SEMI-SUPERVISED MULTI-SENSOR CLASSIFICATION VIA CONSENSUS-BASED MULTI-VIEW MAXIMUM ENTROPY DISCRIMINATION

Tianpei Xie[†], *Nasser M. Nasrabadi*[⋆] *and Alfred O. Hero III*[†]

[†]Dept. of Electrical Eng., System, University of Michigan, Ann Arbor, MI 48109
[⋆] U.S. Army Research Lab., 2800 Powder Mill Road, Adelphi, MD, USA
[†] {tianpei, hero}@umich.edu, [⋆] nasser.m.nasrabadi.civ@mail.mil

ABSTRACT

In this paper, we consider multi-sensor classification when there is a large number of unlabeled samples. The problem is formulated under the multi-view learning framework and a Consensus-based Multi-View Maximum Entropy Discrimination (CMV-MED) algorithm is proposed. By iteratively maximizing the stochastic agreement between multiple classifiers on the unlabeled dataset, the algorithm simultaneously learns multiple high accuracy classifiers. We demonstrate that our proposed method can yield improved performance over previous multi-view learning approaches by comparing performance on three real multi-sensor data sets.

Index Terms— sensor networks, multi-view learning, maximum entropy discrimination, kernel machine

1. INTRODUCTION

In many applications, e.g., in sensor networks, data is collected from multiple sensors and, given that complementary information is present within different sensors, classification using all sensors is expected to yield higher performance as compared to its single-sensor counterpart [1]. Furthermore, as class labeling can be labor intensive, in many situations many training samples may not be labeled. In the machine learning literature, this problem falls under the framework of semi-supervised multi-view learning [2], since the partially-labeled samples are multi-modal in nature and each modality corresponds to one view of physical event.

Most methods to multi-sensor or multi-view classification either rely on feature fusion (early fusion) methods, that find an intermediate joint representation of multiple views [3, 4], or, on decision fusion (late fusion) methods that combine decisions from multiple models to improve the overall performance [5]. Unless the features are optimized for multi-view aggregation, there is no guarantee that feature fusion will lead to good classification performance. In this paper, we pursue a different approach that learns an intermediate model, or a *consensus view* to fuse features from different views, and improves simultaneously the performance of each single-view classifier. Moreover, we propose to train a set of *stochastic classifiers* to handle the large number of unlabeled training samples.

We follow the principle of the disagreement-based multi-view learning [2, 6, 7, 8, 9, 10, 11]. In particular, it is shown in [12] that the error rate of each classifier in the multi-view system is bounded above by the rate of disagreement between multiple view-specific classifiers. In other word, the algorithm that explicitly minimizes the disagreement between multiple view-specific classifiers would learn a set of compatible classifiers with high performance and low sample complexity. In this paper, we propose a Consensus-based Multi-View Maximum Entropy Discrimination (CMV-MED) algorithm that learns a set of classifiers, one for each view, by iteratively maximizing their *stochastic agreement* on the unlabeled training data. Our method is based on the Maximum Entropy Discrimination (MED) by Jaakkola et al. [13]. MED is a Bayesian learning approach that generalizes support vector machine (SVM) classifiers and explicitly incorporate the large-margin training [14] into a unified maximum entropy learning framework. We show the superior performance of our model over previous multi-view learning approaches by comparing performance on three real multi-sensor data sets.

This paper is structured as follows: an overview of the MED model is given in Section 2 and we propose the general model for CMV-MED in Section 3. The algorithm for solving CMV-MED is discussed in Section 4. In Section 5, experiments on a set of real multi-view data sets are discussed.

2. MAXIMUM ENTROPY DISCRIMINATION (MED)

We denote the multi-view data set as $\mathcal{D}_V$. $\mathcal{D}_V$ consists of the labeled part $\{(\mathbf{x}_n, y_n), n \in L\}$ and the unlabeled part $\{\mathbf{x}_m, m \in U\}$, where L and U represent the index set of labeled and unlabeled samples, respectively, and $|L| \ll |U|$. Define the multi-view feature $\mathbf{x}_n = [\mathbf{x}_n^1, \ldots, \mathbf{x}_n^V], \forall n \in L \cup U$, where $\mathbf{x}_n^i \in \mathcal{R}^{d_i}$ are the features extracted from view i and V is the number of views. Here we consider the *binary classification* task, i.e., $y \in |\mathcal{Y}| = \{-1, +1\}$. Let $\mathcal{D}^i$ be the

Acknowledgement: This research was partially supported by US Army Research Office (ARO) grants W911NF-11-1-0391 and WA11NF-11-1-103A1.

set of samples collected from the single view i. In this section, we focus on the single-view MED on labeled subset L.

For a single view $i \in [1, \ldots, V]$, assume the predictive distribution is a generalized log-linear model, i.e., $\log p_i(y|\mathbf{x}^i, \mathbf{w}_i) \propto \frac{1}{2} y \left(\mathbf{w}_i^T \Phi_i(\mathbf{x}^i)\right) \equiv F_i(y, \mathbf{x}; \mathbf{w}_i)$ and $\Phi_i : \mathcal{R}^{d_i} \mapsto \mathcal{R}^{p_i}$ is a prescribed feature map defined in view i. Define the kernel function $K_i : \mathcal{R}^{d_i} \times \mathcal{R}^{d_i} \mapsto \mathcal{R}$ that satisfies $\langle \mathbf{\Phi}_i(\mathbf{x}_n^i), \mathbf{\Phi}_i(\mathbf{x}_m^i) \rangle = K_i(\mathbf{x}_n, \mathbf{x}_m)$, for $\forall \mathbf{x}_n^i, \mathbf{x}_m^i \in \mathcal{D}^i$ in view i and $F_i(y, \mathbf{x}^i; \mathbf{w}_i)$ is the normalized log-likelihood function parameterized by $\mathbf{w}_i$ in the kernel space.

Denote the prior distribution of $\mathbf{w}_i$ as $p_0(\mathbf{w}_i)$. The goal for Maximum Entropy Discrimination [13] is to learn a post-data (posterior) distribution $q(\mathbf{w}_i)$, by solving an entropic regularized risk minimization problem with the prior on model parameter $\mathbf{w}_i$ specified as $p_0(\mathbf{w}_i)$

$$\min_{q(\mathbf{w}_i)} \mathbb{KL}\left(q(\mathbf{w}_i) \| p_0(\mathbf{w}_i)\right) + \sum_{n \in L} \left[1 - \mathbb{E}_{q(\mathbf{w}_i)}\{\Delta F_i(y_n, \mathbf{x}_n^i; \mathbf{w}_i)\}\right]_+, \quad (1)$$

where $[s]_+ = \max\{s, 0\}$. $\mathbb{KL}(p\|q)$ is the *Kullback-Leibler divergence* from distribution p to q, i.e., $\mathbb{KL}\left(q(\mathbf{w}_i)\|p_0(\mathbf{w}_i)\right) = \int_\Theta q(\mathbf{w}_i) \log\left(\frac{q(\mathbf{w}_i)}{p_0(\mathbf{w}_i)}\right) d\mathbf{w}_i$ and $\Delta F_i(y_n, \mathbf{x}_n; \mathbf{w}_i) \equiv F_i(y_n, \mathbf{x}_n^i; \mathbf{w}_i) - F_i(y \neq y_n, \mathbf{x}_n^i; \mathbf{w}_i) = \log\left(\frac{p(y_n|\mathbf{x}_n^i, \mathbf{w}_i)}{p(y \neq y_n|\mathbf{x}_n^i, \mathbf{w}_i)}\right)$ is the log-odds classifier.

The second term in (1) is a hinge-loss that captures the large-margin principle underlying the MED prediction rule,

$$y^* = \text{argmax}_y \, \mathbb{E}_{q(\mathbf{w}_i)}\left[F(y, \mathbf{x}^i; \mathbf{w}_i)\right].$$

If we use a *Gaussian Process* [16] as the prior on $\mathbf{w}_i$, i.e., $p_0(\mathbf{w}_i) = \mathcal{N}(\mathbf{w}_i; 0, \sigma^2 I_{p_i})$, a kernel SVM is obtained by solving (1) in its dual formulation. For multi-view data, it is necessary to learn multiple MEDs simultaneously. For example, in [17], the author applies a joint sparsity prior on $(\mathbf{w}^1, \ldots, \mathbf{w}^V)$ to achieve multi-task feature selection. Instead of assuming a joint prior on all multi-view model parameters, we utilize the available unlabeled samples and require the class prediction of multiple models to agree with each other.

3. CONSENSUS-BASED MULTI-VIEW MED: A GENERAL FRAMEWORK

Define the *consensus view model* as a parameter-free distribution $q(y|\mathbf{x}_n) \in \mathcal{Q}$ on the unlabeled set U, where $\mathbf{x}_n = [\mathbf{x}_n^1, \ldots, \mathbf{x}_n^V], \forall n \in U$, $\mathcal{Q} \equiv \left\{q(x) : q(x) \geq 0, \int q(x)dx = 1\right\}$ and $q(y|\mathbf{x}_n) = \delta\{y = y_n\}$, $n \in L$. In each view i, a joint post-data distribution is obtained as $q_i(y, \mathbf{w}_i|\mathbf{x}) = q(y|\mathbf{x})q(\mathbf{w}_i)$, where $q(y|\mathbf{x})$ is *shared* among all views and the above equality reflects the mean-field approximation.

The *goal* of *Consensus-based Multi-view Maximum Entropy Discrimination* (CMV-MED) is to simultaneously learn the joint post-data distributions $q_i(y, \mathbf{w}_i|\mathbf{x}) = q(y|\mathbf{x})q(\mathbf{w}_i)$, given the priors $p_i(y, \mathbf{w}_i|\mathbf{x}^i) = p_i(y|\mathbf{w}_i, \mathbf{x}^i)p_0(\mathbf{w}_i)$ for $\mathbf{x}^i \in \mathcal{D}^i, \forall i = 1, \ldots, V$. This is accomplished by solving the following optimization problem

$$\min_{\substack{q_i(y, \mathbf{w}^i|\mathbf{x}_n) \in \mathcal{Q}, \\ \forall i = 1, \ldots, V, \; n \in L \cup U}} \sum_{n \in L} \sum_{i=1}^{V} \left[1 - \mathbb{E}_{q_i(y, \mathbf{w}^i|\mathbf{x}_n)}\{\Delta F_i(y, \mathbf{x}_n^i; \mathbf{w}_i)\}\right]_+ + \lambda \sum_{n \in U} \sum_{i=1}^{V} \pi_i \mathbb{KL}\left(q_i(y, \mathbf{w}^i|\mathbf{x}_n) \| p_0(y, \mathbf{w}^i|\mathbf{x}_n^i)\right), \quad (2)$$

where $\pi_i \in \left\{\pi_j : \sum_{j=1}^V \pi_j = 1, \quad \pi_j \geq 0, \forall j\right\}$ is a parameter for view i and $\lambda > 0$ is regularization parameter. Note that $q_i(y, \mathbf{w}_i|\mathbf{x}_n) = \delta\{y = y_n\} q(\mathbf{w}_i)$ on the labeled set L and the second term can be further expanded as

$$\mathbb{KL}\left(q_i(y, \mathbf{w}^i|\mathbf{x}_n) \| p_0(y, \mathbf{w}^i|\mathbf{x}_n^i)\right) = \mathbb{KL}\left(q(\mathbf{w}^i) \| p_0(\mathbf{w}^i)\right) + \mathbb{E}_{q(\mathbf{w}^i)}\left[\mathbb{KL}\left(q(y|\mathbf{x}_n) \| p_i(y|\mathbf{x}_n^i, \mathbf{w}^i)\right)\right], i = 1, \ldots, V. \quad (3)$$

Substituting (3) into (2), we have the following

$$\min_{\substack{q(y|\mathbf{x}_n) \in \mathcal{Q}, n \in U \\ q(\mathbf{w}^i), \forall i = 1, \ldots, V}} \sum_{n \in L} \sum_{i=1}^{V} \left[1 - \mathbb{E}_{q(\mathbf{w}^i)}\{\Delta F_i(y_n, \mathbf{x}_n^i; \mathbf{w}_i)\}\right]_+ + \lambda \sum_{i=1}^{V} \pi_i \mathbb{KL}\left(q(\mathbf{w}^i) \| p_0(\mathbf{w}^i)\right) + \lambda \sum_{n \in U} \sum_{i=1}^{V} \pi_i \mathbb{E}_{q(\mathbf{w}^i)}\left[\mathbb{KL}\left(q(y|\mathbf{x}_n) \| p_i(y|\mathbf{x}_n^i, \mathbf{w}^i)\right)\right]. \quad (4)$$

From (4), we see that the first and second term learn V view-specific MED models $q(\mathbf{w}^i), i = 1, .., V$, simultaneously.

Our main contribution is the third term in (4), which is referred as the *consensus-based disagreement term* on unlabeled set, since it is zero when view-specific predictive models $p_i(y|\mathbf{x}_n^i, \mathbf{w}^i)$ all equal, $i = 1, ..., V$, while it penalizes more when one deviates far from the consensus model $q(y|\mathbf{x})$, which, by construction, is the *center* of these V distributions in the information geometry over the space of probability measures. This center is determined by information projection accomplished by the KL divergence in (4). By incorporating this term, we explicitly require all classifiers to make similar class predictions having similar confidence levels on the unlabeled training samples. The benefit for enforcing the consensus-based disagreement is that the proposed model is sensitive in the case when view-specific classifiers with low confidence agree with each other, while it is lenient when all of them are highly confident and agree. Thus the model is reliable in the situation where the initial view-specifc classifiers only have low confidence results due to the limited size of labeled training set. Fig. 1 is a graphical model representation for the information projection.

4. SOLUTION VIA DETERMINISTIC ANNEALING EXPECTATION MAXIMIZATION

Our solution for CMV-MED in (4) is based on the *deterministic annealing EM* [18]. It is described as the following steps:

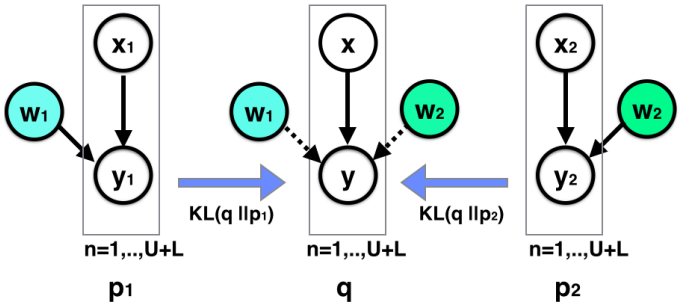

Fig. 1: A graphical model representation for consensus-based multi-view learning via information projection.

1. Set the regularization parameter $\lambda_0 = 0$ in (4) at initialization and train V independent MED classifiers simultaneously to find $q_0(\mathbf{w}^i)$, $i = 1, \ldots, V$. Set the prior distribution $p_0(\mathbf{w}^i) = \mathcal{N}(\mathbf{w}^i : 0, \sigma^2 I)$ and $\pi_i = \frac{1}{V}, \forall i$. Let T be the maximum number of iterations.
2. For $t = 1, \ldots, T$, do
   (a) Given the post-data distribution $q_{t-1}(\mathbf{w}^i), i = 1, \ldots, V$ from MED, find the consensus view on unlabeled data U via information projection, i.e.

$$q_t(y|\mathbf{x}_n) = \text{argmin}_q \frac{1}{V} \sum_{i=1}^{V} \mathbb{E}_{q(\mathbf{w}^i)} \left[\mathbb{KL} \left(q_n(y) \| p_{i,n}(y|\mathbf{w}^i) \right) \right]$$

$$\Rightarrow \log q_t(y|\mathbf{x}_n) = \frac{1}{V} \sum_{i=1}^{V} \log p_{i,n}(y|\hat{\mathbf{w}}_{t-1}^i) - \log Z(\mathbf{x}_n), \quad \forall n \in U,$$

   where $q_n(y) \equiv q(y|\mathbf{x}_n)$, $p_{i,n}(y|\mathbf{w}^i) \equiv p_i(y|\mathbf{x}_n^i, \mathbf{w}^i)$ for $n \in U$, $Z(\mathbf{x}_n)$ is the normalization factor and $\hat{\mathbf{w}}_{t-1}^i$ is the mean of the post-data distribution $q_{t-1}(\mathbf{w}^i), i = 1, \ldots, V$.

   (b) Given the consensus view $q_t(y|\mathbf{x}_n), \forall n \in U$, substitute it into (4) to obtain the following optimization problem

$$\min_{q(\mathbf{w}^i), \forall i=1,\ldots,V} \sum_{n \in L} \sum_{i=1}^{V} \left[1 - \mathbb{E}_{q(\mathbf{w}^i)} \{ \Delta F_i(y_n, \mathbf{x}_n^i; \mathbf{w}_i) \} \right]_+ + \lambda_t \frac{1}{V} \sum_{n \in U} \sum_{i=1}^{V} \mathbb{E}_{q(\mathbf{w}^i)} \left[\mathbb{E}_{q_t(y|\mathbf{x}_n)} \left[-\log p_i(y|\mathbf{x}_n^i, \mathbf{w}^i) \right] \right] + \sum_{i=1}^{V} \pi_i \mathbb{KL} \left(q(\mathbf{w}^i) \| p_0(\mathbf{w}^i) \right)$$

   For each view i, compute the $q_t(\mathbf{w}^i|\mathcal{D}^i, \boldsymbol{\alpha}^i)$ with dual parameter $\boldsymbol{\alpha}^i = [\alpha_1^i, \ldots, \alpha_L^i]^T$ by solving the following dual programming problem, i.e.,

$$\max_{\boldsymbol{\alpha}^i} \mathbf{1}^T \boldsymbol{\alpha}^i - \frac{\sigma^2}{2} (\boldsymbol{\alpha}^i)^T (\widetilde{\mathbf{K}}_i \odot \mathbf{y}\,\mathbf{y}^T) \boldsymbol{\alpha}^i \quad (5)$$
$$\text{s.t. } \mathbf{0} \preceq \boldsymbol{\alpha}^i \preceq \mathbf{1},$$

   where $\mathbf{1} = [1, \ldots, 1]^T$ and $\odot$ is piece-wise product. In (5), *a new kernel* $\widetilde{\mathbf{K}}_i$ is computed via

$$\widetilde{\mathbf{K}}_i = \mathbf{K}_{L,i} - \lambda_t (\mathbf{k}_{UL}^i)^T \left[1/\sigma^2 \mathbf{M}_i^{-1} + \lambda_t \mathbf{K}_{U,i} \right]^{-1} \mathbf{k}_{UL}^i \quad (6)$$
$$\equiv [\langle \widetilde{\Phi}_i(\mathbf{x}_n^i)\,,\ \widetilde{\Phi}_i(\mathbf{x}_m^i) \rangle]_{n,m \in L}, \quad (7)$$

   where $\mathbf{K}_{L,i} = [K_i(\mathbf{x}_n^i, \mathbf{x}_m^i)]_{n,m \in L}$, $\mathbf{K}_{U,i} = [K_i(\mathbf{x}_n^i, \mathbf{x}_m^i)]_{n,m \in U}$ and $\mathbf{k}_{UL}^i = [K_i(\mathbf{x}_n^i, \mathbf{x}_m^i)]_{n \in U, m \in L}$. $\mathbf{M}_i = diag\{\nu_1, \ldots, \nu_U\} \in \mathcal{R}^{|U| \times |U|}$, with $\nu_n \equiv \mathbb{E}_{q_t(y|\mathbf{x}_n)} \left[-\nabla_{\mathbf{w}^i}^2 \log p_i(y|\mathbf{x}_n^i, \hat{\mathbf{w}}_{t-1}^i) \right], n \in U$.

   Then the post-data distribution $q_t(\mathbf{w}^i|\mathcal{D}^i, \boldsymbol{\alpha}^i) = \mathcal{N}(\hat{\mathbf{w}}_t^i, \mathbf{H}_i)$, where the mean is given by $\hat{\mathbf{w}}_t^i = \sum_{m=1}^{L} y_m \alpha_m^i \widetilde{\Phi}_i(\mathbf{x}_n^i)$. The covariance matrix $\mathbf{H}_i = \left(\sigma^2 I + \boldsymbol{\Phi}_i(\mathbf{X}_U)^T \mathbf{M}_i \boldsymbol{\Phi}_i(\mathbf{X}_U) \right)$ with $\boldsymbol{\Phi}(\mathbf{X}_U) \equiv [\Phi_i(\mathbf{x}_1^i), \ldots, \Phi_i(\mathbf{x}_U^i)]^T \in \mathcal{R}^{|U| \times p_i}$.

   (c) Set $\lambda_t = 1 - e^{-0.5t} \to 1$ as t increases.

   (d) $t \leftarrow t + 1$.

3. Finally, make prediction based on consensus view

$$y^* = \text{argmax}_{\hat{y}} \sum_{1 \le i \le V} \mathbb{E}_{q(y, \mathbf{w}_i)} \left[\delta \{ y = \hat{y} \} F(y, \mathbf{x}^i; \mathbf{w}_i) \right].$$

Note that the Step 2(b) can be performed in parallel, as it does not rely on information from other views.

5. EXPERIMENTS

We compare the proposed CMV-MED model with the SVM-2K model proposed by Farquhar et al. [7], the MV-MED model by Sun et al. [11] as well as the conventional MED for each view on several real multi-view data sets. In the following experiments, we focus on two-view learning, i.e. $V = 2$ and use the Gaussian Kernel function $K_i(\mathbf{x}_n^i, \mathbf{x}_m^i) = \exp(c \|\mathbf{x}_n^i - \mathbf{x}_m^i\|^2), i = 1, 2$. For all MED-based methods, a Gaussian Process prior $p_0(\mathbf{w}^i) = \mathcal{N}(\mathbf{0}, \sigma_i^2 I)$ is assigned for view $i = 1, 2$. The view parameter $\pi_1 = \pi_2 = \frac{1}{2}$. All other parameters for each model are obtained by 5-fold-cross-validation. All the experiments are repeated for 20 times, with randomly chosen L and U.

5.1. Footstep Classification

We test on **ARL-Footstep** [19, 20] data, which is a multi-sensor data set that contains acoustic signals collected by four well-synchronized sensors (labeled as Sensor 1,2,3,4) in a natural environment. The task is to discriminate between human footsteps and human-leading animal footsteps. We only use Sensor $1, 2$ in our experiment. It involves 840 segments from human subjects and 660 segments from human-animal subjects. We choose 600 segments from each class as the training set with $|L| = 50$, and the rest is designated as the test set. A 200-dimensional mel-frequency cepstral coefficients (MFCCs) vector is computed from the corresponding segments in all the views, with normalization as in [20].

In Table 1, we see that our CMV-MED outperforms both SVM-2K and MV-MED, and it improves over the single-view MED. This is likely because our method utilizes the confidence as well as decision as a disagreement measure,

Classification Accuracy (%) mean ± standard error					
Dataset.	MED (single views)		SVM-2K	MV-MED	CMV-MED
ARL Footstep (Sensor 1,2, $\|L\| = 50$)	71.1 ± 5.3	62.3 ± 10.2	73.3 ± 5.2	75.6 ± 6.5	$\mathbf{85.5} \pm 6.1$
WebKB4 ($\|L\| = 15$)	76.6 ± 10.2	77.1 ± 10.1	79.0 ± 10.0	77.9 ± 8.7	$\mathbf{91.7} \pm 5.8$
Internet Ads ($\|L\| = 50$)	87.3 ± 0.9	86.2 ± 1.4	82.5 ± 4.3	88.8 ± 2.3	$\mathbf{92.7} \pm 0.7$

Table 1: Classification accuracy with different data set, with the best performance shown in **bold**.

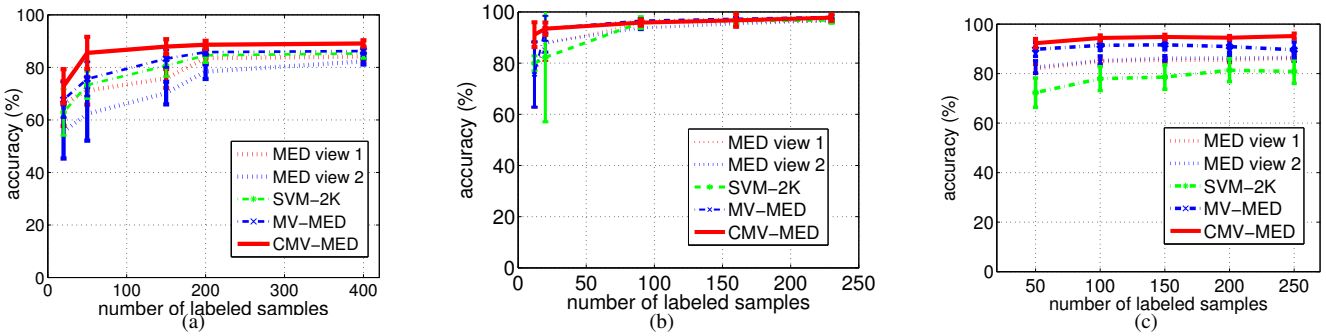

Fig. 2: The classification accuracy vs. the size of labeled set for (a) **ARL-Footstep** data set, (b) **WebKB4** data set and (c) **Internet Ads** data set. The proposed CMV-MED outperforms MV-MED, SVM-2K and two single-view MEDs (view 1 and 2) and it has good stability when the number of labeled samples is small.

In **ARL-Footstep** data, since the signal is contaminated by background noise, the original MED on two single views does not perform well, and both the decision regularization and margin regularization are not as reliable as the confidence regularization implemented by CMV-MED.

Fig. 2(a) shows the accuracy and the standard deviation for the four methods as the size of the labeled set increases. As more ground truth labels are used, the performances of all training methods increases, while CMV-MED shows its superior performance consistently.

5.2. Web-Page Classification

The **WebKB4** [21] data set is widely-used in multi-view learning literature [6, 10]. It consists of 1051 two-view web pages collected from computer science department web sites at four universities. There are 230 course pages and 821 non-course pages. The two natural views are words in a web page and words appearing in the links pointing to that page. We follow the preprocessing step in [10], and extract a 3000-dimensional feature vector via the bag-of-words representation in the page view and a 1840-dimensional feature vector in the link view. Then we compute the term frequency-inverse document frequency weights (TF-IDF) features from the document word matrix. The feature vector is length normalized.

In Table 1, we see that our CMV-MED has significantly better performance as compared to SVM-2K and MV-MED, when the labeled set is small, i.e., $|L| = 15$. Also, according to Fig. 2(b), when more labeled samples are included, all four methods have similarly good performance, even for the single-view MED. The CMV-MED performs better with a few labeled samples because its stability relies on a good estimate of confidence on the unlabeled training samples, which is less affected by the amount of the labeled training samples.

5.3. Internet Advertisement Classification

The **Internet Ads** [22] data set consists of 3279 instances including 458 ads images and 2820 non-ads images. The first view describes the image itself, i.e., words in images' URL and caption, while the other view contains all other features, i.e., words from URLs of pages that contain the image and pages which the image points to. For each view, we extract the bag-of-words representations, which results in a $587-$dimensional vector in view 1 and a $967-$dimension vector in view 2. We set the size of training set as 600 and $|L| = 50$.

From Table 1 and Fig. 2(c) , we see that our CMV-MED still performs better than SVM-2K, MV-MED and single-view MED. It is seen that CMV-MED is more stable as the size of the labeled training set increases, while SVM-2K has much worse stability performance.

6. CONCLUSION

In this paper, we propose a consensus-based multi-view maximum entropy learning model that incorporates large-margin classification and Bayesian learning when a large amount of unlabeled samples from multiple sources are available. The experimental results on three different real data sets show the superiority of the proposed CMV-MED over other multi-view large-margin classification methods in terms of classification accuracy, especially when the number of labeled samples is small compared to the unlabeled ones.

7. REFERENCES

[1] Ning Xiong and Per Svensson, "Multi-sensor management for information fusion: issues and approaches," *Information fusion*, vol. 3, no. 2, pp. 163–186, 2002.

[2] Zhi-Hua Zhou and Ming Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.

[3] Pei Ling Lai and Colin Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural Systems*, vol. 10, no. 05, pp. 365–377, 2000.

[4] Aaron Shon, Keith Grochow, Aaron Hertzmann, and Rajesh P Rao, "Learning shared latent structure for image synthesis and robotic imitation," in *Advances in Neural Information Processing Systems*, 2005, pp. 1233–1240.

[5] Stan Z Li, Long Zhu, ZhenQiu Zhang, Andrew Blake, HongJiang Zhang, and Harry Shum, "Statistical learning of multi-view face detection," in *Computer Vision ECCV 2002*, pp. 67–81. Springer, 2002.

[6] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory (COLT)*. ACM, 1998, pp. 92–100.

[7] Jason Farquhar, David Hardoon, Hongying Meng, John S Shawe-taylor, and Sandor Szedmak, "Two view learning: SVM-2K, theory and practice," in *Advances in neural information processing systems*, 2005, pp. 355–362.

[8] Shipeng Yu, Balaji Krishnapuram, Harald Steck, RB Rao, and Rómer Rosales, "Bayesian co-training," in *Advances in Neural Information Processing Systems*, 2007, pp. 1665–1672.

[9] Kuzman Ganchev, João V Graça, John Blitzer, and Ben Taskar, "Multi-view learning over structured and non-identical outputs," in *Proceedings of the Converence on Uncertainty in Artificial Intelligence (UAI)*, 2008.

[10] Vikas Sindhwani and David S Rosenberg, "An RKHS for multi-view learning and manifold co-regularization," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 976–983.

[11] Shiliang Sun and Guoqing Chao, "Multi-view maximum entropy discrimination," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1706–1712.

[12] Sanjoy Dasgupta, Michael L Littman, and David McAllester, "PAC generalization bounds for co-training," in *Advances in neural information processing systems*. 2002, vol. 1, pp. 375–382, MIT; 1998.

[13] Tommi Jaakkola, Marina Meila, and Tony Jebara, "Maximum entropy discrimination," in *Advances in neural information processing systems*, 1999.

[14] Ben Taskar, Carlos Guestrin, and Daphne Koller, "Max-margin markov networks," *Advances in neural information processing systems*, vol. 16, pp. 25, 2004.

[15] Thomas Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *Communication Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–60, 1967.

[16] Carl Rasmussen and Chris Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[17] Tony Jebara, "Multitask sparsity via maximum entropy discrimination," *The Journal of Machine Learning Research*, vol. 12, pp. 75–110, 2011.

[18] Vikas Sindhwani, S Sathiya Keerthi, and Olivier Chapelle, "Deterministic annealing for semi-supervised kernel machines," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 841–848.

[19] Thyagaraju Damarla, Asif Mehmood, and James Sabatier, "Detection of people and animals using non-imaging sensors," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.

[20] Nam H Nguyen, Nasser M Nasrabadi, and Trac D Tran, "Robust multi-sensor classification via joint sparse representation," *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.

[21] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery, "Learning to construct knowledge bases from the world wide web," *Artificial intelligence*, vol. 118, no. 1, pp. 69–113, 2000.

[22] Nicholas Kushmerick, "Learning to remove internet advertisements," in *Proceedings of the third annual conference on Autonomous Agents*. ACM, 1999, pp. 175–181.