

Two-view Transductive Support Vector Machines

Guangxia Li*

Steven C. H. Hoi†

Kuiyu Chang‡

Abstract

Obtaining high-quality and up-to-date labeled data can be difficult in many real-world machine learning applications, especially for Internet classification tasks like review spam detection, which changes at a very brisk pace. 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 thus propose a novel two-view transductive SVM that takes advantage of both the abundant amount of unlabeled data and their multiple representations to improve the performance of classifiers. The idea is fairly simple: train a classifier on each of the two views of both labeled and unlabeled data, and impose a global constraint that each classifier assigns the same class label to each labeled and unlabeled data. We applied our two-view transductive SVM to the WebKB course dataset, and a real-life review spam classification dataset. Experimental results show that our proposed approach performs up to 5% better than a single view learning algorithm, especially when the amount of labeled data is small. The other advantage of our two-view approach is its significantly improved stability, which is especially useful for noisy real world data.

1 Introduction.

Text classification is an active research problem in data mining and machine learning [21]. The classical text classifier is created by building a machine learning model, e.g., support vector machines (SVM) [15], [27], trained from a collection of labeled data. Unfortunately, in practical problems like review spam classification, up-to-date labeled data are very costly to obtain, 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 unlabeled one. 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 [27]. An intuitive interpretation for the success of transductive SVM is the so-called “cluster assumption” [4]. 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 implement this assumption by exploiting the information of unlabeled data into the SVM optimization procedure.

To improve the performance of the existing transductive SVM, we adopt a multi-view learning approach. In multi-view learning, a classifier is created for each representation or view of the same problem, with each classifier optimized to maximize the overall consensus of their predictions. Where two view representations of the same problem are available, a two-view learning approach typically yields equal or better results than those obtained from either view. In fact, our proposed two-view semi-supervised learning algorithm, called two-view transductive SVM, extends the supervised two-view learning framework of Farquhar et al. [9] to take advantage of the large amount of unlabeled data available. We evaluated the proposed new classification technique on both toy and real-life datasets against single-view semi-supervised and two-view supervised approaches.

Our two-view transductive SVM is practically motivated by the problem of product review filtering, which separates valid product reviews from non-opinioned postings in online forums. This task is in fact a pre-processing step for product review mining, which aims to extract and summarize people’s opinions from product reviews [8]. 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 an analysis of review sentence. Experimental results justified the utility of our method on the product review filtering task together with other general web document classification problem.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents our two-view transductive SVM algorithm. Section 4 gives our experimental results and discussions, and section 5 concludes this paper.

*Nanyang Technological University, ligu0005@ntu.edu.sg

†Nanyang Technological University, chhoi@ntu.edu.sg

‡Nanyang Technological University, askychang@ntu.edu.sg

2 Related Work.

We first review existing work on transductive SVM and multi-view learning algorithms, followed by a brief survey on product review filtering.

2.1 Transductive SVM. The transductive SVM can be viewed as a standard SVM with an extra regularization term defined over the set of unlabeled data [32]. Suppose a training set contains l labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^l$, $y_i = \pm 1$, and u unlabeled examples $\{\mathbf{x}_i\}_{i=l+1}^{l+u}$, $\mathbf{x}_i \in \mathbb{R}^n$. The decision function of SVM has the following form

$$(2.1) \quad f_\theta(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b$$

where $\theta = (\mathbf{w}, b)$ are the parameters of the model, and $\Phi(\cdot)$ is the feature map. The transductive SVM solves the following optimization problem

$$(2.2) \quad \min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l L(y_i f_\theta(\mathbf{x}_i)) + C^* \sum_{i=l+1}^{l+u} L(|f_\theta(\mathbf{x}_i)|)$$

where $L(\cdot) = \max(0, 1 - \cdot)$ is the classical hinge loss for labeled examples as illustrated in Figure 1(a), $L(|\cdot|) = \max(0, 1 - |\cdot|)$ is the symmetric hinge loss for unlabeled examples as illustrated in Figure 1(b), C and C^* are adjustable parameters.

Equation (2.2) is hard to optimize since the objective function is non-convex. To solve this problem, a suite of algorithms have been proposed [3] - [5], [7], [24]. Amongst them, we are particularly interested in the work of Collobert et al. [7], who employed an approximate optimization technique known as the concave convex procedure (CCCP) [29]. CCCP decomposes a non-convex function into convex and concave parts. 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. In CCCP transductive SVM [7], the loss function applied to unlabeled data is called ‘‘ramp loss’’ (Figure 1(c)), which can be expressed as the sum of a hinge loss function (Figure 1(a)) and a concave loss function (Figure 1(d)). 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 $-1 < s \leq 0$.

According to Collobert et al. [7], training a transductive SVM with the CCCP method is equivalent to training a SVM using the hinge loss for labeled data, and the ramp loss for unlabeled data. For a binary classification problem, each unlabeled example is accounted for

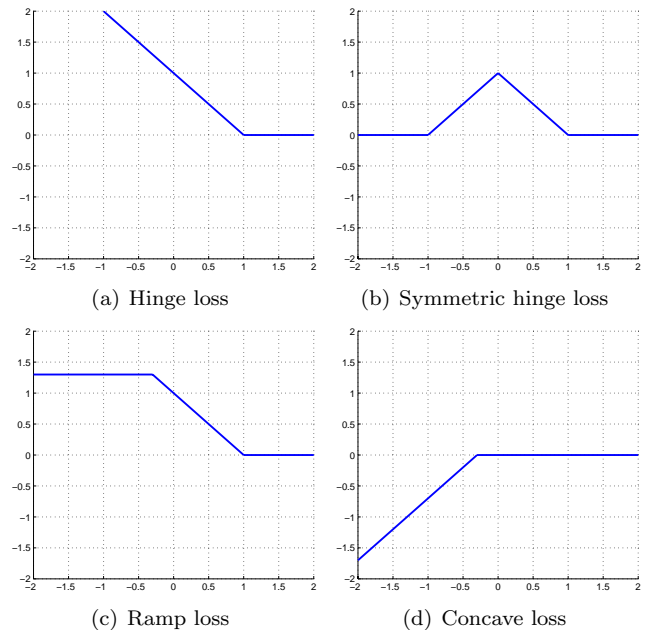


Figure 1: Four kinds of loss function. For ramp loss and concave loss, the parameter s is set to -0.3 .

twice, each time assuming the role of one class, that is, $\{(\mathbf{x}_i, y_i = 1)\}_{i=l+1}^{l+u}$, $\{(\mathbf{x}_i, y_i = -1) : \mathbf{x}_i = \mathbf{x}_{i-u}\}_{i=l+u+1}^{l+2u}$. The corresponding optimization problem of CCCP transductive SVM is given by

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l L(y_i f_\theta(\mathbf{x}_i)) + C^* \sum_{i=l+1}^{l+2u} R_s(y_i f_\theta(\mathbf{x}_i))$$

2.2 Multi-view Learning. Multi-view learning utilizes the agreement among learners trained on different representations of the same problem to improve the overall classification result. The basic idea of using two views with unlabeled data was first proposed by Sa [20]. Blum et al. [2] devised the co-training algorithm that bootstraps a set of classifiers defined on two views by training both with high confidence labeled data. Sindhwani et al. [23] defined a co-regularization approach that learns a multi-view classifier from partially labeled data using a view consensus based regularization term. In particular, we focus on the work of Farquhar et al. [9]. They observed that when two views of the same problem are available, applying the kernel canonical correlation analysis (KCCA) [10] 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 introduced into the optimization is

$$|f_{\theta}^A(\mathbf{x}_i^A) - f_{\theta}^B(\mathbf{x}_i^B)| \leq \eta_i + \varepsilon$$

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

2.3 Product Review Filtering. Both classification and regression based methods have been applied to rate product reviews. Liu et al. [18] detected low-quality product reviews (spam review) with a binary classifier trained from manually annotated data. Some researchers used regression models to predict the utility scores of product reviews crawled from the Amazon website [13], [17], [31]. A multitude of semantic features, term statistics, and metadata of review sentence were extracted to build the regression model. Since the ground-truth of review quality is difficult to obtain, humans were asked to vote on each review’s helpfulness, which were subsequently used to train the ranking models. Jindal et al. [13] solved the problem in a different way: they first tried to recognize duplicate reviews, then treated duplicate reviews as spam reviews to train the model. Their approach was based on the assumption that a large number of duplicate reviews constitute many types of spam reviews if not all. But as Pang et al. [19] stated, the assumption that duplicate reviews constitute some sort of manipulation attempt is weakened by the fact that duplicate reviews in Amazon can be largely attributed to Amazon’s own cross-posting mechanism. We thus propose 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 problem.

3 Two-view Transductive SVM.

3.1 Motivation. We extend the two-view supervised learning algorithm proposed by Farquhar et al. [9] by incorporating unlabeled data, making it 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, and train these classifiers 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 struc-

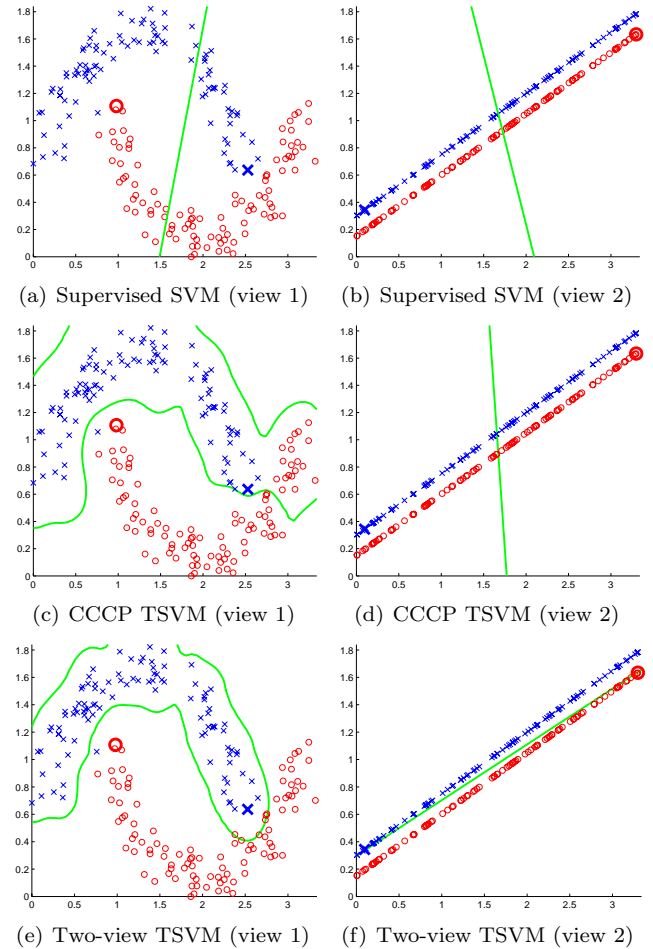


Figure 2: Decision boundaries (denoted by the solid line) obtained by the supervised SVM, CCCP TSVM, and Two-view TSVM. The only two labeled examples are represented by a bold cross and circle. The remaining small points are unlabeled. A Gaussian kernel and linear kernel is used for views 1 and 2, respectively.

ture learnt from each view can reinforce one another. Once trained, a voting or weighting scheme can be used to combine the output from each classifier to classify test samples.

To illustrate the advantage of two view transductive learning, consider a toy dataset in which samples from two classes appear as two moons in one view and two lines in another, as shown in Figure 2 (cross and circle are used to represent the two classes, respectively). Given only two labeled examples (denoted by a bold cross and circle), the solid lines in Figure 2(a) and Figure 2(b) 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 shifts the decision boundary away from dense regions, but still fails to yield a good result in either view (Figure 2(c) and Figure 2(d)). 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 Figure 2(e) and Figure 2(f), in which the solid decision boundary clearly separates the two classes of data.

3.2 Derivation of the Optimization Problem.

Consider a multi-view semi-supervised learning problem containing a set of l labeled examples $\{(\mathbf{x}_i^A, \mathbf{x}_i^B), y_i\}_{i=1}^l$, $y_i = \pm 1$, and a set of u unlabeled examples $\{\mathbf{x}_i^A, \mathbf{x}_i^B\}_{i=l+1}^{l+u}$, $\mathbf{x}_i^{A/B} \in \mathbb{R}^n$. Superscripts A and B denote the two views, respectively. For each view, we aim to find a decision function $f_\theta(\mathbf{x})$ as shown in equation (2.1).

According to Collobert et al. [7], for each view, the CCCP transductive SVM has the following objective function

$$J(\theta^{A/B}) = \frac{1}{2} \|\mathbf{w}^{A/B}\|^2 + C^{A/B} \sum_{i=1}^l \xi_i^{A/B} +$$

$$C^{*A/B} \sum_{i=l+1}^{l+2u} \xi_i^{A/B} + \sum_{i=l+1}^{l+2u} \rho_i^{A/B} y_i f_\theta^{A/B}(\mathbf{x}_i^{A/B})$$

where $\rho_i^{A/B}$ is related to the derivative of the concave loss function mentioned in Section 2.1, written as

(3.3)

$$\rho_i^{A/B} = \begin{cases} C^{*A/B} & \text{if } y_i f_\theta^{A/B}(\mathbf{x}_i^{A/B}) < s \text{ and } i \geq l+1 \\ 0 & \text{otherwise} \end{cases}$$

where s is the parameter of the loss function.

Following Farquhar et al. [9], we add a regularizer that penalizes the decision functions of each view if it deviates from the consensus, and minimize them simultaneously. This leads to the following minimization problem of two-view transductive SVM

$$(3.4a) \quad \min_{\theta^{A/B}, \xi^{A/B}, \eta} J(\theta^A) + J(\theta^B) + D \sum_{i=1}^{l+2u} \eta_i$$

$$(3.4b) \quad \text{s.t.} \quad y_i f_\theta^{A/B}(\mathbf{x}_i^{A/B}) \geq 1 - \xi_i^{A/B}$$

$$(3.4c) \quad \xi_i^{A/B} \geq 0$$

$$(3.4d) \quad |f_\theta^A(\mathbf{x}_i^A) - f_\theta^B(\mathbf{x}_i^B)| \leq \eta_i + \varepsilon$$

$$(3.4e) \quad \eta_i \geq 0$$

$$(3.4f) \quad \frac{1}{u} \sum_{i=l+1}^{l+u} f_\theta^{A/B}(\mathbf{x}_i^{A/B}) = \frac{1}{l} \sum_{i=1}^l y_i$$

where constraint (3.4b) and (3.4c) are the standard SVM constraints, constraint (3.4d) and (3.4e) impose the consensus between the two views, and constraint (3.4f) is a balancing constraint that aims to prevent an extremely skewed classification result caused by assigning all unlabeled examples to only one class. It has been previously used in [4], [7].

By introducing the Lagrange multipliers $\alpha_0^{A/B}$, $\alpha^{A/B}$, $\beta^{+/-}$, $\gamma^{A/B}$ and δ for constraint (3.4f), (3.4b), (3.4d), (3.4c), and (3.4e) respectively, and applying the usual Lagrange multiplier technique, the minimization problem (3.4) is equivalent to the following problem

$$(3.5a) \quad \min_{\tilde{\alpha}^{A/B}, \beta} \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i)(y_j \tilde{\alpha}_j^A + \beta_j) K_{ij}^A + \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i)(y_j \tilde{\alpha}_j^B - \beta_j) K_{ij}^B -$$

$$\sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^A - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^B - \varepsilon \sum_{i=0}^{l+2u} \beta_i^2$$

$$(3.5b) \quad \text{s.t.} \quad 0 \leq \tilde{\alpha}_i^{A/B} \leq C^{A/B} \quad \forall 1 \leq i \leq l$$

(3.5c)

$$-\rho_i^{A/B} \leq \tilde{\alpha}_i^{A/B} \leq C^{*A/B} - \rho_i^{A/B} \quad \forall l+1 \leq i \leq l+2u$$

$$(3.5d) \quad -D \leq \beta_i \leq D \quad \forall 1 \leq i \leq l+2u$$

(3.5e)

$$\sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i) = 0$$

$$\sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i) = 0$$

where $\tilde{\alpha}_i^{A/B} = \alpha_i^{A/B} - \rho_i^{A/B}$, $\beta_i = \beta_i^+ - \beta_i^-$, $y_0 = 1$, $\beta_0 = 0$, $\rho_i^{A/B}$ is defined by equation (3.3), g_i is given by

$$g_i = \begin{cases} \frac{1}{l} \sum_{j=1}^l y_j & \text{if } i = 0 \\ 1 & \text{otherwise} \end{cases}$$

K_{ij} is the kernel matrix of the form

$$K_{ij} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

\mathbf{x}_0 is defined implicitly by

$$\Phi(\mathbf{x}_0) = \frac{1}{u} \sum_{i=l+1}^{l+u} \Phi(\mathbf{x}_i)$$

To solve the minimization problem (3.5), we employ the augmented Lagrangian technique as Farquhar

et al. [9] 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

$$(3.6a) \quad \min_x f(x)$$

$$(3.6b) \quad \text{s.t. } c_i(x) = 0 \quad \forall 1 \leq i \leq n$$

can be written as

$$(3.7) \quad \min_x f(x) - \sum_{i=1}^n \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^n c_i^2(x)$$

where the first two terms correspond to the Lagrangian and the last term is the penalty for violating the constraint. The minimization problem (3.7) can be solved in an iterative way. In each iteration, λ 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 operation with respect to x . In subsequent iterations, λ and μ are updated; and the process is repeated until some stop criteria is reached.

Let us denote the equality constraints (3.5e) and (3.5f) as h_1 and h_2 , and introduce corresponding Lagrange multipliers λ_1 and λ_2 . We can rewrite the minimization problem (3.5) into the augmented Lagrangian form as follows

$$(3.8a) \quad \min_{\tilde{\alpha}^{A/B}, \beta} \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i)(y_j \tilde{\alpha}_j^A + \beta_j) K_{ij}^A + \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i)(y_j \tilde{\alpha}_j^B - \beta_j) K_{ij}^B - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^A - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^B - \varepsilon \sum_{i=0}^{l+2u} \beta_i^2 - \sum_{i=1}^2 \lambda_i h_i + \frac{\mu}{2} \sum_{i=1}^2 \|h_i\|^2$$

$$(3.8b) \quad \text{s.t. } 0 \leq \tilde{\alpha}_i^{A/B} \leq C^{A/B} \quad \forall 1 \leq i \leq l$$

$$(3.8c) \quad -\rho_i^{A/B} \leq \tilde{\alpha}_i^{A/B} \leq C^{*A/B} - \rho_i^{A/B} \quad \forall l+1 \leq i \leq l+2u$$

$$(3.8d) \quad -D \leq \beta_i \leq D \quad \forall 1 \leq i \leq l+2u$$

where μ is the penalty parameter.

Once the minimization problem (3.8) is solved with the augmented Lagrangian method, the decision functions corresponding to the two views can be calculated

as follows

$$(3.9a) \quad f_{\theta}^A(\mathbf{x}^A) = \sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i) K^A(\mathbf{x}_i^A, \mathbf{x}^A) + b^A$$

$$(3.9b) \quad f_{\theta}^B(\mathbf{x}^B) = \sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i) K^B(\mathbf{x}_i^B, \mathbf{x}^B) + b^B$$

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

$$(3.10) \quad f(\mathbf{x}) = \sigma f_{\theta}^A(\mathbf{x}^A) + (1 - \sigma) f_{\theta}^B(\mathbf{x}^B)$$

with $0 \leq \sigma \leq 1$.

Algorithm 1 summarizes the two-view transductive SVM algorithm. The convergence of the CCCP procedure is described in [7]. A detailed convergence analysis of the Lagrange multiplier iteration, which corresponds to the outer loop of Algorithm 1 can be found in [1]. In our experiments, we observed that Algorithm 1 always converges as long as the parameters are selected appropriately.

Algorithm 1 Two-view Transductive SVM

Require: Labeled and unlabeled data of two views.

Initialize $\tilde{\alpha}^{A/B}$, $\rho^{A/B}$, β , λ and μ .

repeat

Solve the following sub-problem.

repeat

Solve the minimization problem (3.8) with fixed λ^k and μ^k .

Compute $f_{\theta}^{A(t+1)}$ and $f_{\theta}^{B(t+1)}$ by equation (3.9) with the solution of the minimization problem (3.8).

Compute $\rho^{A(t+1)}$ and $\rho^{B(t+1)}$ by equation (3.3) with the value of $f_{\theta}^{A(t+1)}$ and $f_{\theta}^{B(t+1)}$.

Update the lower and upper bounds of $\tilde{\alpha}^{A(t+1)}$ and $\tilde{\alpha}^{B(t+1)}$ by equation (3.8b) and (3.8c).

until $\rho^{A(t+1)} = \rho^{A(t)}$ and $\rho^{B(t+1)} = \rho^{B(t)}$

Update the Lagrange multiplier λ by

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

Update the penalty parameter μ by

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

until $\|\mathbf{h}^k\| \leq \epsilon$

return The decision functions corresponding to two views calculated by equation (3.9).

4 Experimental Results.

In this section, we evaluate the classification performance of our two-view transductive SVM on two real-life datasets: the well-known WebKB course dataset and our own product review dataset.

4.1 Product Review Dataset. Our product review dataset was downloaded from two popular online Chinese cell-phone forums¹. Redundant punctuations and stop words were removed and reviews containing less than four characters were eliminated, since they may not hold enough information for a review mining system. We then manually labeled 1000 true reviews and 1000 spam reviews according to the following criteria. A product review is regarded as useful or non-spam 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) towards a product or product feature, and/or the pros and cons analysis of a product or product feature.

To illustrate the idea, consider the following cell-phone forum snippet.

EXAMPLE 4.1. *This cell phone works better than any I’ve ever had. I really like it.*

EXAMPLE 4.2. *I will buy this phone to try it out.*

EXAMPLE 4.3. *How is the battery life of this cell phone?*

Based on our criteria, Example 4.1 is a useful product review with positive opinion on the product, whereas Example 4.2 and 4.3 are spam reviews since they fail to comment on the product.

We treat the product review filtering task as a binary classification problem. To train the classifier, we define two sets of features: one based on the review content (which we call the *lexical view*) and the other based on the characteristics of the review sentences (called *formal view*). For the lexical view, since there is no space separator between Chinese words, raw reviews were preprocessed by a Chinese lexical analyzer — ICTCLAS². ICTCLAS performs 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 below.

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.	–

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

1. Proportion of opinion-bearing phrases in the review sentence. Opinion-bearing phrase refers to adjectives and other terms that are used to express subjective opinions. A review that contains many opinion-bearing phrases is likely to contain comments on the product. We manually crafted a list of dictionary phrases to detect the opinion-bearing phrase in review sentences.
2. Proportion of questioning patterns in the review sentence. Prior work has found that the linguistic style in which reviews are written is a good indicator of spam/non-spam reviews [31]. We extracted the number of questioning patterns from a review. The questioning pattern dictionary includes common Chinese question structures.
3. Proportion of numerical digits in the review sentence. Empirical observations suggest that sentences containing many numerical digits tend to talk about product pricing or refer to a spammer’s contact number rather than product review. This distinction is useful for detecting spam reviews.
4. Percentage of brand mentions in the review sentence. A high percentage of brand mentions in the sentence suggests that the review is a comment on brand only, or an advertisement, both of which are treated as spam reviews.
5. Length of review sentence. This feature is chosen since advertisements (which are regarded as spam reviews) are typically longer than true reviews. For example, the average length of true reviews in our dataset is 21 characters, whereas that of advertisements is 56.

Figure 3 shows each of the five features along with some sample dictionary terms and the dictionary size information.

¹<http://www.club.mobile.163.com> and <http://www.3533.com>

²<http://www.ictclas.org>

Algorithm	Lexical view	Formal view	Hybrid view
SVM (20 labels)	57.63 (18.98)	69.00 (12.10)	-
SVM-2K (20 labels)	74.22 (5.24)	70.80 (4.76)	76.96 (4.77)
CCCP TSVM	76.30 (2.27)	74.39 (6.03)	-
Two-view TSVM	79.17 (4.91)	74.60 (5.20)	80.43 (5.31)

Table 1: Product review classification results showing mean accuracy (in percentage) and its standard deviation (in brackets). The “Hybrid view” classifier uses a linear combination of both classifier outputs.

To evaluate the discriminative power of our proposed high level features against the text corpus features, we trained a supervised SVM to classify product reviews based only on the lexical or formal view. Ten fold cross validation test accuracy was 89.90% and 85.29% for the lexical and formal views, respectively. This indicates that each of the two views contain sufficient information to train a standalone classifier.

We benchmarked our two-view transductive SVM (Two-view TSVM) against the supervised SVM (trained with a few labeled examples from one view), the supervised two-view SVM — SVM-2K [9], and the single-view transductive SVM — CCCP TSVM [7]. To simulate the sparsity of labeled data, we generated 100 random splits of the product review dataset. Each split contains 20 labeled, 1580 unlabeled, and 400 validation examples. We tuned the parameters for each algorithm with respect to their accuracy on the validation dataset, making use of their labels. For the supervised SVM and SVM-2K, the models were trained with only 20 labeled examples. For the semi-supervised CCCP TSVM and Two-view TSVM, models were trained with 20 labeled and 1580 unlabeled examples. For all algorithms, we used the unlabeled data as the test set.

The average accuracy and its standard deviation for each algorithm across the 100 dataset splits is shown in Table 1. It can be seen from Table 1 that our Two-view TSVM achieves the best accuracy compared to all other methods. To assess the statistical significance of Two-view TSVM result, we performed an unpaired t-test at 5% significance level with CCCP TSVM as reference. The results shown in bold in Table 1 are considered to be statistically significant.

To assess the importance of unlabeled data in situations where labeled data are really sparse, we evaluated the performances of CCCP TSVM versus our Two-view TSVM for increasing amount of labeled data starting from 20 up to 1000. Figure 4 gives a plot of accuracy versus number of labeled data for CCCP TSVM and Two-view TSVM. As expected, both algorithms improve with increasing number of labeled examples. The key thing to note here is that the performance of the Two-view TSVM is around 5%

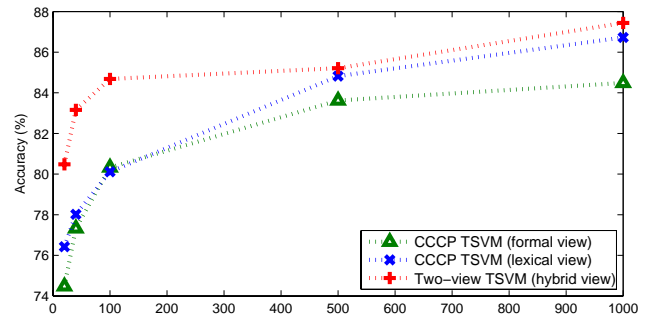


Figure 4: Accuracy versus the number of labeled examples for CCCP TSVM and Two-view TSVM on the product review dataset.

better than the best CCCP TSVM at the lowest amount of labeled data. As the number of labeled data increases, all two 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 outperformed the single-view classifiers as the 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 equal or better results than the classifier trained on a single view.

4.2 WebKB Course Dataset. The WebKB course dataset has been frequently used in the empirical study of multi-view learning since it was first introduced by Blum et al. [2]. The dataset contains 1051 web pages collected from computer science departments of four universities. The task is to classify each page into two classes: course or non-course. 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 that webpage (*link view*). We borrowed a processed WebKB course dataset from Sindhwani et al. [22] and used it in our experiment.

The set of optimal parameters (C and C^*) was chosen from a small range of values, and applied to both the CCCP TSVM and Two-view TSVM. Both algorithms were run over 100 random splits of the

View	Accuracy	Pos F1	Neg F1
Page	87.33 (9.04)	62.76 (33.50)	91.78 (7.07)
Link	90.65 (7.98)	73.76 (20.27)	93.85 (8.88)

Table 2: Average classification performance and standard deviation (in brackets) of CCCP TSVM on the WebKB course dataset.

View	Accuracy	Pos F1	Neg F1
Page	88.36 (4.82)	78.59 (6.86)	91.97 (3.63)
Link	93.53 (2.46)	85.46 (5.60)	95.84 (1.59)
Hybrid	93.55 (2.46)	85.52 (5.59)	95.85 (1.59)

Table 3: Average classification performance and standard deviation (in brackets) of Two-view TSVM on the WebKB course dataset.

WebKB course dataset. Each split contains 12 labeled and 1039 unlabeled examples. Since the distribution of WebKB course dataset is skewed (only 230 of 1051 examples belong to the positive class), we reported the F1-measure besides accuracy. Note that it is always harder for a classifier to achieve a good F1-measure than accuracy on a skewed dataset. The mean and standard deviation results of accuracy and F1-measure on the unlabeled test examples are tallied in Table 2 and Table 3 for CCCP TSVM and Two-view TSVM, respectively.

Compared to CCCP TSVM, Two-view TSVM achieves consistently higher accuracy and F1-measure values. Specifically, performance for the positive class’s F1-measure is more than 11% better (85.52% for Two-view TSVM versus 73.76% for CCCP TSVM). 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. For example, the Two-view TSVM accuracy has a standard deviation of 2.46, versus the standard deviation of 7.98 for CCCP TSVM accuracy. 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.

Figure 5 depicts the detailed F1-measure results of both positive and negative classes over 100 random splits of the test dataset for both the CCCP TSVM and Two-view TSVM algorithms. It can be seen that the performance of CCCP TSVM is rather unstable, which oscillates between zero and non-zero F1-measures. This happens when the CCCP TSVM classifies every test example into one class (despite the balancing constraint (3.4f) is also imposed).

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 learnt 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.

5 Conclusion.

In this paper, we proposed a two-view transductive SVM that is able to take advantage of multiple views of unlabeled data to achieve an improvement in classification performance for problems lacking in labeled data. It is actually an extension of the existing two-view supervised learning algorithm into a semi-supervised setting. In particular, it was motivated by our 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. What’s more, for a general text classification problem, our algorithm is more accurate and stable compared to traditional transductive SVM trained from data in a single view.

For future work, we plan to seek alternative ways to represent the consensus between the two views and examine the performance of our algorithm on more datasets. Further, we are interested to derive the theoretical error bounds of our algorithm, and also study its convergence properties and conditions as well as other efficiency issues. Finally, we may also investigate some state-of-the-art kernel learning techniques [11], [12], [25] for optimizing the kernel functions in building more effective two-view transductive SVM models.

Acknowledgement

We thank David R. Hadoon for generously sharing his SVM-2K source code. The work described in this paper was supported in part by Singapore MOE academic tier-1 research grant (RG67/07).

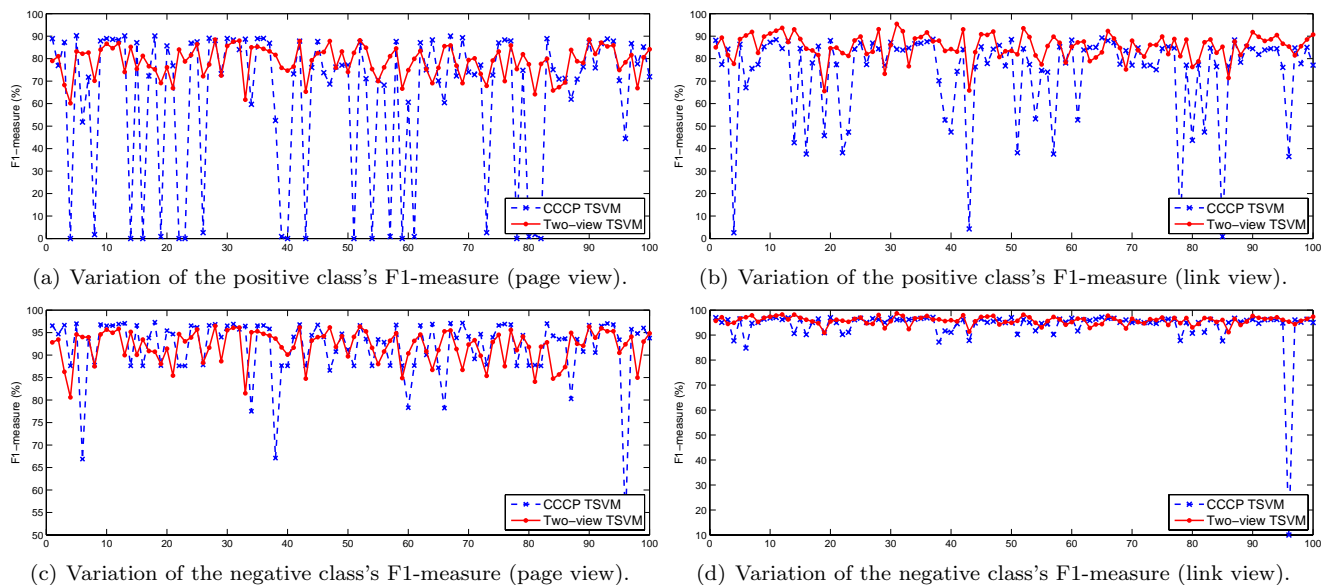


Figure 5: Variation of the positive and negative classes' F1-measures over 100 splits of the WebKB course dataset. This figure shows CCCP TSVM and Two-view TSVM's variation on the positive and negative classes' F1-measures.

References

- [1] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, 1996.
- [2] A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, International Conference on Learning Theory (COLT), 1998, pp. 92–100.
- [3] O. Chapelle, V. Sindhwani, and S. S. Keerthi, *Optimization techniques for semi-supervised support vector machines*, Journal of Machine Learning Research, 9 (2008), pp. 203–233.
- [4] O. Chapelle and A. Zien, *Semi-supervised classification by low density separation*, In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005.
- [5] O. Chapelle, M. Chi, and A. Zien, *A continuation method for semi-supervised SVMs*, International Conference on Machine Learning (ICML), 2006, pp. 185–192.
- [6] C. M. Christoudias, R. Urtasun, and T. Darrell, *Multi-view learning in the presence of view disagreement*, In Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, 2008.
- [7] R. Collobert, F. Sinz, J. Weston, and L. Bottou, *Large scale transductive SVMs*, Journal of Machine Learning Research, 7 (2006), pp. 1687–1712.
- [8] K. Dave, S. Lawrence, and D. M. Pennock, *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*, International World Wide Web Conference (WWW), 2003, pp. 519–528.
- [9] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, *Two view learning: SVM-2K, theory and practice*, Advances in Neural Information Processing Systems (NIPS), 2005.
- [10] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, *Canonical correlation analysis: An overview with application to learning methods*, Neural Computation, 16 (2004), pp. 2639–2664.
- [11] S. C. H. Hoi, M. R. Lyu, and E. Y. Chang, *Learning the unified kernel machines for classification*, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2006, pp. 187–196.
- [12] S. C. H. Hoi, R. Jin, and M. R. Lyu, *Learning non-parametric kernel matrices from pairwise constraints*, International Conference on Machine Learning (ICML), 2007.
- [13] N. Jindal and B. Liu, *Review spam detection*, International World Wide Web Conference (WWW), 2007, pp. 1189–1190.
- [14] N. Jindal and B. Liu, *Opinion spam and analysis*, ACM International Conference on Web Search and Data Mining (WSDM), 2008, pp. 219–230.
- [15] T. Joachims, *Text Categorization with support vector machines: learning with many relevant features*, International Conference on Machine Learning (ICML), 1998, pp. 137–142.
- [16] T. Joachims, *Transductive inference for text classification using support vector Machines*, International Conference on Machine Learning (ICML), 1999, pp. 200–209.
- [17] S. Kim, P. Pantel, T. Chklovski and M. Pennacchiotti, *Automatically assessing review helpfulness*, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006, pp. 423–430.

- [18] J. Liu, Y. Cao, C. Lin, Y. Huang and M. Zhou, *Low-quality product review detection in opinion summarization*, In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 334–342.
- [19] B. Pang and L. Lee, *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
- [20] V. Sa, *Learning classification with unlabeled data*, Advances in Neural Information Processing Systems (NIPS), 1993.
- [21] F. Sebastiani, *Machine learning in automated text categorization*, ACM Computing Surveys, 34(1) (2002), pp. 1–47.
- [22] V. Sindhwani, P. Niyogi, and M. Belkin, *Beyond the point cloud: from transductive to semi-supervised learning*, International Conference on Machine Learning (ICML), 2005, pp. 824–831.
- [23] V. Sindhwani, P. Niyogi, and M. Belkin, *A co-regularization approach to semi-supervised learning with multiple views*, In Proceedings of the 22nd ICML Workshop on Learning with Multiple Views, 2005.
- [24] V. Sindhwani and S. S. Keerthi, *Large scale semi-supervised linear SVMs*, SIGIR, 2006, pp. 477–484.
- [25] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, *Large scale multiple kernel learning*, Journal of Machine Learning Research, 7 (2006), pp. 1531–1565.
- [26] K. Sridharan and S. M. Kakade, *An information theoretic framework for multi-view learning*, International Conference on Learning Theory (COLT), 2008, pp. 403–414.
- [27] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.
- [28] L. Wang, X. Shen, and W. Pan, *On transductive support vector machines*, In J. Verducci, X. Shen, and J. Lafferty, editors, Prediction and Discovery. American Mathematical Society, 2007.
- [29] A. L. Yuille and A. Rangarajan, *The concave-convex procedure (CCCP)*, Advances in Neural Information Processing Systems (NIPS), 2001.
- [30] Y. Yi, D. Xu, F. Nie, J. Luo, and Y. Zhuang, *Ranking with local regression and global alignment for cross media retrieval*, ACM International Conference on Multimedia, 2009, pp. 175–184.
- [31] Z. Zhang and B. Varadarajan, *Utility scoring of product reviews*, ACM Conference on Information and Knowledge Management (CIKM), 2006, pp. 51–57.
- [32] X. Zhu, *Semi-supervised learning literature survey*, Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.
- [33] J. Vittaut, M. Amini, and P. Gallinari, *Learning classification with both labeled and unlabeled Data*, European Conference on Machine Learning (ECML), 2002, pp. 468–479.