Joint Multi-Source Reduction

Lei Zhang[1][0000−0002−5487−557X](✉), Shupeng Wang[1](✉), Xin Jin[2](✉), and Siyu Jia[1](✉)

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
{zhanglei1, wangshupeng, jiasiyu}@iie.ac.cn
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
13911191965@139.com

Abstract. The redundant sources problem in multi-source learning always exists in various real-world applications such as multimedia analysis, information retrieval, and medical diagnosis, in which the heterogeneous representations from different sources always have three-way redundancies. More seriously, the redundancies will cost a lot of storage space, cause high computational time, and degrade the performance of learner. This paper is an attempt to jointly reduce redundant sources. Specifically, a novel Heterogeneous Manifold Smoothness Learning (HMSL) model is proposed to linearly map multi-source data to a low-dimensional feature-isomorphic space, in which the information-correlated representations are close along manifold while the semantic-complementary instances are close in Euclidean distance. Furthermore, to eliminate three-way redundancies, we present a new Correlation-based Multi-source Redundancy Reduction (CMRR) method with 2,1-norm equation and generalized elementary transformation constraints to reduce redundant sources in the learned feature-isomorphic space. Comprehensive empirical investigations are presented that confirm the promise of our proposed framework.

Keywords: Multi-source · redundant · heterogeneous · manifold measure · dimension reduction · sample selection.

1 Introduction

Generally, due to incorrect data storage manner and the like, not all instances are a concise and effective reflection of objective reality, inevitably leading to the redundant sources of multi-source data. Note that different from duplicated data, multi-source heterogeneous redundant data are those which could seriously affect the performance of the learner. Rather, as shown in Fig.1, there is a distinct difference between the redundant sources problem in multi-source learning and mono-source scenario, because multi-source heterogeneous redundant data contain the following three-way redundancies:

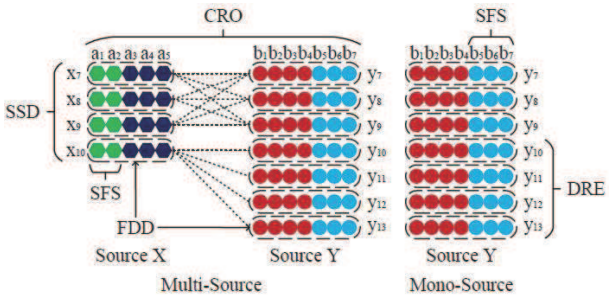

Fig. 1. Multi- and Mono-Source Redundant Data. The x_7, x_8, x_9, and x_{10} denote the redundant representations from Source S_x. Similarly, the y_7, y_8, y_9, y_{10}, y_{11}, y_{12}, and y_{13} are the redundant representations from Source S_y. The a_1, a_2, a_3, a_4, and a_5 represents the features in the representations from Source S_x. The features in the representations from Source S_y are composed of the b_1, b_2, b_3, b_4, b_5, b_6, and b_7. The three-way redundancies are DRE, SFS, and CRO, respectively. The double-level heterogeneities consist of FDD and SSD.

- **Data Representations Excessiveness (DRE).** The existing of multiple unduplicated representations of the same object in the same source leads to taking up too much storage space.
- **Sample Features Superabundance (SFS).** Superabundance caused by curse of dimensionality [4] refers to a high-dimensional space embedding some related or randomized dimensions, resulting in high computational time.
- **Complementary Relationships Overplus (CRO).** One representation from one source has corresponding relationships with multiple heterogeneous descriptions from another source. This overplus will bring about a significant decline in the performances of multi-source representations.

Consequently, due to the existing of three-way redundancies, the redundant sources problem owns double-level heterogeneities, i.e., Feature Dimension Dissimilarity (FDD) and Sample Size Difference (SSD) (see Fig.1). First, different sources use different dimensions and different attributes to represent the same object [29, 10, 14]; besides, there are different number of instances in each source. Even more serious is that these redundancies severely impact the performances of multi-source data, resulting in false analysis, clustering, classification, and retrieval [24, 25, 12]. Therefore, it is extremely necessary to develop an effective reducing method for multi-source heterogeneous redundant data.

For the past few years, to deal with redundancies problem, various machine learning methods have been investigated to reduce computational cost and improve learning accuracy. Up to now, the existing methods involve dimension reduction techniques[18, 17, 24, 13] and sample selection approaches [25, 5, 21, 22].

Dimension Reduction Techniques In [18], Huan et al. investigated a feature extraction approach, called Knowledge Transfer with Low-Quality Data

(KTLQD), to leverage the available auxiliary data sources to aid in knowledge discovery. Nie et al. [17] proposed an Efficient and Robust Feature Selection via Joint $\ell_{2,1}$-Norms Minimization (ERFSJNM) method, which used $\ell_{2,1}$-norm regularization to extract meaningful features and eliminate noisy ones across all data points with joint sparsity. Wang et al. [24] studied a feature selection framework, called Feature Selection via Global Redundancy Minimization (FSGRM), to globally minimize the feature redundancy with maximizing the given feature ranking scores. An unsupervised feature selection scheme, namely, Nonnegative Spectral Analysis with Constrained Redundancy (NSACR), was developed by Li et al. [13] through jointly leveraging nonnegative spectral clustering and redundancy analysis.

Sample Selection Approaches Wang et al. [25] proposed a sample selection mechanism based on the principle of maximal classification ambiguity, i.e., Maximum Ambiguity-based Sample Selection in Fuzzy Decision Tree Induction (MASSFDTI), to select a number of representative samples from a large database. In [5], a Sample Pair Selection with Rough Set (SPSRS) framework was proposed in order to compress the discernibility function of a decision table so that only minimal elements in the discernibility matrix were employed to find reducts. Shahrian and Rajan [21] designed a content-based sample selection method, called Weighted Color and Texture Sample Selection for Image Matting (WCTSSIM), in which color information was leveraged by color sampling-based matting methods to find the best known samples for foreground and background color of unknown pixels. Su et al. [22] developed an Active Correction Propagation (ACP) method using a sample selection criterion for active query of informative samples by minimizing the expected prediction error.

Generally, these existing methods can eliminate only one kind of redundancy, not three kinds of redundancy. Moreover, these methods were designed for single-source data like many other conventional data mining methods. Accordingly, it is impossible for them to eliminate the double-level heterogeneities among different redundant sources. To address the limitations of existing methods, we attempt to explore a multi-source reducing framework to jointly eliminate three-way redundancies and double-level heterogeneities at the same time.

1.1 Organization

The remainder of this paper is organized as follows: A general framework for jointly reducing the redundant sources of multi-source data is proposed in Section 2. Furthermore, the efficient algorithms are provided to solve the proposed framework in Section 3. Section 4 evaluates and analyzes the proposed framework on three multi-source datasets. Finally, our conclusions are presented in Section 5.

1.2 Notations

In Table 2, we describe the notations needed to understand our proposed algorithm.

Table 1. NOTATIONS

Notation	Description
S_x	Source X
S_y	Source Y
$X_N \in \mathbb{R}^{n_1 \times d_x}$	Non-redundant samples in S_x
$Y_N \in \mathbb{R}^{n_1 \times d_y}$	Non-redundant samples in S_y
$L_N \in \mathbb{R}^{n_1 \times m}$	Label indicator matrix
$x_i \in \mathbb{R}^{d_x}$	The i-th sample from S_x
$y_i \in \mathbb{R}^{d_y}$	The i-th sample from S_y
n_1	Number of non-redundant samples
d_x	Dimensionality of S_x
d_y	Dimensionality of S_y
m	Number of labels
(x_i, y_i)	The i-th multi-source datum
$X_R \in \mathbb{R}^{n_2 \times d_x}$	Redundant representations in S_x
$Y_R \in \mathbb{R}^{n_3 \times d_y}$	Redundant representations in S_y
n_2	Number of redundant samples in S_x
n_3	Number of redundant samples in S_y
$\|\| \cdot \|\|_F$	Frobenius norm
$\|\| \cdot \|\|_*$	Trace norm
$\mathbb{S}_+^{k \times k}$	Positive semi-definite matrices
$\nabla f(\cdot)$	Gradient of smooth function $f(\cdot)$
$\| \cdot \|$	Absolute value
$I_k \in \mathbb{R}^{k \times k}$	Identity matrix

2 Reducing Multi-Source Heterogeneous Redundant Data

A general simplifying framework is proposed in this section to jointly reduce the redundant sources of multi-source data. Fig.2 presents an overview of the proposed framework. In this example, a set of multi-source data consists of Source X and Source Y. There are a certain amount of multi-source non-redundant data such as X_N and Y_N. However, some multi-source data X_R and Y_R have three-way redundancies and double-level heterogeneities. For instance, the CRO among different sources brings about that the sample x_7 in Source X is correlated with multiple instances y_7, y_8, and y_9 in Source Y; additionally, there are multiple representations y_{11}, y_{12}, and y_{13} similar to y_{10} due to the DRE in Source Y; furthermore, the representations in the sources are too much superabundant because of the SFS. As a result, the feature dimensions are heterogeneous and there are different number of samples in these sources, i.e., FDD and SSD.

To jointly reduce the redundant sources of multi-source data, HMSL model learns a low-dimensional feature homogeneous subspace, in which the information-correlated representations are close along manifold while the semantic-complementary instances are close in Euclidean distance at the same time. Then, CMRR model

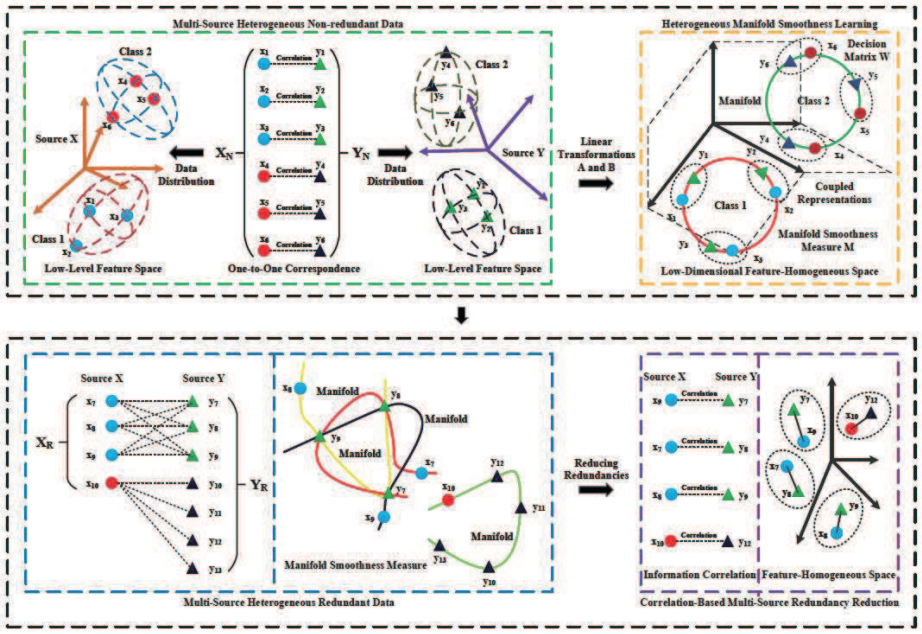

Fig. 2. Framework for Joint Multi-Source Reduction.

removes the three-way redundancies and double-level heterogeneities existing in the multi-source heterogeneous redundant data X_R and Y_R from the feature-homogeneous space under the learned complementarity, correlation, and distributivity restraints.

The following subsections present more details.

2.1 The Proposed HMSL Model

This subsection presents a new HMSL model, which has pseudo-metric constraints, manifold regularization, and leave-one-out validation to correlate different sources. In HMSL model, the existing non-redundant heterogeneous representations X_N and Y_N are utilized to learn two heterogeneous linear transformations A and B, a decision matrix W, and a manifold smoothness measure M to mine the semantic complementarity, information correlation, and distributional similarity among different sources. As a consequence, the heterogeneous representations from different sources are linearly mapped into a low-dimensional feature-homogeneous space, in which the information-correlated samples are close along manifold while the semantic-complementary instances are close in Euclidean distance.

Specifically, the proposed method can be formulated as follows:

$$\begin{aligned} \boldsymbol{\Psi_1}: \underset{A,B,W,M}{min} & f_S(A,B,W) + \alpha g_M(A,B,M) - \beta h_D(A,B) \\ s.t.\ & A^TA = I, \quad B^TB = I, \quad and \quad M \succeq 0, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{d_x \times k}$, $B \in \mathbb{R}^{d_y \times k}$, $k \leq min(d_x, d_y)$, and α and β are two trade-off parameters. The orthogonal constraints $A^T A = I$ and $B^T B = I$ can effectively eliminate the correlation among different features in the same source. The positive semidefinite restraint $M \in \mathbb{S}_+^{k \times k} \succeq 0$ can be used to obtain a well-defined pseudo-metric.

The objective function in Eq.(1) consists of the semantic, correlation, and distributional subfunctions. The semantic subfunction $f_S(A, B, W)$:

$$f_S(A, B, W) = \left\| \begin{bmatrix} X_N A \\ Y_N B \end{bmatrix} W - \begin{bmatrix} L_N \\ L_N \end{bmatrix} \right\|_F^2, \tag{2}$$

is based on multivariant linear regression to capture the semantic complementarity between different sources.

The first term in the objective function is multivariate linear regression based on the semantic function, which is used to capture the semantic complementarity between different sources.

Moreover, we define the new distance metrics as follow to obtain a Mahalanobis distance:

$$\mathcal{D}_{M_X}(x_i, x_j) = (x_i - x_j)^T M_X (x_i - x_j), \tag{3}$$

$$\mathcal{D}_{M_Y}(y_i, y_j) = (y_i - y_j)^T M_Y (y_i - y_j), \tag{4}$$

where $M_X = A^T A$ and $M_Y = B^T B$. Therefore, each pair of co-occurring heterogeneous representations (x_i, y_i) can be embedded by the linear transformations A and B into a feature-homogeneous space.

Accordingly, the motivation of introducing the correlation function $g_M(A, B, M)$:

$$g_M(A, B, M) = \| X_N A M B^T Y_N^T \|_F^2, \tag{5}$$

is to measure the smoothness between A and B to extract the information correlation among heterogeneous representations.

Additionally, $\mathcal{C}_X^t$ and $\mathcal{C}_Y^t$ denote respectively the sample sets of t-th class from the sources V_x and V_y. We assume that each sample x_i selects another sample y_j from another source as its neighbor with the probability p_{ij}. Similarly, q_{ij} refers to the probability that y_i is the neighbor of x_j.

We apply the softmax under the Euclidean distance in the feature-homogeneous space to define p_{ij} and q_{ij} as follows:

$$p_{ij} = \frac{exp(- \| Ax_i - By_j \|^2)}{\sum_k exp(- \| Ax_i - By_k \|^2)}, \tag{6}$$

$$q_{ij} = \frac{exp(- \| By_i - Ax_j \|^2)}{\sum_k exp(- \| By_i - Ax_k \|^2)}. \tag{7}$$

Accordingly, the probabilities p_i and q_i:

$$p_i = \sum_{x_i \in \mathcal{C}_X^t \ \& \ y_j \in \mathcal{C}_Y^t} p_{ij}, \tag{8}$$

$$q_i = \sum_{y_i \in \mathcal{C}_Y^t \ \& \ x_j \in \mathcal{C}_X^t} q_{ij}, \tag{9}$$

represents the odds which the sample i will be correctly classified. Consequently, the distributional similarity subfunction $h_D(A, B)$ based on Mahalanobis distance:

$$h_D(A, B) = \sum p_i + \sum q_i, \tag{10}$$

is a leave-one-out validation, which is used to capture the distributional similarity between different sources.

Section 3.1 presents an efficient algorithm to solve Ψ_1.

2.2 The Proposed CMRR Model

Furthermore, to reduce the three-way redundancies and remove double-level heterogeneities, we propose a new CMRR model with GET constraints and GEC criterion to recover one-to-one complementary relationship between the heterogeneous representations from redundant sources in the learned feature-homogeneous space.

Specifically, assuming (A^*, B^*, W^*, M^*) be the optimal solutions of Ψ_1. Then the proposed approach can be formulated:

$$\begin{aligned} \boldsymbol{\Omega_1}: \ & \min_{P,Q} \|P^T H W^* - Q^T R W^*\|_F^2 + \gamma \|P^T H M^* R^T Q\|_F^2 + \\ & \tau \| \left(P^T H + Q^T R\right)/2 \|_* \\ & s.t. \ P \in \Sigma_{n_2 \times n_4}, Q \in \Sigma_{n_3 \times n_4}, \|P\|_{2,1} = \|Q\|_{2,1} = n_4, \end{aligned} \tag{11}$$

where γ and τ are two regularization parameters, P and Q are two GET matrices, $H = X_R A^*$ and $R = Y_R B^*$ are the redundant matrices in S_x and S_y, $\Sigma_{n_2 \times n_4} \in \mathbb{R}^{n_2 \times n_4}$ and $\Sigma_{n_3 \times n_4} \in \mathbb{R}^{n_3 \times n_4}$ are two set of GET matrices, and $n_4 = min(n_2, n_3)$.

The first term in the objective function uses A^*, B^*, and W^* to build one-to-one complementary relationship between the heterogeneous representations of the same object while removing CRO and eliminating SFS. The second term in the objective function is used to clear DRE in the same source by M^* in order to extract the correlated information between heterogeneous representations. The

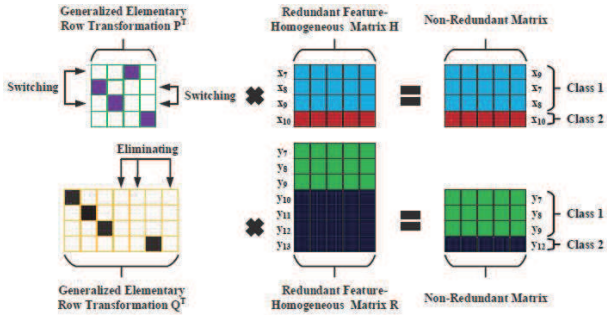

Fig. 3. Correlation-based Multi-source Redundancy Reduction. To establish one-to-one complementary relationship while removing CRO, CMRR model imposes the GET constraints on P and Q to switch the rows in H and R.

low-rank regularization based on trace norm (the third term in the objective function) is used to make the complex representations as linearly-separable as possible. As shown in Fig.3, to switch the rows in H and R, the GET constraints are imposed on P and Q to establish one-to-one complementary relationship while removing CRO. The motivation of introducing the 2,1-norm equality restraint is to clear DRE in H and R through favoring a number of zero rows in P and Q. Note that if there is but only 2,1-norm equality restraint, the P and Q may become a matrix containing only one non-zero row [1]. Thus, it is essential for selecting complementary representations to add the GET constraints on P and Q in Ω_1.

Based on the gradient energy measure [20], the GEC criterion [28] can be used to build a GET matrix effectively. Specifically, every internal element G_{ij} is connected to its four neighbors $G_{i-1,j}$, $G_{i+1,j}$, $G_{i,j-1}$, and $G_{i,j+1}$ in a gradient matrix G obtained by gradient descent method. We can obtain the between-sample energy E_{bs} of G_{ij} according to the ℓ_1-norm gradient magnitude energy [20]:

$$E_{bs} = \frac{\partial}{\partial x}G = \mid G(i+1,j) - G(i,j) \mid + \mid G(i,j) - G(i-1,j) \mid, \tag{12}$$

and the within-sample energy E_{ws} as

$$E_{ws} = \frac{\partial}{\partial y}G = \mid G(i,j+1) - G(i,j) \mid + \mid G(i,j) - G(i,j-1) \mid . \tag{13}$$

We can calculate the global energy of G_{ij} by E_{bs} and E_{ws}:

$$E_{globe} = \delta * E_{bs} + (1-\delta) * E_{ws}, \tag{14}$$

where δ is a trade-off parameter.

The global energy of every element in G can be computed by Eq.(14), and then we can obtain an energy matrix E. As a result, we can compare he global energies of every element. It can be seen that the winner with maximum energy will be set to 1, and the remaining elements in the same row and column will be set to 0. We can repeat the cycle until a GET matrix Q is built.

Section 3.2 presents an efficient algorithm to solve Ω_1.

3 Optimization Technique

In this section, we present an optimization technique to solve the proposed framework.

3.1 An Efficient Solver for Ψ_1

The optimization problem Ψ_1 (see Eq.(1)) can be simplified as follows:

$$\min_{Z\in\mathcal{C}} \quad F(Z), \tag{15}$$

where $F(\cdot) = f_S(\cdot)+\alpha g_M(\cdot)-\beta h_D(\cdot)$ is a smooth function, $Z=[A_Z \quad B_Z \quad W_Z \quad M_Z]$ symbolizes the optimization variables, and the set $\mathcal{C}$ is closed for each variable:

$$\mathcal{C} = \{Z|A_Z^T A_Z = I, B_Z^T B_Z = I, M_Z \succeq 0\}. \tag{16}$$

Because $F(\cdot)$ is continuously differentiable for each variable with Lipschitz continuous gradient L [16], it is appropriate to solve Eq.(15) by Accelerated Projection Gradient (APG) [16] method.

The first-order gradient algorithm APG can accelerate each gradient step and minimize the smooth function, so as to obtain the optimal solution. A solution sequence $\{Z_i\}$ is updated from a search point sequence $\{S_i\}$ in the method.

Due to orthogonal constraints, it is exceedingly difficult for us to optimize the non-convex optimization problem in Eq.(15). However, if Gradient Descent Method with Curvilinear Search (GDMCS) [27] satisfies Armijo-Wolfe conditio,

Algorithm 1: Heterogeneous Manifold Smoothness Learning (**HMSL**)

Input: Z_0=$[A_{Z_0}B_{Z_0}W_{Z_0}M_{Z_0}]$, $F(\cdot)$, $f_S(\cdot)$, $g_M(\cdot)$, $h_D(\cdot)$, X_N, Y_N, $\gamma_1 > 0$, t_0=1, τ_1, $0 < \rho_1 < \rho_2 < 1$, and $maxIter$.
Output: Z^*.
1: Define $F_{\gamma,S}(Z)$=$F(S)$+$\langle\nabla F(S), Z-S\rangle$+$\gamma\|Z-S\|_F^2/2$
2: Calculate $[A_{Z_0}] = \boldsymbol{Schmidt}(A_{Z_0})$.
3: Calculate $[B_{Z_0}] = \boldsymbol{Schmidt}(B_{Z_0})$.
4: Set A_{Z_1}=A_{Z_0}, B_{Z_1}=B_{Z_0}, W_{Z_1}=W_{Z_0}, and M_{Z_1}=M_{Z_0}.
5: for i =1,2,$\cdots$,$maxIter$ do
6: Set $a_i = (t_{i-1} - 1)/t_{i-1}$.
7: Calculate $A_{S_i} = (1 + \alpha_i)A_{Z_i} - \alpha_i A_{Z_{i-1}}$.
8: Calculate $B_{S_i} = (1 + \alpha_i)B_{Z_i} - \alpha_i B_{Z_{i-1}}$.
9: Calculate $W_{S_i} = (1 + \alpha_i)W_{Z_i} - \alpha_i W_{Z_{i-1}}$.
10: Calculate $M_{S_i} = (1 + \alpha_i)M_{Z_i} - \alpha_i M_{Z_{i-1}}$.
11: Set $S_i = [A_{S_i} \quad B_{S_i} \quad W_{S_i} \quad M_{S_i}]$.
12: Calculate $\nabla_{A_S} F(A_{S_i})$, $\nabla_{B_S} F(B_{S_i})$, $\nabla_{W_S} F(W_{S_i})$, and $\nabla_{M_S} F(M_{S_i})$.
13: Define $F_A(A_{Z_i}, B)$ and $F_B(A, B_{Z_i})$.
14: while (true)
15: Calculate $\widehat{A_S} = A_{S_i} - \nabla_{A_S} F(A_{S_i})/\gamma_i$.
16: Calculate $[\widehat{A_S}] = \boldsymbol{Schmidt}(\widehat{A_S})$.
17: Calculate $\widehat{B_S} = B_{S_i} - \nabla_{B_S} F(B_{S_i})/\gamma_i$.
18: Calculate $[\widehat{B_S}] = \boldsymbol{Schmidt}(\widehat{B_S})$.
19: Set $[A_{Z_{i+1}}, B_{Z_{i+1}}] = \boldsymbol{GDMCS}(F(\cdot), \widehat{A_S}, \widehat{B_S}, \tau_1, \rho_1, \rho_2)$.
20: Calculate $W_{Z_{i+1}} = W_{S_i} - \nabla_{W_S} Q(W_{S_i})/\gamma_i$.
21: Calculate $\widehat{M_S} = M_{S_i} - \nabla_{M_S} Q(M_{S_i})/\gamma_i$.

22: Calculate $[M_{Z_{i+1}}] = \boldsymbol{PSP}(\widehat{M_S})$.
23: Set $Z_{i+1} = [A_{Z_{i+1}} \quad B_{Z_{i+1}} \quad W_{Z_{i+1}} \quad M_{Z_{i+1}}]$.
24: if $F(Z_{i+1}) \leq F_{\gamma_i,S_i}(Z_{i+1})$, then break;
25: else Update $\gamma_i = \gamma_i \times 2$.
26: endIf
27: endWhile
28: Update $t_i = \left(1+\sqrt{1+4t_{i-1}^2}\right)/2$, $\gamma_{i+1}=\gamma_i$.
29: endFor
30: Set $Z^* = Z_{i+1}$.

it has been proved by Guo and Xiao in [9] that GDMCS can effectively solve the non-convex problem. We can use the method in [9] to prove that the proposed HMSL algorithm met the using conditions of GDMCS algorithm.

APG projects a given point s onto set $\mathcal{C}$ in the following way:

$$proj_{\mathcal{C}}(s) = arg \min_{z\in\mathcal{C}} \|z-s\|_F^2/2. \tag{17}$$

Positive Semi-definite Projection (PSP) proposed by Weinberger et al. in [26] can remain positive semi-definite constraints, when it minimize a smooth function. It will project optimal variables into a cone of all positive semi-definite matrices after each gradient step. The projection is computed from the diagonalization of optimal variables, which effectively truncates any negative eigenvalues from the gradient step, setting them to zero. PSP can be utilized to solve the problem in Eq.(17).

Finally, when applying APG for solving Eq.(15), a given point S can be projected into the set $\mathcal{C}$ as follows:

$$proj_{\mathcal{C}}(S) = arg \min_{Z\in\mathcal{C}} \|Z-S\|_F^2/2. \tag{18}$$

The problem in Eq.(18) can be solved by the combination of APG, PSP, and GDMCS. The details are given in Algorithm 1, in which the function $\boldsymbol{Schmidt}(\cdot)$ [15] denotes the GramSchmidt process.

3.2 An Efficient Solver for Ω_1

To solve the model Ω_1 (See Section 2.2), an efficient algorithm is given in this subsection. Similarly, the problem in Eq.(11) can be simplified as:

$$\min_{\Theta\in\mathcal{Q}} \quad H(\Theta) = w(\Theta) + \tau t(\Theta), \tag{19}$$

where $w(\cdot) = \|\cdot\|_F^2 + \gamma\|\cdot\|_F^2$ is a smooth subfunction, $t(\cdot) = \|\cdot\|_*$ is an undifferentiable subfunction, $\Theta = [P_\Theta \quad Q_\Theta]$ symbolizes the optimization variables, and set $\mathcal{Q}$ is closed for each variable:

$$\mathcal{Q} = \{\Theta | P_\Theta \in \Sigma_{n_2\times n_4}, Q_\Theta \in \Sigma_{n_3\times n_4}, \|P_\Theta\|_{2,1} = \|Q_\Theta\|_{2,1} = n_4\}. \tag{20}$$

Because $w(\cdot)$ is continuously differentiable for each variable with Lipschitz continuous gradient L [16], it is also appropriate to solve Eq.(19) by APG [16] method.

Similarly, APG projects a given point s onto set $\mathcal{Q}$ in the following way:

$$proj_{\mathcal{Q}}(s) = arg \min_{\theta \in \mathcal{Q}} \|\theta - s\|_F^2/2, \tag{21}$$

The GEC criterion (See Section 2.2) can be used to map the approximate solution of Eq.(21) into the generalized elementary transformation constraint $\mathcal{Q}$. Zhang et al. [28] have successfully used the functions $Energy(\cdot)$ and $Competition(\cdot)$ to implement the GEC criterion according to Eq.(12,13,14).

Algorithm 2: Correlation-based Multi-source Redundancy Reduction (**CMRR**)

Input: $H(\cdot)$, $w(\cdot)$, $t(\cdot)$, $P_{Z_0}=I_{n_2\times n_4}$, $Q_{Z_0}=I_{n_3\times n_4}$, $Z_0=[P_{Z_0}\ Q_{Z_0}]$, X_R, Y_R, δ, $\varepsilon_1 > 0$, $t_0=1$, and $maxIter$.
Output: Z^*.

Define $H_{\varepsilon,S}(Z) = w(S) + \langle \nabla w(S), Z - S\rangle + \varepsilon\|Z - S\|_F^2/2 + \tau t(Z)$.
Set $P_{Z_1} = P_{Z_0}$ and $Q_{Z_1} = Q_{Z_0}$.
for i =1,2,$\cdots$,$maxIter$ do
    Set $a_i = (t_{i-1} - 1)/t_{i-1}$.
    Calculate $P_{S_i} = (1 + \alpha_i)P_{Z_i} - \alpha_i P_{Z_{i-1}}$.
    Calculate $Q_{S_i} = (1 + \alpha_i)Q_{Z_i} - \alpha_i Q_{Z_{i-1}}$.
    Set $S_i = [P_{S_i} \quad Q_{S_i}]$.
    Derive $\nabla_{P_S} w(P_{S_i})$ and $\nabla_{Q_S} w(Q_{S_i})$.
    while (true)
        Calculate $\widehat{P_S} = -\nabla_{P_S} w(P_{S_i})/\varepsilon_i$.
        Calculate $[\widehat{P_{Z_{i+1}}}] = \boldsymbol{Energy}(\widehat{P_S}, \delta)$.
        Calculate $[P_{Z_{i+1}}] = \boldsymbol{Competition}(\widehat{P_{Z_{i+1}}})$.
        Calculate $\widehat{Q_S} = -\nabla_{Q_S} w(Q_{S_i})/\varepsilon_i$.
        Calculate $[\widehat{Q_{Z_{i+1}}}] = \boldsymbol{Energy}(\widehat{Q_S}, \delta)$.
        Calculate $[Q_{Z_{i+1}}] = \boldsymbol{Competition}(\widehat{Q_{Z_{i+1}}})$.
        Set $Z_{i+1} = [P_{Z_{i+1}} \quad Q_{Z_{i+1}}]$.
        if $H(Z_{i+1}) \le H_{\varepsilon_i,S_i}(Z_{i+1})$, then break;
        else Update $\varepsilon_i = \varepsilon_i \times 2$.
        endIf
    endWhile
    Update $t_i = \left(1+\sqrt{1+4t_{i-1}^2}\right)/2$, $\varepsilon_{i+1}=\varepsilon_i$.
endFor
Set $Z^* = Z_{i+1}$.

By combining APG, the function $Energy(\cdot)$, and the function $Competition(\cdot)$, the problem in Eq.(19) can be solved. The Algorithm 2 provides the details.

4 Experimental Results and Analyses

4.1 Datasets and Settings

The three benchmark multi-source datasets, i.e., Wikipedia [19], Corel 5K [8], and MIR Flickr [11], are used to evaluate the proposed framework. The statistics of the datasets are given in Table 2, and brief descriptions of the chosen feature sets in the above-mentioned datasets are listed in Table 3.

These three datasets are divided into train and test subsets. We randomly select 10 percent of multi-source data in the train and test sets, respectively. Then the heterogeneous representations of these multi-source data are rearranged in random order and we manually generated 10 percent of the redundant representations from Source S_y in the data. We use the 5-fold cross-validation to tune some important parameters in all the methods. Additionally, all the experiments take the LIBSVM classifier as the benchmark for classification tasks.

Table 2. Statistics of the Multi-source Datasets

Dataset	Total Attributes	Total Classes	Total Samples
Wikipedia	258	10	2866
Corel 5K	200	260	4999
MIR Flickr	5857	38	25000

Table 3. Brief Descriptions of the Feature Sets

Dataset	Feature Set	Total Attributes	Total Labels	Total Instances
Wikipedia	Image (S_x)	128	10	2866
	Text (S_y)	130	10	2866
Corel 5K	DenseHue (S_x)	100	260	4999
	HarrisHue (S_y)	100	260	4999
MIR Flickr	Image (S_x)	3857	38	25000
	Text (S_y)	2000	38	25000

4.2 Analysis of Manifold Learning Algorithms

To verify the manifold smoothness measure learned by the proposed HMSL method, HMSL is compared in classification performance with other four state-of-the-art manifold learning algorithms such as ESRM [6], EMR [7], MKPLS [2], and DDGR [3]. The MIR FLICKR dataset is used in the experiment, and the best performance is reported. The data in training set are randomly sampled in the ratio $\{25\%, 50\%, 75\%, 100\%\}$, and the size of the test set is fixed. Unlike our framework, before comparing ESRM, EMR, MKPLS and DDGR, we first implement CCA [23] to construct feature-homogeneous space between different sources. We select $min(d_x, d_y)$ as the dimensionality k of the feature-homogeneous space. The setting of the parameters in ESRM, EMR, MKPLS, and DDGR is the same as the original works [6, 7, 2, 3].

In essence, the proposed HMSL model is also a manifold learning method based on manifold regularization. However, there are significant differences between HMSL and the above-mentioned other four methods. The main difference

Table 4. Classification Performance of Manifold Learning Methods in terms of AUC.

Method	Sampling Ratio			
	25%	50%	75%	100%
DDGR	0.5374	0.5963	0.6245	0.6756
MKPLS	0.5481	0.6040	0.6372	0.6824
EMR	0.5171	0.5744	0.6268	0.6654
ESRM	0.5445	0.5978	0.6554	0.6813
HMSL	**0.5991**	**0.6596**	**0.7053**	**0.7494**

between HMSL and ESRM is that ESRM is a mono-source learning algorithm without the ability of handling multi-source problem. Moreover, though MKPLS also use manifold regularization to exploit the correlation among heterogeneous representations, the distributional similarity among different sources is not utilized fully. Additionally, different from EMR and DDGR, HMSL takes full account of the semantic complementarity between different sources.

From Table 4, we can clearly observe that HMSL greatly outperforms other manifold learning methods in classification performance. The results present that HMSL can capture information correlation between different sources more effectively than the comparative methods. In addition, as the number of training samples increases, the performance of HMSL will also be improved. Accordingly, a certain number of existing nonredundant samples is essential for HMSL to learn an excellent manifold smoothness measure.

4.3 Evaluation of Dimension Reduction Techniques

In order to evaluate the possibility of eliminating SFS in the proposed CMRR model, we further compare the effect of dimension reduction among different state-of-the-art methods, such as KTLQD [18], ERFSJNM [17], FSGRM [24], and NSACR [13]. The generalized identity matrices are taken as the initial values of P and Q in Algorithm 2. The regularization parameters γ and τ are tuned among the set $\{10^i | i = -2, -1, 0, 1, 2\}$. The parameter δ in Eq.(14) is set to 0.1. For KTLQD, ERFSJNM, FSGRM, and NSACR, the experimental setups follow the original ones [18, 17, 24, 13], respectively.

In machine learning, the eliminating of sample features superabundance can be divided into feature selection and dimension reduction. It is a key component in building robust machine learning models for analysis, classification, clustering, and retrieval to avoid high computational time. To achieve this goal, CMRR reduces the superabundance of sample features by using the learned multiple heterogeneous linear transformations. Therefore, after eliminating SFS,

the multi-source heterogeneous redundant data is more likely to be separated linearly.

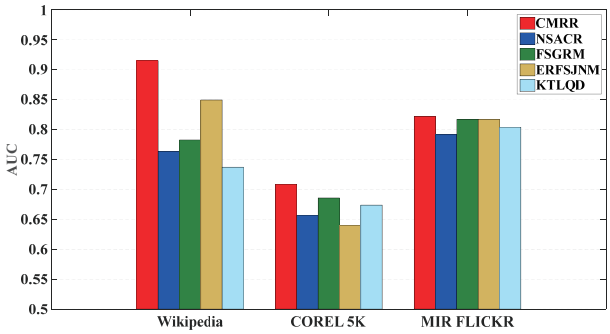

Fig. 4. Comparison of Classification Performance of Dimension Reduction Algorithms.

We can observe from Fig.4 that CMRR has better classification performance than KTLQD, ERFSJNM, FSGRM, and NSACR. This observation further justifies that CMRR can effectively eliminate SFS.

4.4 Comparison of Sample Selection Approaches

To test the performance of the proposed CMRR in different redundancy rates, we further compare the classification performances of CMRR with other sample selection methods such as MASSFDTI [25], SPSRS [5], WCTSSIM [21], and ACP [22] in the larger MIR Flickr dataset. We tune the redundancy rates on the set $\{10\%, 15\%, 20\%, 25\%\}$.

From the view of the function, the proposed CMRR model is also essentially a sample selection method such as MASSFDTI, SPSRS, WCTSSIM, and ACP. However, there are some significant differences between CMRR and other methods. CMRR is based on the correlation among sample representations from different sources. So it will be more favorable to clear DRE and remove CRO for reestablishing the one-to-one complementary relationship among heterogeneous representations.

Just to pursue such a purpose, we first use HMSL to project the multi-source data into a feature-homogeneous space and then apply MASSFDTI, SPSRS, WCTSSIM, ACP, and CMRR to reduce redundant samples. The setting of the parameters in MASSFDTI, SPSRS, WCTSSIM, and ACP is the same as the original works [25, 5, 21, 22].

It can be seen in Fig.5 that CMRR is superior to the other models in the classification performance. This observation further confirms that CMRR has an obvious advantage over other methods in removing FDD and SSD and rebuilding the one-to-one complementary relationship among heterogeneous representations. Nevertheless, with the increasing of redundancy rate, the performance of CMRR will degrade. Thus, CMRR also has some limitations that it needs a certain number of existing nonredundant samples to reduce redundant source.

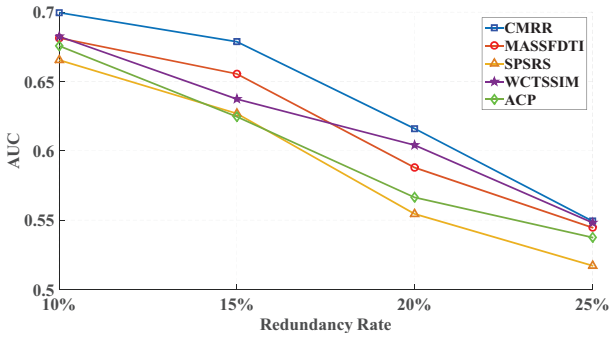

Fig. 5. Comparison of Classification Performance of Sample Selection Approaches in Different Redundancy Rates.

5 Conclusion

This paper investigates the redundant sources problem in multi-source learning. We developed a general simplifying framework to reduce redundant sources of multi-source data. Within this framework, a feature-homogeneous space is learned by the proposed HMSL model to capture the semantic complementarity, information correlation, and distributional similarity among different sources. Meanwhile, we proposed a CMRR method with GET constraints based on GEC criterion to remove the three-way redundancies and double-level heterogeneities in the learned feature-homogeneous space. Finally, we evaluated and verified the effectiveness of the proposed framework on five benchmark multi-source heterogeneous datasets.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (No.61601458).

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. 73(3), 243–272 (2008)
2. Bakry, A., Elgammal, A.: Mkpls: Manifold kernel partial least squares for lipreadingand speaker identification. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 684–691 (2013)
3. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7, 2399–2434 (2006)
4. Bellman, R.: Dynamic programming and lagrange multipliers 42(10), 767 (1956)

5. Chen, D., Zhao, S., Zhang, L., Yang, Y., Zhang, X.: Sample pair selection for attribute reduction with rough set. IEEE Trans. Knowl. Data Eng. 24(11), 2080–2093 (2012)
6. Freedman, D.: Efficient simplicial reconstructions of manifolds from their samples. IEEE Trans. Pattern Anal. Mach. Intell. 24(10), 1349–1357 (2002)
7. Geng, B., Tao, D., Xu, C., Yang, L., Hua, X.: Ensemble manifold regularization. IEEE Trans. Pattern Anal. Mach. Intell. 34(6), 1227–1233 (2012)
8. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 902–909 (2010)
9. Guo, Y., Xiao, M.: Cross language text classification via subspace co-regularized multi-view learning. In: Proc. ACM Int. Conf. Mach. Learn. pp. 915–922 (2012)
10. He, X., Li, L.M., Roqueiro, D., Borgwardt, K.M.: Multi-view spectral clustering on conflicting views. In: Proc. Springer European Conf. Machine Learning and Knowl. Discovery. pp. 826–842 (2017)
11. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proc. ACM Int. Conf. Knowl. Info. Retrieval. pp. 39–43 (2008)
12. Lan, C., Huan, J.: Reducing the unlabeled sample complexity of semi-supervised multi-view learning. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. pp. 627–634 (2015)
13. Li, Z., Tang, J.: Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Trans. Image Process. 24(12), 5343–5355 (2015)
14. Luo, P., Y., P.J., Y., G.Z., Jianping Fan, J.P.: Multi-view semantic learning for data representation. In: Proc. Springer European Conf. Machine Learning and Knowl. Discovery. pp. 367–382 (2015)
15. Meyer, C.D.: Matrix Analysis and Applied Linear Algebra. Siam (2000)
16. Nesterov, Y.: Introductory Lectures on Convex Optimization, vol. 87. Springer Science & Business Media (2004)
17. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Proc. Adv. Neural Inf. Process. Syst. pp. 1813–1821 (2010)
18. Quanz, B., Huan, J., Mishra, M.: Knowledge transfer with low-quality data: A feature extraction issue. IEEE Trans. Knowl. Data Eng. 24(10), 1789–1802 (2012)
19. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proc. ACM Int. Conf. Multimedia. pp. 251–260 (2010)
20. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. ACM Trans. Graphics 27(3), 16 (2008)
21. Shahrian, E., Rajan, D.: Weighted color and texture sample selection for image matting. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 718–725 (2012)
22. Su, H., Yin, Z., Kanade, T., Huh, S.: Active sample selection and correction propagation on a gradually-augmented graph. In: Proc. IEEE Comput. Vis. Pattern Recognit. pp. 1975–1983 (2015)
23. Sun, L., Ji, S., Ye, J.: Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. IEEE Trans. Pattern Anal. Mach. Intell. 33(1), 194–200 (2011)
24. Wang, D., Nie, F., Huang, H.: Feature selection via global redundancy minimization. IEEE Trans. Knowl. Data Eng. 27(10), 2743–2755 (2015)
25. Wang, X., Dong, L., Yan, J.: Maximum ambiguity-based sample selection in fuzzy decision tree induction. IEEE Trans. Knowl. Data Eng. 24(8), 1491–1505 (2012)

26. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10, 207–244 (2009)
27. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. 142(1-2), 397–434 (2013)
28. Zhang, L., Wang, S., Zhang, X., Wang, Y., Li, B., Shen, D., Ji, S.: Collaborative multi-view denoising. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. pp. 2045–2054 (2016)
29. Zhuang, Y., Yang, Y., Wu, F., Pan, Y.: Manifold learning based cross-media retrieval: A solution to media object complementary nature. J. VLSI Signal Process. 46(2-3), 153–164 (2007)