AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ1: Please confirm or add details for any funding or financial support for the research of this article.

AQ2: Please provide the postal code for George Mason University, Fairfax, VA, USA, and King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

AQ3: Please provide the department name for King Abdullah University of Science and Technology.

AQ4: Please provide the complete details and exact format for Reference [1].

AQ5: Please confirm if the location and publisher information for Reference [47] is correct as set.

AQ6: Please provide the educational background for the author C. Domeniconi.

Individuality- and Commonality-Based Multiview Multilabel Learning

Qiaoyu Tan, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang

***Abstract*—In multiview multilabel learning, each object is represented by several heterogeneous feature representations and is also annotated with a set of discrete nonexclusive labels. Previous studies typically focus on capturing the shared latent patterns among multiple views, while not sufficiently considering the diverse characteristics of individual views, which can cause performance degradation. In this article, we propose a novel approach [individuality- and commonality-based multiview multilabel learning (ICM2L)] to explicitly explore the individuality and commonality information of multilabel multiple view data in a unified model. Specifically, a common subspace is learned across different views to capture the shared patterns. Then, multiple individual classifiers are exploited to explore the characteristics of individual views. Next, an ensemble strategy is adopted to make a prediction. Finally, we develop an alternative solution to jointly optimize our model, which can enhance the robustness of the proposed model toward rare labels and reinforce the reciprocal effects of individuality and commonality among heterogeneous views, and thus further improve the performance. Experiments on various real-word datasets validate the effectiveness of ICM2L against the state-of-the-art solutions, and ICM2L can leverage the individuality and commonality information to achieve an improved performance as well as to enhance the robustness toward rare labels.**

***Index Terms*—Commonality, ensemble classification, individuality, multilabel learning, multiview learning.**

I. Introduction

IN MANY real-world applications, data are often associated with several heterogeneous feature representations, each of which gives a different view of the data. For example, a news Web page can be represented by two heterogeneous views, one is from the text (and image) information of the Web page itself, and the other is from the hyperlink to other pages; an image can be described using different features, such as texture descriptors, shape descriptors, color descriptors, surrounding texts, and so on. As a natural formulation for this type of data, multiview learning has attracted a lot of attention in machine learning and in various application domains [1], [2].

Although diverse multiview learning methods have been proposed in the literature over the past years, they still have some limitations. On the one hand, most previous studies often assume that each sample is annotated with a single label [3], [4]. Nevertheless, in real-life applications, individual samples usually have more than one label. For instance, an image can be simultaneously annotated with several labels, such as sea, sky, and seagull; a Web page could be tagged with multiple topics given as labels, such as economics, culture, sports, and politics. On the other hand, the majority of the existing studies are supervised approaches that require a large number of labeled samples [5], [6]. In practice, nevertheless, it is rather difficult and expensive to collect labeled samples, while unlabeled samples are easy to accumulate. Given this, a few semisupervised multiview multilabel learning approaches [7], [8] have been proposed to leverage limited labeled and abundant unlabeled samples. The key motivation behind them is to capture the complementary patterns among multiple views, which can boost the performance of multiview learning [9].

Another limitation of the aforementioned methods is that they do not explicitly account for the distinctive information of individual views, which might degenerate their performance for a variety of reasons. First, with respect to features, since multiview samples have heterogeneous feature representations, in which each representation encodes different properties of the samples, they may fail to capture the global structure of multiview data without exploring the distinctive information hidden in individual views. Second, with respect to labels, since each individual view captures a specific property of data, it is impossible for one view to comprehensively characterize all the relevant labels, especially, when data are annotated with multiple labels. As a result, leveraging the *individuality* of each view, along with the *commonality* of multiple views may further improve the performance of the model, compared to focusing only on the individuality (or commonality) of the views.

Some researchers have already explored the commonality and individuality of multiview data for classification and clustering [10]–[15]. It has been shown that the utilization of individual and shared patterns is beneficial for latent representation learning [11], [12]; multioutput problem [15]; and

Manuscript received April 23, 2019; revised August 29, 2019; accepted October 28, 2019. This work was supported in part by the Natural Science Foundation of China under Grant 61872300 and Grant 61873214, in part by the Fundamental Research Funds for the Central Universities under Grant XDJK2019B024, in part by the Natural Science Foundation of CQ CSTC under Grant cstc2018jcyjAX0228, and in part by the King Abdullah University of Science and Technology, Saudi Arabia. This article was recommended by Associate Editor S. Ventura. *(Corresponding author: Guoxian Yu.)*

Q. Tan, G. Yu, and J. Wang are with the College of Computer and Information Sciences, Southwest University, Chongqing 400715, China (e-mail: qiaoyut@gmail.com; gxyu@swu.edu.cn; kingjun@swu.edu.cn).

C. Domeniconi is with the Department of Computer Science, George Mason University, Fairfax, VA, USA (e-mail: carlotta@cs.gmu.edu).

X. Zhang is with the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia (e-mail: xiangliang.zhang@kaust.edu.sa).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2950560

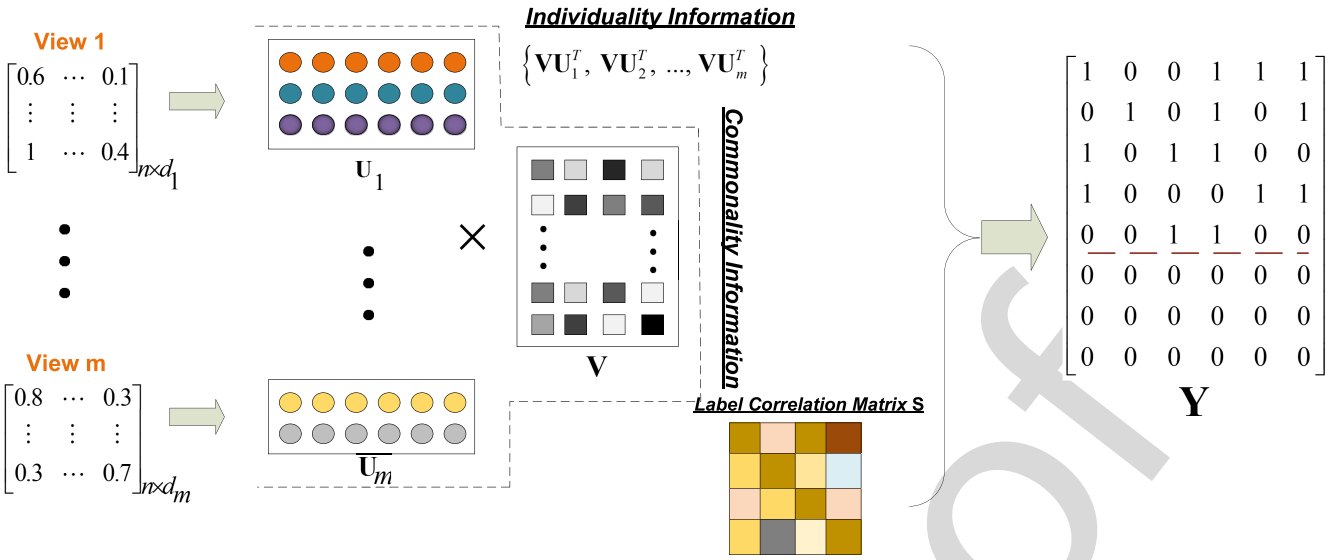


Fig. 1. Method overview. ICM2L jointly leverages commonality and individuality information of multiple data views, as well as label correlations. The commonality information of heterogeneous views is captured by the shared low-dimensional representation $\mathbf{V}$; while the individuality information of multiple views is captured by the reconstructed data matrices $\{\mathbf{VU}_v\}_{v=1}^{m}$.

multilabel classification [10]. In the multiview multilabel classification scenarios, however, they may result in suboptimal classifiers due to the isolated learning of hidden features and multilabel classifier [11], [12], or the lack of capacity to capture label correlations [10], [15]. More important, these methods mainly focus on learning a general classifier for all labels and treat them equally, which ignore the essential label imbalance problem in multilabel learning [16], [17]. To this end, the learned classifier may lose discriminant ability on rare labels, which are widely witnessed in the real-world applications. Motivated by this, in this article, we propose to explore how individual and common patterns of multiview data could be utilized to improve the performance of multilabel classification as well as in advancing the robustness of classifier toward rare labels.

To address the aforementioned issues, we propose a novel approach, called individuality- and commonality-based multiview multilabel learning (ICM2L), to explicitly account for the individual and shared patterns hidden in different views. As shown in Fig. 1, given multiple heterogeneous features of the input data, ICM2L seeks a shared subspace across heterogeneous views, which captures the commonality of different views and adapts an ensemble classifier-based on the shared subspace and on the other individual feature spaces, as well as on the label information in a unified model. In this way, both the shared and view-specific information of different views could be used to boost the performance via a mutually beneficial effect and, thus, further improve the performance of the model. The main contributions of this article are summarized as follows.

1) The proposed ICM2L can explicitly and jointly employ the individuality and commonality information of multiview multilabel data. It learns a shared subspace from different views, label correlations, and an ensemble classifier based on individual and shared feature spaces in a unified model.
2) ICM2L can explore the individuality of multiple views; as a result, it is superior to other methods in discovering rare labels.
3) We develop an alternative optimization solution to iteratively optimize our model. Extensive empirical results on the benchmark datasets demonstrate the superiority of ICM2L with respect to related and competitive methods, such as lrMMC [7], LSML [8], CSMSC [13], MLAN [4], and SMMCL [18].

The remainder of this article is organized as follows. In Section II, we briefly introduce the related work. Section III presents the proposed ICM2L. The experimental results and conclusions are discussed in Sections IV and V, respectively.

II. Related Work

This article is related to two branches of studies: 1) multilabel learning and 2) multiview learning. In this section, we briefly review some related works in these two fields. For more details, please refer to [2] and [19].

A. Multilabel Learning

Different from the binary classification scenarios, where each sample is associated with only one single semantic label, multilabel learning aims at assigning a set of discrete nonexclusive labels to a sample and has received increasing interest in different machine learning tasks [20]. For instance, Huang *et al.* [21] and [22] assumed that fully supervised signals are available and focus on learning multilabel classifiers under supervised setting. Such assumption, however, may not hold in real-world applications, because it requires exhaustive efforts to annotate multilabel samples. To avoid this limitation, researchers have resorted to develop semisupervised multilabel classifiers [23]–[27], in which limited labeled samples, as well as abundant unlabeled samples, are jointly used for training. Besides, considering the fact that labeled data

is tagged by human efforts, they might have some missing or noisy labels [28]–[33], several approaches have been proposed to design multilabel classifiers under weak-label setting [28], [29], [34], [35] or with noisy labels [32], [36]–[38].

Although the aforementioned methods have achieved the state-of-the-art performance for multilabel data, they mainly emphasize on single-view data and are not ready for multiview data. In fact, it has been proved that directly applying the existing multilabel algorithms to multiview data by concatenating multiple feature vectors (views) together will result in a compromised performance [2], [7]. The reason is that such concatenation operation fails to explore the intraview and interview relationships across heterogeneous views, which is very important for successful multiview learning models [39]. In addition, given that label correlation is crucial for the success of multilabel learning [40], how to develop an effective algorithm that can jointly utilize the heterogeneous information as well as important label correlations among multiview multilabel data still remains a challenge.

B. Multiview Learning

Due to the ubiquity of multiview data, multiview learning has been an active research field in many real-world applications [2]. Several multiview learning approaches were proposed to analyze multiview multilabel data recently. For example, Nie *et al.* [4] proposed a nearly parameter-free multiview model MLAN by integrating semisupervised classification and local structure learning simultaneously. Liu *et al.* [7] proposed a matrix factorization-based multiview framework lrMMC, which first seeks a shared representation of multiple views and then conducts classification based on matrix completion on the shared feature space. Nevertheless, lrMMC models the fusion of multiple views and the follow-up prediction tasks as separate objectives, which may lead to suboptimal solution. To avoid such a risk, some unified multiview multilabel learning methods have been proposed [8], [41]. Specifically, Tan *et al.* [41] aimed to improve the multilabel prediction performance by seeking a shared subspace from incomplete views with weak labels, local label correlations, and a predictor in this subspace in a unified model. Zhang *et al.* [8] sought a common feature representation and the corresponding projection model between the learned subspace and labels by simultaneously enforcing the consistence of latent semantic bases among different views in the kernel spaces. However, although such unified models can utilize the shared and complimentary information among different views to some extent, they do not explicitly take into account the individual patterns of heterogeneous views. As a result, they may have a compromised performance.

It has been recognized that exploring individuality and commonality of heterogeneous features can further boost the performance of multiview data mining [10]–[12]. Sagonas *et al.* [11] and Hu *et al.* [12] investigated the benefit of individual and common patterns in multiview data to learn more discriminant low-dimensional representations. However, they both focus on representation learning and cannot be directly applied to the multilabel classification problem. Similarly, [13] and [14] separately learn the individual and common representations of multiview data and then conduct clustering on the merged representations. They not only ignore the important correlations of multilabel data but also suffer from a two-stage fashion that may result in the suboptimal problem. To bridge the gap, Liu *et al.* [10] introduced a unified model to jointly learn compact latent features and multilabel classifier. Specifically, it assumes that the final latent feature vector with size d is composed of multiple view-specific features with size d_s and one shared common feature vector with size d_c, such that $d = md_s + d_c$, where m denotes the number of views. With this assumption, this unified model is able to generate comprehensive feature representations that can capture both individual and common patterns of multiview data. However, it still targets on developing a general multilabel classifier for all labels and may lose discriminant ability toward rare labels, which refers to the label imbalance problem in multilabel classification. Moreover, it also cannot capture the crucial correlations between multiple labels.

One similar work with our ICM2L is SMMCL [18], which is a self-pace-based label propagation method that models the common consensus and individuality of multiple teacher classifiers, each for one view. Although SMMCL can iteratively propagate labels to the most informative candidate unlabeled samples during training, it suffers from the scalable issue and cannot work on a moderate(large)-scale data as shown in our experiments. Besides, SMMCL also ignores the label correlations of multilabeled data, which is the cornerstone for successful multilabel learning algorithms [40]. Another core distinction between SMMCL and our model is that SMMCL uses individual information of multiview data to discover the most informative unlabeled samples, while our method utilizes such information to improve the discriminant ability toward rare labels and the robustness. Extensive empirical results on the benchmark datasets demonstrate the superiority of ICM2L to these related competitive methods [4], [7], [8], [18], [42].

III. Proposed Approach

A. Problem Statement and Notations

Suppose $\mathcal{X} = \{\mathbf{X}_v\}_{v=1}^m$ represents a dataset with n samples and m views, where $\mathbf{X}_v = [\mathbf{x}_v^1, \mathbf{x}_v^2, \ldots, \mathbf{x}_v^n] \in \mathbb{R}^{n \times d_v}$ indicates the full feature space in view v. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T \in \{-1, 1\}^{n \times c}$ is the corresponding label matrix, where $\mathbf{y}_i \in \{-1, 1\}^c$ is the label vector of $\mathbf{x}_i$ and c is the number of distinct labels. $\mathbf{y}_{ic'} = 1$ $(c' = 1, \ldots, c)$ means the c'th label is relevant; otherwise, $\mathbf{y}_{ic'} = -1$. Without loss of generality, we assume that out of n samples, the first l are labeled samples, while the remaining u are unlabeled samples. Our goal is to learn a predictive label matrix $\hat{\mathbf{Y}} \in \{1, -1\}^{n \times c}$ from heterogeneous feature representations $\{\mathbf{X}_v\}_{v=1}^m$ and partial label matrix $\mathbf{Y}$.

B. Problem Formulation

An intuitive strategy to deal with multiview multilabel setting is to concatenate multiview features into a single vector, and then conduct classification based on the classical

multilabel learning algorithms [19]. Such concatenation, however, ignores the fact that features are extracted from different spaces with different statistical properties, and directly applying multilabel methods on the concatenated data may suffer from the overfitting problem, and often leads to high time complexity, which may be unacceptable in the real-world applications. Besides, feature concatenation also ignores the useful diversity information across different views and might lead to a suboptimal problem [8].

In order to take advantage of consensus information, we advocate to seek the shared subspace from different views based on the matrix factorization techniques [43]. Specifically, we resort to non-negative matrix factorization (NMF) [44], [45] for its power in extracting the representations of diverse data [46]. It is worth noting that the major difference between NMF and other matrix factorization methods, such as singular-value decomposition [47], is the non-negative constraints, which helps to obtain a part-based representation, as well as enhancing the interpretability of the learned subspace. Another reason for the choice of NMF is due to the fact that most multiview data are naturally non-negative or can be easily transformed into non-negative ones, without changing the proximity between the original data. Our model can also be adapted to mix-sign data by replacing NMF with other matrix factorization techniques (i.e., semi-NMF [48]). Besides, our method is flexible for basic matrix factorization techniques, thus more powerful algorithms, that is, deep matrix factorization models [49], are expected to further improve the performance. Given heterogeneous representations of data $\{\mathbf{X}_v\}_{v=1}^m$ and the label matrix $\mathbf{Y}$, our unified objective function is given as follows:

$$\begin{aligned}\mathcal{L} &= \sum_{v=1}^{m} \left\|\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T\right\|_F^2 + \alpha\|\mathbf{V}\mathbf{W}_0 - \mathbf{Y}\|_F^2 + \beta\|\mathbf{W}_0\|_F^2 \\ &\text{s.t. } \mathbf{U} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{1}$$

where $\mathbf{U}_v \in \mathbb{R}^{d_v \times k}$ denotes the individual matrix for the vth view; $\mathbf{V} \in \mathbb{R}^{n \times k}$ is the shared subspace, where k is the desired low-rank size; $\|.\|_F$ represents the Frobenius norm; and $\mathbf{U}_v \geq 0$ and $\mathbf{V} \geq 0$ are the non-negative constraints for the matrices. $\mathbf{W}_0 \in \mathbb{R}^{k \times c}$ is the coefficient matrix corresponding to $\mathbf{V}$, and α and β are two tradeoff parameters. The first part of (1) is the consensus term, which aims to seek the common low-dimensional representation of multiview data under the assumption that different views have distinct mapping matrices $\{\mathbf{U}_v\}_{v=1}^m$, but share the same latent feature space $\mathbf{V}$. For this reason, we can make use of the consensus information across multiviews to some extent. The second part of (1) is the standard empirical loss term. It measures the empirical loss on labeled samples and ensures that the predicted label matrix is consistent with the initial ground-truth label assignment. By jointly optimizing the two terms, we can exploit the label information $\mathbf{Y}$ to induce the shared subspace toward a semantic label space. It not only helps to obtain a discriminate subspace but also may alleviate the widely spread semantic gap [50] between the input heterogeneous feature spaces and the semantic label space, since $\mathbf{V}$ can be viewed as a bridge between them. The last part is the widely used ℓ_2-norm regulation term, it is included to avoid the overfitting problem and reduce the impact of noisy features.

By minimizing (1), we can learn a discriminate predictor by jointly using both shared information among views and label information of labeled samples. But the predictor will not be discriminant enough, as we completely ignore another important issue in multiview multilabel learning, that is, the specific characteristics of individual views (*individuality*), which may further improve the performance. As discussed before, individual patterns hidden in multiple views are useful in boosting the robustness of multilabel classifier toward rare labels. In addition, these individual patterns can also boost the robustness of the predictor, due to the known usefulness of diversity for ensemble learning. As such, we need to capture the individual information of different views to further improve the performance of our model.

Considering the fact that distinctive mapping matrices $\{\mathbf{U}_v\}_{v=1}^m$ are actually used to encode the corresponding unique properties of individual views, we define another set of coefficient matrices $\{\mathbf{W}_v\}_{v=1}^m$ to capture the distinctive characteristics of different views, where $\mathbf{W}_v \in \mathbb{R}^{d_v \times c}$ maps the reconstructed feature matrix $\mathbf{V}\mathbf{U}_v^T$ of the vth view to the label space. In this way, we leverage the individual and common information of heterogeneous views and further extend (1) as follows:

$$\begin{aligned}\mathcal{L} = &\sum_{v=0}^{m} \left\|\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T\right\|_F^2 + \alpha\|\mathbf{V}\mathbf{W}_0 - \mathbf{Y}\|_F^2 \\ &+ \frac{1-\alpha}{m} \sum_{v=1}^{m} \left\|\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v - \mathbf{Y}\right\|_F^2 \\ &+ \beta\left(\|\mathbf{W}_0\|_F^2 + \|\mathbf{W}_v\|_F^2\right) \end{aligned} \tag{2}$$

where $\mathbf{X}_0 = \mathbf{V}$ and $\mathbf{U}_0$ is an identity matrix, $\alpha \in (0, 1)$ is a tradeoff parameter which balances the contribution of individuality and commonality among multiview data. Here, $\mathbf{X}_0$ can be regarded as an additionally introduced common view spanned by the shared subspace $\mathbf{V}$. It is worth noting that $\mathbf{V}$ can be regarded as the dictionary matrix for all views, and $\mathbf{U}_v$ represents the view-specific coding coefficients. Therefore, $\mathbf{V}\mathbf{U}_v^T$ targets to reserve the view-specific input signals, while $\mathbf{V}$ captures the shared information for all views. Hence, $\mathbf{V}$ and $\mathbf{V}\mathbf{U}_v^T$ could be regarded as different representations of multiview data. By minimizing (2), we can achieve two goals. On the one hand, since both common and individual information of heterogeneous views are employed in an elegant way, the learned shared subspace is more discriminate. On the other hand, the final predictive model is also enhanced because it absorbs both the public labels appeared in most views as well as the individual labels embedded in several specific views.

Another inherent property of learning from multilabel data is how to utilize label correlations, and this issue has not been addressed in (2). Label correlation has long been regarded as a fundamental challenge that can be used to improve the performance of multilabel learning [19], [40]. To address this limitation, we try to leverage the label correlation among labels

to estimate the predicted likelihood scores as follows:

$$\mathcal{L} = \sum_{v=0}^{m} \left\| \mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T \right\|_F^2 + \alpha \|\mathbf{V}\mathbf{W}_0\mathbf{S} - \mathbf{Y}\|_F^2 + \frac{1-\alpha}{m} \sum_{v=1}^{m} \left\| \mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S} - \mathbf{Y} \right\|_F^2 + \beta\left(\|\mathbf{W}_0\|_F^2 + \|\mathbf{W}_v\|_F^2 \right) \quad (3)$$

where $\mathbf{S} \in \mathbb{R}^{c\times c}$ represents the label correlations matrix, which can be estimated from the known $\mathbf{Y}$. For example, the correlation is often measured by the cosine similarity. However, since labeled samples may be not sufficient, we advocate to learn it by leveraging the features and already known labels of the training data. In experiments, we randomly initialize $\mathbf{S}$ and treat it as a trainable parameter during optimization.

The final prediction label matrix $\hat{\mathbf{Y}}$ is given by majority voting $\hat{\mathbf{Y}} = \mathbf{V}(\sum_{v=1}^{m} \mathbf{U}_v\mathbf{W}_v + \mathbf{W}_0)\mathbf{S}$. Equation (3) not only explicitly considers the common and individual information among multiple views in a principle way but also absorbs label information to induce the shared subspace and enhance its discriminate power. Another advantage of our model is that it exploits ensemble learning to further derive robust results, especially for rare labels. Our following experiments will confirm these advantages.

C. Optimization

The objective function in (3) involves $\{\mathbf{U}_v\}_{v=1}^m$, $\mathbf{V}$, $\mathbf{S}$, $\{\mathbf{W}_v\}_{v=1}^m$ and $\mathbf{W}_0$, and it is not easy to optimize all the variables simultaneously. Given that we adopt an alternating optimization technique to optimize the objective function by alternatively optimizing one variable while fixing other variables.

1) Update $\{\boldsymbol{U}_v\}_{v=1}^m$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{W}_v\}_{v=1}^m$, $\boldsymbol{W_0}$, and $\boldsymbol{S}$: When $\mathbf{V}$, $\mathbf{W}$, $\mathbf{W}_v$, and $\mathbf{S}$ are fixed, the computation of $\mathbf{U}_v$ is independent from $\mathbf{U}_{v'}, v' \neq v$. Thus, for each view v, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{U}_v$:

$$\mathcal{L}(\mathbf{U}_v) = -2\mathbf{X}_v^T\mathbf{V} + 2\mathbf{U}_v\mathbf{V}^T\mathbf{V} - 2(1-\alpha)/m\mathbf{W}_v\mathbf{S}\mathbf{Y}^T\mathbf{V} + 2(1-\alpha)/m\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}. \quad (4)$$

Using the Karush–Kuhn–Tucker (KKT) condition [51], we can derive the following updating rule:

$$(\mathbf{U}_v)_{i,j} \leftarrow (\mathbf{U}_v)_{i,j} \frac{\left(\mathbf{X}_v^T\mathbf{V} + \frac{1-\alpha}{m}\mathbf{W}_v\mathbf{S}\mathbf{Y}_v^T\mathbf{V}\right)_{i,j}}{\left(\mathbf{U}_v\mathbf{V}^T\mathbf{V} + \frac{1-\alpha}{m}\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\right)_{i,j}}. \quad (5)$$

2) Update $\boldsymbol{V}$ With Fixed $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$, $\boldsymbol{W_0}$ and $\boldsymbol{S}$: When fixed $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=1}^m$, $\mathbf{W}_0$, and $\mathbf{S}$, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{V}$ to zero:

$$2\sum_{v=0}^{m}\left(\mathbf{V}\mathbf{U}_v^T\mathbf{U}_v - \mathbf{X}_v\mathbf{U}_v\right) - 2\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + 2\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + 2\alpha\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T\mathbf{W}_0^T - 2\alpha\mathbf{Y}\mathbf{S}^T\mathbf{W}_0^T = 0. \quad (6)$$

Then, we have the following closed-form solution for $\mathbf{V}$, which is updated efficiently:

$$\mathbf{V} = \frac{\sum_{v=0}^{m}\theta_v\mathbf{X}_v\mathbf{U}_v + \alpha\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + \mathbf{Q}_1}{\sum_{v=1}^{m}\theta_v\mathbf{U}_v^T\mathbf{U}_v + \alpha\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T\mathbf{W}_0^T + \mathbf{Q}_2} \quad (7)$$

where $\mathbf{Q}_1 = (1-\alpha/m)\sum_{v=0}^{m}\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v$ and $\mathbf{Q}_2 = (1-\alpha/m)\sum_{v=1}^{m}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v$.

3) Update $\boldsymbol{W}_v$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\boldsymbol{W_0}$, and $\boldsymbol{S}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\mathbf{W}_0$, and $\mathbf{S}$ are fixed, similarly to the update of $\mathbf{U}_v$, we have the following equation by setting the derivative with respect to $\mathbf{W}_v$:

$$2\frac{1-\alpha}{m}\left(\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T - \mathbf{U}_v\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right) + 2\beta\mathbf{W}_v. \quad (8)$$

Based on the KKT condition, we can derive the following updating rule:

$$(\mathbf{W}_v)_{i,j} \leftarrow (\mathbf{W}_v)_{i,j} \frac{\left(\frac{1-\alpha}{m}\mathbf{U}_v\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right)_{i,j}}{\left(\frac{1-\alpha}{m}\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T + \beta\mathbf{W}_v\right)_{i,j}}. \quad (9)$$

4) Update $\boldsymbol{W_0}$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$ and $\boldsymbol{S}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=1}^m$, and $\mathbf{S}$ are fixed, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{W}_0$ to zero:

$$\mathcal{L}(\mathbf{W}_0) = 2\alpha\left(\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T - \mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right) + 2\beta\mathbf{W}_0. \quad (10)$$

Based on the KKT condition, we can derive the following update rule:

$$(\mathbf{W}_0)_{i,j} \leftarrow (\mathbf{W}_0)_{i,j} \frac{\left(\alpha\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right)_{i,j}}{\left(\alpha\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T + \beta\mathbf{W}_0\right)_{i,j}}. \quad (11)$$

5) Update $\boldsymbol{S}$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$ and $\boldsymbol{W_0}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=0}^m$, and $\mathbf{W}_0$ are fixed, similarly to the update of $\mathbf{V}$, we have the following equation for $\mathbf{S}$ by setting the derivative with respect to $\mathbf{S}$ to zero:

$$2\sum_{v=0}^{m}\frac{1-\alpha}{m}\left(\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S} - \mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{Y}\right) + 2\alpha\left(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S} - \mathbf{W}_0^T\mathbf{V}^T\mathbf{Y}\right) = 0. \quad (12)$$

We therefore have the following closed-form solution for $\mathbf{S}$:

$$\mathbf{S} = \left(\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v + \alpha\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\right)^{-1} \times \left(\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{Y} + \alpha\mathbf{W}_0^T\mathbf{V}^T\mathbf{Y}\right). \quad (13)$$

Given the above iterative optimization process, we summarize the main procedure of ICM2L in Algorithm 1.

D. Complexity Analysis

The time complexity of ICM2L is dominated by the matrix multiplication and matrix inverse operations. The time complexity of matrix inverse for $\mathbf{V}$ and $\mathbf{S}$ is relatively small, so the time complexity of ICM2L is mainly driven by the computation for $\mathbf{U}_v$, $\mathbf{W}_v$, and $\mathbf{W}_0$. Concretely, the complexity for solving $\mathbf{U}_v$, $\mathbf{W}_v$, and $\mathbf{W}_0$ is $O(d_{\max}nk + d_{\max}k^2 + d_{\max}ck)$,

Algorithm 1 ICM2L

Input:

$\{\mathbf{X}_v\}_{v=1}^m$: n training samples with m views
$\mathbf{Y}$: Initial label matrix for n samples
k: Dimensionality of the shared subspace
α and β: Trade-off parameters used in Eq. (3)
ε: Convergence threshold
t: Number of iterations

Output:

$\hat{\mathbf{Y}}$: Predicted label likelihood matrix for n samples

Randomly initialize $\mathbf{U}_v$, $\mathbf{W}_v$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{S}$;
Compute $\mathcal{L}_0$ by Eq. (3);
for $i = 1, 2, \cdots, t$ **do**
  for $v = 1, 2, \cdots, m$ **do**
    Update $\mathbf{U}_v$ by Eq. (5);
  end
  Update $\mathbf{V}$ by Eq. (7);
  for $v = 1, 2, \cdots, m$ **do**
    Update $\mathbf{W}_v$ by Eq. (9);
  end
  Update $\mathbf{W}_0$ by Eq. (11);
  Update $\mathbf{S}$ by Eq. (13);
  Update $\mathcal{L}_i$ by Eq. (3);
  If $|\mathcal{L}_i - \mathcal{L}_{i-1}| \leq \varepsilon$, Return.
end

$O(d_{\max}nk)$, and $O(nck)$, respectively, where $d_{\max}$ is the largest dimensionality of the views. Since $n \gg k$ and $n \gg c$, the overall time complexity of ICM2L is $O(d_{\max}nkt)$, where t is the number of iterations to reach convergence. In practice, if $d_{\max} \ll n$, the total complexity of ICM2L scales with the number of samples. In our experiments, we found that t usually does not exceed 80. In addition, some views have sparse feature matrices, so the actual time cost of the above operations can be further reduced.

IV. Experiments

In this section, we conduct extensive experiments over six real-world datasets to evaluate the efficiency and effectiveness of the proposed framework. There are four major questions we aim to answer.

1) How effective is ICM2L compared with other related methods in classifying multiview multilabel data?
2) How robust is ICM2L in discovering rare labels compared with the state-of-the-art methods?
3) What are the impacts of two parameters α and β on ICM2L?
4) How efficient is ICM2L in modeling multiview multilabel learning problem?

A. Experimental Setup

Six multiview datasets that we employed in the experiments are all publicly available. The statistics of them are summarized in Table I. Yeast is a biological dataset with two views [52], one view is the genetic expression and the other is the phylogenetic profile of a gene. Core15k, Pascal07, and ESPGame are the three widely used multiview image datasets.[1] We collected the multiple features of these images from [53], where each image is represented by six representative feature views: HUE, SIFT, GIST, HSV, RGB, and LAB. Each sample in Mirflicker[2] and Nus-wide[3] consists of an image and textual tags, we construct the two views (image and text) according to [54]. For each dataset, we randomly sample 30% data for training and use the remaining 70% data for testing (unlabeled data).

TABLE I
Statistics of Four Multiview Datasets: n Is the Number of Samples; m Is the Number of Views; c Is the Number of Distinct Labels; #avg Is the Average Number of Labels per Sample; $d_{\min}$ Is the Smallest Dimension of All Views

datasets	n	m	c	#avg	d_{min}
Yeast	2417	2	14	4.237	24
Core15k	4999	6	260	3.396	100
Pascal07	9963	6	20	1.465	100
ESPGame	20770	6	268	4.686	100
Mirflicker	25000	2	24	3.794	512
Nus-wide	260648	2	81	2.783	500

Baseline Methods: To study the performance of ICM2L, we compare it with six state-of-the-art methods. In addition, to investigate the contribution of encoding common information, individual information, and using label correlations, we include three variants, namely, ICM2L-c, ICM2L-i, and ICM2L-lc.

1) lrMMC [7] leverages a low-dimensional common representation of all views and matrix completion for multilabel classification.
2) LSML [8] is a recent multiview multilabel learning framework that learns the shared subpace among heterogeneous features as well as the follow-up predictor in a unified objective function.
3) MLAN [4] is another unified multiview learning method and initially focuses on the single-label classification problem; we adapt it for multilabel scenario by assigning multiple labels instead of a single one to unlabeled data.
4) CSMSC [13] is a multiview subspace learning approach, which can jointly extract the consistency and specificity of heterogeneous features for subspace representation learning. We adopt the extracted individual and common representation features as inputs to our model to train the ensemble classifier.
5) SMMCL [18] is a self-paced-based multiview multilabel learning method, which considers both the individuality and commonality characteristics among multiview data.
6) MVMC-LS [42] is a multiview learning approach based on matrix completion, it combines the matrix completion outputs of different views with various weights.
7) ICM2L-c is a variant of ICM2L by excluding individual information and makes prediction by $\mathbf{W}$ (commonality).

[1] http://lear.inrialpes.fr/data/
[2] http://press.liacs.nl/mirflickr/mirdownload.html
[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.html

TABLE II
RESULTS ON FOUR DATASETS WITH $k = 0.5d_{\min}$. $d_{\min}$ REPRESENTS THE MINIMUM DIMENSIONALITY OF MULTIPLE VIEWS

Dataset	metric	lrMMC	MLAN	MVMC-LS	CSMSC	LSML	SMMCL	ICM2L
Yeast	Accuracy	$\mathbf{0.539 \pm 0.001}$	0.381 ± 0.001	0.517 ± 0.001	0.537 ± 0.001	0.535 ± 0.004	$\mathbf{0.542 \pm 0.002}$	0.536 ± 0.004
	1-RL	0.787 ± 0.001	$\mathbf{0.811 \pm 0.002}$	0.761 ± 0.000	0.793 ± 0.001	0.797 ± 0.003	$\mathbf{0.816 \pm 0.001}$	0.788 ± 0.005
	AP	0.703 ± 0.001	0.459 ± 0.001	0.662 ± 0.000	0.698 ± 0.002	0.702 ± 0.004	$\mathbf{0.717 \pm 0.004}$	0.702 ± 0.003
	AUC	0.798 ± 0.001	0.589 ± 0.001	0.778 ± 0.000	0.797 ± 0.001	0.799 ± 0.002	$\mathbf{0.812 \pm 0.003}$	0.799 ± 0.005
Core15k	Accuracy	$\mathbf{0.191 \pm 0.001}$	0.103 ± 0.001	0.172 ± 0.000	$\mathbf{0.193 \pm 0.001}$	$\mathbf{0.193 \pm 0.001}$	$\mathbf{0.192 \pm 0.002}$	$\mathbf{0.194 \pm 0.001}$
	1-RL	0.758 ± 0.001	0.521 ± 0.002	0.750 ± 0.000	0.762 ± 0.001	0.768 ± 0.001	0.771 ± 0.001	$\mathbf{0.795 \pm 0.003}$
	AP	0.236 ± 0.001	0.146 ± 0.001	0.215 ± 0.001	0.432 ± 0.001	0.256 ± 0.001	0.259 ± 0.002	$\mathbf{0.279 \pm 0.004}$
	AUC	0.760 ± 0.001	0.710 ± 0.001	0.752 ± 0.000	0.767 ± 0.001	0.774 ± 0.001	0.778 ± 0.003	$\mathbf{0.797 \pm 0.003}$
Pascal07	Accuracy	0.278 ± 0.000	0.205 ± 0.001	0.264 ± 0.001	0.281 ± 0.002	0.283 ± 0.001	0.285 ± 0.001	$\mathbf{0.296 \pm 0.003}$
	1-RL	0.697 ± 0.001	0.502 ± 0.001	0.692 ± 0.001	0.715 ± 0.001	0.725 ± 0.003	0.730 ± 0.002	$\mathbf{0.756 \pm 0.005}$
	AP	0.429 ± 0.000	0.350 ± 0.002	0.401 ± 0.002	0.424 ± 0.002	0.446 ± 0.001	0.451 ± 0.001	$\mathbf{0.452 \pm 0.001}$
	AUC	0.727 ± 0.000	0.646 ± 0.001	0.725 ± 0.001	0.739 ± 0.003	0.758 ± 0.002	0.763 ± 0.002	$\mathbf{0.785 \pm 0.005}$
ESPGame	Accuracy	0.170 ± 0.000	0.088 ± 0.000	0.134 ± 0.001	0.177 ± 0.000	0.189 ± 0.001	0.192 ± 0.001	$\mathbf{0.206 \pm 0.001}$
	1-RL	0.777 ± 0.000	0.521 ± 0.001	0.764 ± 0.001	0.784 ± 0.001	$\mathbf{0.796 \pm 0.001}$	$\mathbf{0.798 \pm 0.002}$	$\mathbf{0.796 \pm 0.001}$
	AP	0.189 ± 0.000	0.111 ± 0.000	0.167 ± 0.000	0.194 ± 0.001	0.205 ± 0.001	0.207 ± 0.001	$\mathbf{0.220 \pm 0.002}$
	AUC	0.783 ± 0.000	0.642 ± 0.000	0.770 ± 0.001	0.785 ± 0.002	0.789 ± 0.000	0.790 ± 0.002	$\mathbf{0.803 \pm 0.001}$
Mirflickr	Accuracy	0.376 ± 0.001	0.282 ± 0.004	0.355 ± 0.001	0.387 ± 0.002	0.394 ± 0.002	0.412 ± 0.001	$\mathbf{0.436 \pm 0.001}$
	1-RL	0.750 ± 0.002	0.675 ± 0.002	0.736 ± 0.002	0.758 ± 0.001	0.765 ± 0.003	0.773 ± 0.001	$\mathbf{0.796 \pm 0.001}$
	AP	0.466 ± 0.003	0.401 ± 0.001	0.419 ± 0.002	0.471 ± 0.002	0.485 ± 0.001	0.498 ± 0.002	$\mathbf{0.536 \pm 0.002}$
	AUC	0.757 ± 0.001	0.664 ± 0.003	0.737 ± 0.001	0.761 ± 0.001	0.769 ± 0.001	0.774 ± 0.001	$\mathbf{0.790 \pm 0.001}$
Nus-wide	Accuracy	0.249 ± 0.002	0.198 ± 0.001	0.231 ± 0.002	0.253 ± 0.001	0.268 ± 0.003	0.286 ± 0.004	$\mathbf{0.332 \pm 0.001}$
	1-RL	0.791 ± 0.003	0.714 ± 0.004	0.779 ± 0.002	0.804 ± 0.001	0.819 ± 0.002	0.835 ± 0.003	$\mathbf{0.923 \pm 0.002}$
	AP	0.311 ± 0.002	0.243 ± 0.001	0.304 ± 0.003	0.321 ± 0.001	0.338 ± 0.003	0.367 ± 0.004	$\mathbf{0.448 \pm 0.004}$
	AUC	0.812 ± 0.001	0.735 ± 0.002	0.798 ± 0.003	0.823 ± 0.004	0.841 ± 0.002	0.876 ± 0.003	$\mathbf{0.933 \pm 0.002}$

8) ICM2L-i is a variant of ICM2L by excluding common information and makes prediction by integrating $\{\mathbf{W}_v\}_{v=1}^m$ (individuality).
9) ICM2L-lc is a variant of ICML2 by excluding label correlations.

For comparing methods, five-fold cross-validation on the training set is used to select the optimal parameter values from the range as suggested in the original papers. For our method, we selected the parameters α and β in the range of $\{0.1, 0.2, \ldots, 1\}$ and $\{0.1, 0.3, \ldots, 2\}$, respectively. To avoid random effects, all the experiments are independently repeated ten times, and both the mean and standard deviation are reported. For each comparing method, the code is released or provided by corresponding authors. The code of ICM2L is publicly available at http://mlda.swu.edu.cn/codes.php?name=ICM2L.

Evaluation: Four widely used metrics are adopted for performance comparisons: 1) accuracy; 2) ranking loss (RL); 3) average precision (AP); and 4) average AUC. Note that these metrics generally belong to two categories: 1) example-based criterion and 2) label-based criterion [55]. RL, AP, and Accuracy are the example-based metrics, while AUC is a label-based criterion. They evaluate the performance from ranking and classification perspectives [19], in which RL, AP, and AUC are ranking-based metrics, while Accuracy is an example-based classification criteria. Formal definition of the four metrics can be found in [19] and [55]. *Accuracy* requires the predicted label-likelihood vector to be a binary indicator vector. Here, we consider the labels corresponding to the r largest entries of the vector of the ith sample as the predicted labels, where r is determined as the average number of labels (round to next integer) of labeled samples. To maintain consistency with other evaluation metrics, in our experiments, we report 1-RL instead of RL. Thus, as for other metrics, the higher the value of 1-RL, the better the performance is. These metrics evaluate multilabel classification from different points of view, and it is unlikely for a method outperforming all the other techniques across all the metrics.

B. Effectiveness of ICM2L

To investigate the first question stated at the beginning of this section, we compare the classification performance of all methods on the six datasets listed in Table II. Since SMMCL consumes a lot of memory in training, we can only obtain its results on the Yeast dataset with a server (CentOS 6.9, 64-GB RAM, and MATLAB 2014a). For this reason, we independently sample 3000 instances from large datasets 20 times to construct new sampled datasets and report the best results of them. In Table II, the best (or comparable best) results are highlighted in **boldface** using the pairwise t-test at 95% significance level. Besides paired student's t-test, we also apply the Friedman's test [56] with a *post-hoc* Tukey's test [57] to assess the significant difference between ICM2L and other comparing methods, all the p-values are smaller than 10^{-4} and 0.04 for the Friedman's and Tukey's tests, respectively. We implement the test based on the *Friedman* and *multcompare* functions in MATLAB.

From the results reported in Table II, we can observe that ICM2L outperforms other comparing methods in most cases, especially, on the large-scale datasets. Although MVMC-LS and lrMMC are all designed for multiview multilabel data, and MVMC-LS is almost always outperformed by lrMMC. This is mainly because lrMMC exploits the commonality information among multiple views by assuming that different views are generated from a common low-dimensional subspace, while MVMC-LS just utilizes individual information by combining outputs of different views that cannot make full use of complementary information among views. Since MVMC-LS learns view combination coefficients by two-fold cross-validation in the training data, which may result in scarce labeled training samples in our semisupervised setting and

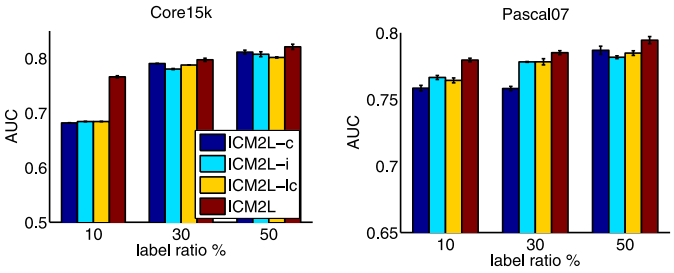


Fig. 2. Results of ICM2L variants under different ratios of labeled samples on Core15k and Pascal07.

impact the combination coefficients learning. Both lrMMC and LSML aim to learn a shared subspace among heterogeneous views, but lrMMC loses to LSML many times. The possible reason is that lrMMC is a two-step method. lrMMC separates the learning process of shared subspace and multilabel classifier, which may result in a suboptimal solution, while LSML can learn the subspace and the follow-up predictor based on the learned representation simultaneously. CSMSC is also a two-step approach, but it outperforms lrMMC in many cases. The main reason is that CSMSC takes advantage of multiple view-specific representations and common representation to train an ensemble classifier, while lrMMC learns a general multilabel classifier for all labels. This comparison justifies our motivation to explore individuality and commonality information to develop the discriminant classifier. Both CSMSC and ICM2L develop an ensemble classifier for classification, but CSMSC loses to ICM2L in most cases. The crucial reason is that ICM2L jointly learn feature representations and ensemble classifier, while CSMSC separates the two learning processes. This comparison indicates the importance to learn feature representation and follow-up classifier jointly.

MLAN is another unified multiview learning method, but it still loses to LSML almost in all cases. The possible reason is that MLAN is naturally designed for single-label problems and it cannot employ label correlations among multiple labels, which is very important in multilabel data as suggested in the literature. Both LSML and ICM2L can utilize label correlations and the commonality information to make prediction; LSML is outperformed by ICM2L in most cases. The principal reason is that ICM2L explicitly utilizes the individuality information of the views. These comparisons justify our motivation to jointly exploit both commonality and individual information of multiple data views. SMMCL is a recent state-of-the-art method that considers the individuality and commonality patterns of multiview data. SMMCL can iteratively propagate labels to the most informative candidate unlabeled samples during training and these samples are then augmented into the training set as labeled data for the next iteration. For this reason, SMMCL, in general, consumes more labeled samples for training, and achieves better performance than other comparing methods in many cases. However, SMMCL is still outperformed by ICM2L in many cases, especially, over relative large-scale datasets, for example, Core15k, Pascal07, ESPGame, Mirflickr, and Nus-wide. The crucial intuition behind is two sides: 1) the label imbalance problem in the large-scale datasets is more serious than that in Yeast and 2) the bonus of augmenting training set is limited with the increased size of dataset, since a number of labeled data is ready to train an effective semisupervised algorithm. These observations further validate our motivation to directly exploit individual information of various views for prediction instead of subspace learning. In addition, another bottleneck of SMMCL lies in its poor scalability, since it needs huge memory for training. These comparisons validate the effectiveness of ICM2L.

C. Effectiveness of ICM2L via Component Analysis

To further justify the effectiveness of our model in capturing the commonality and individuality information of multiple data views, as well as of the label correlations, we conduct additional component analysis experiments on the Core15k and Pascal07 datasets and report the AUC values in Fig. 2. In this figure, we set the ratios of labeled data equal to 10%, 30%, and 50%, respectively.

From the figure, we can see that the performance of all the variants of ICM2L increases as the increasing of labeled training data, and ICM2L outperforms its variants across all the settings. ICM2L-c and ICM2L-i disregard the individuality information and commonality information, respectively, and they are outperformed by ICM2L in many cases. This is mainly because ICM2L utilizes both types of information, and thus improves the final performance. These results corroborate our motivation to explicitly leverage the individual and common information of multiple data views. Both ICM2L and ICML2-lc take advantage of multiview data from individual and common aspects, but ICM2L outperforms ICM2L-lc in many cases across two datasets. The inherent reason is that ICM2L captures label correlations, which are very important in multilabel learning. This fact validates the necessity of capturing label correlations and also proves the effectiveness of the learned label correlation matrix $\mathbf{S}$.

D. Robustness of ICM2L Toward Rare Labels

To answer the second proposed question, we conduct experiments to quantify the benefit of utilizing individual information toward rare labels. Let IR_c denote the imbalance ratio of label c, which is calculated by the ratio between the number of negative samples and that of positive samples for label c. We generate an imbalance dataset from Core15k by first discarding the samples that are annotated with few than three labels. Then, we split the labels of the new dataset as general labels and rare labels based on IR_c. Specifically, we decide label c as a general label if $\mathrm{IR}_c \leq 50$; otherwise, regarding the labels as rare label. In addition, to further investigate the performance of all methods in extreme cases, we divide the rare label into three levels: rare_1, rare_2, and rare_3. rare_1 includes the labels with $50 < \mathrm{IR}_c \leq 100$, rare_2 includes the labels with $100 < \mathrm{IR}_c \leq 150$, and rare_3 includes the labels with $\mathrm{IR}_c \geq 150$. Finally, the new dataset has 4966 samples associated with 119 labels, 45 labels with $\mathrm{IR}_C \leq 50$, and the other 74 rare labels. Specifically, 36 labels for rare_1,

TABLE III
EXPERIMENTAL RESULTS OF COMPARING METHODS ON THE PROCESSED CORE15K DATASET WITH DIFFERENT SCALES OF IMBALANCED LABELS

IR		lrMMC	MLAN	MVMC-LS	CSMSC	LSML	SMMCL	ICM2L
≤ 50 (general)	1-RL	0.714 ± 0.001	0.497 ± 0.001	0.684 ± 0.002	0.724 ± 0.002	0.736 ± 0.001	0.742 ± 0.002	$\mathbf{0.778 \pm 0.002}$
	AP	0.298 ± 0.001	0.213 ± 0.002	0.243 ± 0.000	0.306 ± 0.003	0.312 ± 0.001	0.323 ± 0.001	$\mathbf{0.349 \pm 0.002}$
$50 < IR_c \leq 100$(rare$_1$)	1-RL	0.522 ± 0.001	0.394 ± 0.001	0.481 ± 0.001	0.618 ± 0.001	0.622 ± 0.001	0.633 ± 0.002	$\mathbf{0.681 \pm 0.001}$
	AP	0.246 ± 0.001	0.203 ± 0.001	0.219 ± 0.000	0.278 ± 0.001	0.283 ± 0.001	0.296 ± 0.001	$\mathbf{0.346 \pm 0.001}$
$100 < IR_c \leq 150$(rare$_2$)	1-RL	0.487 ± 0.001	0.326 ± 0.001	0.415 ± 0.001	0.516 ± 0.001	0.519 ± 0.001	0.535 ± 0.001	$\mathbf{0.598 \pm 0.002}$
	AP	0.141 ± 0.001	0.122 ± 0.001	0.137 ± 0.001	0.185 ± 0.001	0.187 ± 0.001	0.197 ± 0.001	$\mathbf{0.263 \pm 0.001}$
$IR_c > 150$(rare$_3$)	1-RL	0.458 ± 0.001	0.218 ± 0.001	0.398 ± 0.001	0.484 ± 0.002	0.486 ± 0.001	0.499 ± 0.002	$\mathbf{0.598 \pm 0.001}$
	AP	0.134 ± 0.001	0.109 ± 0.001	0.121 ± 0.001	0.151 ± 0.001	0.153 ± 0.001	0.168 ± 0.001	$\mathbf{0.225 \pm 0.002}$

27 labels for rare$_2$, and 11 labels for rare$_3$. Table III shows the performance of comparing methods on the general labels (IR$_C \leq 50$) and rare labels (IR$_C > 50$). The experimental configurations are the same with that in Section IV-B.

From the table, we can see that the performance of all methods decreases with IR$_c$ increases, and ICM2L outperforms other comparing methods not only on the general labels but also on all rare label cases, which is consistent with the results in Table II. An interesting observation is that the differences between our model and other comparing methods in rare label cases are larger than those in general labels, especially in the most imbalance situations, for example, rare$_2$ and rare$_3$. These comparisons validate the robustness of our model in classifying rare labels. In addition, although ICM2L, SMMCL, and LSML aim to employ the complementary information, as well as individual information among different views, the last two methods still lose to ICM2L in most cases. The reason behind this fact is that ICM2L utilizes individual information of multiple views in order to capture rare labels hidden in specific views, while the other two approaches exploit individual patterns to enforce subspace learning. These results intuitively justify our motivation to capture individual characteristics of multiple data views. Another interesting observation is that CSMSC performs more similar to LSML in imbalance scenarios, in which the performance gap between them in rare$_2$ (or rare$_3$) is smaller than that in the general case. This result further validates the importance of directly exploring individual and common patterns to construct robust classifier.

E. Parameter Analysis

We now study the third proposed question. ICM2L has two parameters α and β, which control the importance of individual information and regularization terms, respectively. We test the sensitivity of ICM2L with respect to α and β in the range $\{0.1, 0.2, \ldots, 1\}$ and $\{0.1, 0.3, \ldots, 2\}$, respectively. We report Accuracy and AUC on Yeast in Fig. 3; the results for the other datasets and evaluation metrics are similar and lead to similar conclusions.

From Fig. 3, we can observe that ICM2L obtains relatively good performance when α is around 0.6 and β is around 0.7. In addition, when $\alpha \to 0$ or $\alpha \to 1$, the performance of ICM2L is reduced. These results further confirm the contribution of commonality information and individual information. Another interesting observation is that the performance of ICM2L decreases more sharply when $\alpha \approx 1$ than that when α is around 0. Since α controls the importance between individual and common patterns, and $\alpha \approx 1$ means that we discard the individual information of multiview data and only focus on common patterns. In contrast, $\alpha \approx 0$ means that we only utilize individual patterns hidden in each view and discard the common patterns among them. These results again indicate the importance of individual information. When β is close to 0, the Accuracy and AUC values tend to decrease. This fact validates the effectiveness of the l_2 term. In our experiments, we set $\alpha = 0.6$ and $\beta = 0.7$.

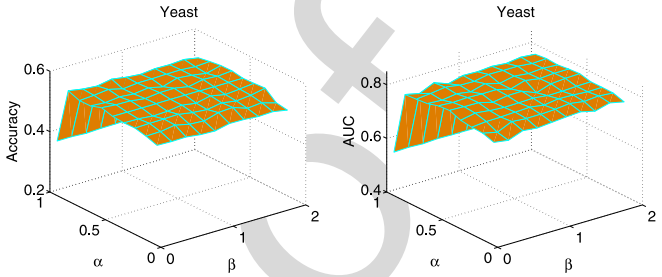


Fig. 3. Results of ICM2L under different input values of α and β.

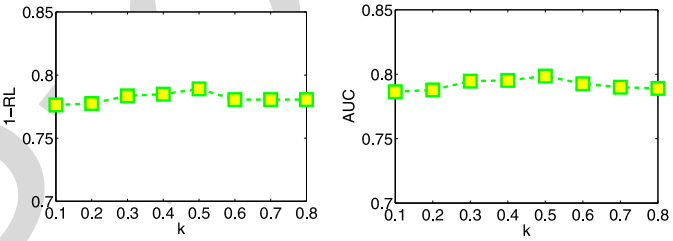


Fig. 4. Results of ICM2L under different input values of k.

In addition, we also conduct experiments to investigate the sensitivity of ICM2L with respect to k. Fig. 4 reports the 1-RL and AUC values of ICM2L on the Yeast dataset with k varying from $0.1d_{\min}$ to $0.8d_{\min}$. As we can see, the performance of ICM2L first increases with k rising, then it decreases when $k > 0.5d_{\min}$. For this reason, we set $k = 0.5d_{\min}$ in experiments. For Mirflickr and Nus-wide, we set $d_{\min} = 100$ for simplicity. For general multilabel datasets, we recommend setting the dimension of features ($k \approx$) #Avg $\log_2(N)$, where #Avg indicates the number of associated labels per sample. The behind intuition is that the minimal number of bits to encode N data points in one class is $\log_2(N)$ in the information theory. For multilabel data, it also needs to consider the statistical property of labels, which could be captured by #Avg. As such, #Avg $\log_2(N)$ bits are needed.

F. Efficiency of ICM2L

To investigate the proposed last question, we conduct experiments on all datasets with the same configuration in

TABLE IV
RUNTIME COMPARISON (IN SECONDS)

	lrMMC	MLAN	MVLC-LS	CSMSC	LSML	SMMCL	ICM2L
Yeast	8.03	197.90	39.57	25.78	23.11	23.20	10.11
Core15k	290.56	357.92	13892.32	437.88	387.24	543.70	325.41
Pascal07	908.49	1604.04	8783.45	1378.44	1023.38	1588.92	768.29
ESPGame	1512.63	9785.80	21748.23	5436.28	4792.04	7332.42	3918.44
Mirflickr	1245.72	7752.33	18428.18	4493.65	3986.13	5856.18	3245.11
Total	3965.43	19697.99	62891.75	11772.03	10211.90	15344.42	8267.36

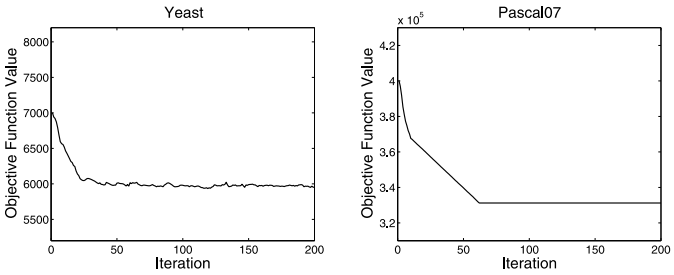


Fig. 5. Convergence trend analysis.

Section IV-B, and report the runtime costs of all methods and the convergence trend of ICM2L. We only report the runtime cost of comparing methods over datasets excluding Nus-wide since the runtime cost on Nus-wide is generally much bigger than other datasets. We observe similar runtime cost trend on Nus-wide. Table IV reports the runtime costs of all the approaches on a server (CentOS 6.9 with Inter Xeon E5-2678, 64-GB RAM, and MATLAB 2014a). From Table IV, we can see that ICM2L is much faster than MLAN, MVLC-LS, CSMSC, and LSML in general. However, lrMMC runs much faster than ICM2L in most cases. This is because lrMMC is a two-step method that learns the shared subspace and the follow-up predictor in two separate steps, while ICM2L has to learn the low-dimensional representations and the multilabel classifier in each iteration. These comparisons corroborate the efficiency of our model.

Fig. 4 shows the convergence curve of ICM2L on the Yeast and Core15k datasets. We can see that ICM2L tends to converge after 70 iterations for the Yeast dataset, and after 60 iterations for the Core15k dataset. The convergence trends on the other datasets are the same as those reported in Fig. 4. Overall, ICM2L converges at most in 80 iterations for the datasets used in the experiments.

V. CONCLUSION

In this article, we investigated how to explore the individuality and commonality of heterogeneous features for effective multiview multilabel classification. To this end, a multiview multilabel framework termed ICM2L is presented. ICM2L learns a shared subspace of heterogeneous views, label correlations, and an ensemble classifier that captures both individuality and commonality information of multiple views in a principled way. Different from the previous works that focus on learning representative hidden representations by capturing the shared and individual patterns across multiple views, we utilize such information to improve the discriminant capacity of classifier toward rare labels. Experiments on several benchmark datasets demonstrate the superiority of the proposed model over related competitive solutions. In the future, we plan to further improve ICM2L by adapting nonlinear mapping functions with deep models.

ACKNOWLEDGMENT

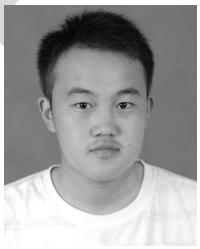
The authors would like to thank the authors who kindly shared their source code and datasets with them for the experiments.

REFERENCES

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
[2] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.
[3] P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via CCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 199–207.
[4] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2408–2414.
[5] Q. Wang, H. Lv, J. Yue, and E. Mitchell, "Supervised multiview learning based on simultaneous learning of multiview intact and single view classifier," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 2293–2301, 2017.
[6] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[7] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2778–2784.
[8] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang, "Latent semantic aware multi-view multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4414–4421.
[9] P. Zhao, Y. Jiang, and Z.-H. Zhou, "Multi-view matrix completion for clustering with side information," in *Proc. 21st Pac.–Asia Conf. Knowl. Disc. Data Min.*, 2017, pp. 403–415.
[10] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1233–1246, Jun. 2015.
[11] C. Sagonas, E. Ververas, Y. Panagakis, and S. Zafeiriou, "Recovering joint and individual components in facial data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2668–2681, Nov. 2018.
[12] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, Sep. 2018.
[13] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3730–3737.
[14] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 586–594.
[15] O. G. R. Reyes and S. Ventura, "Performing multi-target regression via a parameter sharing-based deep network," *Int. J. Neural Syst.*, vol. 29, no. 9, 2019, Art. no. 1950014.
[16] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu, "Towards class-imbalance aware multi-label learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 4041–4047.

AQ4

[17] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1095–1107, May 2014.
[18] C. Gong, "Exploring commonality and individuality for multi-modal curriculum learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1926–1933.
[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[20] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Min.*, vol. 3, no. 3, pp. 1–13, 2007.
[21] J. Huang, G. Li, Q. Huang, and X. Wu, "Joint feature selection and classification for multilabel learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 876–889, Mar. 2018.
[22] S.-J. Huang and Z.-H. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. AAAI Conf. Artif. Intell.*, 2012, pp. 949–955.
[23] X. Kong, M. K. Ng, and Z.-H. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[24] Y. Guo and D. Schuurmans, "Semi-supervised multi-label classification," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Disc. Databases*, 2012, pp. 355–370.
[25] O. G. R. Pupo and S. Ventura, "Evolutionary strategy to perform batch-mode active learning on multi-label data," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, p. 46, 2018.
[26] O. Reyes, C. Morell, and S. Ventura, "Effective active learning strategy for multi-label learning," *Neurocomputing*, vol. 273, no. 1, pp. 494–508, 2018.
[27] A. H. Akbarnejad and M. S. Baghshah, "An efficient semi-supervised multi-label classifier capable of handling missing labels," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 229–242, Feb. 2019.
[28] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in *Proc. AAAI Conf. Artif. Intell.*, 2010, pp. 1862–1868.
[29] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 593–601.
[30] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proc. ACM Multimedia Conf.*, 2018, pp. 700–708.
[31] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.
[32] G. Yu *et al.*, "Feature-induced partial multi-label learning," in *Proc. IEEE Int. Conf. Data Min.*, 2018, pp. 1398–1403.
[33] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4302–4309.
[34] B. Wu, F. Jia, W. Liu, B. Ghanem, and S. Lyu, "Multi-label learning with missing labels using mixed dependency graphs," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 875–896, 2018.
[35] Q. Tan, Y. Yu, G. Yu, and J. Wang, "Semi-supervised multi-label classification using incomplete label information," *Neurocomputing*, vol. 260, pp. 192–202, Oct. 2017.
[36] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1910–1918.
[37] J. Tu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, and M. Guo, "Multi-label answer aggregation based on joint matrix factorization," in *Proc. IEEE Int. Conf. Data Min.*, 2018, pp. 517–526.
[38] C. Zhang, Z. Yu, H. Fu, P. Zhu, L. Chen, and Q. Hu, "Hybrid noise-oriented multilabel learning," *IEEE Trans. Cybern.*, to be published.
[39] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.
[40] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3780–3788.
[41] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Incomplete multi-view weak-label learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2703–2709.
[42] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.
[43] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognit.*, vol. 48, no. 1, pp. 10–19, 2015.
[44] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[45] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.
[46] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2017.
[47] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Heidelberg, Germany: Springer, 1971, pp. 134–151.
[48] C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[49] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2019.
[50] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, p. 5, 2008.
[51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[52] E. L. Gibaja, J. M. Moyano, and S. Ventura, "An ensemble-based approach for multi-view multi-label classification," *Progr. Artif. Intell.*, vol. 5, no. 4, pp. 251–259, 2016.
[53] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 902–909.
[54] X. Liu, G. Yu, C. Domeniconi, J. Wang, Y. Ren, and M. Guo, "Ranking-based deep cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4400–4407.
[55] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Comput. Surveys*, vol. 47, no. 3, p. 52, 2015.
[56] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.
[57] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," in *Encyclopedia of Research Design*. Thousand Oaks, CA, USA: Sage, pp. 1–5, 2010.



Qiaoyu Tan received the Ph.D. degree from the Department of Computer Science, Texas A&M University, College Station, TX, USA.

He was a Research Assistant with the Machine Learning and Data Analysis Laboratory, Southwest University, Chongqing, China. His current research interests include machine learning and data mining.

Guoxian Yu received the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2013.

He is a Professor with the College of Computer and Information Science, Southwest University, Chongqing, China. His current research interests include data mining and bioinformatics.

Prof. Yu has served as PC member for IJCAI, AAAI, KDD, ICDM, SDM, WSDM, and NIPS and the Reviewer for the IEEE TRANSACTIONS JOURNALS and *Bioinformatics*.

Jun Wang received the B.Sc. and M.Eng. degrees in computer science and the Ph.D. degree in artificial intelligence from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively.

She is an Associate Professor with the College of Computer and Information Science, Southwest University, Chongqing, China. Her current research interests include machine learning, and data mining and their applications in bioinformatics.

AQ6

Carlotta Domeniconi is an Associate Professor with the Department of Computer Science, George Mason University, Fairfax, VA, USA. She has published extensively in premier journals and conferences in machine learning and data mining. Her current research interests include machine learning, pattern recognition, and data mining, with applications in text mining and bioinformatics.

Ms. Domeniconi has served as a PC member for KDD, ICDM, SDM, ECML-PKDD, and AAAI. She is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and *Knowledge and Information Systems*.

Xiangliang Zhang received the Ph.D. degree (Hons.) in computer science from INRIA-University Paris-Sud 11, France, in 2010.

She is an Associate Professor and Directs the Machine Intelligence and Knowledge Engineering Laboratory, King Abdullah University of Science and Technology. Her current research interests include diverse areas of machine learning and data mining.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ1: Please confirm or add details for any funding or financial support for the research of this article.

AQ2: Please provide the postal code for George Mason University, Fairfax, VA, USA, and King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

AQ3: Please provide the department name for King Abdullah University of Science and Technology.

AQ4: Please provide the complete details and exact format for Reference [1].

AQ5: Please confirm if the location and publisher information for Reference [47] is correct as set.

AQ6: Please provide the educational background for the author C. Domeniconi.

Individuality- and Commonality-Based Multiview Multilabel Learning

Qiaoyu Tan, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang

***Abstract*—In multiview multilabel learning, each object is represented by several heterogeneous feature representations and is also annotated with a set of discrete nonexclusive labels. Previous studies typically focus on capturing the shared latent patterns among multiple views, while not sufficiently considering the diverse characteristics of individual views, which can cause performance degradation. In this article, we propose a novel approach [individuality- and commonality-based multiview multilabel learning (ICM2L)] to explicitly explore the individuality and commonality information of multilabel multiple view data in a unified model. Specifically, a common subspace is learned across different views to capture the shared patterns. Then, multiple individual classifiers are exploited to explore the characteristics of individual views. Next, an ensemble strategy is adopted to make a prediction. Finally, we develop an alternative solution to jointly optimize our model, which can enhance the robustness of the proposed model toward rare labels and reinforce the reciprocal effects of individuality and commonality among heterogeneous views, and thus further improve the performance. Experiments on various real-word datasets validate the effectiveness of ICM2L against the state-of-the-art solutions, and ICM2L can leverage the individuality and commonality information to achieve an improved performance as well as to enhance the robustness toward rare labels.**

***Index Terms*—Commonality, ensemble classification, individuality, multilabel learning, multiview learning.**

I. Introduction

IN MANY real-world applications, data are often associated with several heterogeneous feature representations, each of which gives a different view of the data. For example, a news Web page can be represented by two heterogeneous views, one is from the text (and image) information of the Web page itself, and the other is from the hyperlink to other pages; an image can be described using different features, such as texture descriptors, shape descriptors, color descriptors, surrounding texts, and so on. As a natural formulation for this type of data, multiview learning has attracted a lot of attention in machine learning and in various application domains [1], [2].

Although diverse multiview learning methods have been proposed in the literature over the past years, they still have some limitations. On the one hand, most previous studies often assume that each sample is annotated with a single label [3], [4]. Nevertheless, in real-life applications, individual samples usually have more than one label. For instance, an image can be simultaneously annotated with several labels, such as sea, sky, and seagull; a Web page could be tagged with multiple topics given as labels, such as economics, culture, sports, and politics. On the other hand, the majority of the existing studies are supervised approaches that require a large number of labeled samples [5], [6]. In practice, nevertheless, it is rather difficult and expensive to collect labeled samples, while unlabeled samples are easy to accumulate. Given this, a few semisupervised multiview multilabel learning approaches [7], [8] have been proposed to leverage limited labeled and abundant unlabeled samples. The key motivation behind them is to capture the complementary patterns among multiple views, which can boost the performance of multiview learning [9].

Another limitation of the aforementioned methods is that they do not explicitly account for the distinctive information of individual views, which might degenerate their performance for a variety of reasons. First, with respect to features, since multiview samples have heterogeneous feature representations, in which each representation encodes different properties of the samples, they may fail to capture the global structure of multiview data without exploring the distinctive information hidden in individual views. Second, with respect to labels, since each individual view captures a specific property of data, it is impossible for one view to comprehensively characterize all the relevant labels, especially, when data are annotated with multiple labels. As a result, leveraging the *individuality* of each view, along with the *commonality* of multiple views may further improve the performance of the model, compared to focusing only on the individuality (or commonality) of the views.

Some researchers have already explored the commonality and individuality of multiview data for classification and clustering [10]–[15]. It has been shown that the utilization of individual and shared patterns is beneficial for latent representation learning [11], [12]; multioutput problem [15]; and

Manuscript received April 23, 2019; revised August 29, 2019; accepted October 28, 2019. This work was supported in part by the Natural Science Foundation of China under Grant 61872300 and Grant 61873214, in part by the Fundamental Research Funds for the Central Universities under Grant XDJK2019B024, in part by the Natural Science Foundation of CQ CSTC under Grant cstc2018jcyjAX0228, and in part by the King Abdullah University of Science and Technology, Saudi Arabia. This article was recommended by Associate Editor S. Ventura. *(Corresponding author: Guoxian Yu.)*

Q. Tan, G. Yu, and J. Wang are with the College of Computer and Information Sciences, Southwest University, Chongqing 400715, China (e-mail: qiaoyut@gmail.com; gxyu@swu.edu.cn; kingjun@swu.edu.cn).

C. Domeniconi is with the Department of Computer Science, George Mason University, Fairfax, VA, USA (e-mail: carlotta@cs.gmu.edu).

X. Zhang is with the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia (e-mail: xiangliang.zhang@kaust.edu.sa).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2950560

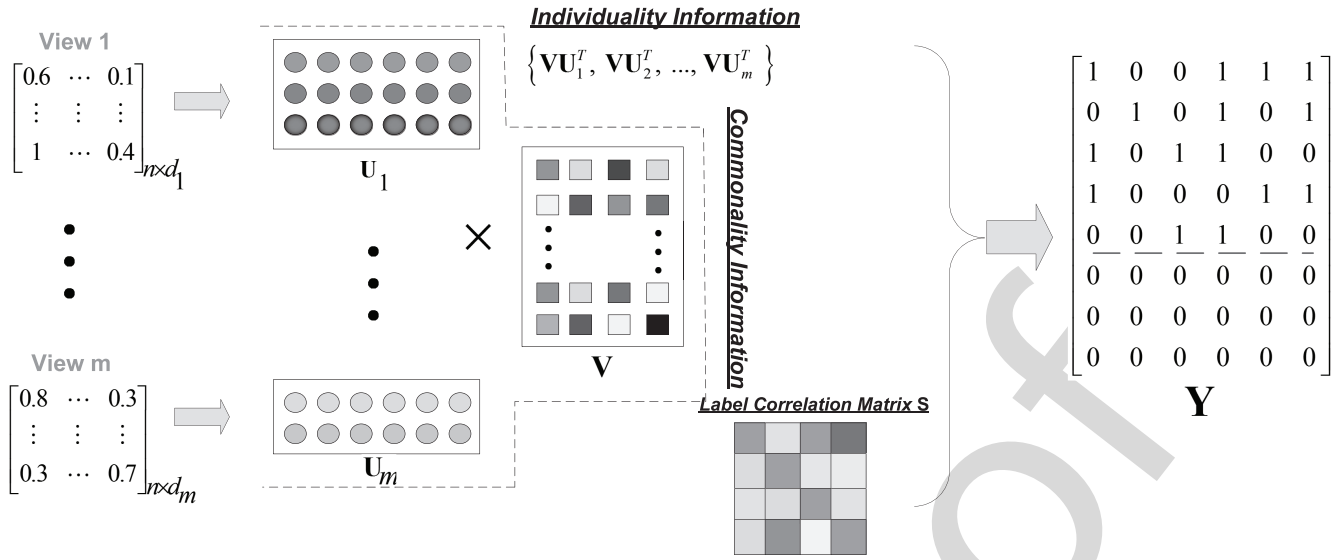


Fig. 1. Method overview. ICM2L jointly leverages commonality and individuality information of multiple data views, as well as label correlations. The commonality information of heterogeneous views is captured by the shared low-dimensional representation $\mathbf{V}$; while the individuality information of multiple views is captured by the reconstructed data matrices $\{\mathbf{VU}_v\}_{v=1}^{m}$.

multilabel classification [10]. In the multiview multilabel classification scenarios, however, they may result in suboptimal classifiers due to the isolated learning of hidden features and multilabel classifier [11], [12], or the lack of capacity to capture label correlations [10], [15]. More important, these methods mainly focus on learning a general classifier for all labels and treat them equally, which ignore the essential label imbalance problem in multilabel learning [16], [17]. To this end, the learned classifier may lose discriminant ability on rare labels, which are widely witnessed in the real-world applications. Motivated by this, in this article, we propose to explore how individual and common patterns of multiview data could be utilized to improve the performance of multilabel classification as well as in advancing the robustness of classifier toward rare labels.

To address the aforementioned issues, we propose a novel approach, called individuality- and commonality-based multiview multilabel learning (ICM2L), to explicitly account for the individual and shared patterns hidden in different views. As shown in Fig. 1, given multiple heterogeneous features of the input data, ICM2L seeks a shared subspace across heterogeneous views, which captures the commonality of different views and adapts an ensemble classifier-based on the shared subspace and on the other individual feature spaces, as well as on the label information in a unified model. In this way, both the shared and view-specific information of different views could be used to boost the performance via a mutually beneficial effect and, thus, further improve the performance of the model. The main contributions of this article are summarized as follows.

1) The proposed ICM2L can explicitly and jointly employ the individuality and commonality information of multiview multilabel data. It learns a shared subspace from different views, label correlations, and an ensemble classifier based on individual and shared feature spaces in a unified model.
2) ICM2L can explore the individuality of multiple views; as a result, it is superior to other methods in discovering rare labels.
3) We develop an alternative optimization solution to iteratively optimize our model. Extensive empirical results on the benchmark datasets demonstrate the superiority of ICM2L with respect to related and competitive methods, such as lrMMC [7], LSML [8], CSMSC [13], MLAN [4], and SMMCL [18].

The remainder of this article is organized as follows. In Section II, we briefly introduce the related work. Section III presents the proposed ICM2L. The experimental results and conclusions are discussed in Sections IV and V, respectively.

II. Related Work

This article is related to two branches of studies: 1) multilabel learning and 2) multiview learning. In this section, we briefly review some related works in these two fields. For more details, please refer to [2] and [19].

A. Multilabel Learning

Different from the binary classification scenarios, where each sample is associated with only one single semantic label, multilabel learning aims at assigning a set of discrete nonexclusive labels to a sample and has received increasing interest in different machine learning tasks [20]. For instance, Huang *et al.* [21] and [22] assumed that fully supervised signals are available and focus on learning multilabel classifiers under supervised setting. Such assumption, however, may not hold in real-world applications, because it requires exhaustive efforts to annotate multilabel samples. To avoid this limitation, researchers have resorted to develop semisupervised multilabel classifiers [23]–[27], in which limited labeled samples, as well as abundant unlabeled samples, are jointly used for training. Besides, considering the fact that labeled data

is tagged by human efforts, they might have some missing or noisy labels [28]–[33], several approaches have been proposed to design multilabel classifiers under weak-label setting [28], [29], [34], [35] or with noisy labels [32], [36]–[38].

Although the aforementioned methods have achieved the state-of-the-art performance for multilabel data, they mainly emphasize on single-view data and are not ready for multiview data. In fact, it has been proved that directly applying the existing multilabel algorithms to multiview data by concatenating multiple feature vectors (views) together will result in a compromised performance [2], [7]. The reason is that such concatenation operation fails to explore the intraview and interview relationships across heterogeneous views, which is very important for successful multiview learning models [39]. In addition, given that label correlation is crucial for the success of multilabel learning [40], how to develop an effective algorithm that can jointly utilize the heterogeneous information as well as important label correlations among multiview multilabel data still remains a challenge.

B. Multiview Learning

Due to the ubiquity of multiview data, multiview learning has been an active research field in many real-world applications [2]. Several multiview learning approaches were proposed to analyze multiview multilabel data recently. For example, Nie *et al.* [4] proposed a nearly parameter-free multiview model MLAN by integrating semisupervised classification and local structure learning simultaneously. Liu *et al.* [7] proposed a matrix factorization-based multiview framework lrMMC, which first seeks a shared representation of multiple views and then conducts classification based on matrix completion on the shared feature space. Nevertheless, lrMMC models the fusion of multiple views and the follow-up prediction tasks as separate objectives, which may lead to suboptimal solution. To avoid such a risk, some unified multiview multilabel learning methods have been proposed [8], [41]. Specifically, Tan *et al.* [41] aimed to improve the multilabel prediction performance by seeking a shared subspace from incomplete views with weak labels, local label correlations, and a predictor in this subspace in a unified model. Zhang *et al.* [8] sought a common feature representation and the corresponding projection model between the learned subspace and labels by simultaneously enforcing the consistence of latent semantic bases among different views in the kernel spaces. However, although such unified models can utilize the shared and complimentary information among different views to some extent, they do not explicitly take into account the individual patterns of heterogeneous views. As a result, they may have a compromised performance.

It has been recognized that exploring individuality and commonality of heterogeneous features can further boost the performance of multiview data mining [10]–[12]. Sagonas *et al.* [11] and Hu *et al.* [12] investigated the benefit of individual and common patterns in multiview data to learn more discriminant low-dimensional representations. However, they both focus on representation learning and cannot be directly applied to the multilabel classification problem. Similarly, [13] and [14] separately learn the individual and common representations of multiview data and then conduct clustering on the merged representations. They not only ignore the important correlations of multilabel data but also suffer from a two-stage fashion that may result in the suboptimal problem. To bridge the gap, Liu *et al.* [10] introduced a unified model to jointly learn compact latent features and multilabel classifier. Specifically, it assumes that the final latent feature vector with size d is composed of multiple view-specific features with size d_s and one shared common feature vector with size d_c, such that $d = md_s + d_c$, where m denotes the number of views. With this assumption, this unified model is able to generate comprehensive feature representations that can capture both individual and common patterns of multiview data. However, it still targets on developing a general multilabel classifier for all labels and may lose discriminant ability toward rare labels, which refers to the label imbalance problem in multilabel classification. Moreover, it also cannot capture the crucial correlations between multiple labels.

One similar work with our ICM2L is SMMCL [18], which is a self-pace-based label propagation method that models the common consensus and individuality of multiple teacher classifiers, each for one view. Although SMMCL can iteratively propagate labels to the most informative candidate unlabeled samples during training, it suffers from the scalable issue and cannot work on a moderate(large)-scale data as shown in our experiments. Besides, SMMCL also ignores the label correlations of multilabeled data, which is the cornerstone for successful multilabel learning algorithms [40]. Another core distinction between SMMCL and our model is that SMMCL uses individual information of multiview data to discover the most informative unlabeled samples, while our method utilizes such information to improve the discriminant ability toward rare labels and the robustness. Extensive empirical results on the benchmark datasets demonstrate the superiority of ICM2L to these related competitive methods [4], [7], [8], [18], [42].

III. Proposed Approach

A. Problem Statement and Notations

Suppose $\mathcal{X} = \{\mathbf{X}_v\}_{v=1}^m$ represents a dataset with n samples and m views, where $\mathbf{X}_v = [\mathbf{x}_v^1, \mathbf{x}_v^2, \ldots, \mathbf{x}_v^n] \in \mathbb{R}^{n \times d_v}$ indicates the full feature space in view v. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T \in \{-1, 1\}^{n \times c}$ is the corresponding label matrix, where $\mathbf{y}_i \in \{-1, 1\}^c$ is the label vector of $\mathbf{x}_i$ and c is the number of distinct labels. $\mathbf{y}_{ic'} = 1$ $(c' = 1, \ldots, c)$ means the c'th label is relevant; otherwise, $\mathbf{y}_{ic'} = -1$. Without loss of generality, we assume that out of n samples, the first l are labeled samples, while the remaining u are unlabeled samples. Our goal is to learn a predictive label matrix $\hat{\mathbf{Y}} \in \{1, -1\}^{n \times c}$ from heterogeneous feature representations $\{\mathbf{X}_v\}_{v=1}^m$ and partial label matrix $\mathbf{Y}$.

B. Problem Formulation

An intuitive strategy to deal with multiview multilabel setting is to concatenate multiview features into a single vector, and then conduct classification based on the classical

multilabel learning algorithms [19]. Such concatenation, however, ignores the fact that features are extracted from different spaces with different statistical properties, and directly applying multilabel methods on the concatenated data may suffer from the overfitting problem, and often leads to high time complexity, which may be unacceptable in the real-world applications. Besides, feature concatenation also ignores the useful diversity information across different views and might lead to a suboptimal problem [8].

In order to take advantage of consensus information, we advocate to seek the shared subspace from different views based on the matrix factorization techniques [43]. Specifically, we resort to non-negative matrix factorization (NMF) [44], [45] for its power in extracting the representations of diverse data [46]. It is worth noting that the major difference between NMF and other matrix factorization methods, such as singular-value decomposition [47], is the non-negative constraints, which helps to obtain a part-based representation, as well as enhancing the interpretability of the learned subspace. Another reason for the choice of NMF is due to the fact that most multiview data are naturally non-negative or can be easily transformed into non-negative ones, without changing the proximity between the original data. Our model can also be adapted to mix-sign data by replacing NMF with other matrix factorization techniques (i.e., semi-NMF [48]). Besides, our method is flexible for basic matrix factorization techniques, thus more powerful algorithms, that is, deep matrix factorization models [49], are expected to further improve the performance. Given heterogeneous representations of data $\{\mathbf{X}_v\}_{v=1}^m$ and the label matrix $\mathbf{Y}$, our unified objective function is given as follows:

$$\begin{aligned}\mathcal{L} &= \sum_{v=1}^{m} \left\|\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T\right\|_F^2 + \alpha\|\mathbf{V}\mathbf{W}_0 - \mathbf{Y}\|_F^2 + \beta\|\mathbf{W}_0\|_F^2 \\ &\text{s.t. } \mathbf{U} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{1}$$

where $\mathbf{U}_v \in \mathbb{R}^{d_v \times k}$ denotes the individual matrix for the vth view; $\mathbf{V} \in \mathbb{R}^{n \times k}$ is the shared subspace, where k is the desired low-rank size; $\|.\|_F$ represents the Frobenius norm; and $\mathbf{U}_v \geq 0$ and $\mathbf{V} \geq 0$ are the non-negative constraints for the matrices. $\mathbf{W}_0 \in \mathbb{R}^{k \times c}$ is the coefficient matrix corresponding to $\mathbf{V}$, and α and β are two tradeoff parameters. The first part of (1) is the consensus term, which aims to seek the common low-dimensional representation of multiview data under the assumption that different views have distinct mapping matrices $\{\mathbf{U}_v\}_{v=1}^m$, but share the same latent feature space $\mathbf{V}$. For this reason, we can make use of the consensus information across multiviews to some extent. The second part of (1) is the standard empirical loss term. It measures the empirical loss on labeled samples and ensures that the predicted label matrix is consistent with the initial ground-truth label assignment. By jointly optimizing the two terms, we can exploit the label information $\mathbf{Y}$ to induce the shared subspace toward a semantic label space. It not only helps to obtain a discriminate subspace but also may alleviate the widely spread semantic gap [50] between the input heterogeneous feature spaces and the semantic label space, since $\mathbf{V}$ can be viewed as a bridge between them. The last part is the widely used ℓ_2-norm regulation term, it is included to avoid the overfitting problem and reduce the impact of noisy features.

By minimizing (1), we can learn a discriminate predictor by jointly using both shared information among views and label information of labeled samples. But the predictor will not be discriminant enough, as we completely ignore another important issue in multiview multilabel learning, that is, the specific characteristics of individual views (*individuality*), which may further improve the performance. As discussed before, individual patterns hidden in multiple views are useful in boosting the robustness of multilabel classifier toward rare labels. In addition, these individual patterns can also boost the robustness of the predictor, due to the known usefulness of diversity for ensemble learning. As such, we need to capture the individual information of different views to further improve the performance of our model.

Considering the fact that distinctive mapping matrices $\{\mathbf{U}_v\}_{v=1}^m$ are actually used to encode the corresponding unique properties of individual views, we define another set of coefficient matrices $\{\mathbf{W}_v\}_{v=1}^m$ to capture the distinctive characteristics of different views, where $\mathbf{W}_v \in \mathbb{R}^{d_v \times c}$ maps the reconstructed feature matrix $\mathbf{V}\mathbf{U}_v^T$ of the vth view to the label space. In this way, we leverage the individual and common information of heterogeneous views and further extend (1) as follows:

$$\begin{aligned}\mathcal{L} = &\sum_{v=0}^{m} \left\|\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T\right\|_F^2 + \alpha\|\mathbf{V}\mathbf{W}_0 - \mathbf{Y}\|_F^2 \\ &+ \frac{1-\alpha}{m}\sum_{v=1}^{m} \left\|\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v - \mathbf{Y}\right\|_F^2 \\ &+ \beta\left(\|\mathbf{W}_0\|_F^2 + \|\mathbf{W}_v\|_F^2\right)\end{aligned} \tag{2}$$

where $\mathbf{X}_0 = \mathbf{V}$ and $\mathbf{U}_0$ is an identity matrix, $\alpha \in (0, 1)$ is a tradeoff parameter which balances the contribution of individuality and commonality among multiview data. Here, $\mathbf{X}_0$ can be regarded as an additionally introduced common view spanned by the shared subspace $\mathbf{V}$. It is worth noting that $\mathbf{V}$ can be regarded as the dictionary matrix for all views, and $\mathbf{U}_v$ represents the view-specific coding coefficients. Therefore, $\mathbf{V}\mathbf{U}_v^T$ targets to reserve the view-specific input signals, while $\mathbf{V}$ captures the shared information for all views. Hence, $\mathbf{V}$ and $\mathbf{V}\mathbf{U}_v^T$ could be regarded as different representations of multiview data. By minimizing (2), we can achieve two goals. On the one hand, since both common and individual information of heterogeneous views are employed in an elegant way, the learned shared subspace is more discriminate. On the other hand, the final predictive model is also enhanced because it absorbs both the public labels appeared in most views as well as the individual labels embedded in several specific views.

Another inherent property of learning from multilabel data is how to utilize label correlations, and this issue has not been addressed in (2). Label correlation has long been regarded as a fundamental challenge that can be used to improve the performance of multilabel learning [19], [40]. To address this limitation, we try to leverage the label correlation among labels

to estimate the predicted likelihood scores as follows:

$$\mathcal{L} = \sum_{v=0}^{m} \left\| \mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T \right\|_F^2 + \alpha \|\mathbf{V}\mathbf{W}_0\mathbf{S} - \mathbf{Y}\|_F^2 + \frac{1-\alpha}{m} \sum_{v=1}^{m} \left\| \mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S} - \mathbf{Y} \right\|_F^2 + \beta\Big(\|\mathbf{W}_0\|_F^2 + \|\mathbf{W}_v\|_F^2 \Big) \quad (3)$$

where $\mathbf{S} \in \mathbb{R}^{c\times c}$ represents the label correlations matrix, which can be estimated from the known $\mathbf{Y}$. For example, the correlation is often measured by the cosine similarity. However, since labeled samples may be not sufficient, we advocate to learn it by leveraging the features and already known labels of the training data. In experiments, we randomly initialize $\mathbf{S}$ and treat it as a trainable parameter during optimization.

The final prediction label matrix $\hat{\mathbf{Y}}$ is given by majority voting $\hat{\mathbf{Y}} = \mathbf{V}(\sum_{v=1}^{m} \mathbf{U}_v\mathbf{W}_v + \mathbf{W}_0)\mathbf{S}$. Equation (3) not only explicitly considers the common and individual information among multiple views in a principle way but also absorbs label information to induce the shared subspace and enhance its discriminate power. Another advantage of our model is that it exploits ensemble learning to further derive robust results, especially for rare labels. Our following experiments will confirm these advantages.

C. Optimization

The objective function in (3) involves $\{\mathbf{U}_v\}_{v=1}^m$, $\mathbf{V}$, $\mathbf{S}$, $\{\mathbf{W}_v\}_{v=1}^m$ and $\mathbf{W}_0$, and it is not easy to optimize all the variables simultaneously. Given that we adopt an alternating optimization technique to optimize the objective function by alternatively optimizing one variable while fixing other variables.

1) Update $\{\boldsymbol{U}_v\}_{v=1}^m$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{W}_v\}_{v=1}^m$, $\boldsymbol{W_0}$, and $\boldsymbol{S}$: When $\mathbf{V}$, $\mathbf{W}$, $\mathbf{W}_v$, and $\mathbf{S}$ are fixed, the computation of $\mathbf{U}_v$ is independent from $\mathbf{U}_{v'}, v' \neq v$. Thus, for each view v, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{U}_v$:

$$\mathcal{L}(\mathbf{U}_v) = -2\mathbf{X}_v^T\mathbf{V} + 2\mathbf{U}_v\mathbf{V}^T\mathbf{V} - 2(1-\alpha)/m\mathbf{W}_v\mathbf{S}\mathbf{Y}^T\mathbf{V} + 2(1-\alpha)/m\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}. \quad (4)$$

Using the Karush–Kuhn–Tucker (KKT) condition [51], we can derive the following updating rule:

$$(\mathbf{U}_v)_{i,j} \leftarrow (\mathbf{U}_v)_{i,j} \frac{\left(\mathbf{X}_v^T\mathbf{V} + \frac{1-\alpha}{m}\mathbf{W}_v\mathbf{S}\mathbf{Y}_v^T\mathbf{V}\right)_{i,j}}{\left(\mathbf{U}_v\mathbf{V}^T\mathbf{V} + \frac{1-\alpha}{m}\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\right)_{i,j}}. \quad (5)$$

2) Update $\boldsymbol{V}$ With Fixed $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$, $\boldsymbol{W_0}$ and $\boldsymbol{S}$: When fixed $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=1}^m$, $\mathbf{W}_0$, and $\mathbf{S}$, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{V}$ to zero:

$$2\sum_{v=0}^{m}\big(\mathbf{V}\mathbf{U}_v^T\mathbf{U}_v - \mathbf{X}_v\mathbf{U}_v\big) - 2\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + 2\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + 2\alpha\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T\mathbf{W}_0^T - 2\alpha\mathbf{Y}\mathbf{S}^T\mathbf{W}_0^T = 0. \quad (6)$$

Then, we have the following closed-form solution for $\mathbf{V}$, which is updated efficiently:

$$\mathbf{V} = \frac{\sum_{v=0}^{m}\theta_v\mathbf{X}_v\mathbf{U}_v + \alpha\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v + \mathbf{Q}_1}{\sum_{v=1}^{m}\theta_v\mathbf{U}_v^T\mathbf{U}_v + \alpha\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T\mathbf{W}_0^T + \mathbf{Q}_2} \quad (7)$$

where $\mathbf{Q}_1 = (1-\alpha/m)\sum_{v=0}^{m}\mathbf{Y}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v$ and $\mathbf{Q}_2 = (1-\alpha/m)\sum_{v=1}^{m}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T\mathbf{W}_v^T\mathbf{U}_v$.

3) Update $\boldsymbol{W}_v$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\boldsymbol{W_0}$, and $\boldsymbol{S}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\mathbf{W}_0$, and $\mathbf{S}$ are fixed, similarly to the update of $\mathbf{U}_v$, we have the following equation by setting the derivative with respect to $\mathbf{W}_v$:

$$2\frac{1-\alpha}{m}\big(\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T - \mathbf{U}_v\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\big) + 2\beta\mathbf{W}_v. \quad (8)$$

Based on the KKT condition, we can derive the following updating rule:

$$(\mathbf{W}_v)_{i,j} \leftarrow (\mathbf{W}_v)_{i,j} \frac{\left(\frac{1-\alpha}{m}\mathbf{U}_v\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right)_{i,j}}{\left(\frac{1-\alpha}{m}\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S}\mathbf{S}^T + \beta\mathbf{W}_v\right)_{i,j}}. \quad (9)$$

4) Update $\boldsymbol{W_0}$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$ and $\boldsymbol{S}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=1}^m$, and $\mathbf{S}$ are fixed, we obtain the following equation by taking the derivative of (3) with respect to $\mathbf{W}_0$ to zero:

$$\mathcal{L}(\mathbf{W}_0) = 2\alpha\big(\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T - \mathbf{V}^T\mathbf{Y}\mathbf{S}^T\big) + 2\beta\mathbf{W}_0. \quad (10)$$

Based on the KKT condition, we can derive the following update rule:

$$(\mathbf{W}_0)_{i,j} \leftarrow (\mathbf{W}_0)_{i,j} \frac{\left(\alpha\mathbf{V}^T\mathbf{Y}\mathbf{S}^T\right)_{i,j}}{\left(\alpha\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\mathbf{S}^T + \beta\mathbf{W}_0\right)_{i,j}}. \quad (11)$$

5) Update $\boldsymbol{S}$ With Fixed $\boldsymbol{V}$, $\{\boldsymbol{U}_v\}_{v=1}^m$, $\{\boldsymbol{W}_v\}_{v=1}^m$ and $\boldsymbol{W_0}$: When $\mathbf{V}$, $\{\mathbf{U}_v\}_{v=1}^m$, $\{\mathbf{W}_v\}_{v=0}^m$, and $\mathbf{W}_0$ are fixed, similarly to the update of $\mathbf{V}$, we have the following equation for $\mathbf{S}$ by setting the derivative with respect to $\mathbf{S}$ to zero:

$$2\sum_{v=0}^{m}\frac{1-\alpha}{m}\big(\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v\mathbf{S} - \mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{Y}\big) + 2\alpha\big(\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S} - \mathbf{W}_0^T\mathbf{V}^T\mathbf{Y}\big) = 0. \quad (12)$$

We therefore have the following closed-form solution for $\mathbf{S}$:

$$\mathbf{S} = \left(\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{V}\mathbf{U}_v^T\mathbf{W}_v + \alpha\mathbf{W}^T\mathbf{V}^T\mathbf{V}\mathbf{W}_0\mathbf{S}\right)^{-1} \times \left(\frac{1-\alpha}{m}\sum_{v=0}^{m}\mathbf{W}_v^T\mathbf{U}_v\mathbf{V}^T\mathbf{Y} + \alpha\mathbf{W}_0^T\mathbf{V}^T\mathbf{Y}\right). \quad (13)$$

Given the above iterative optimization process, we summarize the main procedure of ICM2L in Algorithm 1.

D. Complexity Analysis

The time complexity of ICM2L is dominated by the matrix multiplication and matrix inverse operations. The time complexity of matrix inverse for $\mathbf{V}$ and $\mathbf{S}$ is relatively small, so the time complexity of ICM2L is mainly driven by the computation for $\mathbf{U}_v$, $\mathbf{W}_v$, and $\mathbf{W}_0$. Concretely, the complexity for solving $\mathbf{U}_v$, $\mathbf{W}_v$, and $\mathbf{W}_0$ is $O(d_{\max}nk + d_{\max}k^2 + d_{\max}ck)$,

Algorithm 1 ICM2L

Input:

$\{\mathbf{X}_v\}_{v=1}^m$: n training samples with m views
$\mathbf{Y}$: Initial label matrix for n samples
k: Dimensionality of the shared subspace
α and β: Trade-off parameters used in Eq. (3)
ε: Convergence threshold
t: Number of iterations

Output:

$\hat{\mathbf{Y}}$: Predicted label likelihood matrix for n samples

Randomly initialize $\mathbf{U}_v$, $\mathbf{W}_v$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{S}$;
Compute $\mathcal{L}_0$ by Eq. (3);
for $i = 1, 2, \cdots, t$ **do**
  for $v = 1, 2, \cdots, m$ **do**
    Update $\mathbf{U}_v$ by Eq. (5);
  end
  Update $\mathbf{V}$ by Eq. (7);
  for $v = 1, 2, \cdots, m$ **do**
    Update $\mathbf{W}_v$ by Eq. (9);
  end
  Update $\mathbf{W}_0$ by Eq. (11);
  Update $\mathbf{S}$ by Eq. (13);
  Update $\mathcal{L}_i$ by Eq. (3);
  If $|\mathcal{L}_i - \mathcal{L}_{i-1}| \leq \varepsilon$, Return.
end

$O(d_{\max}nk)$, and $O(nck)$, respectively, where $d_{\max}$ is the largest dimensionality of the views. Since $n \gg k$ and $n \gg c$, the overall time complexity of ICM2L is $O(d_{\max}nkt)$, where t is the number of iterations to reach convergence. In practice, if $d_{\max} \ll n$, the total complexity of ICM2L scales with the number of samples. In our experiments, we found that t usually does not exceed 80. In addition, some views have sparse feature matrices, so the actual time cost of the above operations can be further reduced.

IV. Experiments

In this section, we conduct extensive experiments over six real-world datasets to evaluate the efficiency and effectiveness of the proposed framework. There are four major questions we aim to answer.

1) How effective is ICM2L compared with other related methods in classifying multiview multilabel data?
2) How robust is ICM2L in discovering rare labels compared with the state-of-the-art methods?
3) What are the impacts of two parameters α and β on ICM2L?
4) How efficient is ICM2L in modeling multiview multilabel learning problem?

A. Experimental Setup

Six multiview datasets that we employed in the experiments are all publicly available. The statistics of them are summarized in Table I. Yeast is a biological dataset with two views [52], one view is the genetic expression and the other is the phylogenetic profile of a gene. Core15k, Pascal07, and ESPGame are the three widely used multiview image datasets.[1] We collected the multiple features of these images from [53], where each image is represented by six representative feature views: HUE, SIFT, GIST, HSV, RGB, and LAB. Each sample in Mirflicker[2] and Nus-wide[3] consists of an image and textual tags, we construct the two views (image and text) according to [54]. For each dataset, we randomly sample 30% data for training and use the remaining 70% data for testing (unlabeled data).

TABLE I
Statistics of Four Multiview Datasets: n Is the Number of Samples; m Is the Number of Views; c Is the Number of Distinct Labels; #avg Is the Average Number of Labels per Sample; $d_{\min}$ Is the Smallest Dimension of All Views

datasets	n	m	c	#avg	d_{min}
Yeast	2417	2	14	4.237	24
Core15k	4999	6	260	3.396	100
Pascal07	9963	6	20	1.465	100
ESPGame	20770	6	268	4.686	100
Mirflicker	25000	2	24	3.794	512
Nus-wide	260648	2	81	2.783	500

Baseline Methods: To study the performance of ICM2L, we compare it with six state-of-the-art methods. In addition, to investigate the contribution of encoding common information, individual information, and using label correlations, we include three variants, namely, ICM2L-c, ICM2L-i, and ICM2L-lc.

1) lrMMC [7] leverages a low-dimensional common representation of all views and matrix completion for multilabel classification.
2) LSML [8] is a recent multiview multilabel learning framework that learns the shared subpace among heterogeneous features as well as the follow-up predictor in a unified objective function.
3) MLAN [4] is another unified multiview learning method and initially focuses on the single-label classification problem; we adapt it for multilabel scenario by assigning multiple labels instead of a single one to unlabeled data.
4) CSMSC [13] is a multiview subspace learning approach, which can jointly extract the consistency and specificity of heterogeneous features for subspace representation learning. We adopt the extracted individual and common representation features as inputs to our model to train the ensemble classifier.
5) SMMCL [18] is a self-paced-based multiview multilabel learning method, which considers both the individuality and commonality characteristics among multiview data.
6) MVMC-LS [42] is a multiview learning approach based on matrix completion, it combines the matrix completion outputs of different views with various weights.
7) ICM2L-c is a variant of ICM2L by excluding individual information and makes prediction by $\mathbf{W}$ (commonality).

[1] http://lear.inrialpes.fr/data/
[2] http://press.liacs.nl/mirflickr/mirdownload.html
[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.html

TABLE II
RESULTS ON FOUR DATASETS WITH $k = 0.5d_{\min}$. $d_{\min}$ REPRESENTS THE MINIMUM DIMENSIONALITY OF MULTIPLE VIEWS

Dataset	metric	lrMMC	MLAN	MVMC-LS	CSMSC	LSML	SMMCL	ICM2L
Yeast	Accuracy	$\mathbf{0.539 \pm 0.001}$	0.381 ± 0.001	0.517 ± 0.001	0.537 ± 0.001	0.535 ± 0.004	$\mathbf{0.542 \pm 0.002}$	0.536 ± 0.004
	1-RL	0.787 ± 0.001	$\mathbf{0.811 \pm 0.002}$	0.761 ± 0.000	0.793 ± 0.001	0.797 ± 0.003	$\mathbf{0.816 \pm 0.001}$	0.788 ± 0.005
	AP	0.703 ± 0.001	0.459 ± 0.001	0.662 ± 0.000	0.698 ± 0.002	0.702 ± 0.004	$\mathbf{0.717 \pm 0.004}$	0.702 ± 0.003
	AUC	0.798 ± 0.001	0.589 ± 0.001	0.778 ± 0.000	0.797 ± 0.001	0.799 ± 0.002	$\mathbf{0.812 \pm 0.003}$	0.799 ± 0.005
Core15k	Accuracy	$\mathbf{0.191 \pm 0.001}$	0.103 ± 0.001	0.172 ± 0.000	$\mathbf{0.193 \pm 0.001}$	$\mathbf{0.193 \pm 0.001}$	$\mathbf{0.192 \pm 0.002}$	$\mathbf{0.194 \pm 0.001}$
	1-RL	0.758 ± 0.001	0.521 ± 0.002	0.750 ± 0.000	0.762 ± 0.001	0.768 ± 0.001	0.771 ± 0.001	$\mathbf{0.795 \pm 0.003}$
	AP	0.236 ± 0.001	0.146 ± 0.001	0.215 ± 0.001	0.432 ± 0.001	0.256 ± 0.001	0.259 ± 0.002	$\mathbf{0.279 \pm 0.004}$
	AUC	0.760 ± 0.001	0.710 ± 0.001	0.752 ± 0.000	0.767 ± 0.001	0.774 ± 0.001	0.778 ± 0.003	$\mathbf{0.797 \pm 0.003}$
Pascal07	Accuracy	0.278 ± 0.000	0.205 ± 0.001	0.264 ± 0.001	0.281 ± 0.002	0.283 ± 0.001	0.285 ± 0.001	$\mathbf{0.296 \pm 0.003}$
	1-RL	0.697 ± 0.001	0.502 ± 0.001	0.692 ± 0.001	0.715 ± 0.001	0.725 ± 0.003	0.730 ± 0.002	$\mathbf{0.756 \pm 0.005}$
	AP	0.429 ± 0.000	0.350 ± 0.002	0.401 ± 0.002	0.424 ± 0.002	0.446 ± 0.001	0.451 ± 0.001	$\mathbf{0.452 \pm 0.001}$
	AUC	0.727 ± 0.000	0.646 ± 0.001	0.725 ± 0.001	0.739 ± 0.003	0.758 ± 0.002	0.763 ± 0.002	$\mathbf{0.785 \pm 0.005}$
ESPGame	Accuracy	0.170 ± 0.000	0.088 ± 0.000	0.134 ± 0.001	0.177 ± 0.000	0.189 ± 0.001	0.192 ± 0.001	$\mathbf{0.206 \pm 0.001}$
	1-RL	0.777 ± 0.000	0.521 ± 0.001	0.764 ± 0.001	0.784 ± 0.001	$\mathbf{0.796 \pm 0.001}$	$\mathbf{0.798 \pm 0.002}$	$\mathbf{0.796 \pm 0.001}$
	AP	0.189 ± 0.000	0.111 ± 0.000	0.167 ± 0.000	0.194 ± 0.001	0.205 ± 0.001	0.207 ± 0.001	$\mathbf{0.220 \pm 0.002}$
	AUC	0.783 ± 0.000	0.642 ± 0.000	0.770 ± 0.001	0.785 ± 0.002	0.789 ± 0.000	0.790 ± 0.002	$\mathbf{0.803 \pm 0.001}$
Mirflickr	Accuracy	0.376 ± 0.001	0.282 ± 0.004	0.355 ± 0.001	0.387 ± 0.002	0.394 ± 0.002	0.412 ± 0.001	$\mathbf{0.436 \pm 0.001}$
	1-RL	0.750 ± 0.002	0.675 ± 0.002	0.736 ± 0.002	0.758 ± 0.001	0.765 ± 0.003	0.773 ± 0.001	$\mathbf{0.796 \pm 0.001}$
	AP	0.466 ± 0.003	0.401 ± 0.001	0.419 ± 0.002	0.471 ± 0.002	0.485 ± 0.001	0.498 ± 0.002	$\mathbf{0.536 \pm 0.002}$
	AUC	0.757 ± 0.001	0.664 ± 0.003	0.737 ± 0.001	0.761 ± 0.001	0.769 ± 0.001	0.774 ± 0.001	$\mathbf{0.790 \pm 0.001}$
Nus-wide	Accuracy	0.249 ± 0.002	0.198 ± 0.001	0.231 ± 0.002	0.253 ± 0.001	0.268 ± 0.003	0.286 ± 0.004	$\mathbf{0.332 \pm 0.001}$
	1-RL	0.791 ± 0.003	0.714 ± 0.004	0.779 ± 0.002	0.804 ± 0.001	0.819 ± 0.002	0.835 ± 0.003	$\mathbf{0.923 \pm 0.002}$
	AP	0.311 ± 0.002	0.243 ± 0.001	0.304 ± 0.003	0.321 ± 0.001	0.338 ± 0.003	0.367 ± 0.004	$\mathbf{0.448 \pm 0.004}$
	AUC	0.812 ± 0.001	0.735 ± 0.002	0.798 ± 0.003	0.823 ± 0.004	0.841 ± 0.002	0.876 ± 0.003	$\mathbf{0.933 \pm 0.002}$

8) ICM2L-i is a variant of ICM2L by excluding common information and makes prediction by integrating $\{\mathbf{W}_v\}_{v=1}^m$ (individuality).
9) ICM2L-lc is a variant of ICML2 by excluding label correlations.

For comparing methods, five-fold cross-validation on the training set is used to select the optimal parameter values from the range as suggested in the original papers. For our method, we selected the parameters α and β in the range of $\{0.1, 0.2, \ldots, 1\}$ and $\{0.1, 0.3, \ldots, 2\}$, respectively. To avoid random effects, all the experiments are independently repeated ten times, and both the mean and standard deviation are reported. For each comparing method, the code is released or provided by corresponding authors. The code of ICM2L is publicly available at http://mlda.swu.edu.cn/codes.php?name=ICM2L.

Evaluation: Four widely used metrics are adopted for performance comparisons: 1) accuracy; 2) ranking loss (RL); 3) average precision (AP); and 4) average AUC. Note that these metrics generally belong to two categories: 1) example-based criterion and 2) label-based criterion [55]. RL, AP, and Accuracy are the example-based metrics, while AUC is a label-based criterion. They evaluate the performance from ranking and classification perspectives [19], in which RL, AP, and AUC are ranking-based metrics, while Accuracy is an example-based classification criteria. Formal definition of the four metrics can be found in [19] and [55]. *Accuracy* requires the predicted label-likelihood vector to be a binary indicator vector. Here, we consider the labels corresponding to the r largest entries of the vector of the ith sample as the predicted labels, where r is determined as the average number of labels (round to next integer) of labeled samples. To maintain consistency with other evaluation metrics, in our experiments, we report 1-RL instead of RL. Thus, as for other metrics, the higher the value of 1-RL, the better the performance is. These metrics evaluate multilabel classification from different points of view, and it is unlikely for a method outperforming all the other techniques across all the metrics.

B. Effectiveness of ICM2L

To investigate the first question stated at the beginning of this section, we compare the classification performance of all methods on the six datasets listed in Table II. Since SMMCL consumes a lot of memory in training, we can only obtain its results on the Yeast dataset with a server (CentOS 6.9, 64-GB RAM, and MATLAB 2014a). For this reason, we independently sample 3000 instances from large datasets 20 times to construct new sampled datasets and report the best results of them. In Table II, the best (or comparable best) results are highlighted in **boldface** using the pairwise t-test at 95% significance level. Besides paired student's t-test, we also apply the Friedman's test [56] with a *post-hoc* Tukey's test [57] to assess the significant difference between ICM2L and other comparing methods, all the p-values are smaller than 10^{-4} and 0.04 for the Friedman's and Tukey's tests, respectively. We implement the test based on the *Friedman* and *multcompare* functions in MATLAB.

From the results reported in Table II, we can observe that ICM2L outperforms other comparing methods in most cases, especially, on the large-scale datasets. Although MVMC-LS and lrMMC are all designed for multiview multilabel data, and MVMC-LS is almost always outperformed by lrMMC. This is mainly because lrMMC exploits the commonality information among multiple views by assuming that different views are generated from a common low-dimensional subspace, while MVMC-LS just utilizes individual information by combining outputs of different views that cannot make full use of complementary information among views. Since MVMC-LS learns view combination coefficients by two-fold cross-validation in the training data, which may result in scarce labeled training samples in our semisupervised setting and

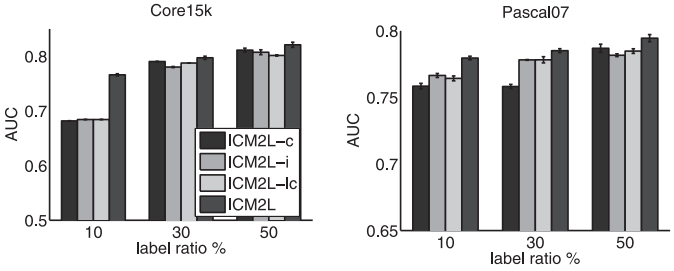


Fig. 2. Results of ICM2L variants under different ratios of labeled samples on Core15k and Pascal07.

impact the combination coefficients learning. Both lrMMC and LSML aim to learn a shared subspace among heterogeneous views, but lrMMC loses to LSML many times. The possible reason is that lrMMC is a two-step method. lrMMC separates the learning process of shared subspace and multilabel classifier, which may result in a suboptimal solution, while LSML can learn the subspace and the follow-up predictor based on the learned representation simultaneously. CSMSC is also a two-step approach, but it outperforms lrMMC in many cases. The main reason is that CSMSC takes advantage of multiple view-specific representations and common representation to train an ensemble classifier, while lrMMC learns a general multilabel classifier for all labels. This comparison justifies our motivation to explore individuality and commonality information to develop the discriminant classifier. Both CSMSC and ICM2L develop an ensemble classifier for classification, but CSMSC loses to ICM2L in most cases. The crucial reason is that ICM2L jointly learn feature representations and ensemble classifier, while CSMSC separates the two learning processes. This comparison indicates the importance to learn feature representation and follow-up classifier jointly.

MLAN is another unified multiview learning method, but it still loses to LSML almost in all cases. The possible reason is that MLAN is naturally designed for single-label problems and it cannot employ label correlations among multiple labels, which is very important in multilabel data as suggested in the literature. Both LSML and ICM2L can utilize label correlations and the commonality information to make prediction; LSML is outperformed by ICM2L in most cases. The principal reason is that ICM2L explicitly utilizes the individuality information of the views. These comparisons justify our motivation to jointly exploit both commonality and individual information of multiple data views. SMMCL is a recent state-of-the-art method that considers the individuality and commonality patterns of multiview data. SMMCL can iteratively propagate labels to the most informative candidate unlabeled samples during training and these samples are then augmented into the training set as labeled data for the next iteration. For this reason, SMMCL, in general, consumes more labeled samples for training, and achieves better performance than other comparing methods in many cases. However, SMMCL is still outperformed by ICM2L in many cases, especially, over relative large-scale datasets, for example, Core15k, Pascal07, ESPGame, Mirflickr, and Nus-wide. The crucial intuition behind is two sides: 1) the label imbalance problem in the large-scale datasets is more serious than that in Yeast and 2) the bonus of augmenting training set is limited with the increased size of dataset, since a number of labeled data is ready to train an effective semisupervised algorithm. These observations further validate our motivation to directly exploit individual information of various views for prediction instead of subspace learning. In addition, another bottleneck of SMMCL lies in its poor scalability, since it needs huge memory for training. These comparisons validate the effectiveness of ICM2L.

C. Effectiveness of ICM2L via Component Analysis

To further justify the effectiveness of our model in capturing the commonality and individuality information of multiple data views, as well as of the label correlations, we conduct additional component analysis experiments on the Core15k and Pascal07 datasets and report the AUC values in Fig. 2. In this figure, we set the ratios of labeled data equal to 10%, 30%, and 50%, respectively.

From the figure, we can see that the performance of all the variants of ICM2L increases as the increasing of labeled training data, and ICM2L outperforms its variants across all the settings. ICM2L-c and ICM2L-i disregard the individuality information and commonality information, respectively, and they are outperformed by ICM2L in many cases. This is mainly because ICM2L utilizes both types of information, and thus improves the final performance. These results corroborate our motivation to explicitly leverage the individual and common information of multiple data views. Both ICM2L and ICML2-lc take advantage of multiview data from individual and common aspects, but ICM2L outperforms ICM2L-lc in many cases across two datasets. The inherent reason is that ICM2L captures label correlations, which are very important in multilabel learning. This fact validates the necessity of capturing label correlations and also proves the effectiveness of the learned label correlation matrix $\mathbf{S}$.

D. Robustness of ICM2L Toward Rare Labels

To answer the second proposed question, we conduct experiments to quantify the benefit of utilizing individual information toward rare labels. Let IR_c denote the imbalance ratio of label c, which is calculated by the ratio between the number of negative samples and that of positive samples for label c. We generate an imbalance dataset from Core15k by first discarding the samples that are annotated with few than three labels. Then, we split the labels of the new dataset as general labels and rare labels based on IR_c. Specifically, we decide label c as a general label if $\mathrm{IR}_c \leq 50$; otherwise, regarding the labels as rare label. In addition, to further investigate the performance of all methods in extreme cases, we divide the rare label into three levels: rare_1, rare_2, and rare_3. rare_1 includes the labels with $50 < \mathrm{IR}_c \leq 100$, rare_2 includes the labels with $100 < \mathrm{IR}_c \leq 150$, and rare_3 includes the labels with $\mathrm{IR}_c \geq 150$. Finally, the new dataset has 4966 samples associated with 119 labels, 45 labels with $\mathrm{IR}_C \leq 50$, and the other 74 rare labels. Specifically, 36 labels for rare_1,

TABLE III
EXPERIMENTAL RESULTS OF COMPARING METHODS ON THE PROCESSED CORE15K DATASET WITH DIFFERENT SCALES OF IMBALANCED LABELS

IR		lrMMC	MLAN	MVMC-LS	CSMSC	LSML	SMMCL	ICM2L
≤ 50 (general)	1-RL	0.714 ± 0.001	0.497 ± 0.001	0.684 ± 0.002	0.724 ± 0.002	0.736 ± 0.001	0.742 ± 0.002	$\mathbf{0.778 \pm 0.002}$
	AP	0.298 ± 0.001	0.213 ± 0.002	0.243 ± 0.000	0.306 ± 0.003	0.312 ± 0.001	0.323 ± 0.001	$\mathbf{0.349 \pm 0.002}$
$50 < IR_c \leq 100$(rare$_1$)	1-RL	0.522 ± 0.001	0.394 ± 0.001	0.481 ± 0.001	0.618 ± 0.001	0.622 ± 0.001	0.633 ± 0.002	$\mathbf{0.681 \pm 0.001}$
	AP	0.246 ± 0.001	0.203 ± 0.001	0.219 ± 0.000	0.278 ± 0.001	0.283 ± 0.001	0.296 ± 0.001	$\mathbf{0.346 \pm 0.001}$
$100 < IR_c \leq 150$(rare$_2$)	1-RL	0.487 ± 0.001	0.326 ± 0.001	0.415 ± 0.001	0.516 ± 0.001	0.519 ± 0.001	0.535 ± 0.001	$\mathbf{0.598 \pm 0.002}$
	AP	0.141 ± 0.001	0.122 ± 0.001	0.137 ± 0.001	0.185 ± 0.001	0.187 ± 0.001	0.197 ± 0.001	$\mathbf{0.263 \pm 0.001}$
$IR_c > 150$(rare$_3$)	1-RL	0.458 ± 0.001	0.218 ± 0.001	0.398 ± 0.001	0.484 ± 0.002	0.486 ± 0.001	0.499 ± 0.002	$\mathbf{0.598 \pm 0.001}$
	AP	0.134 ± 0.001	0.109 ± 0.001	0.121 ± 0.001	0.151 ± 0.001	0.153 ± 0.001	0.168 ± 0.001	$\mathbf{0.225 \pm 0.002}$

27 labels for rare$_2$, and 11 labels for rare$_3$. Table III shows the performance of comparing methods on the general labels (IR$_C \leq 50$) and rare labels (IR$_C > 50$). The experimental configurations are the same with that in Section IV-B.

From the table, we can see that the performance of all methods decreases with IR$_c$ increases, and ICM2L outperforms other comparing methods not only on the general labels but also on all rare label cases, which is consistent with the results in Table II. An interesting observation is that the differences between our model and other comparing methods in rare label cases are larger than those in general labels, especially in the most imbalance situations, for example, rare$_2$ and rare$_3$. These comparisons validate the robustness of our model in classifying rare labels. In addition, although ICM2L, SMMCL, and LSML aim to employ the complementary information, as well as individual information among different views, the last two methods still lose to ICM2L in most cases. The reason behind this fact is that ICM2L utilizes individual information of multiple views in order to capture rare labels hidden in specific views, while the other two approaches exploit individual patterns to enforce subspace learning. These results intuitively justify our motivation to capture individual characteristics of multiple data views. Another interesting observation is that CSMSC performs more similar to LSML in imbalance scenarios, in which the performance gap between them in rare$_2$ (or rare$_3$) is smaller than that in the general case. This result further validates the importance of directly exploring individual and common patterns to construct robust classifier.

E. Parameter Analysis

We now study the third proposed question. ICM2L has two parameters α and β, which control the importance of individual information and regularization terms, respectively. We test the sensitivity of ICM2L with respect to α and β in the range $\{0.1, 0.2, \ldots, 1\}$ and $\{0.1, 0.3, \ldots, 2\}$, respectively. We report Accuracy and AUC on Yeast in Fig. 3; the results for the other datasets and evaluation metrics are similar and lead to similar conclusions.

From Fig. 3, we can observe that ICM2L obtains relatively good performance when α is around 0.6 and β is around 0.7. In addition, when $\alpha \to 0$ or $\alpha \to 1$, the performance of ICM2L is reduced. These results further confirm the contribution of commonality information and individual information. Another interesting observation is that the performance of ICM2L decreases more sharply when $\alpha \approx 1$ than that when α is around 0. Since α controls the importance between individual and common patterns, and $\alpha \approx 1$ means that we discard the individual information of multiview data and only focus on common patterns. In contrast, $\alpha \approx 0$ means that we only utilize individual patterns hidden in each view and discard the common patterns among them. These results again indicate the importance of individual information. When β is close to 0, the Accuracy and AUC values tend to decrease. This fact validates the effectiveness of the l_2 term. In our experiments, we set $\alpha = 0.6$ and $\beta = 0.7$.

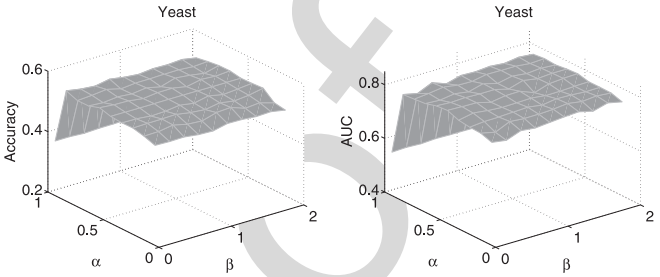


Fig. 3. Results of ICM2L under different input values of α and β.

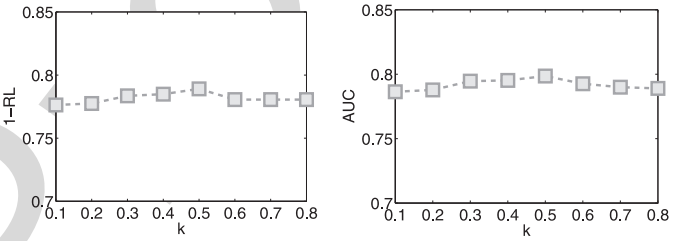


Fig. 4. Results of ICM2L under different input values of k.

In addition, we also conduct experiments to investigate the sensitivity of ICM2L with respect to k. Fig. 4 reports the 1-RL and AUC values of ICM2L on the Yeast dataset with k varying from $0.1d_{\min}$ to $0.8d_{\min}$. As we can see, the performance of ICM2L first increases with k rising, then it decreases when $k > 0.5d_{\min}$. For this reason, we set $k = 0.5d_{\min}$ in experiments. For Mirflickr and Nus-wide, we set $d_{\min} = 100$ for simplicity. For general multilabel datasets, we recommend setting the dimension of features ($k \approx$) #Avg $\log_2(N)$, where #Avg indicates the number of associated labels per sample. The behind intuition is that the minimal number of bits to encode N data points in one class is $\log_2(N)$ in the information theory. For multilabel data, it also needs to consider the statistical property of labels, which could be captured by #Avg. As such, #Avg $\log_2(N)$ bits are needed.

F. Efficiency of ICM2L

To investigate the proposed last question, we conduct experiments on all datasets with the same configuration in

TABLE IV
RUNTIME COMPARISON (IN SECONDS)

	lrMMC	MLAN	MVLC-LS	CSMSC	LSML	SMMCL	ICM2L
Yeast	8.03	197.90	39.57	25.78	23.11	23.20	10.11
Core15k	290.56	357.92	13892.32	437.88	387.24	543.70	325.41
Pascal07	908.49	1604.04	8783.45	1378.44	1023.38	1588.92	768.29
ESPGame	1512.63	9785.80	21748.23	5436.28	4792.04	7332.42	3918.44
Mirflickr	1245.72	7752.33	18428.18	4493.65	3986.13	5856.18	3245.11
Total	3965.43	19697.99	62891.75	11772.03	10211.90	15344.42	8267.36

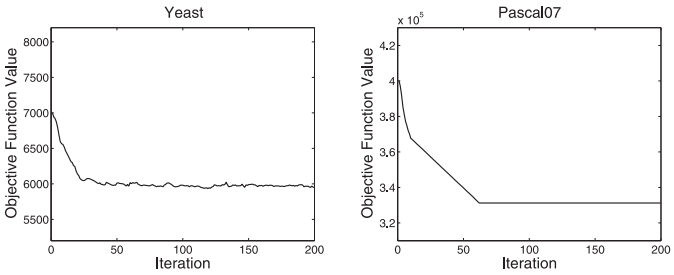


Fig. 5. Convergence trend analysis.

Section IV-B, and report the runtime costs of all methods and the convergence trend of ICM2L. We only report the runtime cost of comparing methods over datasets excluding Nus-wide since the runtime cost on Nus-wide is generally much bigger than other datasets. We observe similar runtime cost trend on Nus-wide. Table IV reports the runtime costs of all the approaches on a server (CentOS 6.9 with Inter Xeon E5-2678, 64-GB RAM, and MATLAB 2014a). From Table IV, we can see that ICM2L is much faster than MLAN, MVLC-LS, CSMSC, and LSML in general. However, lrMMC runs much faster than ICM2L in most cases. This is because lrMMC is a two-step method that learns the shared subspace and the follow-up predictor in two separate steps, while ICM2L has to learn the low-dimensional representations and the multilabel classifier in each iteration. These comparisons corroborate the efficiency of our model.

Fig. 4 shows the convergence curve of ICM2L on the Yeast and Core15k datasets. We can see that ICM2L tends to converge after 70 iterations for the Yeast dataset, and after 60 iterations for the Core15k dataset. The convergence trends on the other datasets are the same as those reported in Fig. 4. Overall, ICM2L converges at most in 80 iterations for the datasets used in the experiments.

V. CONCLUSION

In this article, we investigated how to explore the individuality and commonality of heterogeneous features for effective multiview multilabel classification. To this end, a multiview multilabel framework termed ICM2L is presented. ICM2L learns a shared subspace of heterogeneous views, label correlations, and an ensemble classifier that captures both individuality and commonality information of multiple views in a principled way. Different from the previous works that focus on learning representative hidden representations by capturing the shared and individual patterns across multiple views, we utilize such information to improve the discriminant capacity of classifier toward rare labels. Experiments on several benchmark datasets demonstrate the superiority of the proposed model over related competitive solutions. In the future, we plan to further improve ICM2L by adapting nonlinear mapping functions with deep models.

ACKNOWLEDGMENT

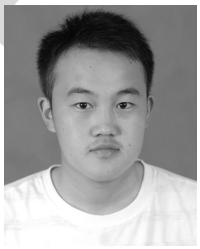
The authors would like to thank the authors who kindly shared their source code and datasets with them for the experiments.

REFERENCES

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.
[2] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.
[3] P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via CCA," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 199–207.
[4] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2408–2414.
[5] Q. Wang, H. Lv, J. Yue, and E. Mitchell, "Supervised multiview learning based on simultaneous learning of multiview intact and single view classifier," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 2293–2301, 2017.
[6] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[7] M. Liu, Y. Luo, D. Tao, C. Xu, and Y. Wen, "Low-rank multi-view learning in matrix completion for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2778–2784.
[8] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang, "Latent semantic aware multi-view multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4414–4421.
[9] P. Zhao, Y. Jiang, and Z.-H. Zhou, "Multi-view matrix completion for clustering with side information," in *Proc. 21st Pac.–Asia Conf. Knowl. Disc. Data Min.*, 2017, pp. 403–415.
[10] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1233–1246, Jun. 2015.
[11] C. Sagonas, E. Ververas, Y. Panagakis, and S. Zafeiriou, "Recovering joint and individual components in facial data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2668–2681, Nov. 2018.
[12] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, Sep. 2018.
[13] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3730–3737.
[14] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 586–594.
[15] O. G. R. Reyes and S. Ventura, "Performing multi-target regression via a parameter sharing-based deep network," *Int. J. Neural Syst.*, vol. 29, no. 9, 2019, Art. no. 1950014.
[16] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu, "Towards class-imbalance aware multi-label learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 4041–4047.

AQ4

[17] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1095–1107, May 2014.
[18] C. Gong, "Exploring commonality and individuality for multi-modal curriculum learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1926–1933.
[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[20] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Min.*, vol. 3, no. 3, pp. 1–13, 2007.
[21] J. Huang, G. Li, Q. Huang, and X. Wu, "Joint feature selection and classification for multilabel learning," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 876–889, Mar. 2018.
[22] S.-J. Huang and Z.-H. Zhou, "Multi-label learning by exploiting label correlations locally," in *Proc. AAAI Conf. Artif. Intell.*, 2012, pp. 949–955.
[23] X. Kong, M. K. Ng, and Z.-H. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[24] Y. Guo and D. Schuurmans, "Semi-supervised multi-label classification," in *Proc. Eur. Conf. Mach. Learn. Principles Pract. Knowl. Disc. Databases*, 2012, pp. 355–370.
[25] O. G. R. Pupo and S. Ventura, "Evolutionary strategy to perform batch-mode active learning on multi-label data," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, p. 46, 2018.
[26] O. Reyes, C. Morell, and S. Ventura, "Effective active learning strategy for multi-label learning," *Neurocomputing*, vol. 273, no. 1, pp. 494–508, 2018.
[27] A. H. Akbarnejad and M. S. Baghshah, "An efficient semi-supervised multi-label classifier capable of handling missing labels," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 229–242, Feb. 2019.
[28] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in *Proc. AAAI Conf. Artif. Intell.*, 2010, pp. 1862–1868.
[29] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 593–601.
[30] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proc. ACM Multimedia Conf.*, 2018, pp. 700–708.
[31] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.
[32] G. Yu *et al.*, "Feature-induced partial multi-label learning," in *Proc. IEEE Int. Conf. Data Min.*, 2018, pp. 1398–1403.
[33] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4302–4309.
[34] B. Wu, F. Jia, W. Liu, B. Ghanem, and S. Lyu, "Multi-label learning with missing labels using mixed dependency graphs," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 875–896, 2018.
[35] Q. Tan, Y. Yu, G. Yu, and J. Wang, "Semi-supervised multi-label classification using incomplete label information," *Neurocomputing*, vol. 260, pp. 192–202, Oct. 2017.
[36] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1910–1918.
[37] J. Tu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, and M. Guo, "Multi-label answer aggregation based on joint matrix factorization," in *Proc. IEEE Int. Conf. Data Min.*, 2018, pp. 517–526.
[38] C. Zhang, Z. Yu, H. Fu, P. Zhu, L. Chen, and Q. Hu, "Hybrid noise-oriented multilabel learning," *IEEE Trans. Cybern.*, to be published.
[39] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.
[40] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3780–3788.
[41] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Incomplete multi-view weak-label learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 2703–2709.
[42] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.
[43] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognit.*, vol. 48, no. 1, pp. 10–19, 2015.
[44] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[45] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.
[46] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2017.
[47] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Heidelberg, Germany: Springer, 1971, pp. 134–151.
[48] C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[49] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2019.
[50] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, p. 5, 2008.
[51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[52] E. L. Gibaja, J. M. Moyano, and S. Ventura, "An ensemble-based approach for multi-view multi-label classification," *Progr. Artif. Intell.*, vol. 5, no. 4, pp. 251–259, 2016.
[53] M. Guillaumin, J. J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 902–909.
[54] X. Liu, G. Yu, C. Domeniconi, J. Wang, Y. Ren, and M. Guo, "Ranking-based deep cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4400–4407.
[55] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Comput. Surveys*, vol. 47, no. 3, p. 52, 2015.
[56] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.
[57] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," in *Encyclopedia of Research Design*. Thousand Oaks, CA, USA: Sage, pp. 1–5, 2010.



Qiaoyu Tan received the Ph.D. degree from the Department of Computer Science, Texas A&M University, College Station, TX, USA.

He was a Research Assistant with the Machine Learning and Data Analysis Laboratory, Southwest University, Chongqing, China. His current research interests include machine learning and data mining.

Guoxian Yu received the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2013.

He is a Professor with the College of Computer and Information Science, Southwest University, Chongqing, China. His current research interests include data mining and bioinformatics.

Prof. Yu has served as PC member for IJCAI, AAAI, KDD, ICDM, SDM, WSDM, and NIPS and the Reviewer for the IEEE TRANSACTIONS JOURNALS and *Bioinformatics*.

Jun Wang received the B.Sc. and M.Eng. degrees in computer science and the Ph.D. degree in artificial intelligence from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively.

She is an Associate Professor with the College of Computer and Information Science, Southwest University, Chongqing, China. Her current research interests include machine learning, and data mining and their applications in bioinformatics.

AQ6

Carlotta Domeniconi is an Associate Professor with the Department of Computer Science, George Mason University, Fairfax, VA, USA. She has published extensively in premier journals and conferences in machine learning and data mining. Her current research interests include machine learning, pattern recognition, and data mining, with applications in text mining and bioinformatics.

Ms. Domeniconi has served as a PC member for KDD, ICDM, SDM, ECML-PKDD, and AAAI. She is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and *Knowledge and Information Systems*.

Xiangliang Zhang received the Ph.D. degree (Hons.) in computer science from INRIA-University Paris-Sud 11, France, in 2010.

She is an Associate Professor and Directs the Machine Intelligence and Knowledge Engineering Laboratory, King Abdullah University of Science and Technology. Her current research interests include diverse areas of machine learning and data mining.