11-2012

Multiview semi-supervised learning with consensus

Guangxia LI
Nanyang Technological University

Kuiyu CHANG
Nanyang Technological University

Steven C. H. HOI
Singapore Management University, CHHOI@smu.edu.sg

DOI: https://doi.org/10.1109/TKDE.2011.160

Published in IEEE Transactions on Knowledge and Data Engineering, Volume 24, Issue 11, 2012, Pages 2040-2051
https://doi.org/10.1109/TKDE.2011.160

Multiview Semi-Supervised Learning with Consensus

Guangxia Li, Kuiyu Chang, and Steven C.H. Hoi

Abstract—Obtaining high-quality and up-to-date labeled data can be difficult in many real-world machine learning applications. Semi-supervised learning aims to improve the performance of a classifier trained with limited number of labeled data by utilizing the unlabeled ones. This paper demonstrates a way to improve the transductive SVM, which is an existing semi-supervised learning algorithm, by employing a multiview learning paradigm. Multiview learning is based on the fact that for some problems, there may exist multiple perspectives, so called views, of each data sample. For example, in text classification, the typical view contains a large number of raw content features such as term frequency, while a second view may contain a small but highly informative number of domain specific features. We propose a novel two-view transductive SVM that takes advantage of both the abundant amount of unlabeled data and their multiple representations to improve classification result. The idea is straightforward: train a classifier on each of the two views of both labeled and unlabeled data, and impose a global constraint requiring each classifier to assign the same class label to each labeled and unlabeled sample. We also incorporate manifold regularization, a kind of graph-based semi-supervised learning method into our framework. The proposed two-view transductive SVM was evaluated on both synthetic and real-life data sets. Experimental results show that our algorithm performs up to 10 percent better than a single-view learning approach, especially when the amount of labeled data is small. The other advantage of our two-view semi-supervised learning approach is its significantly improved stability, which is especially useful when dealing with noisy data in real-world applications.

Index Terms—Artificial intelligence, learning systems, semi-supervised learning, multiview learning, support vector machines

1 Introduction

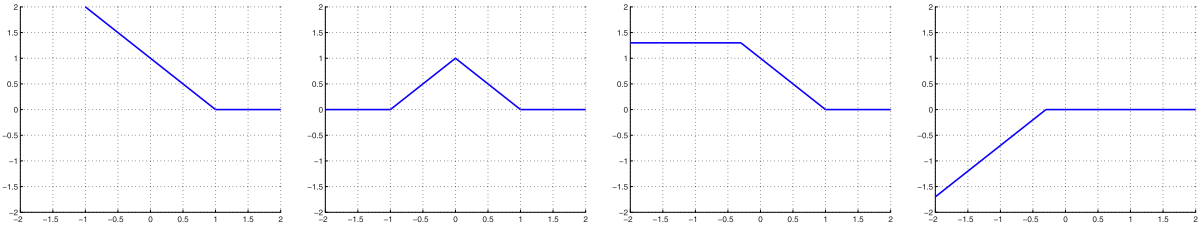
Classification, the task of assigning objects to one of several predefined categories, is an active research problem in data mining and machine learning. The classical classifier is created by building a machine learning model, e.g., support vector machines (SVM) [2], trained from a collection of labeled data. Unfortunately, in real-world applications, labeled training examples are often difficult to obtain, as they require the efforts of human annotators, while unlabeled data are always abundant. We attempt to overcome this limitation with a semi-supervised learning approach, which aims to improve the performance of a classifier trained with limited number of labeled data by utilizing the unlabeled ones. Among various semi-supervised learning algorithms, the transductive support vector machine (TSVM) has drawn a lot of attention since it was first introduced by Vapnik [2]. An intuitive interpretation for the success of transductive SVM is the so-called "cluster assumption" [3]. That is, instead of traversing through high density regions of the data, the decision boundary should always be placed in low density regions. One can incorporate this assumption into the SVM optimization procedure by exploiting the information of unlabeled data.

To improve the performance of transductive SVM, we adopt a multiview learning approach. In multiview learning, a classifier is created for each *representation* or *view* of the same problem, with each classifier optimized to maximize the overall consensus in their predictions, i.e., multiple views of a data sample should be classified into the same category. Where multiple representations of the same problem are available, a multiview learning approach typically yields equal or better results than those obtained from either view alone. Our proposed two-view semi-supervised learning algorithm, called two-view transductive SVM (Two-view TSVM), extends the supervised two-view learning framework of Farquhar et al. [4] to take advantage of the abundance of unlabeled data.

To go a step further, we also apply manifold regularization, which is a kind of graph-based semi-supervised learning approach. Manifold regularization [5], [6] extends many existing supervised learning algorithms to their semi-supervised learning settings by adding a geometrically based regularization term, with the aim of preserving the manifold smoothness. By incorporating a regularizer that records the intrinsic manifold structure of training data, we finally obtain a joint learning framework that combines two types of semi-supervised learning techniques: 1) learning to maximize margin, and 2) learning to explore cluster/manifold structure. The hybrid approach is reasonable since it maximizes the margin on both labeled and unlabeled data and at the same time exploits the manifold structure of the data. We formulate this learning framework into an optimization problem and develop an efficient way to solve it.

We evaluate the proposed classification technique on both synthetic and real-life data sets against single-view/multi-view supervised/semi-supervised approaches, and graph-based method. Specifically, we apply our algorithm to an interesting problem—the product review filtering, which

• *The authors are with the School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798. E-mail: {ligu0005, askychang, chhoi}@ntu.edu.sg.*



(a) Hinge loss (b) Symmetric hinge loss (c) Ramp loss (d) Concave loss

Fig. 1. Four types of loss function. For ramp loss (Fig. 1c) and concave loss (Fig. 1d), the parameter s is set to -0.3.

filters valid product reviews from online forum postings. This task is a preprocessing step for product review mining, which aims to extract and summarize user opinions from product reviews [7]. The vast majority of existing approaches use machine learning models trained on a set of human labeled examples to detect spam reviews. As labeled data are hard to obtain [8], some have come to regard duplicate reviews as spam for training the model [9]. We thus argue that in the absence of user ratings (ground truth), a semi-supervised learning method like our two-view transductive SVM is a better way to tackle the review filtering problem. In particular, we defined the two views for product reviews as 1) a classical text representation based on the word vector model, and 2) a high-level representation based on semantic analysis of review sentence. Experimental results on product review filtering and other general classification data sets justified the utility of our method.

The rest of this paper is organized as follows: Section 2 summarizes related work. Section 3 presents our two-view transductive SVM algorithm. Section 4 gives experimental results and discussions. Finally, Section 5 gives conclusions and discusses future directions.

2 Related Work

We first review existing work on semi-supervised learning, focusing on transductive SVM and graph-based methods, followed by the multiview learning algorithms.

2.1 Semi-Supervised Learning

Semi-supervised learning, i.e., learning from both labeled and unlabeled data, has been extensively studied, leading to several classical approaches. We first give a brief review on the transductive support vector machines, followed by graph-based methods.

2.1.1 Transductive SVM

The transductive SVM can be viewed as a standard SVM with an extra regularization term defined on unlabeled data [10]. Suppose a training set contains ℓ labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$, $y_i \in \{-1, 1\}$, and u unlabeled examples $\{\mathbf{x}_i\}_{i=\ell+1}^{\ell+u}$, $\mathbf{x}_i \in \mathbb{R}^n$. The SVM decision function has the form

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$ are the parameters of the model, and $\Phi(\cdot)$ is the feature map. The transductive SVM adds a regularizer, which is defined over unlabeled data, to the classical SVM optimization function, leading to the following optimization problem:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{\ell} L(y_i f(\mathbf{x}_i)) + C_2 \sum_{i=\ell+1}^{\ell+u} L(|f(\mathbf{x}_i)|),$$

where $L(\cdot) = \max(0, 1 - \cdot)$ is the classical hinge loss for labeled examples as illustrated in Fig. 1a, $L(|\cdot|) = \max(0, 1 - |\cdot|)$ is the symmetric hinge loss for unlabeled examples as illustrated in Fig. 1b. Note that its nonconvex hat shape makes the optimization problem hard to solve.

A suite of algorithms have been proposed to solve the above optimization problem [3], [11], [12], [13], [14]. Particularly, Collobert et al. [12] employed an approximate optimization technique known as the concave convex procedure (CCCP) [15]. It decomposes a nonconvex function into a convex part and a concave part, which are then solved iteratively. In each iteration, the concave part is replaced by its tangential approximation. Then, the sum of the convex part and the tangential approximation is minimized.

For CCCP transductive SVM [12], the loss function applied to unlabeled data is called "ramp loss" (Fig. 1c), which can be expressed as the sum of a hinge loss function (Fig. 1a) and a concave loss function (Fig. 1d). Specifically, the ramp loss function $R_s(\cdot)$ has the form

$$R_s(\cdot) = \min(1 - s, \max(0, 1 - \cdot)) = L(\cdot) + L_s(\cdot),$$

where L is the hinge loss, L_s is the concave loss with the form $L_s(\cdot) = -\max(0, s - \cdot)$, and s is a predefined parameter such that $-1 < s \leq 0$.

Training a transductive SVM with the CCCP method is equivalent to training an SVM using the hinge loss for labeled data, and the ramp loss for unlabeled data [12]. For a binary classification problem, each unlabeled example is accounted for twice, each time assuming the role of one class, that is, $\{(\mathbf{x}_i, y_i = 1)\}_{i=\ell+1}^{\ell+u}, \{(\mathbf{x}_i, y_i = -1) : \mathbf{x}_i = \mathbf{x}_{i-u}\}_{i=\ell+u+1}^{\ell+2u}$. The corresponding optimization problem of CCCP transductive SVM is given by

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{\ell} L(y_i f(\mathbf{x}_i)) + C_2 \sum_{i=\ell+1}^{\ell+2u} R_s(y_i f(\mathbf{x}_i)).$$

2.1.2 Graph-Based Methods

Graph-based semi-supervised learning methods assume that similar examples should be assigned the same class labels. It first defines a graph where labeled and unlabeled data are represented as vertices, with edge weights encoding the similarity between examples. It then estimates a function

over the graph by simultaneously satisfying two conditions [10]: 1) the function should yield assignments similar to the labeled nodes, and 2) it should be smooth throughout the graph. These two conditions can be expressed quantitatively in an optimization framework using a loss function and a regularizer derived from data features.

Thus, existing graph-based methods differ from one another largely in the particular choice of loss functions and regularizers. Blum and Chawla [16] considered semi-supervised learning as a graph min-cut problem. The Gaussian random fields and harmonic function methods [17] introduce a quadratic loss function with infinity weight for labeled data, and incorporate unlabeled data with a regularizer based on the graph combinatorial Laplacian. Zhou et al. [18] proposed the local and global consistency method with a quadratic loss function and the normalized Laplacian as the regularizer. The unification of margin-based and manifold-based regularization has also been explored in [3], [19].

The generative manifold regularization framework [5], [20] exploits the geometry of probability distribution that generates the data and incorporates it as an additional regularization term. Suppose labeled examples are drawn from a probability distribution P, and unlabeled examples are drawn from the marginal distribution $\mathcal{P}_X$ of P. Manifold regularization makes a specific assumption that if two points $\mathbf{x}_1, \mathbf{x}_2 \in X$ are close in the intrinsic geometry of $\mathcal{P}_X$, then the conditional distributions $\mathcal{P}(y|\mathbf{x}_1)$ and $\mathcal{P}(y|\mathbf{x}_2)$ are similar. More specifically, the framework can be expressed as an optimization problem with an arbitrary loss function and two regularizers as shown below:

$$\min \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{x}_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2,$$

where $\|f\|_I^2$ reflects the complexity of the function in the intrinsic geometry of $\mathcal{P}_X$, and can be approximated on the basis of labeled and unlabeled data using the graph Laplacian [21], [22]. That is,

$$\|f\|_I^2 = \mathbf{f}^{\mathrm{T}} L \mathbf{f},$$

where $\mathbf{f}$ is the vector of f evaluation on the labeled and unlabeled data, given by $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{\ell+u})]^{\mathrm{T}}$, and L is the graph Laplacian given by $L = D - W$. The diagonal matrix D is given by $D_{ii} = \sum_{j=1}^{\ell+u} W_{ij}$, where W_{ij} are the edge weights in the data adjacency graph.

2.2 Multiview Learning

Multiview learning utilizes the agreement among learners trained on different representations of the same problem to improve the overall classification performance. The basic idea of using two views with unlabeled data was first introduced in [23]. The well-known cotraining algorithm [24] learns two independent classifiers based on independent attribute sets. These classifiers then predict the unlabeled examples. Their most confident predictions are used to mutually expand the training set. Some theoretical studies and effective variants of cotraining algorithms include [25], [26], [27], [28], [29], [30]. In addition to regular cotraining algorithms, Yu et al. [31] proposed a Bayesian cotraining approach, which defines an undirected graphical model based on a Gaussian process with edge potential functions denoting the internal and external agreement of the views. Sindhwani et al. [32], [33] proposed a coregularization approach to learning a multiview classifier from partially labeled data using a view consensus based on some regularization term. Maillard and Vayatis [34] further analyzed the complexity of coregularization methods for multiview semi-supervised learning. A similar approach has also been adopted for semi-supervised least squares regression [35]. Farquhar et al. [4] observed that when two views of the same problem are available, applying the kernel canonical correlation analysis (KCCA) [36] to the two feature space can improve the performance of the classifier. They also proposed a supervised learning algorithm named SVM-2K, which imposes a similarity constraint between two distinct SVMs, each trained from one view of the data. The constraint they employed is

$$\left| f^{\mathrm{A}}(\mathbf{x}_i^{\mathrm{A}}) - f^{\mathrm{B}}(\mathbf{x}_i^{\mathrm{B}}) \right| \leq \eta_i + \varepsilon,$$

where $f^{\mathrm{A/B}}(\cdot)$ are the SVM decision functions belonging to each of the two views denoted by superscripts A and B, η_i is a variable that enforces consensus between the two views, and ε is a slack variable for allowing some samples to violate the constraint. Combining this constraint with the standard SVM objective functions for each view yields a multiview learning algorithm, which was shown to perform better than the single view approach on an image classification task.

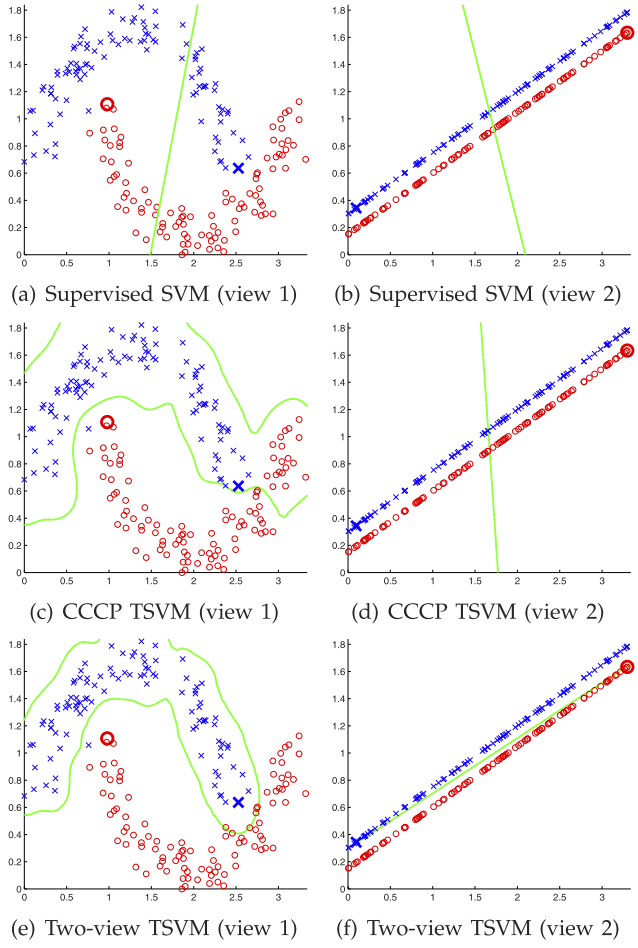
Not restricting to labeled data, Szedmák and Shawe-Taylor [37] went further by exploiting unlabeled data via multiview learning. They required two classifiers to give similar solutions on the unlabeled samples. The similarity is measured by the absolute value of differences between two real-valued predictions of the unlabeled data, and is minimized simultaneously with the error occurring in the estimation of the labeled cases. In their learning framework, the loss function is only defined over the labeled data. In contrast, our proposed method also contains loss functions defined over unlabeled data. The difference between the work of Szedmák and Shawe-Taylor [37] and ours is that our method finds a labeling of the unlabeled data, so that a decision boundary has the maximum margin on both the original labeled data and the (newly labeled) unlabeled data.

3 Two-View Transductive Support Vector Machine

We first use a synthetic data set as an example to illustrate the motivation for two-view, semi-supervised learning. Next, we propose the framework of our two-view transductive support vector machine, followed by the optimization technique and algorithm.

3.1 Motivation

We extend the two-view supervised learning algorithm proposed by Farquhar et al. [4] by incorporating unlabeled data, turning it into a two-view semi-supervised learning approach. The basic idea is to construct two transductive SVM classifiers from both labeled and unlabeled data based on different representations of the original problem. The



(a) Supervised SVM (view 1) (b) Supervised SVM (view 2)
(c) CCCP TSVM (view 1) (d) CCCP TSVM (view 2)
(e) Two-view TSVM (view 1) (f) Two-view TSVM (view 2)

Fig. 2. Decision boundaries (denoted by the solid line) obtained by the supervised SVM, CCCP TSVM, and Two-view TSVM. The only two labeled examples are represented by a bold cross and circle. The remaining points are unlabeled. Gaussian and linear kernels are used for views 1 and 2, respectively.

two classifiers are then trained simultaneously by requiring that they always retain a maximum consensus on their predictions. By enforcing different classifiers trained from different views to agree on both labeled and unlabeled training data, the structure learned from each view can reinforce one another. Once trained, the outputs of two classifiers can be used individually. A voting or weighting scheme can also be applied to combine the classifier outputs to make predictions.

To illustrate the advantage of two-view transductive learning, consider a synthetic data set in which samples from two classes appear as two moons in one view and two lines in another, as shown in Fig. 2 (crosses and circles are used to represent the two classes, respectively). Given only two labeled examples (denoted by a bold cross and circle), the solid lines in Figs. 2a and 2b turn out to be the maximum margin hyperplane of the two training instances. They are clearly suboptimal with respect to the underlying distribution of unlabeled data (denoted by the small crosses and circles). Taking unlabeled data into consideration, a transductive SVM [12] shifts the decision boundary away from dense regions, but still fails to yield a good result in either view (Figs. 2c and 2d). On the contrary, once a consensus between the two views is imposed on both classifiers, a much better decision boundary is obtained in each view. This is shown in Figs. 2e and 2f (the result is obtained by applying our two-view transductive SVM), in which the solid decision boundary clearly separates the two classes of data.

3.2 Problem Formulation

Consider a multiview semi-supervised learning problem on a set of ℓ labeled examples $\{(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_i^{\mathrm{B}}), y_i\}_{i=1}^{\ell}$, $\mathbf{x}_i^{\mathrm{A/B}} \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, and a set of u unlabeled examples $\{\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_i^{\mathrm{B}}\}_{i=\ell+1}^{\ell+u}$. Superscripts A and B denote the two views, respectively. For each view, we aim to find a decision function $f(\mathbf{x})$ with the form shown in (1).

According to Collobert et al. [12], for each view, the CCCP transductive SVM has the following objective function:

$$J = \frac{1}{2}\|\mathbf{w}^{\mathrm{A/B}}\|^2 + C_1^{\mathrm{A/B}} \sum_{i=1}^{\ell} \xi_i^{\mathrm{A/B}} + C_2^{\mathrm{A/B}} \sum_{i=\ell+1}^{\ell+2u} \xi_i^{\mathrm{A/B}} + \sum_{i=\ell+1}^{\ell+2u} \rho_i^{\mathrm{A/B}} y_i f^{\mathrm{A/B}}\big(\mathbf{x}_i^{\mathrm{A/B}}\big),$$

where $\rho_i^{\mathrm{A/B}}$ is related to the derivative of the concave loss function mentioned in Section 2.1.1, written as

$$\rho_i^{\mathrm{A/B}} = \begin{cases} C_2^{\mathrm{A/B}} & \text{if } y_i f^{\mathrm{A/B}}\big(\mathbf{x}_i^{\mathrm{A/B}}\big) < s \text{ and } i \geq \ell + 1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where s is the parameter of the loss function.

In our basic approach, we construct two transductive SVM objective functions based on each of the two views, respectively. We then add a regularizer to penalize the decision function of each view if it deviates from the consensus, and minimize them simultaneously. To go a step further, we also explore the structure of the data manifold by adding two regularizers that penalize any "abrupt changes" of the function values evaluated on neighboring samples in the Laplacian graph. This leads to the following optimization problem of our two-view transductive SVM:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}^{\mathrm{A}}\|^2 + C_1^{\mathrm{A}} \sum_{i=1}^{\ell} \xi_i^{\mathrm{A}} + C_2^{\mathrm{A}} \sum_{i=\ell+1}^{\ell+2u} \xi_i^{\mathrm{A}} \\ & + \frac{1}{2}\|\mathbf{w}^{\mathrm{B}}\|^2 + C_1^{\mathrm{B}} \sum_{i=1}^{\ell} \xi_i^{\mathrm{B}} + C_2^{\mathrm{B}} \sum_{i=\ell+1}^{\ell+2u} \xi_i^{\mathrm{B}} \\ & + \sum_{i=\ell+1}^{\ell+2u} \rho_i^{\mathrm{A}} y_i f^{\mathrm{A}}(\mathbf{x}_i^{\mathrm{A}}) + \sum_{i=\ell+1}^{\ell+2u} \rho_i^{\mathrm{B}} y_i f^{\mathrm{B}}(\mathbf{x}_i^{\mathrm{B}}) \\ & + C_3^{\mathrm{A}} \mathbf{f}^{\mathrm{AT}} \mathrm{L}^{\mathrm{A}} \mathbf{f}^{\mathrm{A}} + C_3^{\mathrm{B}} \mathbf{f}^{\mathrm{BT}} \mathrm{L}^{\mathrm{B}} \mathbf{f}^{\mathrm{B}} + D \sum_{i=1}^{\ell+2u} \eta_i \end{aligned} \tag{4a}$$

$$\begin{aligned} \text{w.r.t.} \quad & \mathbf{w}^{\mathrm{A/B}}, \boldsymbol{\xi}^{\mathrm{A/B}}, \boldsymbol{\eta} \\ \text{s.t.} \quad & y_i f^{\mathrm{A/B}}\big(\mathbf{x}_i^{\mathrm{A/B}}\big) \geq 1 - \xi_i^{\mathrm{A/B}} \end{aligned} \tag{4b}$$

$$\xi_i^{\mathrm{A/B}} \geq 0 \tag{4c}$$

$$\big|f^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}\big) - f^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}\big)\big| \leq \eta_i \tag{4d}$$

$$\eta_i \geq 0 \tag{4e}$$

$$\frac{1}{u}\sum_{i=\ell+1}^{\ell+u} f^{\mathrm{A/B}}\big(\mathbf{x}_i^{\mathrm{A/B}}\big) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i, \tag{4f}$$

where constraints (4b) and (4c) are the standard SVM constraints, constraints (4d) and (4e) impose the consensus between the two views, and constraint (4f) is a balancing constraint, which aims to prevent an extremely skewed classification result caused by assigning all unlabeled examples to only one class. It has been previously used in [3], [12]. The positive parameters $C_2^{\mathrm{A/B}}$ control the influence of unlabeled data on the objective function, while $C_3^{\mathrm{A/B}}$ control the influence of the graph-based regularizers. It is clear that setting $C_2^{\mathrm{A/B}}$ and $C_3^{\mathrm{A/B}}$ to zeros leads to a fully supervised two-view SVM; setting $C_3^{\mathrm{A/B}}$ alone to zero causes the two-view transductive SVM to ignore manifold information of the training samples.

3.3 Derivation of Optimization Problem

We use K interchangeably to denote the kernel function or the Gram matrix. From the Representer theorem, we know that the solution to the problem above has the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell+2u} \alpha_i \mathrm{K}(\mathbf{x}_i, \mathbf{x}).$$

The optimization problem (4) can be rewritten as

$$\begin{aligned}
\min \quad & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{AT}}\mathrm{K}^{\mathrm{A}}\boldsymbol{\alpha}^{\mathrm{A}} + C_1^{\mathrm{A}}\sum_{i=1}^{\ell}\xi_i^{\mathrm{A}} + C_2^{\mathrm{A}}\sum_{i=\ell+1}^{\ell+2u}\xi_i^{\mathrm{A}} \\
& + \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{BT}}\mathrm{K}^{\mathrm{B}}\boldsymbol{\alpha}^{\mathrm{B}} + C_1^{\mathrm{B}}\sum_{i=1}^{\ell}\xi_i^{\mathrm{B}} + C_2^{\mathrm{B}}\sum_{i=\ell+1}^{\ell+2u}\xi_i^{\mathrm{B}} \\
& + \sum_{i=\ell+1}^{\ell+2u}\rho_i^{\mathrm{A}} y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}}\right) \\
& + \sum_{i=\ell+1}^{\ell+2u}\rho_i^{\mathrm{B}} y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}}\right) + D\sum_{i=1}^{\ell+2u}\eta_i \\
& + C_3^{\mathrm{A}}\boldsymbol{\alpha}^{\mathrm{AT}}\mathrm{K}^{\mathrm{AT}}\mathrm{L}^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\boldsymbol{\alpha}^{\mathrm{A}} + C_3^{\mathrm{B}}\boldsymbol{\alpha}^{\mathrm{BT}}\mathrm{K}^{\mathrm{BT}}\mathrm{L}^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\boldsymbol{\alpha}^{\mathrm{B}}
\end{aligned} \tag{5a}$$

$$\begin{aligned}
\text{w.r.t.} \quad & \boldsymbol{\alpha}^{\mathrm{A/B}}, b^{\mathrm{A/B}}, \boldsymbol{\xi}^{\mathrm{A/B}}, \boldsymbol{\eta} \\
\text{s.t.} \quad & y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}}\right) \geq 1 - \xi_i^{\mathrm{A}}
\end{aligned} \tag{5b}$$

$$\xi_i^{\mathrm{A}} \geq 0 \tag{5c}$$

$$y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}}\right) \geq 1 - \xi_i^{\mathrm{B}} \tag{5d}$$

$$\xi_i^{\mathrm{B}} \geq 0 \tag{5e}$$

$$\left|\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}} - \sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) - b^{\mathrm{B}}\right| \leq \eta_i \tag{5f}$$

$$\eta_i \geq 0 \tag{5g}$$

$$\frac{1}{2u}\sum_{i=\ell+1}^{\ell+2u}\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}}\right) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \tag{5h}$$

$$\frac{1}{2u}\sum_{i=\ell+1}^{\ell+2u}\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}}\right) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i. \tag{5i}$$

We apply the Lagrange multiplier technique to solve the optimization problem (5). The assignment between the Lagrange multipliers and the constraints is summarized as follows:

$$\begin{aligned}
& \beta_i^{\mathrm{A}}\colon\ y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}}\right) \geq 1 - \xi_i^{\mathrm{A}} \\
& \beta_i^{\mathrm{B}}\colon\ y_i\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}}\right) \geq 1 - \xi_i^{\mathrm{B}} \\
& \gamma_i^{+}\colon\ \sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}} - \sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) - b^{\mathrm{A}} \geq -\eta_i \\
& \gamma_i^{-}\colon\ \sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}} - \sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) - b^{\mathrm{B}} \geq -\eta_i \\
& \delta^{\mathrm{A}}\colon\ \frac{1}{2u}\sum_{i=\ell+1}^{\ell+2u}\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}_j^{\mathrm{A}}\big) + b^{\mathrm{A}}\right) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \\
& \delta^{\mathrm{B}}\colon\ \frac{1}{2u}\sum_{i=\ell+1}^{\ell+2u}\left(\sum_{j=1}^{\ell+2u}\alpha_j^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}_j^{\mathrm{B}}\big) + b^{\mathrm{B}}\right) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \\
& \zeta_i^{\mathrm{A}}\colon\ \xi_i^{\mathrm{A}} \geq 0 \\
& \zeta_i^{\mathrm{B}}\colon\ \xi_i^{\mathrm{B}} \geq 0 \\
& \sigma\colon\ \eta_i \geq 0.
\end{aligned}$$

Applying the Lagrange multiplier technique, the minimization problem (5) is equivalent to the following problem:

$$\begin{aligned}
\min \quad & \frac{1}{2}\mathrm{M}^{\mathrm{AT}}\mathrm{K}^{\mathrm{A}}\big(\mathrm{I} + 2C_3^{\mathrm{A}}\mathrm{L}^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\big)^{-1}\mathrm{M}^{\mathrm{A}} \\
& + \frac{1}{2}\mathrm{M}^{\mathrm{BT}}\mathrm{K}^{\mathrm{B}}\big(\mathrm{I} + 2C_3^{\mathrm{B}}\mathrm{L}^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\big)^{-1}\mathrm{M}^{\mathrm{B}} \\
& - \sum_{i=1}^{\ell+2u}\big(\widetilde{\beta}_i^{\mathrm{A}} + \widetilde{\beta}_i^{\mathrm{B}}\big) - \frac{\delta^{\mathrm{A}} + \delta^{\mathrm{B}}}{\ell}\sum_{i=1}^{\ell} y_i
\end{aligned} \tag{7a}$$

$$\begin{aligned}
\text{w.r.t.} \quad & \widetilde{\boldsymbol{\beta}}^{\mathrm{A/B}}, \delta^{\mathrm{A/B}}, \boldsymbol{\gamma} \\
\text{s.t.} \quad & 0 \leq \widetilde{\beta}_i^{\mathrm{A}} \leq C_1^{\mathrm{A}} \quad \forall 1 \leq i \leq \ell
\end{aligned} \tag{7b}$$

$$-\rho_i^{\mathrm{A}} \leq \widetilde{\beta}_i^{\mathrm{A}} \leq C_2^{\mathrm{A}} - \rho_i^{\mathrm{A}} \quad \forall \ell+1 \leq i \leq \ell+2u \tag{7c}$$

$$0 \leq \widetilde{\beta}_i^{\mathrm{B}} \leq C_1^{\mathrm{B}} \quad \forall 1 \leq i \leq \ell \tag{7d}$$

$$-\rho_i^{\mathrm{B}} \leq \widetilde{\beta}_i^{\mathrm{B}} \leq C_2^{\mathrm{B}} - \rho_i^{\mathrm{B}} \quad \forall \ell+1 \leq i \leq \ell+2u \tag{7e}$$

$$-D \le \gamma_i \le D \quad \forall 1 \le i \le \ell + 2u \tag{7f}$$

$$\sum_{i=1}^{\ell+2u} \left(\widetilde{\beta}^{\mathrm{A}} y_i - \gamma_i\right) + \delta^{\mathrm{A}} = 0 \tag{7g}$$

$$\sum_{i=1}^{\ell+2u} \left(\widetilde{\beta}^{\mathrm{B}} y_i + \gamma_i\right) + \delta^{\mathrm{B}} = 0, \tag{7h}$$

where $\mathrm{M}^{\mathrm{A/B}} = \mathrm{Y}\widetilde{\boldsymbol{\beta}}^{\mathrm{A/B}} - \boldsymbol{\gamma} + \frac{\delta^{\mathrm{A/B}}}{2u}\mathrm{J}$, $\widetilde{\boldsymbol{\beta}}^{\mathrm{A/B}} = \boldsymbol{\beta}^{\mathrm{A/B}} - \boldsymbol{\rho}^{\mathrm{A/B}}$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{+} - \boldsymbol{\gamma}^{-}$, I is a identity matrix, Y is a diagonal matrix as $\mathrm{Y} = \mathrm{diag}(y_1, \ldots, y_{\ell+2u})$, and J is a $(\ell + 2u) \times 1$ column vector with first ℓ elements equal to zero and last $2u$ elements equal to one.

3.4 Augmented Lagrangian Technique

To solve the minimization problem (7), we employ the augmented Lagrangian [38] technique as Farquhar et al. [4] did. Augmented Lagrangian is a method for solving constrained optimization problems. It reformulates a constrained optimization problem into an unconstrained one by adding Lagrange multipliers and an extra penalty term for each constraint to the original objective function. The augmented Lagrangian function corresponding to the minimization problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & c_i(x) = 0 \quad \forall 1 \le i \le n \end{aligned}$$

can be written as

$$\min_{x} \quad f(x) - \sum_{i=1}^{n} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{n} c_i^2(x), \tag{9}$$

where the first two terms in (9) correspond to the Lagrangian and the last term is the penalty for violating the constraint. The penalty term is positive when the current point x violates the constraint and zero otherwise. It is multiplied by a positive coefficient μ. By making this coefficient larger, we penalize constraint violations more severely, thereby forcing the minimizer to the penalty function to move closer to a feasible region of the constrained problem.

The minimization problem (9) can be solved in an iterative manner. In each iteration, $\boldsymbol{\lambda}$ is fixed to some estimate of the optimal Lagrange multiplier and the penalty parameter μ is set to some positive value, then one can perform minimization with respect to x. In subsequent iterations, $\boldsymbol{\lambda}$ and μ are updated, and the process is repeated until some stopping criterion is reached. It has been shown that convergence of the augmented Lagrangian method is assured provided that μ does not increase indefinitely [38].

3.5 Two-View Transductive SVM Algorithm

Let us denote the equality constraints (7g) and (7h) as h_1 and h_2, and introduce corresponding Lagrange multipliers λ_1 and λ_2. We can rewrite the minimization problem (7) into the augmented Lagrangian form as follows:

$$\begin{aligned} \min \quad & \frac{1}{2}\mathrm{M}^{\mathrm{AT}}\mathrm{K}^{\mathrm{A}}\left(\mathrm{I} + 2C_3^{\mathrm{A}}\mathrm{L}^{\mathrm{A}}\mathrm{K}^{\mathrm{A}}\right)^{-1}\mathrm{M}^{\mathrm{A}} - \sum_{i=1}^{\ell+2u} \widetilde{\beta}_i^{\mathrm{A}} \\ & + \frac{1}{2}\mathrm{M}^{\mathrm{BT}}\mathrm{K}^{\mathrm{B}}\left(\mathrm{I} + 2C_3^{\mathrm{B}}\mathrm{L}^{\mathrm{B}}\mathrm{K}^{\mathrm{B}}\right)^{-1}\mathrm{M}^{\mathrm{B}} - \sum_{i=1}^{\ell+2u} \widetilde{\beta}_i^{\mathrm{B}} \\ & - \frac{\delta^{\mathrm{A}} + \delta^{\mathrm{B}}}{\ell} \sum_{i=1}^{\ell} y_i - \sum_{i=1}^{2} \lambda_i h_i + \frac{\mu}{2} \sum_{i=1}^{2} \|h_i\|^2 \end{aligned} \tag{10a}$$

$$\begin{aligned} \text{w.r.t.} \quad & \widetilde{\boldsymbol{\beta}}^{\mathrm{A/B}}, \delta^{\mathrm{A/B}}, \boldsymbol{\gamma} \\ \text{s.t.} \quad & 0 \le \widetilde{\beta}_i^{\mathrm{A}} \le C_1^{\mathrm{A}} \qquad \forall 1 \le i \le \ell \end{aligned} \tag{10b}$$

$$-\rho_i^{\mathrm{A}} \le \widetilde{\beta}_i^{\mathrm{A}} \le C_2^{\mathrm{A}} - \rho_i^{\mathrm{A}} \quad \forall \ell + 1 \le i \le \ell + 2u \tag{10c}$$

$$0 \le \widetilde{\beta}_i^{\mathrm{B}} \le C_1^{\mathrm{B}} \qquad \forall 1 \le i \le \ell \tag{10d}$$

$$-\rho_i^{\mathrm{B}} \le \widetilde{\beta}_i^{\mathrm{B}} \le C_2^{\mathrm{B}} - \rho_i^{\mathrm{B}} \quad \forall \ell + 1 \le i \le \ell + 2u \tag{10e}$$

$$-D \le \gamma_i \le D \qquad \forall 1 \le i \le \ell + 2u, \tag{10f}$$

where $\mu > 0$ is the penalty parameter.

Once the minimization problem (10) is solved with the augmented Lagrangian method, the decision functions corresponding to the two views can be calculated as follows:

$$f^{\mathrm{A}}(\mathbf{x}^{\mathrm{A}}) = \sum_{i=1}^{\ell+2u} \left(y_i \widetilde{\beta}_i^{\mathrm{A}} + \gamma_i\right) \mathrm{K}^{\mathrm{A}}\left(\mathbf{x}_i^{\mathrm{A}}, \mathbf{x}^{\mathrm{A}}\right) + b^{\mathrm{A}} \tag{11a}$$

$$f^{\mathrm{B}}(\mathbf{x}^{\mathrm{B}}) = \sum_{i=1}^{\ell+2u} \left(y_i \widetilde{\beta}_i^{\mathrm{B}} - \gamma_i\right) \mathrm{K}^{\mathrm{B}}\left(\mathbf{x}_i^{\mathrm{B}}, \mathbf{x}^{\mathrm{B}}\right) + b^{\mathrm{B}}. \tag{11b}$$

A hybrid decision function can be written as a linear combination of the two classifiers as

$$f(\mathbf{x}) = \omega f^{\mathrm{A}}(\mathbf{x}^{\mathrm{A}}) + (1 - \omega) f^{\mathrm{B}}(\mathbf{x}^{\mathrm{B}}) \tag{12}$$

with $0 \le \omega \le 1$.

Algorithm 1 summarizes the two-view transductive SVM algorithm. The convergence of the CCCP procedure is described in [12]. A detailed convergence analysis of the Lagrange multiplier iteration, which corresponds to the outer loop of Algorithm 1 can be found in [39]. In our experiments, we set the maximum number of Lagrange multiplier iterations to five. We observe that the algorithm converges before reaching the maximum number of iterations in most cases.

Algorithm 1. Two-view Transductive SVM
Require: Labeled and unlabeled data of two views.
Initialize $\widetilde{\boldsymbol{\beta}}^{\mathrm{A/B}}$, $\boldsymbol{\rho}^{\mathrm{A/B}}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$ and μ.
repeat
Solve the following sub-problem.
repeat
Solve the minimization problem (10) with fixed $\boldsymbol{\lambda}^k$ and μ^k.
Compute $f^{\mathrm{A(t+1)}}$ and $f^{\mathrm{B(t+1)}}$ via (11) using the solution

TABLE 1
Summary of Data Sets in Our Experiments

	View		Sample Count			
	Name	#Dimension	#Positive	#Negative	#Total	#Labeled
Synthetic	line	2	100	100	200	2
	moon	2				
Ads	img url	457	459	2820	3279	32
	dest url	472				
	alt	111				
WebKB	page	3000	230	821	1051	12
	link	1840				
Product review	lexical	2759	1000	1000	2000	20
	formal	5				

to the minimization problem (10).
Compute $\boldsymbol{\rho}^{\mathrm{A}(t+1)}$ and $\boldsymbol{\rho}^{\mathrm{B}(t+1)}$ via (3) with the value of $f^{\mathrm{A}(t+1)}$ and $f^{\mathrm{B}(t+1)}$.
Update the lower and upper bounds of $\widetilde{\boldsymbol{\beta}}^{\mathrm{A}(t+1)}$ and $\widetilde{\boldsymbol{\beta}}^{\mathrm{B}(t+1)}$ with (10c) and (10e).
until $\boldsymbol{\rho}^{\mathrm{A}(t+1)} = \boldsymbol{\rho}^{\mathrm{A}(t)}$ and $\boldsymbol{\rho}^{\mathrm{B}(t+1)} = \boldsymbol{\rho}^{\mathrm{B}(t)}$
Update the Lagrange multiplier $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^{k} + \mu^{k}\boldsymbol{h}^{k}$$

Update the penalty parameter μ as

$$\mu^{k+1} = \phi\mu^{k}$$

until $\|\boldsymbol{h}^{k}\| \leq \epsilon$
return Decision functions corresponding to the two views calculated by (11).

4 Experimental Results

4.1 Experimental Testbed

We evaluate the performance of our two-view transductive SVM on one synthetic data set and three real-life data sets: the ads data set [40], the WebKB course data set [6], and our product review data set. The characteristics of the data sets including the number of dimensions, class distribution, and portion of labeled examples, are summarized in Table 1.

The synthetic data set contains 200 samples evenly drawn from two classes, distributed in the silhouette of two "moons" in one view and two "lines" in the other. Each view has only two labeled samples (1 positive, 1 negative), with the remaining 198 unlabeled.

The ads data set was first used by Kushmerick [40] to study methods that automatically remove advertisement images from webpages. Each example in the data set corresponds to an image on the web, and the task is to predict whether an image is used for advertisement or not. The ads data set consists of more than two views. We adopted three views in our experiments, including *image URL view* (457 features related to the image server name), *destination URL view* (472 features related to the image URL), and *alt view* (111 features related to "alternate" words in the HTML image tag). Among 3,279 examples in the data set, 459 examples belong to the positive class (ads) and the remaining examples are negative (non-ads).

The WebKB course data set has been frequently used in the empirical study of multiview learning. It comprises 1,051 webpages collected from the computer science departments of four universities. The task is to classify each page into two classes: course or noncourse. The two views are the textual content of a webpage (*page view*) and the words that occur in the hyperlinks of other webpages pointing to it (*link view*), respectively. We used a processed version of the WebKB course data set [6] in our experiment.

Feature Type	Sample Dictionary Terms	Dictionary Size
Opinion phrases	漂亮 (beautiful), 难看 (ugly), 贵 (expensive), 好用 (easy to use), 清晰 (vivid) ...	551 phrases
Question patterns	?, 哪里 (where), 怎么样 (how about), 谁 (who), 为何 (why), 什么 (what) ...	32 patterns
Digits	0 - 9	10 characters
Brands	摩托罗拉 (Motorola), 多普达 (Dopod), 诺基亚 (Nokia), 三星 (Samsung) ...	12 brands
Length of review	Counts the total number of Unicode Characters (Chinese, English letters, and digits) in the review, excluding punctuations.	–

Fig. 3. The five extracted features and their sample dictionary terms where applicable.

Our product review data set was created by crawling two popular online Chinese cell-phone forums.[1] Redundant punctuations and stop words were removed and reviews containing less than four characters were eliminated. We manually labeled 1,000 true reviews and 1,000 spam reviews. A product review is regarded as useful or nonspam if 1) it contains a declarative sentence (all questions are regarded as spam reviews), and 2) it expresses opinions on a product or product feature. Opinions include the reviewer's personal sentiment (positive or negative) about a product or product feature, and/or the pros and cons analysis of a product or product feature.

We treat the product review filtering task as a classification problem. To train the classifier, we define two sets of features: one based on the review content (*lexical view*) and the other based on the characteristics of the review sentences (*formal view*). For the lexical view, since there are no space separators between Chinese words, raw reviews were preprocessed by a Chinese lexical analyzer—ICTCLAS.[2] ICTCLAS performs Chinese word segmentation and part-of-speech tagging. Each sentence was converted to a word vector using the standard TF-IDF (term frequency-inverse document frequency) representation. For the formal view, five types of features are enumerated as follows:

1. Proportion of opinion-bearing phrases in a review sentence.
2. Proportion of questioning patterns in a review sentence.
3. Proportion of numerical digits in a review sentence.
4. Proportion of brand mentions in a review sentence.
5. Length of review sentence.

Fig. 3 shows each of the five features along with some sample dictionary terms and dictionary size. To evaluate the discriminative power of the features, we trained a supervised SVM to classify product reviews based only on the lexical or formal view. The 10-fold cross-validation accuracy was 89.90 and 85.29 percent for the lexical and formal views, respectively. This indicates that each of the

1. http://www.club.mobile.163.com and http://www.3533.com.
2. http://www.ictclas.org.

two views contains sufficient information that is enough to train a good classifier, individually.

4.2 Experimental Setup

We compare the proposed two-view transductive SVM against the standard supervised SVM (LIBSVM [41] trained with a few labeled examples), the two-view supervised learning algorithm—SVM-2K [4] (trained with only labeled examples), the two-view semi-supervised learning algorithm—cotraining [24], the graph-based semi-supervised learning algorithm—Laplacian SVM [6], and the single-view transductive SVM—CCCP TSVM [12].

For the two single-view baselines (standard SVM and CCCP-TSVM), besides reporting their performances on two different views, respectively, we also concatenate the input feature vectors from each view to form a larger feature set, and report the results. We denote this alternative approach as the "Hybrid View" in the following tables. The "Hybrid view" for Laplacian SVM uses the sum of graph Laplacians in each view for regularization (see [6] for details), while the "Hybrid view" for SVM-2K and our Two-view TSVM uses a linear combination of both view's outputs. The weight variable ω in (12) that controls each view's influence on the output is set to 0.5 in all experiments. Note that it is better to set a higher weight for a classifier that is more accurate. In practice, better results can be obtained by tuning the mixing weight.

We generated 10 random splits of the ads data set, and 100 random splits of the WebKB course data set and product review data set. Each split contains a proportion of labeled and unlabeled examples (as shown in Table 1). For all algorithms, the unlabeled data were used as the test set. Since the distribution of some data set is skewed (e.g., 459 of 3,279 examples in the ads data set belong to the positive class), we report the F1-measure in addition to accuracy. The F1-measure is the harmonic mean of precision and recall. It is typically harder for a classifier to achieve a good F1-measure compared to accuracy on a highly skewed data set.

We manually tuned and found the best parameters for each algorithm, using the positive class F1-measure on the unlabeled data set. For simplicity and fairness, we first tuned the parameters (C_1 and C_2) for CCCP TSVM and used the same values for our Two-view TSVM. We only chose the penalty for disagreement (D) from a small range of values for the Two-view TSVM. Note that in the following tables, the notation "Two-view TSVM" denotes the two-view transductive SVM algorithm without manifold regularization (i.e., parameters C_3^{A} and C_3^{B} are set to zero), and the notation "Two-view TSVM with Laplacian" includes manifold regularization. To assess the statistical significance of the Two-view TSVM result, we performed an unpaired t-test at 5 percent significance level with CCCP TSVM as a baseline. Results shown in bold are considered statistically significant.

4.3 Performance Evaluation

4.3.1 Synthetic Data Set

Fig. 2 depicts the classification results of supervised SVM (trained with only two labeled samples), CCCP TSVM, and our Two-view TSVM on the synthetic data set. The superiority of Two-view TSVM is self-evident by comparing the contours of the various generated decision boundaries. The failure of supervised SVM is not surprising due to insufficient labeled training data. The substandard performance of CCCP TSVM shown in Fig. 2d may be ascribed to the fact that the two lines are too close, blurring the boundary between the two classes. We found experimentally that unlabeled data can affect the decision boundary of CCCP TSVM if the gap between the two lines were enlarged.

4.3.2 Ads Data Set

The average accuracy, class-specific F1-measures, and their standard deviations on the unlabeled test examples for each algorithm across the 10 random splits of the ads data set are shown in Table 2 for different combination of views.

From Table 2, we can see that most algorithms achieve high accuracies, but many of the them score remarkably low F1-measure for the positive class. Considering the minority of positive class in the ads data set (14 percent of samples belong to the positive class), one can conclude that algorithms with low positive class F1-measure actually fail to make the right prediction on the test set. The poor performance of plain SVM is as expected since the model is only trained with a few labeled data in the traditional supervised sense. The failure of cotraining is probably because the ads data set violates its key assumption that the subfeatures are sufficiently good and conditionally independent.

Compared to other algorithms, our Two-view TSVM achieves consistently higher accuracy and F1-measures in most cases. It is also noted that Two-view TSVM with Laplacian graph regularization yields better results than the plain Two-view TSVM. The improvement on the positive class's F1-measure is significant, e.g., 70.10 percent for Two-view TSVM with Laplacian graph regularization versus 45.15 percent for CCCP TSVM for Image URL view in Table 2a. These results show that the proposed two-view semi-supervised learning algorithm not only performs more accurately, but also achieves considerably more stable results than the regular single-view learning approach.

4.3.3 WebKB Course Data Set

All the compared algorithms were run over 100 random splits of the WebKB course data set. Each split contains 12 labeled and 1,039 unlabeled examples. The test results on the unlabeled examples are shown in Table 3.

The results for the various methods are similar to those of the ads data set. Specifically, the positive F1 measure of Two-view TSVM is about 10 percent better than that of the runner up (e.g., 83.28 percent for Two-view TSVM versus 73.76 percent for CCCP TSVM on the link view). Further, the variation (standard deviation shown in brackets) in all of the results for Two-view TSVM is on average three to four times lower than that of the CCCP TSVM. These results show that the proposed Two-view TSVM performs not only more accurately but also achieves considerably more stable results than the regular single-view approach.

Fig. 4 depicts the detailed F1-measures of both positive and negative classes over 100 random splits of the WebKB course data set for both the CCCP TSVM, Two-view TSVM, and Laplacian SVM. It can be seen that the performance of CCCP TSVM and Laplacian SVM is rather unstable, oscillating between zero and nonzero F1 values. This

TABLE 2
Ads Data Set Result Showing Mean Accuracy, F1-Measure (Percent) and Their Standard Deviation (in Brackets)

(a) Image URL View and Destination URL View

	Image URL View			Destination URL View			Hybrid View		
	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1
SVM (32 labeled)	87.01 (2.22)	16.90 (24.33)	92.93 (1.14)	90.02 (2.06)	49.44 (19.02)	94.44 (1.14)	90.41 (1.99)	49.53 (19.99)	94.69 (1.02)
SVM-2K (32 labeled)	83.83 (12.70)	41.91 (17.57)	89.88 (10.11)	92.03 (1.10)	64.82 (5.79)	95.50 (0.62)	91.91 (0.84)	64.13 (6.51)	95.44 (0.45)
Co-training	81.87 (4.43)	39.05 (10.06)	89.29 (2.83)	83.98 (3.92)	43.20 (9.76)	90.63 (2.46)	83.15 (4.01)	42.95 (9.94)	90.87 (2.84)
Laplacian SVM	87.59 (1.63)	25.02 (15.48)	93.23 (0.86)	89.07 (1.27)	35.73 (12.35)	94.02 (0.65)	89.35 (1.01)	36.07 (13.46)	92.55 (1.91)
CCCP TSVM	88.01 (3.58)	45.15 (20.05)	93.20 (2.23)	90.91 (1.66)	61.35 (7.94)	94.83 (0.99)	91.69 (2.02)	60.49 (13.60)	95.34 (1.13)
Two-view TSVM	**90.17 (7.08)**	**66.27 (11.38)**	**94.14 (4.78)**	**91.82 (6.56)**	**72.38 (10.20)**	**95.08 (4.51)**	**92.13 (6.71)**	**73.99 (11.33)**	95.27 (4.54)
Two-view TSVM with Laplacian	**92.53 (1.73)**	**70.10 (5.62)**	**95.73 (1.02)**	**93.69 (1.93)**	**75.19 (6.46)**	**96.38 (1.14)**	**93.81 (2.09)**	**76.31 (7.00)**	**96.44 (1.23)**

(b) Destination URL View and Alt View

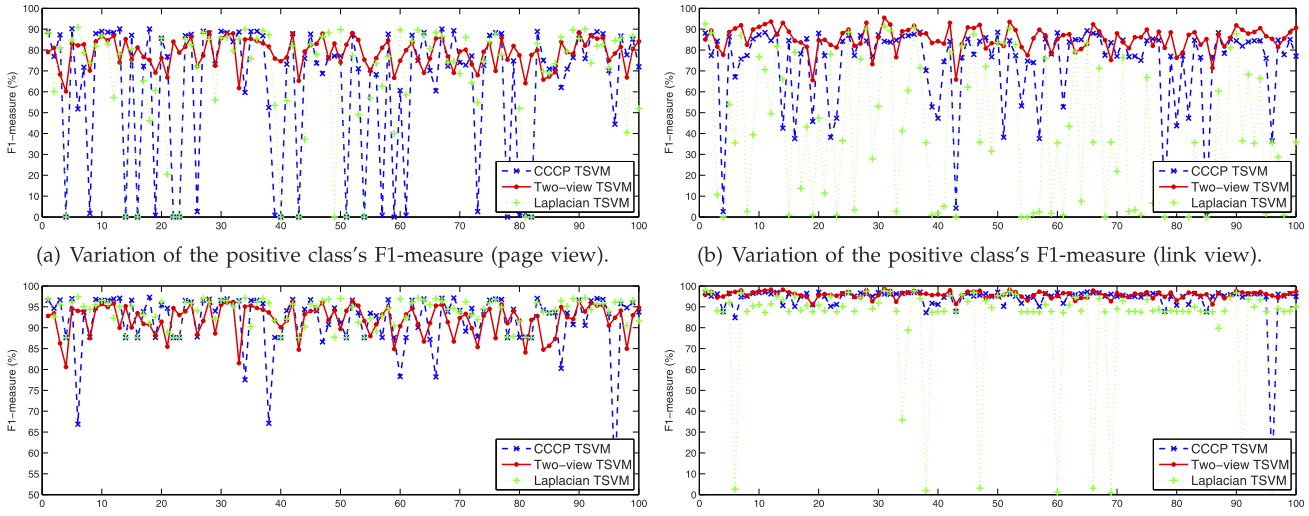
	Destination URL View			Alt View			Hybrid View		
	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1
SVM (32 labeled)	90.02 (2.06)	49.44 (19.02)	94.44 (1.14)	87.73 (1.47)	28.48 (20.54)	93.27 (0.72)	90.48 (2.05)	52.14 (20.02)	94.69 (1.06)
SVM-2K (32 labeled)	92.59 (1.02)	68.11 (4.76)	95.81 (0.59)	87.98 (0.87)	45.80 (1.78)	93.23 (0.55)	91.24 (1.03)	66.04 (4.58)	94.96 (0.61)
Co-training	81.64 (4.07)	44.30 (5.95)	88.97 (2.68)	83.27 (3.38)	37.37 (5.50)	90.31 (2.17)	83.87 (4.19)	40.55 (5.21)	90.18 (2.87)
Laplacian SVM	89.07 (1.27)	35.73 (12.35)	94.02 (0.65)	88.96 (0.41)	39.48 (5.14)	93.92 (0.20)	90.23 (1.51)	37.54 (11.01)	94.67 (0.57)
CCCP TSVM	90.91 (1.66)	61.35 (7.94)	94.83 (0.99)	88.65 (0.95)	44.30 (1.30)	93.67 (0.60)	90.69 (1.94)	61.39 (7.51)	94.69 (1.16)
Two-view TSVM	**91.06 (1.52)**	**67.27 (2.97)**	94.81 (0.97)	86.78 (1.72)	42.37 (3.32)	92.52 (1.08)	**90.97 (1.79)**	**68.25 (3.17)**	94.73 (1.15)
Two-view TSVM with Laplacian	**92.64 (0.64)**	**71.02 (2.23)**	**95.78 (0.38)**	88.57 (0.97)	**45.02 (2.04)**	93.62 (0.59)	**92.27 (0.89)**	**71.18 (2.28)**	**95.53 (0.55)**

happens when every test example is classified into one class (despite the balancing constraint (4f) is also imposed on CCCP TSVM).

On the contrary, by simultaneously training two transductive SVMs based on two views, the Two-view TSVM successfully overcomes this problem. In fact, the F1-measure for Two-view TSVM remains relatively stable, regardless of changes in the training/test data. Since the amount of labeled data in semi-supervised learning is relatively small, there are always variations in the small training set. The variability among training examples is considered one of the primary sources of errors in a classifier. By requiring two classifiers to agree with each other, the structure learned from each view can reinforce one another, and the effect of large variations in the training set can be reduced. Further, the hybrid classifier output is a weighted sum of the individual classifier outputs, which effectively reduces the probability of large swings; any major disagreement between the two view classifiers is essentially averaged out after the linear combination.

TABLE 3
WebKB Course Data Set Result Showing Mean Accuracy, F1-Measure (percent) and Their Standard Deviation (in Brackets)

	Page View			Link View			Hybrid View		
	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1
SVM (12 labeled)	71.17 (17.22)	5.74 (15.88)	80.65 (19.35)	73.81 (17.20)	6.47 (16.05)	82.13 (20.98)	75.00 (14.71)	10.41 (21.87)	83.75 (16.56)
SVM-2K (12 labeled)	88.90 (5.83)	73.07 (13.07)	92.76 (5.79)	88.73 (10.03)	72.68 (11.47)	92.03 (13.12)	90.87 (9.10)	78.56 (11.43)	93.60 (10.69)
Co-training	62.46 (17.30)	22.05 (19.81)	70.98 (20.27)	62.94 (16.41)	21.84 (20.16)	72.03 (17.83)	63.80 (15.28)	22.85 (19.77)	73.84 (18.27)
Laplacian SVM	89.49 (5.77)	64.82 (29.46)	93.73 (3.16)	78.90 (18.42)	39.71 (33.72)	83.20 (24.66)	87.95 (8.17)	52.52 (39.95)	93.02 (4.51)
CCCP TSVM	87.33 (9.04)	62.76 (33.50)	91.78 (7.07)	90.65 (7.98)	73.76 (20.27)	93.85 (8.88)	91.22 (8.10)	77.17 (20.83)	94.22 (6.61)
Two-view TSVM	**88.67 (5.46)**	**78.71 (7.99)**	**92.23 (4.07)**	**93.04 (2.84)**	**83.28 (7.13)**	**95.60 (1.78)**	**92.49 (3.09)**	**83.80 (5.90)**	**95.10 (2.10)**
Two-view TSVM with Laplacian	**89.94 (5.31)**	**80.79 (7.92)**	**93.14 (3.93)**	**93.64 (2.68)**	**84.89 (6.68)**	**95.97 (1.69)**	**93.23 (2.74)**	**85.34 (5.35)**	**95.59 (1.84)**



(a) Variation of the positive class's F1-measure (page view).

(b) Variation of the positive class's F1-measure (link view).

(c) Variation of the negative class's F1-measure (page view).

(d) Variation of the negative class's F1-measure (link view).

Fig. 4. Variation of the positive and negative classes' F1-measures over 100 splits of the WebKB course data set.

TABLE 4
Product Rev Data Set Result Showing Mean Accuracy, F1-Measure (percent) and Their Standard Deviation (in Brackets)

	Lexical View			Formal View			Hybrid View		
	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1	Acc	Pos F1	Neg F1
SVM (20 labeled)	68.10 (10.97)	65.88 (13.91)	69.01 (10.68)	55.02 (20.05)	51.43 (28.56)	56.12 (13.74)	72.95 (7.52)	72.42 (10.51)	72.55 (7.43)
SVM-2K (20 labeled)	74.22 (5.24)	73.97 (6.13)	73.63 (5.66)	70.80 (4.76)	66.32 (9.99)	72.74 (5.61)	75.18 (4.95)	74.91 (6.17)	74.87 (5.64)
Co-training	70.15 (12.98)	70.86 (13.42)	69.36 (12.52)	66.83 (9.07)	64.95 (23.37)	62.88 (14.91)	78.65 (13.80)	69.01 (24.35)	68.04 (15.22)
Laplacian SVM	58.92 (7.88)	40.48 (25.61)	60.15 (23.10)	69.36 (7.49)	68.08 (8.30)	69.86 (8.84)	66.01 (5.79)	48.82 (19.98)	65.27 (7.81)
CCCP TSVM	76.30 (2.27)	74.50 (2.48)	77.85 (2.21)	74.39 (6.03)	73.74 (6.10)	74.78 (6.60)	75.92 (2.95)	77.21 (2.32)	74.30 (4.21)
Two-view TSVM	**79.98 (4.80)**	**79.35 (6.70)**	**80.23 (5.05)**	**74.78 (5.35)**	**74.39 (5.29)**	**75.03 (5.77)**	**77.43 (5.41)**	77.29 (5.51)	**77.46 (5.74)**
Two-view TSVM with Laplacian	**80.42 (3.99)**	**80.14 (4.18)**	**80.57 (4.31)**	**74.97 (5.36)**	**74.58 (5.26)**	**75.22 (5.84)**	**77.55 (5.37)**	77.48 (5.27)	**77.52 (5.80)**

4.3.4 *Product Review Data Set*

The experimental results on the product review data set are summarized in Table 4. From the table, we can observe that the proposed Two-view TSVM algorithm achieves the best accuracy among all the compared algorithms. To assess the importance of unlabeled data in situations where labeled data are really sparse, we evaluate the performances of CCCP TSVM versus Two-view TSVM by varying the number of labeled data instances from 20 to 1,000. Fig. 5 plots accuracy versus number of labeled data for CCCP TSVM and Two-view TSVM. As expected, both algorithms improve with increasing number of labeled examples. Further, we found that when the number of labeled data is very small, e.g., 20, the performance of the Two-view TSVM is significantly better (around 5 percent better accuracy) than the best CCCP TSVM. When the amount of labeled data increases, both algorithms performed more or less in the same ballpark. From the figure, we can conclude that the Two-view TSVM shines when the amount of labeled data is very small, but it also slightly outperforms the single-view classifiers as the amount of labeled data increases. Therefore, it is safe to employ the Two-view TSVM regardless of the amount of labeled data at hand, as it always produces comparable or better results than a classifier trained on a single view.

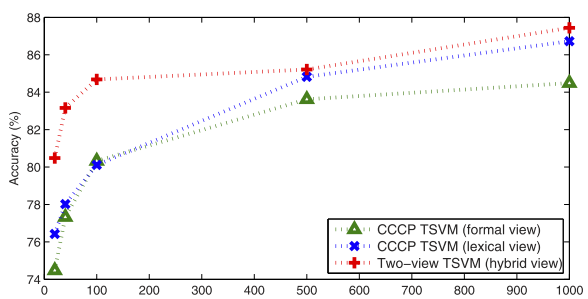


Fig. 5. Accuracy versus the number of labeled examples for CCCP TSVM and Two-view TSVM on the product review data set.

5 Conclusion

In this paper, we investigate the problem of multiview semi-supervised learning, and propose a two-view transductive SVM technique, which is able to take advantage of multiple representations of the same problem to achieve an improvement in classification performance for problems lacking labeled data. Our technique was motivated by

extending the existing two-view supervised learning algorithm into a semi-supervised learning setting. We also incorporate the idea of graph-based semi-supervised learning into our algorithm by utilizing the intrinsic manifold structure of the data. We formulate our learning framework into an optimization problem and present an effective way to solve it. Experimental results on both synthetic and real-life data sets validate the efficacy of the proposed two-view transductive learning algorithm when comparing with the state-of-the-art single-view/multiview supervised/semi-supervised learning approaches. In particular, the proposed technique always makes the classifier more accurate and stable by requiring two views to maintain a maximum consensus on both labeled and unlabeled data. Further, incorporating additional manifold information can also help to improve the classification results.

Our two-view transductive SVM was also partially motivated by the need to detect spam product reviews from online forums. Experimental results were promising on the review spam detection task: a model trained with a few labeled data using our algorithm is comparable to one trained on a significantly larger amount of labeled data using the supervised learning approach. The task of product review mining can be enhanced by applying our method to detect and filter spam reviews.

Many interesting open questions remain. For example, it is unknown in what conditions multiview learning approach is to be preferred to a concatenated hybrid-view learning approach. Given examples represented by a set of features, how to split features into two or multiple views so that multiview learning approach can achieve better result than single-view learning based on the original feature set. Lastly, alternative ways to enforce or balance the consensus between the two views can be further studied.

ACKNOWLEDGMENTS

This research was supported in part by Singapore Ministry of Education's Academic Research Fund Tier 1 grant RG 30/09, and Tier 2 grant (T208B2203). A short version of this paper [1] appeared in the Proceedings of the SIAM International Conference on Data Mining (SDM '10).

REFERENCES

[1] G. Li, S.C.H. Hoi, and K. Chang, "Two-View Transductive Support Vector Machines," *Proc. 10th SIAM Int'l Conf. Data Mining (SDM '10),* pp. 235-244, 2010.

[2] V.N. Vapnik, *Statistical Learning Theory.* Wiley-Interscience, 1998.

[3] O. Chapelle and A. Zien, "Semi-Supervised Classification by Low Density Separation," *Proc. 10th Int'l Workshop Artificial Intelligence and Statistics (AISTATS '05),* pp. 57-64, 2005.

[4] J.D.R. Farquhar, D.R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák, "Two View Learning: Svm-2k, Theory and Practice," *Proc. Advances in Neural Information Processing Systems (NIPS),* 2005.

[5] M. Belkin, P. Niyogi, and V. Sindhwani, "On Manifold Regularization," *Proc. 10th Int'l Workshop Artificial Intelligence and Statistics (AISTAT '05),* pp. 17-24, 2005.

[6] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the Point Cloud: From Transductive to Semi-Supervised Learning," *Proc. 22nd Int'l Conf. Machine Learning (ICML '05),* pp. 824-831, 2005.

[7] K. Dave, S. Lawrence, and D.M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. 12th Int'l World Wide Web Conf. (WWW '03),* pp. 519-528, 2003.

[8] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07),* pp. 334-342, 2007.

[9] N. Jindal and B. Liu, "Opinion Spam and Analysis," *Proc. Int'l Conf. Web Search and Web Data Mining (WSDM '08),* pp. 219-230, 2008.

[10] X. Zhu, "Semi-Supervised Learning Literature Survey," Technical Report 1530, Computer Sciences, Univ. of Wisconsin-Madison, 2005.

[11] O. Chapelle, V. Sindhwani, and S.S. Keerthi, "Optimization Techniques for Semi-Supervised Support Vector Machines," *J. Machine Learning Research,* vol. 9, pp. 203-233, 2008.

[12] R. Collobert, F.H. Sinz, J. Weston, and L. Bottou, "Large Scale Transductive Svms," *J. Machine Learning Research,* vol. 7, pp. 1687-1712, 2006.

[13] O. Chapelle, M. Chi, and A. Zien, "A Continuation Method for Semi-Supervised Svms," *Proc. 23rd Int'l Conf. Machine Learning (ICML '06),* pp. 185-192, 2006.

[14] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. 16th Int'l Conf. Machine Learning (ICML '99),* pp. 200-209, 1999.

[15] A.L. Yuille and A. Rangarajan, "The Concave-Convex Procedure (cccp)," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 1033-1040, 2001.

[16] A. Blum and S. Chawla, "Learning from Labeled and Unlabeled Data Using Graph Mincuts," *Proc. 18th Int'l Conf. Machine Learning (ICML '01),* pp. 19-26, 2001.

[17] X. Zhu, Z. Ghahramani, and J.D. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," *Proc. 20th Int'l Conf. Machine Learning (ICML '03),* pp. 912-919, 2003.

[18] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 321-328, 2004.

[19] M. Karlen, J. Weston, A. Erkan, and R. Collobert, "Large Scale Manifold Transduction," *Proc. 25th Int'l Conf. Machine Learning (ICML '08),* pp. 448-455, 2008.

[20] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Machine Learning Research,* vol. 7, pp. 2399-2434, 2006.

[21] M. Belkin and P. Niyogi, "Using Manifold Structure for Partially Labeled Classification," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 929-936, 2002.

[22] F.R.K. Chung, *Spectral Graph Theory,* no. 92. Am. Math. Soc., (CBMS Regional Conf. Series in Math.), 1997.

[23] V.R. de Sa, "Learning Classification with Unlabeled Data," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 112-119, 1994.

[24] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. 11th Ann. Conf. Computational Learning Theory (COLT '98),* pp. 92-100, 1998.

[25] K. Nigam and R. Ghani, "Analyzing the Effectiveness and Applicability of Co-Training," *Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM '00),* pp. 86-93, 2000.

[26] U. Brefeld and T. Scheffer, "Co-em Support Vector Learning," *Proc. 21st Int'l Conf. Machine Learning (ICML '04),* pp. 121-128, 2004.

[27] W.W. 0028 and Z.-H. Zhou, "A New Analysis of Co-Training," *Proc. 27th Int'l Conf. Machine Learning (ICML '10),* pp. 1135-1142, 2010.

[28] Ü. Güz, S. Cuendet, D. Hakkani-Tür, and G. Tür, "Multi-View Semi-Supervised Learning for Dialog Act Segmentation of Speech," *IEEE Trans. Audio, Speech and Language Processing,* vol. 18, no. 2, pp. 320-329, Feb. 2010.

[29] C. Christoudias, R. Urtasun, and T. Darrell, "Multi-View Learning in the Presence of View Disagreement," *Proc. 24th Conf. Uncertainty in Artificial Intelligence (UAI '08),* pp. 88-96, 2008.

[30] I. Muslea, S. Minton, and C.A. Knoblock, "Active + Semi-Supervised Learning = Robust Multi-View Learning," *Proc. 19th Int'l Conf. Machine Learning (ICML '02),* pp. 435-442, 2002.

[31] S. Yu, B. Krishnapuram, R. Rosales, H. Steck, and R.B. Rao, "Bayesian Co-Training," *Proc. Advances in Neural Information Processing Systems (NIPS),* pp. 1665-1672, 2007.

[32] V. Sindhwani and P. Niyogi, "A Co-Regularized Approach to Semi-Supervised Learning with Multiple Views," *Proc. ICML Workshop Learning with Multiple Views,* pp. 74-79, 2005.

[33] V. Sindhwani and D.S. Rosenberg, "An Rkhs for Multi-View Learning and Manifold Co-Regularization," *Proc. 25th Int'l Conf. Machine Learning (ICML '08)*, pp. 976-983, 2008.
[34] O.-A. Maillard and N. Vayatis, "Complexity versus Agreement for Many Views," *Proc. 20th Int'l Conf. Algorithmic Learning Theory (ALT '09)*, pp. 232-246, 2009.
[35] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel, "Efficient Co-Regularised Least Squares Regression," *Proc. 23rd Int'l Conf. Machine Learning (ICML '06)*, pp. 137-144, 2006.
[36] D.R. Hardoon, S. Szedmák, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, vol. 16, no. 12, pp. 2639-2664, 2004.
[37] S. Szedmák and J. Shawe-Taylor, "Synthesis of Maximum Margin and Multiview Learning Using Unlabeled Data," *Neurocomputing*, vol. 70, nos. 7-9, pp. 1254-1264, 2007.
[38] J. Nocedal and S.J. Wright, *Numerical Optimization*. Springer, 2000.
[39] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, first ed. Athena Scientific, (Optimization and Neural Computation Series), 1996.
[40] N. Kushmerick, "Learning to Remove Internet Advertisements," *Proc. Third Int'l Conf. Autonomous Agents*, pp. 175-181, 1999.
[41] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *Science*, vol. 2, pp. 1-39, http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

Guangxia Li received the bachelor's degree in computer science from Northwestern Polytechnical University, Xi'an, P.R. China, in 2006. He is currently working toward the PhD degree in the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests are information retrieval, data mining, and statistical machine learning. His work currently focuses on statistical and graph models, and their applications to real-world problems in text and web mining, sentiment analysis, and social network analysis.

Kuiyu Chang received the BS degree from National Taiwan University, MS degree from the University of Hawaii at Manoa, and the PhD degree from the University of Texas at Austin, all in electrical/computer engineering. He is an assistant professor of computer engineering at Nanyang Technological University, Singapore. Prior to that, he served as senior risk management analyst for ClearCommerce (now eFunds). From 2000 to 2002, he was a member of technical staff at Interwoven (now Autonomy). He has served as program cochair for PAISI (2006-2008), and publications chair for PAKDD 2006. He has published more than 50 papers, including in top journals (IEEE PAMI) and conferences (SIGIR, IJCAI, ICDE, ICDM, SDM), and is also recipient of two international best paper awards. Since 2005, he has been leading the Review and Opinion Search Engine (ROSE) project, which aggregates online Chinese sentiments. He consults regularly for IT companies in China, Singapore, and Malaysia.

Steven C.H. Hoi received the bachelor's degree from Tsinghua University, PR China, in 2002, and the PhD degree in computer science and engineering from The Chinese University of Hong Kong in 2006. He is currently an assistant professor of the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests are machine learning and its applications to multimedia information retrieval (image and video retrieval), web and social media search and mining, pattern recognition, and computational finance. He has published more than 80 referred papers in top conferences and journals in his areas. He has served as general cochair for ACM SIGMM Workshops on Social Media (WSM '09, WSM '10, WSM '11), program cochair for The fourth Asian Conference on Machine Learning (ACML '12), book editor for "*Social Media Modeling and Computing*," guest editor for *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, technical program committee member for many international conferences, and external reviewer for many top journals and worldwide funding agencies, including the US National Science Foundation (NSF) and RGC in Hong Kong.