Semi-Supervised Multi-Modal Clustering and Classification with Incomplete Modalities

Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, *Senior Member, IEEE* and Yuan Jiang,

Abstract—In this paper, we propose a novel Semi-supervised Learning with Incomplete Modality (SLIM) method considering the modal consistency and complementarity simultaneously, and Kernel SLIM (SLIM-K) based on matrix completion for further solving the modal incompleteness. As is well known, most realistic data have multi-modal representations, multi-modal learning refers to the process of learning a precise model for complete modalities. However, due to the failures of data collection, self-deficiencies or other various reasons, multi-modal examples are usually with incomplete modalities, which generate utility obstacle using previous methods. In this paper, SLIM integrates the intrinsic consistency and extrinsic complementary information for prediction and cluster simultaneously. In detail, SLIM forms different modal classifiers and clustering learner consistently in a unified framework, while using the extrinsic complementary information from unlabeled data against the insufficiencies brought by the incomplete modal issue. Moreover, in order to deal with missing modality in essence, we propose the SLIM-K, which takes the complemented kernel matrix into the classifiers and the cluster learner respectively. Thus SLIM-K can solve the defects of missing modality in result. Finally, we give the discussion of generalization of incomplete modalities. Experiments on 13 benchmark multi-modal datasets and 2 real-world incomplete multi-modal datasets validate the effectiveness of our methods.

Index Terms—Semi-supervised Learning, Incomplete Multi-Modal Learning, Modal consistency, Modal Complementarity, Matrix Completion.

1 Introduction

This paper investigates an essential problem focusing on semi-supervised incomplete multi-modal learning. Nowadays, multi-modal learning becomes attractive with the development of data collection, and is widely used in relative applications, e.g., biological data with gene expression, array-comparative genomic hybridization, single-nucleotide polymorphism, and methylation. Most multi-modal learning approaches aim to utilize the consistency or complementarity principle among multiple modalities, to improve the generalization ability of the whole learner. E.g., [1] handled multiple modal information in semi-supervised scenario while extracting informative features of weak modality by feature learning; [2] studied the partial least square problem as a stochastic optimization problem in the big data setting; [3] proposed probabilistic latent variable models for multi-modal anomaly detection; [4], [5], [6] generalized novel deep cross-modal hash methods, which can effectively capture the intrinsic relationships between modalities. It is notable that these mainstream multi-modal learning approaches assume that all the examples are with complete modalities.

Nevertheless, the completeness hypothesis is too excessive, since many reasons could lead to incompleteness, including data collection failures from the damage of data sensors, data corruption by network communication, and data privacy policies. E.g., in web page classification with document/image representations, documents and images are two modalities, yet some web pages only have document or image information; in cross-network user identification, the user profile features, content information and 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 be directly applied to the incomplete modal situation, unless with some straightforward strategies, e.g., removing the examples with partial modalities, filling the incomplete modalities with missing data techniques. Straightforward strategies can execute the foundations for current multi-modal learning approaches, while the process causes information lose and introduces extra noises.

Aiming at the incomplete modal issue, there are some preliminary investigations. [7] studied a partial modal approach, which completes the missing modal similarity matrix using Laplacian regularization; [8] proposed the collective kernel learning to infer hidden instance similarities from multiple incomplete modalities. Yet these methods require at least one modality that contains all the examples, this is impractical in real application. Therefore, [9] learned an online multi-modal clustering algorithm OMVC to learn the latent feature matrices; [10] handled the incomplete multi-modal data well by transforming the original and incomplete data to a new and complete representation in a latent space. However, these methods mainly focus on making full use of the *inherent information*, i.e., the consistencies among multiple modalities. In this paper, we consider the defects of insufficient information caused by the incompleteness, can be remedied by supplementing *extrinsic information*, i.e., complementary modal information. Transductive multi-modal learning methods are proposed for utilizing extrinsic information, e.g., [11] proposed a new

- *Yang Yang, De-Chuan Zhan, Yi-Feng Wu and Yuan Jiang are with National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China.*
  E-mail: yangy, zhandc, wuyf, jiangy@lamda.nju.edu.cn
- *Zhi-Bin Liu is with Tencent WXG, ShenZhen 518057, China.*
  E-mail: lewiszbliu@tencent.com
- *Hui Xiong is with Rutgers University, Newark, NJ07102, United States.*
  E-mail: hxiong@rutgers.edu

De-Chuan Zhan is the corresponding author.

method SMGI, which integrates multiple graphs for label propagation; [12] proposed a novel approach to integrate heterogeneous features by performing multi-modal semi-supervised classification on unlabeled instances. However, these transductive methods are difficult to extend to classification under the incomplete modal setting with unseen data.

Different from solutions mentioned above, we propose a novel Semi-supervised Learning with Incomplete Modalities (SLIM). SLIM utilizes the intrinsic modal consistency for learning discriminative modal predictors, while considering the extrinsic unlabeled multi-modal information for clustering. In result, SLIM can perform in both transductive and inductive configurations. Essentially, SLIM only considers the extrinsic modal complementary against the incompleteness, while ignores the incomplete examples in learning modal predictors. We further improve SLIM with matrix completion to the Kernel SLIM (SLIM-K). Specifically, SLIM-K uses the latent consistently prediction to complement each modal kernel matrix, then takes the complemented kernel matrix into the classifiers and the cluster learner respectively, thus solving the defects of missing modality in result. Finally, we give out the discussion of generalization of incomplete modalities, and point out that it is better with more incomplete instances for better generalization. In conclusion, more discriminative classifiers and robust clustering learner can be achieved with SLIM and SLIM-K. In other words, SLIM/SLIM-K has wider applicable range in both classification and clustering tasks

The main contributions of this paper are summarized in the following points:

- A novel unified Semi-Supervised Learning with Incomplete Modalities (SLIM) method, which utilizes the intrinsic modal consistencies and extrinsic complementary information in one unified framework to perform transductive and inductive configurations;
- A square-root loss is utilized to calibrate modal similarity matrix by considering the different noise levels of all modal features, without learning the weights;
- A Kernel SLIM (SLIM-K) method, which takes the complemented kernel matrix into the classifiers and the cluster learner respectively, thus solves the defects of missing modality essentially;
- A discussion of generalization of incomplete modalities, finding that it is better with more incomplete instances;
- A superior performance on real-world applications, and obtaining consistently superior performances stably.

In the following parts, we start with a brief review of related works. Then we propose SLIM/SLIM-K approaches and the theoretical analysis. After that, we give the experimental results. Finally, 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 consistent and extrinsic complementary information in a semi-supervised scenario with incomplete data, and the proposed method can acquire each modal classifiers and overall clustering learner. Therefore, our work is related to both multi-modal learning and semi-supervised transductive learning.

Most of the previous multi-modal methods assume that training examples have complete modalities. However, multi-modal examples are usually with incomplete feature representation in real applications. Therefore, many researches have devoted to handle the incomplete modal data. E.g., [13] established a latent subspace where the instances corresponding to the same example in different modalities are close to each other; [14] proposed the multi incomplete modal clustering, which learns the latent feature matrices for all the modalities, and generates a consensus matrix to minimize the difference between each modality and the consensus matrix; [15] studied an effective algorithm to accomplish multi-modal learning with incomplete modalities by assuming that different modalities are generated from a shared subspace. These algorithms exploited the connections between multiple modalities, and enabled 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 consistency among multiple modalities. In this paper, we consider that the defects of insufficient information caused by the incompleteness among modalities, should be remedied by *extrinsic information* instead, i.e., using the complementarity of modal structure information.

Transductive multi-modal learning, as a matter of fact, utilizes the extrinsic information from test sets. E.g., [16] proposed a constrained clustering that can operate with an incomplete mapping, and can propagate given pairwise constraints using a local similarity measure; [17] gave a novel subspace learning framework for incomplete and unlabeled multi-modal data, the learning algorithm directly optimizes the class indicator matrix, so that the inter-modal and intra-modal data similarities are preserved to enhance the model. These approaches incorporate with the semi-supervised learning techniques can relax the issues introduced by modal incompleteness partially. However, these approaches are under the configuration of transductive learning and are difficult to extend on unseen test data, i.e., cannot build classifiers for prediction.

3 The SLIM Approach

The incomplete multi-modal learning problem in this paper focuses on following problems: 1) process the incomplete multi-modal data, rather than remove the incomplete modalities or fill in with the complete modal average values; 2) use the latent consistent predictions to complement each modal kernel representation, thus the predictors can take more advantage of the incomplete modalities. 3) generalize discussion in the incomplete modal scenario. Consequently, we can perform both transductive and inductive configurations in one unified framework.

3.1 Problem Formulation

In multiple modal learning, an instance is characterized by multiple modal representations with one unified label. Suppose we are given a dataset possessing N examples with V modalities. The $i-$th instance $\mathbf{x}_i$ of $v-$th modality can be represented as $\mathbf{x}_{i^v} \in \mathbb{R}^{d_v}$, where d_v is the dimension of the $v-$th modality. In the complete modal setting, all the instances have V modal representations. On the contrary, partial instances exist missing modalities in incomplete modal scenario. For example, as shown in Fig. 1, each instance may have complete or partial modalities, i.e., incomplete image/content pairs only have text or image information. It is worth noting that incomplete modalities exist in both labeled and unlabeled instances, this is more suitable for realistic application.

Without any loss of generality, under the semi-supervised scenario, we assume that there are l labeled examples including

Fig. 1. An illustration of the Incomplete Multi-Modal Data in real-world application as Wiki data.

complete or incomplete modalities, the labeled example sets can be represented as Θ_l. For labeled examples, the ground truth can be represented as $\mathbf{y}_i \in \{1, \cdots, C\}$, C represents the number of class. For the incomplete representation perspective, suppose we have N_c homogeneous examples with complete modal features, i.e., $X_c = \{(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_c})\} \in \mathbb{R}^{N_c \times d}$, where $\mathbf{x}_i$ represents the instance with complete modalities, i.e., $\mathbf{x}_i = \{\mathbf{x}_{i^1}, \mathbf{x}_{i^2}, \cdots, \mathbf{x}_{i^V}\} \in R^d$, $d = d_1 + d_2 + \cdots + d_V$. The remaining $N_{in} = N - N_c$ instances are with incomplete modalities. $X_{in} = \{(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_{N_{in}})\} \in \mathbb{R}^{N_{in}}$, where $\hat{\mathbf{x}}$ denotes the instance with incomplete modalities missing one or more modalities. Therefore, $X_v = \{X_{c^v}, X_{in^v}\}$ denotes the representation of $v-$th modality and ignores the incomplete instances in this modality. Consequently, the incomplete multi-modal learning problem can be defined as:

Definition 1. (Semi-supervised Learning with Incomplete Modalities) Given $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_c}, \hat{\mathbf{x}}_{N_c+1}, \cdots, \hat{\mathbf{x}}_N\}$ and the ground truth Y of labeled examples, $v-$th modality can be denoted as X_v. The task is to learn a function set for various modalities: $\mathcal{H} = \{h_1, h_2, \cdots, h_V\}$, where $h_v : X_v \to Y$ represents the classifier for $v-$th modality. While the task also aims to get the $\hat{Y}$, which represents the cluster results for all the unlabeled instances.

3.2 The Proposed Approach

In this section, we will describe the SLIM and SLIM-K in detail. In incomplete modal learning, SLIM utilizes the intrinsic modal consistencies to learn more discriminative predictors, while considers the extrinsic complementary modal structure information against the incompleteness for clustering. It is easy to find that the SLIM ignores the missing modality during predicting, to solve this problem, SLIM-K uses the modal kernel matrix instead of the similarity matrix, and the latent consistent prediction across various modalities can complement each modal kernel matrix conversely. Therefore, we include the complemented kernel matrix in predictors to overcome the missing modalities.

3.2.1 SLIM Method

Specifically, SLIM can be decomposed into two targets: first, we aim to learn the predictors to classify each modality accurately. Second, we wish to cluster the unlabeled instances by modeling a joint transformed matrix factorization problem, with respect to each modal similarity matrix and the shared learned predictions. Therefore, SLIM can be defined as:

$$\begin{aligned} &\min_{f_v, \hat{Y}} \sum_{v=1}^{V} (\hat{L}_v(f_v(X_v), \hat{Y}) + \frac{\lambda_2}{2} \tilde{L}_v(X_v, \hat{Y})) \\ &s.t. \quad \hat{Y}^{\Theta_l} = Y \end{aligned} \tag{1}$$

There are V modalities, the first term $\hat{L}_v(f_v(X_v), \hat{Y})$ denotes the loss of classification of $v-$th modality, which indicates the intrinsic consistency among different modalities. $f_v(X_v)$ is the classification result, $\hat{Y}$ is the label to be learned. In multi-class cases, we expand the label Y to a vector with C elements, where $Y_{i,j} = 1$ indicates the $i-$th instance with label j, otherwise, $y_{i,j} = 0$. The constraint $\hat{Y}^{\Theta_l} = Y$ restricts the prediction on labeled data as same as the ground truth to avoid collapsing of predictions, Θ_l here is the set of labeled data. The second term $\tilde{L}_v(X_v, \hat{Y})$ considers extrinsic complementary modal structure information for clustering. More specifically, it models a joint transformed matrix factorization problem, here $X_v \in \mathbb{R}^{d_v}$ is the $v-$th modality with missing rows filling with zeros. In other words, we treat each modal similarity matrix and the learned consistent predictions as a transformed matrix factorization problem, and $\lambda_2 > 0$ is the balance parameter.

Predictor for Each Modality. Objective function $\hat{L}_v$ of the $v-$th modality in Eq. 1 can be generally represented as the form:

$$\min_{f_v} \ell(f_v(X_v), \hat{Y}) + \frac{\lambda_1}{2} r(f_v). \tag{2}$$

Here $r(f_v)$ is the regularization for modal-specific classifier. $\frac{\lambda_1}{2}$ is a scalar coefficient to balance the weights of the two terms. Without any loss of generality, the f_v can take linear or non-linear classifier here, for simplicity, here we use linear function for SLIM:

$$F_v = X_v W_v + \mathbf{1}\mathbf{b}_v^\top \odot P_v \tag{3}$$

Where $F_v = \{f_v(x_{1^v}), f_v(x_{2^v}), \cdots, f_v(x_{N^v})\} \in \mathbb{R}^{N \times C}$, missing rows are filled with zeros. $W_v \in \mathbb{R}^{d_v \times C}$ is the linear classifier, $\mathbf{b}_v \in \mathbb{R}^C$ is the bias for current predictor, $\mathbf{1}$ is the all one vector, $\odot$ represents element wise product operator, $P_v \in \mathbb{R}^{N \times C}$ is the indicator matrix, where $[P_v]_{i,\cdot} = 1$ iff $i-$th instance is complete on $v-$th modality, otherwise $[P_v]_{i,\cdot} = 0$ indicates the incomplete modal scenario. The loss function $\ell(\cdot)$ can take any convex forms, we use square loss here. As a result, combining with Eq. 3, the loss function can be rewritten as:

$$\min_{W_v, b_v} \frac{1}{2\eta_v} \|F_v - \hat{Y} \odot P_v\|_{\mathrm{F}}^2 + \frac{\lambda_1}{2} \|W_v\|_{\mathrm{F}}^2 \tag{4}$$

Here $\hat{Y} \in \mathbb{R}^{N \times C}$ denotes the predictions of all instances, η_v is the number of the complete examples of $v-$th modality for balance weight.

Cluster learner for All Modalities. The extrinsic information, i.e., the complementarity among different modalities, is one of the most prominent information to relieve the modal incompleteness. We refer the complete modalities to complement the incomplete modalities. Therefore, as to the supervised case, the 2nd term can be defined as:

$$\tilde{L}_v = \|\mathcal{R}_\Omega(M_v) - \mathcal{R}_\Omega(YY^\top)\|_{\mathrm{F}}^2$$

Where $M_v \in \mathbb{R}^{N_l \times N_l}$ is the Laplacian matrix of the $v-$th modal labeled examples, and we define $[M_v]_{i,j} = [D_v]_{i,j} - [S_v]_{i,j}$,

$[S_v]_{i,j}$ is similarity matrix of $v-$th modal instances, denotes as $exp(-\frac{1}{2\delta^2}(\mathbf{x}_{i^v}-\mathbf{x}_{j^v})^\top(\mathbf{x}_{i^v}-\mathbf{x}_{j^v}))$ according to [18]. $[D_v]_{i,j}$ is the diagonal matrix induced from $[S_v]_{i,j}$, $[D_v]_{i,i}=\sum_j[S_v]_{i,j}$. $[\mathcal{R}_\Omega(M_v)]_{i,j}=[M_v]_{i,j}$ iff $i-$th instance and $j-$th instance have complete entries on $v-$th modality, otherwise $[\mathcal{R}_\Omega(M_v)]_{i,j}=0$, 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 labeled and unlabeled data, with learned labels as $\hat{Y}$, and $\tilde{L}_v$ can be reformulated as:

$$\begin{aligned}&\min_{\hat{Y}}\frac{1}{\eta_v^2}\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}^2\\&s.t.\quad \hat{Y}^{\Theta_l}=Y, 0\le\hat{Y}\le 1,\end{aligned}\tag{5}$$

note that here each similarity matrix only has $\eta_v\times\eta_v$ real-valued entries and we fill the rest entries with zeros. In addition, we constrain the predicted values into the same range as true labels by $0\le\hat{Y}\le 1$ to maintain the intrinsic consistency. It is notable that the Eq. 5 relates to the kernel k-means and laplacian-based spectral clustering closely in a wild condition [19], which implies that the whole approach can also be applied in clustering tasks.

Considering Modal Insufficiencies. It is also notable that real world data always contain noise and outlier 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 caused by noises. However, in semi-supervised learning scenario, few labeled data can be used for parameters tuning. In this paper, we further employ the square-root loss function instead of the least square function in Eq. 5 to reduce the affections from noisy data. This solution can be regarded as a weighted regularized least square form of the original one, where the weight for each modality is: $\frac{1}{\eta_v\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}}$ according to [20]. This modification can calibrate each modality by considering the different noise levels, and increase the robustness of the 2nd term in SLIM:

$$\begin{aligned}&\min_{\hat{Y}}\frac{1}{\eta_v}\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}\\&s.t.\quad \hat{Y}^{\Theta_l}=Y, 0\le\hat{Y}\le 1.\end{aligned}\tag{6}$$

We can combine Eq. 4 and Eq. 6 in a unified framework and yield the whole SLIM model:

$$\begin{aligned}\min_{W_v,b_v,\hat{Y}}&\sum_{v=1}^{V}\frac{1}{2\eta_v}\|F_v-\hat{Y}\odot P_v\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\|F_v\|_{\mathrm{F}}^2\\&+\frac{\lambda_2}{\eta_v^2}\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}\\&s.t.\quad 0\le\hat{Y}\le 1,\hat{Y}^{\Theta_l}=Y\end{aligned}\tag{7}$$

3.2.2 Kernel SLIM (SLIM-K)

SLIM utilizes the complementarity among different modalities to avoid the influence of incompleteness. However, SLIM ignores the missing modal features when build the predictor for each modality in Eq. 3, which only considers the complete modal information. Consequently, it may lead weak predictors when there are a large number of missing modalities. To consider the intrinsic and extrinsic information in learning cluster learner and predictors simultaneously, we propose a kernel extension for building the predictors. Specifically, inspired from [21], we apply kernel tricks

Algorithm 1 The pseudo code of SLIM

Input:
- Dataset: $\mathcal{D}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_{N_c},\hat{\mathbf{x}}_{N_c+1},\cdots,\hat{\mathbf{x}}_N\}$
- Parameter: λ_1, λ_2
- maxIter: T, learning rate: $\{\alpha_t\}_{t=1}^T$

Output:
- Classifiers: $W_v,\mathbf{b}_v, v=1,2,\cdots,V$
- Clustering Result: $\hat{Y}$

```
Initialize M_v ← X_v
Initialize Ŷ ← Y
while ‖Func_obj^{t+1} − Func_obj^t‖ > ϵ and t < T do
    Calculate g_t ← Eq. 15
    Ŷ^{t+1} = Ŷ^t − α_t g_t
    Ŷ^{t+1}^{Θ_l} = Y
    Func_obj^{t+1} ← calculate obj. value in Eq. 14 with Ŷ^{t+1}
end while
Solve W_v, b_v from Ŷ using Eq. 13, Eq. 11
```

to build f_v, in which the inner products of two data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ in the new random space approximates the kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$, as a result, the kernel SLIM method (named SLIM-K) is given as:

$$\hat{F}_v=\alpha_v{}^\top K_v+\mathbf{1}\mathbf{b}_v^\top\tag{8}$$

where $\hat{F}_v=\{\hat{f}_v(x_{1^v}),\hat{f}_v(x_{2^v}),\cdots,\hat{f}_v(x_{N^v})\}\in\mathbb{R}^{C\times N}$. $\alpha_v=\{\alpha_{j^v}^c, j=1,2,\cdots,N\}\in\mathbb{R}^{N\times C}$, C is the class number, $K_v(\cdot,\cdot)$ is the $v-$th modal kernel matrix for both labeled and unlabeled data, where $[K_v]_{i,j}=\phi(\mathbf{x}_{i^v})\phi(\mathbf{x}_{j^v})$, and $\phi(\cdot)$ is the defined kernel function. Consequently, loss function can be reformulated as:

$$\min_{\alpha_v,\mathbf{b}_v,K_v}\frac{1}{2\eta_v}\|\hat{F}_v-\hat{Y}\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\|\hat{F}_v\|_2^2+\frac{\lambda_2}{\eta_v^2}\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(K_v)\|_{\mathrm{F}}\tag{9}$$

$\|\hat{F}_v\|_2^2$ is the structure risk of $v-$th modal predictor in the function space. K_v is a variable for optimization, $\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(K_v)\|_{\mathrm{F}}$ controls the intrinsic and extrinsic consistency among different modal kernel matrixes and observed similarity matrix. $M_v\in\mathbb{R}^{N\times N}$ is the $v-$th modal similarity matrix of the labeled and unlabeled examples. Furthermore, to insure the consistency between the latent indicator matrix and each modal kernel matrix, we can define a new regularization as following:

$$\begin{aligned}&\min_{K_v,\hat{Y}}\|K_v-\hat{Y}\hat{Y}^\top\|_{\mathrm{F}}^2\\&s.t.\quad \hat{Y}^{\Theta_l}=Y, 0\le\hat{Y}\le 1,\end{aligned}$$

Here the regularization aims to learn more discriminative kernel matrix and indicator matrix, note that this term also acts as the matrix completion. Thereby we can acquire more robust predictors and cluster learner. And the SLIM-K can be formulated as:

$$\begin{aligned}\min_{\alpha_v,\mathbf{b}_v,K_v,\hat{Y}}&\sum_{v=1}^{V}\frac{1}{2\eta_v}\|\hat{F}_v-\hat{Y}\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\|\hat{F}_v\|_{\mathrm{F}}^2\\&+\frac{\lambda_2}{\eta_v^2}\|\mathcal{R}_\Omega(M_v)-\mathcal{R}_\Omega(K_v)\|_{\mathrm{F}}+\|K_v-\hat{Y}\hat{Y}^\top\|_{\mathrm{F}}^2\\&s.t.\quad 0\le\hat{Y}\le 1,\hat{Y}^{\Theta_l}=Y\end{aligned}\tag{10}$$

3.3 Optimization

In this section, we mainly focus on the methodology of addressing the optimization of SLIM and SLIM-K. These two methods share a similar optimization process. Without any loss of generalization, we take SLIM as an example. SLIM is convex to $W_v, \mathbf{b}_v$ yet not a jointly convex problem. An alternative descent algorithm is considered to solve this problem. Nevertheless, further derivations successfully show that alternative descent approach is with closed-form solutions for some key parameters, i.e., $W_v, \mathbf{b}_v$.

3.3.1 Fix W_v and $\hat{Y}$, Update $\mathbf{b}_v$

First, the optimal solution of $\mathbf{b}_v$ is with closed-form when W_v and $\hat{Y}$ are fixed,

$$\mathbf{b}_v = \frac{1}{\eta_v}(\hat{Y} \odot P_v - X_v W_v)^\top \mathbf{1} \tag{11}$$

3.3.2 Fix $\mathbf{b}_v$ and $\hat{Y}$, Update W_v

Substitute Eq. 11 into Eq. 7, we can simplify Eq. 7 as:

$$\begin{aligned}\min_{W_v,\hat{Y}} \sum_{v=1}^{V} \frac{1}{2\eta_v}\|C_v X_v W_v - C_v(\hat{Y} \odot P_v)\|_{\mathrm{F}}^2 + \frac{\lambda_1}{2}\|W_v\|_{\mathrm{F}}^2 \\ + \frac{\lambda_2}{2}\frac{1}{\eta_v}\|\mathcal{R}_\Omega(M_v) - \mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}} \\ s.t. \quad 0 \le \hat{Y} \le 1, \hat{Y}^{\Theta_l} = Y,\end{aligned} \tag{12}$$

where $C_v = I - \frac{1}{\eta_v}\mathbf{1}\mathbf{1}^\top \odot P_v$. Then we can find that the optimal solution of W_v is also with closed-form when $\hat{Y}$ is fixed:

$$W_v = A_v B_v C_v(\hat{Y} \odot P_v), \tag{13}$$

where $A_v = ({X_v}^\top C_v^\top C_v X_v + \eta_v \lambda_1 I)^{-1}$, $B^v = {X_v}^\top C_v^\top$.

3.3.3 Fix $\mathbf{b}_v$ and W_v, Update $\hat{Y}$

Combining Eq. 13 and Eq. 12, we can rewrite the Eq. 12 as:

$$\min_{\hat{Y}} tr(\hat{Y}^\top H \hat{Y}) + \lambda_2 \sum_{v=1}^{V} \frac{1}{2\eta_v}\|\mathcal{R}_\Omega(M_v) - \mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}, \tag{14}$$

where $tr(\cdot)$ is the matrix trace operator, $H = \sum_{v=1}^{V} \Pi_{\Gamma_v}[C_v C_v^\top B_v^\top A_v^\top(\frac{\lambda_1}{2}A_v B_v + \frac{1}{2\eta_v}B_v B_v^\top A_v B_v - \frac{1}{\eta_v}B_v) + \frac{1}{2\eta_v}C_v^\top C_v]$, where $\Gamma_v = \{\gamma_1, \gamma_2, \cdots, \gamma_{\eta_v}\}$ represents the index set of the complete instances of $v-$th modality. $\Pi_{\Gamma_v}(A)$ represents the rows and columns in Γ_v of matrix A are 0. And we can use the project sub-gradient method to optimize Eq. 14 for simplicity.

$$g = \begin{cases} H\hat{Y}, & \bar{L} = 0, \\ H\hat{Y} + \lambda_2 \sum_{v=1}^{V} \dfrac{\mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top) - \mathcal{R}_\Omega(M_v)}{\|\mathcal{R}_\Omega(M_v) - \mathcal{R}_\Omega(\hat{Y}\hat{Y}^\top)\|_{\mathrm{F}}}\hat{Y}, & \text{Otherwise} \end{cases} \tag{15}$$

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

With the parameters W_v and $\mathbf{b}_v$ in closed-form, we can solve the $\hat{Y}$ with the projected sub-gradient of Eq. 15. The whole procedure is summarized in Algorithm 1.

3.4 Discussion

In the incomplete multi-modal scenario, assuming that we miss the $v-$th modality of $i-$th instance with the probability p, the probability of missing the $i-$th instance will be p^V obviously. In other words, we may miss an instance of all modalities with equal probability of p^V (V is the number of modality), while other instances are complete. This brings out the dilemma between "more incomplete instances missing partial modalities" and "more complete instances missing all modalities".

In Eq. 7, the 3rd term can be regarded as a variant of the consistency principle of co-regularize method, and our method can be degenerated as a CoRLS style method. Considering the binary classification for two modalities, [22] gives the generalization bounds for the class in terms of the empirical Rademacher complexity. Analogously, under the "more complete instances" scenario, we remove the instances with the probability of p^V. Thus, the generalization bounds is give as:

Theorem 1. Suppose that L: $\mathcal{Y}^2 \to [0, 1]$ satisfies the uniform Lipschitz condition: for all $y \in \mathcal{Y}$ and all $\bar{y}_1, \bar{y}_2 \in \mathcal{Y}$ with $\bar{y}_1 \neq \bar{y}_2$,

$$\frac{\|L(\bar{y}_1, y) - L(\bar{y}_2, y)\|}{\bar{y}_1 - \bar{y}_2} \le B$$

where B is the size of the labeled kernel sub-matrices. Then conditioned on the unlabeled data, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the sample of labeled points drawn i.i.d. from $\mathcal{D}$, we have that for any predictor $f \in \mathcal{J}$:

$$\begin{aligned}E_{\mathcal{D}} L(\varphi(X), Y) \le & \frac{1}{(1-p^V)l} L(\varphi(X), Y) + 2B\hat{R}_l(J) \\ & + \frac{1}{\sqrt{(1-p^V)l}}(2 + 3\sqrt{ln(2/\delta)/2})\end{aligned}$$

The Loss function $L(\bar{y}, y)$ is defined the same as [22].

Theorem 2. For the CoRLS style function class $\mathcal{J}$,

$$\frac{1}{\sqrt[4]{2}}\frac{U}{(1-p^V)l} \le \hat{R}_l(J) \le \frac{U}{(1-p^V)l}$$

Where U can be calculated by the unlabeled kernel sub-matrices, labeled kernel sub-matrices and cross-terms in the kernel matrix K, and can be defined similar to [22].

Given the incomplete kernel matrix, "more incomplete instances" can be regarded as a matrix decomposition problem, which computes the low rank approximation $\hat{Y}$ of a given matrix K by using the actual rows and columns of the matrix. [23] gives the matrix approximation theorem,

Theorem 3. Assume $rank(K) \le r$, $d \ge 7\mu(r)r(t + lnr)$ is the number of columns and rows uniformly sampled at random, and the complete size $\|\Omega\| \ge 7\mu^2(r)r^2(t + 2lnr)$. Then, with a probability at least $1 - 5e^{-t}$, we have $K = \hat{K}$, where $\hat{K}$ is a approximation estimation.

Theorem 3 shows that under the incoherent condition, a rank r incomplete matrix can be perfectly recovered with $\mathcal{O}(nrlnr)$ observed entries. With Theorem 3, we can firstly predict a well approximated kernel matrix K, then get a more tight generalization bound with N instances in Theorem 1. Thus, "more incomplete instances" in our setting is better than the "more complete instances" in previous works.

TABLE 1
Clustering comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on 14 benchmark datasets only missing one modality for each instance, the ratio of the multiple incomplete modal data is 90%. 2 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller the better of a criterion.

Data	Putity ↑					NMI ↑				
	SLIM-K	SLIM	ConvexSub	PVC	MIC	SLIM-K	SLIM	ConvexSub	PVC	MIC
Mov.	**.266±.014**	.247±.009	.123±.004	.193±.003	.172±.001	**.387±.049**	.353±.010	.361±.010	.309±.015	.365±.007
Cite.	**.524±.052**	.490±.010	.218±.003	.472±.014	.202±.003	**.384±.039**	.379±.011	.250±.008	.376±.014	.325±.004
Cora	**.606±.064**	.587±.015	.214±.002	.225±.013	.201±.009	**.476±.080**	.454±.014	.264±.004	.294±.045	.341±.004
Corn.	**.476±.034**	.458±.041	.340±.051	.449±.051	.313±.022	**.423±.048**	.386±.039	.231±.044	.272±.057	.290±.026
Texas	**.697±.082**	.694±.053	.428±.030	.554±.074	.433±.033	**.432±.050**	.406±.071	.234±.031	.264±.067	.298±.028
Wash.	**.628±.065**	.586±.029	.406±.055	.583±.055	.359±.020	**.452±.051**	.401±.032	.264±.059	.332±.048	.282±.029
Wis.	**.609±.030**	.545±.065	.378±.043	.568±.063	.355±.021	**.462±.047**	.408±.046	.240±.050	.301±.063	.286±.031
M2	**.839±.032**	.791±.030	.547±.016	-	.530±.006	**.502±.057**	.479±.056	.159±.051	-	.176±.030
M5	**.630±.031**	.617±.026	.265±.017	-	.228±.003	**.590±.038**	.506±.029	.241±.039	-	.288±.011
M10	**.476±.027**	.401±.024	.159±.007	-	.117±.002	**.476±.029**	.416±.028	.260±.024	-	.339±.010
NG1	**.806±.032**	.773±.032	.535±.012	-	.531±.008	**.469±.090**	.448±.059	.141±.068	-	.176±.033
NG2	**.675±.024**	.635±.019	.246±.007	-	.225±.002	**.577±.054**	.522±.021	.230±.024	-	.300±.009
NG3	.565±.034	**.566±.012**	.178±.015	-	.144±.002	**.569±.039**	.518±.014	.274±.023	-	.335±.006
Reut.	**.510±.041**	.472±.014	.198±.002	-	.200±.002	**.411±.034**	.376±.014	.252.±.006	-	.341±.007

TABLE 2
Classification comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on 14 benchmark datasets only missing one modality for each instance, the ratio of the multiple incomplete modal data is 90%. 2 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller the better of a criterion.

Data	Accuracy ↑					F1 ↑				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
Mov.	**.229±.027**	.211±.055	.149±.040	.203±.042	.134±.043	**.139±.013**	.131±.035	.113±.008	.118±.033	.129±.006
Cite.	**.529±.037**	.510±.028	.287±.142	.457±.076	.486±.019	**.364±.039**	.347±.027	.259±.009	.303±.029	.343±.017
Cora	**.646±.096**	.617±.020	.436±.154	.537±.119	.536±.022	**.444±.023**	.433±.021	.295±.016	.381±.068	.379±.019
Corn.	**.509±.081**	.502±.094	.492±.097	.441±.091	.493±.076	**.484±.062**	.431±.064	.384±.055	.373±.065	.412±.056
Texas	**.630±.063**	.625±.065	.623±.077	.591±.043	.568±.050	**.577±.037**	.560±.043	.548±.057	.498±.065	.532±.067
Wash.	.600±.090	**.612±.046**	.552±.026	.586±.086	.584±.074	**.578±.040**	.539±.034	.491±.076	.511±.068	.467±.058
Wis.	**.627±.081**	.611±.079	.554±.019	.570±.056	.574±.054	**.569±.068**	.525±.076	.492±.069	.461±.054	.502±.086
M2	.733±.013	**.743±.071**	.651±.039	.705±.030	.692±.049	**.695±.051**	.673±.042	.570±.022	.586±.027	.609±.018
M5	**.600±.079**	.573±.056	.337±.045	.504±.044	.571±.052	**.402±.036**	.401±.041	.298±.025	.327±.026	.326±.015
M10	.353±.090	**.365±.048**	.275±.039	.351±.029	.251±.025	**.210±.013**	.207±.027	.179±.003	.182±.022	.182±.006
NG1	**.731±.087**	.726±.066	.679±.071	.687±.043	.712±.071	**.666±.013**	.642±.038	.575±.035	.583±.023	.619±.010
NG2	**.700±.099**	.660±.040	.349±.020	.552±.040	.597±.053	**.490±.057**	.489±.039	.291±.020	.365±.037	.324±.009
NG3	**.600±.067**	**.600±.024**	.325±.083	.471±.030	.474±.029	**.400±.079**	.367±.025	.221±.001	.266±.020	.225±.005
Reut.	**.440±.073**	.434±.053	.433±.136	.394±.072	.439±.058	**.301±.017**	.285±.015	.237±.035	.246±.027	.280±.013

4 EXPERIMENTS

Data Sets: In this paper, we conduct experiments on 8 two modalities datasets and 8 multiple modalities datasets. In detail, two modal datasets come from: Movie dataset is extracted from IMDB, which has 617 movies of 17 genres, with two modalities describing the same movies, i.e., keywords matrix and actors matrix. The goal is to find the genre of the movies. Citeseer dataset [24] is originally made with 4 modalities, i.e., content, inbound, outbound, citation. Cora dataset [24] has the same structure as Citeseer, i.e., the content modality and the citation modality are used in our experiment as well [25]. WebKB [24] is described with two modalities: content and citation. In this paper, we seperate WebKB into 4 sub-datasets grouped by universities: Cornell, Texas, Wisconsin and Washington, each has 5 categories, i.e., student, project, course, stuff and faculty. Multiple modal datasets include: NewsGroup [25] has 3 modalities, which are constructed by different preprocessing methods for texts, i.e., partitioning around medoids, supervised mutual information and unsupervised mutual information. NewsGroup dataset [25] is of 6 groups extracted from 20 Newsgroup datasets, i.e., M2, M5, M10, NG1, NG2, NG3. Every group contains 10 subsets, 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. Reuters dataset [25] is built from the Reuters RCV1/RCV2 multilingual test collection, multi-modal information is created from different languages, i.e., English, French, German, Italian and Spanish.

We also conduct experiments on with 2 realistic incomplete multi-modal datasets. 3-Source Text data (3Sources) [1] is collected from three online news sources, i.e., BBC, Reuters, and Guardian, each source can be seen as one modality for the news reports. In total, there are 948 news articles covering 416 distinct news reports. In these reports, 169 were reported with three sources,

[1] http://mlg.ucd.ie/datasets/3sources.html

TABLE 3
Classification comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on different modalities (only missing one modality for each instance), the ratio of the multiple incomplete modal data is 90%. The best performance for each criterion is bolded.

Data	Accuracy ↑					F1 ↑				
	Modality_1					Modality_1				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
Mov.	**.143±.028**	.139±.012	.046±.029	.130±.026	.055±.013	**.129±.013**	.127 ±.011	.114±.007	.105±.006	.113±.009
Cite.	**.386±.044**	.314±.023	.239±.016	.307±.017	.207±.009	**.321±.056**	.317 ±.034	.258±.014	.255±.009	.214±.003
Cora	**.406±.070**	.403±.016	.331±.017	.402±.101	.297±.012	**.400±.095**	.303 ±.008	.298±.011	.304±.022	.221±.007
Corn.	**.441±.061**	.437±.083	.390±.089	.268±.065	.395±.111	**.440±.049**	.417 ±.046	.373±.061	.338±.026	.345±.040
Texas	**.601±.038**	.554±.054	.382±.108	.440±.150	.568±.050	**.555±.031**	.553 ±.044	.460±.043	.500±.062	.506±.109
Wash.	**.437±.058**	.432±.062	.384±.065	.433±.093	.257±.136	**.478±.040**	.477 ±.041	.391±.042	.415±.054	.412±.049
Wis.	**.571±.052**	.517±.058	.430±.074	.395±.088	.271±.150	**.486±.051**	.485 ±.052	.421±.050	.393±.047	.370±.033
Data	Modality_2					Modality_2				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
Mov.	**.200±.025**	.170±.046	.039±.012	.135±.026	.078±.007	**.132±.032**	.130 ±.019	.114±.005	.106±.009	.123±.010
Cite.	**.486±.057**	.474±.094	.210±.015	.349±.065	.468±.028	**.311±.069**	.289 ±.052	.264±.002	.269±.010	.243±.009
Cora	**.521±.083**	.506±.013	.285±.032	.350±.059	.484±.016	**.403±.054**	.305 ±.009	.285±.021	.274±.015	.239±.006
Corn.	**.489±.097**	.488±.053	.397±.098	.397±.119	.480±.068	.423±.047	**.425 ±.057**	.404±.042	.369±.053	.427±.057
Texas	**.621±.072**	.611±.070	.581±.050	.477±.086	.604±.069	.552±.045	**.558 ±.050**	.557±.044	.420±.078	.497±.078
Wash.	.556±.088	**.559±.076**	.443±.049	.426±.079	.532±.115	**.539±.034**	.496 ±.045	.496±.049	.409±.031	.490±.058
Wis.	.600±.050	**.608±.029**	.491±.102	.463±.067	.466±.077	**.528±.039**	.518 ±.073	.448±.099	.416±.052	.512±.075
Data	OverAll					OverAll				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
Mov.	**.229±.027**	.211±.055	.149±.040	.203±.042	.134±.043	.139±.013	**.193±.038**	.116±.003	.164±.043	.160±.007
Cite.	**.529±.037**	.510±.028	.287±.142	.457±.076	.486±.019	**.364±.039**	.347±.027	.259±.009	.303±.029	.343±.017
Cora	**.646±.096**	.617±.020	.436±.154	.537±.119	.536±.022	**.444±.023**	.433±.021	.295±.016	.381±.068	.379±.019
Corn.	**.509±.081**	.502±.094	.492±.097	.441±.091	.493±.076	**.484±.062**	.431±.064	.384±.055	.373±.065	.412±.056
Texas	**.630±.063**	.625±.065	.623±.077	.591±.043	.568±.050	**.577±.037**	.575±.042	.575±.075	.535±.078	.565±.074
Wash.	.600±.090	**.612±.046**	.552±.026	.586±.086	.584±.074	**.578±.040**	.539±.034	.491±.076	.511±.068	.467±.058
Wis.	**.627±.081**	.611±.079	.554±.019	.570±.056	.574±.054	**.569±.068**	.525±.076	.492±.069	.461±.054	.502±.086

TABLE 4
Dataset description, datasets with two modalities or multiple modalities are separated with a horizontal line.

Datasets	C	N	V	$d_v(v = 1, 2, \cdots, V)$
Movie (Mov.)	17	617	2	1878, 1398
Citeseer (Cite.)	6	3264	2	3703, 3264
Cora	7	2708	2	1433, 2708
Cornell (Corn.)	5	195	2	1703, 195
Texas	5	185	2	1703, 185
Washington (Wash.)	5	217	2	1703, 217
Wisconsin (Wis.)	5	262	2	1703, 262
WKG	2	6,500	2	1024,300
News-M2 (M2)	2	1200	3	2000, 2000, 2000
News-M5 (M5)	5	500	3	2000, 2000, 2000
News-M10 (M10)	10	500	3	2000, 2000, 2000
News-NG1 (NG1)	2	500	3	2000, 2000, 2000
News-NG2 (NG2)	5	400	3	2000, 2000, 2000
News-NG3 (NG3)	8	1000	3	2000, 2000, 2000
Reuters (Reut.)	6	1600	5	2000, 2000, 2000, 2000, 2000
3Sources	6	416	3	3560, 3631, 3068

194 with two sources, and 53 appeared with single news source. Each report is manually annotated with one of the 6 topical labels: business, entertainment, health, politics, sport, and technology. We also collect data from WKG Game-Hub, which consists 13,750 articles collected from the Game-Hub of " Strike of Kings", we take the largest two classes as a binary balance problem. Each article contains several images and content paragraphs. The text is represented as a 300-dimensional word2vector vector, the image is represented as 1024 deep output feature. The description sketches of datasets, including the number of classes, the number of examples and modalities as well as the feature numbers, are summarized in Table 4.

We run each compared method 30 times for the 16 datasets. In each datasets, we randomly select 70% for training and the remains are for testing. For both the training and testing set, we randomly select 10% to 90% examples in each split, with 20% as interval, as homogeneous examples with complete modality, and remains are incomplete instances as in [13], i.e., in WebKB datasets, they are described by either content or citation modality. For all the examples, we randomly choose 30% as the labeled data, and the left 70% are unlabeled data. In the training phase, the parameters λ_1 and λ_2 are selected by 5-fold cross validation from $\{10^{-5}, 10^{-4}, \cdots, 10^4, 10^5\}$, there is no overlap between the test set and the validation set. Empirically, when the variation between the objective values of Eq. 14 is less than 10^{-6} in iteration, we treat SLIM/SLIM-K converged. The average mean and std. of predictions are recorded for indicating the classification performance, and NMI and Purity are recorded for cluster performance. For compared methods, the parameters are tuned respected to the original paper suggested.

Compared Approaches: Our method solves the problem of semi-supervised clustering and classification with incomplete modality.

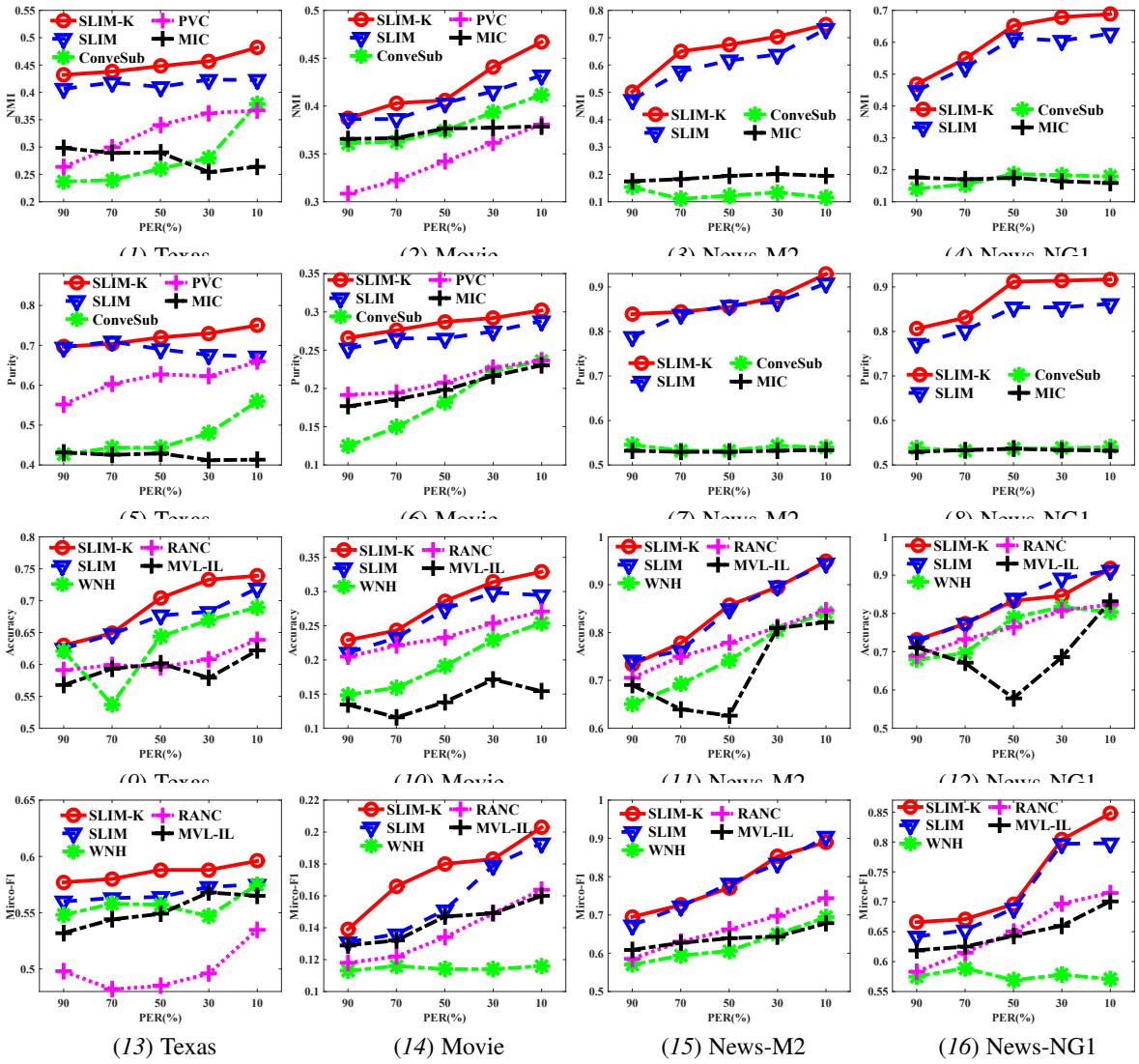


Fig. 2. The NMI, Purity, Accuracy, Mirco-F1 results of the Texas, Movie, News-M2, News-NG1 with only missing one modality for each instance. PER (partial example ratio) is the ratio of incomplete examples.

Thus, to evaluate the performance of our proposed approaches. We choose 3 state-of-the-art multi-modal methods to evaluate semi-supervised clustering task: ConvexSub [26]; PVC [13]; MIC [14]. Considering the limitation of the compared clustering method, we first learn a latent representation of the original data, and use the semi-supervised K-means to get clustering result. For classification task, we compare the WNH [27], RANC [28] and MVL-IL [15]. Since some methods cannot handle incomplete examples, i.e., ConvexSub, WNH, RANC, for fair comparison, we facilitate with the ALM (Augmented Lagrange Multipliers) [29] matrix completion method to fill in the missing information of the partial examples. In detail, **ConvexSub**: construct a subspace-based multi-modal clustering; **PVC**: establish a latent subspace to make different incomplete modalities are close to each other; **MIC**: learn the latent feature matrices for different incomplete modalities and a consensus matrix, by minimizing the difference between each modal matric and the consensus matrix; **WNH**: combine all modal values together and then uses $l_{2,1}$-norm to regularize the modal selection process; **RANC**: convert each modal predicted values into an accumulated prediction matrix with low-rank constraint; **MVL-IL**: exploit the connections among multiple modalities to handle the incomplete modalities, and estimates the incomplete modalities by integrating the information from the other observed modalities through this subspace.

4.1 Experiment Results

First, we evaluate our algorithm with fix incomplete ratio, then evaluate the influence of incomplete ratio.

4.1.1 *Semi-Supervised Clustering/Classification*

To demonstrate the effectiveness of our proposed method, we designed two missing settings. First, we set each incomplete instance only miss one modality, and we record both the clustering results in Table 1 and the classification results in Table 2. Moreover, we record each modal classification performance of binary datasets in Table 3. Meanwhile we set each incomplete instance missing modalities with the probability of $\frac{1}{K}$ randomly, and record the clustering and classification results in Table 5, Table 6. For all datasets, we fix the incomplete ratio as 90%. To further validate the effectiveness of the proposed methods, we experiment on 2 real-world datasets, i.e., 3Sources (3S.) and WKG, and record the

TABLE 5
Clustering comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on 14 benchmark datasets with missing modalities randomly, the ratio of the multiple incomplete modal data is 90%. 2 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller the better of a criterion.

Data	Purity ↑					NMI ↑				
	SLIM-K	SLIM	ConvexSub	PVC	MIC	SLIM-K	SLIM	ConvexSub	PVC	MIC
Mov.	**.343±.017**	.341±.018	.184±.010	.275±.016	.150±.012	**.372±.047**	.358 ±.018	.168±.010	.330±.013	.149±.008
Cite.	**.398±.029**	.362±.024	.342±.032	.356±.043	.250±.031	**.188±.046**	.143 ±.013	.133±.043	.096±.027	.081±.032
Cora	**.410±.073**	.408±.035	.383±.042	.378±.036	.328±.018	**.196±.076**	.195 ±.032	.177±.043	.120±.038	.067±.015
Corn.	**.501±.052**	.499±.058	.472±.029	.477±.031	.420±.017	**.181±.047**	.138 ±.035	.082±.023	.111±.024	.074±.018
Texas	**.687±.042**	.675±.025	.583±.026	.605±.030	.555±.011	**.188±.070**	.151 ±.087	.085±.053	.139±.055	.068±.021
Wash.	**.664±.086**	.653±.045	.534±.048	.627±.065	.495±.017	**.234±.069**	.219 ±.042	.012±.056	.202±.072	.079±.030
Wis.	**.673±.053**	.654±.018	.527±.050	.595±.041	.529±.048	**.223±.073**	.193 ±.075	.110±.057	.179±.044	.138±.028
M2	**.644±.032**	.621±.038	.500±.018	-	.510±.024	**.063±.046**	.050 ±.030	.003±.004	-	.022±.014
M5	**.353±.042**	.335±.029	.243±.016	-	.250±.020	**.077±.046**	.063 ±.026	.028±.017	-	.052±.019
M10	**.400±.036**	.386±.019	.246±.020	-	.199±.010	**.175±.049**	.156 ±.027	.140±.017	-	.129±.017
NG1	.532±.017	**.591±.040**	.504±.003	-	.500±.008	**.086±.029**	.073 ±.033	.059±.004	-	.060±.008
NG2	**.372±.032**	.337±.027	.223±.022	-	.244±.011	**.091±.051**	.084 ±.020	.034±.016	-	.032±.008
NG3	**.413±.028**	.395±.028	.230±.020	-	.196±.017	**.198±.044**	.183 ±.036	.150±.026	-	.072±.012
Reut.	**.502±.055**	.474±.032	.284±.041	-	.310±.038	**.192±.048**	.169 ±.029	.113±.020	-	.119±.031

TABLE 6
Classification comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on 14 benchmark datasets with missing modalities randomly, the ratio of the multiple incomplete modal data is 90%. 2 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller the better of a criterion.

Data	Accuracy ↑					F1 ↑				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
Mov.	**.204±.032**	.188±.044	.107±.042	.116±.031	.082±.022	**.144±.027**	.126 ±.018	.104±.019	.094±.047	.115±.004
Cite.	**.614±.040**	.596±.015	.383±.114	.457±.076	.333±.025	**.441±.078**	.410 ±.017	.299±.012	.303±.029	.349±.016
Cora	**.571±.058**	.557±.013	.450±.085	.368±.056	.281±.057	**.424±.077**	.389 ±.014	.326±.038	.259±.036	.309±.024
Corn.	**.583±.026**	.554±.079	.446±.138	.446±.091	.427±.069	**.478±.098**	.471 ±.070	.422±.093	.372±.065	.440±.041
Texas	**.693±.071**	.674±.041	.518±.142	.557±.073	.435±.164	**.656±.073**	.643 ±.037	.541±.066	.469±.068	.512±.082
Wash.	**.750±.063**	.696±.042	.509±.232	.584±.075	.441±.125	.616±.086	**.622±.043**	.527±.056	.505±.065	.513±.075
Wis.	**.714±.091**	.679±.058	.529±.190	.570±.056	.433±.089	**.588±.053**	.585±.079	.502±.098	.461±.054	.479±.065
M2	**.719±.083**	.717±.072	.691±.058	.687±.122	.559±.036	**.665±.056**	.647 ±.040	.614±.035	.594±.079	.614±.015
M5	**.524±.109**	.490±.115	.407±.045	.447±.067	.299±.024	**.381±.035**	.376 ±.043	.301±.033	.292±.047	.333±.008
M10	**.381±.014**	.338±.060	.155±.072	.244±.062	.141±.020	**.237±.036**	.201 ±.025	.170±.011	.164±.030	.180±.006
NG1	**.737±.031**	.702±.044	.698±.055	.642±.075	.563±.033	**.648±.043**	.629 ±.019	.621±.032	.547±.040	.612±.015
NG2	**.636±.083**	.585±.069	.337±.096	.409±.144	.271±.039	**.467±.058**	.430 ±.058	.307±.020	.313±.069	.320±.012
NG3	**.500±.094**	.470±.050	.205±.059	.319±.065	.205±.023	**.288±.040**	.279 ±.035	.206±.023	.202±.019	.253±.009
Reut.	**.429±.031**	.417±.061	.383±.054	.304±.074	.248±.022	**.320±.027**	.306 ±.026	.283±.030	.211±.049	.303±.006

results in Table 7. Note that PVC method can only leverage two modalities, we did not compare our methods with PVC for multi-modalities.

Table 1 and Table 5 reveal that for both two modal and multiple modal datasets in two incomplete setting, SLIM/SLIM-K almost consistently achieve the significant superior clustering performance on either purity or NMI, except for Wisconsin on purity. It is owing to that in SLIM/SLIM-K, the similarity matrices are initialized with cosine similarity for more robust generalization, rather than task-specific similarity matrix construction method. Besides, Table 2 and Table 6 show that SLIM/SLIM-K also achieve the best prediction performance on all datasets. In Table 7, it further reveals that SLIM/SLIM-K are superior than other compare methods in real applications.

These phenomenons reveal the effectiveness of considering the high order consistencies between different modal similarity matrixes and the learned prediction. On the other hand, Table 3 reveals that SLIM/SLIM-K can get superior performance whether on single modality or overall result. It is notable that SLIM-K is better than SLIM with the complemented kernel matrix for classification and clustering. The phenomenon indicates that SLIM-K can learn more discriminative classifiers and cluster learner.

4.1.2 *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 and SLIM-K are set 1, while the ratio of the incomplete data varies in $\{90\%, 70\%, \cdots, 10\%\}$ with 20% intervals. Due to the page limits, results on 4 datasets, i.e., Texas, Movie, News-M2, News-NG1, and the results of NMI, purity, accuracy, and Mirco-F1 with two incomplete setting are recorded in Fig. 2 and 3. These figures clearly show that SLIM-K achieves the

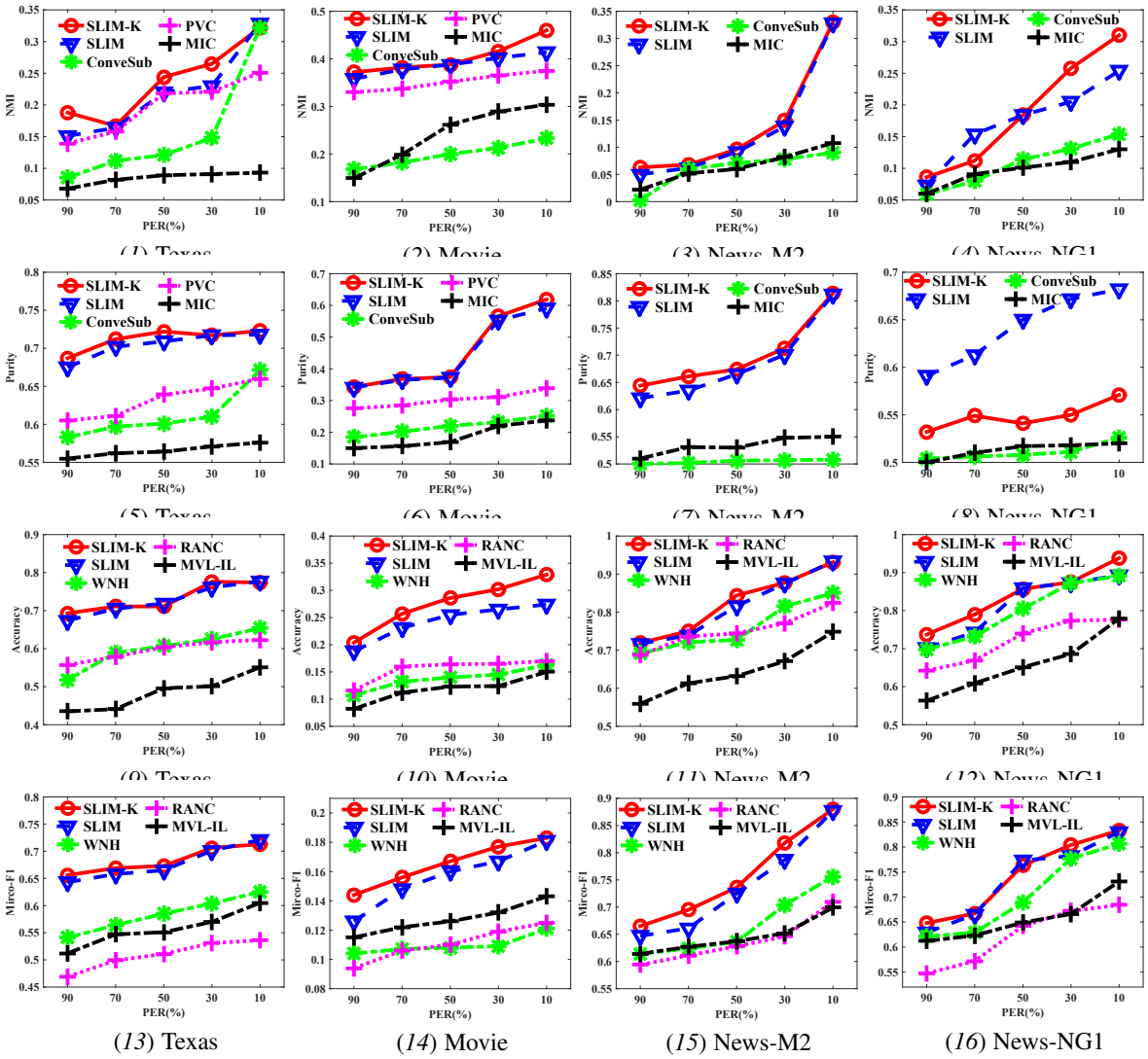


Fig. 3. The NMI, Purity, Accuracy, Mirco-F1 results of the Texas, Movie, News-M2, News-NG1 with missing modalities randomly. PER (partial example ratio) is the ratio of incomplete examples.

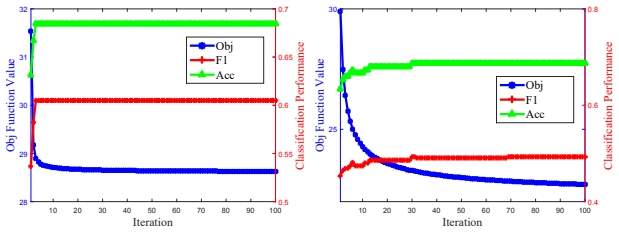


Fig. 5. Objective function value convergence and corresponding classification performance (Accuracy, F1) vs. number of iterations.

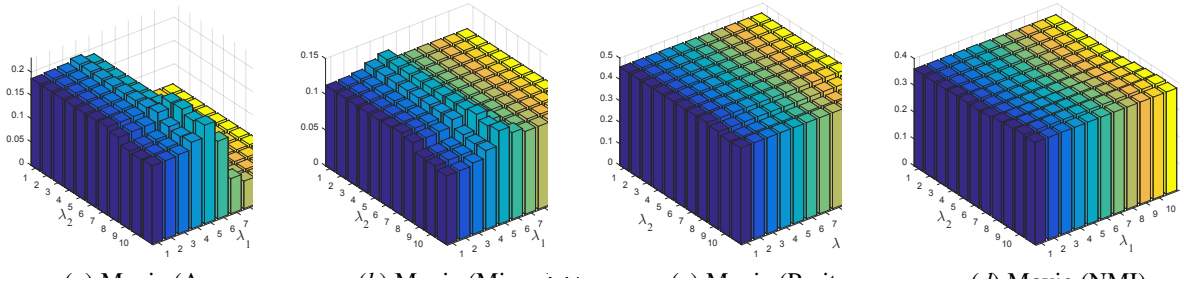
competitive on almost all datasets except purity on News-NG1 dataset. We also find that SLIM-K achieves superiorities from high incomplete ratio, and SLIM-K increases faster when incomplete ratio decreases.

4.1.3 Empirical Investigation on Convergence

To investigate the convergence empirically, the objective function value, i.e., the value of Eq. 7 and the classification performance in each iteration are recorded. Due to the page limits, we plot results of only 2 datasets in Fig. 5. It clearly reveals that the objective function value decreases as the iterations increase, and the classification performance becomes stable after several iterations on different datasets. Moreover, additional experiments result indicates that our methodes converge very fast, i.e., on most datasets, SLIM-K converges after 10 rounds.

4.1.4 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 $\{10^{-5}, 10^{-4}, \cdots, 10^{4}, 10^{5}\}$, then we fix the λ_2 while tuning λ_1 in $\{10^{-5}, 10^{-4}, \cdots, 10^{4}, 10^{5}\}$, and record the average performance in Fig. 4. Due to the page limits, we only list 2 datasets for verification, i.e., Movie, News-M2. These figures show that SLIM-K achieves a stable performance on each dataset except accuracy with large λ_1, i.e., (a) and (e) in Fig. 4, which indicates the insensitivity of SLIM-K to parameters.



(*e*) News-M2 (Accuracy) (*f*) News-M2 (Mirco-F1) (*g*) News-M2 (Purity) (*h*) News-M2 (NMI)

Fig. 4. Influence of the parameters λ_1, λ_2 on the 2 datasets, i.e., Movie, News-M2, the ratio of the multiple incomplete modal data is 90%.

TABLE 7
Clustering/Classification comparison results (mean and std.) of SLIM and SLIM-K with both compared methods on real-world incomplete datasets.

Data	Purity ↑					NMI ↑				
	SLIM-K	SLIM	ConvexSub	PVC	MIC	SLIM-K	SLIM	ConvexSub	PVC	MIC
3S.	**.880±.020**	.858±.014	.282±.009	-	.389±.019	**.849±.026**	.801±.019	.236±.010	N/A	.401±.019
WKG	**.673±.019**	.669 ±.054	.543±.008	.565 ±.011	.532 ±.009	**.149±.050**	.098 ±.052	.041 ±.028	.016±.006	.021 ±.005
Data	Accuracy ↑					F1 ↑				
	SLIM-K	SLIM	WNH	RANC	MVL-IL	SLIM-K	SLIM	WNH	RANC	MVL-IL
3S.	**.850±.025**	.828±.040	.735±.083	.546±.144	.263±.044	.849±.029	**.854±.014**	.608±.100	.424±.031	.337±.019
WKG	.673±.022	**.678 ±.042**	.647±.155	.648 ±.019	.470 ±.028	**.739±.017**	.711 ±.029	.674 ±.095	.626±.009	.630 ±.043

5 Conclusion

This paper focuses on the issues of incomplete multi-modal learning, which extends our preliminary research [30]. Previous mainstream solutions alleviated 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 complementary information from extrinsic data, and thus form transductive solutions. In this paper, we proposed novel semi-supervised incomplete multi-modal approach, with more extrinsic information exploited from unlabeled data, and yielded an inductive learner which consequently can be applied in general multi-modal circumstances. With the complemented kernel matrix, SLIM-K can get higher performance. Therefore, SLIM and SLIM-K can be easily adopted to either classification or clustering tasks. Experimental evaluations on real-world applications demonstrate the superiority of our proposed method.

Appendix A
Comparison with Different Feature Embedding

It is notable that the Citeseer and Cora datasets are graph datasets, while in this paper, following [13], we only use the primitive features of these two datasets. To explore effectiveness of the feature embedding designed for special raw data, i.e., graph convolution network for the graph datasets, we conduct more experiments comparing with the state-of-the-art semi-supervised GCN style methods: GCN [31] and MixHop [32]. GCN utilized the convolutional networks for graph data considering the structure information on each hidden layer, MixHop learned a general class of neighborhood mixing relationships with specific sparsity regularization, and achieved the top results on these datasets. Similarly, in each split, we randomly select 10% to 90% of examples with incomplete modalities, with 20% as interval, and the remains are complete instances. Note that we use the GCN feature embedding as the input for our methods, and the results are recorded in Table 8, Table 9, Table 10 and Table 11. The results reveal that with more discriminative feature embedding, the clustering/classification performances of SLIM and SLIM-K are better than primitive features, while are worse than the MixHop method on some criteria, especially in the case of low missing rates, i.e., 10%, 30%, 50%. This is because MixHop can better consider the neighbor relationships in the case of lower missing rate, while SLIM/SLIM-K simply use linear kernel to represent the neighbor relationships between instances. On the other hand, GCN methods can learn more discriminative features by using end-to-end deep network, while our methods concentrate on building

TABLE 8
Clustering/Classification comparison results of SLIM and SLIM-K with GCN style methods on Citeseer dataset.

Ratio	Purity ↑			NMI ↑		
	SLIM-K	SLIM	GCN	SLIM-K	SLIM	GCN
10%	**.833±.009**	.829±.007	.785±.008	**.610±.015**	.605±.012	.532±.015
30%	**.823±.011**	.820±.012	.774±.010	**.594±.019**	.589±.019	.515±.014
50%	.825±.010	**.828±.005**	.773±.008	.596±.015	**.600±.011**	.511±.010
70%	**.828±.006**	.825±.004	.773±.007	**.598±.010**	.594±.006	.509±.009
90%	**.824±.107**	.818±.073	.764±.003	**.592±.101**	.585±.071	.497±.007
Ratio	Accuracy ↑			F1 ↑		
	SLIM-K	SLIM	GCN	SLIM-K	SLIM	GCN
10%	**.719±.012**	.712±.011	.709±.007	**.566±.016**	.560±.016	.552±.012
30%	**.706±.010**	.697±.014	.694±.012	**.547±.016**	.539±.020	.532±.017
50%	**.693±.009**	.686±.013	.687±.010	**.534±.011**	.525±.015	.526±.013
70%	**.682±.011**	.677±.011	.674±.013	**.516±.014**	.511±.015	.506±.019
90%	**.678±.015**	.667±.092	.664±.019	**.510±.017**	.503±.017	.500±.023

TABLE 9
Clustering/Classification comparison results of SLIM and SLIM-K with GCN style methods on Citeseer dataset.

Ratio	Purity ↑			NMI ↑		
	SLIM-K	SLIM	mixhop	SLIM-K	SLIM	mixhop
10%	.765±.012	**.801±.008**	.743±.013	.500±.019	**.553±.012**	.469±.016
30%	.739±.009	**.776±.009**	.689±.015	.475±.013	**.515±.015**	.404±.019
50%	**.697±.017**	.678±.035	.630±.010	**.434±.013**	.409±.039	.354±.013
70%	**.652±.015**	.609±.032	.552±.012	**.377±.017**	.359±.034	.315±.023
90%	**.551±.014**	.480±.018	.387±.013	**.282±.026**	.218±.022	.143±.027
Ratio	Accuracy ↑			F1 ↑		
	SLIM-K	SLIM	mixhop	SLIM-K	SLIM	mixhop
10%	.575±.035	.647±.028	**.652±.016**	.396±.033	.473±.034	**.481±.022**
30%	.592±.021	.603±.019	**.604±.015**	.408±.022	.419±.020	**.422±.016**
50%	**.542±.040**	.540±.022	.537±.021	**.363±.037**	.359±.017	.356±.017
70%	**.476±.031**	.465±.037	.454±.015	**.319±.021**	.317±.015	.305±.009
90%	**.406±.029**	.367±.015	.341±.008	**.280±.013**	.277±.011	.268±.016

TABLE 10
Clustering/Classification comparison results of SLIM and SLIM-K with GCN style methods on Cora dataset.

Ratio	Purity ↑			NMI ↑		
	SLIM-K	SLIM	GCN	SLIM-K	SLIM	GCN
10%	**.896±.008**	.893±.008	.879±.012	**.741±.014**	.735±.014	.710±.023
30%	**.892±.006**	.879±.008	.866±.019	**.716±.012**	.706±.015	.689±.034
50%	**.860±.013**	.860±.009	.844±.024	**.676±.020**	.675±.015	.645±.046
70%	.819±.028	**.849±.008**	.798±.023	.611±.029	**.653±.017**	.563±.040
90%	.779±.048	**.824±.016**	.722±.047	.568±.039	**.617±.011**	.470±.046
Ratio	Accuracy ↑			F1 ↑		
	SLIM-K	SLIM	GCN	SLIM-K	SLIM	GCN
10%	**.853±.018**	.839±.013	.824±.014	.653±.031	**.727±.019**	.706±.021
30%	**.822±.032**	.793±.032	.808±.030	**.706±.050**	.657±.048	.688±.038
50%	**.803±.037**	.797±.044	.778±.045	**.684±.051**	.663±.062	.638±.061
70%	.722±.035	**.729±.056**	.721±.046	.560±.037	**.569±.069**	.567±.059
90%	**.684±.030**	.669±.029	.658±.022	**.508±.033**	.484±.031	.481±.035

TABLE 11
Clustering/Classification comparison results of SLIM and SLIM-K with GCN style methods on Cora dataset.

Ratio	Purity ↑			NMI ↑		
	SLIM-K	SLIM	mixhop	SLIM-K	SLIM	mixhop
10%	.802±.009	**.835±.016**	.786±.015	.573±.016	**.637±.021**	.549±.023
30%	.775±.011	**.788±.029**	.740±.020	.536±.015	**.569±.025**	.494±.025
50%	**.745±.010**	.708±.044	.695±.021	.494±.016	**.500±.023**	.451±.025
70%	**.684±.011**	.611±.014	.598±.034	**.423±.010**	.412±.018	.372±.033
90%	**.621±.017**	.512±.004	.496±.011	**.336±.024**	.333±.006	.259±.019
Ratio	Accuracy ↑			F1 ↑		
	SLIM-K	SLIM	mixhop	SLIM-K	SLIM	mixhop
10%	.643±.022	.575±.022	**.706±.015**	.463±.025	.384±.021	**.548±.023**
30%	.615±.019	.560±.023	**.661±.021**	.435±.021	.374±.020	**.487±.024**
50%	.585±.012	.561±.025	**.608±.022**	.412±.020	.375±.021	**.421±.025**
70%	**.550±.016**	.521±.026	.517±.029	**.392±.017**	.347±.020	.342±.025
90%	**.485±.042**	.472±.030	.451±.008	**.346±.019**	.327±.023	.325±.005

predictors with given feature representations. On the contrary, in the case of high missing ratio, we have achieved better results by using the complementarity and consistency among the modalities. Actually, how to expand SLIM and SLIM-K on specific data structures is an interesting future research direction.

ACKNOWLEDGMENT

This research was supported by National Key R&D Program of China (2018YFB1004300), NSFC (61773198, 61632004, 61751306), NSFC-NRF Joint Research Project under Grant 61861146001, NSF IIS-1814510, and Collaborative Innovation Center of Novel Software Technology and Industrialization, Postgraduate Research & Practice Innovation Program of Jiangsu province (KYCX18-0045).

REFERENCES

[1] Y. Yang, H.-J. Ye, D.-C. Zhan, and Y. Jiang, "Auxiliary information regularized machine for multiple modality feature learning," in *IJCAI*, Buenos Aires, Argentina, 2015, pp. 1033–1039.

[2] R. Arora, P. Mianjy, and T. V. Marinov, "Stochastic optimization for multiview representation learning using partial least squares," in *ICML*, New York City, NY, 2016, pp. 1786–1794.

[3] T. Iwata and M. Yamada, "Multi-view anomaly detection via robust probabilistic latent variable models," in *NIPS*, Barcelona, Spain, 2016, pp. 1136–1144.

[4] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *AAAI*, San Francisco, California, 2017, pp. 1618–1625.

[5] X. Shen, F. Shen, Q. Sun, Y. Yang, Y. Yuan, and H. T. Shen, "Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval," *IEEE Trans. Cybernetics*, vol. 47, no. 12, pp. 4275–4288, 2017.

[6] H. Liu, M. Lin, S. Zhang, Y. Wu, F. Huang, and R. Ji, "Dense auto-encoder hashing for robust cross-modality retrieval," in *MM*, Seoul, Republic of Korea, 2018, pp. 1589–1597.

[7] A. Trivedi, P. Rai, H. Daume, and S. L. Duvall, "Multiview clustering with incomplete views," 2010.

[8] W. Shao, X. Shi, and S. Y. Philip, "Clustering on multiple incomplete datasets via collective kernel learning," in *ICDM*, Dallas, TX, 2013, pp. 1181–1186.

[9] W. Shao, L. He, C. Lu, and P. S. Yu, "Online multi-view clustering with incomplete views," in *IEEE Big Data*, Washington DC, 2016, pp. 1012–1017.

[10] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping," in *IJCAI*, New York, NY, 2016, pp. 2392–2398.

[11] M. Karasuyama and H. Mamitsuka, "Multiple graph label propagation

by sparse integration," *NN*, vol. 24, no. 12, pp. 1999–2012, 2013.
[12] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *ICCV*, Sydney, Australia, 2013, pp. 1737–1744.
[13] S. Li, Y. Jiang, and Z. Zhou, "Partial multi-view clustering," in *AAAI*, Quebec, Canada, 2014, pp. 1968–1974.
[14] W. Shao, L. He, and P. S. Yu, "Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization," in *ECML/PKDD*, Porto, Portugal, 2015, pp. 318–334.
[15] C. Xu, D. Tao, and C. Xu, "Multi-view learning with incomplete views," *TIP*, vol. 24, no. 12, pp. 5812–5825, 2015.
[16] E. Eaton, M. desJardins, and S. Jacob, "Multi-view clustering with constraint propagation for learning with an incomplete mapping between views," in *CIKM*, Ontario, Canada, 2010, pp. 389–398.
[17] Q. Yin, S. Wu, and L. Wang, "Unified subspace learning for incomplete and unlabeled multi-view data," *PR*, vol. 67, pp. 313–327, 2017.
[18] A. Mollaysa, P. Strasser, and A. Kalousis, "Regularising non-linear models using feature side-information," in *ICML*, 2017, pp. 2508–2517.
[19] C. H. Q. Ding and X. He, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *ICDM*, Newport Beach, CA, 2005, pp. 606–610.
[20] H. Liu, L. Wang, and T. Zhao, "Multivariate regression with calibration," in *NIPS*, Quebec, Canada, 2014, pp. 127–135.
[21] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *JMLR*, vol. 12, pp. 1149–1184, 2011.
[22] D. S. Rosenberg and P. L. Bartlett, "The rademacher complexity of co-regularized kernel classes," in *AISTATS*, San Juan, Puerto Rico, 2007, pp. 396–403.
[23] M. Xu, R. Jin, and Z. Zhou, "CUR algorithm for partially observed matrices," in *ICML*, Lille, France, 2015, pp. 1412–1421.
[24] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, vol. 29, no. 3, pp. 93–106, 2008.
[25] G. Bisson and C. Grimal, "Co-clustering of multi-view datasets: A parallelizable approach," in *ICDM*, Brussels, Belgium, 2012, pp. 828–833.
[26] Y. Guo, "Convex subspace representation learning from multi-view data," in *AAAI*, Bellevue, Washington, 2013, pp. 387–393.
[27] H. Wang, F. Nie, and H. Huang, "Multi-View Clustering and Feature Learning via Structured Sparsity," in *ICML*, Atlanta, GA, 2013, pp. 352–360.
[28] H. Ye, D. Zhan, Y. Miao, Y. Jiang, and Z. Zhou, "Rank consistency based multi-view learning: A privacy-preserving approach," in *CIKM*, Melbourne, Australia, 2015, pp. 991–1000.
[29] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *CoRR*, vol. abs/1009.5055, 2010.
[30] Y. Yang, D. Zhan, X. Sheng, and Y. Jiang, "Semi-supervised multi-modal learning with incomplete modalities," in *IJCAI*, Stockholm, Sweden, 2018, pp. 2998–3004.
[31] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[32] S. Abu-El-Haija, B. Perozzi, A. Kapoor, H. Harutyunyan, N. Alipourfard, K. Lerman, G. V. Steeg, and A. Galstyan, "Mixhop: Higher-order graph convolution architectures via sparsified neighborhood mixing," *arXiv preprint arXiv:1905.00067*, 2019.

Yang Yang is working towards the PhD degree with the National Key Lab for Novel Software Technology, the Department of Computer Science & Technology in Nanjing University, China. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining.

De-Chuan Zhan received the Ph.D. degree in computer science, Nanjing University, China in 2010. At the same year, he became a faculty member in the Department of Computer Science and Technology at Nanjing University, China. He is currently an Associate Professor with the Department of Computer Science and Technology at Nanjing University. His research interests are mainly in machine learning, data mining and mobile intelligence. He has published over 30 papers in leading international journal/conferences. He serves as an editorial board member of IDA and IJAPR, and serves as SPC/PC in leading conferences such as IJCAI, AAAI, ICML, NIPS, etc.

Yi-Feng Wu is working towards the M.Sc. degree with the National Key Lab for Novel Software Technology, the Department of Computer Science & Technology in Nanjing University, China. His research interests lie primarily in machine learning and data mining, including multi-modal learning.

Zhi-Bin Liu received the Ph.D. degree and M.S. degree in control science and engineering from Tsinghua Universtiy, Beijing, China, in 2010, and the B.S. degree in automatic control engineering from Central South University, Changsha, China, in 2004. His research interests are in big data minning, machine learning, AI, NLP, computer vision, information fusion and etc.

Hui Xiong (SM'07) is currently a full professor at Rutgers, the State University of New Jersey, where he received the ICDM-2011 Best Research Paper Award, and the 2017 IEEE ICDM Outstanding Service Award. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published prolifically in refereed journals and conference proceedings (4 books, more than 80 journal papers, and more than 100 conference papers). He is a co-editor-in-chief of Encyclopedia of GIS, an associate editor of the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Big Data, the ACM Transactions on Knowledge Discovery from Data, and the ACM Transactions on Management Information Systems. He has served regularly on the organization and program committees of numerous conferences, including as a program cochair of the Industrial and Government Track for KDD-2012, a program co-chair for ICDM-2013, a general co-chair for ICDM-2015, and a program co-chair of the Research Track for KDD-2018. For his outstanding contributions to data mining and mobile computing, he was elected an ACM distinguished scientist in 2014. He is a senior member of the IEEE.

Yuan Jiang received the PhD degree in computer science from Nanjing University, China, in 2004. At the same year, she became a faculty member in the Department of Computer Science & Technology at Nanjing University, China and currently is a Professor. She was selected in the Program for New Century Excellent talents in University, Ministry of Education in 2009. Her research interests are mainly in artificial intelligence, machine learning, and data mining. She has published over 50 papers in leading international/national journals and conferences.