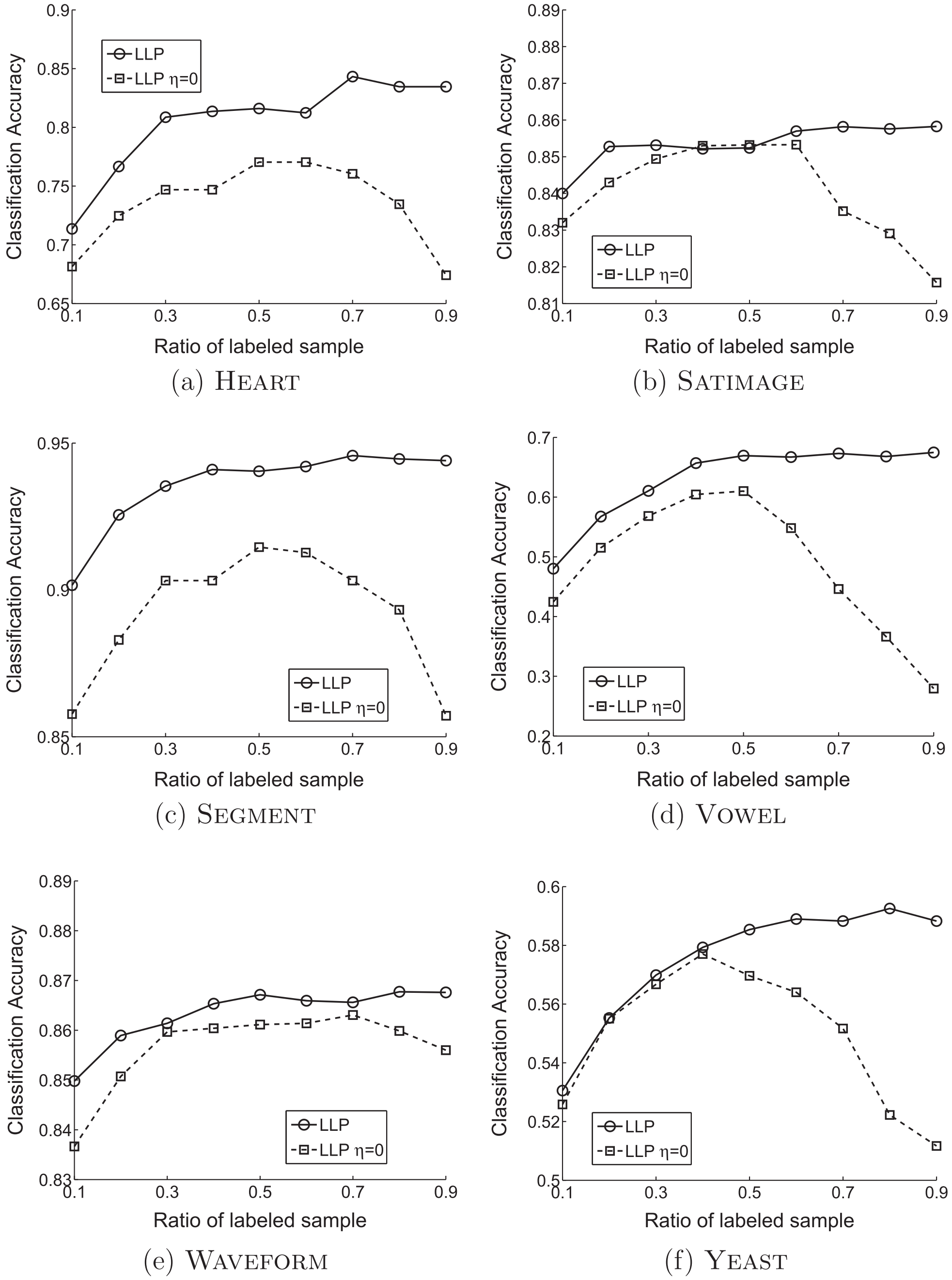
First, in order to demonstrate the effectiveness of the proposed method (LLP), we applied LLP to a toy example consisting of two classes: samples in one class are generated according to the Gaussian distribution and those in the other class obey the curved distribution, as shown in Fig. 3. We assigned the class labels only to three samples in each class (filled markers in Fig. 3). For comparison, the method of logistic regression (LR) in the supervised learning framework is applied. The sample distribution and the classifiers obtained by LLP and LR are shown in Fig. 3. LR utilizes only the labeled samples and thus yields the classifier that cannot properly classify the unlabeled samples as shown in Fig. 3(a). On the other hand, LLP learns the classifier using both the labeled and the unlabeled samples, and all the samples are correctly classified as shown in Fig. 3(b). Through the graph Laplacian using similarities between samples, the underlying manifold structure of the whole samples are effectively extracted, as in graph embedding method (Yan et al., 2007). This result shows the effectiveness of the proposed method in the framework of semi-supervised learning.

4.2. Benchmark dataset

Next, we applied the method to several benchmark datasets collected from UCI repository; HEART, SATIMAGE, SEGMENT, VOWEL, WAVEFORM and YEAST. By using these datasets, we evaluated the performances by threefold cross validation. In each fold, the training set is further randomly split into labeled and unlabeled samples. The performances are evaluated on various ratios of labeled samples: $|\mathcal{L}|/(|\mathcal{L}|+|\mathcal{U}|) = 0.1 \sim 0.9$. The unlabeled samples, which are not assigned with any class labels, are used for training in the semi-supervised learning methods, and the performance is measured on the test samples with excluding the training samples. At each ratio, the trial is repeated three times. Finally, we report the averaged performances across the three times trials for the threefold cross validations.

As a performance study, we compare LLP to the (fixed) LLP with $\eta = 0$,[2] demonstrating the effectiveness of the logistic regression term in Eq. (6). Note that in LLP, the parameter values $\eta_l(l\in\mathcal{L})$ are

[2] The preliminary formulation in (Watanabe et al., 2010) corresponds to the LLP with $\eta = 0$.



(a) Heart

(b) Satimage

(c) Segment

(d) Vowel

(e) Waveform

(f) Yeast

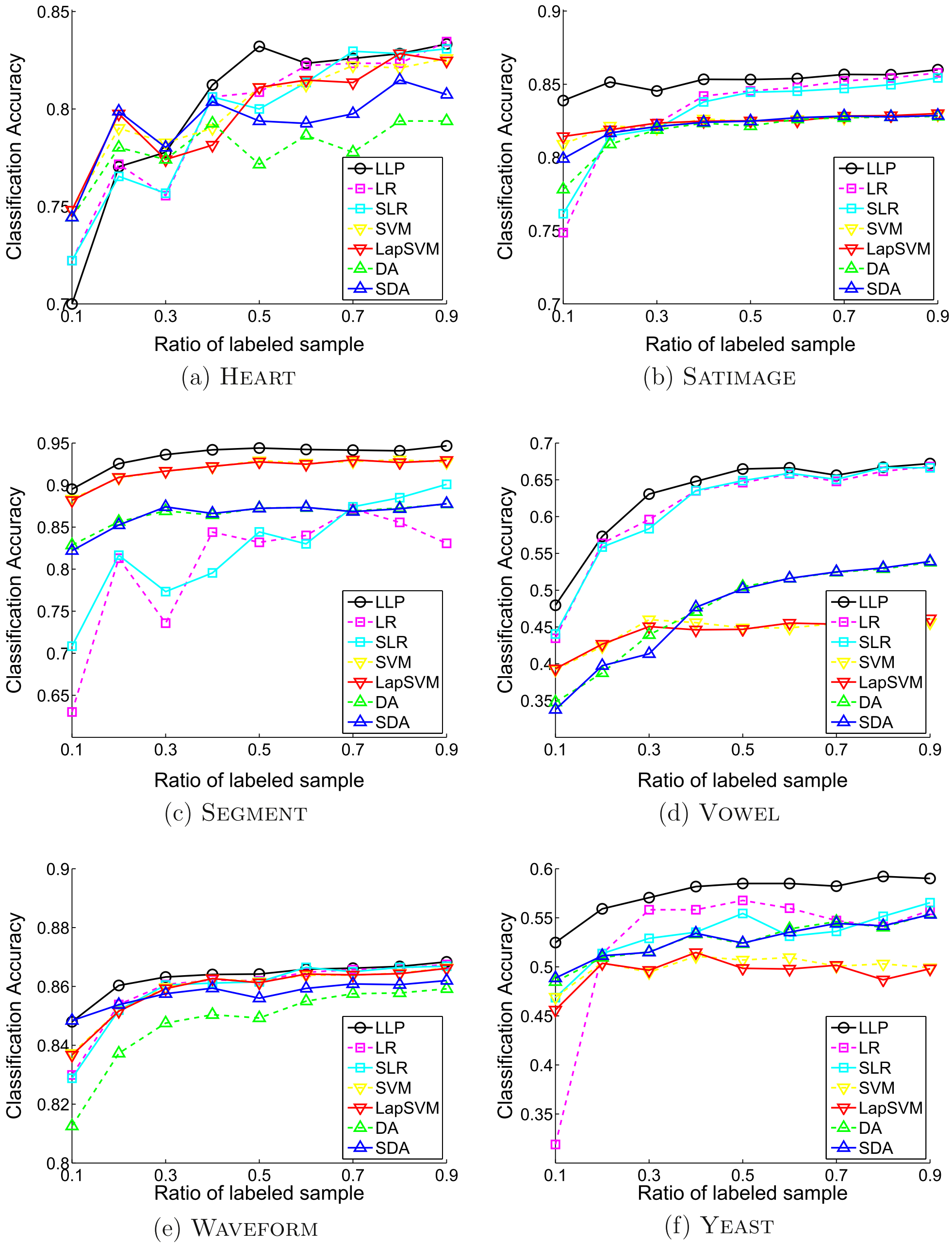
Fig. 4. Performance study of LLP by using benchmark datasets.

Table 1
Classification accuracies for various balancing parameters at $|\mathcal{L}|/(|\mathcal{L}| + |\mathcal{U}|) = 0.4$. The numbers in the parentheses indicate the ranks of the performances.

	Heart	Satimage	Segment	Vowel	Waveform	Yeast
(a) Similarity by Wang and Zhang (2007) *and* Wang et al. (2009)						
Ours	0.8136 (1)	0.8522 (2)	0.9410 (1)	0.6569 (1)	0.8653 (1)	0.5793 (1)
$\eta = 1$	0.8136 (2)	0.8523 (1)	0.9410 (2)	0.6495 (2)	0.8637 (3)	0.5750 (4)
$\eta = 5$	0.8012 (3)	0.8413 (3)	0.9401 (3)	0.6360 (3)	0.8643 (2)	0.5768 (2)
$\eta = 10$	0.7975 (4)	0.8392 (4)	0.9392 (4)	0.6293 (4)	0.8634 (4)	0.5757 (3)
(b) Gaussian kernel similarity						
Ours	0.7704 (1)	0.8204 (1)	0.8492 (1)	0.6202 (1)	0.8451 (1)	0.5613 (1)
$\eta = 1$	0.6889 (4)	0.7233 (4)	0.4724 (4)	0.3943 (4)	0.7605 (4)	0.4393 (4)
$\eta = 5$	0.7259 (3)	0.7285 (3)	0.5306 (3)	0.5172 (3)	0.7796 (3)	0.4827 (3)
$\eta = 10$	0.7630 (2)	0.7348 (2)	0.5856 (2)	0.5684 (2)	0.7938 (2)	0.5135 (2)

Table 2
Objective cost values (J) of the proposed and random initialization at $|\mathcal{L}|/(|\mathcal{L}|+|\mathcal{U}|)=0.4$.

	Heart	Satimage	Segment	Vowel	Waveform	Yeast
Ours	−1.4120	−111.9236	−99.7839	72.1649	−35.3092	148.4519
Rand (100 trials)	−1.4120	−110.3993	−96.7556	72.1649	−35.3092	148.4793



(a) Heart

(b) Satimage

(c) Segment

(d) Vowel

(e) Waveform

(f) Yeast

Fig. 5. Classification performances compared to the other methods on benchmark datasets.

adaptively determined as described in Section 3.2.2. The performance results are shown in Fig. 4. LLP exhibits superior performances in all datasets, and thus we can say that the performances are improved by incorporating the logistic regression term. The performances of the LLP with $\eta = 0$ are maximized around the half ratio of labeled samples on most of the datasets. The reason is as follows. When only a few samples are labeled, such label information is insufficient to discriminate the other samples. On the other hand, for larger amount of labeled samples, the LLP with $\eta = 0$ is learnt from only a few unlabeled samples in contrast (see the first two terms in Eq. (6)), resulting in less generalization performance. Thus, the half ratio of labeled samples is a trade-off point where the performance is maximized. The proposed LLP effectively incorporates the labeled samples in the logistic regression cost (the last term in Eq. (6)) with adaptively determined parameter value η_l, and thus the performance is improved as the number of labeled samples increases.

Then, we investigated the sensitivities to the balancing parameter η and show the effectiveness of the adaptive determination of the

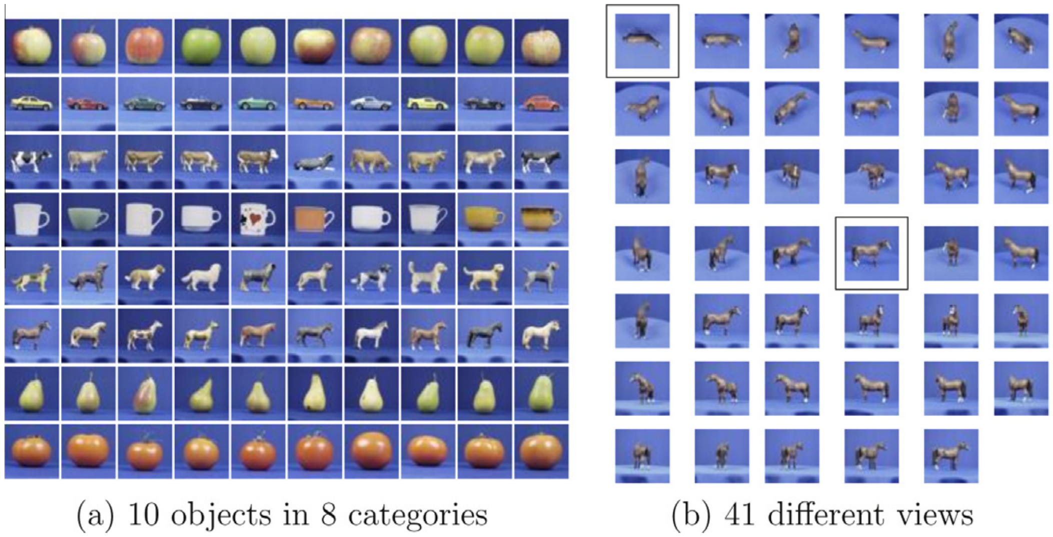


Fig. 6. ETH-80 dataset (Leibe and Schiele, 2003). (a) There are eight categories (row) and ten objects (column) in each category. (b) In each object, 41 images are captured from various camera angles. The labeled images (views) are indicated by boxes.

Table 3
Classification accuracy on ETH-80 dataset.

Method	Acc.
LLP	**0.6308**
LR	0.1302
SLR	0.1302
SVM	0.4829
LapSVM	0.4841
DA	0.5125
SDA	0.6198

parameter described in Section 3.2.2. For comparison, we applied the constant parameter values η with the similarity (Wang and Zhang, 2007; Wang et al., 2009) that we use in the experiments and Gaussian kernel similarity, and the classification accuracies are shown in Table 1. As shown in Table 1(a), the proposed method exhibits favorable performances, while all the parameter values, especially $\eta = 1$, are comparable due to the discriminative power of the similarity (Wang and Zhang, 2007; Wang et al., 2009). In the Gaussian kernel similarity, the performances of the proposed method are significantly superior to the others. This experimental result shows that the proposed parameter determination adaptively works on various types of similarities.

We further investigated the optimality of the LLP with the proposed initialization described in Section 3.2.1. Table 2 shows the objective cost values J in Eq. (6) for the proposed initialization and the random initialization. In the random initialization, we report the best performance over the 100 times random initializations. The cost values by the proposed initialization are lower than (or equal to) those by the random initialization. We can say that the proposed initialization is effective to obtain 'good' solution, though the obtained solution is not surely global optimum due to that the proposed formulation is not necessarily convex as described in Section 3.2.1.

For a comparative study, we applied the other methods: logistic regression (LR), semi-supervised logistic regression (SLR) (Amini and Gallinari, 2002), SVM (Vapnik, 1998), Laplacian SVM (LapSVM) (Belkin and Niyogi, 2006), discriminant analysis (DA), and semi-supervised discriminant analysis (SDA) (Cai et al., 2007). The methods of DA and SDA construct the lower-dimensional discriminant space into which the sample vectors are mapped, and the linear classifier is learnt by applying SVM to labeled samples in that space for a fair comparison; all the methods are based on linear classifiers. The methods of LLP, SLR, LapSVM and SDA are semi-supervised learning methods. The performance results in Fig. 5 show that LLP exhibits favorable performances compared to the others in almost all datasets and ratios of labeled samples. The logistic regression (LR) in supervised learning framework is susceptible to over-fitting. The proposed method (LLP), however, incorporates not only labeled but also unlabeled samples and thus it possibly avoids such over-fitting problem.

4.3. ETH-80 dataset

The ETH-80 database (Leibe and Schiele, 2003) contains 80 objects from eight categories, as shown in Fig. 6(a). Each object is represented in 41 images from equally spaced viewpoints over the upper viewing hemisphere, as shown in Fig. 6(b). The performances are measured by fivefold cross validation for objects; the images of 64 objects are used for training and those of the remained 16 objects are for test. In the training, we labeled only two images (views) for each object, as shown in Fig. 6(b). In this case of quite a few labeled samples, we constructed the graph Laplacian by using all images both of training and test ones. The bag-of-features (Csurka et al., 2004) are extracted from the image by using SIFT features (Lowe, 2004) on grid points and then by clustering them via hierarchical k-means (Nister and Stewenius, 2006) with 11 branches and four depths. All the methods are based on linear classifiers in the same manner as Section 4.2. The performance results are shown in Table 3. LLP produces the superior performance to the others even in such case of a few sparsely labeled sample. SLR falls to a local minimum corresponding to the initial point produced by LR in this case.

5. Concluding remarks

We have proposed the method of logistic label propagation (LLP) in the framework of semi-supervised learning. The proposed LLP can effectively incorporate unlabeled samples as well as labeled ones via similarities as in label propagation, and in addition, the label values are estimated by using the logistic functions which are used in logistic regression. In LLP, these methods of label propagation and logistic regression are integrated in terms of posterior probabilities. LLP can estimate the label values of newly input samples (vectors) by using the learnt logistic function as posterior probabilities without the exhaustive computation for solving linear equations of the whole graph Laplacian unlike the standard label propagation. Another merit of the proposed semi-supervised learning method is that it can estimate the labels from a small amount of labeled samples, which is favorable in practical situations where labeling samples is a exhaustive task. In addition, we suggested the ways for the parameter setting and the initialization in the optimization of the proposed method, and as a result, the proposed method has no parameter that users have to determine. In the experiments on classification in the semi-supervised learning framework, the proposed method produced favorable performances compared to the other semi-supervised learning methods including Laplacian support vector machine and semi-supervised discriminant analysis, etc.

Appendix A. Gradients in LLP

The derivative of the cost function in Eq. (6) with respect to $\boldsymbol{w}_c$ is

$$\frac{\partial J}{\partial \boldsymbol{w}_c} = \sum_{q=1}^{C}\left(\sum_{u\in\mathcal{U}} \frac{\partial J}{\partial y_{uq}}\frac{\partial y_{uq}}{\partial \boldsymbol{w}_c} + \sum_{l\in\mathcal{L}} \frac{\partial J}{\partial y_{lq}}\frac{\partial y_{lq}}{\partial \boldsymbol{w}_c}\right). \tag{A.1}$$

The respective forms of the derivative that appear in Eq. (A.1) are written as follows:

$$\forall q, \quad \forall u\in\mathcal{U}, \quad \frac{\partial J}{\partial y_{uq}} = 2\sum_{i\in\mathcal{U}}(D_{ui}-s_{ui})\hat{y}_{iq} - 2\sum_{l\in\mathcal{L}} s_{ul}y_{lq} \triangleq 2\epsilon_{uq}, \tag{A.2}$$

$$\forall q, \quad \forall l\in\mathcal{L}, \quad \frac{\partial J}{\partial y_{lq}} = -\eta_l \frac{y_{lq}}{\hat{y}_{lq}}, \tag{A.3}$$

$$\forall q, \quad \forall i, \quad \frac{\partial y_{iq}}{\partial \boldsymbol{w}_c} = (\delta(q,c)\hat{y}_{ic} - \hat{y}_{iq}\hat{y}_{ic})\boldsymbol{x}_i, \tag{A.4}$$

where $\delta(q,c)$ is a delta function which equals to 1 only for $q = c$ and 0 otherwise. By substituting these equations into Eq. (A.1), we obtain

$$\frac{\partial J}{\partial \boldsymbol{w}_c} = 2\sum_{u\in\mathcal{U}}\sum_{q=1}^{C}\epsilon_{uq}(\delta(q,c)\hat{y}_{uc} - \hat{y}_{uc}\hat{y}_{uq})\boldsymbol{x}_u - \sum_{l\in\mathcal{L}}\sum_{q=1}^{C}\eta_l\frac{y_{lq}}{\hat{y}_{lq}}(\delta(q,c)\hat{y}_{lc} - \hat{y}_{lq}\hat{y}_{lc})\boldsymbol{x}_l, \tag{A.5}$$

$$= 2\sum_{u\in\mathcal{U}}\left(\hat{y}_{uc}\epsilon_{uc} - \hat{y}_{uc}\sum_{q=1}^{C}\hat{y}_{uq}\epsilon_{uq}\right)\boldsymbol{x}_u + \sum_{l\in\mathcal{L}}\eta_l(\hat{y}_{lc} - y_{lc})\boldsymbol{x}_l. \tag{A.6}$$

References

Amini, M.-R., Gallinari, P., 2002. Semi-supervised logistic regression. In: European Conference on Artificial Intelligence, pp. 390–394.

Belkin, M., Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, pp. 346–358.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. (15), 1373–1396.

Belkin, M., Niyogi, P., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach .Learn. Res. (48), 1–36.

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. Springer.

Bishop, C. M., Ulusoy, I., 2005. Object recognition via local patch labeling. In: Workshop on Machine Learning.

Cai, D., He, X., Han, J., 2007. Semi-supervised discriminant analysis. In: International Conference on Computer Vision.

Cheng, H., Liu, Z., Yang, J., 2009. Sparsity induced similarity measure for label propagation. In: International Conference on Computer Vision.

Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74.

Fujino, A., Ueda, N., Saito, K., 2008. Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. IEEE Trans. Pattern Anal. Machine Intell. 30 (3), 424–437.

Grady, L., Schiwietz, T., Aharon, S., 2005. Random walks for interactive alpha-matting. In: Visualization, Imaging and Image Processing.

He, X., Niyogi, P., 2004. Locality preserving projections. Adv. Neural Inform. Process. Syst. 16.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning, pp. 200–209.

Kang, F., Jin, R., Sukthankar, R., 2006. Correlated label propagation with application to multi-label learning. In: IEEE Conference on Computer Vision and Pattern Recognition.

Lasserre, J.A., Bishop, C.M., Minka, T.P., 2006. Principled hybrids of generative and discriminative models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 87–94.

Leibe, B., Schiele, B., 2003. Analyzing appearance and contour based methods for object categorization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–415.

Levin, A., Lischinski, D., Weiss, Y., 2008. A closed form solution to natural image matting. IEEE Trans. Pattern Anal. Machine Intell. 30 (2), 228–242.

Lowe, D., 2004. Distinctive image features from scale invariant features. Int. J. Comput. Vision 60, 91–110.

Ng, A., Jordan, M., 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. Adv. Neural Inform. Process. Syst. 14, 841–848.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161–2168.

Nocedal, J., Wright, S.J., 1999. Numerical Optimization. Springer.

Schölkopf, B., Smola, A., 2001. Learning with Kernels. MIT Press.

Tomioka, R., Aihara, K., Muller, K.-R., 2007. Logistic regression for single trial eeg classification. Adv. Neural Inform. Process. Syst. 19, 1377–1384.

Vapnik, V., 1998. Statistical Learning Theory. Wiley.

Wang, F., Zhang, C., 2007. Label propagation through linear neighborhoods. IEEE Trans. Knowledge Data Eng. 20 (1), 55–67.

Wang, J., Wang, F., Zhang, C., Shen, H.C., Quan, L., 2009. Linear neighborhood propagation and its applications. IEEE Trans. Pattern Anal. Machine Intell. 31 (9), 1600–1615.

Watanabe, K., Kobayashi, T., Otsu, N., 2010. Logistic label propagation for semi-supervised learning. In: International Conference on Neural Information Processing, pp. 462–469.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S., 2007. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Machine Intell. 29 (1), 40–51.

Zhang, Y., Yeung, D.-Y., 2008. Semi-supervised discriminant analysis using robust path-based similarity. In: IEEE Conference on Computer Vision and Pattern Recognition.

Zhu, J., Hastie, T., 2005. Kernel logistic regression and the import vector machine. J. Comput. Graph. Stat. 14 (1), 185–205.

Zhu, X., Ghahramani, Z., Lafferty, J., 2003. Semi-supervised learning using gaussian fields and harmonic functions. In: International Conference on Machine Learning.