Alternative Multi-View Maximum Entropy Discrimination

Guoqing Chao, Shiliang Sun

***Abstract*—Maximum entropy discrimination (MED) is a general framework for discriminative estimation based on maximum entropy and maximum margin principles, and can produce hard-margin support vector machines (SVMs) under some assumptions. Recently, the multi-view version of MED multi-view maximum entropy discrimination (MVMED) was proposed. In this paper, we try to explore a more natural MVMED framework by assuming two separate distributions $p_1(\Theta_1)$ over the first view classifier parameter Θ_1 and $p_2(\Theta_2)$ over the second view classifier parameter Θ_2. We name the new MVMED framework as alternative MVMED (AMVMED) which enforces the posteriors of two view margins to be equal. The proposed AMVMED is more flexible than the existing MVMED, because compared with MVMED which optimizes one relative entropy, AMVMED assigns one relative entropy term to each of the two views, thus incorporating a tradeoff between the two views. We give the detailed solving procedure which can be divided into two steps. The first step is solving our optimization problem without considering the equal margin posteriors from two views, and then in the second step we consider the equal posteriors. Experimental results on multiple real-world data sets verify the effectiveness of the AMVMED, and comparisons with MVMED are also reported.**

***Index Terms*—Multi-view learning, maximum entropy discrimination, support vector machine, maximum margin.**

I. Introduction

MAXIMUM entropy discrimination (MED) is an effective approach to discriminative training of model parameters, which embodies the Bayesian integration of prior information with maximum margin constraints on observations, and has achieved a success in a large number of machine learning problems. It is a learning paradigm for combining the discriminative learning and generative learning mechanisms.

MED was first presented by Jaakkola et al. [1] in 1999. Instead of looking for a single classifier parameter $\boldsymbol{\Theta}$ (the classifier can be $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{X} + b$ for which $\boldsymbol{\Theta}$ will be divided into $\boldsymbol{\theta}$ and b), MED considers a more general problem of finding a distribution $p(\boldsymbol{\Theta})$ over the classifier parameter $\boldsymbol{\Theta}$. The distribution $p(\boldsymbol{\Theta})$ can be obtained by seeking a joint distribution $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ over the classifier (non-margin) parameter $\boldsymbol{\Theta}$ and margin parameter $\boldsymbol{\gamma}$ and then marginalizing out $\boldsymbol{\gamma}$. MED regularizes $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ by minimizing its relative entropy (also known as Kullback-Leibler divergence) towards some prior target distribution $p_0(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ under certain large margin constraints. MED was applied successfully to classification problems. It can even be well applied to the case when the labels in the training set are uncertain or incomplete. By introducing a selector variable into the discrimination function, Jebara and Jaakkola [2] employed MED for feature selection. Jebara [3] [4] further extended MED to the problem of multi-task feature and kernel selection. On the theoretical side, Long and Wu [5] established a mistake bound for an ensemble method for MED and provided a refined bound that leads to a nearly optimal algorithm for learning disjunctions based on the maximum entropy principle. In recent years, Zhu and Xing [6] proposed an MED Markov network which combines MED and structure learning. By adopting a Laplace prior, Zhu et al. [7] obtained a Laplace maximum margin Markov network which is a sparse model suitable for learning complex structures. In order to deal with the situation where latent variables exist, Zhu et al. [8] further presented a partially observed MED Markov network.

Multi-view learning (MVL) is the learning task that uses multiple representations of the data. These views or representations may be obtained from multiple feature sets or different sources. MVL is a rapidly growing direction in machine learning with well theoretical underpinnings and great practical success. Its popularity is mainly motivated by the fact that many real-world data have multiple views [9]–[14]. For instance, 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. As another example, in content-based web-image retrieval, an object can be described by visual features from the image and at the same time by the text surrounding the image. A noteworthy fact for MVL is that when there are no natural multiple views, using manually generated multiple views can still improve the performance [15]. Ando and Zhang [16] presented a view generating method which assumes that the views share the same low dimensional manifold. The current MVL methods can be divided into two major categories: co-training style algorithms [9], [17]–[23] and co-regularization style algorithms [24]–[32]. For a comprehensive survey on MVL, refer to [33].

Inspired by the recent success of MVL, Farquhar et al. [34] presented a two-view version of support vector machine (SVM) called SVM-2K and inspected its Rademacher com-

Manuscript received November 06, 2013; revised September 30, 2014 and March 24, 2015; accepted June 1, 2015.

This work is supported by the National Natural Science Foundation of China under Project 61370175, Shanghai Knowledge Service Platform Project (No. ZF1213), and the Science and Technology Commission of Shanghai Municipality under Research Grant No. 14DZ2260800.

Guoqing Chao and Shiliang Sun (corresponding author) are with Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China (e-mail: guoqingchao10@gmail.com; slsun@cs.ecnu.edu.cn).

plexity. By making a weak conditional independence assumption that multi-view observations and response variables are independent given a set of latent variables, Chen et al. [35] proposed a large margin approach for predicting subspace learning for multi-view data, which is based on an undirected latent space Markov network. With regard to MED, recently a multi-view maximum entropy discrimination (MVMED) method was proposed [36]. Different from existing MVL styles, MVMED exploits the multiple views in a new fashion named margin consistency and compares an instantiation with SVM-2K. MVMED employs a single joint distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ over the first view classifier parameter $\boldsymbol{\Theta}_1$ and the second view classifier parameter $\boldsymbol{\Theta}_2$ (the joint distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ will be augmented as $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ with the margin distribution $p(\boldsymbol{\gamma})$). In this paper, we propose a more natural MVMED framework named alternative MVMED (AMVMED), which utilizes two separate distributions $p_1(\boldsymbol{\Theta}_1)$ over $\boldsymbol{\Theta}_1$ and $p_2(\boldsymbol{\Theta}_2)$ over $\boldsymbol{\Theta}_2$ (the distributions $p_1(\boldsymbol{\Theta}_1)$ and $p_2(\boldsymbol{\Theta}_1)$ will be augmented as $p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})$ and $p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ with the margin distributions $p_1(\boldsymbol{\gamma})$ and $p_2(\boldsymbol{\gamma})$). Especially MVMED optimizes one relative entropy, whereas AMVMED assigns one relative entropy term to each of the two views, hence incorporating a tradeoff between the two views. It is thus very interesting to compare the new method and the previous MVMED.

The main contributions of this paper include the following two points:

1) We propose a new MVMED framework AMVMED, which utilizes two separate Kullback-Leibler (KL) divergences $\mathrm{KL}\big(p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})\big)$, $\mathrm{KL}\big(p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})\big)$ in the objective function. By balancing the two KL divergences, AMVMED demonstrates its flexible property, which is intriguing and meaningful.
2) We implement one approximate version of AMVMED with a two-step procedure, and investigate how it performs compared with its single-view versions and MVMED.

The rest of this paper is organized as follows. Section II briefly reviews MED. Section III describes the existing MVMED. Section IV introduces our proposed AMVMED and gives an instantiation. In Section V, we investigate the relationship of AMVMED to MVMED and SVM-2K. Section VI reports experiments on multiple real-world data sets and makes comparisons. Finally, we give conclusions and point out future work directions in Section VII.

II. Maximum entropy discrimination

MED is a general framework for discriminative estimation which integrates the principles of the maximum entropy and large margin. It is similar to Bayesian learning in the sense that the posterior of model parameters requires to be inferred, but it may not need the formulation of generative distributions of data.

Consider a two-class problem where labels $y_t \in \{+1, -1\}$ are assigned to the examples X_t, $t = 1, ..., N$. We need to find one discriminant function $L(X_t|\boldsymbol{\Theta})$ whose sign is an estimate of the label y_t. The discriminant function $L(X_t|\boldsymbol{\Theta})$ is specified by the parameter $\boldsymbol{\Theta}$. Common classifiers such as SVM will seek a specific $\boldsymbol{\Theta}$ value, but MED seeks for a distribution $p(\boldsymbol{\Theta})$ over $\boldsymbol{\Theta}$ such that the expected value of the discriminant function under this distribution agrees with the corresponding label. It is this characteristic that makes it straightforward to augment the solution distributions to be joint densities over parameters of several classifiers. In MED, the distribution $p(\boldsymbol{\Theta})$ is augmented as $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ over the classifier parameter $\boldsymbol{\Theta}$ and margin parameter $\boldsymbol{\gamma}$. MED regularizes the distribution $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ by minimizing its relative entropy (KL divergence) towards some prior target distribution $p_0(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ under the expected large margin constraints. Thus, MED can be formulated as the following constrained optimization problem

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

where $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_N\}$ specifies the desired classification margins which reflect the maximum margin principle as in SVMs. In this framework, one joint distribution $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ over the classifier parameter $\boldsymbol{\Theta}$ and margin parameter $\boldsymbol{\gamma}$ is used. $p_0(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ is the assumed prior distribution that $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$ tends to be close to. By marginalizing out $p(\boldsymbol{\gamma})$, the distribution $p(\boldsymbol{\Theta})$ can be obtained to predict the label of a new example. Instead of using one single discriminant function, MED utilizes a convex combination of discriminant functions, i.e., $\int p(\boldsymbol{\Theta}) L(X_t|\boldsymbol{\Theta}) d\boldsymbol{\Theta}$ to get model averaging for decisions. As Domingos [37] proved, model averaging can improve the classification performance by means of alleviating the overfitting problem.

Since the KL divergence is convex with respect to $p(\boldsymbol{\Theta}, \boldsymbol{\gamma})$, and the large margin constraints are intrinsically linear, problem (1) is convex. The solution to MED can be given by the usual maximum entropy method [1]

$$p(\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\lambda})} p_0(\boldsymbol{\Theta}, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_t [y_t L(X_t|\boldsymbol{\Theta}) - \gamma_t]}, \tag{2}$$

where $Z(\boldsymbol{\lambda})$ is the normalization constant 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 jointly concave objective function $J(\boldsymbol{\lambda}) = -\log Z(\boldsymbol{\lambda})$.

The solution is sparse with only a few non-zero Lagrange multipliers, because many classification constraints become irrelevant once the constraints are enforced for a small subset of examples. Sparsity leads to immediate generalization guarantees expressed in terms of the number of non-zero Lagrange multipliers [1].

III. Multi-view maximum entropy discrimination

Based on MED and MVL, Sun and Chao [36] presented an MVMED approach exploiting multiple views in a style named "margin consistency". They enforced the margins from two views to be identical, which means that the classification

confidences from different views are deemed to match each other exactly.

Given the multi-view data set $\{X_t^1, X_t^2, y_t\}$ with N examples where X_t^1 and X_t^2 indicate the tth input from view 1 and view 2, respectively, and $y_t \in \{\pm 1\}$ denotes the corresponding label. MVMED considers a joint distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ over the first view classifier parameter $\boldsymbol{\Theta}_1$ and the second view classifier parameter $\boldsymbol{\Theta}_2$. It uses the augmented joint distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ with the common margin $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_N\}$. The MVMED framework is formulated as

$$\begin{cases} \min\limits_{p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})} \mathrm{KL}\big(p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})\big) \\ \text{s.t.} \int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad \int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad 1 \leq t \leq N, \end{cases} \tag{3}$$

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 expected large margin constraints are enforced on two views.

The solution to the MVMED problem relies on the following theorem [1].

Theorem 1 *The solution to the MVMED problem has the following general form*

$$p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)} p_0(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \\ e^{\left(\sum_{t=1}^N \lambda_{1t}[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] + \sum_{t=1}^N \lambda_{2t}[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t]\right)}, \tag{4}$$

where $Z(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ is the normalization constant and $\boldsymbol{\lambda}_1 = \{\lambda_{11}, ..., \lambda_{1N}\}$, $\boldsymbol{\lambda}_2 = \{\lambda_{21}, ..., \lambda_{2N}\}$ 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

$$J(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = -logZ(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2). \tag{5}$$

After $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are obtained, the distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ will be specified accordingly. By marginalizing out $\boldsymbol{\gamma}$, we will get the distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ to further predict the label of a new example (X^1, X^2) from view 1 and view 2 with the following two formulae

$$\hat{y}_1 = \mathrm{sign}\Big(\int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) L_1(X^1|\boldsymbol{\Theta}_1) d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big), \tag{6}$$

$$\hat{y}_2 = \mathrm{sign}\Big(\int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) L_2(X^2|\boldsymbol{\Theta}_2) d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big). \tag{7}$$

The prediction formula can also be made by using the two views together

$$\hat{y} = \mathrm{sign}\Big(\frac{1}{2}\int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)\big(L_1(X^1|\boldsymbol{\Theta}_1) + L_2(X^2|\boldsymbol{\Theta}_2)\big) \\ d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2\Big). \tag{8}$$

IV. Alternative multi-view maximum entropy discrimination

The settings are the same with MVMED. However, different from MVMED, AMVMED considers two separate distributions $p_1(\boldsymbol{\Theta}_1)$ over the first view classifier parameter $\boldsymbol{\Theta}_1$ and $p_2(\boldsymbol{\Theta}_2)$ over the second view classifier parameter $\boldsymbol{\Theta}_2$. Similar to MVMED, AMVMED augments them with the margin parameter $\boldsymbol{\gamma}$. Thus we utilize two augmented distributions $p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})$ and $p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ in the new framework. Moreover, we enforce the margins of two views to be the same and at the same time make the posteriors of the two view margins be equal, achieving the purpose that the classification confidences from different views are deemed to match each other. Our AMVMED framework is formulated as follows:

$$\begin{cases} \min\limits_{p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}),\ p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})} \rho \mathrm{KL}\big(p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})\big) \\ \qquad\qquad + (1-\rho)\mathrm{KL}\big(p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})\big) \\ \text{s.t.} \quad \int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \geq 0 \\ \quad \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t] d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad \int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) d\boldsymbol{\Theta}_1 = \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) d\boldsymbol{\Theta}_2 \\ \quad 1 \leq t \leq N. \end{cases} \tag{9}$$

The parameter ρ in the objective function indicates the tradeoff of the two KL terms (also known as two view relative entropies). It can be tuned to emphasize one view against the other. The constraints consist of two view maximum margin constraints and one equal posterior constraint for two views. It is noted that $\rho \in [0, 1]$. If $\rho = 0$ AMVMED is equivalent to single-view MED on view 2, and if $\rho = 1$ it degenerates to single-view MED on view 1.

A. The Solution to AMVMED

With respect to how to solve the above optimization problem, we propose to use a two-step procedure. The first step is to solve the problem without considering the equal posteriors of the two view margins, and the second step is to make the posteriors of the two view margins the same. Otherwise, the optimization problem will be tricky to solve.

In other words, we will consider the case that the optimization can be divided into the above two steps. The other case is an open problem that will be our future research direction. The specific implementations are detailed below. Firstly, we will solve the following problem,

$$\begin{cases} \min\limits_{p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}),\ p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})} \rho \mathrm{KL}\big(p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})\big) \\ \qquad\qquad + (1-\rho)\mathrm{KL}\big(p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \parallel p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})\big) \\ \text{s.t.} \int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \geq 0 \\ \quad \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2) - \gamma_t] d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0 \\ \quad 1 \leq t \leq N. \end{cases} \tag{10}$$

The Lagrangian of the optimization problem is

$$\begin{aligned} L &= \rho \int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) \log \frac{p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})} d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \\ &+ (1-\rho) \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \log \frac{p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})}{p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})} d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \\ &- \sum_{t=1}^{N} \rho \int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) \lambda_{1t} [y_t L_1(X_t^1 | \boldsymbol{\Theta}_1) - \gamma_t] d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \\ &- \sum_{t=1}^{N} (1-\rho) \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) \lambda_{2t} [y_t L_2(X_t^2 | \boldsymbol{\Theta}_2) - \gamma_t] d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}, \end{aligned} \tag{11}$$

where $\boldsymbol{\lambda}_1 = \{\lambda_{11}, \ldots, \lambda_{1N}\}$ and $\boldsymbol{\lambda}_2 = \{\lambda_{21}, \ldots, \lambda_{2N}\}$ are the non-negative Lagrange multipliers for view 1 and view 2, respectively. Making the partial derivatives of L to $p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})$ and $p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ be zero, we get the solutions to (10) as follows:

$$p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) = \frac{1}{Z_1(\boldsymbol{\lambda}_1)} p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{1t} [y_t L_1(X_t^1 | \boldsymbol{\Theta}_1) - \gamma_t]}, \tag{12}$$

$$p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) = \frac{1}{Z_2(\boldsymbol{\lambda}_2)} p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{2t} [y_t L_2(X_t^2 | \boldsymbol{\Theta}_2) - \gamma_t]}, \tag{13}$$

where $Z_1(\boldsymbol{\lambda}_1)$ and $Z_2(\boldsymbol{\lambda}_2)$ are the normalization constants.

Accordingly,

$$Z_1(\boldsymbol{\lambda}_1) = \int p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{1t} [y_t L_1(X_t^1 | \boldsymbol{\Theta}_1) - \gamma_t]} d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma}, \tag{14}$$

$$Z_2(\boldsymbol{\lambda}_2) = \int p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{2t} [y_t L_2(X_t^2 | \boldsymbol{\Theta}_2) - \gamma_t]} d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}. \tag{15}$$

Secondly, we will enforce the posteriors of the two view margins to be equal

$$\int p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) d\boldsymbol{\Theta}_1 = \int p_2(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) d\boldsymbol{\Theta}_2. \tag{16}$$

Herein, we suppose

$$p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) = p_0(\boldsymbol{\Theta}_1) p_0(\boldsymbol{\gamma}) = p_0(\boldsymbol{\theta}_1) p_0(b_1) p_0(\boldsymbol{\gamma}), \tag{17}$$

$$p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) = p_0(\boldsymbol{\Theta}_2) p_0(\boldsymbol{\gamma}) = p_0(\boldsymbol{\theta}_2) p_0(b_2) p_0(\boldsymbol{\gamma}), \tag{18}$$

where $p_0(b_1)$, $p_0(b_2)$ approach the 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 $\boldsymbol{I}$, and the prior over the margin constraints $\boldsymbol{\gamma}$ is assumed to be fully factored

$$p_0(\boldsymbol{\gamma}) = \prod_{t=1}^{N} p_0(\gamma_t), \tag{19}$$

with $p_0(\gamma_t) = ce^{-c(1-\gamma_t)}$, and $\gamma_t \leq 1$. In addition, we will use the usual linear classifier assumptions, that is,

$$L_1(X_t^1 | \boldsymbol{\Theta}_1) = {\boldsymbol{\theta}_1}^{\mathrm{T}} X_t^1 + b_1 \tag{20}$$

and

$$L_2(X_t^2 | \boldsymbol{\Theta}_2) = {\boldsymbol{\theta}_2}^{\mathrm{T}} X_t^2 + b_2. \tag{21}$$

Theorem 2 *By making the above assumptions, we can obtain that* $\lambda_{1t} = \lambda_{2t},\ \forall t \in \{1, ..., N\}$.

Proof: By substituting (12) and (13) into (16), we get

$$\begin{aligned} &\int \frac{1}{Z_1(\boldsymbol{\lambda}_1)} p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{1t} [y_t L_1(X_t^1 | \boldsymbol{\Theta}_1) - \gamma_t]} d\boldsymbol{\Theta}_1 = \\ &\int \frac{1}{Z_2(\boldsymbol{\lambda}_2)} p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{2t} [y_t L_2(X_t^2 | \boldsymbol{\Theta}_2) - \gamma_t]} d\boldsymbol{\Theta}_2. \end{aligned} \tag{22}$$

Then substituting (17) and (20) into the left hand side of equation (22), and substituting (18) and (21) into the right hand side, results in

$$\begin{aligned} &\int \frac{1}{Z_1(\boldsymbol{\lambda}_1)} p_0(\boldsymbol{\theta}_1) p_0(b_1) p_0(\boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{1t} [y_t ({\boldsymbol{\theta}_1}^{\mathrm{T}} X_t^1 + b_1) - \gamma_t]} d\boldsymbol{\Theta}_1 = \\ &\int \frac{1}{Z_2(\boldsymbol{\lambda}_2)} p_0(\boldsymbol{\theta}_2) p_0(b_2) p_0(\boldsymbol{\gamma}) e^{\sum_{t=1}^{N} \lambda_{2t} [y_t ({\boldsymbol{\theta}_2}^{\mathrm{T}} X_t^2 + b_2) - \gamma_t]} d\boldsymbol{\Theta}_2. \end{aligned} \tag{23}$$

Integrating the probability distributions $p_0(\boldsymbol{\theta}_1)$, $p_0(\boldsymbol{\theta}_2)$, $p_0(b_1)$ and $p_0(b_2)$, and then putting the variables irrelevant to $\boldsymbol{\gamma}$ into $C(\boldsymbol{\lambda}_1)$ and $D(\boldsymbol{\lambda}_2)$ ($C(\boldsymbol{\lambda}_1)$ and $D(\boldsymbol{\lambda}_2)$ can be considered as the normalization constants of $p_0(\boldsymbol{\gamma}) e^{-\sum_{t=1}^{N} \lambda_{1t} \gamma_t}$ and $p_0(\boldsymbol{\gamma}) e^{-\sum_{t=1}^{N} \lambda_{2t} \gamma_t}$), we will obtain

$$\frac{p_0(\boldsymbol{\gamma}) e^{-\sum_{t=1}^{N} \lambda_{1t} \gamma_t}}{C(\boldsymbol{\lambda}_1)} = \frac{p_0(\boldsymbol{\gamma}) e^{-\sum_{t=1}^{N} \lambda_{2t} \gamma_t}}{D(\boldsymbol{\lambda}_2)}. \tag{24}$$

Cancelling out the common item $p_0(\boldsymbol{\gamma})$ on both sides of equation (24), and making some simple transformations, we get the following equation

$$e^{-\sum_{t=1}^{N} (\lambda_{1t} - \lambda_{2t}) \gamma_t} = \frac{C(\boldsymbol{\lambda}_1)}{D(\boldsymbol{\lambda}_2)}. \tag{25}$$

Since $C(\boldsymbol{\lambda}_1)$ and $D(\boldsymbol{\lambda}_2)$ are irrelevant to $\boldsymbol{\gamma}$, we reach the conclusion that $\lambda_{1t} = \lambda_{2t}$, $\forall t = 1, ..., N$. ■

Let $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \boldsymbol{\lambda}$. Substituting (12) and (13) into (11) results in the Lagrange dual objective function that needs to be maximized

$$\begin{aligned} J(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) &= -\rho \log Z_1(\boldsymbol{\lambda}_1) - (1-\rho) \log Z_2(\boldsymbol{\lambda}_2) \\ &= -\rho \log Z_1(\boldsymbol{\lambda}) - (1-\rho) \log Z_2(\boldsymbol{\lambda}). \end{aligned} \tag{26}$$

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

$$\hat{y}_1 = \mathrm{sign}\Big(\int p_1(\boldsymbol{\Theta}_1) L_1(X^1 | \boldsymbol{\Theta}_1) d\boldsymbol{\Theta}_1 \Big), \tag{27}$$

$$\hat{y}_2 = \mathrm{sign}\Big(\int p_2(\boldsymbol{\Theta}_2) L_2(X^2 | \boldsymbol{\Theta}_2) d\boldsymbol{\Theta}_2 \Big), \tag{28}$$

$$\begin{aligned} \hat{y} = \mathrm{sign}\Big(&\rho \int p_1(\boldsymbol{\Theta}_1) L_1(X^1 | \boldsymbol{\Theta}_1) d\boldsymbol{\Theta}_1 \\ &+ (1-\rho) \int p_2(\boldsymbol{\Theta}_2) L_2(X^2 | \boldsymbol{\Theta}_2) d\boldsymbol{\Theta}_2 \Big). \end{aligned} \tag{29}$$

B. Instantiation of AMVMED and MVMED

The priors $p_0(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})$ and $p_0(\boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ play an important role in our AMVMED framework as shown in (12) and (13). Now we instantiate our AMVMED with two concrete prior formulations of (17) and (18), and the prior over the margin constraints $\boldsymbol{\gamma}$ is assumed to be fully factored as (19). 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 to the use of slack variables and additive penalties in SVMs. It allows some slackness to handle the non-separable case, which is analogous to soft-margin SVMs. Restricted to the margin prior, the slackness is less flexible than soft-margin SVMs, i.e., irrespective of the c value chosen, 40% of the probability mass lies to the left of the mean value and the remaining 60% lies to the right. The ratio 40/60 will regularize the slackness that is allowed. But the margin prior indeed provides an approach to deal with the non-separable case. After making the prior assumptions, (14) becomes

$$\begin{aligned} Z_1(\boldsymbol{\lambda}_1) = & \int \mathcal{N}(\boldsymbol{\theta}_1|\mathbf{0},\mathbf{I})\mathcal{N}(b_1|\mathbf{0},{\boldsymbol{\sigma}_1}^2)\prod_{t=1}^{N} ce^{-c(1-\gamma_t)} \\ & e^{\left(\sum_{t=1}^{N}\lambda_{1t}[y_tL_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]\right)}d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \\ & = e^{\left(\frac{1}{2}\sum_{t,\tau=1}^{N}\lambda_{1t}\lambda_{1\tau}y_ty_\tau {X_t^1}^{\mathrm{T}}X_\tau^1+\frac{{\boldsymbol{\sigma}_1}^2}{2}\left(\sum_{t=1}^{N}\lambda_{1t}y_t\right)^2\right)} \\ & \prod_{t=1}^{N}\Big(\frac{c}{c-\lambda_{1t}}e^{-\lambda_{1t}}\Big), \end{aligned} \tag{30}$$

and (15) becomes

$$\begin{aligned} Z_2(\boldsymbol{\lambda}_2) = & \int \mathcal{N}(\boldsymbol{\theta}_2|\mathbf{0},\mathbf{I})\mathcal{N}(b_2|\mathbf{0},{\boldsymbol{\sigma}_2}^2)\prod_{t=1}^{N} ce^{-c(1-\gamma_t)} \\ & e^{\left(\sum_{t=1}^{N}\lambda_{2t}[y_tL_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]\right)}d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \\ & = e^{\left(\frac{1}{2}\sum_{t,\tau=1}^{N}\lambda_{2t}\lambda_{2\tau}y_ty_\tau {X_t^2}^{\mathrm{T}}X_\tau^2+\frac{{\boldsymbol{\sigma}_2}^2}{2}\left(\sum_{t=1}^{N}\lambda_{2t}y_t\right)^2\right)} \\ & \prod_{t=1}^{N}\Big(\frac{c}{c-\lambda_{2t}}e^{-\lambda_{2t}}\Big), \end{aligned} \tag{31}$$

where we have used (20) and (21). We substitute (30), (31) into (26), and thus get

$$\begin{aligned} J(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) = & \rho\Big(\sum_{t=1}^{N}\big[\lambda_{1t}+\log(1-\frac{\lambda_{1t}}{c})\big]\Big) \\ & -\frac{1}{2}\rho\Big(\sum_{t,\tau=1}^{N}\lambda_{1t}\lambda_{1\tau}y_ty_\tau {X_t^1}^{\mathrm{T}}X_\tau^1\Big) \\ & -\frac{\boldsymbol{\sigma}_1^2}{2}\rho\big(\sum_{t=1}^{N}\lambda_{1t}y_t\big)^2 \\ & +(1-\rho)\Big(\sum_{t=1}^{N}\big[\lambda_{2t}+\log(1-\frac{\lambda_{2t}}{c})\big]\Big) \\ & -\frac{1}{2}(1-\rho)\Big(\sum_{t,\tau=1}^{N}\lambda_{2t}\lambda_{2\tau}y_ty_\tau {X_t^2}^{\mathrm{T}}X_\tau^2\Big) \\ & -\frac{\boldsymbol{\sigma}_2^2}{2}(1-\rho)\big(\sum_{t=1}^{N}\lambda_{2t}y_t\big)^2. \end{aligned} \tag{32}$$

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 and b_2, the above dual objective function requires the constraints $\sum_{t=1}^{N}\lambda_{1t}y_t = 0$ and $\sum_{t=1}^{N}\lambda_{2t}y_t = 0$. Note that $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \boldsymbol{\lambda}$. Thus we have the following dual optimization problem

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}} \sum\limits_{t=1}^{N}\Big(\lambda_t+\log\big(1-\frac{\lambda_t}{c}\big)\Big) \\ \qquad -\frac{1}{2}\rho\sum\limits_{t,\tau=1}^{N}\lambda_t\lambda_\tau y_ty_\tau {X_t^1}^{\mathrm{T}}X_\tau^1 \\ \qquad -\frac{1}{2}(1-\rho)\sum\limits_{t,\tau=1}^{N}\lambda_t\lambda_\tau y_ty_\tau {X_t^2}^{\mathrm{T}}X_\tau^2 \\ \text{s.t. } \boldsymbol{\lambda} \geq \mathbf{0} \\ \qquad \sum\limits_{t=1}^{N}\lambda_t y_t = 0. \end{cases} \tag{33}$$

The Lagrange multipliers $\boldsymbol{\lambda}$ is recovered by solving the convex optimization problem (33), whose non-zero values indicate support vectors. ${X_t^v}^{\mathrm{T}}X_\tau^v$ $(v=1,2)$ can be replaced with $\kappa(X_t^v, X_\tau^v)$ $(v=1,2)$ so that a nonlinear classifier can be obtained by using some kernel function such as Gaussian and polynomial.

The time complexity of AMVMED is $O(N^3)$, which is about the same with SVMs, MED and MVMED. For large-scale data, some specific speedup strategies are needed to make AMVMED scalable.

Substituting equations (17) (19) (20) into equation (12), we will get the solution distribution $p_1(\boldsymbol{\Theta}_1, \boldsymbol{\gamma})$, from which we have the following expected quantities

$$\hat{\boldsymbol{\theta}}_1 = \int p_1(\boldsymbol{\theta}_1)\boldsymbol{\theta}_1 d\boldsymbol{\theta}_1 = \sum_{t=1}^{N}\lambda_{1t}y_tX_t^1 = \sum_{t=1}^{N}\lambda_t y_t X_t^1, \tag{34}$$

$$\hat{\boldsymbol{\gamma}}_t = \int p_1(\boldsymbol{\gamma})\gamma_t d\boldsymbol{\gamma} = 1-\frac{1}{c-\lambda_t}. \tag{35}$$

The prediction formula on a new example (X^1, X^2) from equation (27) can be given as

$$\hat{y}_1 = \mathrm{sign}\Big(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}X^1 + \hat{b}_1\Big). \tag{36}$$

Putting equation (34) into (36), the prediction rule using view 1 is given as

$$\hat{y}_1 = \mathrm{sign}\Big(\sum_{t=1}^{N}\lambda_t y_t {X_t^1}^{\mathrm{T}}X^1 + \hat{b}_1\Big). \tag{37}$$

Similarly, the prediction rule using view 2 is given as

$$\hat{y}_2 = \mathrm{sign}\Big(\sum_{t=1}^{N}\lambda_t y_t {X_t^2}^{\mathrm{T}}X^2 + \hat{b}_2\Big). \tag{38}$$

If classifiers from the two views are combined together to make predictions, the prediction rule can be given as

$$\begin{aligned} \hat{y} = \mathrm{sign}\Big(\rho\big(\sum_{t=1}^{N}\lambda_t y_t {X_t^1}^{\mathrm{T}}X^1 + \hat{b}_1\big) \\ +(1-\rho)\big(\sum_{t=1}^{N}\lambda_t y_t {X_t^2}^{\mathrm{T}}X^2 + \hat{b}_2\big)\Big). \end{aligned} \tag{39}$$

Next, we will give the procedure on how to solve $\hat{b}_1$. $\hat{b}_2$ can be obtained similarly.

For the approximate AMVMED primal problem (10), the Karush-Kuhn-Tucker (KKT) conditions can be stated as:

$$\frac{\partial L}{p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})}=0, \tag{40}$$

$$\frac{\partial L}{p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})}=0, \tag{41}$$

$$\int p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma}\geq 0,\quad t=1,...,N, \tag{42}$$

$$\int p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}\geq 0,\quad t=1,...,N, \tag{43}$$

$$\lambda_{1t}\geq 0,\quad t=1,...,N, \tag{44}$$

$$\lambda_{2t}\geq 0,\quad t=1,...,N, \tag{45}$$

$$\int p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})\lambda_{1t}[y_t L_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma}\geq 0,\ \ t=1,...,N, \tag{46}$$

$$\int p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\lambda_{2t}[y_t L_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}\geq 0,\ \ t=1,...,N. \tag{47}$$

For our approximate AMVMED problem, since its objective function is convex and the constraints give a convex feasible region, the KKT conditions will be necessary and sufficient for $p_1(\boldsymbol{\Theta}_1)$, $p_2(\boldsymbol{\Theta}_2)$ to be a solution [38]. Now, we focus on how to solve $\hat{b}_1$ and $\hat{b}_2$. We can get $\hat{b}_1$ with equation (46), the KKT "complementarity" condition, by choosing any t for which $\lambda_{1t}\neq 0$. From (46), we can get

$$y_s(\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}X_s^1+\hat{b}_1)-\hat{\gamma}_t=0, \tag{48}$$

from which we obtain

$$\hat{b}_1=\frac{\hat{\gamma}_t}{y_s}-\hat{\boldsymbol{\theta}}_1^{\mathrm{T}}X_s^1. \tag{49}$$

Note that y_s and X_s^1 denote the label and view 1 feature of the data corresponding to a non-zero λ_t in (48). Putting equations (34) and (35) into (49), we will get

$$\hat{b}_1=\frac{1-\frac{1}{c-\lambda_t}}{y_s}-\sum_{t=1}^{N}\lambda_t y_t {X_t^1}^{\mathrm{T}}X_s^1. \tag{50}$$

Similarly, we will obtain $\hat{b}_2$ as

$$\hat{b}_2=\frac{1-\frac{1}{c-\lambda_t}}{y_s}-\sum_{t=1}^{N}\lambda_t y_t {X_t^2}^{\mathrm{T}}X_s^2. \tag{51}$$

Here, we used the result $\lambda_t=\lambda_{1t}=\lambda_{2t}$.

In order to better understand the procedure of the AMVMED instantiation, we give the algorithm of AMVMED in Algorithm 1.

Algorithm 1 AMVMED

Input:

Data sets $\{X_t^1, X_t^2, y_t\}$, sample size N, parameter c, tradeoff parameter ρ.

Initialize $\boldsymbol{\lambda}$.

Solve the following optimization problem

$$\begin{cases}\max\limits_{\boldsymbol{\lambda}} \sum\limits_{t=1}^{N}\Big(\lambda_t+\log(1-\frac{\lambda_t}{c})\Big) \\ \qquad -\frac{1}{2}\rho\sum\limits_{t,\tau=1}^{N}\lambda_t\lambda_\tau y_t y_\tau {X_t^1}^{\mathrm{T}}X_\tau^1 \\ \qquad -\frac{1}{2}(1-\rho)\sum\limits_{t,\tau=1}^{N}\lambda_t\lambda_\tau y_t y_\tau {X_t^2}^{\mathrm{T}}X_\tau^2 \\ \text{s.t. } \boldsymbol{\lambda}\geq\mathbf{0} \\ \qquad \sum\limits_{t=1}^{N}\lambda_t y_t=0.\end{cases}$$

Once the Lagrange multiplier $\boldsymbol{\lambda}$ is obtained, use any of the formulae (37), (38) and (39) to make prediction for a new example.

Correspondingly, the instantiation of MVMED is given below. Suppose

$$\begin{aligned}p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})&=p_0(\boldsymbol{\Theta}_1)p_0(\boldsymbol{\Theta}_2)p_0(\boldsymbol{\gamma})\\&=p_0(\boldsymbol{\theta}_1)p_0(b_1)p_0(\boldsymbol{\theta}_2)p_0(b_2)p_0(\boldsymbol{\gamma}),\end{aligned} \tag{52}$$

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 $\boldsymbol{I}$, and the prior over the margin constraints $\boldsymbol{\gamma}$ is assumed to be fully factored as (19). Then the normalization constant $Z(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$ in (4) can be obtained as

$$\begin{aligned}&Z(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)=\int p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\\&e^{\Big(\sum_{t=1}^{N}\lambda_{1t}\left[y_t L_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t\right]+\sum_{t=1}^{N}\lambda_{2t}\left[y_t L_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t\right]\Big)}\\&d\boldsymbol{\Theta}_1 d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma}\\&=e^{\left(\frac{1}{2}\sum_{t,\tau=1}^{N}\lambda_{1,t}\lambda_{1,\tau}y_t y_\tau {X_t^1}^{\mathrm{T}}X_\tau^1+\frac{1}{2}\sum_{t,\tau=1}^{N}\lambda_{2,t}\lambda_{2,\tau}y_t y_\tau {X_t^2}^{\mathrm{T}}X_\tau^2\right)}\\&e^{\left(\frac{{\sigma_1}^2}{2}\left(\sum_{t=1}^{N}\lambda_{1t}y_t\right)^2+\frac{{\sigma_2}^2}{2}\left(\sum_{t=1}^{N}\lambda_{2t}y_t\right)^2\right)}\\&\prod_{t=1}^{N}\Big(\frac{c}{c-\lambda_{1t}-\lambda_{2t}}e^{-\lambda_{1t}-\lambda_{2t}}\Big).\end{aligned} \tag{53}$$

By maximizing (5), we have the following dual optimization problem

$$\begin{cases}\max\limits_{\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2} \sum\limits_{t=1}^{N}\Big(\lambda_{1t}+\lambda_{2t}+\log\big(1-\frac{\lambda_{1t}+\lambda_{2t}}{c}\big)\Big) \\ \qquad -\frac{1}{2}\sum\limits_{t,\tau=1}^{N}\lambda_{1t}\lambda_{1\tau}y_t y_\tau {X_t^1}^{\mathrm{T}}X_\tau^1 \\ \qquad -\frac{1}{2}\sum\limits_{t,\tau=1}^{N}\lambda_{2t}\lambda_{2\tau}y_t y_\tau {X_t^2}^{\mathrm{T}}X_\tau^2 \\ \text{s.t. } \boldsymbol{\lambda}_1\geq\mathbf{0},\boldsymbol{\lambda}_2\geq\mathbf{0} \\ \qquad \sum\limits_{t=1}^{N}\lambda_{1t}y_t=0,\sum\limits_{t=1}^{N}\lambda_{2t}y_t=0.\end{cases} \tag{54}$$

After obtaining the Lagrange multipliers $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, the prediction rules for view 1 and view 2 on a new example (X^1, X^2) are respectively

$$\hat{y}_1 = \text{sign}\Big(\sum_{t=1}^{N} \lambda_{1t} y_t {X_t^1}^{\mathrm{T}} X^1 + \hat{b}_1\Big), \tag{55}$$

$$\hat{y}_2 = \text{sign}\Big(\sum_{t=1}^{N} \lambda_{2t} y_t {X_t^2}^{\mathrm{T}} X^2 + \hat{b}_2\Big), \tag{56}$$

where $\hat{b}_1$ and $\hat{b}_2$ are given by the KKT conditions using support vectors. If classifiers from two views are combined together to make predictions, the prediction rule can be given analogously as (8).

V. Relationship to MVMED and SVM-2K

AMVMED has a great relevance to other classification methods such as MVMED and SVM-2K. We will investigate the relationship of AMVMED to MVMED and SVM-2K in this section. To facilitate the comparison and analysis, we put the dual formula (33) of AMVMED and the dual formula (54) of MVMED together as follows:

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}} \sum\limits_{t=1}^{N} \Big(\lambda_t + \log\big(1 - \dfrac{\lambda_t}{c}\big)\Big) \\ \qquad - \dfrac{1}{2}\rho \sum\limits_{t,\tau=1}^{N} \lambda_t \lambda_\tau y_t y_\tau {X_t^1}^{\mathrm{T}} X_\tau^1 \\ \qquad - \dfrac{1}{2}(1-\rho) \sum\limits_{t,\tau=1}^{N} \lambda_t \lambda_\tau y_t y_\tau {X_t^2}^{\mathrm{T}} X_\tau^2 \\ \text{s.t. } \boldsymbol{\lambda} \geq \mathbf{0} \\ \qquad \sum\limits_{t=1}^{N} \lambda_t y_t = 0. \end{cases} \tag{57}$$

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2} \sum\limits_{t=1}^{N} \Big(\lambda_{1t} + \lambda_{2t} + \log\big(1 - \dfrac{\lambda_{1t} + \lambda_{2t}}{c}\big)\Big) \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} \lambda_{1t} \lambda_{1\tau} y_t y_\tau {X_t^1}^{\mathrm{T}} X_\tau^1 \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} \lambda_{2t} \lambda_{2\tau} y_t y_\tau {X_t^2}^{\mathrm{T}} X_\tau^2 \\ \text{s.t. } \boldsymbol{\lambda}_1 \geq \mathbf{0}, \boldsymbol{\lambda}_2 \geq \mathbf{0} \\ \qquad \sum\limits_{t=1}^{N} \lambda_{1t} y_t = 0, \sum\limits_{t=1}^{N} \lambda_{2t} y_t = 0. \end{cases} \tag{58}$$

By comparison, we can see that when $\rho = 0.5$ in (57) and $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$ in (58), the two optimization problems are somewhat alike. In addition, the ρ in (57) makes AMVMED more flexible, but compared to (58), there is only one $\boldsymbol{\lambda}$ in (57).

In order to analyze the relationship of MVMED and SVM-2K, we rewrite (58) as (59) by replacing ${X_t^1}^{\mathrm{T}} X_\tau^1$, ${X_t^2}^{\mathrm{T}} X_\tau^2$ with Mercer kernel functions $\kappa(X_t^1, X_\tau^1)$, $\kappa(X_t^2, X_\tau^2)$ and setting $g_{1,t} = \lambda_{1,t} y_t$, $g_{2,t} = \lambda_{2,t} y_t$.

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2} \sum\limits_{t=1}^{N} \Big(\lambda_{1,t} + \lambda_{2,t} + \log\big(1 - \dfrac{\lambda_{1,t} + \lambda_{2,t}}{c}\big)\Big) \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} g_{1,t} g_{1,\tau} \kappa(X_t^1, X_\tau^1) \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} g_{2,t} g_{2,\tau} \kappa(X_t^2, X_\tau^2) \\ \text{s.t. } g_{1,t} = \lambda_{1,t} y_t, \ g_{2,t} = \lambda_{2,t} y_t, \quad 1 \leq t \leq N \\ \qquad \sum\limits_{t=1}^{N} g_{1,t} = 0 = \sum\limits_{t=1}^{N} g_{2,t} \\ \qquad \boldsymbol{\lambda}_1 \geq \mathbf{0}, \boldsymbol{\lambda}_2 \geq \mathbf{0}. \end{cases} \tag{59}$$

Following the paper [34], we can write out the original SVM-2K optimization problem as

$$\begin{cases} \min\limits_{\boldsymbol{w}_A, b_A, \boldsymbol{w}_B, b_B} \dfrac{1}{2} \|\boldsymbol{w}_A\|^2 + \dfrac{1}{2} \|\boldsymbol{w}_B\|^2 + C^A \sum\limits_{t=1}^{N} \xi_i^A \\ \qquad + C^B \sum\limits_{t=1}^{N} \xi_i^B + D \sum\limits_{t=1}^{N} \eta_i \\ \text{s.t. } |\langle \boldsymbol{w}_A, \Phi_A(x_i)\rangle + b_A - \langle \boldsymbol{w}_B, \Phi_B(x_i)\rangle - b_B| \leq \eta_i + \epsilon \\ \qquad y_i(\langle \boldsymbol{w}_A, \Phi_A(x_i)\rangle + b_A) \geq 1 - \xi_i^A \\ \qquad y_i(\langle \boldsymbol{w}_B, \Phi_B(x_i)\rangle + b_B) \geq 1 - \xi_i^B \\ \qquad \xi_i^A \geq 0, \ \xi_i^B \geq 0, \ \eta_i \geq 0 \\ \qquad 1 \leq i \leq N, \end{cases} \tag{60}$$

where $\boldsymbol{w}_A, b_A$ are the weight and threshold of the first view SVM, and $\boldsymbol{w}_B, b_B$ are the weight and threshold of the second view SVM. $|\langle \boldsymbol{w}_A, \Phi_A(x_i)\rangle + b_A - \langle \boldsymbol{w}_B, \Phi_B(x_i)\rangle - b_B| \leq \eta_i + \epsilon$ is the ϵ-insensitive 1-norm constraint where slack variables are used to measure the amount of how points fail to meet the ϵ similarity.

Applying the usual Lagrange multiplier techniques, we get the following dual problem:

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-} \sum\limits_{t=1}^{N} (\lambda_{1,t} + \lambda_{2,t}) \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} g_{1,t} g_{1,\tau} \kappa(X_t^1, X_\tau^1) \\ \qquad - \dfrac{1}{2} \sum\limits_{t,\tau=1}^{N} g_{2,t} g_{2,\tau} \kappa(X_t^2, X_\tau^2) \\ \text{s.t. } g_{1,t} = \lambda_{1,t} y_t - \beta_t^+ + \beta_t^- \\ \qquad g_{2,t} = \lambda_{2,t} y_t + \beta_t^+ - \beta_t^- \\ \qquad \sum\limits_{t=1}^{N} g_{1,t} = 0 = \sum\limits_{t=1}^{N} g_{2,t} \\ \qquad 0 \leq \lambda_{1,t} \leq C^A, \ 0 \leq \lambda_{2,t} \leq C^B \\ \qquad 0 \leq \beta_t^{+/-}, \ \beta_t^+ + \beta_t^- \leq D \\ \qquad 1 \leq t \leq N. \end{cases} \tag{61}$$

Comparing (59) with (61), we can find that (59) has an additional term $\log\Big(1-\frac{\lambda_{1,t}+\lambda_{2,t}}{c}\Big)$ in the objective function, while (61) 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 but in different forms. If we set $c \to \infty$ in (59) and set $\beta_t^+ = \beta_t^- = 0$, $C^A \to \infty$, $C^B \to \infty$ in (61), the two formulae will be exactly identical. On the one hand, MVMED is restricted by some prior assumptions; on the other hand, MVMED is flexible by obtaining different instantiations with different prior specifications.

In order to analyze the relationship of AMVMED and SVM-2K, we rewrite (57) as (62) by replacing $X_t^{1\mathrm{T}} X_\tau^1$, $X_t^{2\mathrm{T}} X_\tau^2$ with Mercer kernel functions $\kappa(X_t^1, X_\tau^1)$, $\kappa(X_t^2, X_\tau^2)$ and setting $g_{1,t} = \lambda_{1,t} y_t$, $g_{2,t} = \lambda_{2,t} y_t$. Using the conclusion $\lambda_{1,t} = \lambda_{2,t} = \lambda_t$, we will obtain

$$\begin{cases} \max\limits_{\boldsymbol{\lambda}} \sum\limits_{t=1}^{N} \Big(\lambda_t + \log\big(1-\frac{\lambda_t}{c}\big)\Big) \\ \qquad -\frac{1}{2}\rho \sum\limits_{t,\tau=1}^{N} g_t g_\tau \kappa(X_t^1, X_\tau^1) \\ \qquad -\frac{1}{2}(1-\rho) \sum\limits_{t,\tau=1}^{N} g_t g_\tau \kappa(X_t^2, X_\tau^2) \\ \text{s.t.} \quad g_t = \lambda_t y_t, \quad g_\tau = \lambda_\tau y_\tau, \quad 1 \le t \le N \\ \quad \boldsymbol{\lambda} \ge \mathbf{0} \\ \quad \sum\limits_{t=1}^{N} \lambda_t y_t = 0. \end{cases} \tag{62}$$

Comparing (62) with (61), we can see that they are very different in the formulation. But if we set $\rho = 0.5$, $c \to \infty$ in (62) and set $\lambda_{1,t} = \lambda_{2,t}$, $\beta_t^+ = \beta_t^- = 0$, $C^A \to \infty$, $C^B \to \infty$ in (61), the two formulae will be somewhat alike. Although AMVMED is restricted by some prior assumptions as MVMED, AMVMED is more flexible than SVM-2K. It can tune the parameter ρ to balance the two views and obtain different instantiations with different prior specifications.

VI. Experiments

We performed experiments with our proposed AMVMED on three real-world data sets: web-page classification, ionosphere classification and advertisement classification.

For all the experiments, we explore multiple values of the parameters ρ and c for AMVMED and multiple values of parameter c for single-view MED1, single-view MED2, and MVMED. Given a division of the training and test set, we use one half of the test set as the validation set for parameter selection and the other half for test. The values of ρ and c are determined on the validation set and then tested on the unseen test set. All the experiments are run for ten times. The average accuracies obtained by ten random divisions of the training and test sets are reported. In addition, we initialize $\boldsymbol{\lambda} = 0.5 \times \boldsymbol{I}_N$ throughout the experiments and use the linear kernel for the instantiations of AMVMED and MVMED.

The MVL methods MVMED and SVM-2K are also used for comparison. Two single-view methods corresponding to AMVMED named single-view MED1 and single-view MED2 are employed to compare with our AMVMED. In addition, the ρ for AMVMED is chosen from $\{0, 0.1, 0.2, ..., 0.9, 1.0\}$. For MVMED, AMVMED 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 $\text{sign}\big((f_1+f_2)/2\big)$ for MVMED and SVM-2K, $\text{sign}\big(\rho f_1 + (1-\rho) f_2\big)$ for AMVMED, and the one with the highest validation accuracy will be selected.

We report the average accuracies and standard deviations of the above five methods on three data sets with half of the data as the training set in Table I. Then we decrease and increase the training set sizes gradually, and show their performances in Fig. 1, Fig. 2 and Fig. 3. Moreover, we report the sensitivity of parameter ρ on the three data sets in Fig. 5.

To make the comparison of three multi-view multi-hyperplane learning methods SVM-2K, AMVMED and MVMED more sufficient, we perform an additional experiment on AMVMED vs. MVMED and multi-hyperplane SVM-2K. The experimental results are shown in Table II with classification accuracies and standard deviations Fig. 4 using the receiver operating characteristic (ROC) curve, and Table III with area under roc curve (AUC) values.

A. 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 [9] [24]. The dimensions of the two views are 2333 and 87, respectively. For convenience and effectiveness, we reduce the dimension of view 1 from 2333 to 500 via principal component analysis (PCA). The parameters C^A and C^B in SVM-2K and c in AMVMED and MVMED are independently chosen by validation from $\{2^{-5}, 2^{-4}, ..., 2^5\}$.

Clearly, Table I indicates that AMVMED and MVMED are superior to single-view MED1, single-view MED2 and SVM-2K. Fig. 1 with varying training sizes also shows that our AMVMED consistently outperforms the other three methods except MVMED. We can also find that SVM-2K does not perform well at the first half part and gradually behaves better than single-view MED1 and single-view MED2 at last. Moreover, our AMVMED performs competitively with MVMED especially with more training data.

B. Ionosphere Classification

The ionosphere data set which origins from UCI,[1] was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. Good radar returns are those showing evidence of some type of structure

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

TABLE I
THE AVERAGE CLASSIFICATION ACCURACIES AND STANDARD DEVIATIONS (%) OF FIVE METHODS ON THREE DATA SETS.

Data	single-view MED1	single-view MED2	SVM-2K	MVMED	AMVMED
Web-page	90.72±1.57	92.47±1.59	90.42±2.44	**92.93**±2.07	92.74±1.46
Ionosphere	92.95±1.29	100±0	98.19±1.27	100±0	100±0
Advertisement	93.47±1.03	93.20±1.83	93.57±1.73	**94.60**±0.81	94.47±1.30

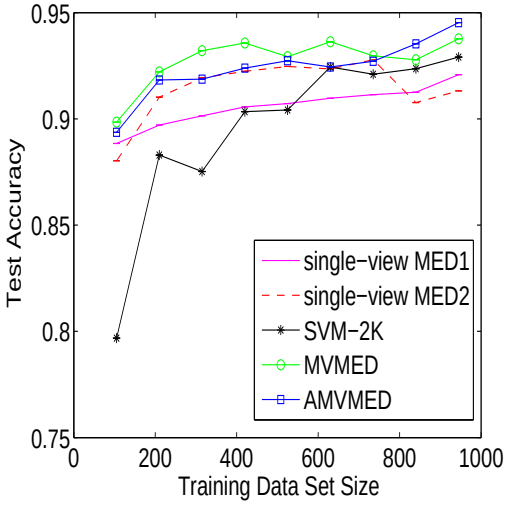


Fig. 1. Comparison of five methods on web-page data with increasing training sizes.

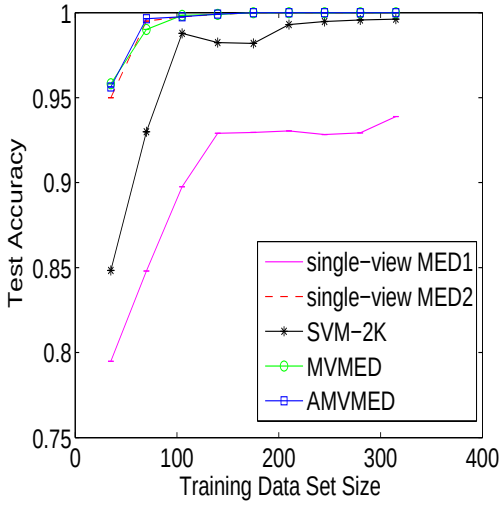


Fig. 2. Comparison of five methods on ionosphere data with increasing training sizes.

in the ionosphere. Bad returns are those that do not and their signals pass through the ionosphere. The data set 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 parameters C^A and C^B in SVM-2K and c in AMVMED and MVMED are independently chosen by validation from $\{2^1, 2^2, ..., 2^{30}\}$.

The superiority of MVMED, AMVMED and single-view MED2 over other methods is demonstrated in Table I and Fig. 2. In addition, Table I and Fig. 2 both show that SVM-2K performs better than single-view MED1 but worse than single-view MED2.

C. Advertisement Classification

The data set consists of 3279 examples including 459 ads images (positive examples) and 2820 non-ads images (negative examples). 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). The dimensions of two views are 587 and 967, respectively. Here, we randomly select 600 examples therein to form the used data set. The parameters C^A and C^B in SVM-2K and c in AMVMED and MVMED are independently chosen by validation from $\{2^1, 2^2, ..., 2^{15}\}$.

From Table I and Fig. 3, we can find that our method AMVMED performs better than all the other methods except MVMED. We can also find that single-view MED1 and single-view MED2 perform worse than SVM-2K, MVMED and AMVMED.

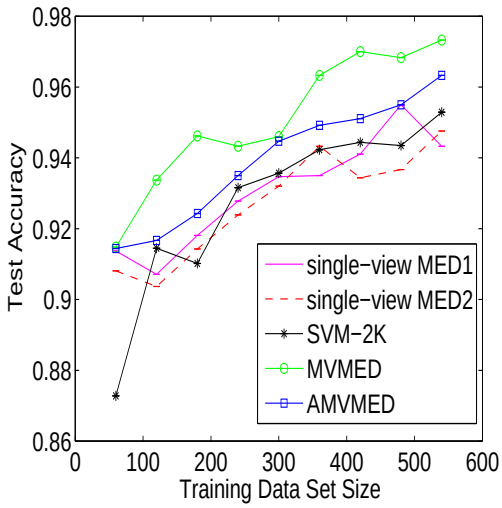


Fig. 3. Comparison of five methods on advertisement data with increasing training sizes.

D. AMVMED vs. MVMED and Multi-Hyperplane SVM-2K

We know that MED-based methods produce many classifiers (hyperplanes) and then make posterior weighted decisions. To make their comparisons with SVM-2K complete, we

TABLE II
THE AVERAGE CLASSIFICATION ACCURACIES AND STANDARD DEVIATIONS (%) OF FOUR MULTI-VIEW MULTI-HYPERPLANE METHODS ON THREE DATA SETS.

Data	MVMED	AMVMED	aveSVM-2K	voteSVM-2K
Web-page	**92.93**±2.07	92.74±1.46	86.03±1.74	88.38±0
Ionosphere	100±0	100±0	100±0	100±0
Advertisement	**94.60**±0.81	94.47±1.30	93.87±1.68	92.00±0

TABLE III
THE AUC VALUES (%) OF THREE MULTI-VIEW MULTI-HYPERPLANE METHODS ON THREE DATA SETS.

Data	MVMED	AMVMED	aveSVM-2K
Web-page	97.23	**97.93**	93.21
Ionosphere	**99.92**	**99.92**	99.51
Advertisement	95.52	**95.80**	92.25

add the experiment comparison with multi-hyperplane SVM-2K which produces many SVM-2K classifiers and then makes a collective decision. With half of each data set as the training set (T data points), when performing validation, we take each subset of $T-1$ points of the training set to train SVM-2K. By this we obtain T SVM-2Ks and at the same time with the T training points we train one AMVMED or MVMED. With T SVM-2Ks, if we make decisions for a test data point by inputting their average discriminant value into $\text{sign}(\cdot)$ function, this SVM-2K is named as aveSVM-2K; if the test data point is classified by voting among the decisions of these SVM-2Ks, we name it as voteSVM-2K.

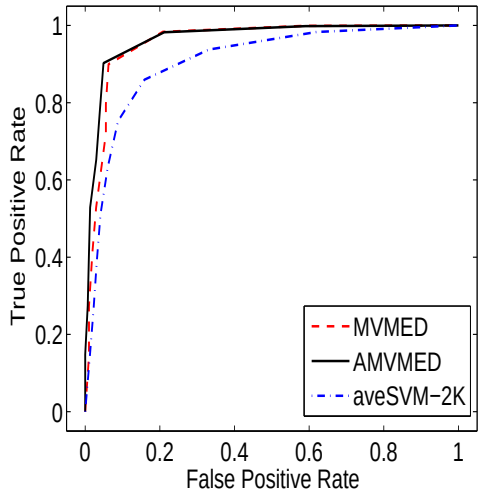
The experimental settings are the same with the above experiments. We first report the accuracies and standard deviations of AMVMED, MVMED and multi-hyperplane SVM-2K (both aveSVM-2K and voteSVM-2K) on all the three data sets in Table II. For AMVMED, MVMED and aveSVM-2K, we also show their performance with the ROC curve in Fig. 4 and with the AUC value in Table III. As we have mentioned, we produce many SVM-2Ks by the independent validation procedure, and then we average their predictions. With the average prediction, we create its ROC curve.

From Table II and Fig. 4, we can see that both AMVMED and MVMED perform better than multi-hyperplane SVM-2K on each data set. Table II demonstrates that AMVMED performs comparable with MVMED, while Fig. 4 and Table III show that for the AUC measure, AMVMED performs a little better than or as well as MVMED and they both perform better than multi-hyperplane SVM-2K.

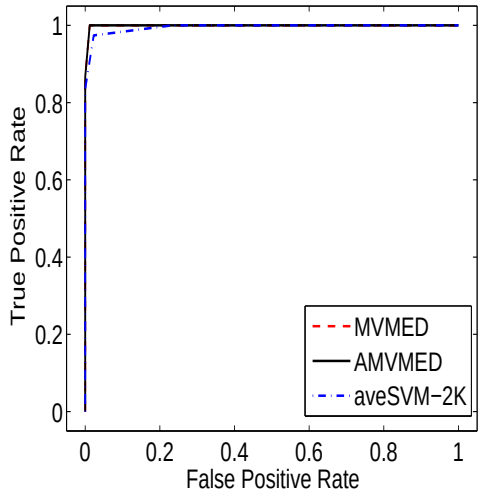
This experiment shows that for many-hyperplanes-to-many-hyperplanes comparison, MED-based methods illustrate a better performance than SVM-based methods. The effectiveness of the AMVMED is also further verified.

E. Sensitivity Analysis of Parameter ρ

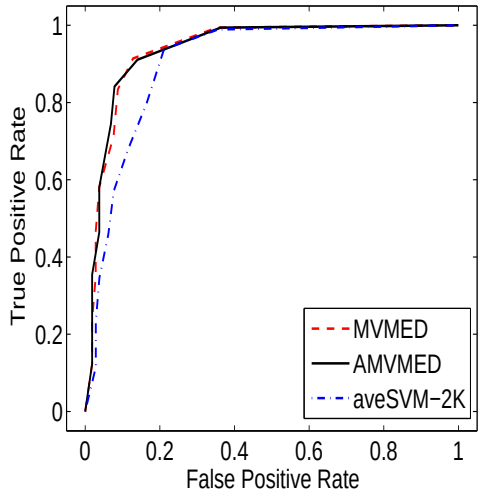
We test the sensitivity of AMVMED to different choices of the parameter ρ. Throughout the experiment, the ρ varies in $\{0, 0.1, ..., 1\}$ and we give the experimental results on three data sets with half of the data as the training set. It is noted that AMVMED with $\rho = 0$ corresponds to single-view MED2 while AMVMED with $\rho = 1$ corresponds to single-view



(a) Web-page classification

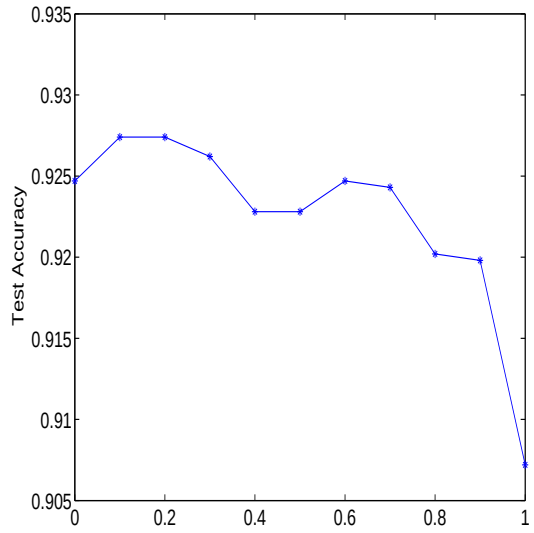


(b) Ionosphere classification

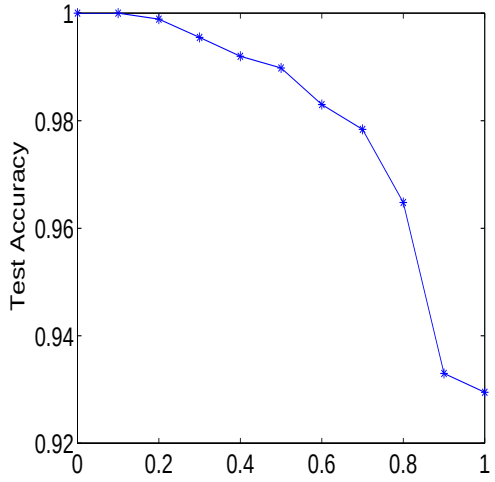


(c) Advertisement classification

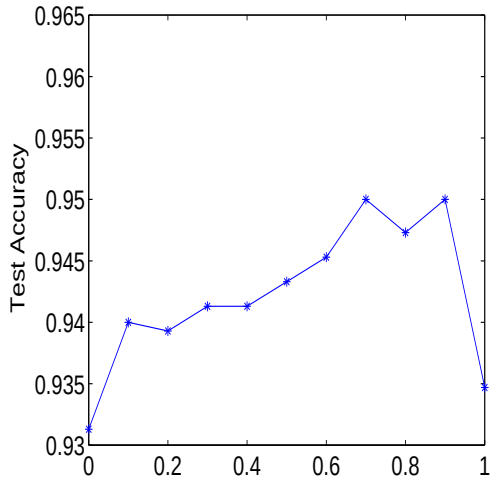
Fig. 4. The ROC curve of the methods on three real-world data sets.



(a) Web-page classification



(b) Ionosphere classification



(c) Advertisement classification

Fig. 5. The classification accuracy sensitivity with respect to ρ on three real-world data sets.

MED1. When ρ takes other values, AMVMED will combine the two views at the same time and make a balance between both views.

Fig. 5(a) shows the variation of the classification accuracy with varying ρ on web-page data. We can find that the accuracies of $\rho = 0.1$ and $\rho = 0.2$ are the highest, which demonstrate the effectiveness of AMVMED. At the same time, we can also see that the AMVMED with $\rho = 0$ (single-view MED2) performs better than the case of $\rho = 1$ (single-view MED1) and the AMVMED with $\rho = 1$ (single-view MED1) performs the worst.

In Fig. 5(b), the accuracy of AMVMED on ionosphere data decreases with the increasing ρ. The AMVMED with $\rho = 0$ shows the best performance while that with $\rho = 1$ behaves the worst. That is to say, the single-view MED2 performs the best while the single-view MED1 performs the worst. Though AMVMED with two views does not demonstrate its advantage, AMVMED indeed possesses its unique merit since it can include the two single-view MEDs as its special cases.

The performance of AMVMED with increasing ρ on advertisement data is demonstrated in Fig. 5(c). The AMVMED with $\rho = 1$ (single-view MED1) performs better than the case of $\rho = 0$ (single-view MED2) and the AMVMED with $\rho = 0$ (single-view MED2) performs the worst. The best performance occurs at $\rho = 0.7$ and $\rho = 0.9$, which demonstrates the effectiveness of AMVMED.

From Fig. 5, we can conclude that the experimental results on three data sets indeed demonstrate the effectiveness of AMVMED and the parameter ρ is important for the sake of performance. Especially when there exists some complementary property in multiple views, AMVMED usually combines the information of different views well.

F. Summary

From the above experimental results, we can conclude that on these data sets usually MVMED performs the best, our AMVMED performs a little worse than MVMED (see Table I), but sometimes AMVMED can behave better than MVMED (see the end points of curves in Fig. 1). Both MVMED and AMVMED perform better than the other methods including multi-hyperplane SVM-2K. By the sensitivity analysis of parameter ρ, the effectiveness of AMVMED for MVL is verified. However, it should be noted that our method is only an approximation of the original problem, because perfectly solving the original problem is tricky. But it may achieve better results, and thus it is valuable to explore the perfect solution of the original problem.

VII. Conclusion and Future Work

We have proposed an AMVMED framework which is the alternative version of MVMED. Different from MVMED, we not only enforce the margins of two views to be equal, but also make the posteriors of the two view margins be the same. Compared with MVMED which optimizes one relative entropy, AMVMED assigns one relative entropy term to each of the two views, hence incorporating a tradeoff between the two views. We also provide its approximate solution and give

an instantiation of the AMVMED framework with a factored prior. Experimental results on real-world applications webpage classification, ionosphere classification and advertisement classification validate the effectiveness of the proposed AMVMED.

It is worthy to further investigate how to reach the perfect solution of the original problem. In present, in order to make the solution of the AMVMED feasible, we use a two-step procedure. However, there may exist a gap between the resultant solution and the perfect solution to the original problem. Thus, we believe that analyzing the difference between our approximate solving procedure and the perfect solution, and exploring a better solving procedure are both meaningful. In addition, it is interesting to apply our method to large-scale data sets. In this paper, we have just utilized the standard toolbox to optimize AMVMED instead of specifically designing a speedup optimization algorithm for large-scale data sets. Thus it is important to investigate how to specifically deal with large-scale data sets in the future.

REFERENCES

[1] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *Proceedings of the 13th Annual Conference on Neural Information Processing Systems*, Denver, Colorado, Nov. 1999, pp. 470–476.

[2] T. Jebara and T. Jaakkola, "Feature selection and dualities in maximum entropy discrimination," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, Jul. 2000, pp. 291–300.

[3] T. Jebara, "Multi-task feature and kernel selection for SVMs," in *Proceedings of the 21st International Conference on Machine Learning*, New York, NY, Jul. 2004, pp. 55–62.

[4] ——, "Multitask sparsity via maximum entropy discrimination," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 75–110, 2011.

[5] P. M. Long and X. Wu, "Mistake bounds for maximum entropy discrimination," in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2004, pp. 833–840.

[6] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks," *Journal of Machine Learning Research*, vol. 10, no. 11, pp. 2531–2569, 2009.

[7] J. Zhu, E. P. Xing, and B. Zhang, "Laplace maximum margin Markov networks," in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, Jul. 2008, pp. 1256–1263.

[8] ——, "Partially observed maximum entropy discrimination Markov networks," in *Proceedings of the 22th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2008, pp. 1977–1984.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory*, New York, NY, Jul. 1998, pp. 92–100.

[10] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[11] E. Vatikiotis-Bateson and H. C. Yehia, "Speaking mode variability in multimodal speech production," *IEEE Transactions on Neural Networks*, vol. 13, pp. 894–899, Jul. 2002.

[12] G. F. Tzortzis and C. Likas, "Multiple view clustering using a weighted combination of exemplar-based mixture models," *IEEE Transactions on Neural Networks*, vol. 21, pp. 1925–1938, Dec. 2010.

[13] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *Journal of Machine Learning Research*, vol. 12, no. 2, pp. 491–523, 2011.

[14] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 412–424, Mar. 2012.

[15] 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*, New York, NY, Nov. 2000, pp. 86–93.

[16] R. K. Ando and T. Zhang, "Two-view feature generation model for semi-supervised learning," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, Oregon, Jun. 2007, pp. 25–32.

[17] S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao, "Bayesian co-training," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2649–2680, 2011.

[18] S. Sun and F. Jin, "Robust co-training," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 1, pp. 1113–1126, 2011.

[19] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proceedings of the 4th IEEE International Conference on Data Mining*, vol. 4, Brighton, UK, Nov. 2004, pp. 19–26.

[20] A. Kumar and H. D. III, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, Jun. 2011, pp. 393–400.

[21] S. Dasgupta, M. L. Littman, and D. McAllester, "PAC generalization bounds for co-training," in *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2002, pp. 375–382.

[22] M.-F. Balcan, A. Blum, and K. Yang, "Co-training and expansion: Towards bridging theory and practice," in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2004, pp. 89–96.

[23] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 1135–1142.

[24] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proceedings of the 22th International Conference on Machine Learning Workshop on Learning with Multiple Views*, Bonn, Germany, Aug. 2005, pp. 74–79.

[25] A. Kumar, P. Rai, and H. D. III, "Co-regularized multi-view spectral clustering," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 2011, pp. 1413–1421.

[26] S. Sun, "Multi-view Laplacian support vector machines," in *Proceedings of the 7th International Conference on Advanced Data Mining and Applications*, vol. 24, Beijing, China, Dec. 2011, pp. 209–222.

[27] M. White, X. Zhang, D. Schuurmans, and Y. Yu, "Convex multi-view subspace learning," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, United States, Dec. 2012, pp. 1682–1690.

[28] D. S. Rosenberg and P. L. Bartlett, "The Rademacher complexity of co-regularized kernel classes," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, Jun. 2007, pp. 396–403.

[29] 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*, Helsinki, Finland, Jul. 2008, pp. 976–983.

[30] S. Sun and J. Shawe-Taylor, "Sparse semi-supervised learning using conjugate functions," *Journal of Machine Learning Research*, vol. 11, no. 1, pp. 2423–2455, 2010.

[31] P. Xie and E. P. Xing, "Multi-modal distance metric learning," in *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, Beijing, China, Aug. 2013, pp. 1806–1812.

[32] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 709–722, May. 2013.

[33] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 2, pp. 1–8, 2013.

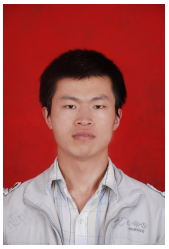
[34] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2005, pp. 355–362.

[35] N. Chen, J. Zhu, and E. P. Xing, "Predictive subspace learning for multi-view data: A large margin approach," in *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2010, pp. 361–369.

[36] S. Sun and G. Chao, "Multi-view maximum entropy discrimination," in *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, Beijing, China, Aug. 2013, pp. 1706–1712.

[37] P. Domingos, "Bayesian averaging of classifiers and the overfitting problem," in *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA, Jun. 2000, pp. 223–230.

[38] R. Fletcher, *Practical Methods of Optimization (2nd Ed.)*. New York, NY: Wiley-Interscience, 1987.



Guoqing Chao received the B.S. degree from Xinyang Normal University, Xinyang, China, in 2009. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, East China Normal University, Shanghai, China.

His current research interests include machine learning and pattern recognition.

Shiliang Sun received the B.E. degree in automatic control from the Department of Automatic Control, Beijing University of Aeronautics and Astronautics in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Department of Automation and the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China, in 2007. In 2004, he was entitled Microsoft Fellow.

He is currently a professor at the Department of Computer Science and Technology and the head of the Pattern Recognition and Machine Learning Research Group, East China Normal University. From 2009 to 2010, he was a visiting researcher at the Department of Computer Science, University College London, working within the Centre for Computational Statistics and Machine Learning. From March to April 2012, he was a visiting researcher at the Department of Statistics, Rutgers University. He is a member of the PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) network of excellence, and on the editorial boards of multiple international journals including Neurocomputing and IEEE Transactions on Intelligent Transportation Systems. His research interests include approximate inference, learning theory, sequential modeling, kernel methods, and their applications.