Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Semi-Supervised Multi-Modal Learning with Incomplete Modalities

Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, Yuan Jiang
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{yangy, zhandc, shengxr, jiangy}@lamda.nju.edu.cn

Abstract

In real world applications, data are often with multiple modalities. Researchers proposed the multi-modal learning approaches for integrating the information from different modalities. Most of the previous multi-modal methods assume that training examples are with complete modalities. However, due to the failures of data collection, self-deficiencies and other various reasons, multi-modal examples are usually with incomplete feature representation in real applications. In this paper, the incomplete feature representation issues in multi-modal learning are named as incomplete modalities, and we propose a semi-supervised multi-modal learning method aimed at this incomplete modal issue (SLIM). SLIM can utilize the extrinsic information from unlabeled data against the insufficiencies brought by the incomplete modal issues in a semi-supervised scenario. Besides, the proposed SLIM forms the problem into a unified framework which can be treated as a classifier or clustering learner, and integrates the intrinsic consistencies and extrinsic unlabeled information. As SLIM can extract the most discriminative predictors for each modality, experiments on 15 real world multi-modal datasets validate the effectiveness of our method.

1 Introduction

Multi-modal learning becomes attractive as the development of data collection technicals, and can be widely used in applications with relatively independent data sources, e.g., biological data with gene expression, array-comparative genomic hybridization, single-nucleotide polymorphism and methylation. Multi-modal learning approaches can utilize information from multiple modalities, among which information from each single modality can complement each other to improve the generalization ability of the whole learner, e.g., clustering [Arora *et al.*, 2016; Iwata and Yamada, 2016], classification [Yang *et al.*, 2015; 2016], regression [Yang *et al.*, 2017]. It is notable that the mainstream multi-modal learning approaches assume that training examples are with complete modalities.

Nevertheless, the assumption mentioned above is excessive, since there are many reasons for insufficiencies or incompleteness, including data collection failures from the damage of data sensors, data corruption by network communication, data privacy policies, etc., e.g., in web pages classification with document/image representations, documents and images are two modalities, yet some web pages only have document or image information; for user identification in cross-network, the user profile features, content information or linkage information can be regarded as multiple modalities, yet some users only have one or partial modalities due to personal preference or privacy issues. Existing multi-modal learning approaches cannot directly be applied on the incomplete modal situation. With some straight forward strategies, e.g., removing the examples only with partial modal features, filling in the incomplete modal features with missing data techniques, current multi-modal learning approaches can be executed, yet the model trained will clearly loses information and introduces extra noises.

Aiming at the incomplete modal issues, there are some preliminary investigations. Trivedi *et al.* [2010] proposed a partial modal approach, which uses one modal kernel matrix as the similarity matrix and completes the missing modal kernel using Laplacian regularization; Shao *et al.* [2016] proposed an online multi-modal clustering algorithm OMVC to learn the latent feature matrices for each individual incomplete modality and pushes them towards a common consensus; Zhao *et al.* [2016] proposed an unsupervised method which well handles the incomplete multi-modal data by transforming the original and incomplete data to a new and complete representation in a latent space. These methods mainly focus on making full use of the *inherent information*, i.e., the consistencies between multiple modalities. In this paper, we consider the defects of insufficient information caused by the incompleteness among modalities should be remedied by supplementing *extrinsic information*.

Transductive multi-modal learning methods are proposed for utilizing extrinsic information from test sets, e.g., Karasuyama and Mamitsuka [2013] proposed a new method SMGI, integrating multiple graphs for label propagation, which appeals to the sparsity of graph weights and can easily eliminate irrelevant graphs; Cai *et al.* [2013] proposed a novel approach to integrate heterogeneous features by performing multi-modal semi-supervised classification on unla-

beled as well as unsegmented instances, which learns a common shared class indicator matrix and weights for different modalities. However, these transductive methods are difficult to extend to classification under the incomplete modal setting with unseen test data.

Different to above solutions, we propose a novel Semi-supervised multi-modal Learning approach with the Incomplete Modal data (SLIM). SLIM utilizes the intrinsic modal consistencies and extrinsic unlabeled information in one unified framework as well as can perform in both transductive and inductive configurations. Consequently, SLIM is with wider applicable range and can be applied in both classification and clustering tasks. Besides, SLIM can also learn the most discriminative classifiers for each modality separately. Meanwhile, noting that real world datasets are always with noise and outliers resulting in unreliable solution, the square-root loss is used for robust weight learning for each modality. Finally, more discriminative classifiers and robust clustering performance can be achieved in SLIM. We empirically investigate the effectiveness of SLIM, and it achieves significantly better performance on various tasks.

In the following of this paper, we start with a brief review of related works. Then the proposed SLIM approach and the experimental results. After that, we conclude the paper.

2 Related Work

The exploitation of multiple modal learning has attracted much attention recently. In this paper, our method integrates the intrinsic consistencies and extrinsic unlabeled information in a semi-supervised scenario with incomplete multiple modal data, which can be treated as a classifier or clustering learner. Therefore, we consider our work related to multi-modal learning and semi-supervised learning.

Most of the previous multi-modal methods assume that training examples are with complete modalities. However, multi-modal examples are usually with incomplete feature representation in real applications. Therefore, many researchers have devoted to handling the incomplete modal data recently. Li *et al.* [2014] established a latent subspace where the instances corresponding to the same example in different modalities are close to each other, and similar instances in the same modality should be well grouped; Shao *et al.* [2015] proposed the MIC (Multi-Incomplete-view Clustering), an algorithm based on weighted nonnegative matrix factorization with $L_{2,1}$ regularization, which learns the latent feature matrices for all the modalities and generating a consensus matrix so that minimize the difference between each modality and the consensus matrix; Xu *et al.* [2015] proposed an effective algorithm to accomplish multi-modal learning with incomplete modalities by assuming that different modalities are generated from a shared subspace, which exploits the connections between multiple modalities, enabling the incomplete modalities to be restored with the help of the complete modalities. However, these methods mainly focus on the *inherent information*, i.e., the consistencies between multiple modalities or the data structures among multiple modalities. In this paper, we consider that the defects of insufficient information caused by the incompleteness among modalities should be remedied by supplementing *extrinsic information* instead.

Transductive multi-modal learning, as a matter of fact, utilizes the extrinsic information from test sets. Eaton *et al.* [2010] proposed a constrained clustering that can operate with an incomplete mapping, which propagates given pairwise constraints using a local similarity measure to those instances that can be mapped to other modalities; Yin *et al.* [2017] proposed a novel subspace learning framework for incomplete and unlabeled multi-modal data, which directly optimizes the class indicator matrix, the inter-modal and intra-modal data similarities are preserved to enhance the model. These multi-modal learning approaches with incomplete modal information partially incorporate with the semi-supervised learning techniques to relax the issues introduced by modality incompleteness. However, these approaches are under the configuration of transductive learning and are difficult to extend on unseen test data.

In this paper, we propose a novel approach named Semi-supervised multi-modal Learning with the Incomplete Modal information (SLIM), which utilizes the intrinsic modal consistencies and extrinsic unlabeled information in one unified framework, and can perform under both transductive and inductive configurations. As a consequence, SLIM can be applied in both classification and clustering tasks. Besides, considering that different modalities have various noise levels, we utilize the square-root loss rather than learning the weight for each modality. Finally, more discriminative classifiers and robust clustering performance can be achieved in SLIM.

3 Proposed Method

In our incomplete multiple modal learning setting, an instance is characterized by multiple modal features while only with one unified label. Suppose we are given a dataset with N examples and K modalities. The $i-$th instance $\mathbf{x}_i$ of $k-$th modality can be represented as $\mathbf{x}_{i^k} \in \mathbb{R}^{d_k}$, where d_k is the dimension of the $k-$th modality. Each instance may has complete or partial modalities as shown in Fig. 1. Without any loss of generality, suppose we have N_c homogeneous examples with complete modal features, meanwhile, we have N_k heterogeneous instances for each modality. Thus, the incomplete modal example set can be represented as $\mathcal{D} = \{\mathbf{X}_c, X_1, X_2, \cdots, X_K\}$, where $\mathbf{X}_c = \{(\mathbf{x}_{i^1}, \mathbf{x}_{i^2}, \cdots, \mathbf{x}_{i^K})\}_{i=1}^{N_c} \in \mathbb{R}^{N_c \times d}$ denotes the examples present in all modalities, $d = d_1 + d_2 + \cdots + d_K$, $X_k \in \mathbb{R}^{N_k \times d_k}$ denotes the incomplete examples present in the $k-$th modality. While under the semi-supervised learning scenario, we assume that there are l labeled examples including complete or incomplete examples. For labeled examples, the label of example $\mathbf{x}_i$ can be represented as $y_i \in \{1, \cdots, C\}$, C is the class number, and the labeled example sets can be represented as Θ_l. The goal of SLIM is to cluster the N examples into their corresponding clusters, while learning discriminative predictors for each modal prediction.

3.1 The Formulation of SLIM

In this section, we will describe the SLIM in detail. In incomplete modal learning, SLIM aims to utilize the intrinsic modal consistencies and extrinsic unlabeled information in

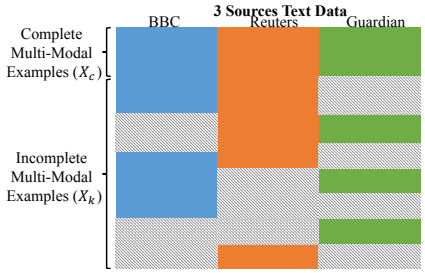


Figure 1: An illustration of the Incomplete Multi-Modal Data in real-world application as 3-Source Text data. 3 Sources are collected from three online news sources: BBC, Reuters, and Guardian (represented as blue, orange, green rectangles), where each news source can be seen as one modality for the news reports. The missing modalities are denoted as clinodiagonal rectangles.

one unified framework, and can perform in both transductive and inductive configurations. Specifically, SLIM can be decomposed into two targets: first, with the incomplete modal examples, we aim to learn the predictors to classify the test instances of each modality accurately. Besides, we wish to cluster the unlabeled instances by modeling a joint transformed matrix factorization problem with respect to each modal similarity matrix and the learned predictions, and pushes them towards a consensus. Thus, SLIM can be defined as:

$$\min_{W_k,b_k,F} \sum_{k=1}^{K} (\hat{L}_k(F_k, F) + \frac{\lambda_2}{2} \tilde{L}_k(\hat{X}_k, F)) \tag{1}$$

There are K modalities, the first term $\hat{L}_k(F_k, F)$ denotes the loss of classification of the $k-$th modality, F_k and F are the classification results and the labels of all the instances on $k-$th modality to be learned. The second term, $\tilde{L}_k(\hat{X}_k, F)$ considers both the intrinsic and extrinsic information for incomplete modalities. More specifically, it models with a joint transformed matrix factorization problem, here $\hat{X}_k \in \mathbb{R}^{N \times d_k}$ is the matrix of the $k-$th modality with missing rows filling with zeros. In detail, we treat the incomplete similarity matrix of each modal and the learned predictions as a transformed matrix factorization problem, and wish to keep the consistency between them, $\lambda_2 > 0$ is a balance parameter.

Specifically, objective function $\hat{L}_k$ on the $k-$th modality in Eq. 1 can be generally represented as the form: $\min_{F_k} \hat{L}_k(F_k, F) = \ell(F_k, F) + \frac{\lambda_1}{2} r(F_k)$. Here $r(F_k)$ is the regularizer for modal-specific classifier. $\frac{\lambda_1}{2}$ is a scalar coefficient to balance the weights of the two terms. Eq. 1 indicates the classifier in $\ell(\cdot)$ and the prediction results F for instances are connected. Without any loss of generality, the loss function $\ell(F_k, F)$ can take any convex forms, and we use square loss and linear classifier here for simplicity:

$$\min_{W_k,b_k} \frac{1}{2\eta_k} \|\hat{X}_k W_k + \mathbf{1}\mathbf{b_k}^\top \odot P_k - F \odot P_k\|_{\mathrm{F}}^2 + \frac{\lambda_1}{2} \|W_k\|_{\mathrm{F}}^2 \tag{2}$$

Where $W_k \in \mathbb{R}^{d_k \times C}$ is the linear classifier, $\mathbf{b_k} \in \mathbb{R}^C$ is the bias for current predictor, $\mathbf{1}$ is the all one vector, $\odot$ represents element wise product operator, $P_k \in \mathbb{R}^{N \times C}$ is the indicator matrix, where $[P_k]_{i,\cdot} = 1$ iff $i-$th instance is complete on $k-$th modality, otherwise $[P_k]_{i,\cdot} = 0$. In multi-class cases, we expand the label y_i for instance $\mathbf{x}_i$ to a vector with C elements, where $y_{i,j} = 1$ indicates the $i-$th instance is with label j, otherwise, $y_{i,j} = 0$, similarly, $F \in \mathbb{R}^{N \times C}$ denotes the predictions of all instances to be learned, η_k is the number of the complete examples of $k-$th modality.

The intrinsic information, i.e., the consistencies between the indicator matrix of different modalities are one of the most prominent information for relief the insufficiencies from modality incompletenesses. Therefore, the consistency loss can be treated as the main component for the 2nd term $\tilde{L}_k$: $\|\mathcal{R}_\Omega(M_k) - \mathcal{R}_\Omega(YY^\top)\|_{\mathrm{F}}^2$, where $M_k \in \mathbb{R}^{N \times N}$ is the similarity matrix of the labeled examples of $k-$th modality. $[\mathcal{R}_\Omega(M_k)]_{i,j} = [M_k]_{i,j}$ iff $i-$th instance and $j-$th instance have complete entries on $k-$th modality, otherwise $[\mathcal{R}_\Omega(M_k)]_{i,j} = 0$, and $\mathcal{R}_\Omega(YY^\top)$ has the same definition, Y denotes the label matrix of the labeled examples.

However, in the semi-supervised scenario, more extrinsic information can be involved for better modeling. In this paper, we treat all examples, both label and unlabeled data, with labels as F, and $\tilde{L}_k$ can be reformulated as:

$$\begin{aligned} &\min_F \frac{1}{\eta_k^2} \|\mathcal{R}_\Omega(M_k) - \mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}^2 \\ &s.t. \quad F^{\Theta_l} = Y, 0 \le F \le 1, \end{aligned} \tag{3}$$

note that here each similarity matrix only has $\eta_k \times \eta_k$ real-valued entries and we fill the rest entries with zeros. The constraint $F^{\Theta_l} = Y$ restricts the prediction on labeled data as same as the ground truth to avoid collapsing of predictions, Θ_l here is the index set of the labeled data. In addition, we constrain the predicted values into the same range as true labels by $0 \le F \le 1$ to maintain the intrinsic consistencies. It is notable that the Eq. 3 closely relates to a kernel Kmeans and laplacian-based spectral clustering in a wild condition [Ding and He, 2005], which implies that the whole approach (consisted with this term) can be also applied in clustering tasks.

In addition, it is also notable that real world data always contain noise and outlying entries that result in the unreliable similarity matrix, which will impair the final performance. Previous multi-modal learning methods usually weight different modalities or instances against the affections introduced by noises for ensuring the robustness. However, in semi-supervised learning scenario, there are only insufficient number of labeled data for weighting parameters tuning, and as a matter of fact, the affections of noises become one of the barriers for modeling robustly. In this paper, we further employ the square-root loss function instead of the least squares function in Eq. 3 to reduce the affections from noisy data. This solution can be regarded as a weighted regularized least squares form of the original one, where the weight for each modality is: $\frac{1}{\eta_k \|\mathcal{R}_\Omega(M_k) - \mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}}$ due to [Liu *et al.*, 2014]. This modification can calibrate each modality by considering the different noise levels of all modalities and increases the robustness of the 2nd term in SLIM:

$$\begin{aligned} &\min_F \frac{1}{\eta_k} \|\mathcal{R}_\Omega(M_k) - \mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}} \\ &s.t. \quad F^{\Theta_l} = Y, 0 \le F \le 1. \end{aligned} \tag{4}$$

Without loss of generality, we can combine Eq. 2 and Eq. 4 in a unified framework and yield the whole SLIM model:

$$\begin{aligned}\min_{W_k,b_k,F}\sum_{k=1}^{K}(\frac{1}{2\eta_k}\|\hat{X}_kW_k+\mathbf{1}\mathbf{b}_k^T\odot P_k-F\odot P_k\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\|W_k\|_{\mathrm{F}}^2)\\+\frac{\lambda_2}{2}\sum_{k=1}^{K}\frac{1}{\eta_k}\|\mathcal{R}_\Omega(M_k)-\mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}\\ s.t.\quad 0\le F\le 1, F^{\Theta_l}=Y\end{aligned} \tag{5}$$

3.2 Solutions

In this section, we mainly focus on the methodology for addressing the optimization of SLIM represented as Eq. 5 which is convex to $W_k, \mathbf{b}_k$ yet not a jointly convex problem. An alternative descent algorithm is considered to be utilized for solving this problem, nevertheless, further derivations successfully show that the alternative descent approach is with closed form solutions for some key parameters.

First, it clearly shows that the optimal solution of $\mathbf{b}_k$ is with closed-form when W_k and F are fixed,

$$\mathbf{b}_k=\frac{1}{\eta_k}(F\odot P_k-\hat{X}_kW_k)^\top\mathbf{1} \tag{6}$$

Substitute Eq. 6 into Eq. 5, we can simplify Eq. 5 as:

$$\begin{aligned}\min_{W_k,F}\sum_{k=1}^{K}(\frac{1}{2\eta_k}\|C_k\hat{X}_kW_k-C_k(F\odot P_k)\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\|W_k\|_{\mathrm{F}}^2)\\+\frac{\lambda_2}{2}\sum_{k=1}^{K}\frac{1}{\eta_k}\|\mathcal{R}_\Omega(M_k)-\mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}\\ s.t.\quad 0\le F\le 1, F^{\Theta_l}=Y,\end{aligned} \tag{7}$$

where $C_k=I-\frac{1}{\eta_k}\mathbf{1}\mathbf{1}^\top\odot P_k$. Then we can find that the W_k is also with closed-form when F is fixed:

$$W_k=A_kB_kC_k(F\odot P_k), \tag{8}$$

where $A_k=(\hat{X}_k^\top C_k^\top C_k\hat{X}_k+\eta_k\lambda_1 I)^{-1}$, $B^k=\hat{X}_k^\top C_k^\top$. Combining Eq. 8 and Eq. 7, we can rewrite the Eq. 7 as:

$$\min_F tr(F^THF)+\lambda_2\sum_{k=1}^{K}\frac{1}{2\eta_k}\|\mathcal{R}_\Omega(M_k)-\mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}, \tag{9}$$

where $tr(\cdot)$ is the matrix trace operator, $H=\sum_{k=1}^{K}\Pi_{\Gamma_k}[C_kC_k^\top B_k^\top A_k^\top(\frac{\lambda_1}{2}A_kB_k+\frac{1}{2\eta_k}B_kB_k^\top A_kB_k-\frac{1}{\eta_k}B_k)+\frac{1}{2\eta_k}C_k^\top C_k]$, where $\Gamma_k=\{\gamma_1,\gamma_2,\cdots,\gamma_{\eta_k}\}$ represents the index set of the complete instances of $k-$th modality. $\Pi_{\Gamma_k}(A)$ represents the rows and columns in Γ_k of matrix A are set as 0. And we can use the project sub-gradient method to optimize Eq. 9 for simplicity.

$$g=\begin{cases}HF, & \bar{L}=0,\\ HF+\lambda_2\sum_{k=1}^{K}\frac{\mathcal{R}_\Omega(FF^\top)-\mathcal{R}_\Omega(M_k)}{\|\mathcal{R}_\Omega(M_k)-\mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}}F, & \text{Otherwise}\end{cases} \tag{10}$$

where $\bar{L}=\|\mathcal{R}_\Omega(M_k)-\mathcal{R}_\Omega(FF^\top)\|_{\mathrm{F}}$.

With the parameters W_k, b_k solved in closed-form, we can solve the F with the projected sub-gradient of Eq. 10.

Datasets	C	N	K	$d_k(k=1,2,\cdots,K)$
Movie	17	617	2	1878, 1398
Citeseer	6	3264	2	3703, 3264
Cora	7	2708	2	1433, 2708
Cornell	5	195	2	1703, 195
Texas	5	185	2	1703, 185
Washington	5	217	2	1703, 217
Wisconsin	5	262	2	1703, 262
News-M2	2	1200	3	2000, 2000, 2000
News-M5	5	500	3	2000, 2000, 2000
News-M10	10	500	3	2000, 2000, 2000
News-NG1	2	500	3	2000, 2000, 2000
News-NG2	5	400	3	2000, 2000, 2000
News-NG3	8	1000	3	2000, 2000, 2000
Reuters	6	1600	5	2000, 2000, 2000, 2000, 2000
3Sources	6	416	3	3560, 3631, 3068

Table 1: Dataset description, datasets with two modalities or multiple modalities are separated with a horizontal line.

Datasets	SLIM	ConvexSub	PVC	MIC
Movie	**.247±.009**	.123±.004	.193±.003	.172±.001
Citeseer	**.490±.010**	.218±.003	.472±.014	.202±.003
Cora	**.587±.015**	.214±.002	.225±.013	.201±.009
Cornell	**.458±.041**	.340±.051	.449±.051	.313±.022
Texas	**.694±.053**	.428±.030	.554±.074	.433±.033
Washington	**.586±.029**	.406±.055	.583±.055	.359±.020
Wisconsin	.545±.065	.378±.043	**.568±.063**	.355±.021
News-M2	**.791±.030**	.547±.016	-	.530±.006
News-M5	**.617±.026**	.265±.017	-	.228±.003
News-M10	**.401±.024**	.159±.007	-	.117±.002
News-NG1	**.773±.032**	.535±.012	-	.531±.008
News-NG2	**.635±.019**	.246±.007	-	.225±.002
News-NG3	**.566±.012**	.178±.015	-	.144±.002
Reuters	**.472±.014**	.198±.002	-	.200±.002
3Sources	**.858±.014**	.282±.009	-	.389±.019

Table 2: Clustering results in terms of purity (mean and std.), the ratio of the multiple incomplete modal data is 90%.

4 Experiments

Data Sets: In this paper, we conduct experiments on 7 two modalities datasets and 8 multiple modalities datasets. In detail, two modal datasets include: Movie dataset is extracted from IMDb, which has 617 movies of 17 genres, and there are two data matrices describing the same movies, i.e., keywords matrix and actors matrix. The main goal is to find the genre of the movies; Citeseer dataset [Sen *et al.*, 2008] is originally made of 4 modalities, i.e., content, inbound, outbound, cites, on the same documents. We follow [Bisson and Grimal, 2012] to choose the content and cites modalities in our experiment. WebKB dataset [Sen *et al.*, 2008] contains webpages collected from 4 universities: Cornell, Texas, Wisconsin and Washington, which have 5 categories, i.e., student, project, course, stuff and faculty. Multiple modal datasets include: NewsGroup dataset [Bisson and Grimal, 2012] is of 6 groups extracted from the 20 Newsgroup datasets, i.e., M2, M5, M10, NG1, NG2, NG3. Every group contains 10 sub-

Datasets	SLIM	ConvexSub	PVC	MIC
Movie	.353±.010	.361±.010	.309±.015	**.365±.007**
Citeseer	**.379±.011**	.250±.008	.376±.014	.325±.004
Cora	**.454±.014**	.264±.004	.294±.045	.341±.004
Cornell	**.386±.039**	.231±.044	.272±.057	.290±.026
Texas	**.406±.071**	.234±.031	.264±.067	.298±.028
Washington	**.401±.032**	.264±.059	.332±.048	.282±.029
Wisconsin	**.408±.046**	.240±.050	.301±.063	.286±.031
News-M2	**.479±.056**	.159±.051	-	.176±.030
News-M5	**.506±.029**	.241±.039	-	.288±.011
News-M10	**.416±.028**	.260±.024	-	.339±.010
News-NG1	**.448±.059**	.141±.068	-	.176±.033
News-NG2	**.522±.021**	.230±.024	-	.300±.009
News-NG3	**.518±.014**	.274±.023	-	.335±.006
Reuters	**.376±.014**	.252.±.006	-	.341±.007
3Sources	**.801±.019**	.236±.010	-	.401±.019

Table 3: Clustering results in terms of NMI (mean and std.), the ratio of the multiple incomplete modal data is 90%.

Datasets	SLIM	WNH	RANC	MVL-IL
Movie	**.211±.055**	.149±.040	.203±.042	.134±.043
Citeseer	**.510±.028**	.287±.142	.457±.076	.486±.019
Cora	**.617±.020**	.436±.154	.537±.119	.536±.022
Cornell	**.502±.094**	.492±.097	.441±.091	.493±.076
Texas	**.625±.065**	.623±.077	.591±.043	.568±.050
Washington	**.612±.046**	.552±.026	.586±.086	.584±.074
Wisconsin	**.611±.079**	.554±.019	.570±.056	.574±.054
News-M2	**.743±.071**	.651±.039	.705±.030	.692±.049
News-M5	**.573±.056**	.337±.045	.504±.044	.571±.052
News-M10	**.365±.048**	.275±.039	.351±.029	.251±.025
News-NG1	**.726±.066**	.679±.071	.687±.043	.712±.071
News-NG2	**.660±.040**	.349±.020	.552±.040	.597±.053
News-NG3	**.600±.024**	.325±.083	.471±.030	.474±.029
Reuters	.434±.053	.433±.136	.394±.072	**.439±.058**
3Sources	**.828±.040**	.735±.083	.546±.144	.263±.044

Table 4: Classification results in terms of accuracy (mean and std.), the ratio of the multiple incomplete modal data is 90%.

sets, and we choose the first subset for all 6 groups in our experiment, i.e., News-M2, News-M5, News-M10, News-NG1, News-NG2 and News-NG3, respectively. 3-Source Text data (3Sources)(http://mlg.ucd.ie/datasets/3sources.html) is collected from three online news sources: BBC, Reuters, and Guardian. The description sketches of datasets, including the number of classes, the number of examples and modalities together with the feature numbers are summarized in Table 1.

We run each compared method 30 times for the 15 datasets. For all datasets, we randomly select 70% for training and the remains are for test. For both the training set and test set. As in [Li *et al.*, 2014], in each split, we randomly select 10% to 90% examples, with 20% as interval, as homogeneous examples with complete modality, and the remains are incomplete instances, i.e., in WebKB datasets, they are described by either the content or the citation modality, but not both. For all the examples, we randomly choose 30% as the labeled data, and the left 70% as unlabeled ones. In the training phase, the parameters λ_1 and λ_2 are selected by 5-fold cross validation from $\{10^{-5}, 10^{-4}, \cdots, 10^4, 10^5\}$ with further splittings on the training datasets only, i.e., there is no overlap between the test set and the validation set for parameter picking up. Empirically, when the variations between the objective value of Eq. 9 is less than 10^{-6} in iteration, we treat SLIM converged. The average accuracy and std. of predictions are recorded for indicating the classification performance, and the NMI and Purity are recorded as clustering performance. For all compared methods, the parameters are tuned best.

Compared Approaches: Our method solves the problem of semi-supervised clustering and classification with incomplete modality. Thus, to evaluate the performance of our proposed approach, for semi-supervised clustering task, we choose 3 state-of-the-art multi-modal methods: ConvexSub [Guo, 2013]; PVC [Li *et al.*, 2014]; MIC [Shao *et al.*, 2015], considering the limitation of the clustering compared method, we first learn a latent representation of the original data and then using the semi-supervised K-means to get the clustering result. For classification task, we compare with the WNH [Wang *et al.*, 2013], RANC [Ye *et al.*, 2015], MVL-IL [Xu *et al.*, 2015]. For compared methods which can't handle incomplete examples, i.e., ConvexSub, WNH, RANC, for fair comparison, we are facilitated with the ALM (Augmented Lagrange Multipliers) [Lin *et al.*, 2010] matrix completion method by first filling in the missing information.

4.1 Experiment Results

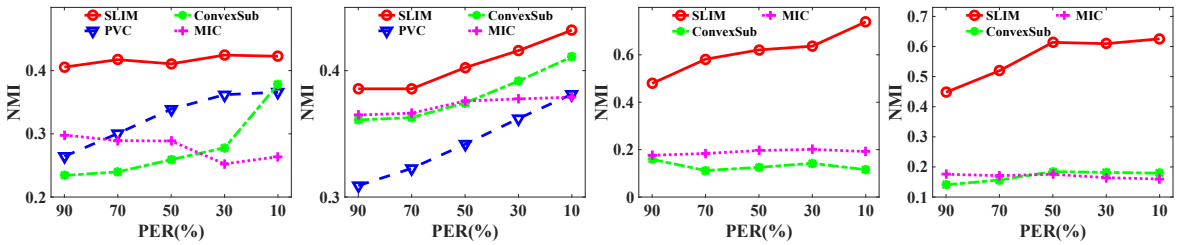
Semi-Supervised Clustering/Classification

To demonstrate the effectiveness of our proposed method. For all datasets, we fix the incomplete ratio of the multi-modal data as 90%, and record the clustering results NMI, purity of the SLIM and compared methods in Table 2, Table 3, and prediction accuracies (avg.± std.) in Table 4. It is notable that PVC method can only leverage two modalities, so we have not compared with PVC for multi-modal datasets.

From the Table 2 and Table 3, it reveals that for both two modal datasets and multiple modal datasets, SLIM almost consistently achieve the significant superior clustering performance on either purity or NMI comparing to all other methods, except for Wisconsin on purity and Movie on NMI. It can be owing to that in SLIM, the similarity matrices of all datasets are initialized with cosine similarity for more robust generalization, rather than task specific similarity matrix construction method. Besides, from the Table 4, it can be observed that SLIM also achieve the best performance on most datasets except Reuters. This phenomenon clearly reveals the effectiveness of considering the high order consistencies between each modal similarity matrix and the learned predictions, consequently, we can learn the most discriminative predictors for each modality, while acquiring better results.

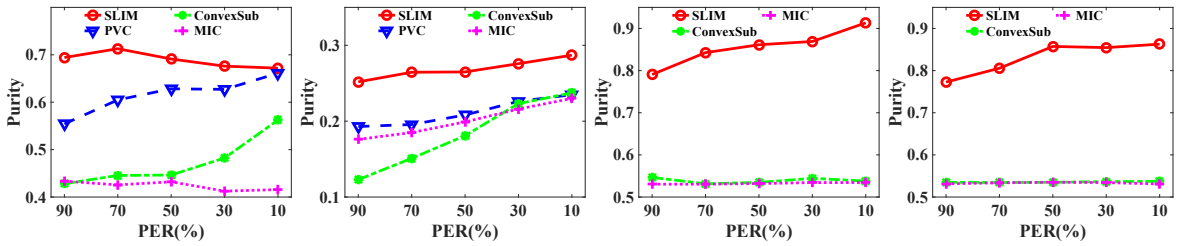
Influence of Number of Incomplete Multi-Modal Data

In order to explore the influence of the ratio of the incomplete modalities on performance, extensive experiments are conducted. In this section, the parameters in each investigation are fixed as the optimal values selected in above investigations, the λ_1 and λ_2 in SLIM are set 1, while the ratio of the incomplete data varies in $\{90\%, 70\%, \cdots, 10\%\}$, with 20%



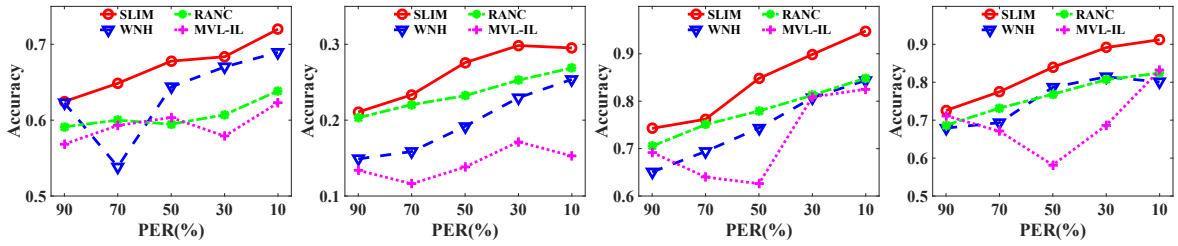
(*a*) Texas (*b*) Movie (*c*) News-M2 (*d*) News-NG1

Figure 2: The NMI results of the Texas, Movie, News-M2, News-NG1. PER (partial example ratio) is the ratio of incomplete examples.



(*a*) Texas (*b*) Movie (*c*) News-M2 (*d*) News-NG1

Figure 3: The purity results of the Texas, Movie, News-M2, News-NG1. PER (partial example ratio) is the ratio of incomplete examples.



(*a*) Texas (*b*) Movie (*c*) News-M2 (*d*) News-NG1

Figure 4: The accuracy results of the Texas, Movie, News-M2, News-NG1. PER (partial example ratio) is the ratio of incomplete examples.

as interval. Due to the page limits, results on 4 datasets, i.e., Texas, Movie, News-M2, News-NG1, and the results of NMI, purity and accuracy are recorded in Fig. 2, Fig. 3 and Fig. 4. From these figures, it clearly shows that SLIM achieves the best on all datasets. Besides, we can also find that SLIM achieves superiorities from high incomplete ratio, and the performance of SLIM increases faster than compared methods as incomplete ratio decreasing.

5 Conclusion

This paper focus on the issues of incomplete modalities in multi-modal learning. Previous mainstream solutions alleviate the affections of incomplete modal issues via utilizing the intrinsic information from the data structures or prediction consistencies among multiple modalities. A few of multi-modal learning methods consider making use of the auxiliary information from test data, and thus form transductive solutions which cannot be applied on unseen data. In this paper, we proposed a novel multi-modal learning approach, SLIM, with more extrinsic information exploited from unlabeled data in a semi-supervised scenario, and yielded an inductive learner which consequently can be applied in general multi-modal circumstances. By leveraging the intrinsic and extrinsic information together, SLIM possesses a unified framework which is closely related to classification and semi-supervised clustering. Therefore, SLIM can be easily adopted to either classification or clustering tasks. Besides, by incorporating in square-root loss, SLIM becomes less sensitive to data noise and ensures the robustness of the whole solution. Experimental evaluations on real-world applications demonstrate the superiority of our proposed method over the compared methods. How to extend the scalability with improved performance and theoretical analysis on incomplete multi-modal learning can be interesting future work.

Acknowledgments

This work was supported by National Key R&D Program of China (SQ2018YFB100002) NSFC (61773198, 61632004, 6163000043).

References

[Arora *et al.*, 2016] Raman Arora, Poorya Mianjy, and Teodor V. Marinov. Stochastic optimization for multiview representation learning using partial least squares. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 1786–1794, New York City, NY, 2016.

[Bisson and Grimal, 2012] Gilles Bisson and Clement Grimal. Co-clustering of multi-view datasets: A parallelizable approach. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 828–833, Brussels, Belgium, 2012.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1737–1744, Sydney, Australia, 2013.

[Ding and He, 2005] Chris H. Q. Ding and Xiaofeng He. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610, Newport Beach, CA, 2005.

[Eaton *et al.*, 2010] Eric Eaton, Marie desJardins, and Sara Jacob. Multi-view clustering with constraint propagation for learning with an incomplete mapping between views. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, pages 389–398, Ontario, Canada, 2010.

[Guo, 2013] Yuhong Guo. Convex subspace representation learning from multi-view data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 387–393, Bellevue, Washington, 2013.

[Iwata and Yamada, 2016] Tomoharu Iwata and Makoto Yamada. Multi-view anomaly detection via robust probabilistic latent variable models. In *Advances in Neural Information Processing Systems 29*, pages 1136–1144, Barcelona, Spain, 2016.

[Karasuyama and Mamitsuka, 2013] Masayuki Karasuyama and Hiroshi Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE Trans. Neural Netw. Learning Syst.*, 24(12):1999–2012, 2013.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1968–1974, Quebec, Canada, 2014.

[Lin *et al.*, 2010] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *CoRR*, abs/1009.5055, 2010.

[Liu *et al.*, 2014] Han Liu, Lie Wang, and Tuo Zhao. Multivariate regression with calibration. In *Advances in Neural Information Processing Systems 27*, pages 127–135, Quebec, Canada, 2014.

[Sen *et al.*, 2008] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

[Shao *et al.*, 2015] Weixiang Shao, Lifang He, and Philip S. Yu. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *Proceedings of the European Conference Machine Learning and Knowledge Discovery in Databases*, pages 318–334, Porto, Portugal, 2015.

[Shao *et al.*, 2016] Weixiang Shao, Lifang He, Chun-Ta Lu, and Philip S. Yu. Online multi-view clustering with incomplete views. In *Proceedings of the IEEE International Conference on Big Data*, pages 1012–1017, Washington DC, 2016.

[Trivedi *et al.*, 2010] Anusua Trivedi, Piyush Rai, Hal Daume, and Scott L Duvall. Multiview clustering with incomplete views. 2010.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-View Clustering and Feature Learning via Structured Sparsity. In *Proceedings of the 30th International Conference on Machine Learning*, pages 352–360, Atlanta, GA, 2013.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. *IEEE Trans. Image Processing*, 24(12):5812–5825, 2015.

[Yang *et al.*, 2015] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1033–1039, Buenos Aires, Argentina, 2015.

[Yang *et al.*, 2016] Yang Yang, De-Chuan Zhan, and Yuan Jiang. Learning by actively querying strong modal features. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2280–2286, New York, NY, 2016.

[Yang *et al.*, 2017] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1618–1625, San Francisco, California, 2017.

[Ye *et al.*, 2015] Han-Jia Ye, De-Chuan Zhan, Yuan Miao, Yuan Jiang, and Zhi-Hua Zhou. Rank consistency based multi-view learning: A privacy-preserving approach. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 991–1000, Melbourne, Australia, 2015.

[Yin *et al.*, 2017] Qiyue Yin, Shu Wu, and Liang Wang. Unified subspace learning for incomplete and unlabeled multi-view data. *Pattern Recognition*, 67:313–327, 2017.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2392–2398, New York, NY, 2016.