Multi-view Maximum Entropy Discrimination

Shiliang Sun and Guoqing Chao
Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, China
Email: slsun@cs.ecnu.edu.cn, guoqingchao10@gmail.com

Abstract

Maximum entropy discrimination (MED) is a general framework for discriminative estimation based on the well known maximum entropy principle, which embodies the Bayesian integration of prior information with large margin constraints on observations. It is a successful combination of maximum entropy learning and maximum margin learning, and can subsume support vector machines (SVMs) as a special case. In this paper, we present a multi-view maximum entropy discrimination framework that is an extension of MED to the scenario of learning with multiple feature sets. Different from existing approaches to exploiting multiple views, such as co-training style algorithms and co-regularization style algorithms, we propose a new method to make use of the distinct views where classification margins from these views are required to be identical. We give the general form of the solution to the multi-view maximum entropy discrimination, and provide an instantiation under a specific prior formulation which is analogical to a multi-view version of SVMs. Experimental results on real-world data sets show the effectiveness of the proposed multi-view maximum entropy discrimination approach.

1 Introduction

Maximum entropy discrimination (MED) is an effective approach to discriminative training of model parameters, and relies on the maximum entropy or minimum relative entropy principle and the maximum margin principle. It can make full use of the merits of generative and discriminative modeling, and has been successfully applied to a large number of machine learning problems.

MED was first presented by Jaakkola et al. [1999], and was applied to anomaly detection and classification involving partially labeled examples. Jebara and Jaakkola [2000] employed MED for feature selection by introducing a selector variable into the discrimination function. Jebara [2004; 2011] extended MED to the problem of multi-task feature and kernel selection. On the theoretical side, Long and Wu [2004] established a mistake bound for an ensemble method for MED and proved a more refined bound that leads to a nearly optimal algorithm for learning disjunctions based on the maximum entropy principle.

Recently, Zhu and Xing [2009] applied MED to structure learning which posses the advantages of probabilistic models and the maximum margin approach. By adopting a Laplace prior, Zhu et al. [2008a] obtained a Laplace maximum margin Markov network which is a sparse model suitable for learning complex structures. Zhu et al. [2008b] also presented a partially observed MED Markov network to deal with the situation where latent variables exist.

Multi-view learning (MVL) is an emerging direction which considers learning with multiple feature sets. Its popularity is mainly motivated by the fact that many real-world data have multiple representations. For example, a web page can be described by words appearing on the web page itself and words underlying all links pointing to the web page from other pages. In multimedia content understanding, multimedia segments can be simultaneously described by their video signals and audio signals. But it should be noted that when there are no natural multiple views, manually generated multiple views can also be helpful [Nigam and Ghani, 2000]. Among the current MVL methods, we can identify two main categories: co-training style algorithms and co-regularization style algorithms.

Co-training style algorithms are inspired by the co-training algorithm [Blum and Mitchell, 1998], which iteratively runs the following procedure untill a termination condition is satisfied: Two learners are separately obtained from two views initially, and then the most confident examples identified by one learner are fed to the other to improve their learning performance. Recently, Yu et al. [2011] proposed a Bayesian undirected graphical model for co-training. Sun and Jin [2011] proposed a robust co-training algorithm, which integrates canonical correlation analysis to examine the predictions of co-training on unlabeled data. In addition to classification, the idea of co-training was later used for clustering [Bickel and Scheffer, 2004; Kumar and Daumé III, 2011]. Theoretical work also attracted many researchers. For example, Dasgupta et al. [2001] gave a PAC-style bound on the generalization error, Balcan et al. [2005] presented a weaker assumption called ϵ-expansion to guarantee the success of co-training, and Wang and Zhou [2010] provided a sufficient and necessary condition for co-training to succeed.

The core idea of co-regularization style algorithms is that minimizing the distinction between the functions of two views acts as one part of the objective function. Representative methods include [Sindhwani *et al.*, 2005; Kumar *et al.*, 2011; Sun, 2011; White *et al.*, 2012]. There are also some theoretical research on co-regularization style algorithms. Rosenberg and Bartlett [2007] provided a tight bound on the Rademacher complexity of the co-regularized hypothesis class in terms of the kernel matrices from each reproducing kernel Hilbert space (RKHS). Sindhwani and Rosenberg [2008] constructed a single, new RKHS which can transform standard supervised algorithms to multi-view semi-supervised algorithms. Sun and Shawe-Taylor [2010] characterized the generalization error of a multi-view SVM in terms of the margin bound and derived the empirical Rademacher complexity of the considered function class.

Inspired by the recent success of MVL, in this paper we extend MED to the MVL setting. But different from the previous two kinds of methods combining multiple views, we propose a new approach to exploiting multiple views. We enforce the margins from two views to be identical to yield a new MVL mechanism, which extends MED to our new framework multi-view maximum entropy discrimination (MVMED). Then we derive the solution to the optimization problem of MVMED, and give one instantiation by using a specific prior formulation.

The rest of the paper is organized as follows. Section 2 briefly reviews MED. Section 3 describes our proposed MVMED. Section 4 reports the experiments on three real-world data sets and makes comparisons. Finally, we give conclusions and point out possible future work in Section 5.

2 Maximum Entropy Discrimination

MED is similar to Bayesian learning in the sense that the posterior of model parameters requires to be inferred. But it integrates the large margin principle and may not need the formulation of generative distributions for data.

Now, we introduce the general learning setup of MED. Suppose we have a data set $\{X_t, y_t\}$ with N examples where X_t indicates the tth input, y_t indicates the corresponding output, and $y_t \in \{\pm 1\}$. If we have two class-conditional probability distributions over the examples, i.e., $p(X_t|\theta_{y_t})$ with parameters θ_{y_t}, the decision rule follows the sign of the discriminant function

$$L(X_t|\boldsymbol{\Theta}) = \log\frac{p(X_t|\theta_1)}{p(X_t|\theta_{-1})} + b, \tag{1}$$

where $\boldsymbol{\Theta} = \{\theta_1, \theta_{-1}, b\}$ includes the model parameters and b is a bias term that can be expressed as a log-ratio of class priors $b = \log\left(p_+/(1-p_+)\right)$ with p_+ being the prior of the positive class. Alternatively, the discriminant function can be directly described by a parameter formulation without any reference to probability models. MED applies well to any of the above two cases.

The general MED is formulated as follows:

$$\begin{cases} \min_{p(\boldsymbol{\Theta},\boldsymbol{\gamma})} \mathrm{KL}\big(p(\boldsymbol{\Theta},\boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta},\boldsymbol{\gamma})\big) \\ \text{s.t.} \int p(\boldsymbol{\Theta},\boldsymbol{\gamma})[y_t L(X_t|\boldsymbol{\Theta}) - \boldsymbol{\gamma}_t] d\boldsymbol{\Theta} d\boldsymbol{\gamma} \geq 0 \\ \quad 1 \leq t \leq N, \end{cases} \tag{2}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_N\}$ specifies the desired classification margins which reflect the large margin principle as in SVMs. Here, instead of seeking a single parameter estimation, MED considers a more general problem of finding a distribution $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ over the parameters and margins, from which we can get the parameter distribution $p(\boldsymbol{\Theta})$ by marginalization. Correspondingly, it uses a convex combination of discriminant functions, i.e., $\int p(\boldsymbol{\Theta}) L(X_t|\boldsymbol{\Theta}) d\boldsymbol{\Theta}$ rather than one single discriminant function to make model averaging for decisions. As Domingos [2000] proved, model averaging can improve the classification performance by means of alleviating the overfitting problem. In addition, the solution of MED is unique as long as it exists since the optimization problem in (2) is convex with respect to $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ [Jaakkola *et al.*, 1999].

After adding a set of dual variables, one for each constraint, the Lagrangian of the optimization problem can be written as

$$\begin{aligned} L = & \int p(\boldsymbol{\Theta},\boldsymbol{\gamma}) \log\frac{p(\boldsymbol{\Theta},\boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta},\boldsymbol{\gamma})} d\boldsymbol{\Theta} d\boldsymbol{\gamma} \\ & - \sum_{t=1}^{N} \int p(\boldsymbol{\Theta},\boldsymbol{\gamma}) \lambda_t [y_t L(X_t|\boldsymbol{\Theta}) - \gamma_t] d\boldsymbol{\Theta} d\boldsymbol{\gamma}. \end{aligned} \tag{3}$$

In order to find the solution to (2), we require

$$\begin{aligned} \frac{\partial L}{\partial p(\boldsymbol{\Theta},\boldsymbol{\gamma})} &= \log\frac{p(\boldsymbol{\Theta},\boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta},\boldsymbol{\gamma})} + 1 \\ &\quad - \sum_{t=1}^{N} \lambda_t [y_t L(X_t|\boldsymbol{\Theta}) - \gamma_t] \\ &= 0, \end{aligned} \tag{4}$$

which results in the following theorem [Jaakkola *et al.*, 1999].

Theorem 2.1 *The solution to the MED problem has the following general form*

$$p(\boldsymbol{\Theta},\boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\lambda})} p_0(\boldsymbol{\Theta},\boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_t [y_t L(X_t|\boldsymbol{\Theta}) - \gamma_t]}, \tag{5}$$

where $Z(\boldsymbol{\lambda})$ is the normalization constant (partition function) and $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_N\}$ defines a set of non-negative Lagrange multipliers, one for each classification constraint. $\boldsymbol{\lambda}$ is set by finding the unique maximum of the following jointly concave objective function

$$J(\boldsymbol{\lambda}) = -logZ(\boldsymbol{\lambda}). \tag{6}$$

Whether the solution to MED can be found depends entirely on whether the partition function $Z(\boldsymbol{\lambda})$ can be evaluated in a closed form, which is given as

$$Z(\boldsymbol{\lambda}) = \int p_0(\boldsymbol{\Theta},\boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_t [y_t L(X_t|\boldsymbol{\Theta}) - \gamma_t]} d\boldsymbol{\Theta} d\boldsymbol{\gamma}. \tag{7}$$

After $\boldsymbol{\lambda}$ is obtained, the following formula is used to predict the label of a new example X

$$\hat{y} = \text{sign}\big(\mathbb{E}_{p(\boldsymbol{\Theta})}[L(X|\boldsymbol{\Theta})]\big). \tag{8}$$

3 Multi-view Maximum Entropy Discrimination

As we have mentioned above, MED incorporates the principles of maximum entropy and maximum margin, which can provide a good justification for its successful applications. In addition, the recent MVL shows that simultaneously using multiple feature sets can further improve the performance over using a single feature set. But as far as we know, there is no research on MVMED yet. Our work in this paper aims to fill the gap and investigate the feasibility of MVMED. We also propose a novel approach to exploiting multiple views which is completely different from existing approaches.

We enforce the margins from two views to be identical, which is a new attempt to combine multiple views. This means that the classification confidences from different views are deemed to match each other exactly. Another benefit of this kind of "margin consistency" is that the solution to MVMED will be convenient to be computed.

Suppose the data set is $\{X_t^1, X_t^2, y_t\}$ with N examples where X_t^1 and X_t^2 indicate the tth input from view 1 and view 2, respectively, and $y_t \in \{\pm 1\}$ is the label. Our MVMED considers a joint distribution of $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$ and $\boldsymbol{\gamma}$ where $\boldsymbol{\Theta}_1 = \{\theta_1, b_1\}$, $\boldsymbol{\Theta}_2 = \{\theta_2, b_2\}$, and the common margin vector $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_N\}$. Formally, the MVMED framework is formulated as follows:

$$\begin{cases} \min_{p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} \text{KL}\big(p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\big) \\ \text{s.t.} \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad 1 \leq t \leq N, \end{cases} \tag{9}$$

where $L_1(X_t^1|\boldsymbol{\Theta}_1)$ and $L_2(X_t^2|\boldsymbol{\Theta}_2)$ are discriminant functions from view 1 and view 2, respectively. The Lagrangian of the optimization problem is

$$\begin{aligned} L = & \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) \log \frac{p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \\ & - \sum_{t=1}^{N} \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) \lambda_{1,t}[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \\ & - \sum_{t=1}^{N} \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) \lambda_{2,t}[y_t L_2(X_t^2|\Theta_2) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}, \end{aligned} \tag{10}$$

where $\boldsymbol{\lambda}_1 = \{\lambda_{1,1}, \ldots, \lambda_{1,N}\}$ and $\boldsymbol{\lambda}_2 = \{\lambda_{2,1}, \ldots, \lambda_{2,N}\}$ are Lagrange multipliers for view 1 and view 2, respectively. In order to find the solution to (9), we require

$$\begin{aligned} \frac{\partial L}{\partial p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} = & \log \frac{p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} + 1 \\ & - \sum_{t=1}^{N} \lambda_{1,t}[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] \\ & - \sum_{t=1}^{N} \lambda_{2,t}[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t] \\ = & \, 0, \end{aligned} \tag{11}$$

which results in the following theorem.

Theorem 3.1 *The solution to the MVMED problem has the following general form*

$$p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)} p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) e^{\left(\sum_{t=1}^{N} \lambda_{1,t}[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] + \sum_{t=1}^{N} \lambda_{2,t}[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t]\right)}, \tag{12}$$

where $Z(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ is the normalization constant and $\boldsymbol{\lambda}_1 = \{\lambda_{1,1}, ..., \lambda_{1,N}\}$, $\boldsymbol{\lambda}_2 = \{\lambda_{2,1}, ..., \lambda_{2,N}\}$ define two sets of non-negative Lagrange multipliers, one for each classification constraint. $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are set by finding the unique maximum of the following jointly concave objective function

$$J(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = -logZ(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2). \tag{13}$$

After $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are obtained, the following two formulae are used to predict the label of a new example (X^1, X^2) from view 1 and view 2, respectively

$$\hat{y}_1 = \text{sign}\Big(\int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) L_1(X^1|\boldsymbol{\Theta}_1) d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big), \tag{14}$$

$$\hat{y}_2 = \text{sign}\Big(\int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) L_2(X^2|\boldsymbol{\Theta}_2) d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big). \tag{15}$$

We can also make predictions by using the two views together

$$\hat{y} = \text{sign}\Big(\frac{1}{2} \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)\big(L_1(X^1|\boldsymbol{\Theta}_1) + L_2(X^2|\boldsymbol{\Theta}_2)\big) d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big). \tag{16}$$

3.1 Instantiation of MVMED

The prior $p_0(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ plays an important role in our MVMED framework as shown in (12). Now we instantiate our MVMED with a concrete prior formulation. Suppose

$$\begin{aligned} p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}) &= p_0(\boldsymbol{\Theta}_1) p_0(\boldsymbol{\Theta}_2) p_0(\boldsymbol{\gamma}) \\ &= p_0(\boldsymbol{\theta}_1) p_0(b_1) p_0(\boldsymbol{\theta}_2) p_0(b_2) p_0(\boldsymbol{\gamma}), \end{aligned} \tag{17}$$

where $p_0(b_1)$, $p_0(b_2)$ approach a non-informative Gaussian prior, $p_0(\boldsymbol{\theta}_1)$, $p_0(\boldsymbol{\theta}_2)$ are both Gaussian distributed with mean $\mathbf{0}$ and identity covariance $\mathbf{I}$, and the prior over the margin constraints $\boldsymbol{\gamma}$ is assumed to be fully factored

$$p_0(\boldsymbol{\gamma}) = \prod_{t=1}^{N} p_0(\gamma_t), \tag{18}$$

with $p_0(\gamma_t) = ce^{-c(1-\gamma_t)}$ and $\gamma_t \leq 1$. A penalty is incurred for margins smaller than $1 - 1/c$ (the prior mean of γ_t) while vanishes otherwise. In fact, this choice of the margin prior corresponds closely to the use of slack variables and additive penalties in SVMs.

Then the normalization constant in (12) can be obtained as

$$
\begin{aligned}
&Z(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \int p_0(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \\
&e^{\Big(\sum_{t=1}^N \lambda_{1,t} \big[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t \big] + \sum_{t=1}^N \lambda_{2,t} \big[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t \big] \Big)} \\
&d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \\
&= \int \mathcal{N}(\boldsymbol{\theta}_1|\mathbf{0}, \mathbf{I}) \mathcal{N}(b_1|\mathbf{0}, {\boldsymbol{\sigma}_1}^2) \mathcal{N}(\boldsymbol{\theta}_2|\mathbf{0}, \mathbf{I}) \mathcal{N}(b_2|\mathbf{0}, {\boldsymbol{\sigma}_2}^2) \\
&\prod_{t=1}^N c e^{-c(1-\gamma_t)} \\
&e^{\Big(\sum_{t=1}^N \lambda_{1,t} \big[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t \big] + \sum_{t=1}^N \lambda_{2,t} \big[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t \big] \Big)} \\
&d\theta_1 d\theta_2 db_1 db_2 d\boldsymbol{\gamma} \\
&= e^{\Big(\frac{1}{2} \sum_{t,\tau=1}^N \lambda_{1,t} \lambda_{1,\tau} y_t y_\tau X_t^1 X_\tau^1 + \frac{1}{2} \sum_{t,\tau=1}^N \lambda_{2,t} \lambda_{2,\tau} y_t y_\tau X_t^2 X_\tau^2 \Big)} \\
&e^{\Big(\frac{{\boldsymbol{\sigma}_1}^2}{2} \big(\sum_{t=1}^N \lambda_{1,t} y_t \big)^2 + \frac{{\boldsymbol{\sigma}_2}^2}{2} \big(\sum_{t=1}^N \lambda_{2,t} y_t \big)^2 \Big)} \\
&\prod_{t=1}^N \Big(\frac{c}{c - \lambda_{1,t} - \lambda_{2,t}} e^{-\lambda_{1,t} - \lambda_{2,t}} \Big),
\end{aligned}
\tag{19}
$$

where we have used the fact that $L_1(X_t^1|\boldsymbol{\Theta}_1) = \boldsymbol{\theta}_1 X_t^1 + b_1$ and $L_2(X_t^2|\boldsymbol{\Theta}_2) = \boldsymbol{\theta}_2 X_t^2 + b_2$. We substitute (19) into (13) and get

$$
\begin{aligned}
J(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = &\sum_{t=1}^N \Big(\lambda_{1,t} + \lambda_{2,t} + \log\big(1 - \frac{\lambda_{1,t} + \lambda_{2,t}}{c}\big) \Big) \\
&- \frac{1}{2} \sum_{t,\tau=1}^N \lambda_{1,t} \lambda_{1,\tau} y_t y_\tau X_t^1 X_\tau^1 \\
&- \frac{1}{2} \sum_{t,\tau=1}^N \lambda_{2,t} \lambda_{2,\tau} y_t y_\tau X_t^2 X_\tau^2 \\
&- \frac{{\boldsymbol{\sigma}_1}^2}{2} \Big(\sum_{t=1}^N \lambda_{1,t} y_t \Big)^2 - \frac{{\boldsymbol{\sigma}_2}^2}{2} \Big(\sum_{t=1}^N \lambda_{2,t} y_t \Big)^2.
\end{aligned}
\tag{20}
$$

Here, $\boldsymbol{\lambda}_1 \geq \mathbf{0}$, $\boldsymbol{\lambda}_2 \geq \mathbf{0}$. Since ${\boldsymbol{\sigma}_1}^2 \to \infty$ and ${\boldsymbol{\sigma}_2}^2 \to \infty$ correspond to using non-informative priors on the bias terms $b_{1,t}$ and $b_{2,t}$, the above dual objective function requires the constraints $\sum_{t=1}^N \lambda_{1,t} y_t = 0$ and $\sum_{t=1}^N \lambda_{2,t} y_t = 0$. Thus we have the following dual optimization problem

$$
\begin{cases}
\max\limits_{\lambda_1, \lambda_2} \sum\limits_{t=1}^N \Big(\lambda_{1,t} + \lambda_{2,t} + \log\big(1 - \frac{\lambda_{1,t} + \lambda_{2,t}}{c}\big) \Big) \\
\qquad - \frac{1}{2} \sum\limits_{t,\tau=1}^N \lambda_{1,t} \lambda_{1,\tau} y_t y_\tau X_t^1 X_\tau^1 \\
\qquad - \frac{1}{2} \sum\limits_{t,\tau=1}^N \lambda_{2,t} \lambda_{2,\tau} y_t y_\tau X_t^2 X_\tau^2 \\
\text{s.t. } \boldsymbol{\lambda}_1 \geq \mathbf{0}, \boldsymbol{\lambda}_2 \geq \mathbf{0} \\
\qquad \sum\limits_{t=1}^N \lambda_{1,t} y_t = 0, \sum\limits_{t=1}^N \lambda_{2,t} y_t = 0.
\end{cases}
\tag{21}
$$

The Lagrange multipliers $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are recovered by solving the convex optimization problem (21), whose non-zero values indicate support vectors. Then the prediction rules for view 1 and view 2 on a new example (X^1, X^2) are respectively

$$
\hat{y}_1 = \text{sign}\Big(\sum_{t=1}^N \lambda_{1,t} y_t X_t^1 X^1 + \hat{b}_1 \Big), \tag{22}
$$

$$
\hat{y}_2 = \text{sign}\Big(\sum_{t=1}^N \lambda_{2,t} y_t X_t^2 X^2 + \hat{b}_2 \Big), \tag{23}
$$

where $\hat{b}_1$ and $\hat{b}_2$ are given by the Karush-Kuhn-Tucker (KKT) conditions using support vectors. If classifiers from two views are combined together to make predictions, the prediction rule can be given analogously from (16).

3.2 Relationship to SVM-2K

This section will discuss the relationship between our instantiation of MVMED and an MVL algorithm SVM-2K [Farquhar *et al.*, 2005]. In order to facilitate the comparison and analysis, we rewrite (21) as (24) by replacing $X_t^1 X_\tau^1$, $X_t^2 X_\tau^2$ with Mercer kernel functions $\kappa(X_t^1, X_\tau^1)$, $\kappa(X_t^2, X_\tau^2)$ and setting $g_{1,t} = \lambda_{1,t} y_t$, $g_{2,t} = \lambda_{2,t} y_t$, and the dual form of SVM-2K is given in (25).

$$
\begin{cases}
\max \sum\limits_{t=1}^N \Big(\lambda_{1,t} + \lambda_{2,t} + \log\big(1 - \frac{\lambda_{1,t} + \lambda_{2,t}}{c}\big) \Big) \\
\qquad - \frac{1}{2} \sum\limits_{t,\tau=1}^N g_{1,t} g_{1,\tau} \kappa(X_t^1, X_\tau^1) \\
\qquad - \frac{1}{2} \sum\limits_{t,\tau=1}^N g_{2,t} g_{2,\tau} \kappa(X_t^2, X_\tau^2) \\
\text{s.t. } g_{1,t} = \lambda_{1,t} y_t, \ g_{2,t} = \lambda_{2,t} y_t, \ 1 \leq t \leq N \\
\qquad \sum\limits_{t=1}^N g_{1,t} = 0 = \sum\limits_{t=1}^N g_{2,t} \\
\qquad \boldsymbol{\lambda}_1 \geq \mathbf{0}, \boldsymbol{\lambda}_2 \geq \mathbf{0}.
\end{cases}
\tag{24}
$$

$$
\begin{cases}
\max & \sum_{t=1}^{N}(\lambda_{1,t}+\lambda_{2,t}) \\
& -\frac{1}{2}\sum_{t,\tau=1}^{N} g_{1,t}g_{1,\tau}\kappa(X_t^1, X_\tau^1) \\
& -\frac{1}{2}\sum_{t,\tau=1}^{N} g_{2,t}g_{2,\tau}\kappa(X_t^2, X_\tau^2) \\
\text{s.t.} & g_{1,t} = \lambda_{1,t}y_t - \beta_t^+ + \beta_t^-,\ g_{2,t} = \lambda_{2,t}y_t + \beta_t^+ - \beta_t^- \\
& \sum_{t=1}^{N} g_{1,t} = 0 = \sum_{t=1}^{N} g_{2,t} \\
& 0 \le \lambda_{1,t} \le C^1,\ 0 \le \lambda_{2,t} \le C^2 \\
& 0 \le \beta_t^{+/-},\ \beta_t^+ + \beta_t^- \le D \\
& 1 \le t \le N.
\end{cases}
\tag{25}
$$

By inspection, we find that compared to each other, (24) has an additional term $\log\left(1 - \frac{\lambda_{1,t}+\lambda_{2,t}}{c}\right)$ in the objective function, while (25) has additional $\beta_t^+ - \beta_t^-$ in $g_{1,t}$ and $g_{2,t}$. In fact, they both play the role of combining two views, though in different fashions. If we set $c \to \infty$ in (24) and set $\beta_t^+ = \beta_t^- = 0, C^1 \to \infty, C^2 \to \infty$ in (25) , the two formulae will be exactly identical. However, it should be noted that our MVMED framework is much more flexible than SVM-2K, since we can reach different instantiations in terms of different prior specifications.

4 Experiments

In this section, we evaluate our proposed MVMED on three real-world data sets: web-page classification, ionosphere classification and advertisement classification.

For all the experiments, given a division of the training and test set, we use one half of the test set as a validation set for parameter selection and the other half for test. The average accuracies obtained by ten random divisions of the training and test sets are reported. The parameter c in MVMED is chosen from $\{2^{-5}, 2^{-4}, ..., 2^5\}$. Two single-view methods MED1 and MED2 are employed to compare with our MVMED. In addition, the MVL method SVM-2K is also used for comparison. For MVMED and SVM-2K, besides the prediction functions $\text{sign}(f_1)$ and $\text{sign}(f_2)$ from the separate views, we also consider the hybrid prediction function $\text{sign}\big((f_1+f_2)/2\big)$ and the one with the highest validation accuracy will be selected.

We first report the average accuracies and standard deviations of the four methods on data sets with only one tenth of the data as the training set in Table 1. Then we increase the training set sizes gradually, and show their performances in Figure 1.

4.1 Web-Page Classification

The data set for this experiment consists of 1051 two-view web pages collected from computer science department web sites at four universities: Cornell University, University of Washington, University of Wisconsin, and University of Texas. There are 230 course pages and 821 non-course pages. The two natural views are words occurring in a web page and words appearing in the links pointing to that page [Blum and Mitchell, 1998; Sindhwani *et al.*, 2005]. The dimensions of the two views are 2333 and 87, respectively. For convenience, we reduce the dimension of view 1 from 2333 to 500 via principal component analysis (PCA).

Clearly, Table 1 indicates that MVMED is superior to single-view MED1, single-view MED2 and SVM-2K. We can also find that SVM-2K performs better than single-view MED1 but worse than single-view MED2. Figure 1(a) with varying training sizes also shows that our MVMED consistently outperforms the other three methods.

4.2 Ionosphere Classification

The ionosphere data set origins from UCI,[1] and includes 351 instances in total which are divided into 225 "good" (positive) instances and 126 "bad" (negative) instances. This data set has only one view, but we generate the other view through PCA. Now, the two views have 35 and 24 dimensions, respectively.

The experimental results are shown in Table 1 and Figure 1(b), from which we can see that MVMED performs the best among all the methods. SVM-2K performs better than single-view MED1 but worse than single-view MED2. From Figure 1(b), we can also find that although MVMED and single-view MED2 need fewer training data to reach a high accuracy, MVMED is more stable than single-view MED2.

4.3 Advertisement Classification

The data set consists of 3279 examples including 459 ads images (positive examples) and 2820 non-ads images (negative examples) [Kushmerick, 1999]. The first view describes the image itself (words in the image's URL, alt text and caption), while the other view contains all other features (words from the URLs of the pages that contain the image and the image points to). Here, we randomly select 600 examples therein to form the used data set.

From Table 1 and Figure 1(c), we can find that our method MVMED performs better than all the other methods. SVM-2K performs the worst in the beginning, but then performs nearly as well as single-view MED1, though still behaves worse than single-view MED2 and MVMED.

4.4 Summary

From the above experiments, we can find that our MVMED performs the best. Another MVL method SVM-2K performs not as good as the MVMED, and sometimes is even worse than single-view MEDs. In a nutshell, the new MVMED framework is effective.

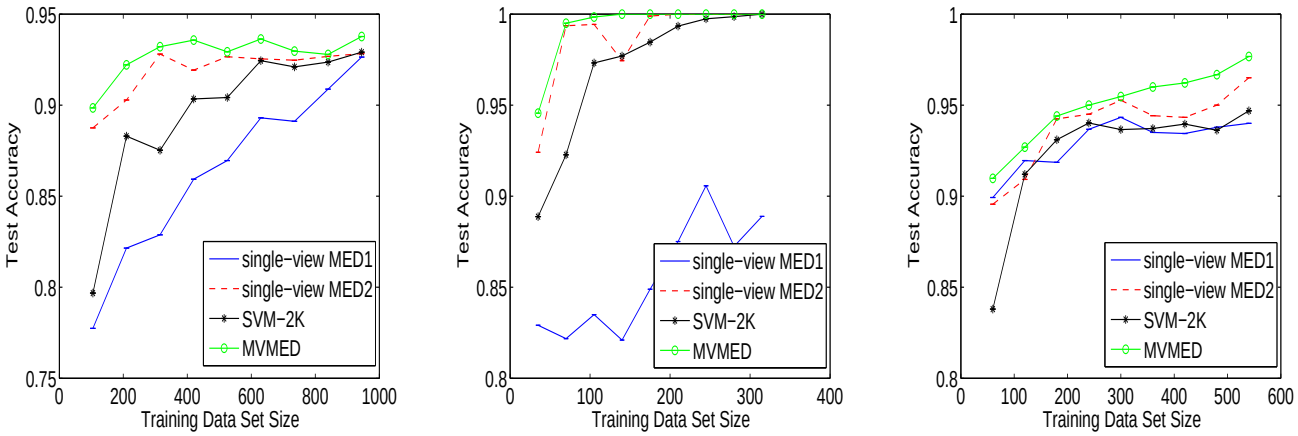
5 Conclusion and Future Work

We have proposed an MVMED framework which is an extension of MED to the scenario of learning with multiple views

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

Data	single-view MED1	single-view MED2	SVM-2K	MVMED
Web-page	77.74±1.43	88.75±1.90	79.68±5.83	**89.85**±2.04
Ionosphere	82.91±4.37	92.41±3.87	88.88±14.54	**94.56**±3.22
Advertisement	89.93±1.53	89.56±1.16	83.80±9.27	**90.98**±1.60

Table 1: The average classification accuracies and standard deviations (%) of four methods.



(a) Web-page Classification (b) Ionosphere Classification (c) Advertisement Classification

Figure 1: Comparison of four methods on different data sets with increasing training sizes.

and therefore integrates the merits of MVL and MED. Different from existing approaches to exploiting multiple views, we propose to use "margin consistency" to perform MVL, and apply it to MVMED. We also give an instantiation of the MVMED framework with a factorized prior, which is further shown to be related to the multi-view method SVM-2K. Experimental results on real-world applications web-page classification, ionosphere classification and advertisement classification validate the effectiveness of the proposed MVMED.

Interesting future work directions include extending our MVMED framework to more than two views, and applying it to other learning scenarios such as semi-supervised learning and structure learning.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Project 61075005, and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

[Balcan *et al.*, 2005] M.F. Balcan, A. Blum, and Y. Ke. Co-training and expansion: Towards bridging theory and practice. *Advances in Neural Information Processing Systems*, 18:89–96, 2005.

[Bickel and Scheffer, 2004] S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the 4th International Conference on Data Mining*, pages 19–26, 2004.

[Blum and Mitchell, 1998] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[Dasgupta *et al.*, 2001] S. Dasgupta, M.L. Littman, and D. McAllester. PAC generalization bounds for co-training. *Advances in Neural Information Processing Systems*, 14:375–382, 2001.

[Domingos, 2000] P. Domingos. Bayesian averaging of classifiers and the overfitting problem. In *Proceedings of the 7th International Conference on Machine Learning*, pages 223–230, 2000.

[Farquhar *et al.*, 2005] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. *Advances in Neural Information Processing Systems*, 18:355–362, 2005.

[Jaakkola *et al.*, 1999] T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. *Advances in Neural Information Processing Systems*, 12:470–476, 1999.

[Jebara and Jaakkola, 2000] T. Jebara and T. Jaakkola. Feature selection and dualities in maximum entropy discrimination. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 291–300, 2000.

[Jebara, 2004] T. Jebara. Multi-task feature and kernel selection for SVMs. In *Proceedings of the 21st International Conference on Machine Learning*, pages 55–62, 2004.

[Jebara, 2011] T. Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research*, 12:75–110, 2011.

[Kumar and Daumé III, 2011] A. Kumar and H. Daumé III. A co-training approach for multi-view spectral clustering.

In *Proceedings of the 27th International Conference on Machine Learning*, pages 393–400, 2011.

[Kumar *et al.*, 2011] A. Kumar, P. Rai, and H. Daumé III. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, 24:1413–1421, 2011.

[Kushmerick, 1999] N. Kushmerick. Learning to remove Internet advertisements. In *Proceedings of the 3rd annual Conference on Autonomous Agents*, pages 175–181, 1999.

[Long and Wu, 2004] P.M. Long and X. Wu. Mistake bounds for maximum entropy discrimination. *Advances in Neural Information Processing Systems*, 17:833–840, 2004.

[Nigam and Ghani, 2000] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, pages 86–93, 2000.

[Rosenberg and Bartlett, 2007] D. Rosenberg and P.L. Bartlett. The Rademacher complexity of co-regularized kernel classes. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 396–403, 2007.

[Sindhwani and Rosenberg, 2008] V. Sindhwani and D.S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 976–983, 2008.

[Sindhwani *et al.*, 2005] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*, pages 74–79, 2005.

[Sun and Jin, 2011] S. Sun and F. Jin. Robust co-training. *International Journal of Pattern Recognition and Artificial Intelligence*, 25:1113–1126, 2011.

[Sun and Shawe-Taylor, 2010] S. Sun and J. Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *Journal of Machine Learning Research*, 11:2423–2455, 2010.

[Sun, 2011] S. Sun. Multi-view Laplacian support vector machines. *Lecture Notes in Artificial Intelligence*, 7121:2423–2455, 2011.

[Wang and Zhou, 2010] W. Wang and Z.H. Zhou. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1135–1142, 2010.

[White *et al.*, 2012] M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. *Advances in Neural Information Processing Systems*, 25:1682–1690, 2012.

[Yu *et al.*, 2011] S. Yu, B. Krishnapuram, R. Rosales, and R.B. Rao. Bayesian co-training. *Journal of Machine Learning Research*, 12:2649–2680, 2011.

[Zhu and Xing, 2009] J. Zhu and E.P. Xing. Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research*, 10:2531–2569, 2009.

[Zhu *et al.*, 2008a] J. Zhu, E.P. Xing, and B. Zhang. Laplace maximum margin Markov networks. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1256–1263, 2008.

[Zhu *et al.*, 2008b] J. Zhu, E.P. Xing, and B. Zhang. Partially observed maximum entropy discrimination Markov networks. *Advances in Neural Information Processing Systems*, 21:1977–1984, 2008.