Multi-label Classification with High-rank and High-order Label Correlations

Chongjie Si, Yuheng Jia, *Member, IEEE*, Ran Wang, *Senior Member, IEEE*, Min-Ling Zhang, *Senior Member, IEEE*, Yanghe Feng, Chongxiao Qu

Abstract—Exploiting label correlations is important to multi-label classification. Previous methods capture the high-order label correlations mainly by transforming the label matrix to a latent label space with low-rank matrix factorization. However, the label matrix is generally a full-rank or approximate full-rank matrix, making the low-rank factorization inappropriate. Besides, in the latent space, the label correlations will become implicit. To this end, we propose a simple yet effective method to depict the high-order label correlations explicitly, and at the same time maintain the high-rank of the label matrix. Moreover, we estimate the label correlations and infer model parameters simultaneously via the local geometric structure of the input to achieve mutual enhancement. Comparative studies over twelve benchmark data sets validate the effectiveness of the proposed algorithm in multi-label classification. The exploited high-order label correlations are consistent with common sense empirically. **Our code is publicly available at https://github.com/Chongjie-Si/HOMI.**

Index Terms—High-rank matrix approximation, high-order label correlations, multi-label classification.

1 Introduction

Recently, multi-label classification has attracted a lot of attention, aiming to solve real-world tasks with rich semantics [1], [2], [3], [4]. Specifically, in multi-label classification, one instance may be associated with several labels. For example, an image may be associated with a set of tags [5], and a piece of news may belong to several topics. Different from the traditional single-label classification problem which can be regarded as a degenerated version of multi-label classification, the overwhelming size of output space makes multi-label classification a much more challenging task.

Exploiting label correlations is of great importance, as label correlations can provide valuable semantic relationship for the output of multi-label classification. For instance, if two labels, "rainforest" and "soccer" are assigned to a sample, then the label "Brazil" may be also assigned to it. Similarly, if "teacher" and "blackboard" are present, it is very likely that label "classroom" will also be present. Based on how to explore the label correlations, multi-label classification methods can be roughly divided into three families: first-order, second-order and high-order. For first-order methods, the label correlations are not considered. For example, binary relevance (BR) [5] transformed multi-label classification into a set of independent binary classification problems and solved them separately. The second-order methods take the pairwise relationship of labels into considerations. For example, multi-label classification with Label specIfic FeaTures (LIFT) [6] employed clustering techniques to find second-order correlations between labels. However, in real-world scenarios, the label correlations may be much more complex than first-order and second-order relations. To this end, many high-order label correlations exploiting methods were proposed. For example, classifier chains (CC) [7] transformed the multi-label classification into a chain of binary classification problems. Random-k-labelsets (RAKEL) [8] converted multi-label classification into a set of multi-class classification problems over k randomly-chosen class labels. Some approaches assumed the labels were correlated in a hierarchical structure [9]. All the mentioned approaches specify the high-order correlations of labels manually, which will depress the classification performance if the manual setting is inappropriate.

Recently, some high-order approaches were proposed to automatically explore the high-order label correlations [10], [11]. They generally decomposed the label matrix to a latent space by low-rank matrix factorization [12], and then assumed the latent labels may capture the higher level semantic concepts [13]. *However, as can be seen from Table 1, the rank of the label matrix usually equals to or approximately equals to the number of labels, which means the label matrix is full-rank or approximate full-rank, making the low-rank matrix assumption inappropriate. Besides, in the latent space, the label correlations become indirect and semantically unclear.*

- *C. Si is with the Chien-Shiung Wu College, Southeast University, Nanjing 210096, China, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: chongjiesi@sjtu.edu.cn.*
- *Y. Jia is with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, and with the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China, and also with School of Computing & Information Sciences, Caritas Institute of Higher Education, Hong Kong. E-mail: yhjia@seu.edu.cn.*
- *M. Zhang is with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China e-mail: zhangml@seu.edu.cn.*
- *R. Wang is with the School of Mathematical Science, Shenzhen University, Shenzhen 518060, China, and also with Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Shenzhen University, Shenzhen 518060, China. E-mail: wangran@szu.edu.cn.*
- *Y. Feng is with the College of Systems Engineering, National University of Defense Technology.E-mail: fengyanghe@nudt.edu.cn.*
- *C. Qu is with the 52nd Research Institute of China Electronics Technology Group.E-mail:quchongxiao@163.com.*

TABLE 1. The rank of the label matrix on some commonly used multi-label data sets.

Data set	mediamill	CAL500	emotions	enron	bibtex	delicious	language log	birds	yeast	scene	corel5k	corel16k
Number of labels	101	174	6	53	159	983	75	19	14	6	374	153
Number of samples	43097	502	593	1702	7395	16105	1460	645	2417	2407	5000	13766
Size of label matrix	43907×101	502×174	593×6	1702×53	7395×159	16105×983	1460×75	645×19	2417×17	2407×6	5000×374	13766×153
Rank of label matrix	100	174	6	52	159	983	75	19	14	6	371	153

To solve these issues, in this paper, we propose a novel approach called HOMI, with High-Rank and **H**igh-**O**rder **M**ult**I**-Label learning. Specifically, we argue that if a label is highly correlated to a set of other labels, it can be easily reconstructed by the linear combination of that set of labels. Therefore, we propose to use self-representation to exploit the high-order label correlations for multi-label classification. Note that it can keep the rank of the label matrix unchanged and indicate the high-order correlations among labels explicitly. Moreover, the local geometric structure is also beneficial to multi-label classification, as if two samples are similar to each other in the feature space, they are likely to share similar labels. Here, we adopt an s-nearest-neighbors (SNN) graph to depict the local geometric structure of the input samples and incorporate the local geometric structure by a graph Laplacian regularization. Besides, the proposed approach naturally unifies high-order label correlations learning and multi-label prediction into a joint model via the graph Laplacian regularization, such that those two separate processes can be well interacted with each other to achieve mutual enhancement. Comprehensive experiments substantiate that HOMI outperforms the state-of-the-art multi-label classification methods significantly, and reasonable high-order label correlations can be constructed by HOMI.

The rest of the paper is organized as follows. We first review some related works in exploiting label correlations and explain why the label matrix should be full-rank or approximate full-rank in multi-label classification in Section 2. Then, we introduce the proposed approach as well as the numerical solution in Section 3, and present the experimental results and analysis in Section 4. Finally, conclusion is given in Section 5.

2 Related Work

2.1 Exploiting Label Correlations in Multi-label Classification

In multi-label classification, an instance is associated with a set of labels. In recent years, this new machine learning paradigm has made great progress and has been widely used in image classification [14], [15], [16], automatic annotation [17], [18], [19], web mining [20], [21], [22], audio recognition [23], [24], [25], [26] and information retrieval [27], [28], etc.

However, the task of inducing multi-label classification functions is challenging, as the classifier's output space is exponential in size to the number of possible class labels, i.e., $|2^{\mathcal{Y}}|$, where $\mathcal{Y}$ denotes the number of possible labels. A useful way to cope with this issue is to exploit label correlations to simplify the learning procedure. Based on the degree of label correlations used, the algorithms of multi-label classification can be divided into three categories [29]: first-order, second-order and high-order.

First-order methods do not take label correlations into consideration and assume that all the labels are independent. BR [5] is a prevailing first-order approach, transforming the original multi-label classification into a set of independent binary classification tasks. ML-KNN [30] is also a popular first-order algorithm based on k-nearest-neighbour classification. The major advantage of first-order approaches is the conceptual simplicity and high efficiency, for they are easy to understand and operate. Nevertheless, they ignore the label correlations, which results in performance degeneration.

Second-order approaches focus on pairwise label relations. For instance, calibrated label ranking (CLR) [31] and LIFT [6] are two representative approaches, transforming original multi-label classification into pairwise ranking problems. Second-order approaches are relatively more effective than first-order ones in exploiting label correlations. However, in real-world applications, the relationship of labels may be quite complex and sophisticated, such that the pairwise relationship cannot describe the real-world label correlations very well.

High-order approaches aim to dig the high-order label correlations. CC [7], for instance, converted the multi-label task into a chain of independent binary classification problems, with the ground-truth labels decoded into the features each time. RAKEL [8] reformulated the multi-label classification into several sets of multi-class classification tasks.

Recently, some low-rank based approaches proposed to learn high-order label correlations based on the assumption that the label matrix is low-rank, for there are correlations among labels in multi-label classification. For example, Zhu et al. [10] used low-rank decomposition in multi-label learning, exploiting global and local label correlations simultaneously, through learning a latent label representation and optimizing label manifolds [32]. Wang et al. [33] controlled the sparsity of the coefficient matrix to filter out label-specific features and applied low-rank constraints to the label matrix to mine the local correlations of class labels. Xu et al. [34] proposed an integrated framework that learns the correlations among labels while training the multi-label model simultaneously, and specifically adopted a low-rank structure to capture the complex correlations among labels. Moreover, Yu et al. [35] proposed an approach learning a linear instance-to-label mapping with low-rank structure, and implicitly taking advantage of global label correlations.

The general strategy of the above mentioned methods to capture label semantics is to decompose the label matrix to a latent label space by low-rank matrix factorization, [36], [37]. Specifically, denote the label matrix $\mathbf{Y} \in \mathbb{R}^{n \times l}$, with n and l being the number of samples and labels, respectively. The low-rank based approaches usually decompose $\mathbf{Y}$ into two smaller matrices $\mathbf{U}$ and $\mathbf{V}$, i.e., $\mathbf{Y}$ can be written as the product

$$\min_{\mathbf{UV}} \|\mathbf{Y} - \mathbf{UV}\|_F^2, \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{h\times l}$ is the latent label matrix exploiting higher level label semantics, and $\mathbf{U} \in \mathbb{R}^{n\times h}$ is a basis matrix relating the original labels to the latent labels. As h is smaller than l and n, the rank of $\mathbf{UV}$ is smaller than $\mathbf{Y}$, i.e., Eq. (1) approximates the label matrix by low-rank factorization.

2.2 Why is the label matrix full-rank?

However, we believe that the label matrix is full-rank or approximately full-rank, and accordingly the low-rank matrix assumption is not the best choice for multi-label learning. The reasons are as follows. First, the size of the label matrix $\mathbf{Y}$ is $n \times l$, as $l \ll n$ in general, the rank of $\mathbf{Y}$ is very likely to be or close to l. Second, although there are some connections among labels, the connections are usually not determinated. For example, if label A is related to label B, in other words, if A appears on a sample, we can infer that B has high possibility to be also appeared on that sample. But we cannot conclude that B will absolutely appear, as there will always be samples that only have label A or B. Therefore, those connections cannot reduce the rank of the label matrix, and likewise cannot result in a low-rank label matrix. Last but not least, as shown in the Table 1, the commonly used multi-label data sets are always full-rank or approximately full-rank, which further empirically validates that the label matrix of multi-label classification should be high-rank rather than low-rank.

As a summary, the label matrix in multi-label learning is usually a full-rank matrix, which cannot be well approximated by low-rank decomposition. Besides, in the latent space, the label correlations become indirect and semantically unclear. To solve these issues, in the next section, a new approach named HOMI is proposed, which can keep the rank of the label matrix unchanged, and indicate label correlations directly in the label space.

2.3 Deep-learning Based Multi-label Classification

Due to its robust learning capability, deep learning has emerged as an important technique for achieving multi-label classification. In those methods, effectively leveraging deep learning is critical for capturing intricate label dependencies. To exploit the underlying intricate label structure, Cisse et al. (2016) proposed ADIOS [38], which employs a novel deep architecture that partitions the labels into a Markov blanket chain, capitalizing on this partition to enhance classification performance. Wang et al. (2016) introduced CNN-RNN [39], which leverages recurrent neural networks (RNNs) to better model higher-order label dependencies. CNN-RNN learns a unified image-label embedding that encapsulates both semantic label dependencies and image-label relevance. Notably, this approach enables end-to-end training from scratch. Moreover, Nam et al. proposed an alternative technique to the traditional classifier chain method [40]. Their approach employs RNNs to convert the MLC problem into a sequential prediction task, with initially arbitrary label ordering. This method offers the advantage of focusing on predicting positive labels exclusively, significantly reducing the prediction space compared to the complete set of labels.

These methodologies exemplify diverse strategies for harnessing deep learning in MLC, each addressing label dependencies through distinctive means. However, different from these methods, HOMI explicitly reveals the label dependency based on the full-rank assumption, which further improves the MLC problem.

3 The Proposed Approach

Motivated by the fact that the label matrix is generally full-rank, in this section, we introduce HOMI to exploit the high-order label correlations for multi-label classification. Prior to that, we first briefly summarize the notations used in this paper. Formally, let $\mathcal{X} = \mathbb{R}^m$ denote the m-dimensional feature space and $\mathcal{Y} = \{c_1, c_2, .., c_l\}$ denote the label space of l labels, where $c_i \in \{0, 1\}$ stands for the ith label, multi-label classification learns a function $f : \mathcal{X} \to 2^{\mathcal{Y}}$ from the training data set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)|1 \le i \le n\}$, where $\mathbf{x}_i$ stands for the ith instance and $\mathbf{y}_i$ stands for the corresponding label set, and n is the number of instances. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^\mathsf{T} \in \mathbb{R}^{n\times m}$ denote the instance matrix and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^\mathsf{T} \in \{0, 1\}^{n\times l}$ denote the label matrix with l labels. Note that the value of the labels is binary, and we have $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{il}]$ where $y_{ij} = 1$ (resp. $y_{ij} = 0$) if the sample $\mathbf{x}_i$ has (resp. does not have) the jth label.

3.1 Basic Model

First, HOMI uses a weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l] \in \mathbb{R}^{m\times l}$ to map the instance to the labels by minimizing the following least squares loss:

$$\min_{\mathbf{W},\mathbf{z}} \|\mathbf{Y} - \mathbf{XW} - \mathbf{1}_n\mathbf{z}^\mathsf{T}\|_F^2 + \lambda(\|\mathbf{W}\|_F^2 + \|\mathbf{z}\|_2^2), \tag{2}$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ stand for the Frobenius norm and ℓ_2 norm of a matrix and a vector, respectively. $\mathbf{z} = [z_1, z_2, ..., z_l]^\mathsf{T} \in \mathbb{R}^l$ is the bias term and $\mathbf{1}_n \in \mathbb{R}^n$ is an all one vector. The first term in Eq. (2) measures the predictive error of the model, and the second term is the regularization of the weight matrix $\mathbf{W}$ and the bias $\mathbf{z}$, trying to control the complexity of the whole model, and $\lambda \ge 0$ is a hyper-parameter to balance those two terms.

3.2 Exploiting High-Order Label Information

As mentioned earlier, if a label is highly correlated with other labels, it can be easily reconstructed by those labels. Thus, we propose to adopt the self-representation strategy to dig the high-order correlations between labels, which can be mathematically formulated as

$$\min_{\mathbf{B},\mathbf{t}} \|\mathbf{Y} - \mathbf{YB} - \mathbf{1}_n\mathbf{t}^\mathsf{T}\|_F^2 + \lambda(\|\mathbf{B}\|_F^2 + \|\mathbf{t}\|_2^2), \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^{l\times l}$ records the high-order label correlations, and similar to Eq. (2), $\mathbf{t} \in \mathbb{R}^l$ is the bias term for better self-regression. We also introduce an additional penalty term on

$\mathbf{B}$ and $\mathbf{t}$ to avoid the trivial solution (i.e., $\mathbf{B} = \mathbf{I}$ with $\mathbf{I}$ being an identity matrix) and over-fitting.

The previous methods use low-rank matrix factorization to decompose the label matrix $\mathbf{Y}$ to a latent space to exploit the high-order correlations, however, the label matrix is usually a full-rank matrix. Technically, this is because although correlations exist in labels, the correlated labels also have a chance to exist alone on some samples. Table 1 also empirically verifies this observation. The full-rank property of the label matrix makes the low-rank factorization-based methods unreasonable. Differently, the adopted self-representation approach can keep the rank of label matrix unchanged [41]. Moreover, in the latent space, the label correlations can only be captured implicitly, while on the contrary, the elements in $\mathbf{B}$ can explicitly indicate the correlations between two labels.

3.3 Incorporating Local Geometric Structure

HOMI also takes the local geometric structure of instances into consideration, i.e., if $\mathbf{x}_i$ is similar to $\mathbf{x}_j$, the predictive label sets of them may have some labels in common. Specifically, we first calculate the Pearson correlation coefficient of samples, i.e.,

$$\mathbf{R}_{ij} = \frac{\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}},$$

where $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance of $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\sigma_{\mathbf{x}_i}$ is the standard deviation of $\mathbf{x}_i$. Afterwards, we construct an s-nearest-neighbors graph to capture the local geometric structure of the input, i.e.,

$$\mathbf{S}_{ij} = \begin{cases} \mathbf{R}_{ij} & \text{if}(i,j) \in \mathcal{N}_{si} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathcal{N}_{si}$ is the set of s nearest instances of the ith instance, and we choose the ones with the top s values of $\mathbf{R}_{ij}$ of the ith sample as its neighbors. Then, the local geometric structure is incorporated by minimizing

$$\sum_{i,j} \mathbf{S}_{ij} \|\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_j)\|_2^2 = \mathrm{tr}(\mathbf{K}^\mathsf{T}\mathbf{L}\mathbf{K}), \tag{5}$$

where $\mathbf{L} = \mathbf{Q} - \mathbf{S}$ is the graph Laplacian matrix with $\mathbf{Q}$ being the diagonal degree matrix of $\mathbf{S}$. In order to learn the local structural information more reasonably, we make $\mathbf{L}$ symmetric, i.e., $\mathbf{L} = \frac{1}{2}(\mathbf{L} + \mathbf{L}^\mathsf{T})$. $\mathbf{g}$ is the discriminative function, i.e. $\mathbf{g}(\mathbf{x}_i) = (\mathbf{x}_i\mathbf{W} + \mathbf{z}^\mathsf{T})\mathbf{B} + \mathbf{t}^\mathsf{T}$, which takes the high-order label correlations into account. $\mathbf{K} = [\mathbf{g}(\mathbf{x}_1), \mathbf{g}(\mathbf{x}_2), ..., \mathbf{g}(\mathbf{x}_n)]^\mathsf{T} \in \mathbb{R}^{n\times l}$ (i.e. $\mathbf{K} = (\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{z}^\mathsf{T})\mathbf{B} + \mathbf{1}_n\mathbf{t}^\mathsf{T}$). If Eq. (5) is minimized, two similar instances will have similar predictive label sets.

3.4 Model Formulation

Based on the above discussion, the objective function of HOMI is finally formulated as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{z},\mathbf{t}} & \|\mathbf{Y} - \mathbf{X}\mathbf{W} - \mathbf{1_n}\mathbf{z}^\mathsf{T}\|_F^2 + \gamma\mathrm{tr}(\mathbf{K}^\mathsf{T}\mathbf{L}\mathbf{K}) \\ & + \beta\|\mathbf{Y} - \mathbf{Y}\mathbf{B} - \mathbf{1_n}\mathbf{t}^\mathsf{T}\|_F^2 + \lambda(\|\mathbf{W}\|_F^2 \\ & + \|\mathbf{z}\|_2^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{t}\|_2^2), \end{aligned} \tag{6}$$

where γ, β, and λ denote different hyper-parameters to balance different terms. As will be illustrated in the experiments, HOMI is quite robust to those hyper-parameters. It is also worth pointing out that HOMI integrates high-order correlations exploitation and model prediction into a joint model via the graph Laplacian regularization term. It is able to simultaneously enhance those two processes via the joint learning.

3.5 Prediction

For an unseen instance $\mathbf{x}$, the discriminative function $\mathbf{g}$ of HOMI is obtained by

$$\mathbf{g}(\mathbf{x}) = (\mathbf{x}\mathbf{W} + \mathbf{z}^\mathsf{T})\mathbf{B} + \mathbf{t}^\mathsf{T}, \tag{7}$$

and the predictive label set is obtained by

$$\mathbf{y}_{pre} = f(\mathbf{x}) = \{c_i | \mathbf{g}_i(\mathbf{x}) > 0.5, 1 \le i \le l\}, \tag{8}$$

where $\mathbf{g}_i(\mathbf{x})$ is the ith element of $\mathbf{g}(\mathbf{x})$.

3.6 Numerical Solution to Eq. (6)

The problem in Eq. (6) is not convex in all the variables together, but it is convex to each variable with the remaining variables fixed. Therefore, we solved it by the following alternating optimization procedure.

Update W With fixed $\mathbf{B}$, $\mathbf{z}$ and $\mathbf{t}$, Eq. (6) becomes:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W} - \mathbf{1_n}\mathbf{z}^\mathsf{T}\|_F^2 + \gamma\mathrm{tr}(\mathbf{K}^\mathsf{T}\mathbf{L}\mathbf{K}) + \lambda\|\mathbf{W}\|_F^2. \tag{9}$$

Taking the gradient of Eq. (9) w.r.t. $\mathbf{W}$, we have

$$\begin{aligned} \nabla_{\mathbf{W}} = & \mathbf{X}^\mathsf{T}(\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{z}^\mathsf{T} - \mathbf{Y}) + \lambda\mathbf{W} \\ & + \gamma\mathbf{X}^\mathsf{T}\mathbf{L}(\mathbf{X}\mathbf{W}\mathbf{B} + \mathbf{1}_n\mathbf{z}^\mathsf{T}\mathbf{B} + \mathbf{1}_n\mathbf{t}^\mathsf{T})\mathbf{B}^\mathsf{T}. \end{aligned} \tag{10}$$

The optimal solution of Eq. (9) is achieved when $\nabla_{\mathbf{W}} = 0$, and accordingly we have the following Sylvester equation

$$\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{E} = \mathbf{Q} \tag{11}$$

where $\mathbf{A} = \frac{1}{\gamma}(\mathbf{X}^\mathsf{T}\mathbf{L}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})$, $\mathbf{E} = \mathbf{B}\mathbf{B}^\mathsf{T}$, $\mathbf{Q} = \frac{1}{\gamma}(\mathbf{X}^\mathsf{T}\mathbf{L}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{Y} - \mathbf{X}^\mathsf{T}\mathbf{1}_n\mathbf{z}^\mathsf{T} - \gamma\mathbf{X}^\mathsf{T}\mathbf{L}\mathbf{1_n}\mathbf{t}^\mathsf{T}\mathbf{B}^\mathsf{T} - \gamma\mathbf{X}^\mathsf{T}\mathbf{L}\mathbf{1_n}\mathbf{z}^\mathsf{T}\mathbf{B}\mathbf{B}^\mathsf{T})$, which can be efficiently solved according to [42].

Update B Fixing $\mathbf{W}$,$\mathbf{z}$ and $\mathbf{t}$, the $\mathbf{B}$-subproblem becomes

$$\min_{\mathbf{B}} \beta\|\mathbf{Y} - \mathbf{Y}\mathbf{B} - \mathbf{1}_n\mathbf{t}^\mathsf{T}\|_F^2 + \gamma\mathrm{tr}(\mathbf{K}^\mathsf{T}\mathbf{L}\mathbf{K}) + \lambda\|\mathbf{B}\|_F^2,$$

which is a quadratic optimization problem, and the solution is obtained by setting its derivative to 0, i.e.,

$$\begin{aligned} \mathbf{B} = & (\beta\mathbf{Y}^\mathsf{T}\mathbf{Y} + \lambda\mathbf{I} + \gamma(\mathbf{W}^\mathsf{T}\mathbf{X}^\mathsf{T} + \mathbf{z}\mathbf{1_n}^\mathsf{T})\mathbf{L}(\mathbf{X}\mathbf{W} + \mathbf{1}_n\mathbf{z}^\mathsf{T}))^{-1} \\ & (\beta\mathbf{Y}^\mathsf{T}\mathbf{Y} - \beta\mathbf{Y}^\mathsf{T}\mathbf{1}_n\mathbf{t}^\mathsf{T} - \gamma(\mathbf{W}^\mathsf{T}\mathbf{X}^\mathsf{T} + \mathbf{z}\mathbf{1_n}^\mathsf{T})\mathbf{L}\mathbf{1_n}\mathbf{t}^\mathsf{T}). \end{aligned} \tag{12}$$

Update z With fixed $\mathbf{W}$, $\mathbf{B}$ and $\mathbf{t}$, the $\mathbf{z}$-subproblem is rewritten as

$$\min_{\mathbf{z}} \|\mathbf{Y} - \mathbf{X}\mathbf{W} - \mathbf{1}_n\mathbf{z}^\mathsf{T}\|_F^2 + \gamma\mathrm{tr}(\mathbf{K}^\mathsf{T}\mathbf{L}\mathbf{K}) + \lambda\|\mathbf{z}\|_2^2,$$

which is also a quadratic optimization problem, and the optimal solution is achieved when the derivative approaches zero, i.e.,

$$\mathbf{z} = ((n+\lambda)\mathbf{I} + \gamma \mathbf{1}_n{}^\top \mathbf{L}\mathbf{1}_n \mathbf{B}\mathbf{B}^\top)^{-1}(\mathbf{Y}^\top \mathbf{1_n} - \mathbf{W}^\top \mathbf{X}^\top \mathbf{1_n} - \gamma \mathbf{B}(\mathbf{B}^\top \mathbf{W}^\top \mathbf{X}^\top + \mathbf{t}\mathbf{1_n}{}^\top)\mathbf{L}\mathbf{1_n}). \tag{13}$$

Update t With other variables fixed, the $\mathbf{t}$-subproblem is reformulated as

$$\min_{\mathbf{t}} \beta \|\mathbf{Y} - \mathbf{Y}\mathbf{B} - \mathbf{1_n}\mathbf{t}^\top\|_F^2 + \gamma \mathrm{tr}(\mathbf{K}^\top \mathbf{L}\mathbf{K}) + \lambda \|\mathbf{t}\|_2^2,$$

which is also a quadratic optimization problem, and the solution of it is achieved when $\nabla_{\mathbf{t}} = 0$, i.e.,

$$\mathbf{t} = \frac{1}{(\beta n + \gamma \mathbf{1}_n{}^\top \mathbf{L}\mathbf{1}_n + \lambda)}(\beta \mathbf{Y}^\top \mathbf{1}_n - \beta \mathbf{B}^\top \mathbf{Y}^\top \mathbf{1}_n - \gamma \mathbf{B}^\top (\mathbf{W}^\top \mathbf{X}^\top + \mathbf{z}\mathbf{1_n}{}^\top)\mathbf{L}\mathbf{1}_n). \tag{14}$$

In summary, HOMI first randomly initialize $\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t}$, and then iteratively and alternatively update these four variables. The iteration stops when the difference between two consecutive loss is less than 0.001. Finally the label set $\mathbf{y}_{pre}$ for an unseen $\mathbf{x}$ is predicted according to Eq. (8). The whole schedule is concluded in algorithm 1.

Algorithm 1 HOMI.

Require: Training data set $\mathcal{D}$; s; β; γ; λ, max iteration number $iter$; an unseen instance $\mathbf{x}$.
Ensure: Predicted label set $\mathbf{y}_{pre}$ for $\mathbf{x}$.
1: Initialize $\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t}$ to $\mathbf{0}$;
2: Calculate $\mathbf{L}$ by Eqs. (4)-(5);
3: **Repeat:**
4: Update $\mathbf{W}$ according to Eq. (11);
5: Update $\mathbf{B}$ according to Eq. (12);
6: Update $\mathbf{z}$ according to Eq. (13);
7: Update $\mathbf{t}$ according to Eq. (14);
8: **If** number of iteration $\geq iter$: **break**
9: **Until** Convergence;
10: **return** $\mathbf{y}_{pre}$ according to Eq. (8).

3.7 Complexity Analysis

HOMI iteratively solves four optimization problems. To solve $\mathbf{W}$, HOMI needs to solve a Sylvester equation, which can be computed in $O(\max(m^3, l^3, mn^2, m^2 n, mnl))$; the rest three optimization problems are all quadratic optimization problems with the complexity of $O(\max(l^2 n, lmn, l^3))$, $O(\max(l^3, lmn, n^2))$ and $O(\max(l^3, lmn, n^2))$. In summary, the overall complexity of HOMI is $O(\max(m^3, l^3, mn^2, m^2 n, mnl))$ in one iteration.

3.8 Convergence Analysis

The proposed numerical solution in Algorithm 1 is guaranteed to converge theoretically. Specifically, define the optimal function in Eq. (6) as F, the problem in Eq. (6) is not convex in all the variables together, but it is convex to each variable with the remaining variables fixed. Therefore, we solved it by an alternating optimization procedure. Specifically, we transform the original problem into four subproblems, where each subproblems can be solved efficiently. We need to minimize $F(\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t})$ with four variables $\mathbf{W}, \mathbf{B}, \mathbf{z}$ and $\mathbf{t}$. We transform the original problem into four subproblems $\min_{\mathbf{W}} F(\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t})$, $\min_{\mathbf{B}} F(\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t})$, $\min_{\mathbf{z}} F(\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t})$ and $\min_{\mathbf{t}} F(\mathbf{W}, \mathbf{B}, \mathbf{z}, \mathbf{t})$ and solve them alternatively and iteratively. When solving the $\mathbf{W}$-subproblem $\min_{\mathbf{W}} F(\mathbf{W}_{k-1}, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1})$ at the kth iteration, the variables $\mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1}$ are fixed, and we try to find optimal $\mathbf{W}_k$ to minimize the corresponding function value. It is obvious that $F(\mathbf{W}_{k-1}, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1}) \geq F(\mathbf{W}_k, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1})$. Similarly, when solving the $\mathbf{B}$-subproblem $\min_{\mathbf{B}} F(\mathbf{W}_k, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1})$ at the kth iteration, the variables $\mathbf{W}_k, \mathbf{z}_{k-1}, \mathbf{t}_{k-1}$ are fixed, and $F(\mathbf{W}_k, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1}) \geq F(\mathbf{W}_k, \mathbf{B}_k, \mathbf{z}_{k-1}, \mathbf{t}_{k-1})$, and the same to $\mathbf{z}_k$ and $\mathbf{t}_k$. Therefore, we can get $F(\mathbf{W}_{k-1}, \mathbf{B}_{k-1}, \mathbf{z}_{k-1}, \mathbf{t}_{k-1}) \geq F(\mathbf{W}_k, \mathbf{B}_k, \mathbf{z}_k, \mathbf{t}_k)$, i.e., in each iteration, the value of the loss function is decreased. As the loss function has a lower bound ($F \geq 0$), the above alternating algorithm will surely be converged.

4 Experiments

4.1 Data Sets

In this section, comparative studies were conducted on twelve commonly used benchmark multi-label data sets. Table 2 summarizes the detailed characteristics of each data set $\mathcal{D}$, with the number of examples (n), the dimension of features (m), the number of class labels (l), label cardinality, i.e., average number of relevant labels per example ($LCard(\mathcal{D})$), label density, i.e. label cardinality over the number of class labels ($LDen(\mathcal{D})$), and number of distinct label sets ($DL(\mathcal{D})$) in $\mathcal{D}$. Those data sets are publicly available at https://mulan.sourceforge.net/datasets-mlc.html

4.2 Compared Methods

We compared HOMI with the following ten state-of-the-art multi-label classification approaches.

- ECC (Ensemble of classifier chains) [7] is an ensemble-based multi-label classification approach, building an ensemble of N classifier chains to solve multi-label classification. [hyper-parameter configuration: N = 5];
- BR [5] is a classical algorithm in multi-label classification, trying to decompose the original multi-label classification task into a set of binary classification tasks. [hyper-parameter C = 1];
- ML-KNN [30] is a popular first-order multi-label learning algorithm based on k-nearest-neighbour classification. [hyper-parameter configuration: k=10];
- WRAP [43] tires to generate label specific features in an embedded feature space to deal with multi-label classification. [hyper-parameter configuration: step_size = 1, λ = 0.1, α = 0.9, $d = \alpha \min(m, l)$];
- MLSF [44] generates label-specific features by analyzing local and global feature-to-label correlations. [hyper-parameter configuration: K = $l/10$, ϵ = 0.01, α = 0.8, γ = 0.01];
- LFLC [45] generates label tailored features by analyzing local and global feature-to-label correlations.

TABLE 2. Characteristics of the experimental data sets.

Data set	n	m	l	$LCard(\mathcal{D})$	$LDen(\mathcal{D})$	$DL(\mathcal{D})$	Domain
mediamill	43097	120	101	4.376	0.043	6555	video
CAL500	502	68	174	26.044	0.150	502	music
emotions	593	72	6	1.869	0.311	27	music
enron	1702	1001	53	3.378	0.064	753	text
bibtex	7395	1836	159	2.402	0.015	2856	text
delicious	16105	500	983	19.020	0.019	15806	text
language log	1460	1004	75	1.180	0.016	286	text
birds	645	260	19	1.014	0.053	133	audio
yeast	2417	103	14	4.237	0.303	198	biology
scene	2407	294	6	1.074	0.179	15	images
corel5k	5000	499	374	3.52	0.009	3175	images
corel16k	13766	500	153	2.859	0.019	4803	images

TABLE 3. The comparison of different methods on all the data sets with respect to Hamming Loss. • / ○ indicates whether HOMI is superior or same / inferior to the compared algorithm.

Hamming loss ↓												
Approach	mediamill	CAL500	emotions	enron	bibtex	delicious	language log	birds	yeast	scene	corel5k	corel16k1
ECC	0.031 ± 0.000 •	0.147 ± 0.002 •	0.226 ± 0.011 •	0.050 ± 0.001 •	0.013 ± 0.000 •	0.019 ± 0.000 •	0.015 ± 0.001 •	0.050 ± 0.004 ○	0.211 ± 0.008 •	0.122 ± 0.003 •	0.010 ± 0.000 •	0.020 ± 0.000 •
BR	0.029 ± 0.000 •	0.165 ± 0.005 •	0.207 ± 0.021 •	0.061 ± 0.002 •	0.016 ± 0.000 •	0.018 ± 0.000 •	0.019 ± 0.001 •	0.070 ± 0.010 •	0.202 ± 0.006 •	0.131 ± 0.005 •	0.011 ± 0.000 •	0.019 ± 0.000 •
ML-KNN	0.029 ± 0.001 •	0.147 ± 0.003 •	0.200 ± 0.007 ○	0.055 ± 0.002 •	0.015 ± 0.000 •	0.019 ± 0.000 •	0.051 ± 0.003 •	0.051 ± 0.006 •	0.199 ± 0.008 ○	0.086 ± 0.005 ○	0.010 ± 0.000 •	0.020 ± 0.000 •
WRAP	0.030 ± 0.000 •	0.136 ± 0.002 •	0.200 ± 0.007 ○	0.047 ± 0.001 ○	0.012 ± 0.000 •	0.019 ± 0.000 •	0.015 ± 0.001 •	0.045 ± 0.048	0.199 ± 0.005 ○	0.110 ± 0.006 ○	0.009 ± 0.000 •	0.019 ± 0.001 •
MLSF	0.030 ± 0.000 •	0.138 ± 0.004 •	0.224 ± 0.021 •	0.051 ± 0.001 •	0.012 ± 0.000 •	0.028 ± 0.000 •	0.015 ± 0.001 •	0.050 ± 0.003 ○	0.210 ± 0.005 •	0.128 ± 0.007 •	0.009 ± 0.000 •	0.019 ± 0.000 •
LFLC	0.030 ± 0.000 •	0.136 ± 0.004 •	0.197 ± 0.004 ○	0.046 ± 0.001 ○	0.012 ± 0.000 •	0.024 ± 0.000 •	0.015 ± 0.001 •	0.045 ± 0.003 ○	0.197 ± 0.004 ○	0.107 ± 0.005 ○	0.009 ± 0.000 •	0.019 ± 0.000 •
BILAS	0.029 ± 0.001 •	0.138 ± 0.000 •	0.233 ± 0.000 •	0.045 ± 0.000 ○	0.011 ± 0.000 •	0.044 ± 0.000 •	0.018 ± 0.002 •	0.052 ± 0.000 •	0.198 ± 0.000 ○	0.097 ± 0.000 ○	0.009 ± 0.000 •	0.018 ± 0.000 •
GLOCAL	0.045 ± 0.000 •	0.143 ± 0.008 •	0.323 ± 0.020 •	0.064 ± 0.001 •	0.014 ± 0.000 •	0.022 ± 0.002 •	0.018 ± 0.000 •	0.097 ± 0.004 •	0.300 ± 0.003 •	0.179 ± 0.000 •	0.009 ± 0.000 •	0.019 ± 0.001 •
ML-LRC	0.042 ± 0.000 •	0.150 ± 0.003 •	0.273 ± 0.006 •	0.050 ± 0.013 •	0.010 ± 0.000 •	0.019 ± 0.000 •	0.072 ± 0.004 •	0.055 ± 0.006 •	0.296 ± 0.003 •	0.138 ± 0.002 •	0.009 ± 0.000 •	0.019 ± 0.000 •
CLML	0.049 ± 0.001 •	0.192 ± 0.006 •	0.315 ± 0.011 •	0.051 ± 0.002 •	0.016 ± 0.000 •	0.027 ± 0.001 •	0.020 ± 0.002 •	0.046 ± 0.005 ○	0.304 ± 0.005 •	0.179 ± 0.003 •	0.015 ± 0.000 •	0.019 ± 0.000 •
HOMI	0.029 ± 0.000	0.135 ± 0.001	0.205 ± 0.023	0.049 ± 0.012	0.009 ± 0.000	0.018 ± 0.000	0.015 ± 0.000	0.051 ± 0.006	0.200 ± 0.005	0.112 ± 0.005	0.009 ± 0.000	0.018 ± 0.000

TABLE 4. The comparison of different methods on all the data sets with respect to Ranking Loss. • / ○ indicates whether HOMI is superior or same / inferior to the compared algorithm.

Ranking loss ↓												
Approach	mediamill	CAL500	emotions	enron	bibtex	delicious	language log	birds	yeast	scene	corel5k	corel16k1
ECC	0.092 ± 0.002 •	0.214 ± 0.007 •	0.183 ± 0.017 •	0.088 ± 0.005 •	0.088 ± 0.002 •	0.141 ± 0.002 ○	0.112 ± 0.011 ○	0.097 ± 0.008 ○	0.202 ± 0.009 •	0.138 ± 0.004 •	0.148 ± 0.003 •	0.180 ± 0.002 •
BR	0.036 ± 0.000 ○	0.180 ± 0.006 •	0.162 ± 0.014 ○	0.085 ± 0.004 •	0.086 ± 0.003 •	0.121 ± 0.001 ○	0.113 ± 0.004 ○	0.104 ± 0.007 ○	0.174 ± 0.011 •	0.115 ± 0.013 •	0.123 ± 0.004 ○	0.162 ± 0.002 •
ML-KNN	0.046 ± 0.000 •	0.220 ± 0.005 •	0.214 ± 0.009 •	0.111 ± 0.003 •	0.252 ± 0.006 •	0.166 ± 0.002	0.156 ± 0.006 •	0.125 ± 0.023 •	0.221 ± 0.007 •	0.105 ± 0.007 ○	0.149 ± 0.004 •	0.196 ± 0.001 •
WRAP	0.048 ± 0.001 •	0.175 ± 0.006 ○	0.157 ± 0.028 ○	0.080 ± 0.006 •	0.102 ± 0.003 •	0.118 ± 0.002 ○	0.157 ± 0.015 •	0.088 ± 0.011 ○	0.170 ± 0.004 ○	0.103 ± 0.006 ○	0.142 ± 0.004 ○	0.145 ± 0.001 ○
MLSF	0.054 ± 0.005 •	0.204 ± 0.004 •	0.181 ± 0.014 •	0.084 ± 0.005 •	0.092 ± 0.015 •	0.127 ± 0.002 ○	0.118 ± 0.019 ○	0.107 ± 0.027 ○	0.201 ± 0.008 •	0.122 ± 0.010 •	0.143 ± 0.004 ○	0.183 ± 0.006 •
LFLC	0.047 ± 0.001 •	0.177 ± 0.004 ○	0.170 ± 0.008 •	0.080 ± 0.007 •	0.098 ± 0.003 •	0.169 ± 0.002 •	0.143 ± 0.013 ○	0.086 ± 0.010 ○	0.171 ± 0.009 •	0.084 ± 0.006 ○	0.162 ± 0.008 •	0.171 ± 0.004 •
BILAS	0.048 ± 0.001 •	0.181 ± 0.000 •	0.217 ± 0.000 •	0.065 ± 0.000 ○	0.077 ± 0.003 •	0.129 ± 0.001 ○	0.166 ± 0.007 •	0.355 ± 0.000 •	0.162 ± 0.000 ○	0.075 ± 0.000 ○	0.201 ± 0.004 •	0.172 ± 0.004 •
GLOCAL	0.055 ± 0.009 •	0.182 ± 0.016 •	0.260 ± 0.009 •	0.039 ± 0.004 ○	0.093 ± 0.010 •	0.243 ± 0.002 •	0.297 ± 0.004 •	0.293 ± 0.021 •	0.362 ± 0.088 •	0.177 ± 0.004 •	0.227 ± 0.010 •	0.189 ± 0.005 •
ML-LRC	0.206 ± 0.002 •	0.485 ± 0.013 •	0.190 ± 0.023 •	0.106 ± 0.024 •	0.257 ± 0.002 •	0.238 ± 0.004 •	0.336 ± 0.015 •	0.145 ± 0.021 •	0.353 ± 0.015 •	0.110 ± 0.011 •	0.314 ± 0.002 •	0.291 ± 0.003 •
CLML	0.449 ± 0.009 •	0.180 ± 0.005 •	0.197 ± 0.014 •	0.082 ± 0.004 •	0.071 ± 0.003 •	0.135 ± 0.002 ○	0.128 ± 0.004 ○	0.099 ± 0.022 ○	0.356 ± 0.012 •	0.111 ± 0.014 •	0.162 ± 0.004 •	0.282 ± 0.005 •
HOMI	0.043 ± 0.001	0.179 ± 0.005	0.169 ± 0.026	0.078 ± 0.015	0.024 ± 0.000	0.170 ± 0.003	0.147 ± 0.002	0.110 ± 0.016	0.171 ± 0.006	0.108 ± 0.012	0.145 ± 0.002	0.155 ± 0.003

TABLE 5. The comparison of different methods on all the data sets with respect to One-error. • / ○ indicates whether HOMI is superior or same / inferior to the compared algorithm.

One-error ↓												
Approach	mediamill	CAL500	emotions	enron	bibtex	delicious	language log	birds	yeast	scene	corel5k	corel16k1
ECC	0.176 ± 0.005 •	0.299 ± 0.042 •	0.312 ± 0.033 •	0.267 ± 0.026 •	0.402 ± 0.015 •	0.467 ± 0.010 ○	0.750 ± 0.016 ○	0.712 ± 0.035 •	0.262 ± 0.023 •	0.338 ± 0.018 •	0.757 ± 0.014 •	0.693 ± 0.003 •
BR	0.162 ± 0.006 ○	0.119 ± 0.045 •	0.268 ± 0.024 ○	0.271 ± 0.015 •	0.399 ± 0.011 •	0.343 ± 0.008 ○	0.750 ± 0.008 ○	0.712 ± 0.045 •	0.233 ± 0.033	0.317 ± 0.030 •	0.667 ± 0.017 ○	0.738 ± 0.010 •
ML-KNN	0.182 ± 0.004 •	0.158 ± 0.014 •	0.372 ± 0.004 •	0.468 ± 0.018 •	0.837 ± 0.006 •	0.573 ± 0.012 •	0.905 ± 0.016 •	0.837 ± 0.029 •	0.254 ± 0.023 •	0.317 ± 0.013 •	0.791 ± 0.009 •	0.805 ± 0.007 •
WRAP	0.156 ± 0.003 ○	0.113 ± 0.033 •	0.265 ± 0.025 ○	0.220 ± 0.021 •	0.363 ± 0.012 •	0.340 ± 0.010 ○	0.730 ± 0.032 ○	0.660 ± 0.031 ○	0.223 ± 0.008 ○	0.272 ± 0.008 ○	0.621 ± 0.032 ○	0.635 ± 0.005 ○
MLSF	0.193 ± 0.008 •	0.128 ± 0.040 •	0.319 ± 0.021 •	0.283 ± 0.028 •	0.402 ± 0.013 •	0.358 ± 0.009 ○	0.755 ± 0.012 ○	0.701 ± 0.055 •	0.261 ± 0.022 •	0.345 ± 0.010 •	0.667 ± 0.010 ○	0.743 ± 0.019 •
LFLC	0.158 ± 0.004 ○	0.120 ± 0.022 •	0.227 ± 0.031 ○	0.231 ± 0.009 •	0.351 ± 0.007 •	0.343 ± 0.008 ○	0.721 ± 0.025 ○	0.656 ± 0.037 ○	0.225 ± 0.025 ○	0.250 ± 0.023 ○	0.631 ± 0.018 ○	0.645 ± 0.010 ○
BILAS	0.166 ± 0.003 •	0.115 ± 0.000 •	0.357 ± 0.000 •	0.233 ± 0.000 •	0.361 ± 0.006 •	0.543 ± 0.013 ○	0.877 ± 0.015 •	0.868 ± 0.000 •	0.223 ± 0.000 ○	0.229 ± 0.000 ○	0.776 ± 0.000 •	0.724 ± 0.003 •
GLOCAL	0.171 ± 0.005 •	0.118 ± 0.001 •	0.432 ± 0.000 •	0.077 ± 0.010 ○	0.433 ± 0.031 •	0.672 ± 0.003 •	0.881 ± 0.004 •	0.684 ± 0.008 ○	0.319 ± 0.040 •	0.479 ± 0.021 •	0.670 ± 0.220 ○	0.644 ± 0.070 ○
ML-LRC	0.322 ± 0.008 •	0.603 ± 0.048 •	0.291 ± 0.036 •	0.141 ± 0.034 ○	0.141 ± 0.007 •	0.403 ± 0.018 ○	0.879 ± 0.017 •	0.719 ± 0.031 •	0.368 ± 0.009 •	0.285 ± 0.022 •	0.683 ± 0.015 ○	0.674 ± 0.012 •
CLML	0.305 ± 0.011 •	0.115 ± 0.026 •	0.286 ± 0.007 •	0.220 ± 0.014 •	0.348 ± 0.006 •	0.386 ± 0.005 ○	0.697 ± 0.015 ○	0.721 ± 0.023 •	0.382 ± 0.013 •	0.294 ± 0.020 •	0.635 ± 0.005 ○	0.695 ± 0.031 •
HOMI	0.165 ± 0.001	0.099 ± 0.035	0.276 ± 0.052	0.183 ± 0.035	0.132 ± 0.000	0.571 ± 0.002	0.873 ± 0.003	0.699 ± 0.022	0.242 ± 0.016	0.284 ± 0.026	0.691 ± 0.038	0.651 ± 0.009

TABLE 6. The comparison of different methods on all the data sets with respect to Marco-averaging AUC. • / ○ indicates whether HOMI is superior or same / inferior to the compared algorithm.

Macro-averaging AUC ↑												
Approach	mediamill	CAL500	emotions	enron	bibtex	delicious	language log	birds	yeast	scene	corel5k	corel16k1
ECC	0.776 ± 0.004 •	0.497 ± 0.013 •	0.815 ± 0.013 ○	0.650 ± 0.019 •	0.870 ± 0.004 •	0.702 ± 0.002 •	0.544 ± 0.052 •	0.732 ± 0.050 •	0.650 ± 0.017 •	0.878 ± 0.004 •	0.531 ± 0.018 •	0.658 ± 0.005 •
BR	0.839 ± 0.004 •	0.502 ± 0.007 •	0.826 ± 0.023 ○	0.619 ± 0.028 •	0.874 ± 0.003 •	0.738 ± 0.003 •	0.553 ± 0.020 •	0.718 ± 0.051 •	0.629 ± 0.013 •	0.886 ± 0.011 •	0.524 ± 0.015 •	0.673 ± 0.003 •
ML-KNN	0.933 ± 0.000 •	0.691 ± 0.020 •	0.802 ± 0.010 ○	0.836 ± 0.032 •	0.938 ± 0.002 •	0.918 ± 0.002 •	0.703 ± 0.027 •	0.847 ± 0.004 •	0.709 ± 0.008 •	0.938 ± 0.007 ○	0.790 ± 0.023 •	0.922 ± 0.001 •
WRAP	0.851 ± 0.003 •	0.555 ± 0.011 •	0.835 ± 0.029 ○	0.635 ± 0.014 •	0.870 ± 0.004 •	0.755 ± 0.004 •	0.540 ± 0.040 •	0.815 ± 0.041 •	0.688 ± 0.008 •	0.906 ± 0.005 ○	0.581 ± 0.010 •	0.744 ± 0.008 •
MLSF	0.842 ± 0.003 •	0.528 ± 0.012 •	0.816 ± 0.020 ○	0.503 ± 0.024 •	0.873 ± 0.006 •	0.730 ± 0.004 •	0.558 ± 0.036 •	0.660 ± 0.038 •	0.631 ± 0.008 •	0.892 ± 0.010 ○	0.520 ± 0.004 •	0.669 ± 0.007 •
LFLC	0.832 ± 0.007 •	0.553 ± 0.010 •	0.834 ± 0.013 ○	0.686 ± 0.028 •	0.875 ± 0.005 •	0.757 ± 0.003 •	0.557 ± 0.004 •	0.835 ± 0.040 •	0.684 ± 0.009 •	0.924 ± 0.005 ○	0.568 ± 0.012 •	0.726 ± 0.004 •
BILAS	0.889 ± 0.006 •	0.533 ± 0.000 •	0.758 ± 0.000 •	0.844 ± 0.000 •	0.920 ± 0.004 •	0.796 ± 0.005 •	0.559 ± 0.006 •	0.495 ± 0.000 •	0.650 ± 0.000 •	0.920 ± 0.000 ○	0.655 ± 0.037 •	0.686 ± 0.024 •
GLOCAL	0.605 ± 0.009 •	0.509 ± 0.030 •	0.750 ± 0.032 •	0.787 ± 0.022 •	0.825 ± 0.021 •	0.637 ± 0.004 •	0.613 ± 0.023 •	0.513 ± 0.027 •	0.732 ± 0.040 •	0.826 ± 0.004 •	0.518 ± 0.043 •	0.581 ± 0.032 •
ML-LRC	0.957 ± 0.003 •	0.849 ± 0.003 •	0.726 ± 0.006 •	0.949 ± 0.013 •	0.989 ± 0.000 •	0.980 ± 0.001 •	0.927 ± 0.004 •	0.944 ± 0.006 •	0.704 ± 0.000 •	0.861 ± 0.002 •	0.990 ± 0.000 •	0.981 ± 0.000 •
CLML	0.950 ± 0.008 •	0.534 ± 0.010 •	0.684 ± 0.019 •	0.664 ± 0.031 •	0.990 ± 0.000 •	0.980 ± 0.000 •	0.905 ± 0.041 •	0.944 ± 0.008 •	0.695 ± 0.004 •	0.820 ± 0.003 •	0.892 ± 0.014 •	0.981 ± 0.000 •
HOMI	0.969 ± 0.000	0.865 ± 0.006	0.795 ± 0.023	0.950 ± 0.012	0.990 ± 0.000	0.981 ± 0.002	0.985 ± 0.001	0.948 ± 0.006	0.800 ± 0.004	0.887 ± 0.005	0.991 ± 0.000	0.981 ± 0.000

[hyper-parameter configuration: grid search for $\lambda \in \{1, 3, ..., 19\}$ with step-size 2, $\eta \in \{1e-10, ..., 1e-5\}$

with a multiple of e at each step, $\beta = 10^4$];
- BILAS [46] generates a group of tailored features for a pair of class labels with heuristic prototype selection and embedding. [hyper-parameter configuration: t=0.1, ratio=0.5];
- GLOCAL [10] uses low-rank factorization to dig the global and local label correlations at the same time, through learning a latent label representation and optimizing label manifolds. [hyper-parameter configuration: $\lambda = 1$];
- ML-LRC [33] is a low-rank approach applying low-rank constraints to the label matrix to mine the local correlations of class labels. [hyper-parameter configuration: grid search for $\alpha, \beta \in \{2^{-10}, 2^{-9}, 2^{-8}, ..., 2^{10}\}$, $\gamma = 0.1$ and $\tau = 0.5$];
- CLML [47] is a approach that learns common and label-specific features based on the correlation information from labels and instances. [hyper-parameter configuration: grid search for $\alpha, \beta, \lambda_1, \lambda_2 \in \{2^{-10}, 2^{-9}, 2^{-8}, ..., 2^{10}\}$ with step 2^2].

In brief, BR [5] and ML-KNN [30] belong to first-order approaches, BILAS [46] is a second-order approach, and WRAP [43], ECC [7], GLOCAL [10] and CLML [47] are high-order approaches. *Particularly, GLOCAL and ML-LRC [33] are two low-rank based approaches.* MLSF [44] and LFLC [45] are two label-specific approaches. The hyper-parameter configurations for different methods are suggested by their original papers.

4.3 Experimental Settings

The hyper parameters of HOMI are set as follows: $\beta = 2$, $\gamma = 1$, $\lambda = 1$, $iter = 100$ and $s = 10^1$. Following [43], five-fold cross-validation is performed on each data set, with mean metric and standard deviation recorded.

4.4 Evaluation Metrics

Let C_i^+, C_i^- be the sets of positive and negative labels corresponding to the ith instance, and T_i^+, T_i^- be the sets of positive and negative instances corresponding to the ith label, p is the number of the test instances. We chose the following four popular metrics to evaluate the performance of the proposed method and the methods under comparison.

- Hamming loss (Hloss) evaluates the rate of the mistook labels. Hloss = $\frac{1}{p}\sum_i f(\mathbf{x}_i)\Delta\mathbf{y_i}$, Δ stands for the symmetric difference between two sets, i.e., $a\Delta b = 1$ (resp. 0) if $a = b$ (resp. a $\neq$ b).
- Ranking loss (Rloss) calculates the fraction that a negative label is ranked higher than a positive label. Specifically, for instance i, suppose $M_i = \{(j^{'}, j^{''})|\mathbf{g}_{j'}(\mathbf{x}_i) \leq \mathbf{g}_{j''}(\mathbf{x}_i), (j^{'}, j^{''}) \in C_i^+ \times C_i^-\}$, Rloss = $\frac{1}{p}\sum_{i=1}^{p}\frac{|M_i|}{|C_i^+||C_i^-|}$.
- One-error evaluates the fraction of examples whose top-ranked label is not in the relevant label set. One-error = $\frac{1}{p}\sum_{i=1}^{p}[c_{argmax_j f_j(\mathbf{x}_i)} \notin \mathbf{y_i}]$.
- Average Area Under the ROC Curve (Macro-averaging AUC) denotes the fraction that a positive instance is ranked higher than a negative instance averaged over all labels. Suppose $N_i = \{(i^{'}, i^{''})|\mathbf{g}_j(x_{i'}) \geq \mathbf{g}_j(x_{i''}), (x_{i'}, x_{j''} \in T_j^+ \times T_j^-)\}$, Macro-averaging AUC = $\frac{1}{l}\sum_{j=1}^{l}\frac{|N_i|}{|T_i^+||T_i^-|}$.

For hamming loss, ranking loss and one-error, the lower the better, while for macro-averaging AUC, the higher the better. All the metrics lie in the range of $[0, 1]$.

4.5 Experimental Analysis

Tables 3-6 show the experimental results of the proposed method and the compared baselines on twelve data sets with respect to four metrics. Additionally, the widely-used *Friedman test* [48] is used for statistical analysis of the performance among all the methods on the benchmark data sets. Suppose k denotes the number of comparing algorithms, N denotes the number of data sets and r_i^j denotes the rank of the jth approach on the ith data set. Suppose $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$ denotes the average rank of the jth method on all the data sets. The Friedman statistic F_F, which is distributed according to the F-distribution with $(k-1)$ numerator degrees of freedom and $(k-1)(N-1)$ denominator degrees of freedom, is defined as

$$F_F = \frac{(N-1)\mathcal{X}_F^2}{N(k-1) - \mathcal{X}_F^2},$$

where

$$\mathcal{X}_F^2 = \frac{12N}{k(k+1)}(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}).$$

Table 7 reports the detailed statistics over all evaluation metrics as well as the related critical value at 0.05 significance level for HOMI ($k = 11, N = 12$). We can observe that the F_F value is larger than the critical value w.r.t. all evaluation metrics. Therefore, the null hypothesis of equal performance among comparing approaches is clearly rejected.

Evaluation Metric	F_F	Critical Value
Hamming Loss	6.0589	1.9178
Ranking Loss	6.8014	
One-error	7.2527	
Macro-averaging AUC	10.1662	

TABLE 7. Friedman test statistics over each evaluation metrics and the critical value at 0.05 significance level ($k = 11, N = 12$).

In order to verify whether HOMI significantly outperforms other algorithms, we employ *Holm's procedure* [48] as the post-hoc test by treating HOMI as the control approach. Without loss of generality, we take HOMI as the first comparing approach $\mathcal{A}_1$, and for the other $k-1$ approaches, we let $\mathcal{A}_j$ $(2 \leq j \leq k)$ denote the one with the $(j-1)$th largest average rank. Then, the test statistic for comparing $\mathcal{A}_1$ and $\mathcal{A}_j$ is defined as follows:

$$z_j = (R_1 - R_j)/\sqrt{\frac{k(k+1)}{6N}} \quad (2 \leq j \leq k),$$

Accordingly, let p_j denote the p-value of z_j under normal distribution, and the Holm's procedure sequentially checks whether p_j is below $\alpha/(k-j+1)$ in ascending order of

1. For emotions and bibtex, s is set to be 2.

j at significance level α. Specifically, the Holm's procedure is supposed to terminate at j^* where j^* is the first j that satisfying $p_j \geq \alpha/(k-j+1)$[2]. Then HOMI is deemed to perform significantly different compared with $\mathcal{A}_j$ where $j \in \{2, ..., j^*-1\}$.

Table 8 reports the statistics by taking Holm's procedure as post-hoc test at 0.05 significance level, where HOMI is treated as the control approach. We can have the following observations based on the experimental results:

TABLE 8. Comparison of HOMI against other comparing approaches with *Holm's procedure* w.r.t. all evaluation metrics at significance level $\alpha = 0.05$

Hamming Loss				
j	approach	z_j	p_j	$\alpha/k-j+1$
2	CLML	-4.801	1.582e-6	0.005
3	GLOCAL	-4.308	1.645e-4	0.006
4	BR	-3.261	1.106e-3	0.006
5	ML-LRC	-3.200	1.372e-3	0.007
6	ECC	-3.077	2.088e-3	0.008
7	ML-KNN	-2.646	8.133e-3	0.010
8	MLSF	-2.646	8.133e-3	0.013
9	WRAP	-1.600	1.095e-1	0.017
10	BILAS	-1.538	1.238e-1	0.025
11	LFLC	-0.677	4.984e-1	0.050
Ranking Loss				
j	approach	z_j	p_j	$\alpha/k-j+1$
2	ML-LRC	-4.185	2.850e-5	0.005
3	GLOCAL	-3.508	4.513e-4	0.006
4	ML-KNN	-2.646	8.133e-3	0.006
5	CLML	-1.846	6.483e-2	0.007
6	ECC	-1.784	7.428e-1	0.008
7	MLSF	-1.354	1.757e-1	0.010
8	BILAS	-1.107	2.679e-1	0.013
9	LFLC	-0.430	6.665e-1	0.017
10	BR	0.123	1.000	0.025
11	WRAP	0.492	1.000	0.050
One-error				
j	approach	z_j	p_j	$\alpha/k-j+1$
2	ML-KNN	-3.754	1.738e-4	0.005
3	ECC	-2.277	2.277e-2	0.006
4	MLSF	-2.154	3.123e-2	0.006
5	GLOCAL	-1.907	5.641e-2	0.007
6	ML-LRC	-1.723	8.483e-1	0.008
7	BILAS	-1.354	1.757e-1	0.010
8	CLML	-0.861	3.889e-1	0.013
9	BR	-0.800	4.237e-1	0.017
10	LFLC	1.107	1.000	0.025
11	WRAP	1.538	1.000	0.050
Macro-averaging AUC				
j	approach	z_j	p_j	$\alpha/k-j+1$
2	GLOCAL	-5.046	4.494e-7	0.005
3	ECC	-5.046	4.494e-7	0.006
4	BR	-4.615	3.913e-6	0.006
5	MLSF	-4.554	5.254e-6	0.007
6	WRAP	-3.200	1.372e-3	0.008
7	BILAS	-3.077	2.089e-3	0.010
8	LFLC	-2.831	4.638e-3	0.013
9	CLML	-2.154	3.123e-2	0.017
10	ML-KNN	-1.354	1.757e-1	0.025
11	ML-LRC	-1.292	1.962e-1	0.050

- HOMI performs better than all the first-order and second-order methods. For example, HOMI performs significantly better than BR according to Table 8. It performs nearly 2 times better than BR on bibtex in average with respect to hamming loss, 1.6 times better than BILAS with respect to macro-averaging AUC on CAL500, and more than 6 times better than ML-KNN on bibtex with respect to one-error, which validates the importance of involving high-order information.
- HOMI also significantly outperforms all the high-order methods including the low-rank based approaches. Specifically, the improvements of HOMI over GLOCAL and ML-LRC are significant according to Table 8, which are two state-of-the-art multi-label classification methods based on low-rank factorization. For example, HOMI performs nearly 2 times better than GLOCAL on corel16k001 regarding macro-averaging AUC and nearly 4 times better w.r.t. ranking loss on bibtex. It outperforms ML-LRC more than 2 times on CAL500 w.r.t. ranking loss. *This observation verifies the rationality of our basic assumption that the label matrix of multi-label classification should be high-rank rather than low-rank.*
- LFLC and BILAS perform well on scene, while they are not apt to deal with enron and bibtex. Besides, LFLC performs well on emotions, but does not on mediamill. However, HOMI is adept in almost all kinds of data set especially on text and image data sets, which validates the robustness of HOMI to different types of data sets.
- HOMI is also robust to different evaluation metrics compared with the other approaches. HOMI achieves especially outstanding performance on macro-averaging AUC, for it enables positive labels to rank higher than negative ones effectively.
- The performance of WRAP and LFLC seems comparable to HOME with respect to the Hamming Loss and the One-error, while HOMI is obvious superior to the other methods with respect to the Macro-averaging AUC. The reason is that HOMI uses self-representation to exploit the high-order label correlations while keeping the label matrix full-rank, leading to more effective label-wise discrimination and aggregation. Therefore, HOMI outperforms the other two methods significantly w.r.t. Macro-averaging AUC.
- In general, HOMI performs superior or at least comparable to the other algorithms in 85%, 71.7%, 65.8%, 90.8% cases in terms of hamming loss, ranking loss, one-error and macro-averaging AUC which validates that HOMI is a promising approach in multi-label classification.

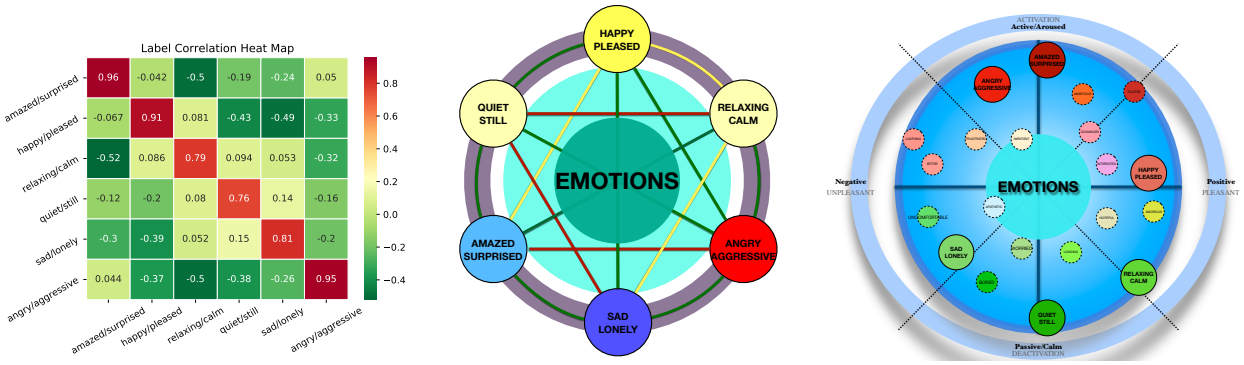
The success of HOMI is partially credited to the information of high-order label correlations exploited from label space, and partially credited to the incorporation of the local geometric structure of instances to achieve joint learning of high-order label correlations and model prediction.

4.6 Further Analysis

4.6.1 High-order correlations exploited by HOMI

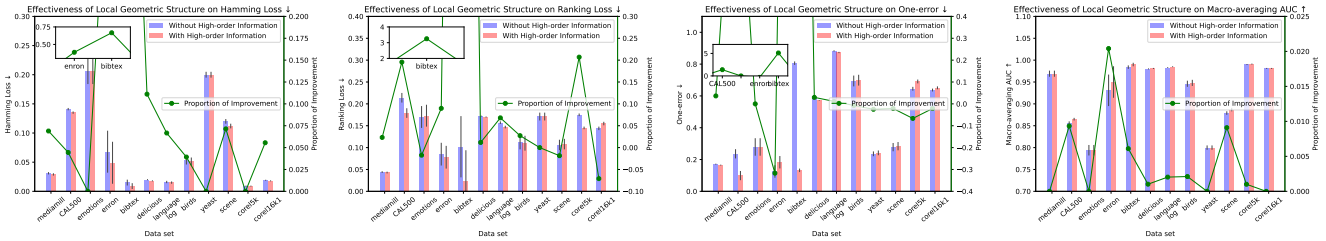
HOMI exploits high-order information during the training process. In order to know what information HOMI has learned, we recorded the high-order correlation matrix $\mathbf{B}$ trained on emotions. Emotions is a multi-label data set that

2. If $p_j < \alpha/(k-j+1)$ holds for all j, j^* is set to be $k+1$.



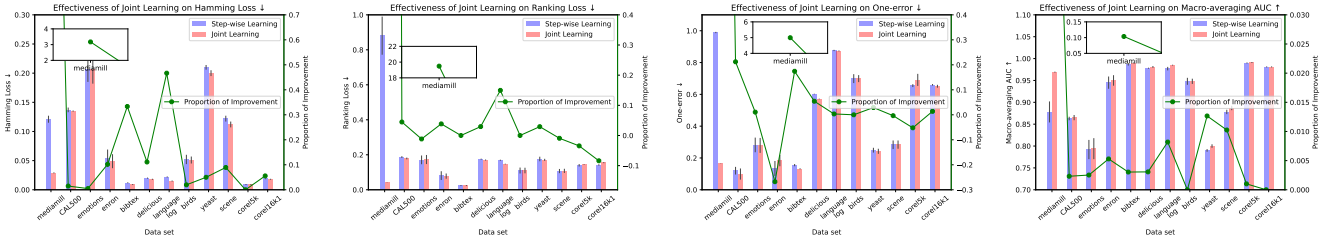
(a) (b) (c)

Fig. 1: (a) The normalized matrix $\mathbf{B}$ learned by HOMI on emotions data set. Redder color block represents larger value, while greener color block represents smaller value. (b) Label correlations based on matrix $\mathbf{B}$. The green line between two labels indicates that the two labels are negatively correlated, while red one indicates those are positively related, and yellow one means that the correlation is uncertain or very weak. (c) Russell's emotion circumplex [49]. The large solid circles represent the labels in data set emotions, and the small dotted circles represent other representative emotions do not exist in the emotions data set.



(a) Hamming Loss (b) Ranking Loss (c) One-error (d) Macro-averaging AUC

Fig. 2: Usefulness of high-order information. The bar in each sub-figure stands for the value of each metric on different data sets based on the left y-axis. The black vertical line on the top of the bars denotes the standard deviation on each metric. The green line in each sub-figure stands for the proportion of improvement based on the right y-axis.



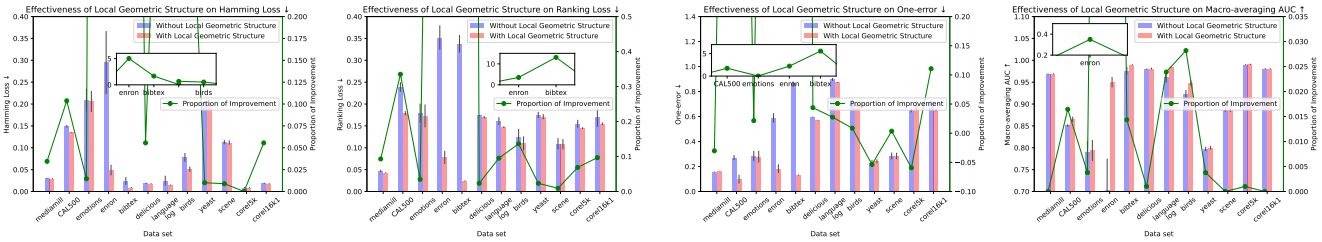
(a) Hamming Loss (b) Ranking Loss (c) One-error (d) Macro-averaging AUC

Fig. 3: Effectiveness of joint learning. The bar in each sub-figure stands for the value of each metric on different data sets based on the left y-axis. The black vertical line on the top of the bars denotes the standard deviation on each metric. The green line in each sub-figure stands for the proportion of improvement based on the right y-axis.

describes the emotions of different music. It contains six common emotions, such as happy, relaxing and angry.

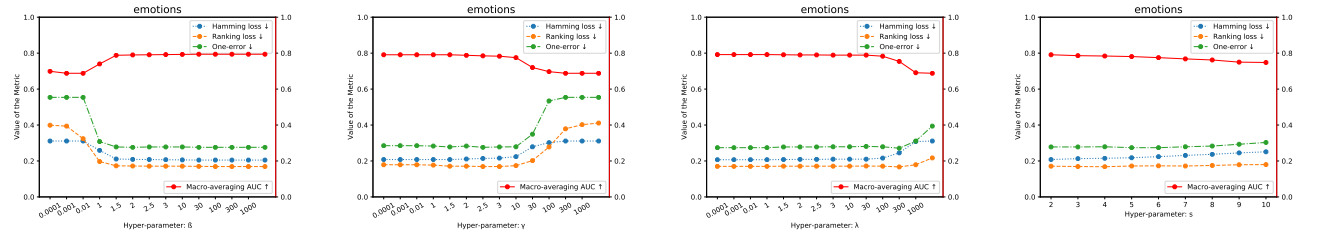
Fig. 1(a) shows the matrix $\mathbf{B}$ that HOMI learns. The value of each column represents the contribution of the corresponding row label to the column label. In order to better display the learned high-order label correlations, the matrix is normalized, and the relationships among labels extracted from $\mathbf{B}$ is visually represented in Fig. 1(b). It is obvious that the diagonal element of $\mathbf{B}$ is large, which means in prediction, a label should be dominated by itself. Moreover, it should also be influenced by correlated labels. We can clearly observe that label "relaxing/calm" and "quiet/still" are positively related, while "amazed/surprised" and "sad/lonely" are negatively correlated, and "happy/pleased" and "sad/lonely" are negatively correlated, too. Additionally, label "angry/aggressive" has negative contribution to all the other labels except "amazed/surprised". Those label correlations learned by HOMI are consistent with common sense.

We can further verify the correctness and validity of the excavated higher-order information based on the Russell's emotion circumplex theory [49]. Russell thought that



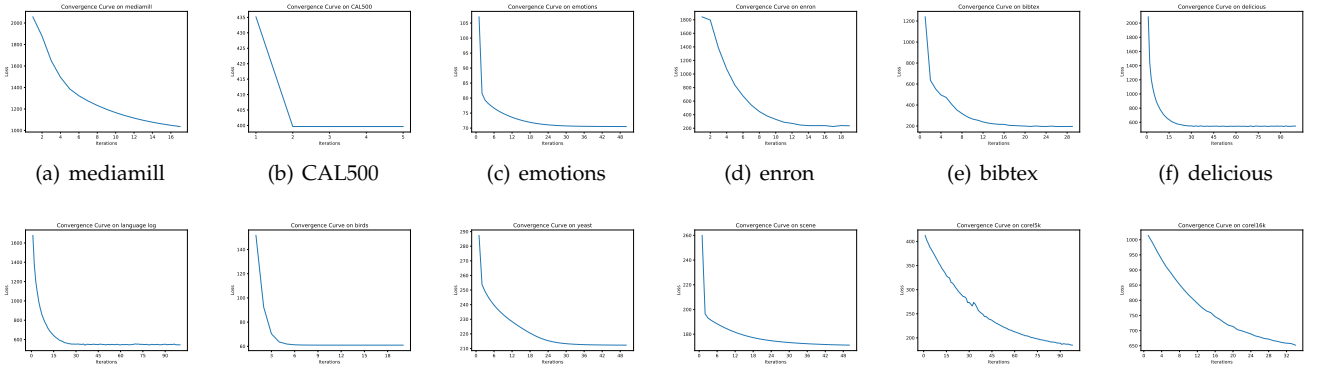
(a) Hamming Loss (b) Ranking Loss (c) One-error (d) Macro-averaging AUC

Fig. 4: Usefulness of local geometric structure. The bar in each sub-figure stands for the value of each metric on different data sets based on the left y-axis. The black vertical line on the top of the bars denotes the standard deviation on each metric. The green line in each sub-figure stands for the proportion of improvement based on the right y-axis.



(a) β (b) γ (c) λ (d) s

Fig. 5: Sensitivity analysis. The value of macro-averaging AUC is based on the right y-axis, while the other three metrics are based on the left one.



(a) mediamill (b) CAL500 (c) emotions (d) enron (e) bibtex (f) delicious

(g) language log (h) birds (i) yeast (j) scene (k) corel5k (l) corel16k

Fig. 6: Convergence Curve.

emotions can be measured in two dimensions, i.e., pleasure–displeasure and degree-of-arousal. We show a simplified Russell's emotion circumplex in Fig. 1(c). We assume that the correlations between two emotions can be determined by their cosine similarity, i.e, if the similarity is larger than 0, the two emotions are positively connected and a larger cosine similarity means more correlation; ; otherwise, they are negatively correlated. From Fig. 1(c), it can be observed that "happy/pleased" and "sad/lonely"are negatively related, while "angry" and "relaxing/calm" are negatively correlated, which is also learned by HOMI.

We can conclude that the correlations exploited by HOMI are reasonable as they are confirmed by both common sense and the Russell's emotion circumplex theory.

4.6.2 *Usefulness of High-order Information*

In order to validate the effectiveness of using high-order information, we compared HOMI with its degenerated version that makes predictions without considering the high-order label correlations, i.e., $\mathbf{g}(\mathbf{x}) = \mathbf{x}\mathbf{W} + \mathbf{1}_n\mathbf{z}^\top$ on all the twelve data sets with respect to hamming loss, ranking loss, one-error and macro-averaging AUC. The detailed comparisons are shown in Fig. 2.

From Fig. 2 we can observe that the performance of HOMI with high-order information outperforms its degenerated version in 83.3%, with total 48 cases (12 data sets $\times$ 4 evaluation metrics). Especially on the data set bibtex with respect to the one-error, HOMI performs 5 times better than its degenerated version. As the matrix $\mathbf{B}$ captures the high-order correlations among labels, it is effective in improving

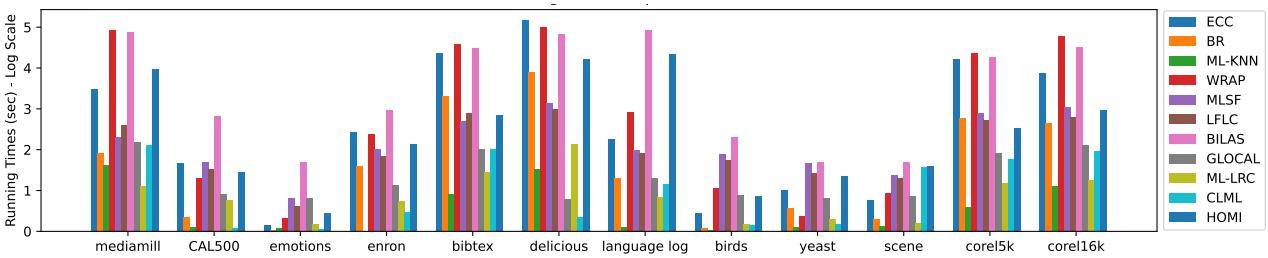


Fig. 7: Running time comparison.

prediction accuracy. We thus can conclude that considering high-order label information when making predictions enable HOMI to perform better and more stable in general.

4.6.3 Effectiveness of Joint Learning

HOMI exploits the high-order label correlations and make predictions in a joint manner. To show the effectiveness of joint learning, we also designed a compared method with a step-wise learning fashion that first learns high-order information by minimizing $\|\mathbf{Y}-\mathbf{YB}-\mathbf{1}_n\mathbf{t}^\mathsf{T}\|_F^2$ and then makes predictions by $(\mathbf{xW}+\mathbf{1}_n\mathbf{z}^\mathsf{T})\mathbf{B}+\mathbf{1}_n\mathbf{t}^\mathsf{T}$.

The comparison between joint learning and step-wise learning is shown in Fig. 3, where we can observe that the joint learning manner outperforms the step-wise learning significantly. Specifically, the performance of joint learning model is relatively superior to step-wise learning in 85.4%, with total 48 cases (12 data sets $\times$ 4 evaluation metrics). Especially on the data set mediamill, the Ranking Loss of joint learning manner is 20 times lower than that of the step-wise one.

4.6.4 Usefulness of Local Geometric Structure

Aiming to validate the effectiveness of local geometric structure, we evaluated the performance of HOMI without local geometric structure (i.e., $\gamma = 0$) in Fig. 4, where we can find that the local geometric structure is important to the proposed approach. For example, HOMI performs 13 times better than its simplified version on bibtex with respect to ranking loss. Generally, the one with local geometric structure significantly outperforms the one without that in 95.9% cases.

4.6.5 Sensitivity Analysis

In Fig. 5, we investigate the sensitivity of HOMI with respect to β, γ, λ and the number of nearest instances (s) on emotions. It is evident that the performance of HOMI is relatively stable and excellent as the value of the parameters change within a reasonable wide range, validating the robustness of HOMI to the hyper-parameters, which serves as a desirable property in practice.

4.6.6 Convergence Analysis

Fig. 6 shows the convergence property of the proposed approach on twelve data sets where we can observe that the objective function decreases significantly and converges in about 50 iterations on nearly all the data sets.

4.6.7 Running time

We also compare the running time of each algorithm on each data set, which is shown in Fig. 7. It is obvious that HOMI is usually faster than BILAS, WRAP and ECC. Additionally, the running speed of HOMI is also comparable to other methods, which is also a promising property in practice.

5 Conclusion

In this paper, we have presented an effective multi-label classification approach. Different from the traditional low-rank based methods, we argue that the label matrix of multi-label classification is full-rank or approximately full-rank, and thus we propose to keep the rank of label matrix unchanged by self-representation. Moreover, by incorporating the local geometric structure of the input, the proposed method can simultaneously make predictions and exploit high-order label correlations. Extensive experiments validate the effectiveness of the proposed approach in learning high-order label correlations, and in incorporating the local geometric structure. As the proposed model can explicitly indicate the high-order label correlations, we have verified HOMI can learn meaningful label correlations. Besides, HOMI also significantly outperforms the state-of-the-art multi-label classification approaches, serving as a promising solution for multi-label classification. In the future, it is interesting to investigate how to extend HOMI to a non-linear version.

Appendix A
The Matrix $\mathbf{B}$ HOMI Has Learned

In section 4.6.1, we have recorded the normalized matrix $\mathbf{B}$ trained by HOMI on emotions. The original value of the matrix $\mathbf{B}$ is shown in Fig. 8.

References

[1] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine learning*, vol. 88, no. 1-2, pp. 157–208, 2012.

[2] F. Sun, J. Tang, H. Li, G.-J. Qi, and T. S. Huang, "Multi-label image categorization with sparse factor representation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1028–1037, 2014.

[3] X. Wang and G. Sukthankar, "Multi-label relational neighbor classification using social context features," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 464–472.

[4] R. Wang, S. Kwong, X. Wang, and Y. Jia, "Active k-labelsets ensemble for multi-label classification," *Pattern Recognition*, vol. 109, p. 107583, 2021.

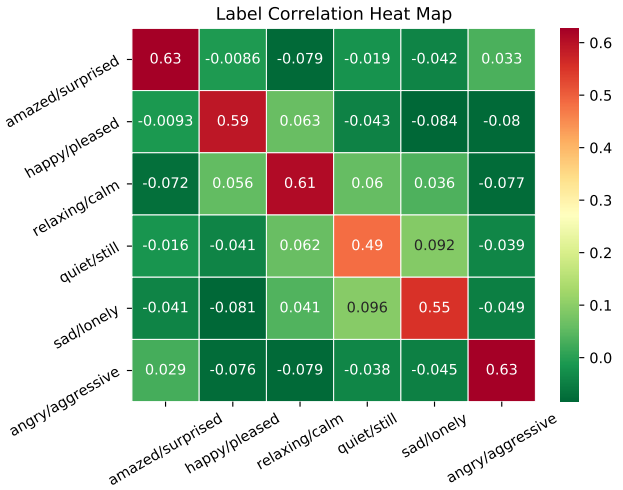


Fig. 8: Matrix $\mathbf{B}$ learned by HOMI on the emotion data set.

[5] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[6] M.-L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.

[7] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, 2009.

[8] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[9] K. Punera, S. Rajan, and J. Ghosh, "Automatically learning document taxonomies for hierarchical classification," in *Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005, pp. 1010–1011.

[10] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1081–1094, 2018.

[11] L. Feng, B. An, and S. He, "Collaboration based multi-label learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3550–3557.

[12] Y. Jia, H. Liu, J. Hou, S. Kwong, and Q. Zhang, "Multi-view spectral clustering tailored tensor low-rank representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4784–4797, 2021.

[13] J. K. Valadi, P. T. Ovhal, and K. J. Rathore, "A simple method of solution for multi-label feature selection," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1–4.

[14] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in neural information processing systems*, 2006, pp. 1609–1616.

[15] L. Yong, T. Liu, D. Tao, and X. Chao, "Multi-view matrix completion for multi-label image classification," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2355–2368, 2015.

[16] M. Wang, X. Zhou, and T. S. Chua, "Automatic image annotation via local multi-label classification," in *ACM*, 2008, p. 17.

[17] C. Wang, S. Yan, Z. Lei, and H. J. Zhang, "Multi-label sparse coding for automatic image annotation," in *IEEE Computer Society Conference on Computer Vision Pattern Recognition*, 2009.

[18] W. Fei, Y. Han, T. Qi, and Y. Zhuang, "Multi-label boosting for image annotation by structural grouping sparsity," in *Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010*, 2010.

[19] S. Feng and D. Xu, "Transductive multi-instance multi-label learning algorithm with application to automatic image annotation," *Expert Systems with Applications*, vol. 37, no. 1, pp. 661–670, 2010.

[20] X. Tao, Y. Li, R. Lau, and W. Hua, "Unsupervised multi-label text classification using a world knowledge ontology," in *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, 2012.

[21] K. M. Ozonat and D. Young, "Towards a universal marketplace over the web: Statistical multi-label classification of service provider forms with simulated annealing." *ACM*, 2009.

[22] B. Parlak and A. K. Uysal, "Classification of medical documents according to diseases," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, 2015, pp. 1635–1638.

[23] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 117–126.

[24] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive stacking for audio tag annotation and retrieval," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2308–2311.

[25] F. Pachet and P. Roy, "Improving multilabel analysis of music titles: A large-scale validation of the correction approach," *IEEE Transactions on Audio Speech & Language Processing*, vol. 17, no. 2, pp. 335–343, 2009.

[26] A. Wieczorkowska, P. Synak, and Z. W. Ra, "Multi-label classification of emotions in music," *Intelligent Information Processing and Web Mining*, 2006.

[27] Y. Yang and S. Gopal, "Multilabel classification with meta-level features in a learning-to-rank framework," *Machine Learning*, vol. 88, no. 1-2, pp. 47–68, 2012.

[28] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems With Applications*, vol. 57, no. Sep., pp. 1–11, 2016.

[29] M. L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," *ACM*, 2010.

[30] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[31] J. Fürnkranz, E. Hüllermeier, E. Loza?Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.

[32] Y. Jia, H. Liu, J. Hou, and S. Kwong, "Pairwise constraint propagation with dual adversarial manifold regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5575–5587, 2020.

[33] X. Wang, J. Xie, L. Yu, and X. Tao, "Ml-lrc: Low-rank-constraint-based multi-label learning with label noise," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 2020, pp. 129–136.

[34] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen, "Learning low-rank label correlations for multi-label classification with missing labels," in *2014 IEEE international conference on data mining*. IEEE, 2014, pp. 1067–1072.

[35] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale multi-label learning with missing labels," in *International conference on machine learning*. PMLR, 2014, pp. 593–601.

[36] K. H. Huang and H. T. Lin, "Cost-sensitive label embedding for multi-label classification," *Machine Learning*, 2017.

[37] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent space for multi-label classification," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[38] M. Cissé, M. Al-Shedivat, and S. Bengio, "Adios: Architectures deep in output space," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2770–2779.

[39] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[40] J. Nam, E. Loza Mencía, H. J. Kim, and J. Fürnkranz, "Maximizing subset accuracy with recurrent neural networks in multi-label classification," *Advances in neural information processing systems*, vol. 30, 2017.

[41] E. Elhamifar, "High-rank matrix completion and clustering under self-expressive models," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[42] G. Golub, S. Nash, and C. Van Loan, "A hessenberg-schur method for the problem ax + xb= c," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 909–913, 1979.

[43] Z.-B. Yu and M.-L. Zhang, "Multi-label classification with label-specific feature generation: A wrapped approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[44] L. Sun, M. Kudo, and K. Kimura, "Multi-label classification with meta-label-specific features," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 1612–1617.
[45] J. Ma, H. Zhang, and T. W. S. Chow, "Multilabel classification with label-specific features and classifiers: A coarse- and fine-tuned framework," *IEEE Transactions on Cybernetics*, vol. 51, no. 2, pp. 1028–1042, 2021.
[46] M.-L. Zhang, J.-P. Fang, and Y.-B. Wang, "Bilabel-specific features for multi-label classification," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 1, pp. 1–23, 2021.
[47] J. Li, P. Li, X. Hu, and K. Yu, "Learning common and label-specific features for multi-label classification with correlation information," *Pattern Recognition*, vol. 121, p. 108259, 2022.
[48] J. Demiar and D. Schuurmans, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.
[49] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.