Neurocomputing 371 (2020) 67–77

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Global and local multi-view multi-label learning

Changming Zhu[a,b,*], Duoqian Miao[b], Zhe Wang[c], Rigui Zhou[a], Lai Wei[a], Xiafen Zhang[a]

[a] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, PR China
[b] College of Electronics and Information Engineering, Tongji University, Shanghai 200092, PR China
[c] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 4 November 2018
Revised 25 June 2019
Accepted 5 September 2019
Available online 13 September 2019

Communicated by Dr. Chenping Hou

Keywords:
Multi-label
Label correlation
Multi-view

ABSTRACT

In order to process multi-view multi-label data sets, we propose global and local multi-view multi-label learning (GLMVML). This method can exploit global and local label correlations of both the whole data set and each view simultaneously. What's more, GLMVML introduces a consensus multi-view representation which encodes the complementary information from different views. Related experiments on three multi-view data sets, fourteen multi-label data sets, and one multi-view multi-label data set have validated that (1) GLMVML has a better average AUC and precision and it is superior to the classical multi-view learning methods and multi-label learning methods in statistical; (2) the running time of GLMVML won't add too much; (3) GLMVML has a good convergence and ability to process multi-view multi-label data sets; (4) since the model of GLMVML consists of both the global label correlations and local label correlations, so parameter values should be moderate rather than too large or too small.

1. Introduction

1.1. Background: three classical data sets

In real-world applications, there are three kinds of data sets are always encountered, i.e., multi-view data sets, multi-label data sets, and multi-view multi-label data sets [1–9].

(1) A multi-view data set consists of multiple instances with different views. Each view denotes information of instances in a certain area. Taking a web page data set X as an example, if this data set consists of n web pages and each page is an instance x_i $(i = 1, 2, \dots, n)$, then each instance can be represented by v forms including text, image, video. Each form x_i^j is a view of x_i and $X^j = \{x_i^j\}_{i=1}^n$ represents j-th view $(j = 1, 2, \dots, v)$. Under this definition, $X = \{X^j\}_{j=1}^v = \{x_i\}_{i=1}^n$ is a multi-view data set [10].
(2) A multi-label data set consists of multiple instances with multiple class labels. For example, as [11] said, a scene image can be annotated with several tags [12], a document may belong to multiple topics [13], and a piece of music may be associated with different genres [14].
(3) A multi-view multi-label data set consists of instances which have multiple views and class labels. Namely, for each instance, it has respective class label if observed and moreover, each instance has multiple views and different views will also bring different class labels if still observed. For example, if there is a stereoscopic artistic works and people can appreciate it from multiple aspects including the text introduction, picture introduction, voice introduction, and video introduction. Now text, picture, voice, video can be regarded as different views. Then from different views, this stereoscopic artistic works can be labeled as different labels. Maybe according to the text introduction, this works can be treated as a historical, classical one and according to picture introduction, this works can be treated as a nature, rural one. Here, historical, classical, nature, rural are labels. Moreover, in real-world applications, some instances are difficult to be labeled with the lack of information. Thus, sometimes, in a view, some instances lost some labels.

1.2. Background: traditional solutions

In order to process multi-view data sets, multi-view learning has been developed and widely used in many fields including multi-view clustering [9,15], handwritten digit recognition [16], human gait recognition [17], image recognition [18,19], dimensionality reduction [6,7] and so on [20]. For example, Zhang et al. have proposed a latent multi-view subspace clustering (LMSC) [9] to improve the clustering performances. Sun et al. have proposed a multiple-view multiple-learner (MVML) [16] to enhance the ability of handwritten digit recognition. Deng et al. have developed a ro-

* Corresponding author at: College of Information Engineering, Shanghai Maritime University, Shanghai 201306, PR China.
E-mail address: cmzhu@shmtu.edu.cn (C. Zhu).

https://doi.org/10.1016/j.neucom.2019.09.009
0925-2312/

bust gait recognition method using multiple views fusion and deterministic learning so as to encourage recognition accuracy of human gait characteristics [17]. Wu et al. find that the existing multi-view dictionary learning (DL) methods suffer from the problem of performance degeneration when large noise exists in multiple views and propose a multi-view low-rank DL (MLDL) to overcome such a problem [18]. Yuan et al. have developed a fractional-order embedding canonical correlation analysis (FECCA), which is based on fractional-order within-set and between-set scatter matrices for multi-view dimensionality reduction [7] and Hou et al. propose a multiple view semi-supervised dimensionality reduction (MVSSDR) to solve semi-supervised multi-view data sets whose training instances consist of some labeled ones and some unlabeled ones [6].

In order to process multi-label data sets, multi-label learning has also been developed and many methods are proposed. For example, Weng et al. have developed a multi-label learning based on label-specific features and local pairwise label correlation (LF-LPLC) [21], Kumar et al. have developed a multi-label classification machine with hierarchical embedding (MLCHE) to process practical multi-label classification problems such as image annotation, text categorization and sentiment analysis [22], and Zhu et al. have proposed a multi-label learning named GLOCAL which takes both global and local label correlation into consideration [11].

Since multi-view learning and multi-label learning have a strong pertinence and they cannot process multi-view multi-label data sets, thus some scholars develop corresponding solutions. First, for the multi-view multi-label data sets, most existing multi-label learning methods do not sufficiently consider the complementary information among multiple views which leads to unsatisfying performance, thus Zhang et al. develop latent semantic aware multi-view multi-label learning (LSA-MML) to fully take advantage of multiple views of data and well learn the common representations by simultaneously enforcing the consistence of latent semantic bases among different views in kernel spaces. Experiments validate that the superiority of LSA-MML for multi-view multi-label classification [8]. Second, in terms of the features among the multi-view multi-label data sets, Luo et al. introduce multi-view vector-valued manifold regularization (MV^3MR) to integrate multiple features and exploit the complementary property of different features and discover the intrinsic local geometry of the compact support shared by different features under the theme of manifold regularization. Doing so, MV^3MR can process multi-view multi-label image classification tasks well [5]. Third, Zhu et al. conduct a hierarchical feature selection for the multi-view multi-label learning and develop a block-row sparse multi-view multi-label learning framework (BrSMVML). BrSMVML effectively conducts image classification by avoiding the adverse impact of both the redundant views and the noisy features [4]. Besides those learning methods, multi-view based multi-label propagation (MVMLP) [23] and semi-supervised dimension reduction for multi-label and multi-view learning (SSDR-MML) [24] are also the widely used methods. Generally speaking, the above multi-view multi-label learning methods can effectively process many multi-view multi-label data sets.

1.3. Problems

Although many learning methods are proposed for these different kinds of data sets, but with the further analysis, it is found that these methods neglect two factors. The first factor is the ignorance of the exploitation of global and local label correlations simultaneously. As is known to all, labels of instances exist some correlations, for example, label 'nature' and label 'rural' have a subordinate relationship. But for the above methods, they always assume that the label correlations are global and shared by all instances or that the label correlations are local and shared only by a data subset, namely, they cannot exploit global and local label correlations simultaneously. The second factor is that some multi-view multi-label learning methods cannot reflect the consensus or complementary principle of multi-view learning. For some methods including LSA-MML, LMSC [8,9], they introduce a consensus multi-view representation which encodes the complementary information from different views. But for other methods, they neglect that.

1.4. Proposal

In order to introduce these two factors into a multi-view multi-label learning method simultaneously, in this work, we take GLOCAL which can exploit global and local label correlations simultaneously as a basic method and extend the model to multi-view problem. Then we also introduce a consensus multi-view representation into this new model so as to reflect the complementary information from different views. The proposed model is named global and local multi-view multi-label learning (GLMVML).

1.5. Novelty and contributions

The novelty of GLMVML is that in the filed of multi-view multi-label learning, it is the first trial for the combination of global and local label correlations and the complementary information from different views. Different from the basic learning method GLOCAL which is a multi-label single-view learning, the proposed GLMVML is the multi-view version of GLOCAL and GLMVML can reflect the complementary information from different views.

The contributions of GLMVML are (1) it can take advantage of the complementary information from different views; (2) it can reflect the global and local label correlations simultaneously; (3) it has a better ability to process multi-view multi-label data sets; (4) it pushes on the research of multi-view multi-label learning further.

1.6. Framework

The framework of the rest paper is given below. Section 2 shows the framework of the developed GLMVML. Section 3 gives the experimental results. The conclusion and future work are given in Section 4.

2. Global and local multi-view multi-label learning

2.1. Data preparation

Suppose there is a multi-view multi-label data set X, it has v views and its dimensionality is $d \times n$ (see Fig. 1). d is the dimensionality of each instance and n is the total number of instances. Then, ith instance $x_i \in \mathbb{R}^d$ can be represented as

$$x_i = \begin{pmatrix} x_i^1 \\ x_i^2 \\ . \\ . \\ . \\ x_i^v \end{pmatrix}$$

where $x_i^j \in \mathbb{R}^{d_j \times 1}$ denotes jth view of ith instance and $d = \sum_{j=1}^{v} d_j$, d_j is the number of features of x_i^j. Here

$$x_i^j = \begin{pmatrix} x_{i1}^j \\ x_{i2}^j \\ . \\ . \\ . \\ x_{id_j}^j \end{pmatrix}$$

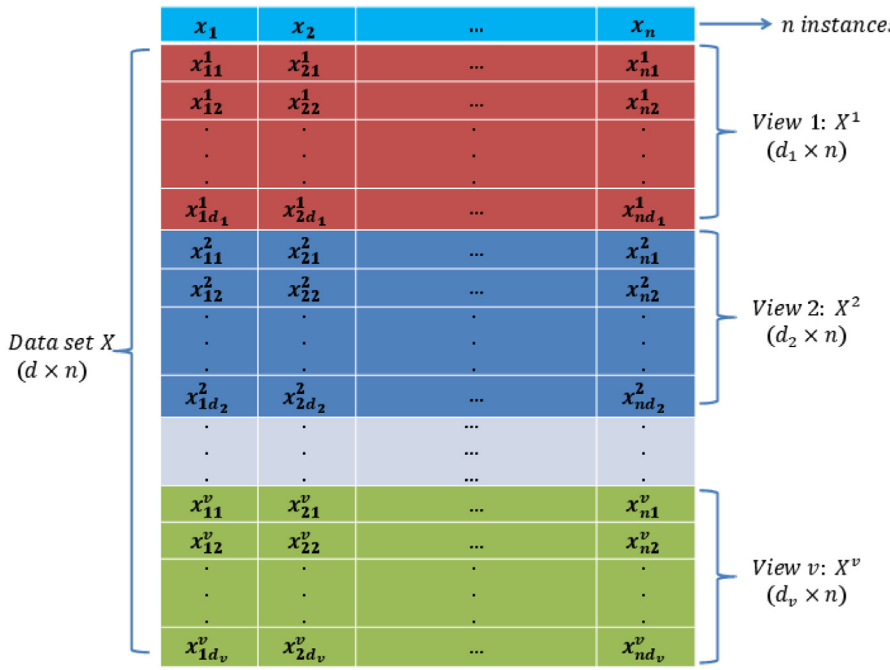


Fig. 1. The expression of a multi-view data set.

and x_{it}^j denotes the tth feature of x_i^j. According to these definitions, the jth view of this data set can be written as $X^j = (x_1^j, x_2^j, \dots, x_n^j)$ and the dimensionality of X^j is $d_j \times n$. Now, we have

$$X = \begin{pmatrix} X^1 \\ X^2 \\ \cdot \\ \cdot \\ \cdot \\ X^v \end{pmatrix} = (x_1, x_2, \dots, x_n).$$

Then for X, since it is a multi-view multi-label data set and in real-world applications, some labels can be observed while some cannot. Thus, we suppose X has l class labels, i.e., $C = \{c_1, c_2, \dots, c_l\}$ and we define the ground-truth label vector of x_i^j is $\widetilde{y}_i^j \in \mathbb{R}^{l\times 1} \subseteq \{-1, 1\}^l$, where $[\widetilde{y}_i^j]_t = 1$ if x_i^j is labeled as c_t, and -1 otherwise. Alike, we define the observed label vector of x_i^j is $y_i^j \in \mathbb{R}^{l\times 1} \subseteq \{-1, 1, 0\}^l$, where $[y_i^j]_t = 0$ if class label c_t is not labeled for x_i^j, and $[y_i^j]_t = [\widetilde{y}_i^j]_t$ otherwise. With the same definition, for an instance x_i, its ground-truth label vector is $\widetilde{y}_i \in \mathbb{R}^{l\times 1} \subseteq \{-1, 1\}^l$ and its observed label vector is $y_i \in \mathbb{R}^{l\times 1} \subseteq \{-1, 1, 0\}^l$. Here, $[\widetilde{y}_i]_t = 1$ if x_i is labeled as c_t, and -1 otherwise, $[y_i]_t = 0$ if class label c_t is not labeled for x_i, and $[y_i]_t = [\widetilde{y}_i]_t$ otherwise.

Now, we can get the ground-truth label matrix of the whole data set, i.e, $\widetilde{Y} = (\widetilde{y}_1, \widetilde{y}_2, \dots, \widetilde{y}_n) \in \mathbb{R}^{l\times n}$, the observed label matrix of the whole data set, i.e., $Y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^{l\times n}$, the ground-truth label matrix of the jth view, i.e, $\widetilde{Y}^j = (\widetilde{y}_1^j, \widetilde{y}_2^j, \dots, \widetilde{y}_n^j) \in \mathbb{R}^{l\times n}$, the observed label matrix of the jth view, i.e., $Y^j = (y_1^j, y_2^j, \dots, y_n^j) \in \mathbb{R}^{l\times n}$.

2.2. Framework

According to the above contents, we know if $\widetilde{Y}$ is low-rank, it can be written as the low-rank decomposition, i.e., $\widetilde{Y} = UV$, where $U \in \mathbb{R}^{l\times k}$, $V \in \mathbb{R}^{k\times n}$, and $rank(\widetilde{Y}) = k < l$. U has a function to project the original labels to the latent label space while V can be treated as the latent labels that are more compact and more semantically abstract than the original labels. For $\widetilde{Y}^j$, it can also be decomposed by $\widetilde{Y}^j = U^j V^j$ where $U^j \in \mathbb{R}^{l\times k_j}$, $V^j \in \mathbb{R}^{k_j\times n}$, and $rank(\widetilde{Y}^j) = k_j < l$. Since in real-world applications, labels are only partially observed, so we always want to minimize the reconstruction error on the observed labels, i.e., $\min\limits_{U,V,U^j,V^j} ||\amalg_\Omega(Y - UV)||_F^2 + \sum_{j=1}^v \left|\left|\amalg_{\Omega^j}(Y^j - U^j V^j)\right|\right|_F^2$. Here, $||\star||_F^2$ represents the square of Frobenius norm for $\star$, Ω (Ω^j) consists of indices of the observed labels in Y (Y^j), $[||\amalg_\Omega(A)||]_{ij} = A_{ij}$ if $(i, j) \in \Omega$, and 0 otherwise (similar to Ω^j case).

After that, we adopt a linear mapping $W \in \mathbb{R}^{d\times k}$ ($W^j \in \mathbb{R}^{d_j\times k_j}$) to map instances to the latent labels and W (W^j) is learned by

$$\min_{W,V,W^j,V^j} \left|\left|V - W^T X\right|\right|_F^2 + \sum_{j=1}^{v} \left|\left|V^j - {W^j}^T X^j\right|\right|_F^2.$$

Moreover, in order to introduce the local label correlation, we divide the data set into several groups with a clustering method (in our work, k-means is used. Although there are many clustering methods can be used, but we only select k-means just for convenience and selecting another one won't influence our conclusion). Namely, for X, it is partitioned into g groups, i.e., $X = \{X_1, X_2, \dots, X_g\}$ and each part $X_m \in \mathbb{R}^{d\times n_m}$ where n_m is the number of instances in X_m. Then under j-th view, X^j is also divided into g^j groups, i.e., $X^j = \{X_1^j, X_2^j, \dots, X_{g^j}^j\}$ and mth group of X^j is $X_m^j \in \mathbb{R}^{d_j\times n_m^j}$. Then since the prediction on instance x_i is $sign(f(x_i))$ where $f(x_i) = UW^T x_i \in \mathbb{R}^{l\times 1}$, so $F_0 = [f(x_1), f(x_2), \dots, f(x_n)] = UW^T X$ represents the classifier output matrix of X. Similarly, $F_0^j = U^j {W^j}^T X^j$, $F_m = UW^T X_m$, $F_m^j = U^j {W^j}^T X_m^j$ represent the classifier output matrices of X^j, X_m, X_m^j respectively. The dimensions of F_0, F_0^j, F_m, F_m^j are $l \times n$, $l \times n$, $l \times n_m$, $l \times n_m^j$, respectively.

Then, on the base of X, X_m, X^j, X_m^j and their corresponding observed label matrices, we compute the label correlation matrices. Take X as instance, $S_0 = \{[S_0]_{pq}\}$ denotes global label correlation matrix and $[S_0]_{pq} = \frac{y_{p,:} y_{q,:}^T}{||y_{p,:}|| \, ||y_{q,:}||}$ represents the global label correlation of p-th label with respect to qth label and $y_{p,:}$ is the pth row of Y. Then we let L_0 be the Laplacian matrix of S_0. Similarly, for X_m, $S_m = \{[S_m]_{pq}\}$ is the corresponding local label correlation matrix and L_m is its Laplacian matrix. Then, under j-th view, for X^j and X_m^j $S_0^j = \{[S_0^j]_{pq}\}$ and $S_m^j = \{[S_m^j]_{pq}\}$ are the corresponding global label correlation matrix and local label correlation matrix, L_0^j and L_m^j are their corresponding Laplacian matrices. Dimensions of S_0, L_0, S_m, L_m, S_0^j, L_0^j, S_m^j, L_m^j are both $l \times l$.

According to the above definitions, we want the classifier outputs can be closer if two labels are more positively correlated and as [28] said, we should minimize $tr(F_0^T L_0 F_0) + \sum_{m=1}^{g} tr(F_m^T L_m F_m) + \sum_{j=1}^{v} (tr(F_0^{j^T} L_0^j F_0^j) + \sum_{m=1}^{g^j} tr(F_m^{j^T} L_m^j F_m^j))$ where $tr(A)$ represents the trace of A.

Furthermore, refer to LSA-MML and LMSC [8,9] which introduce a consensus multi-view representation to encode the complementary information from different views, we adopt the same way. Suppose P is a latent representation matrix (i.e., consensus multi-view representation), B^j is the basic matrix corresponding to jth view, then $\sum_{j=1}^{v} ||X^j - B^j P||_F^2$ searches a comprehensive multi-view representation and $\sum_{j \neq t} IND(B^j, B^t)$ is used to measure the independence between different views where $IND(B^j, B^t) = -HSIC(B^j, B^t)$ and $HSIC$ is a Hilbert-Schmidt independence criterion estimator [8].

So according to the above contents, our goal is to solve the following optimization problem.

$$
\begin{aligned}
&\min_{U,W,V,U^j,W^j,V^j} ||\amalg_\Omega (Y - UV)||_F^2 \\
&+ \lambda_0 ||V - W^T X||_F^2 + \lambda_1 \Re\left(U, V, W, U^j, V^j, W^j, P, B^j\right) \\
&+ \sum_{j=1}^{v} \left(\lambda_2 ||\amalg_{\Omega^j} (Y^j - U^j V^j)||_F^2 + \lambda_3 ||V^j - W^{j^T} X^j||_F^2 \right) \\
&+ \lambda_4 tr\left(F_0^T L_0 F_0\right) + \lambda_5 \sum_{m=1}^{g} tr\left(F_m^T L_m F_m\right) \\
&+ \sum_{j=1}^{v} \left(\lambda_6^j tr(F_0^{j^T} L_0^j F_0^j) + \lambda_7^j \sum_{m=1}^{g^j} tr\left(F_m^{j^T} L_m^j F_m^j\right) \right) \\
&+ \lambda_8 \sum_{j=1}^{v} ||X^j - B^j P||_F^2 + \lambda_9 \sum_{j \neq t} IND(B^j, B^t) \qquad (1)
\end{aligned}
$$

where λs are tradeoff parameters, λ^js are tradeoff parameters corresponding to j-th views, $\Re(U, V, W, U^j, V^j, W^j, P, B^j)$ is the regularizer.

Furthermore, for convenience of computation, we use $J = \{[J]_{pq}\} \in \mathbb{R}^{l \times n}$ and $J^j = \{[J^j]_{pq}\} \in \mathbb{R}^{l \times n}$ where $J_{pq} = 1$ ($J_{pq}^j = 1$) if $(p, q) \in \Omega$ ($(p, q) \in \Omega^j$), and 0 (0) otherwise. We treat J and J^j as observation indicator matrices. Moreover, it can be easily found that $S_0 = \sum_{m=1}^{g} \frac{n_m}{n} S_m$ and $S_0^j = \sum_{m=1}^{g} \frac{n_m^j}{n} S_m^j$, so $L_0 = \sum_{m=1}^{g} \frac{n_m}{n} L_m$ and $L_0^j = \sum_{m=1}^{g} \frac{n_m^j}{n} L_m^j$. Since Laplacian matrices are symmetric positive definite, so we can decompose them by $L_m = Z_m Z_m^T$ and $L_m^j = Z_m^j Z_m^{j^T}$ where $Z_m \in \mathbb{R}^{l \times k}$ and $Z_m^j \in \mathbb{R}^{l \times k_j}$. In order to avoid the $Z_m = 0$ and $Z_m^j = 0$ during the procedure of optimization, we add the constraint $diag(Z_m Z_m^T) = 1$ and $diag(Z_m^j Z_m^{j^T}) = 1$ where $diag(A)$ represents a vector which contains the diagonal entries in A.

Finally, the optimization problem Eq. (1) can be rewritten as below.

$$
\begin{aligned}
&\min_{Z_m,U,V,W,Z_m^j,U^j,V^j,W^j} ||J \circ (Y - UV)||_F^2 \\
&+ \lambda_0 ||V - W^T X||_F^2 + \lambda_1 \Re(U, V, W, U^j, V^j, W^j, P, B^j) \\
&+ \sum_{j=1}^{v} \left(\lambda_2 ||J^j \circ (Y^j - U^j V^j)||_F^2 + \lambda_3 ||V^j - W^{j^T} X^j||_F^2 \right) \\
&+ \sum_{m=1}^{g} \left(\lambda_4 \frac{n_m}{n} tr(F_0^T Z_m Z_m^T F_0) + \lambda_5 tr\left(F_m^T Z_m Z_m^T F_m\right) \right) \\
&+ \sum_{j=1}^{v} \left(\sum_{m=1}^{g^j} \left(\lambda_6^j \frac{n_m^j}{n} tr\left(F_0^{j^T} Z_m^j Z_m^{j^T} F_0^j\right) + \lambda_7^j tr\left(F_m^{j^T} Z_m^j Z_m^{j^T} F_m^j\right) \right) \right) \\
&+ \lambda_8 \sum_{j=1}^{v} ||X^j - B^j P||_F^2 + \lambda_9 \sum_{j \neq t} IND(B^j, B^t) \\
&s.t. \quad diag(Z_m Z_m^T) = 1 \qquad m \in \{1, 2, \ldots, g\} \\
&\qquad diag(Z_m^j Z_m^{j^T}) = 1 \qquad m \in \{1, 2, \ldots, g^j\} \qquad (2)
\end{aligned}
$$

where $\circ$ represents Hadamard (element-wise) product and we let $\Re(U, V, W, U^j, V^j, W^j, P, B^j) = ||U||_F^2 + ||V||_F^2 + ||W||_F^2 + ||U^j||_F^2 + ||V^j||_F^2 + ||W^j||_F^2 + ||P||_F^2 + ||B^j||_F^2$ in this work.

Table 1
Algorithm: GLMVML.

Input: X and X^j and their corresponding group partition which are given with k-means, label matrix Y and Y^j, and observation indicator matrices J and J^j
Output: Z_m, U, W, Z_r^j, U^j, W^j where $j = 1, 2, \ldots, v$, $m = 1, 2, \ldots, g$, $r = 1, 2, \ldots, g^j$
1. initialize Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j;
2. **repeat**
3. for $m = 1, 2, \ldots, g$, $j = 1, 2, \ldots, v$, $r = 1, 2, \ldots, g^j$
4. update one of the $A \in \{Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j\}$ and fix the others simultaneously;
5. end for
6. **until** convergence or maximum number of iterations.

2.3. Solution

In order to solve the problem Eq. (2), we adopt alternating optimization. Namely, in each iteration, we update one of the variables in $\{Z_m, U, V, W, Z_m^j, U^j, V^j, W^j, P, B^j\}$ with gradient descent and leave the others fixed. Here, take the updating of Z_m as instance. In the each iteration, in order to update Z_m, we fix others and problem Eq. (2) can be reduced to

$$
\begin{aligned}
&\min_{Z_m} \lambda_4 \frac{n_m}{n} tr\left(F_0^T Z_m Z_m^T F_0\right) + \lambda_5 tr\left(F_m^T Z_m Z_m^T F_m\right) \qquad (3) \\
&s.t. \quad diag(Z_m Z_m^T) = 1 \qquad m \in \{1, 2, \ldots, g\}
\end{aligned}
$$

Then compute the gradient of the Eq. (3) with respect to Z_m with the following equation.

$$
\nabla_{Z_m} = \lambda_4 \frac{n_m}{n} U W^T X X^T W U^T Z_m + \lambda_5 U W^T X_m X_m^T W U^T Z_m \qquad (4)
$$

After we get the ∇_A where $A \in \{Z_m, U, V, W, Z_m^j, U^j, V^j, W^j, P, B^j\}$, we can use $A := A - \eta \nabla_A$ to update A where η is the step size.

Table 1 gives the summary of the algorithm GLMVML. After we get the optimal matrices, the $UW^T X$ can be used to compute the classifier outputs for X. For X^j and the group partitions X_m^j and X_m, the outputs can be gotten with the corresponding optimal matrices including U^js, W^js and so on.

2.4. Computational complexity

In order to solve the problem Eq. (2) and optimize the GLMVML, in each iteration, we update one of the variables in $\{Z_m, U, V, W, Z_m^j, U^j, V^j, W^j, P, B^j\}$ with gradient descent and leave the others fixed. Thus, the computational complexity of GLMVML

Table 2
Detailed information of Mfeat data set.

View	No. instances	No. features	No. digits
fac	2000	216	10
fou	2000	76	10
kar	2000	64	10
pix	2000	240	10
zer	2000	47	10
mor	2000	6	10

Table 3
Detailed information of Reuters data set.

View	No. documents	Vocabulary size
EN	18,758	21,513
FR	26,648	24,839
GR	29,953	34,279
SP	12,342	11,547
IT	24,039	15,506
Topic	No. documents	Per(%)
C15	18,816	16.84
CCAT	21,426	19.17
E21	13,701	12.26
ECAT	19,198	17.18
GCAT	19,178	17.16
M11	19,421	17.39

Table 4
Detailed information of Corel data set.

View	No. instances	No. features	No. categories
Col-h	1000	32	10
Col-hl	1000	32	10
Col-m	1000	9	10
Coo-t	1000	16	10

Table 5
Detailed information of multi-label data sets.

Data set	No. instances	No. features	No. labels	label/instance
Arts	5000	462	26	1.64
Computers	5000	681	33	1.51
Entertainment	5000	640	21	1.42
Recreation	5000	606	22	1.42
Science	5000	743	40	1.45
Society	5000	636	27	1.69
Corel5k	5000	499	374	3.52
Business	5000	438	30	1.59
Education	5000	550	33	1.46
Health	5000	612	32	1.66
Reference	5000	793	33	1.17
Social	5000	1047	39	1.28
Enron	1702	1001	53	3.37
Image	2000	294	5	1.24

is depended on the ones of the update of these parameters. What's more, since the computational complexity of matrix multiplication is much larger than matrix subtraction, thus the computational complexity of the update for a variable is mainly depended on the computation of ∇_A rather than the computation of $A := A - \eta\nabla_A$. So, we can say that the computational complexity of GLMVML is finally depended on the computation of ∇_As.

Here, for convenience, we take the update of Z_m as instance. According to Eq. (4), the computational complexity of $UW^TXX^TWU^TZ_m$ is $O(lk^2d^2n^2d^2k^2l^2k)$ and the one of $UW^TX_mX_m^TWU^TZ_m$ is $O(lk^2d^2n_m^2d^2k^2l^2k)$, since in generally, n and n_m is always larger than l, d, k, thus the computational complexity of update of Z_m is $O(n^2+n_m^2)$. Then, for others, we can get the similar operations and the computational complexity of the update of U, V, W, Z_m^j, U^j, V^j, W^j, P, B^j are $O(n^2+n_m^2)$, $O(kn)$, $O(n^2+n_m^2)$, $O(n^2+n_m^{j\ 2})$, $O(n^2+n_m^{j\ 2})$, $O(k_jn_m^j)$, $O(n^2+n_m^{j\ 2})$, $O(n)$, $O(n^2+n_m^{j\ 2})$, respectively. Then the computational complexity of GLMVML is mostly equal to $O(3(n^2+\sum_{m=1}^{g}n_m^2)+kn+3\sum_{m=1}^{gj}\sum_{j=1}^{v}v(n^2+n_m^{j\ 2})+vn+v\sum_{m=1}^{gj}\sum_{j=1}^{v}k_jn_m^j)$ and since quadratic term is always larger than the linear term, thus the computational complexity of GLMVML can be written as $O(3(n^2+\sum_{m=1}^{g}n_m^2)+3\sum_{m=1}^{gj}\sum_{j=1}^{v}v(n^2+n_m^{j\ 2}))$. If the number of clusters are larger and n_m, n_m^j are smaller simultaneously, then the computational complexity of GLMVML will be reduced to $O(Gn^2)$ where G is a constant. In generally, this computational complexity is smaller than $O(n^3)$ which is the computational complexity of many traditional methods.

3. Experiments

In order to validate the performance of GLMVML, we adopt some benchmark data sets for experiments and the related experimental results are shown in Section 3.2.

3.1. Experimental setting

3.1.1. Data set

In our experiments, the used data sets include multi-view data sets, multi-label data sets, and multi-view multi-label data sets.

In terms of the used multi-view data sets, we adopt three classical ones. They are Mfeat, Reuters, and Corel [10]. (1) Mfeat[1] consists of hand written digits (0–9) [25] and each instance consists of six views, i.e., Fourier coefficients of the character shapes (fou), profile correlations (fac), Karhunen-Love coefficients (kar), pixel averages in 2×3 windows (pix), Zernike moments(zer), and morphological features (mor). Details of Mfeat can be found in Table 2. (2) Reuters[2] consists of machine translated documents which are written in five different languages and these languages are treated as five views [26,27]. These five languages are English (EN), French (FR), German (GR), Italian (IT), and Spanish (SP) and each document can be translated from one language to another language. Moreover, the documents are also categorized into six different topics. Details of Reuters can be shown in Table 3. (3) Corel[3] is extracted from a Corel image collection [25] and it consists of 68,040 photos from various categories. In our experiments, we randomly select 1000 photos from 10 categories and each category has 100 photos. The 10 categories are C0-Africa, C1-Beach, C2-Building, C3-Buses, C4-Dinosaurs, C5-Elephants, C6-Flowers, C7-Horses, C8-Mountains and C9-Food. For this data set, four views are adopted. They are color histogram (abbr. Col-h), color histogram layout (abbr. Col-hl), color moments (abbr. Col-m), and co-occurrence texture (abbr. Coo-t). Each view represents a feature set. Information of this data set is given in Table 4.

In terms of multi-label data sets, we adopt the same ones in [11], i.e., text data sets including eleven Yahoo data sets[4] (Arts, Business, Computers, Education, Entertainment, Health, Recreation, Reference, Science, Social and Society) and Enron data set[5] and image data sets[6] including Corel5k and Image. Table 5 shows infor-

[1] http://archive.ics.uci.edu/ml/datasets/Multiple+Features.
[2] http://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual%2C+Multiview+Text+Categorization+Test+collection
[3] http://archive.ics.uci.edu/ml/datasets/Corel+Image+Features.
[4] http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.
[5] http://mulan.sourceforge.net/datasets-mlc.html.
[6] http://cse.seu.edu.cn/people/zhangml/files/Image.rar.

mation of these multi-label data sets and here *label*/*instance* represents the average number of labels possessed by each instance.

For the multi-view multi-label data sets, we adopt NUS-WIDE data set [29]. This data set has six views, they are color histogram (dimensionality: 64), color correlogram (dimensionality: 144), edge direction histogram (dimensionality: 73), wavelet texture (dimensionality: 128), block-wise color moments extracted over 5×5 fixed grid partitions (dimensionality: 255), and bag of words based on SIFT descriptions (dimensionality: 500). Then this data set has 81 labels and 810 images are adopted as the instances in our experiments. Details can be found in [23,29].

3.1.2. Compared method

The compared methods include multi-view learning methods MVML [16], LMSC [9] and MLDL [18], multi-label learning methods LF-LPLC [21], MLCHE [22], and GLOCAL [11], multi-view multi-label learning methods MVMLP [23], SSDR-MML [24], and LSA-MML [8].

3.1.3. Parameter setting

For the compared methods, the parameter settings can be found in related work and in terms of the proposed GLMVML, in order to divide the data set into several groups, we adopt k-means and its parameter K which also determine the g and g^j is selected from the set $\{1, 2, 3, \ldots, 10\} \times l$ where l is the number of classes. For the Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j, we initialize them according to the X, X^j and their corresponding groups. Here, for GLOCAL which also need to divide the data set into several groups, we adopt the same setting with GLMVML. For parameters λ_0, λ_1, λ_2, and λ_3 which are used for the Frobenius norm regularizer are selected from the set $\{2^{-5}, 2^{-4}, \ldots, 2^0\}$. For the λ_4 and λ_5, (λ_6s and λ_7s) which correspond to the manifold regularizer for global and local label correlations, respectively are selected from the set $\{10^{-6}, 10^{-5}, \ldots, 10^0\}$. For λ_8 and λ_9 which are used for the reflection of complementary information from different views are selected from the set $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$.

In order to get the optimal results and according to the compared methods' demands, for each data set, we randomly select $\{10\%, 20\%, \ldots, 60\%\}$ for training and the rest for test. Then for multi-label data sets and NUS-WIDE, each instance or each view, we randomly remove $10\% \sim 30\%$ labels so as to get the observed label matrices. Then we repeat the experiments with each parameter combination for ten times and get the average results and the corresponding standard deviation. The best parameters are the ones whose average precision is the best. Then, the other performance indexes including the AUC (i.e., the area under the receiver operating characteristic (ROC) curve), running time and convergence are given with the optimal parameters. Here, we should notice that for each data set, different methods should process same data.

3.1.4. Experimental environment

Our experiments are conducted with the following environment. All the computations are performed on a node of compute cluster with 16 CPUs (Intel Core Due 3.0GHz) running RedHat Linux Enterprise 5 with 48GB main memory. The coding environment is MATLAB 2016. Furthermore, the maximum number of iterations is set to be 1000.

3.2. Experimental results

3.2.1. AUC And precision

In the experiments, we adopt AUC and precision to measure the performance of GLMVML and we know a higher AUC and precision denote a better performance. Tables 6 and 7 give the testing average AUC and precision respectively. From these tables, it is found that in terms of testing average AUC, our proposed GLMVML is better than other compared methods in average. Moreover, the win/tie/loss counts show that GLMVML is clearly superior to these multi-view learning methods and multi-label learning methods, as it wins for most times and never loses. For the other multi-view multi-label learning methods, they performs better than GLMVML sometimes. In terms of the average precision, we can get similar results.

3.2.2. Running time

Moreover, we want to show the running time of these methods. Table 8 shows the related experimental results and Avg. (mv) represents the average running time for multi-view data sets while Avg. (ml) represents the average running time for multi-label data sets. From this table, we find that our proposed GLMVML cost a little more running time which is also accepted by us.

3.2.3. Convergence

Since in GLOCAL, the related scholars empirically studied the convergence of GLOCAL [11], thus in this work, we adopt the same operation. Fig. 2 shows the objective value with respect to the number of iterations. For convenience and due to the lack of space, we only show the results on multi-view data set Mfeat, multi-label data sets Computers and Science, multi-view multi-label data set NUS-WIDE. As can be seen, the objective (i.e., Eq. (2)) converges quickly in a few iterations (less than 25). A similar phenomenon can be observed on the other data sets.

3.2.4. Influence of parameters

In our experiments, for the optimization problem Eq. (2), there are many parameters including λs and K should be chosen and tuned. Here, K (this parameter determine the number of g and g^j) is selected from $\{1, 2, 3, \ldots, 10\} \times l$ where l is the number of classes; for the Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j, they are initialized according to the X, X^j and the corresponding groups; λ_0, λ_1, λ_2, and λ_3 are selected from the set $\{2^{-5}, 2^{-4}, \ldots, 2^0\}$; λ_4, λ_5, λ_6s, and λ_7s are selected from the set $\{10^{-6}, 10^{-5}, \ldots, 10^0\}$; λ_8 and λ_9 are selected from the set $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. Since according to these parameters, we have too many parameter combinations and how to tune them is a challenge, especially, in real-world applications. Thus here, we show the influence of parameters and deal with the following questions. (1) Whether these parameters are data-dependent or not; (2) can we tune all these parameters well in practical experiments; (3) for different cases, how are the parameters Ks determined and what are the optimal values; (4) according to the parameters, in optimization problem Eq. (2), which part plays an important role for the performance of GLMVML; (5) what is the influence of the initialization of Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j on the performance of GLMVML.

We know among these parameter combinations, there is an optimal combination and this combination brings the best precision. Then under this combination, we can also get the corresponding AUC, convergence, and running time. Now we adopt other combinations and change the values of these different parameters so that we can see the changing of performances. Table 9 shows the influence of parameters on four performance indexes, i.e., AUC, precision, convergence, and running time. For convenience, we only show the influence on data set NUS-WIDE since for other data sets, we find the results are similar. In this table, ‘range’ indicates that when we change the value of a parameter, how the corresponding performance of an index changes. 'std' shows the standard deviation of values in the 'range'. 'gap' means the span of the 'range'. For example, for AUC and K, we can see the 'range' is [0.698,0.849]. This means expect for the optimal parameter combination, when we change the value of K, the AUC falls in the range [0.698,0.849], namely, the worst AUC is 0.698 and the best one is 0.849. Then the 'gap' is $0.849 - 0.698 = 0.151$ and the 'std' (0.052) is derived from [0.698,0.849].

Table 6
Testing average AUC (mean ± std.) of GLMVML with compared methods. •/◦ indicates that GLMVML is significantly better/worse than the corresponding method (pairwise t-tests at 95% significance level). The best average AUC for each data set is shown in bold. / represents no result since the related method cannot process that data set.

data sets	GLMVML	LMSC	MVML	MLDL	LF-LPLC
Mfeat	**0.796** ± **0.015**	0.751 ± 0.020 •	0.775 ± 0.018 •	0.778 ± 0.005 •	/
Reuters	**0.948** ± **0.015**	0.631 ± 0.011 •	0.894 ± 0.020 •	0.886 ± 0.028 •	/
Corel	**0.836** ± **0.001**	0.761 ± 0.013 •	0.798 ± 0.015 •	0.717 ± 0.014 •	/
Arts	0.887 ± 0.007	/	/	/	0.838 ± 0.005 •
Business	0.967 ± 0.003	/	/	/	0.926 ± 0.003 •
Computers	**0.928** ± **0.003**	/	/	/	0.842 ± 0.002
Education	**0.887** ± **0.008**	/	/	/	0.870 ± 0.006 •
Entertainment	**0.924** ± **0.007**	/	/	/	0.874 ± 0.005 •
Health	0.977 ± 0.010	/	/	/	0.914 ± 0.007 •
Recreation	**0.879** ± **0.000**	/	/	/	0.812 ± 0.000 •
Reference	**0.939** ± **0.004**	/	/	/	0.857 ± 0.004
Science	**0.868** ± **0.013**	/	/	/	0.831 ± 0.010 •
Social	**0.985** ± **0.005**	/	/	/	0.911 ± 0.005 •
Society	**0.891** ± **0.008**	/	/	/	0.800 ± 0.006 •
Enron	**0.927** ± **0.007**	/	/	/	0.861 ± 0.005
Corel5k	0.852 ± 0.006	/	/	/	0.796 ± 0.005 •
Image	**0.865** ± **0.013**	/	/	/	0.777 ± 0.009
NUS-WIDE	0.850 ± 0.027	/	/	/	/
win/tie/loss		3 / 0 / 0	3 / 0 / 0	3 / 0 / 0	10 / 4 / 0
data sets	MLCHE	GLOCAL	MVMLP	SSDR-MML	LSA-MML
Mfeat	/	/	0.726 ± 0.017	0.749 ± 0.017	0.712 ± 0.017 •
Reuters	/	/	0.820 ± 0.019 •	0.813 ± 0.020 •	0.739 ± 0.019 •
Corel	/	/	0.676 ± 0.014 •	0.711 ± 0.014 •	0.698 ± 0.014 •
Arts	0.851 ± 0.005 •	0.855 ± 0.005 •	0.865 ± 0.005	**0.897** ± **0.005**	0.881 ± 0.005
Business	0.885 ± 0.003 •	0.957 ± 0.003 •	0.937 ± 0.003	0.968 ± 0.003 ◦	**0.973** ± **0.003**
Computers	0.882 ± 0.002	0.883 ± 0.002 •	0.886 ± 0.002	0.897 ± 0.002 •	0.814 ± 0.002 •
Education	0.873 ± 0.006 •	0.874 ± 0.006 •	0.875 ± 0.006 •	0.872 ± 0.006 •	0.878 ± 0.006
Entertainment	0.799 ± 0.005 •	0.880 ± 0.005	0.879 ± 0.005	0.893 ± 0.005	0.832 ± 0.005 •
Health	0.868 ± 0.006 •	0.927 ± 0.007 •	0.952 ± 0.007 •	0.938 ± 0.007 •	**0.988** ± **0.008**
Recreation	0.764 ± 0.000 •	0.831 ± 0.000	0.832 ± 0.000 •	0.833 ± 0.000 •	0.812 ± 0.000 •
Reference	0.808 ± 0.004	0.893 ± 0.004 •	0.893 ± 0.004 •	0.904 ± 0.004 •	0.864 ± 0.004 •
Science	0.799 ± 0.009	**0.853** ± **0.010**	0.852 ± 0.010	0.843 ± 0.010 •	0.841 ± 0.010
Social	0.841 ± 0.005 •	0.924 ± 0.005	0.941 ± 0.005 •	0.939 ± 0.005	0.941 ± 0.005 •
Society	0.815 ± 0.006 •	0.839 ± 0.006 •	0.868 ± 0.006 •	0.869 ± 0.006 •	0.871 ± 0.006 •
Enron	0.799 ± 0.005	0.875 ± 0.005	0.891 ± 0.005	0.861 ± 0.005 •	0.866 ± 0.005 •
Corel5k	0.745 ± 0.005 •	0.818 ± 0.005 •	0.809 ± 0.005	0.811 ± 0.005 •	**0.877** ± **0.010**
Image	0.758 ± 0.009	0.810 ± 0.009	0.821 ± 0.008 •	0.826 ± 0.008 •	0.794 ± 0.008 •
NUS-WIDE	/	/	0.812 ± 0.031	0.815 ± 0.056	**0.893** ± **0.056** ◦
	9 / 5 / 0	8 / 6 / 0	9 / 9 / 0	12 / 5 / 1	11 / 6 / 1

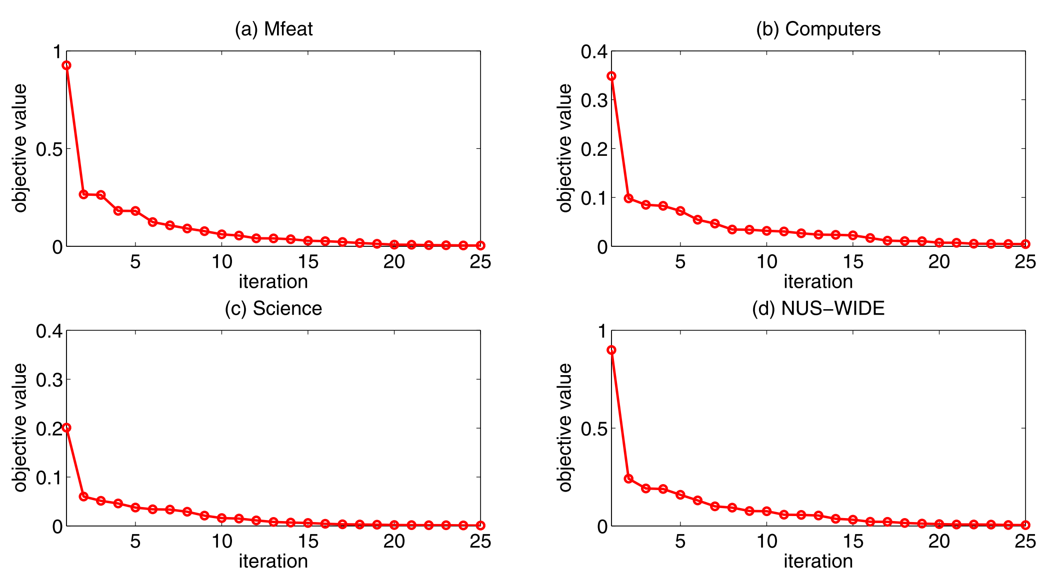


Fig. 2. Convergence of GLMVML on data sets Mfeat, Computers, Science, and NUS-WIDE.

Table 7
Testing average precision (mean ± std.) of GLMVML with compared methods. •/○ indicates that GLMVML is significantly better/worse than the corresponding method (pairwise t-tests at 95% significance level). The best average precision for each data set is shown in bold. / represents no result since the related method cannot process that data set.

data sets	GLMVML	LMSC	MVML	MLDL	LF-LPLC
Mfeat	**0.818 ± 0.024**	0.712 ± 0.011 •	0.725 ± 0.017 •	0.754 ± 0.005 •	/
Reuters	**0.972 ± 0.022**	0.674 ± 0.013 •	0.894 ± 0.020 •	0.850 ± 0.027 •	/
Corel	**0.864 ± 0.014**	0.776 ± 0.015 •	0.788 ± 0.014 •	0.696 ± 0.013 •	/
Arts	**0.647 ± 0.008**	/	/	/	0.579 ± 0.005 •
Business	**0.910 ± 0.004**	/	/	/	0.837 ± 0.004 •
Computers	0.717 ± 0.005	/	/	/	0.695 ± 0.004 •
Education	**0.643 ± 0.010**	/	/	/	0.597 ± 0.008 •
Entertainment	**0.701 ± 0.009**	/	/	/	0.665 ± 0.008 •
Health	0.785 ± 0.001	/	/	/	0.753 ± 0.001 •
Recreation	**0.654 ± 0.005**	/	/	/	0.588 ± 0.004 •
Reference	**0.739 ± 0.009**	/	/	/	0.677 ± 0.007 •
Science	**0.607 ± 0.010**	/	/	/	0.571 ± 0.008 •
Social	**0.823 ± 0.010**	/	/	/	0.714 ± 0.008 •
Society	**0.666 ± 0.013**	/	/	/	0.624 ± 0.009
Enron	0.674 ± 0.009	/	/	/	0.621 ± 0.006 •
Corel5k	**0.420 ± 0.005**	/	/	/	0.192 ± 0.004 •
Image	**0.824 ± 0.010**	/	/	/	0.760 ± 0.007 •
NUS-WIDE	**0.871 ± 0.011**	/	/	/	/
	win/tie/loss	3 / 0 / 0	3 / 0 / 0	3 / 0 / 0	13 / 1 / 0

data sets	MLCHE	GLOCAL	MVMLP	SSDR-MML	LSA-MML
Mfeat	/	/	0.700 ± 0.016 •	0.699 ± 0.016 •	0.686 ± 0.016 •
Reuters	/	/	0.773 ± 0.018 •	0.732 ± 0.018 •	0.733 ± 0.017 •
Corel	/	/	0.807 ± 0.013 •	0.748 ± 0.013 •	0.694 ± 0.012 •
Arts	0.600 ± 0.005 •	0.609 ± 0.005 •	0.614 ± 0.005 •	0.619 ± 0.004	0.645 ± 0.004
Business	0.827 ± 0.004 •	0.881 ± 0.004	0.871 ± 0.004 •	0.876 ± 0.004 •	0.909 ± 0.004
Computers	0.681 ± 0.004	0.702 ± 0.004 •	0.694 ± 0.004 •	0.704 ± 0.004	**0.739 ± 0.004** ○
Education	0.562 ± 0.008 •	0.621 ± 0.008 •	0.625 ± 0.007	0.611 ± 0.008	0.577 ± 0.008 •
Entertainment	0.616 ± 0.007 •	0.676 ± 0.008 •	0.684 ± 0.008 •	0.689 ± 0.008	0.652 ± 0.008 •
Health	0.748 ± 0.001	0.782 ± 0.001	0.782 ± 0.001	0.772 ± 0.001 •	**0.790 ± 0.001**
Recreation	0.582 ± 0.004 •	0.618 ± 0.004 •	0.641 ± 0.004 •	0.612 ± 0.004 •	0.645 ± 0.004
Reference	0.655 ± 0.007 •	0.694 ± 0.007 •	0.716 ± 0.007 •	0.677 ± 0.007 •	0.692 ± 0.007 •
Science	0.535 ± 0.008 •	0.587 ± 0.009 •	0.600 ± 0.009	0.599 ± 0.008 •	0.600 ± 0.008
Social	0.735 ± 0.007 •	0.750 ± 0.008 •	0.778 ± 0.008 •	0.766 ± 0.007 •	0.817 ± 0.007 •
Society	0.611 ± 0.008 •	0.631 ± 0.009 •	0.645 ± 0.009 •	0.652 ± 0.009 •	0.658 ± 0.009
Enron	0.616 ± 0.006	0.651 ± 0.006 •	0.659 ± 0.006 •	0.673 ± 0.006	**0.682 ± 0.006** ○
Corel5k	0.190 ± 0.004 •	0.198 ± 0.004	0.200 ± 0.004 •	0.201 ± 0.004 •	0.419 ± 0.014
Image	0.720 ± 0.007 •	0.798 ± 0.007 •	0.808 ± 0.007	0.802 ± 0.007 •	0.813 ± 0.006 •
NUS-WIDE	/	/	0.721 ± 0.015 •	0.803 ± 0.011 •	0.863 ± 0.011
	11 / 3 / 0	11 / 3 / 0	14 / 4 / 0	13 / 5 / 0	8 / 8 / 2

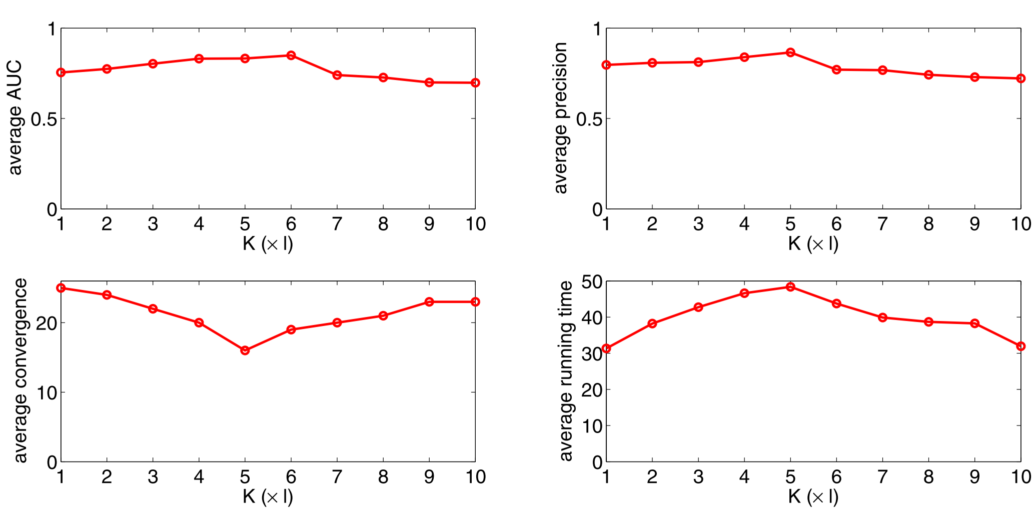


Fig. 3. Influence of K on NUS-WIDE with GLMVML.

Table 8
Running time (in seconds) of GLMVML with compared methods. / represents no result since the related method cannot process that data set.

data sets	GLMVML	LMSC	MVML	MLDL	LF-LPLC
Mfeat	24.19	20.59	20.30	22.24	/
Reuters	309.03	251.73	294.45	305.06	/
Corel	9.13	7.64	8.35	9.07	/
Avg. (mv)	114.12	93.32	107.70	112.12	/
Arts	53.77	/	/	/	47.13
Business	54.18	/	/	/	44.84
Computers	58.91	/	/	/	49.67
Education	48.88	/	/	/	42.44
Entertainment	58.41	/	/	/	54.77
Health	73.14	/	/	/	69.18
Recreation	55.92	/	/	/	49.90
Reference	84.46	/	/	/	76.51
Science	76.26	/	/	/	70.94
Social	95.87	/	/	/	79.50
Society	43.31	/	/	/	39.23
Enron	68.56	/	/	/	62.88
Corel5k	449.68	/	/	/	370.10
Image	15.06	/	/	/	13.94
Avg. (ml)	88.31	/	/	/	76.50
NUS-WIDE	41.03	/	/	/	/
data sets	MLCHE	GLOCAL	MVMLP	SSDR-MML	LSA-MML
Mfeat	/	/	20.82	22.46	24.07
Reuters	/	/	269.56	286.79	281.91
Corel	/	/	8.17	8.52	8.25
Avg. (mv)	/	/	99.51	105.92	104.74
Arts	44.46	49.02	49.28	49.07	51.05
Business	47.44	49.52	51.37	51.18	50.07
Computers	49.90	52.99	54.02	54.49	54.23
Education	44.71	44.94	45.07	47.01	45.60
Entertainment	56.16	56.33	56.40	59.05	58.10
Health	63.81	69.23	71.74	71.04	70.81
Recreation	49.43	51.62	54.19	53.22	52.88
Reference	77.26	81.23	85.22	83.80	82.76
Science	71.51	72.03	73.74	72.46	73.06
Social	78.81	86.50	88.35	90.15	88.62
Society	38.27	40.84	40.84	42.33	42.39
Enron	59.97	65.63	66.12	66.63	66.55
Corel5k	372.42	403.92	423.75	406.00	421.74
Image	14.19	14.93	15.56	14.94	14.95
Avg. (ml)	76.31	81.34	83.98	82.95	83.77
NUS-WIDE	/	/	36.47	37.63	39.39

Table 9
Influence of parameters in terms of the performance ranges, standard deviation, and the gap on AUC, precision, convergence, running time.

	AUC		Precision	
Parameter	Range / std	gap	range / std	Gap
K	[0.698,0.849] / ±0.052	0.151	[0.722,0.866] / ±0.043	0.144
λ_0	[0.830,0.850] / ±0.007	0.020	[0.852,0.869] / ±0.005	0.017
λ_1	[0.830,0.850] / ±0.007	0.020	[0.852,0.866] / ±0.005	0.014
λ_2	[0.834,0.849] / ±0.005	0.015	[0.851,0.868] / ±0.006	0.017
λ_3	[0.831,0.849] / ±0.005	0.018	[0.852,0.870] / ±0.006	0.018
λ_4	[0.767,0.848] / ±0.026	0.081	[0.773,0.869] / ±0.035	0.096
λ_5	[0.757,0.846] / ±0.026	0.089	[0.772,0.870] / ±0.032	0.098
λ_6	[0.760,0.850] / ±0.028	0.090	[0.793,0.865] / ±0.023	0.072
λ_7	[0.760,0.849] / ±0.026	0.089	[0.776,0.867] / ±0.031	0.091
λ_8	[0.831,0.848] / ±0.004	0.017	[0.852,0.870] / ±0.006	0.018
λ_9	[0.834,0.849] / ±0.004	0.015	[0.853,0.871] / ±0.006	0.018
Z_m	[0.830,0.850] / ±0.006	0.020	[0.855,0.870] / ±0.005	0.015
U	[0.838,0.850] / ±0.004	0.012	[0.852,0.870] / ±0.006	0.018
V	[0.831,0.848] / ±0.006	0.017	[0.851,0.871] / ±0.006	0.020
W	[0.830,0.849] / ±0.006	0.019	[0.856,0.870] / ±0.004	0.014
Z_r^j	[0.831,0.846] / ±0.005	0.015	[0.852,0.871] / ±0.005	0.019
U^j	[0.831,0.847] / ±0.006	0.016	[0.852,0.870] / ±0.006	0.018
V^j	[0.830,0.848] / ±0.005	0.018	[0.851,0.867] / ±0.006	0.016
W^j	[0.831,0.849] / ±0.005	0.018	[0.852,0.869] / ±0.006	0.017
P	[0.831,0.848] / ±0.006	0.017	[0.852,0.869] / ±0.005	0.017
B^j	[0.831,0.850] / ±0.007	0.019	[0.852,0.871] / ±0.006	0.019
	convergence		running time	
parameter	range / std	gap	range / std	gap
K	[16,25] / ±2.98	9	[31.34,48.38] / ±5.61	17.040
λ_0	[18,22] / ±1.76	4	[38.54,43.55] / ±2.02	5.010
λ_1	[18,21] / ±1.37	3	[38.05,43.88] / ±2.60	5.830
λ_2	[18,21] / ±1.21	3	[38.66,43.06] / ±1.60	4.400
λ_3	[20,22] / ±0.89	2	[38.31,43.52] / ±2.05	5.210
λ_4	[17,25] / ±3.42	8	[31.45,47.38] / ±6.38	15.930
λ_5	[16,25] / ±3.95	9	[37.14,49.46] / ±4.08	12.320
λ_6	[15,24] / ±3.51	9	[32.58,46.71] / ±4.87	14.130
λ_7	[15,25] / ±3.55	10	[31.25,47.11] / ±5.44	15.860
λ_8	[19,22] / ±1.64	3	[38.21,43.62] / ±2.61	5.410
λ_9	[18,22] / ±1.52	4	[39.13,42.99] / ±1.47	3.860
Z_m	[18,21] / ±1.17	3	[38.19,42.97] / ±1.33	4.780
U	[20,22] / ±0.67	2	[39.10,42.33] / ±1.04	3.230
V	[18,22] / ±1.15	4	[38.57,42.98] / ±1.46	4.410
W	[18,21] / ±1.03	3	[38.16,42.43] / ±1.36	4.270
Z_r^j	[19,22] / ±1.26	3	[38.24,43.96] / ±1.78	5.720
U^j	[18,22] / ±1.57	4	[38.34,43.16] / ±1.87	4.820
V^j	[18,22] / ±1.37	4	[38.53,43.17] / ±1.75	4.640
W^j	[18,22] / ±1.26	4	[41.42,43.76] / ±0.87	2.340
P	[18,21] / ±1.26	3	[38.80,42.59] / ±1.56	3.790
B^j	[18,22] / ±1.76	4	[38.15,43.74] / ±1.84	5.590

Now according to Table 9, we can see the influence of K, λ_4, λ_5, λ_6, and λ_7 is larger, so we adopt Table 10 and Fig. 3 for further revelation. In Fig. 3, $K \times l$ represents the label value A equals to $A \times l$. Of course, in Table 10 and Fig. 3, the results are also not derived from the optimal parameter combination so as to reveal the influence well. Then according to the experimental results given in Tables 9, 10, and Fig. 3, we can draw the following conclusions and answer the above mentioned questions. (1) The influence of K, λ_4, λ_5, λ_6, and λ_7 is larger than others since for other parameters, no matter which initial value is adopted, the performance gaps for AUC and precision is always not larger than 2%, and the ones for convergence and running time are also smaller. In other words, for λ_0, λ_1, λ_2, λ_3, λ_8, λ_9, Z_m, U, V, W, Z_r^j, U^j, V^j, W^j, P, B^j, when we change their values, the performances of AUC, precision, convergence, and running time won't change too much. But for others, the influence is a little larger; (2) according to the Table 10, it is found that except for the running time, with the increasement of parameters λ_4, λ_5, λ_6, and λ_7, the performance of GLMVML improves (for convergence, the smaller the iteration number is, the better the performance is) while when these parameters are very large, the performance deteriorates due to considering too many or too few local label correlations will both deteriorate the performance. Our GLMVML is a model with both global and local label correlations, so considering balanced global and local label correlations is important to improve its performance. Thus in this table, we can draw a conclusion that setting λ_4, λ_5, λ_6, and λ_7 be 10^{-3} is much more feasible. Although in this case, setting them be 10^{-3} brings a longer running time, but combining the performances of AUC, precision, convergence, a little more running time is also accepted by us; (3) according to Fig. 3, it is found that expect for the running time, with more clusters (i.e., larger K), performance improves as more local label correlations are taken into account. While if the number of clusters is too large, this makes each cluster possess very few instances, and the local label correlations cannot be reliably estimated. As a result, the performance starts to deteriorate. Thus, according to this figure, we can draw another conclusion that setting K be $5 \times l$ is feasible. Similar with the previous conclusion, although in this case, running time is increased, but for the purpose of classification, this increasement is also accepted by us; (4) according to the previous two conclusions, we can see that although there are too many parameters should be tuned, we can still tune them well in practical experiments due to some parameters have small influence and some parameters can be set a feasible initial value; (5) although we only show the in-

Table 10
Influence of parameters $\lambda_4 \sim \lambda_7$ in terms of AUC, precision, convergence, running time.

AUC	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
λ_4	0.813	0.820	0.832	0.848	0.791	0.782	0.767
λ_5	0.801	0.813	0.824	0.846	0.795	0.784	0.757
λ_6	0.811	0.830	0.844	0.850	0.808	0.785	0.760
λ_7	0.800	0.814	0.823	0.849	0.795	0.790	0.760
Precision	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
λ_4	0.844	0.858	0.867	0.869	0.814	0.789	0.773
λ_5	0.835	0.845	0.847	0.870	0.819	0.789	0.772
λ_6	0.841	0.850	0.856	0.865	0.837	0.822	0.793
λ_7	0.832	0.837	0.858	0.867	0.828	0.792	0.776
Convergence	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
λ_4	25	24	21	18	20	22	24
λ_5	25	23	20	16	17	18	25
λ_6	24	23	18	15	16	17	21
λ_7	25	21	18	15	16	20	22
Running time	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
λ_4	32.05	42.47	45.06	47.38	40.06	31.45	34.63
λ_5	39.73	42.33	42.39	49.46	46.49	43.11	37.14
λ_6	32.58	34.94	40.40	46.71	43.66	39.00	37.53
λ_7	31.25	32.95	39.45	47.11	35.33	33.93	33.53

Table 11
Detailed comparison between our proposed method and other compared ones in terms of the increase proportion.

Multi-view			
index	LMSC	MVML	MLDL
AUC	14.56%	3.75%	6.63%
Precision	16.40%	8.22%	11.81%
Running time	22.28%	5.96%	1.78%
multi-label			
index	LF-LPLC	MLCHE	GLOCAL
AUC	6.17%	9.20%	3.97%
precision	6.71%	8.08%	4.38%
running time	15.44%	15.73%	8.58%
multi-view multi-label			
Index	MVMLP	SSDR-MML	LSA-MML
AUC	4.83%	4.26%	5.16%
Precision	5.65%	6.12%	4.01%
Running time	7.22%	6.79%	6.12%

fluence on data set NUS-WIDE, for other data sets, we can still get similar results. This means that the setting of these parameters are data-independent; (6) according to these tables, we know due to the influence of K, λ_4, λ_5, λ_6, and λ_7 is larger and K always decides the number of clusters which also influences the local label correlation, thus in optimization problem Eq. (2), global and local label correlations plays an important role for the performance of GLMVML. Of course, generally speaking, since we can get the feasible selections for these parameters, thus in practical experiments, tuning parameters is not a difficult task.

3.2.5. Summary of increase proportion of performance

Table 11 shows the detailed comparison between our proposed method and other compared ones in terms of the increase proportion. The value in this table means that for an index, compared with the other method, how many increase proportion does the GLMVML get. For example, for LMSC and AUC, 14.56% indicates that compared with LMSC, our GLMVML brings a better AUC and the increase proportion is 14.56%. Then according to this summary table, we can see that with the consideration of global and local label correlation and the complementary information from different views, our GLMVML can bring a better performance. Although GLMVML needs a longer running time, but in practice experiments, a higher classification accuracy is always the priority target. What's more, if we improve the experimental environment and adopt GPU, we can imagine that the increase proportion of running time will be smaller.

4. Conclusion and future work

Multi-view multi-label learning methods are developed to process multi-view multi-label data sets which are widely used in our real-world applications. But for the global and local label correlations of the data sets, these traditional methods cannot exploit simultaneously. Furthermore, these methods cannot reflect the complementary information from different views. Thus, in this paper, we propose global and local multi-view multi-label learning (GLMVML). In GLMVML, it divides the whole data set and each view into several clusters with some clustering methods so as to consider both the global and local label correlations. Moreover, GLMVML introduces a consensus multi-view representation to encode the complementary information from different views. Experiments on some multi-view, multi-label, and multi-view multi-label data sets have validated that (1) in terms of AUC and precision, GLMVML is better than other compared methods in average; (2) according to the win/tie/loss counts, GLMVML is superior to the compared multi-view learning methods and multi-label learning methods; (3) although the model of GLMVML is more complicate than others, only a small amount of running time is added; (4) the optimization of GLMVML can be converged in a few iterations; (5) some parameters of GLMVML have influence on the performance of GLMVML while some haven't. Setting a moderate parameter value (not too large and not too small) is important to get a better performance and the tuning of parameters is data-independent and not difficult.

But according to the optimization and experiments of our GLMVML, it is found that we show the convergence empirically rather than theoretically. Indeed, proving the GLMVML and some other similar models convergence in theory is difficult and complex, and many scholars always adopt an alternative, i.e., empirically show the convergence [8,9,11]. Although in some references, for example [4], some scholars prove similar parts of Eq. (2), i.e., $||J \circ (Y - UV)||_F^2 + \lambda_0 \left|\left|V - W^T X\right|\right|_F^2 + \lambda_1 \Re(U, V, W, U^j, V^j, W^j, P, B^j) + \sum_{j=1}^{v} (\lambda_2 \left|\left|J^j \circ (Y^j - U^j V^j)\right|\right|_F^2 + \lambda_3 \left|\left|V^j - {W^j}^T X^j\right|\right|_F^2)$ can converge to a global optimum in theory, but for the whole Eq. (2), it is hard to prove its convergence in theory. Thus, here, we also empirically show the convergence. In the future work, we will find a method to show the convergence in theory.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is sponsored by 'Chenguang Program' supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission under grant number 18CG54. Furthermore, this work is also supported by National Natural Science Foundation of China (CN) under grant number 61602296, Natural Science Foundation of Shanghai (CN) under grant number

16ZR1414500, Project funded by China Postdoctoral Science Foundation under grant number 2019M651576, and the authors would like to thank their supports.

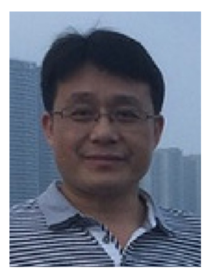
References

[1] H. Jiang, J. Yi, S. Chen, X. Zhu, A multi-objective algorithm for task scheduling and resource allocation in cloud-based disassembly, J. Manuf. Syst. 41 (2016) 239–255.
[2] Q.Q. Fan, W.L. Wang, X.F. Yan, Multi-objective differential evolution with performance-metric-based self-adaptive mutation operator for chemical and biochemical dynamic optimization problems, Appl. Soft Comput. 59 (2017) 33–44.
[3] P. Fan, R.G. Zhou, N. Jing, H.S. Li, Geometric transformations of multidimensional color images based on NASS, Inf. Sci. 340 (2016) 191–208.
[4] X.F. Zhu, X.L. Li, S.C. Zhang, Block-row sparse multiview multilabel learning for image classification, IEEE Trans. Cybern. 46 (2) (2016) 450–461.
[5] Y. Luo, D.C. Tao, C. Xu, C. Xu, H. Liu, Y.G. Wen, Multiview vector-valued manifold regularization for multilabel image classification, IEEE Trans. Neural Netw. Learn. Syst. 24 (5) (2013) 709–722.
[6] C.P. Hou, C.S. Zhang, Y. Wu, F.P. Nie, Multiple view semi-supervised dimensionality reduction, Pattern Recognit. 43 (3) (2010) 720–730.
[7] Y.H. Yuan, Q.S. Sun, H.W. Ge, Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition, Pattern Recognit. 47 (3) (2014) 1411–1424.
[8] C.Q. Zhang, Z.W. Yu, Q.H. Hu, P.F. Zhu, X.W. Liu, X.B. Wang, Latent semantic aware multi-view multi-label classification, in: Proceedings of the thirty-second AAAI Conference on Artificial Intelligence, 2018, pp. 4414–4421.
[9] C.Q. Zhang, H.Z. Fu, Q.H. Hu, X.C. Cao, Y. Xie, D.C. Tao, D. Xu, Generalized latent multi-view subspace clustering, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, doi:10.1109/TPAMI.2018.2877660.
[10] Y.M. Xu, C.D. Wang, J.H. Lai, Weighted multi-view clustering with feature selection, Pattern Recognit. 53 (2016) 25–35.
[11] Y. Zhu, J.T. Kwok, Z.H. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 99 (2017) 1–24.
[12] M. Boutell, J. Luo, X. Shen, C. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.
[13] N. Ueda, K. Saito, Parametric mixture models for multi-labeled text, Adv. Neural Inf. Process. Syst. 15 (2002) 721–728.
[14] D. Turnbull, L. Barrington, D. Torres, C. Lanckriet, Semantic annotation and retrieval of music and sound effects., IEEE Trans. Audio Speech Lang. Process. 16 (2) (2008) 467–476.
[15] G. Tzortzis, A. Likas, Kernel-based weighted multi-view clustering, in: Proceedings of the 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 675–684.
[16] S.L. Sun, Q.J. Zhang, Multiple-view multiple-learner semi-supervised learning, Neural Process. Lett. 34 (2011) 229–240.
[17] M.Q. Deng, C. Wang, Q.F. Chen, Human gait recognition based on deterministic learning through multiple views fusion, Pattern Recognit. Lett. 78 (C) (2016) 56–63.
[18] F. Wu, X.Y. Jing, X.G. You, D. Yue, R.M. Hu, J.Y. Yang, Multi-view low-rank dictionary learning for image classification, Pattern Recognit. 50 (2016) 143–154.
[19] S.H. Zhu, X. Sun, D.L. jin, Multi-view semi-supervised learning for image classification, Neurocomputing 208 (2016) 136–142.
[20] H.Y. Wang, X. Wang, J. Zheng, J.R. Deller, H.Y. Peng, L.Q. Zhu, W.G. Chen, X.L. Li, R.J. Liu, H.J. Bao, Video object matching across multiple non-overlapping camera views based on multi-feature fusion and incremental learning, Pattern Recognit. 47 (12) (2014) 3841–3851.
[21] W. Weng, Y.J. Lin, S.X. Wu, Y.W. Li, Y. Kang, Multi-label learning based on label-specific features and local pairwise label correlation, Neurocomputing 273 (2018) 385–394.
[22] V. Kumar, A.K. Pujari, V. Padmanabhan, S.K. Sahu, V.R. Kagita, Multi-label classification using hierarchical embedding, Expert Syst. Appl. 91 (2018) 263–269.
[23] Z.Y. He, C. Chen, J.J. Bu, P. Li, D. Cai, Multi-view based multi-label propagation for image annotation, Neurocomputing 168 (2015) 853–860.
[24] B.Y. Qian, X. Wang, J.P. Ye, I. Davidson, A reconstruction error based framework for multi-label and multi-view learning, in: Proceedings of the IEEE Transactions on Knowledge and Data Engineering, volume 27, 2015, pp. 594–607.
[25] A. Asuncion, D. Newman, UCI machine learning repository, 2007, http://archive.ics.uci.edu/ml/.
[26] M.R. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views an application to multilingual text categorization, Neural Inf. Process. Syst. (NIPS) (2009) 28–36.
[27] http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm.
[28] D. Luo, C. Ding, H. Huang, T. Li, Non-negative Laplacian embedding, in: Proceedings of the 9th IEEE International Conference on Data Mining, 2009, pp. 337–346.
[29] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-wide: a real-world web image database from national university of Singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, p. 48.

Changming Zhu gets his Ph.D. degree at East China University of Science and Technology in 2015. Now he is a Professor in Shanghai Maritime University. His research interesting include pattern recognition, image process, and multi-view learning.

Duoqian Miao is a Professor in Tongji University, Vice President of International Rough Set Society (IRSS), Chair of CAAI Granular Computing and Knowledge Discovery Technical Committee, Vice President of Shanghai Computer Federation, etc. He has hold many projects including National Natural Science Foundation of China, Research Fund for the Doctoral Program of Higher Education of China. His research interesting include pattern recognition, image process, and Rough Set Theory.

Zhe Wang is a Professor in East China University of Science and Technology and he is also the Shanghai Shuguang Scholar. His research interesting also include pattern recognition, image process, and multi-view learning.



Rigui Zhou is a Professor in Shanghai Maritime University and an Excellent Talent of the Ministry of Education for the New Century. His research interesting include pattern recognition, image process, and quantum processing.

Lai Wei is a Professor in Shanghai Maritime University and his research interesting include pattern recognition and granular computing.

Xiafen Zhang is Professor in Shanghai Maritime University and her research interesting include calligraphy recognition.