Neurocomputing 139 (2014) 130–137

Contents lists available at ScienceDirect

Neurocomputing

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

# Semi-supervised classification with pairwise constraints

Chen Gong [a,b], Keren Fu [a], Qiang Wu [b], Enmei Tu [a], Jie Yang [a,*]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[b] School of Computing and Communications, University of Technology, Sydney, Australia

ARTICLE INFO

*Article history:*
Received 19 September 2013
Received in revised form
26 November 2013
Accepted 7 February 2014
Communicated by D. Tao
Available online 5 April 2014

*Keywords:*
Semi-supervised learning
Pairwise constraints
Smoothness regularizer

ABSTRACT

Graph-based semi-supervised learning has been intensively investigated for a long history. However, existing algorithms only utilize the similarity information between examples for graph construction, so their discriminative ability is rather limited. In order to overcome this limitation, this paper considers both similarity and dissimilarity constraints, and constructs a signed graph with positive and negative edge weights to improve the classification performance. Therefore, the proposed algorithm is termed as Constrained Semi-supervised Classifier (CSSC). A novel smoothness regularizer is proposed to make the "must-linked" examples obtain similar labels, and "cannot-linked" examples get totally different labels. Experiments on a variety of synthetic and real-world datasets demonstrate that CSSC achieves better performances than some state-of-the-art semi-supervised learning algorithms, such as Harmonic Functions, Linear Neighborhood Propagation, LapRLS, LapSVM, and Safe Semi-supervised Support Vector Machines.



## 1. Introduction

Semi-supervised learning (SSL) is widely adopted in many situations where the labeled examples are insufficient while the unlabeled examples are extremely abundant. Though these massive unlabeled examples do not have explicit labels, they provide the prior of underlying data distribution, which can support accurate classifications along with the labeled examples.

However, the unlabeled examples should be used properly with certain assumptions, otherwise they may hurt the performance instead. Two commonly adopted assumptions are cluster assumption and manifold assumption [1]. Cluster assumption assumes that the examples of different classes form several well-separated clusters, and the decision boundary falls into the low density area in the feature space. Representative algorithms include Transductive Support Vector Machines (TSVM, [2]), Multiple Kernel TSVM [3], concaVe Semi-supervised Support Vector Machine (VS3VM, [4]), Structural Regularized Support Vector Machines (SRSVM, [5]), and Safe Semi-supervised Support Vector Machines (S4VM, [6]), etc. Methods above are the variants of traditional supervised Support Vector Machines (SVM). The only differences are on the definition of loss function, since the hinge loss employed by traditional SVM cannot be directly applied to the semi-supervised settings.

Manifold assumption postulates that the geometry of data distribution is usually supported by an underlying manifold (*e.g.* Riemannian manifold). The manifold can be described by a graph, of which the examples are represented by vertices and their similarities are measured by weighted edges. Therefore, manifold assumption requires that the labels should vary smoothly on the graph. In other words, if two examples are connected by a strong edge, they tend to share similar labels. Under this assumption, many graph-based semi-supervised learning algorithms have been developed. Zhu et al. proposed Harmonic Functions (HF, [7]) and related it to random walks, electric networks, and spectral graph theory. Zhou et al. developed Local and Global Consistency (LGC, [8]), in which the smoothness of labels are defined by the normalized *graph Laplacian*. Moreover, Spectral graph partitioning [9] formulates SSL as a graph cut problem, which aims to find a partitioning that minimizes the defined objective function. Wang et al. proposed Linear Neighborhood Propagation (LNP, [10]) that assumes that each data point in the graph can be optimally reconstructed by its neighbors. By introducing the manifold regularizer, Belkin et al. proposed the Laplacian Support Vector Machines (LapSVM) and Laplacian Regularized Least Squares (LapRLS). The idea of manifold regularization was successfully adapted to multi-label classification by multiview vector-valued manifold regularization (MV$^3$MR, [11]) and manifold regularized multitask learning (MRMTL, [12]) algorithms. Other typical manifold assumption-based approaches include AnchorGraph [13], Graph Transduction via Alternative Minimization (GTAM, [14]), and Label Propagation through Sparse Neighborhood (LPSN, [15]), etc. In recent years, some hypergraph-based manifold learning algorithms were developed and adopted to solve the critical

* Corresponding author.
*E-mail address:* jieyang@sjtu.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.neucom.2014.02.053
0925-2312/

problems in computer vision, such as image classification [16–18] and cartoon animation [19].

However, the graph established in the methods above only contains nonnegative edge weights. That is, only the similarities between examples are considered for classification, and the dissimilarity information is ignored. However, we believe that the dissimilarity information is important for improving the discriminative ability of semi-supervised classifiers. Therefore, this paper aims to design a novel semi-supervised classifier that incorporates both similarity and dissimilarity constraints between examples. In contrast to the traditional graph-based methods which require edge weights to be nonnegative, the weights in our algorithm are in the range $[-1,1]$. The positive weights representing "must-links" describe how similar the two connected examples are, and the negative weights standing for "cannot-links" evaluate the dissimilarity between the pairwise examples.

Actually, pairwise constraints including "must-links" and "cannot-links" have been widely adopted by various constrained clustering [20–22], dimensional reduction [23] and metric learning algorithms [24,25]. However, they are seldom employed to solve the semi-supervised classification problems. In this paper, pairwise constraints are adopted in order to improve the performance of traditional graph-based SSL algorithms, and the proposed classifier is named as Constrained Semi-supervised Classifier (CSSC). The most relevant work is [26], which also incorporates the dissimilarity into the framework of manifold regularization. However, the negative edges in this method should be manually generated among the unlabeled examples, which is different from CSSC that automatically constructs the graph of signed edges without any manual assistance.

The main contributions of this paper are summarized below:

1. A novel semi-supervised classification algorithm is proposed by incorporating both similarity and dissimilarity constraints.
2. The graph is built via similarity/dissimilarity propagation, in which the constraints imbalance is particularly considered.
3. A convex regularization framework is developed, so that the obtained solution is globally optimal.

The remainder of this paper is organized as follows: Section 2 constructs the signed graph with positive and negative edge weights. Section 3 derives the regularization framework of CSSC based on the established graph. We prove the convexity of the proposed model in Section 4, and present the empirical validations of CSSC and other experimental results in Section 5. Finally, a conclusion is drawn in Section 6.

## 2. Graph construction

For convenience, some important notations used in the rest of the paper are listed in Table 1. Given $l$ labeled examples $\mathcal{L}=\{(\mathbf{x}_1,\mathbf{Y}_1),(\mathbf{x}_2,\mathbf{Y}_2),\ldots,(\mathbf{x}_l,\mathbf{Y}_l)\}\in\mathbb{R}^d\times\mathbb{R}^C$ and $u$ unlabeled examples $\mathcal{U}=\{(\mathbf{x}_{l+1},\mathbf{Y}_{l+1}),(\mathbf{x}_{l+2},\mathbf{Y}_{l+2}),\ldots,(\mathbf{x}_n,\mathbf{Y}_n)\}\in\mathbb{R}^d\times\mathbb{R}^C$ $(n=l+u)$ drawn from the same distribution, the task of SSL is to propagate the labels $\{\mathbf{Y}_i\}_{i=1}^{l}\in\mathbb{R}^{1\times C}$ in $\mathcal{L}$, to the unknown labels $\{\mathbf{Y}_i\}_{i=l+1}^{l+u}\in\mathbb{R}^{1\times C}$ in $\mathcal{U}$. Here $C$ is the total number of classes. Then the $c'$-th $(1\le c'\le C)$ element of label vector $\{\mathbf{Y}_i\}_{i=1}^{n}$ is defined as $(\mathbf{Y}_i)_{c'}=1$ if $\mathbf{x}_i$ belongs to the $c'$-th class, and $(\mathbf{Y}_i)_{c'}=0$ otherwise. Consequently, a graph $\mathcal{G}=\langle\mathcal{V},\mathcal{E}\rangle$ can be built where $\mathcal{V}$ is the vertex set composed of all the elements in $\mathcal{L}\bigcup\mathcal{U}$, and $\mathcal{E}$ is the edge set describing the similarity/dissimilarity between pairs of examples.

Table 1
Important notations used in this paper.

| Notation | Description | Notation | Description |
|---|---|---|---|
| $\mathbf{x}_i$ | The $i$th example | $\widehat{\mathbf{W}}$ | The adjacency matrix of $\mathcal{G}$ |
| $\mathbf{Y}_i$ | The label vector of $\mathbf{x}_i$ | $\mathbf{W}$ | The matrix recording the values of $\lvert(\widehat{\mathbf{W}})_{ij}\rvert$ |
| $K$ | The number of neighborhoods | $\mathbf{S}$ | Indicator matrix |
| $\tilde{\mathcal{G}}$ | Unsigned graph | $\mathbf{I}$ | Identity matrix |
| $\mathcal{G}$ | Signed graph | $\mathbf{F}$ | The obtained label matrix |
| $\mathbf{M}$ | The adjacency matrix of $\tilde{\mathcal{G}}$ | $\mathbf{H}$ | Hessian matrix |
| $\overline{\mathbf{M}}$ | Normalized $\mathbf{M}$ | $\tilde{\mathbf{L}}$ | Generalized graph Laplacian |

Traditionally, there are two ways to compute the nonnegative edge weight between two examples. One is the 0–1 weight, which simply takes the binary value from $\{0,1\}$ to indicate whether an edge exists between the two vertices or not. The other is to use the RBF kernel, which produces a real value within $[0,1]$, to represent the similarity of examples. However, these two methods only generate nonnegative weights, so they are not suitable to represent both "must-link" and "cannot-link" constraints. Below we introduce a two-step approach called "balanced constraints propagation" to explicitly construct a graph with edge weights in the range of $[-1,1]$.

In the first step, we establish a traditional unsigned graph $\tilde{\mathcal{G}}$ with nonnegative edge weights. $K$ nearest neighborhood ($K$-NN) graph is adopted because sparse graph usually leads to better performance [27]. The edge weights $m_{ij}$ $(1\le i,j\le n)$ of $\tilde{G}$ are computed by using the RBF kernel $m_{ij}=\exp(\|\mathbf{x}_i-\mathbf{x}_j\|^2/(2\sigma^2))$ ($\sigma$ is the kernel width), and thus we have the adjacency matrix $\mathbf{M}$ of $\tilde{\mathcal{G}}$ with $(\mathbf{M})_{ij}=m_{ij}$. Moreover, we define a diagonal matrix $\tilde{\mathbf{D}}$ in which the $i$-th diagonal element $\tilde{d}_{ii}$ is calculated as $\tilde{d}_{ii}=\sum_{j=1}^{n}m_{ij}$. Therefore, $\mathbf{M}$ can be further normalized by $\overline{\mathbf{M}}=\tilde{\mathbf{D}}^{-1/2}\mathbf{M}\tilde{\mathbf{D}}^{-1/2}$, so that the elements $\overline{m}_{ij}$ of $\overline{\mathbf{M}}$ satisfy $\sum_{j=1}^{n}\overline{m}_{ij}=1$ for $1\le i\le n$ [8].

In the second step, we aim to build a signed graph $\mathcal{G}$ that incorporates both positive and negative constraints based on $\overline{\mathbf{M}}$. It is obvious that $l(l-1)/2$ definitely correct constraints are already available based on the $l$ labeled examples, and they are recorded by the similarity set $\mathcal{S}$ and dissimilarity set $\mathcal{D}$:

$$\mathcal{S}=\{(\mathbf{x}_i,\mathbf{x}_j)|\mathbf{x}_i \text{ and } \mathbf{x}_j \text{ come from the same class}\}$$

$$\mathcal{D}=\{(\mathbf{x}_i,\mathbf{x}_j)|\mathbf{x}_i \text{ and } \mathbf{x}_j \text{ come from different classes}\}.$$

The aim of our proposal is to propagate the limited available elements in $\mathcal{S}$ and $\mathcal{D}$, to the remaining pairs of examples. This process is called "balanced constraints propagation".

To facilitate the mathematical manipulations, we use the matrix $\widehat{\mathbf{W}}^{(0)}\in\mathbb{R}^{n\times n}$ to encode the pairwise constraints in $\mathcal{S}$ and $\mathcal{D}$, namely

$$(\widehat{\mathbf{W}}^{(0)})_{ij}=\widehat{\mathbf{W}}_{ij}^{(0)}=\begin{cases}1 & (\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S} \text{ or } i=j\\ -\gamma & (\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}\\ 0 & (\mathbf{x}_i,\mathbf{x}_j) \text{ is not specified}\end{cases}. \quad (1)$$

In (1), $\gamma=a/b$ where $a=|\mathcal{S}|+n$, $b=|\mathcal{D}|$ and $|\cdot|$ represents the size of a set. Note that we set $\widehat{\omega}_{ij}^{(0)}=-\gamma$ rather than $-1$ to avoid the constraints imbalance. In fact, if $|\mathcal{D}|$ is very small, the element "1" in $\widehat{\mathbf{W}}^{(0)}$ will be much more than the element "0" because all the diagonal elements are 1s, thus the "must-links" may dominate the propagation process, which significantly weakens the propagation "strength" of the "cannot-links". Alternatively, more negative constraints can be added to $\widehat{\mathbf{W}}^{(0)}$ based on the prior knowledge, so the negative constraints can significantly outnumber the positive constraints sometimes. Therefore, $\gamma$ assigns larger weight

on the minority constraint between "must-link" and "cannot-link" so that their "strengths" are comparable. This is the main difference between the proposed "balanced constraints propagation" and the existing "similarity propagation" [24]. Liu et al. [24] only propagates the positive edges, while our method transmits both positive edges and negative edges in a balanced way. The necessity of handling the imbalanced constrains is demonstrated in Section 5.1.

Similar to [8] and [10], the propagation process can be simulated by an iterative process. In the $t$-th iteration, the similarities between the $i$-th example and other examples can be modeled as a convex combination of two factors: one is its initial similarities with other examples denoted by $\widehat{\mathbf{W}}_i^{(0)}$ ($\widehat{\mathbf{W}}_i^{(0)}$ means the $i$-th row of matrix $\widehat{\mathbf{W}}^{(0)}$), and the other is the influence of similarities of other examples in the previous iteration. Therefore, we have

$$\widehat{\mathbf{W}}_i^{(t)} = (1-\alpha)\widehat{\mathbf{W}}_i^{(0)} + \alpha \sum_{j=1}^{n} \overline{m}_{ij} \widehat{\mathbf{W}}_j^{(t-1)}, \tag{2}$$

in which $\alpha \in (0,1)$ is a parameter governing the relative weight between $\widehat{\mathbf{W}}_i^{(0)}$ and the weighted sum of $\widehat{\mathbf{W}}_j^{(t-1)}$ $(1 \le j \le n)$. (2) can be presented concisely by

$$\widehat{\mathbf{W}}^{(t)} = (1-\alpha)\widehat{\mathbf{W}}^{(0)} + \alpha\overline{\mathbf{M}}\widehat{\mathbf{W}}^{(t-1)}. \tag{3}$$

Therefore, by iteratively using (3), we obtain

$$\widehat{\mathbf{W}}^{(t)} (\alpha\overline{\mathbf{M}})^t \widehat{\mathbf{W}}^{(0)} + (1-\alpha) \sum_{i=0}^{t-1} (\alpha\overline{\mathbf{M}})^i \widehat{\mathbf{W}}^{(0)}. \tag{4}$$

Note that $\overline{\mathbf{M}}$ is similar to the stochastic matrix $\mathbf{\Omega} = \tilde{\mathbf{D}}^{-1}\mathbf{M} = \tilde{\mathbf{D}}^{-1/2}\overline{\mathbf{M}}\tilde{\mathbf{D}}^{1/2}$, so all the eigenvalues of $\overline{\mathbf{M}}$ are within $[-1,1]$ [8]. Consequently, from *Perron–Frobenius theorem* [28] we know that when $t \to \infty$, (4) will converge to

$$\widehat{\mathbf{W}} = (1-\alpha)(\mathbf{I} - \alpha\overline{\mathbf{M}})^{-1}\widehat{\mathbf{W}}^{(0)}. \tag{5}$$

$\widehat{\mathbf{W}}$ is a symmetrical matrix in which the elements are in the range $[-1,1]$. Therefore, it incorporates both similarity and dissimilarity constraints. After zeroing out the entries smaller than $\varepsilon$ and implementing normalization (the normalization process is similar to $\mathbf{M}$ in the first step), we finally obtain the adjacency matrix $\widehat{\mathbf{W}}$ of signed graph $\mathcal{G}$.

## 3. Regularization framework

As mentioned in the Introduction, graph-based SSL requires the labels of examples varying smoothly on the graph. This is usually achieved through the regularization techniques, such as [8–10]. However, traditional regularization framework for the unsigned graph is not applicable to the signed graph, so a novel smoothness regularizer is to be developed.

To facilitate the following explanations, we use the nonnegative $\mathbf{W} \in \mathbb{R}^{n\times n}$ to store the absolute values of $\widehat{\mathbf{W}}$'s elements, and use the indicator matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ to record the signs of the corresponding elements in $\widehat{\mathbf{W}}$. That is, $(\mathbf{W})_{ij} = \omega_{ij} = |\omega_{ij}^{\wedge}|$ ($\omega_{ij}^{\wedge}$ is the $(i,j)$-th element of $\widehat{\mathbf{W}}$), and $(\mathbf{S})_{ij} = s_{ij} = 1, 0, -1$ if $\omega_{ij}^{\wedge}$ is positive, zero and negative, respectively. Besides, we stack the initial label vectors $\{\mathbf{Y}_i\}_{i=1}^n$ into a matrix $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_n^T)^T \in \mathbb{R}^{n\times C}$, and define the final label matrix by $\mathbf{F} = (\mathbf{F}_1^T, \mathbf{F}_2^T, \ldots, \mathbf{F}_n^T)^T$ in which $\{\mathbf{F}_i\}_{i=1}^n \in \mathbb{R}^{1\times n}$ are final soft label vectors with the elements taking values from a real range $[0,1]$. $\mathbf{F}_i$ determines the class of $\mathbf{x}_i$ as $c_i = \arg\max_{\mathbf{F}_{ic'}}$.

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by a "must-link", we have $\omega_{ij} = \omega_{ij}^{\wedge}$, and the smoothness regularizer is the same as the one utilized by other existing methods [7,10], namely

$$Q_1(\mathbf{F}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{ij}\|\mathbf{F}_i - \mathbf{F}_j\|^2. \tag{6}$$

Eq. (6) suggests that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are very similar, their labels $\mathbf{F}_i$ and $\mathbf{F}_j$ should not differ significantly.

If $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by a "cannot-link", then $\omega_{ij} = -\omega_{ij}^{\wedge}$, and the smoothness regularizer is defined by

$$Q_2(\mathbf{F}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{ij}\|\mathbf{1}^T - \mathbf{F}_i - \mathbf{F}_j\|^2, \tag{7}$$

where $\mathbf{1} = (1,1,\ldots,1)^T \in \mathbb{R}^C$ is a $C$-dimensional all-one vector. Because all the entries in $\mathbf{F}_i$ and $\mathbf{F}_j$ are within $[0,1]$, (7) indicates that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are very dissimilar, then their labels should be completely different. For example, if the edge weight between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $\omega_{ij}^{\wedge} = -1$, these two examples definitely belong to different classes. If $\mathbf{x}_i$'s label vector $\mathbf{F}_i = (0.1, 0.7)$, the desirable label of $\mathbf{x}_j$ should be $\mathbf{F}_j = (0.9, 0.3)$.

By using the matrix $\mathbf{S}$ defined above, (6) and (7) can be cast into a unified framework:

$$Q(\mathbf{F}) = \frac{1}{2}\sum_i\sum_j \omega_{ij}\|\frac{1}{2}(1-s_{ij})\cdot\mathbf{1}^T + s_{ij}\mathbf{F}_i - \mathbf{F}_j\|^2. \tag{8}$$

We observe that if the traditional unsigned graph is adopted (*i.e.* $s_{ij} = 1$ for $1 \le i, j \le n$), (8) will degenerate into the existing smoothness regularizer [7], which has the same formation as (6). However, (8) can also handle the negative weights, which is the main innovation of this paper. By denoting $(1-s_{ij})/2 = p_{ij}$ in (8), the complete regularization framework of CSSC is expressed as

$$\min_{\mathbf{F}} H(\mathbf{F}) = \frac{1}{2}\left[\sum_i\sum_j \omega_{ij}\|p_{ij}\cdot\mathbf{1}^T + s_{ij}\mathbf{F}_i - \mathbf{F}_j\|^2 + \beta\sum_i\|\mathbf{F}_i - \mathbf{Y}_i\|^2\right], \tag{9}$$

in which the first term in the bracket is the proposed *smoothness term*, and the second *fitting term* means that the classification results should be well consistent with the examples' initial states. The regularization parameter $\beta \in (0,1)$ controls the trade-off between smoothness term and fitting term.

By computing the derivative of $H(\mathbf{F})$ with respect to $\mathbf{F}$, and enforcing the result to $\mathbf{0}$, we have

$$\frac{\partial H}{\partial \mathbf{F}} = 2\tilde{\mathbf{L}}\mathbf{F} + \mathbf{v}\bullet\mathbf{E} + \beta(\mathbf{F} - \mathbf{Y}) = \mathbf{0}. \tag{10}$$

In (10), $\mathbf{E} \in \mathbb{R}^{n\times C}$ represents an all-one matrix, and $\mathbf{v} \in \mathbb{R}^{1\times n}$ is an $n$-dimensional row vector with the $i$-th element represented by $\mathbf{v}_i = \sum_j (s_{ij}-1)\omega_{ij}p_{ij}$. $\mathbf{v}\bullet\mathbf{E}$ defines a matrix of the same size as $\mathbf{E}$, of which the $i$-th row is $\mathbf{v}_i \cdot \mathbf{E}_i$. $\tilde{\mathbf{L}}$ has the formation as

$$\tilde{\mathbf{L}} = \begin{pmatrix} \sum_j \omega_{1j} & -\omega_{12}s_{12} & \cdots & -\omega_{1n}s_{1n} \\ -\omega_{21}s_{21} & \sum_j \omega_{2j} & \cdots & -\omega_{2n}s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\omega_{n1}s_{n1} & -\omega_{n2}s_{n2} & \cdots & \sum_j \omega_{nj} \end{pmatrix} = \mathbf{L} \odot \mathbf{S}, \tag{11}$$

in which $\mathbf{L} = \mathbf{D} - \mathbf{W}$ (the definition of $\mathbf{D}$ is identical to $\tilde{\mathbf{D}}$) is the traditional *graph Laplacian* [29], and "$\odot$" is the Hadamard product. Note that if $\mathbf{S}$ is an all-one matrix, $\tilde{\mathbf{L}}$ will degenerate into the traditional *graph Laplacian*. Therefore, $\tilde{\mathbf{L}}$ can be regarded as the generalized *graph Laplacian* that is also applicable to the signed graph.

The closed-form solution of (9) can be obtained by solving (10), which is

$$\mathbf{F} = (2\tilde{\mathbf{L}} + \beta\mathbf{I})^{-1}(\beta\mathbf{Y} - \mathbf{v}\bullet\mathbf{E}). \tag{12}$$

$\mathbf{I} \in \mathbb{R}^{n\times n}$ in (12) is an identity matrix. Because $2\tilde{\mathbf{L}} + \beta\mathbf{I}$ is always invertible, so (9) is guaranteed to have a meaningful solution. Note

that the only computational burden for obtaining $\mathbf{F}$ in (12) is to finding the inverse of $2\tilde{\mathbf{L}}+\beta\mathbf{I}$, which usually has a cubic time complexity $O(n^3)$. However, this complexity can be decreased by adopting some approximation techniques, *e.g.* Nystrom approximation [30].

## 4. Convexity analysis

It is worth pointing out that the regularization framework (9) of CSSC is convex. By calculating the Hessian matrix $\mathbf{H}$ of (9), we have

$$\mathbf{H}=2\tilde{\mathbf{L}}+\beta\mathbf{I}, \tag{13}$$

in which $\tilde{\mathbf{L}}$ is defined by (11). It is obvious that $\mathbf{H}$ is diagonally dominant, so it is a positive definite matrix. Therefore, (9) defines a convex optimization problem, and its solution $\mathbf{F}$ is guaranteed to be globally optimal.

## 5. Experimental results

In this section, we validate the proposed CSSC on two synthetic datasets, and compare CSSC with some state-of-the-art SSL algorithms on a number of real-world computer vision collections. HF [7], LNP [10], LapRLS [31], LapSVM [31], S4VM with linear kernel [6] and S4VM with RBF kernel [6] serve as the baselines for evaluating the performance of CSSC. In all the experiments below, we choose $\beta=100$, $\alpha=0.9$ for CSSC, $C_1=100$, $C_2=0.1$ for S4VM, and $\gamma_A=\gamma_I=1$ for LapRLS and LapSVM. Other parameters in the compared algorithms are also properly tuned to obtain the optimal performances.

### 5.1. Toy data

Two synthetic datasets, *Ring&Triangular* and *DoubleMoon*, are adopted to visualize the results of graph construction and classification performance of CSSC. In *Ring&Triangular*, a triangular representing the negative class is surrounded by a ring that forms the positive class (see Fig. 1(a)). Both the outer ring and inner triangular are centered at $(0,0)$, and the radius of the ring is 2. The difficulty of this dataset is that the intersection area between triangular and ring may cause the mutual transmission of positive and negative labels. *DoubleMoon* consists of 640 examples, which are equally divided into two moons (see Fig. 1(b)). This dataset is contaminated by the Gaussian noise with standard variance 0.15, and each class has only one labeled example.

In these two datasets, the available negative constraints in $\mathcal{D}$ is very sparse compared with the "1" elements in $\widehat{\mathbf{W}}^{(0)}$, so the $\gamma$ in (1) plays an important role in handling the constraints imbalance. We set $K=10$, $\sigma=2$ in *Ring&Triangular*, and choose $K=5$, $\sigma=2$ for *DoubleMoon*. Fig. 1(c) and (e) respectively plot the established graphs for *Ring&Triangular* and *DoubleMoon*, by simply setting $\widehat{\omega}_{ij}^{(0)}=-1$ when $(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}$. Fig. 1(d) and (f) is the results of balanced constraints propagation by adopting the adaptive $\gamma$. It can be observed that more "cannot-link" edges are generated by considering the constraints imbalance, which is very important to deal with the ambiguous points in the intersection regions of different classes.

In Fig. 1(g), a fraction of negative labels are erroneously propagated to the outer ring because the "cannot-link" edges are not sufficient to prevent the negative labels being transmitted to

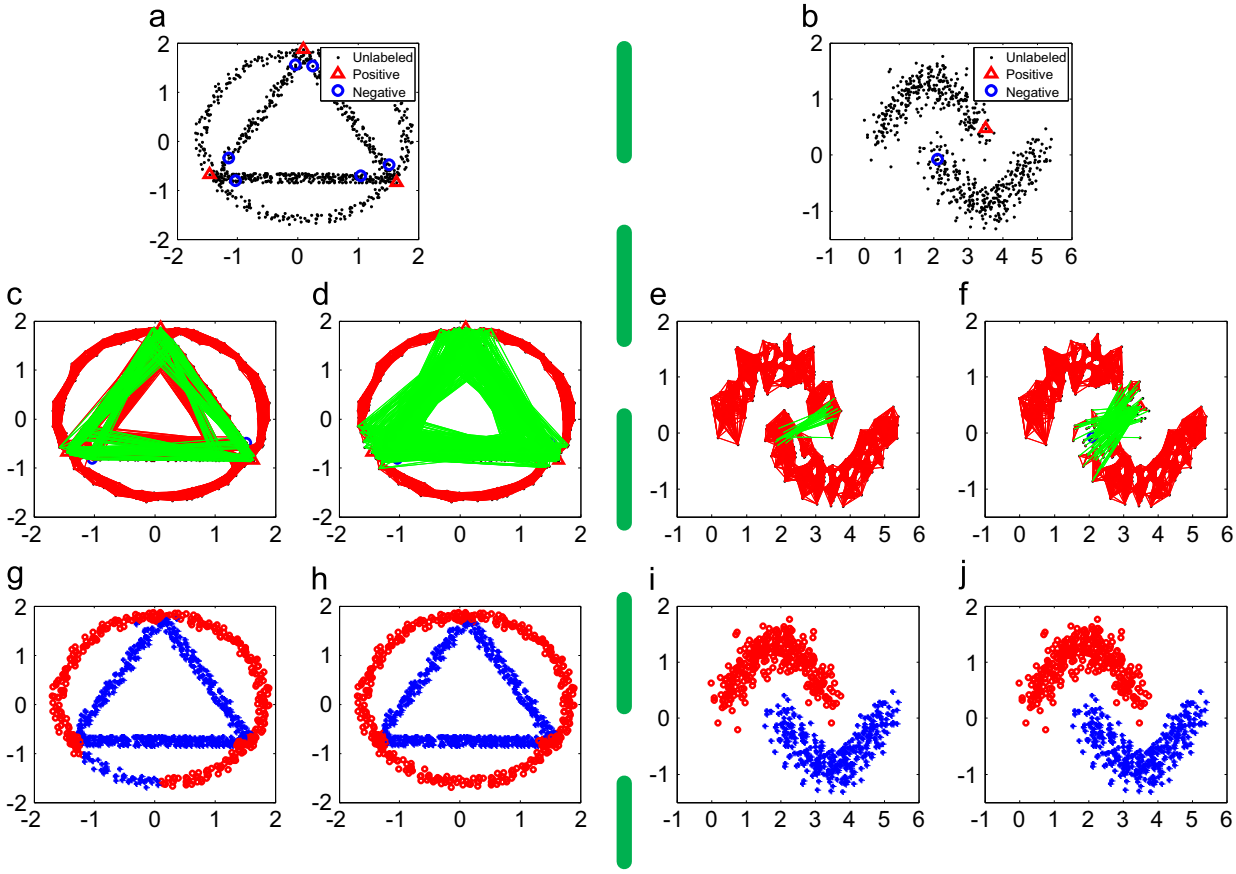



**Fig. 1.** Performances of CSSC on *Ring&Triangular* and *DoubleMoon* datasets, in which the left part is *Ring&Triangular* and the right part is *DoubleMoon*. The first row [(a) and (b)] illustrates the initial labeled and unlabeled examples. The second row [(c)–(f)] shows the constructed sighed graphs with and without considering the constraints imbalance, in which the red lines are "must-links" and the green lines are "cannot-links". The third row [(g)-(j)] presents the final classification results. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

**Table 2**
Classification accuracies (%) of algorithms on four benchmark datasets. (The reported results are of the format "accuracy ± standard variation".)

| | *Iris* | | *Wine* | | *Seeds* | | *USPS* | |
|---|---|---|---|---|---|---|---|---|
| Algorithms | $l=12$ | $l=24$ | $l=12$ | $l=24$ | $l=12$ | $l=24$ | $l=10$ | $l=100$ |
| HF | 0.532 ± 0.179 | 0.832 ± 0.064 | 0.693 ± 0.174 | 0.818 ± 0.121 | 0.371 ± 0.063 | 0.414 ± 0.005 | 0.802 ± 0.005 | 0.813 ± 0.002 |
| LNP | 0.807 ± 0.110 | 0.854 ± 0.107 | 0.698 ± 0.060 | 0.688 ± 0.089 | 0.803 ± 0.081 | 0.841 ± 0.077 | 0.776 ± 0.058 | 0.799 ± 0.004 |
| LapRLS | 0.903 ± 0.001 | 0.939 ± 0.001 | 0.903 ± 0.001 | 0.931 ± 0.001 | 0.892 ± 0.001 | 0.895 ± 0.001 | 0.657 ± 0.002 | 0.774 ± 0.001 |
| LapSVM | 0.824 ± 0.001 | 0.898 ± 0.001 | 0.861 ± 0.001 | 0.893 ± 0.001 | 0.870 ± 0.001 | 0.897 ± 0.001 | 0.673 ± 0.002 | 0.873 ± 0.001 |
| S4VM (RBF) | 0.913 ± 0.022 | 0.919 ± 0.023 | 0.912 ± 0.041 | 0.927 ± 0.010 | 0.881 ± 0.041 | 0.881 ± 0.010 | 0.685 ± 0.066 | 0.742 ± 0.029 |
| S4VM (linear) | 0.756 ± 0.030 | 0.760 ± 0.053 | **0.936 ± 0.024** | 0.923 ± 0.037 | 0.838 ± 0.049 | 0.861 ± 0.031 | 0.677 ± 0.068 | 0.688 ± 0.036 |
| CSSC | **0.945 ± 0.029** | **0.955 ± 0.023** | 0.933 ± 0.016 | **0.945 ± 0.011** | **0.904 ± 0.013** | **0.909 ± 0.011** | **0.844 ± 0.050** | **0.953 ± 0.007** |

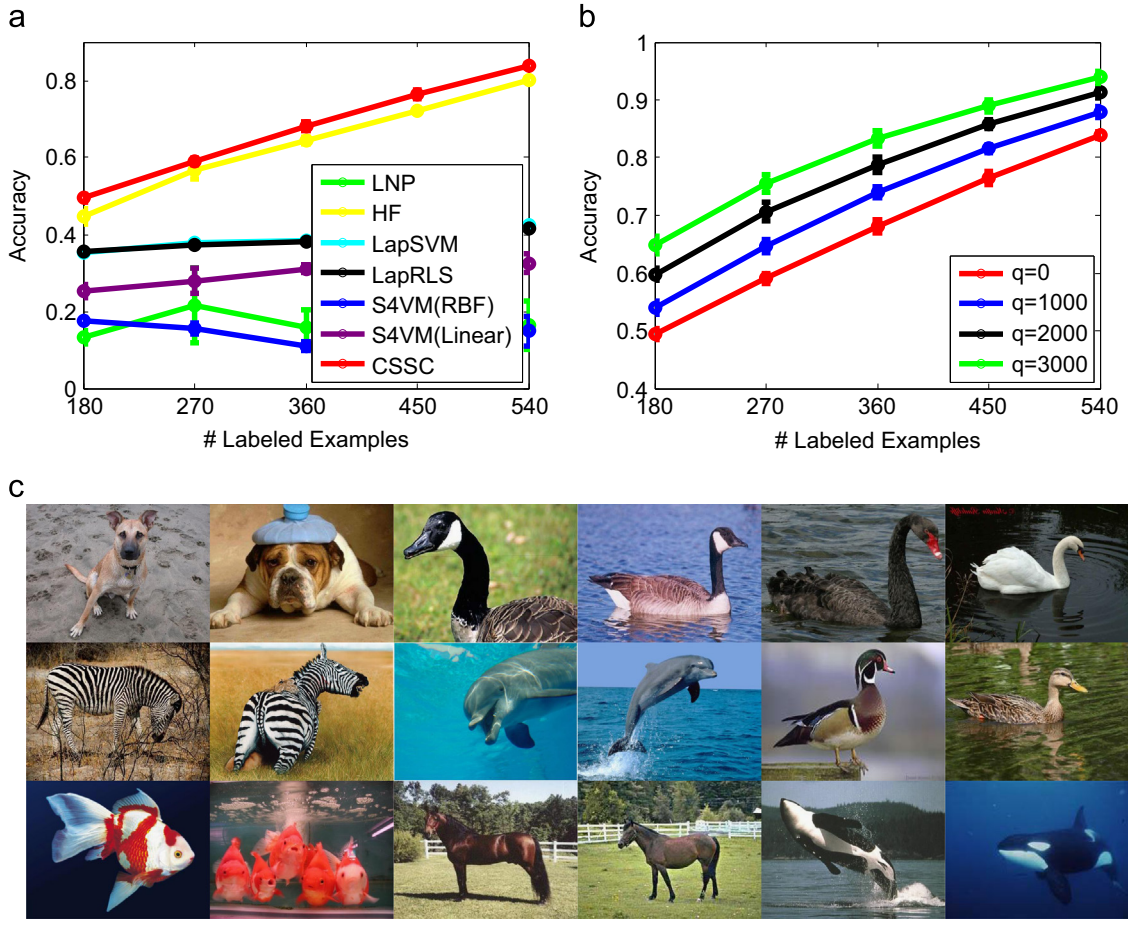



**Fig. 2.** Experimental results on *Caltech 256* dataset. (a) The performances of CSSC and baselines are compared. (b) The performances of the proposed method when different numbers of extra constrains are incorporated are shown. (c) Some representative examples are shown.

the positive class. As a result, CSSC is confused at the intersection region and fails to achieve the perfect classification. Comparatively, a better result is obtained in Fig. 1(h), for the reason that more negative edges are generated in the intersection regions. Therefore, the effectiveness of the proposed method for preventing the imbalanced constraints is validated. For *DoubleRing*, perfect performances are achieved on both graphs (see Fig. 1 (i) and (j)), because the obtained negative weights in Fig. 1(e) are sufficient to make a clear discrimination between the two classes.

Moreover, the experimental results in Fig. 1(g)–(j) reveal that CSSC is able to successfully discover the geometric structure of the data distribution. This is another rationale why the proposed method can achieve encouraging performance.

### 5.2. Real benchmark data

In this section, we use three UCI datasets[1] including *Iris*, *Wine* and *Seeds*, and *USPS* dataset in Chapelle's book [32] to compare the performances of CSSC and all the baselines. There are totally 150, 178, 210 and 1500 examples in the above four datasets, which are attributed to 4 classes, 13 classes, 7 classes and 2 classes, respectively. Particularly, *USPS* is a handwritten digit recognition collection which contains 150 images (examples) of each of the ten digits. The digits "2" and "5" form the positive class, and all the other digits constitute

[1] http://archive.ics.uci.edu/ml/

the negative class. Therefore, the sizes of examples are imbalanced and the size ratio of two classes is 1:4. The feature of every digit image is represented by a 241-dimensional vector with elements representing the pixel-wise gray values.

In each dataset, we implement all the algorithms under different $l$ as listed in Table 2, and the reported results are averaged over 20 independent implementations for a certain $l$. For each implementation, the labeled sets are randomly chosen from the entire dataset, and at least one example is guaranteed in every class.

The RBF kernel width $\sigma$ is set to 1 in *Iris*, *Wine* and *Seeds*, and optimally tuned to 5 in *USPS*. The parameter $K$ in all the four datasets is adjusted to 10, and we zero out the weak edges in $\widehat{\mathbf{W}}$ with weights smaller than $\varepsilon = 0.001$. In Table 2, the best results are marked in bold, which reveal that the proposed CSSC outperforms other baselines overall. This is because CSSC utilizes not only "must-links", but also the "cannot-links" that are not incorporated by the traditional SSL algorithms. Therefore, CSSC shows stronger discriminative ability than other baselines. Moreover, we note that CSSC achieves very impressive performance on the *USPS* dataset, which demonstrates that CSSC can perfectly handle the data imbalanced situations.

### 5.3. Image classification

Classifying objects under natural scenes is a challenging problem because of various viewing angles, complicated background and unexpected noises. We extract a subset from the original *Caltech 256* dataset [33], to test the ability of CSSC and baselines on classifying nine animals, *i.e.* dog, goose, swan, zebra, dolphin, duck, goldfish, horse, and whale. Each animal has 80 images in our subset, and a few typical examples are illustrated in Fig. 2(c). It is shown that even two examples belonging to the same class may look very differently, which demonstrates the challenges to the accurate classification. Images are characterized by a concatenation [34] of four image descriptors, including PHOG [35], SIFT Descriptor [36], Region Covariance [37], and LBP [38]. The parameters for graph construction are $K=10$ and $\sigma=2$. We zero out the elements in $\mathbf{W}$ of which the absolute values are smaller than $\varepsilon=0.001$. The accuracies of algorithms vs. different sizes of labeled set are evaluated, and Fig. 2(a) shows the result. It is observed that CSSC always achieves higher accuracy compared with the baselines with the best performance of 80% accuracy rate.

As discussed in Section 2, certain prior knowledge can be used to further boost the classification performance of CSSC. To validate this argument, we randomly add a number of extra pairwise constraints according to the groundtruth. We use these manually added constraints, along with the constraints in $\mathcal{S}$ and $\mathcal{D}$, to carry out the classifications on these four animals again. We gradually add the extra constraints with $q=1000$, 2000 and 3000 ($q$ is the amount of extra constraints), and the accuracy vs. the increasing pairwise constraints is plotted in Fig. 2(b). It can be observed that the accuracy can be improved significantly by incorporating more constraints. Therefore, the effectiveness of integrating "must-link" and "cannot-link" in CSSC is demonstrated.

### 5.4. Violent behavior detection

The proposed CSSC can also be applied to processing video data. Detection of violent behaviors, such as fighting and robbery, is an important application of intelligent surveillance. The *HockeyFight* dataset contains 1000 video clips collected in ice hockey competitions, of which 500 clips contain fight behavior and 500 clips are non-fight sequences. Our purpose is to accurately identify the fighting clips. Fig. 3 shows a few typical frames extracted from the fighting and non-fighting clips. Similar to [39], we adopt the space–time interest points (STIP) and motion SIFT (MoSIFT) as action descriptors, and use the Bag-of-Words (BoW) approach to represent each video clip as a histogram over 100 visual words. Therefore, every clip in the dataset is characterized by a 100-dimensional feature vector.

For fair comparison, a 5-NN graph with $\sigma=2$ is built for all the graph-based comparators to evaluate the relevant classification performances. The accuracies of various algorithms under $l=40$, 80, 120 and 160 are particularly observed, and all the algorithms are implemented 20 times independently for each $l$. The accuracies and standard variances are plotted in Fig. 4, which indicates that the performances of all the compared methods can be substantially improved by increasing $l$. Comparatively, S4VM with RBF kernel and the proposed CSSC perform better than other methods. Particularly, CSSC can achieve more than 85% accuracy when $l$ is larger than 120, which is a very encouraging result for fight behavior detection.

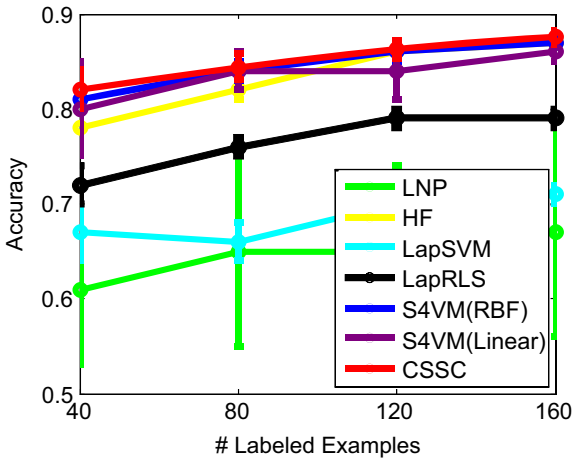



**Fig. 4.** Performance comparison on fight behavior detection based on *HockeyFight* dataset.

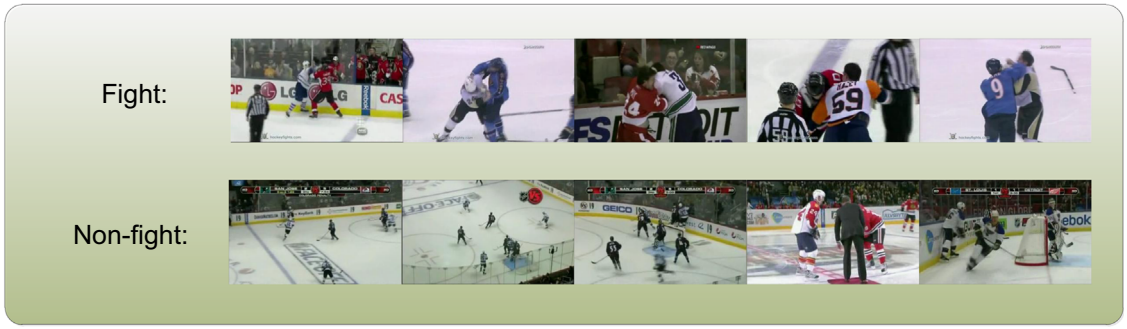



**Fig. 3.** Example frames containing fight and non-fight behaviors.

Moreover, we observe that the standard variance of CSSC is not large when changing $l$, so the performance of proposed CSSC is not sensitive to the choice of initial labeled examples.

## 6. Conclusion

This paper proposed a graph-based SSL algorithm called Constrained Semi-supervised Classifier (CSSC). In CSSC, we established a signed graph considering both similarity and dissimilarity constraints between pairs of examples. Then a novel regularization framework was developed to adapt the traditional smoothness regularizer to the signed graph. A specific technique was developed to avoid the constraints imbalance, which significantly improves the classification accuracy as revealed by comprehensive empirical studies. Moreover, CSSC is demonstrated to be very effective in dealing with several challenging computer vision tasks, such as image classification and fight behavior detection.

A number of open issues remain. Firstly, current approach for graph construction is computationally expensive, so it cannot handle the huge datasets; secondly, the parameters $\sigma$ and $K$ are adjusted empirically, so a systematic way for choosing the optimal parameters is to be investigated; and thirdly, CSSC will be evaluated more broadly on more practical problems.

## Acknowledgements

This research is supported by NSFC, China (No. 6127325 861375048), and Ph.D. Programs Foundation of Ministry of Education of China (No. 20120073110018).

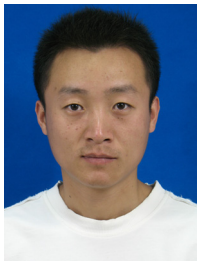
## References

[1] X. Zhu, B. Goldberg, Introduction to Semi-Supervised Learning, 2009.
[2] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the International Conference on Machine Learning, 1999, pp. 200–209.
[3] X. Tian, G. Gasso, S. Canu, A multiple kernel framework for inductive semi-supervised SVM learning, Neurocomputing 90 (2012) 46–58.
[4] G. Fung, O. Mangasarian, Semi-supervised support vector machines for unlabeled data classification, Optim. Methods Softw. 15 (2001) 29–44.
[5] H. Xue, S. Chen, Q. Yang, Structural regularized support vector machine: a framework for structural large margin classifier, IEEE Trans. Neural Netw. 22 (4) (2011) 573–587.
[6] Y. Li, Z. Zhou, Towards making unlabeled data never hurt, in: Proceedings of the International Conference on Machine Learning, 2011, pp. 1081–1088.
[7] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: Proceedings of the International Conference on Machine Learning, Washington, DC, USA, 2003, pp. 912–919.
[8] D. Zhou, O. Bousquet, Learning with local and global consistency, in: Advances in Neural Information Processing Systems, Vancouver, Canada, 2003, pp. 321–328.
[9] T. Joachims, Transductive learning via spectral graph partitioning, in: Proceedings of the International Conference on Machine Learning, 2003, pp. 290–297.
[10] J. Wang, F. Wang, C. Zhang, Linear neighborhood propagation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1600–1615.
[11] Y. Luo, D. Tao, C. Xu, H. Liu, Y. Wen, Multiview vector-valued manifold regularization for multilabel image classification, IEEE Trans. Neural Netw. Learn. Syst. 24 (2013) 709–722.
[12] Y. Luo, D. Tao, B. Geng, C. Xu, S. Maybank, Manifold regularized multi-task learning for semi-supervised multi-label image classification, IEEE Trans. Image Process. 22 (2013) 523–536.
[13] W. Liu, J. He, S. Chang, Large graph construction for scalable semi-supervised learning, in: Proceedings of the International Conference on Machine Learning, Haifa, Israel, 2010, pp. 679–686.
[14] J. Wang, T. Jebara, S. Chang, Graph transduction via alternating minimization, in: Proceedings of the International Conference on Machine Learning, Helsinki, Finland, 2008, pp. 1144–1151.
[15] F. Zang, J. Zhang, Label propagation through sparse neighborhood and its applications, Neurocomputing 97 (2012) 267–277.
[16] J. Yu, D. Tao, M. Wang, Adaptive hypergraph learning and its application in image classification, IEEE Trans. Image Process. 21 (7) (2012) 3262–3272.
[17] Y. Huang, Q. Liu, F. Lv, Y. Gong, D. Metaxas, Unsupervised image categorization by hypergraph partition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1266–1273.
[18] C. Hong, J. Yu, J. Li, X. Chen, Multi-view hypergraph learning by patch alignment framework, Neurocomputing 118 (22) (2013) 79–86.
[19] J. Yu, D. Tao, Modern Machine Learning Techniques and Their Applications in Cartoon Animation Research, John Wiley & Sons, 2013, Available at: 〈http://as.wiley.com/WileyCDA/WileyTitle/productCd-1118115147,subjectCd-CSC0.html〉.
[20] B. Soleymani, S. Bagheri, Kernel-based metric learning for semi-supervised clustering, Neurocomputing 73 (7) (2010) 1352–1361.
[21] M. Baghshah, S. Shouraki, Learning low-rank kernel matrices for constrained clustering, Neurocomputing 74 (12) (2011) 2201–2211.
[22] J. Yu, D. Liu, D. Tao, H. Seah, Complex object correspondence construction in two-dimensional animation, IEEE Trans. Image Process. 20 (11) (2011) 3257–3269.
[23] J. Yu, D. Tao, Y. Rui, J. Cheng, Pairwise constraints based multiview features fusion for scene classification, Pattern Recognit. 46 (2) (2013) 483–496.
[24] W. Liu, X. Tian, D. Tao, Constrained metric learning via distance gap maximization, in: The AAAI Conference on Artificial Intelligence (AAAI), Atlanta, USA, 2010.
[25] J. Yu, M. Wang, D. Tao, Semi-supervised multiview distance metric learning for cartoon synthesis, IEEE Trans. Image Process. 21 (11) (2012) 4636–4648.
[26] A. Goldberg, Z. Zhu, S. Wright, Dissimilarity in graph-based semi-supervised classification, in: International Conference on Artificial Intelligence and Statistics, 2007, pp. 155–162.
[27] T. Jebara, J. Wang, S. Chang, Graph construction and b-matching for semi-supervised learning, in: Proceedings of the International Conference on Machine Learning, 2009, pp. 441–448.
[28] G. Golub, C. Van Loan, Matrix Computations, 3rd, Johns Hopkins University Press, 1996, Available at: 〈http://www.cs.cornell.edu/courses/cs621/Books/GVL/index.htm〉.
[29] F. Chung, Spectral Graph Theory, AMS Bookstore, 1997, Available at: 〈http://www.math.ucsd.edu/~fan/research/revised.html〉.
[30] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral grouping using the Nystrom method, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2) (2004) 214–225.
[31] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[32] O. Chapelle, B. Scholkopf, A. Zien, Semi-Supervised Learning, vol. 2, MIT Press, Cambridge, MA, 2006.
[33] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Technical Report 7694, California Institute of Technology, 2007, URL 〈http://authors.library.caltech.edu/7694〉.
[34] T. Tommasi, F. Orabona, B. Caputo, Safety in numbers: Learning categories from few examples with multi model knowledge transfer, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3081–3088.
[35] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: 6th ACM International Conference on Image and Video Retrieval, 2007.
[36] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[37] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on Riemannian manifolds, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.
[38] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[39] E. Nievas, O. Suarez, G. García, R. Sukthankar, Violence detection in video using computer vision techniques, in: Computer Analysis of Images and Patterns, Springer, 2011, pp. 332–339. Available at: 〈http://link.springer.com/chapter/10.1007/978-3-642-23678-5_39〉.

**Chen Gong** received his B.Sc. degree from East China University of Science and Technology (ECUST) in 2010. Currently he is a Ph.D. candidate at Shanghai Jiao Tong University (SJTU) in the Institute of Image Processing and Pattern Recognition under the supervision of Professors Dacheng Tao and Jie Yang. His research interests mainly include machine learning, object detection and tracking. He has also served as reviewer for several international journals, such as IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Intelligent Transportation Systems, Neurocomputing, and Neural Processing Letters.

**Keren Fu** received the B.Sc. degree in automation from Huazhong University of Science and Technology, Hubei, China, in 2011. Currently, he is pursuing the Ph.D. degree at the Image Processing and Pattern Recognition Laboratory, Shanghai Jiao Tong University, Shanghai, China. His current research interests include object detection, saliency detection, visual tracking, and machine learning.

**Qiang Wu** received the B.Eng. and M.Eng. degrees in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree in computing science from the University of Technology Sydney, Sydney, Australia, in 2004. He is currently a Senior Lecturer with the School of Computing and Communications, University of Technology Sydney. His major research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. He has published more than 70 refereed papers in these areas, including those published in prestigious journals and top international conferences. He has been a Guest Editor of several international journals, such as Pattern Recognition Letters and the International Journal of Pattern Recognition and Artificial Intelligence. He has served as a Chair and/or a Program Committee Member for a number of international conferences. He has also served as a Reviewer for several international journals, such as the IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, the IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition, Pattern Recognition Letters, the International Journal of Pattern Recognition and Artificial Intelligence, and the EURASIP Journal on Image and Video Processing.



**Enmei Tu** was born in Anhui, China. He received his B.Sc. and M.Sc. degrees from University of Electronic Science and Technology of China (UESTC) in 2007 and 2010, respectively. He is now a Ph.D. candidate in the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. His research interests are machine learning, computer vision and remote sensing data processing.

**Jie Yang** received the Ph.D. degree from the Department of Computer Science, Hamburg University, Hamburg, Germany, in 1994. Currently, he is a professor at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. He has led many research projects (*e.g.* National Science Foundation, 863 National High Tech. Plan), had one book published in Germany, and authored more than 200 journal papers. His current research interests include object detection and recognition, data fusion and data mining, and medical image processing.