Auxiliary Information Regularized Machine for Multiple Modality Feature Learning*

Yang Yang, Han-Jia Ye, De-Chuan Zhan, Yuan Jiang
National Key Laboratory for Novel Software Technology, Nanjing University,
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing, Jiangsu, China
{yangy, yehj, zhandc, jiangy}@lamda.nju.edu.cn

Abstract

In real world applications, data are often with multiple modalities. Previous works assumed that each modality contains sufficient information for target and can be treated with equal importance. However, it is often that different modalities are of various importance in real tasks, e.g., the facial feature is weak modality and the fingerprint feature is strong modality in ID recognition. In this paper, we point out that different modalities should be treated with different strategies and propose the Auxiliary information Regularized Machine (ARM), which works by extracting the most discriminative feature subspace of weak modality while regularizing the strong modal predictor. Experiments on binary and multi-class datasets demonstrate the advantages of our proposed approach ARM.

1 Introduction

With the development of data collection techniques, multiple modal data can be acquired from many applications, e.g., modern mobile phones with types of sensors can collect sensor signals from multiple channels. In order to utilize the information from different modalities, more and more attentions have been paid to multi-modal learning. Ngiam *et al.*[2011] applied deep networks to learn features over multiple modal data; Zhai *et al.*[2013] used multi-modal method for efficient large-scale similarity search; Nguyen *et al.*[2013] proposed the M3LDA to annotate image regions and provide a promising way to understand the relation between input patterns and output semantics. The main assumption behind these methods is that each modality contains sufficient information for target tasks and is with equal importance.

Nevertheless, different modalities are of various importance under specific circumstance. In this paper, we denote the important modalities with particular tasks by strong modality. On the contrary, the modalities with less importance are weak modality. E.g., in identification systems, recognizer can more easily tell the ID with the fingerprints provided, so we denote the fingerprints as strong modality, while features like face descriptors and gaits are usually less accurate, they are weak modalities in this task. It is notable that strong modal features can lead to a better performance, nevertheless, are more expensive, therefore a group of serialized feature extraction methods were proposed. These methods extract weak modal features firstly, and then extract more strong modal features gradually to improve the performance and reduce the overall cost as well. Marcialis *et al.*[2010] proposed a serial fusion technique for multiple biometric modal features through extracting gaits information and face information step by step; Zhang *et al.*[2014] addressed the serialized multi-modal learning techniques in a semi-supervised learning scenario. These methods handle strong and weak modalities independently while leaving the fact of unsatisfied performance on weak modality unexplained. In this paper, we consider the issue of unsatisfied classification performance on weak modality, and attribute the phenomena into two different reasons:

- Weak modality can be with insufficient information, e.g., face with occluded illumination conditions;
- Weak modality contains helpful information, but these information are 'hidden' behind other irrelevant factors, e.g., noise, irrelevant features.

In the second situation, how to extract the discriminative features of weak modality by feature learning and improve the multi-modal learning classification performance are the urgent problem. In this paper, we will mainly focus on this situation for discussion.

We therefore propose the ARM (Auxiliary information Regularized Machine) approach to extract the most discriminative features of weak modality with the auxiliary information of strong modality in multi-modal learning. In supervised learning, lots of feature extraction methods have been proposed, we can easily use sufficient supervised information in these methods, e.g., face recognition with supervised feature extraction methods. However, here is usually in semi-supervised scenarios when performing multi-modal learning, and it is difficult to improve the performance with a limited amount of labeled data, while it is easy to gather the strong modal information corresponding to weak modality. In this paper, we consider the auxiliary information provided by strong modality, and utilize them to help weak modality extract more discriminative features. Meanwhile, through the

*This research was supported by NSFC(61273301, 61305067).

consistency of classifiers on strong and weak modality, we want to get better multi-modal learning performance. The effectiveness of the proposed ARM approach is validated in our empirical investigations.

Section 2 is related work. Section 3 presents our approach, and the algorithm is presented in Section 4. Section 5 reports our experiments. Finally, Section 6 concludes.

2 Related Work

Our ARM approach can handle multiple modal information in semi-supervised scenarios as well as extracting informative features of weak modality by feature learning. Therefore our work is closely related to: multi-modal learning, semi-supervised learning and feature learning.

The exploitation of multiple modalities has attracted many attentions recently. The mainstream of the multi-modal learning models handle multi-modal information in parallel or serialized style, especially for biometric learning tasks, e.g., [Hong and Jain, 1998], [Zhou *et al.*, 2005], [Zhang and Li, 2014]. However, these methods handle each modality independently first and then fuse them with late fusion strategies, without considering the correlation between different modalities, while in recent, Kiros *et al.*[2014] jointly learn word representations and image features by convolutional network.

Many studies have made efforts on the exploitation of multi-modal learning with unlabeled data. Modern multi-modal semi-supervised learning methods mainly derive from co-training methods. Co-training [Blum and Mitchell, 1998] is the most famous multi-modal semi-supervised learning method, which trains two classifiers separately on two modalities and then uses them to label unlabeled instances for each other. Co-training style semi-supervised learning approaches have been well developed in decades, e.g., [Kiritchenko and Matwin, 2011], [Zhang and Zhou, 2011]. However, previous methods mainly focus on how to label the unlabeled data between different modalities while leaving the importance of different modalities unconsidered.

In feature learning research, conventional feature learning methods liking PCA [Jolliffe, 2005], LPP [He and Partha, 2003], LDA [Fisher, 1936] etc. aim to extract features in a low dimension space. Feature learning methods are also used in many applications, e.g., [Xu *et al.*, 2010], [Guo and Xue, 2013], [Xie *et al.*, 2014]. However, most existing feature learning methods require supervised information. Yet it is hard to get enough labeled examples in practice.

To the best of our knowledge, previous multi-modal methods assume that each modality would be sufficient for classification without differences of 'strong' or 'weak'. But in practice, it is difficult to build a classifier just using weak modality with limited labeled data, so we turn to utilize auxiliary information provided by the strong modality. In this paper, we proposed an Auxiliary information Regularized Machine (ARM) which utilizes the multi-modal information in semi-supervised scenario. Moreover, ARM can exploit the weak modal features more extensively, extract the useful feature subspaces from weak modalities. Therefore, with ARM we can only extract/utilize the relatively inexpensive weak modal features, and obtain an improved weak modal prediction performance during the test phase.

3 Proposed Method

Suppose we have N examples, where labeled examples are denoted by $\{(x_i, y_i)\}_{i=1}^{N_1}$ and unlabeled instances are $\{x_j\}_{j=N_1+1}^{N}$, where $x_i \in \mathcal{R}^d$, and $y_i \in \{0, 1\}$ which denotes the class label of the i-th sample. d is the feature dimension. Meanwhile, in multi-modal learning, instance space can be denoted as two parts without overlap, $V = \{V_1^{d_1}, V_2^{d_2}\}$, where $V_1^{d_1}$ is the feature space derived from weak modality and $V_2^{d_2}$ is the features of strong modality, $d = d_1 + d_2$. Consequently, each instance x_i can be denoted as (x_{i,v_1}, x_{i,v_2}).

Auxiliary Information Regularized Machine(ARM)

Auxiliary information Regularized Machine (ARM) aims to improve the multi-modal learning performance while extracting the most discriminative weak modal feature subspace at the same time. As a consequence, we can get more accurate predictions with either only the weak or the strong modal features provided in the test phase.

In detail, ARM can be decomposed into two targets: first, with the feature subspace learned by ARM on weak modality, we can predict a instance only with the raw features from weak modality, and this requires extracting the most discriminative feature subspace in weak modality. Conventional feature learning methods often require lots of labeled data to supervise the learning process, while in multi-modal learning scenario, the volume of data is usually larger than that of single modal learning, nevertheless, the labeled examples are very limited. Therefore, in order to extract more discriminative feature from the weak modality, ARM needs to make full use of information from strong modality. Besides, it is also expected ARM can achieve even better performance when the weak as well as the strong modal data are provided in the test phase, and this requires the consistency of predictions built on the strong and weak modality. Thus, we can formulate the ARM model as:

$$\begin{aligned} \arg\min_{F_{v_2},\omega} \quad & \|F_{v_2}\|_2^2 + \lambda_1 F_{v_2}^\top L_{v_1} F_{v_2} + \lambda_2 \|\tilde{X}_{v_1}^\top \omega - Y\|_F^2 \\ s.t. \quad & y_i f_{v_2}(x_{i,v_2}) \geq 1, \qquad \forall i \in \{1, \ldots, N_1\}, \end{aligned} \tag{1}$$

where $F_{v_2} = \{f_{v_2}(x_{1,v_2}), f_{v_2}(x_{2,v_2}), \ldots, f_{v_2}(x_{N,v_2})\} \in \mathcal{R}^N$, $f_{v_2}(x_{i,v_2})$ is the strong modal prediction value for x_{i,v_2}. L_{v_1} is the Laplacian matrix on weak modality. $\tilde{X}_{v_1} = \{x_{i,v_1}, i = 1, \cdots, N_1\}$ is the matrix of weak modal labeled examples. The label information is encoded in the vector $Y = [y_1, y_2, \ldots, y_{N_1}] \in \mathcal{R}^{N_1}$, where $y_i = 1$ if x_i belongs to positive class, and $y_i = 0$ is negative. ω is the feature extraction matrix for weak modality. λ_1 and λ_2 are the balance factors, and constraints $y_i f_{v2} \geq 1$ is a hard margin for strong modal labeled examples.

The $\|F_{v_2}\|_2^2$ is the structure risk of strong modal predictor f_{v_2} in the function space, and $\|\tilde{X}_{v_1}^\top \omega - Y\|_F^2$ acts as linear dimensionality reduction on weak modality [Ye, 2007].

The main targets of ARM are closely related to the 2nd term of Eq. 1. In this term, L_{v_1} is defined as following:

$$L_{v_1} = D_{v_1} - W_{v_1}, \tag{2}$$

where without any loss of generality, we represent the weak modal instance matrix $X_{v_1} \in \mathcal{R}^{d_1 \times N}$ as $X_{v_1}^\top \omega$ in the discriminant feature space, which is expected to extract the most informative features for classification, and W_{v_1} is similarity matrix of weak modal instances, denoted by the inner product of instances in the discriminant feature space: $W_{v_1} = \langle X_{v_1}^\top \omega, X_{v_1}^\top \omega \rangle$. D_{v_1} is the diagonal matrix induced from the W_{v_1}, $D_{v_1,i}$ is the diagonal entry. $D_{v_1,i} = \sum_j W_{v_1,i,j}$, where $W_{v_1,i,j}$ is the (i,j) element of matrix W_{v_1}. When ω is provided, the 2nd term is equivalent to:

$$\lambda_1 \sum_{i,j} (f_{v_2}(x_{i,v_2}) - f_{v_2}(x_{j,v_2}))^2 W_{v_1,i,j}, \tag{3}$$

which reveals that similar weak modal instances in the projected feature space should have similar predictions with strong modal predictor.

By reciting the 2nd term of Eq. 1, it can be found that when f_{v_2} is provided or obtained with high accuracy, we can get a better projection matrix ω by minimizing the 2nd term. In this way, the strong modal information acts like a supervisor, to guide the weak modal feature extraction. The labels of weak modal instances should also be considered, and thus we introduce the 3rd term which is a variant of linear discriminant analysis [Ye, 2007]. On the other hand, when a better feature space for weak modality is provided, we can trust the similarities calculated on the weak modal feature space, and then the 2nd term of Eq. 1 actually becomes a manifold regularizer for the strong modal predictor according to [Belkin *et al.*, 2006].

As a matter of fact, we not only encounter with binary problems in many practical applications, and the ARM formulation can be easily extended into multi-class case where we have c classes:

$$\begin{aligned} \underset{F_{v_2}^c, \omega}{\arg\min} \quad & \|F_{v_2}^c\|_2^2 + \lambda_1 \operatorname{tr}\left(F_{v_2}^{c\top} L_{v_1} F_{v_2}^c\right) \\ & + \lambda_2 \|\tilde{X}_{v_1}^\top \omega - (YY^\top)^{-\frac{1}{2}} Y\|_F^2 \\ s.t. \quad & y_i^c \circ f_{v2}^c(x_{i,v_2}) \geq \mathbf{1}, \ \ \forall i \in \{1, \ldots, N_1\}, \end{aligned} \tag{4}$$

where $F_{v_2}^c = \left\{ f_{v_2}^c(x_{1,v_2}) \ldots f_{v_2}^c(x_{N,v_2}) \right\} \in \mathcal{R}^{N \times c}$, $y_i^c \in \mathcal{R}^c$ is a multi-class label vector, $y_i^c(j) = 1$ indicates the i-th instance belongs to the j-th class, $\circ$ is element wise product operator, $\mathbf{1}$ is the all one vector, $\omega \in \mathcal{R}^{d_1 \times c}$, $\operatorname{tr}(\cdot)$ is the trace operator. The constraints $y_i^c \circ F_{v2}^c \geq \mathbf{1}$ are still hard margins, since strong modal examples should be classified correctly. While herein Eq. 4, the normalized label matrix can be represented by $(YY^\top)^{-\frac{1}{2}} Y$ [Ye, 2007]. Inspired by [Melacci and Belkin, 2011], the strong modal predictor can be denoted as:

$$F_{v2}^c(x_{i,v_2}) = \sum_j^N \alpha_j^{c\top} K(x_{j,v_2}, x_{i,v_2}) + \mathbf{b}$$

where $\alpha = \{\alpha_j^c, \ j = 1, \cdots, N\} \in \mathcal{R}^{N \times c}$, $K(\cdot, x_{i,v_2})$ is the i-th column of strong modal kernel matrix for both labeled and unlabeled data, $\alpha_j^c \in \mathcal{R}^c$. The Eq. 4 is reformed as:

$$\begin{aligned} \underset{F_{v_2}^c, \omega}{\arg\min} \quad & \operatorname{tr}\left(\alpha^\top K \alpha\right) + \lambda_1 \operatorname{tr}\left(\alpha^\top K L_{v_1} K \alpha\right) \\ & + \lambda_2 \|\tilde{X}_{v_1}^\top \omega - (YY^\top)^{-\frac{1}{2}} Y\|_F^2 \\ s.t. \quad & y_i^c \circ (\alpha^\top K(\cdot, x_{i,v_2}) + \mathbf{b}) \geq \mathbf{1}, \ \forall i \in \{1, \ldots, N_1\}. \end{aligned} \tag{5}$$

4 Algorithm

The 2nd term in Eq. 5 involves the product of the weak modal feature extraction matrix ω and the strong modal predictor α, which makes the formulation not joint convex. Consequently, the formulation cannot be optimized easily. The alternative descent algorithm is utilized for solving this problem, and we will reveal the physical meanings of each step in corresponding subsection. Specifically, we first optimize the objective function respect to ω when α is fixed, then optimize α while making ω fixed. We provide the optimization process below:

Fix α, Update ω

We define $\widehat{F}_{i,j} = \|(f_{v_2}^c(x_{i,v_2}) - f_{v_2}^c(x_{j,v_2}))\|_2^2$, and $M = X_{v_1} \widehat{F} X_{v_1}^\top$. According to [Ye, 2007], the 3rd term of Eq. 5 can be written as LDA. When α is fixed, the term $\operatorname{tr}(\alpha^\top K \alpha)$ is not related to ω, thus Eq. 5 can be equivalently written as:

$$\begin{aligned} \underset{\omega}{\arg\min} \quad & \lambda_1 \operatorname{tr}(\omega^\top M \omega) + \lambda_2 \operatorname{tr}(\omega^\top S_W \omega) \\ s.t. \quad & \omega^\top S_B \omega = I, \end{aligned} \tag{6}$$

where S_W is the within-class variance, and S_B is the between-class variance.

However, $\widehat{F}$ may not always be positive definite and $\widehat{F}$ is not guaranteed to be reversible. To overcome this, we define $\widetilde{F}$ to be the inner product of strong modal instances, which can be represented by $\widetilde{F}_{i,j} = f_{v_2}^c(x_{i,v_2})^\top f_{v_2}^c(x_{j,v_2})$ and define $\widetilde{M} = X_{v_1} \widetilde{F} X_{v_1}^\top$, and yield the main target function, which aims to get a better feature projection with the help of strong modality:

$$\begin{aligned} \underset{\omega}{\arg\max} \quad & \lambda_1 \operatorname{tr}(\omega^\top \widetilde{M} \omega) + \lambda_2 \operatorname{tr}(\omega^\top S_B \omega) \\ s.t. \quad & \omega^\top S_W \omega = I. \end{aligned} \tag{7}$$

Eq. 7 has closed-form solution for ω, or ω can be obtained by the generalized eigenvalue of $(\lambda_1 \widetilde{M} + \lambda_2 S_B)\omega = \lambda S_W \omega$.

From the aspect of dimensionality reduction or feature extraction, we treat the extra strong modal information as supervision to help learning the discriminative projection matrix ω for weak modality, i.e., we consider the predictor f_{v_2} as pseudo labels and make kernel alignments between the strong modal predictions and the weak modality in the projected feature space in this step.

Fix ω, Update α

When ω is fixed, note that the 3rd term in Eq. 5 is not related to the predictor f_{v_2}. So Eq. 5 can be addressed as the following sub-problems:

$$\begin{aligned} \underset{\alpha}{\arg\min} \quad & \operatorname{tr}\left(\alpha^\top K \alpha\right) + \lambda_1 \operatorname{tr}\left(\alpha^\top K L_{v_1} K \alpha\right) \\ s.t. \quad & y_i^c \circ (\alpha^\top K(\cdot, x_{i,v_2}) + \mathbf{b}) \geq \mathbf{1}, \ \forall i \in \{1, \ldots, N_1\}, \end{aligned} \tag{8}$$

where we can solve it via efficient quadratic programming(QP) method.

When ω is fixed, in the 2nd term of Eq. 8, the strong modal predictor α treats the $KL_{v_1}K$, which is a hybrid of K and Laplacian matrix L_{v_1} as a regularizer, and we expect the Laplacian regularizer defined on weak modality can boost the performance of strong modality.

Above procedures are repeated iteratively until convergence, the ARM algorithm is shown in Algorithm 1. The algorithm updates the parameters, and experiments show the objective value can be decreased gradually.

Prediction

In the prediction procedure, from Algorithm 1, it can be found that we can predict either with the strong modal predictor α or by *kNN* with the weak modal linear projection matrix ω. In detail, for strong modality, we can use

$$f^c_{v_2}(x_{i,v_2}) = \sum_j^N \alpha^c_j K(x_{j,v_2}, x_{i,v_2}) + \mathbf{b}$$

for predicting x_{i,v_2}. For weak modality, we first use the ω to extract the latent feature representation of both training data and testing data on weak modality, then use *kNN* for classification on weak modality, since most feature extraction methods employ *kNN* as the classifier.

Algorithm 1 The ARM method

Require: X^l_{v1},X^u_{v1},X^l_{v2},X^u_{v2},λ_1, λ_2, Y;
1: Initialize $\omega^0 \leftarrow I$;
2: Initialize $\alpha^0 \leftarrow$ Eq. 8 with fixed $\omega = I$;
3: **while** true **do**
4: $\quad \mathrm{Func}^t_{\mathrm{obj}} \leftarrow$ calculate obj. value in Eq. 5 with α^t, ω^t;
5: $\quad$ Fix α^t,update $\omega^{t+1} \leftarrow$ Eq. 7;
6: $\quad$ Fix ω^{t+1},update $\alpha^{t+1} \leftarrow$ Eq. 8;
7: $\quad \mathrm{Func}^{t+1}_{\mathrm{obj}} \leftarrow$ calculate obj. in Eq. 5 with $\alpha^{t+1}, \omega^{t+1}$;
8: $\quad$ **if** $|\mathrm{Func}^{t+1}_{\mathrm{obj}} - \mathrm{Func}^t_{\mathrm{obj}}| \leq \mathrm{threshold}$ **then**
9: $\quad\quad$ Break;
10: $\quad$ **end if**
11: **end while**
12: **return** α, ω;

5 Experiment

In this section, we first introduce the datasets in brief and then give the empirical results of ARM and compared methods.

Datasets and Configurations

ARM can be adopted for many applications with multi-modal features. In this paper, we use the datasets from image categorization , webpage classification and the biometric tasks in our empirical investigations.

For image categorization, two real-world image datasets are used, i.e., *Nus* [Chua *et al.*, 2009] and *Msra* [Wang *et al.*, 2009]. *Nus* subset contains 9,109 images of 10 categories, and has 6 groups of features extracted. *Msra* subset contains 10,680 images of 9 categories, and has 7 groups of features are extracted. We partition all the features into strong modal features and weak modal features. More specifically, in *Nus*, color histogram features are weak modality while the rest are strong modal features. Examples from Professional and Movie are selected for balanced binary classification; In *Msra*, HSV color histogram are weak modal features and the rest are strong modal features, and 2 balanced categories are selected for classification as well.

The *WebKB* dataset contains webpages collected from 4 universities: Wisconsin, Washington, Cornell and Texas. These webpages are about five categories, i.e., student, project, course, stuff and faculty, and described with two modalities: the *content* and the *citation*. We consider the *content* as weak modality and the *citation* as strong modality. Five subsets are constructed from *WebKB*, i.e., *Winsconsin*; *Washington*; *Cornell* and *Texas* (denoted as *Wins.*, *Wash.*, *Corn.* and *Texas* in tables), which use the instances from each single university, and *WebKB* that combines all universities data is also tested in our experiments. The first 4 datasets are for binary classification, which aim to tell the differences between student and stuff vs. the rest categories. While *WebKB* is a multi-class classification task, which aims to identify the differences between each university.

In Biometric task, we construct the virtual faces and gaits dataset, in which faces are strong modality and gaits are weak modality. The CMU PIE is used for face modality construction and the gaits dataset is collected by SDUML [Zhang *et al.*, 2014]. PIE contains more than 750,000 images of 337 people. In the constructed faces and gaits dataset, 25 virtual users are picked out by assigning the same identity to the users from PIE and SDUML. Gait sequences contain 25 users with 40 gait sequences per user. For each gait sequences, we choose the first 50 frames in each video to represent the gait sequence. For each of the 25 virtual users, 150 frames are randomly chosen and the contour features of gait sequences are extracted; for face modality, 150 images are randomly selected per person. This dataset is denoted as *Biom.* in tables.

For all datasets in our experiments, we randomly select 66% instances for training, and the remains are used for testing. The labeled ratio is set to 30% according to [Zhang *et al.*, 2014]. We repeat this for 30 times, the average accuracy and std. of predictions are recorded as classification performance. The parameter λ_1 and λ_2 in the training phase is tuned in $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. Empirically, when the variations between the objective value of Eq. 5 is less than 10^{-5} in iteration, we treat ARM converges. HIK kernel is used in ARM.

Comparing with Multi-Modal Learning Methods

ARM is essentially a multi-modal learning algorithm. It should be compared with the state of the art multi-modal learning methods. In our empirical investigations, 4 multi-modal learning methods are compared:

KCCA (kernel CCA) [Hardoon *et al.*, 2004] is first used to extract the latent feature representation of both the strong modality and weak modality, and then *kNN* is performed for the final classification;

Co-Training is a famous multi-modal semi-supervised learning method, and co-training usually employs *Naive Bayes* as the base learner;

CoTRADE As a Co-Training style algorithms, CoTRADE

Table 1: The accuracy (avg.±std.) of compared multi-modal methods on weak and strong modality. For ARM and KCCA the performance of weak and strong modality are listed separately. The best classification performance of strong modality is bolded while that of weak modality is marked with black dots.

	ARM		KCCA		Co-Training	CoTRADE	Co-Regularization
	Strong Modality	Weak Modality	Strong Modality	Weak Modality			
Nus	**.898±.012**	.756±.012	.762±.016	.694±.020	.725±.016	.728±.016	.788±0.013
Msra	**.936±.006**	.814±.007•	.825±.056	.740±.011	.592±.015	.590±.019	.769±0.003
Wins.	**.847±.057**	.725±.043•	.694±.030	.625±.070	.653±.045	.662±.049	.657±0.042
Wash.	**.895±.041**	.733±.056	.764±.046	.672±.087	.658±.114	.685±.098	.744±0.045
Corn.	**.863±.049**	.696±.040•	.699±.030	.628±.113	.642±.091	.653±.069	.688±0.004
Texas	**.845±.052**	.781±.049•	.725±.017	.758±.058	.667±.058	.702±.058	.721±0.007
WebKB	**.644±.029**	.494±.041	.545±.023	.446±.049	.420±.067	N/A	.546±.026
Biom.	**.973±.005**	.683±.007•	.848±.008	.566±.005	.519±.042	N/A	.668±.020

chooses the predictions with authentic high confidence for labeling information communication based on a particular designed data editing techniques;

Co-Regularization In Co-Regularization [Sindhwani and Niyogi, 2005], it aims to reduce the divergence of two predictors on different modalities. We follow [Sindhwani and Niyogi, 2005] and use the Laplacian SVM as the classifier.

Table 1 records the prediction accuracies (avg.± std.) of weak/strong modalities of the ARM and compared methods. ARM is tested with a kernel classifier f_{v_2} on strong modality as well as *kNN* on weak modality. As a consequence, we can either directly use the kernel classifier or employ *kNN* on the reduced dimensions(feature spaces) as the classification method during the test phase. The former result is denoted by 'strong modality' in Table 1, while the latter is denoted by 'weak modality' in the same table.

Table 1 clearly reveals that on all 6 binary classification datasets, the ARM average accuracies of 'strong modality' are the best. While comparing to KCCA (weak modality), Co-Training, CoTRADE, Co-Regularization, on *Msra*, *Winsconsin*, *Cornell* and *Texas*, i.e., 4 of the 6 binary classification datasets, the 'weak modality' performance of ARM are also the best. On multi-class classification, the average accuracies of ARM (on both the weak modality and strong modality) are the best among all of the compared methods except for the weak modality performance on *WebKB*.

Comparing with Dimensionality Reduction Methods on Weak Modality

ARM employs the information from strong modality to help extract a better feature space of weak modality, thus, it is closely related to dimensionality reduction from the aspect of weak modality. To validate the effectiveness of utilizing strong modality for weak modal feature extraction. We compare ARM with dimensionality reduction methods, i.e.,

Linear dimensionality reduction methods such as LPP, LDA, PCA, CSFS [Chang *et al.*, 2014] are used to extract the reduced subspace, and then *kNN* is performed for obtaining the final classification result. Table 2 records the weak modal performance of ARM and compared methods. It clearly reveals that on 6 binary classification datasets and 2 multi-class datasets, the performance of ARM are the best.

Table 2: The accuracy (avg.±std.) of ARM (weak modality) compare with other dimensionality reduction methods on classification tasks. The best classification performance on each dataset is bolded.

	ARM	LPP	LDA	PCA	CSFS
Nus	**.756±.012**	.691±.012	.690±.012	.578±.016	.742±.014
Msra	**.814±.007**	.752±.012	.752±.013	.672±.011	.799±.008
Wins.	**.725±.043**	.653±.072	.653±.073	.582±.071	.659±.065
Wash.	**.733±.056**	.701±.073	.701±.073	.662±.046	.732±.050
Corn.	**.696±.040**	.603±.130	.600±.127	.573±.108	.605±.121
Texas	**.781±.049**	.754±.049	.754±.049	.683±.076	.748±.061
WebKB	**.494±.041**	.425±.028	.425±.028	.407±.028	.475±.043
Biom.	**.683±.007**	.664±.015	.644±.015	.492±.013	.634±.011

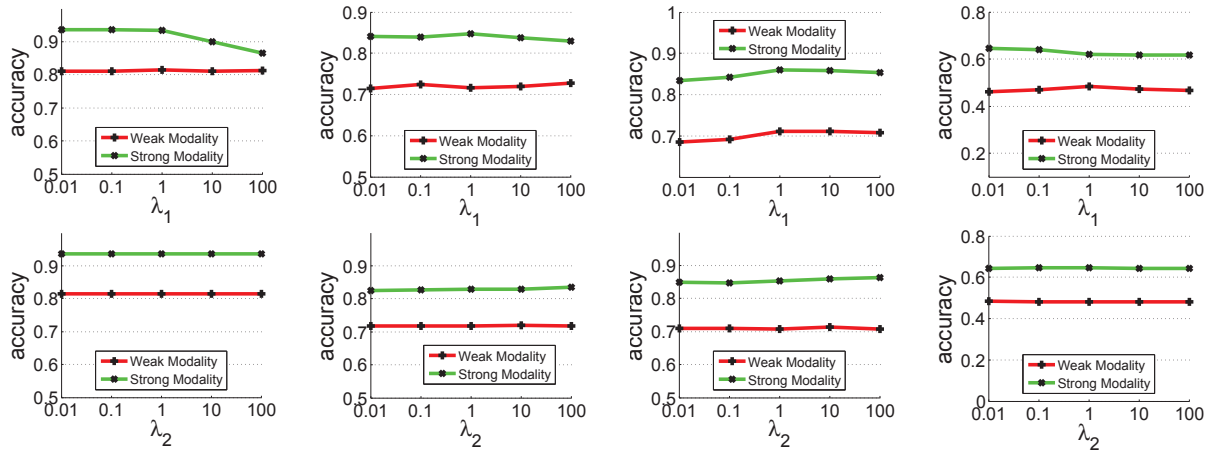
Table 3: The accuracy (avg.±std.) of ARM (strong modality) compare with kernel methods on classification tasks. The best classification performance on each dataset is bolded.

	ARM(Strong)	LibSVM_1	LibSVM_2	LapSVM
Nus	**.898±.012**	.895±.013	.837±.031	.890±.010
Msra	**.936±.006**	.933±.004	.769±.000	.932±.003
Wins.	.847±.057	**.869±.045**	.781±.111	.798±.037
Wash.	**.895±.041**	.865±.032	.785±.105	.781±.050
Corn.	**.863±.049**	.859±.034	.801±.082	.689±.004
Texas	.845±.052	**.860±.038**	.812±.067	.729±.008
WebKB	**.644±.029**	.618±.025	.485±.000	.606±.033
Biom.	**.973±.005**	.914±.000	.960±.000	.800±.000

Comparing with Kernel Methods on Strong Modality

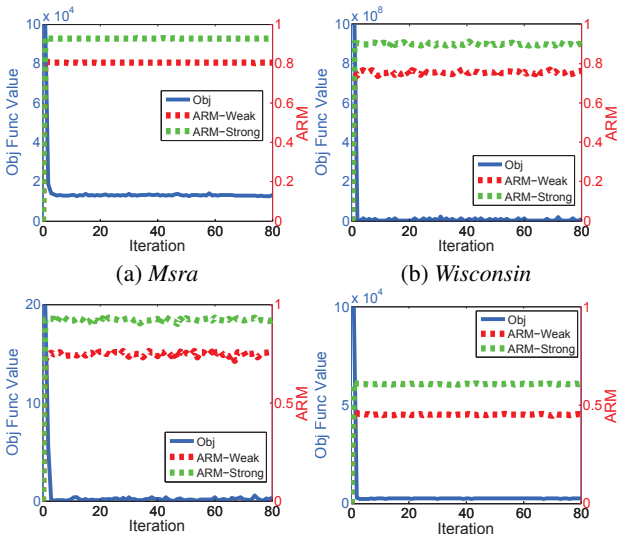
ARM trains a kernel classifier on strong modality. As a matter of fact, if the test examples are only with strong modality, we can employ the kernel classifier trained by ARM to predict the instances. As a consequence, ARM should be compared with the kernel methods. In our empirical study, 3 kernel methods are compared, i.e.,

Kernel methods, such as LibSVM [Chang and Lin, 2011], LapSVM [Belkin *et al.*, 2006], are compared with ARM in additional experiments. It is notable that here LibSVM_1 rep-



(a) *MSRA* (b) *Wisconsin* (c) *Cornell* (d) *WebKB*

Figure 1: Influence of the parameters λ_1, λ_2 on the 4 datasets with labeled data ratio at 30%



(a) *Msra* (b) *Wisconsin* (c) *Cornell* (d) *WebKB*

Figure 2: Objective function value convergence and corresponding classification accuracy vs. number of iterations of ARM with labeled data ratio at 30%

resents LibSVM with HIK kernel and LibSVM$_2$ represents LibSVM with polynomial kernel. Table 3 records the strong modal performance of the ARM and compared methods. It clearly reveals that on *Msra*, *Washington*, *Cornell* and *Texas*, i.e., 4 binary datasets and all multi-class datasets, ARM has achieved the best performance.

Investigation on Stability of Parameter

In order to explore the influence of parameters λ_1 and λ_2, more experiments are conducted. We first fix the λ_1 while tuning λ_2 in $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and record the average accuracy in the first row of Fig. 1, then we fix the λ_2 while tuning λ_1 in $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and get the average accuracy in the second row of Fig. 1. Due to the page limits, we only list 4 datasets for verification, i.e., *Msra*, *Winsconsin*, *Cornell* and *WebKB*. From these figures, we can find that ARM achieves a stable performance on each dataset, which indicates the insensitivity of ARM to parameters.

Empirical Investigation on Convergence

To investigate the convergence of ARM iterations empirically. The objective function value, i.e., the value of Eq. 5 and the classification performance of ARM in each iteration are recorded. Due to the page limits, results on only 4 dataset are plotted in Fig. 2. It clearly reveals that the objective function value decreases as the iterations increase, and the classification performance is stable after several iterations on different datasets in Fig. 2. Moreover, these additional experiments result indicates that our ARM can converge very fast, i.e., on most datasets, ARM converges after 3 rounds.

6 Conclusion

In this paper, we focus on multi-modal classification and present a novel method ARM. We first analyze the phenomenon of the unsatisfied classification performance on weak modality and attribute the reasons into lack of information or disturbance of informative weak modal features. We claim that different modalities should be treated with different strategies, and consequently proposed the ARM approach. ARM can perform feature extraction on weak modal features by treating the auxiliary information from strong modality as supervision, meanwhile it can also improve the classification performance on strong modality by organizing the weak modal features as a regularizer. Empirical results on real world datasets clearly validate the effectiveness of ARM, and show that ARM can extract the most discriminative feature subspace on weak modality while successfully regularize the strong modal predictor at the same time. In the current setting of ARM, only two modalities are considered. How to distinguish more than two modalities by classification capacities and integrate multiple weak or strong modalities in the ARM framework can be an interesting future work.

References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI., 1998.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):Article No. 27, 2011.

[Chang *et al.*, 2014] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *Proceedings of the 24th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 1171–1177, Atlanta, Georgia, 2014.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page Article No.48, Santorini, Greece, 2009.

[Fisher, 1936] Ronald A Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of eugenics*, 7(2):179–188, 1936.

[Guo and Xue, 2013] Yuhong Guo and Wei Xue. Probabilistic Multi-label Classification with Sparse Feature Learning. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1373–1379, Beijing, China, 2013.

[Hardoon *et al.*, 2004] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

[He and Partha, 2003] Xiaofei He and Niyogi Partha. Locality Preserving Projections. In *Proceedings of the 17th Conference Advances in Neural Information Processing Systems*, pages 153–160, Whistler, Canada, 2003.

[Hong and Jain, 1998] Lin Hong and Anil Jain. Integrating Faces and Fingerprints for Personal Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1295–1307, 1998.

[Jolliffe, 2005] Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, Hoboken, 2005.

[Kiritchenko and Matwin, 2011] Svetlana Kiritchenko and Stan Matwin. Email Classification with Co-Training. In *Proceedings of the 2001 Conference of the Center for Advanced Studies on Collaborative Research*, pages 301–312, Toronto, Canada, 2011.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. Multimodal Neural Language Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 595–603, Beijing, China, 2014.

[Marcialis *et al.*, 2010] Gian Luca Marcialis, Paolo Mastinu, and Fabio Roli. Serial Fusion of Multi-modal Biometric Systems. In *Proceedings of the Workshop on Biometric Measurements and Systems for Security and Medical Applications*, pages 1–7, Taranto, Italy, 2010.

[Melacci and Belkin, 2011] Stefano Melacci and Mikhail Belkin. Laplacian Support Vector Machines Trained in The Primal. *The Journal of Machine Learning Research*, 12:1149–1184, 2011.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, Bellevue, Washington, 2011.

[Nguyen *et al.*, 2013] Cam-Tu Nguyen, De-Chuan Zhan, and Zhi-Hua Zhou. Multi-modal Image Annotation with Multi-Instance Multi-Label LDA. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1558–1564, Beijing, China, 2013.

[Sindhwani and Niyogi, 2005] Vikas Sindhwani and Partha Niyogi. A Co-Regularized Approach to Semi-supervised Learning with Multiple Views. In *Proceedings of Workshop on Learning with Multiple Views joint with the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[Wang *et al.*, 2009] Meng Wang, Linjun Yang, and Xian-Sheng Hua. MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval. Technical report, Microsoft Research Asia, Microsoft, 2009.

[Xie *et al.*, 2014] Wenxuan Xie, Yuxin Peng, and Jianguo Xiao. Cross-View Feature Learning for Scalable Social Image Analysis. In *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 201–207, Quebec, Canada, 2014.

[Xu *et al.*, 2010] Yong Xu, David Zhang, and Jing-Yu Yang. A Feature Extraction Method for Use with Bimodal Biometrics. *Pattern Recognition*, 43(3):1106–1115, 2010.

[Ye, 2007] Jieping Ye. Least Squares Linear Discriminant Analysis. In *Proceedings of the 24th International Conference on Machine learning*, pages 1087–1093, Corvallis, OR., 2007.

[Zhai *et al.*, 2013] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, and Wen Gao. Parametric Local Multimodal Hashing for Cross-View Similarity Search. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2754–2760, Beijing, China, 2013.

[Zhang and Li, 2014] Dongqing Zhang and Wu-Jun Li. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization. In *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 2177–2183, Quebec, Canada, 2014.

[Zhang and Zhou, 2011] Min-Ling Zhang and Zhi-Hua Zhou. CoTrade: Confident Co-Training With Data Editing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(6):1612–1626, 2011.

[Zhang *et al.*, 2014] Qing Zhang, Yilong Yin, De-Chuan Zhan, and Jingliang Peng. A Novel Serial Multimodal Biometrics Framework Based on Semi-Supervised Learning Techniques. *IEEE Transactions on Information Forensics and Security*, 9(10):1681–1694, 2014.

[Zhou *et al.*, 2005] Xiaoli Zhou, Bir Bhanu, and Ju Han. Human Recognition at a Distance in Video by Integrating Face Profile and Gait. *IEEE Transactions on Face Biometrics for Personal Identification*, 3546(5):533–543, 2005.