Information Sciences 594 (2022) 200–216

Contents lists available at ScienceDirect

Information Sciences

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

Weak multi-label learning with missing labels via instance granular discrimination

Anhui Tan [a,b], Xiaowan Ji [b,c], Jiye Liang [b,*], Yuzhi Tao [a,c], Wei-Zhi Wu [a,c], Witold Pedrycz [d]

[a] School of Information Engineering, Zhejiang Ocean University, Zhoushan, Zhejiang 316022, PR China
[b] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, PR China
[c] Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhoushan, Zhejiang 316022, PR China
[d] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6R 2V4, AB, Canada

ARTICLE INFO

Article history:
Received 23 October 2021
Received in revised form 17 December 2021
Accepted 9 February 2022
Available online 22 February 2022

Keywords:
Granular computing
Granular discrimination
Multi-label learning
Incomplete label

ABSTRACT

In multi-label learning, each training instance is associated with multiple class labels. It is typical in reality that relevant labels are partially missing and only a part of labels are valid, resulting in the problem of weak multi-label learning with missing labels. It is still an evident challenge to estimate the ground-truth label matrix and to generate a prediction function, especially on the multi-label data with a large number of missing labels. In this paper, we propose a multi-label learning framework within which feature structure and label manifold are both utilized to recover the incomplete label matrix and to train the classification model simultaneously. To mitigate the imbalanced risks brought by the sparse label issue, a self-adaptive penalty factor is imposed on the deviated predictions of different labels. Moreover, instance granular discrimination is incorporated in the framework to characterize the approximate distribution structure of data. Matrix vectorization, cave-convex programming (CCCP), and block coordinate descent techniques are employed to solve the proposed optimization problem. Extensive experiments illustrate the superiority of the proposed method against some well-established methods.

1. Introduction

In recent years, multi-label learning has a wide range of real-world applications, such as image annotation, text categorization and facial action unit (AU) recognition [16]. In multi-label learning, an instance is usually associated with multiple class labels simultaneously rather than a single one, and the main purpose of learning is to determine a model that can accurately predict the possible label set for test instances [22,42].

Conventional multi-label learning assumes that training data sets are completely labeled. However, this assumption is usually infeasible in practice due to the high costs of annotation and the ambiguity among class labels. For example, in image and video annotations, annotators with different backgrounds may ignore the labels they do not know or are of little interest, especially when the label space is too large. In this case, the labels of examples are partially available and many labels are missing, which is realized as the multi-label learning problem with missing labels [31]. In particular, annotators may only give complete labels for a part of training examples and many training examples are unlabeled, which is realized as

* Corresponding author.
E-mail addresses: tananhui86@163.com (A. Tan), jxwan549@163.com (X. Ji), ljy@sxu.edu.cn (J. Liang), shujujiegouwang@1626.com (Y. Tao), wuwz@zjou.edu.cn (W.-Z. Wu), wpedrycz@ualberta.ca (W. Pedrycz).

https://doi.org/10.1016/j.ins.2022.02.011
0020-0255/

semi-supervised multi-label learning problem [1,23]; annotators may only give partial relevant labels for training examples. In this case, many relevant labels are missing, which is realized as conventional multi-label learning problem with missing labels [31]; annotators may only give partial relevant and irrelevant labels for training examples. In this case, relevant and irrelevant labels of training examples are can be both missing, which is realized as general label learning problem with missing labels [41,47].

Recently, various types of approaches have been developed to conduct weak multi-label learning with missing labels. In general, those methods can be categorized into two categories: algorithm adaptation and problem transformation. The algorithm adaptation methods transform and upgrade existing learning algorithms so as to adapt them to various scenarios. The problem transformation methods transform the multi-label learning problem into other well-established learning tasks rather than directly learning from multi-label data. However, existing multi-label algorithms mainly utilize an identical data representation in the discrimination of all instances of all class labels, while they can not learn the instance distribution and discriminate the instance granular structure of data. Moreover, many studies ignore the class imbalance and sparse annotation problems and treat relevant and irrelevant labels equally. This may drive the trained models to be more adaptable to the structure of irrelevant labels rather than the relevant labels. Hence, the imbalance and sparsity of class labels could have a large influence on the predictive performance.

In this paper, we tackle the practical yet challenging problem, i.e., weak multi-label learning with missing labels, and propose a new multi-label method that can learn the granular structure of instance spaces and can discriminate the class labels of instance granules. In particular, the idea of Granular computing is introduced to partition the data into small sample granules and the label discrimination between different sample granules is formalized to increase the interclass distance and to minimize the innerclass distance of sample granules. In the method, the learning of classifiers and the recovery of the label matrix are performed simultaneously, and the instance and label manifolds are jointly utilized to reconstruct the label manifold and to train the feature label mapping. Moreover, the self-adaptive penalty factor is introduced to make a trade-off between the losses of the outputs for different labels. To accelerate the convergence of the optimization algorithm, the objective function w.r.t. each variable is converted to a vector form and is solved by concave-convex programming (CCCP) [40] and block coordinate descent algorithm [35]. Finally, extensive experiments demonstrate that the proposed method is superior to some state-of-the-art multi-label learning algorithms in multi-label classification with missing labels.

The remainder of this article is organized as follows. Section 2 reviews related works on multi-label learning with missing labels and Section 3 introduces the notations. In Sections 4, the proposed approach is formulated in detail, and the optimization solution is presented. The experimental results are analyzed in Section 5. Finally, Section 6 covers some concluding remarks.

2. Related work

Multi-label learning refers to the problem that each example has multiple class labels simultaneously and comprehensive reviews on this topic are available in excellent surveys [22,42]. Traditional supervised learning requires completely labeled training examples which may not be easy to obtain in many real world applications. Multi-label learning with incomplete labels has emerged in various application scenarios and has resulted in widespread attention in recent years. In some of early works, missing labels are usually treated as negative labels, and then the multi-label data are converted to complete data. For instance, Sun et al. [31] presented a weak label learning method by estimating the label information of training examples by considering the feature information. Chen et al. [7] defined a regularization framework including two regularization terms corresponding to the instance graph and category graph and predicted the labels of the unlabeled instances by solving a Sylvester equation. Yu et al. [39] proposed a generic low-rank empirical risk minimization framework for multi-label learning with missing labels. Bucak et al. [5] presented a multi-label learning framework by considering the errors in ranking the assigned class labels against the unassigned class labels. The idea of treating missing labels indiscriminately as negative labels is based on the assumption of annotation sparsity, which may give rise to poor and unstable performance, especially when too many ground-truth positive labels are recognized as negative labels.

To alleviate the influence of incomplete labels, matrix completion is an efficient scheme to handle weak multi-label classification tasks. Because of label-level/instance-level correlation, the Low-rank assumption is widely adopted to recover the label matrix of examples. For example, Zhu et al. [46] proposed a label refinement formulation that comprehensively considered the low rank and sparsity of the label matrix. Luo et al. [24] employed multiview matrix completion to linearly combine the predicted labels to approximate the ground-truth labels. Wu et al. [37] treated the testing instances as a part of the unlabeled data and predicted the unlabeled items based on instance-level smoothness and class-level smoothness. In this work, the prediction process is simultaneously accomplished in the training process, and no classifier is constructed, which may limit the inductive capability. Liu et al. [23] proposed a low-rank multiview matrix completion model by utilizing multiple features taken from different views. Recently, Ma et al. [25] explored the low-rank and sparse structures of label sets and transformed the feature matrix via low-dimensional embedding with a projection.

Label recovery and label mapping can also be implemented in a mutually benefit manner. For instance, Jing et al. [15] developed semisupervised multi-label learning methods for handling cases with a small number of labeled instances and a large number of unlabeled instances. Lin et al. [21] proposed the FalE algorithm via feature-aware implicit label space embedding. Li et al. [18] proposed a prediction model to handle missing labels by conducting ranking-preserving

low-rank factorization. Huang et al. [13] introduced a single framework within which missing label set recovery and the label-specific feature learning are combined. Note that these methods captured only the global low-rank label structure, while ignoring the local label structure. To solve this issue, Zhu et al. [47] exploited the global and local label structures of label sets simultaneously and learned a mapping from the feature space to the latent low-rank label space. To apply label similarity and instance similarity to the complement of missing labels, Dong et al. [10] constructed an ensemble of two independent classifiers to improve the robustness of label prediction. By considering the low rank of label spaces from global and local aspects, Ma et al. [26] developed a discriminative multi-label learning model by jointly capturing the local low-rank structure and the global high-rank structure of label spaces. However, the classifier was learned by minimizing the losses of the false outputs only on the observed labels which may limit the predictive effectiveness especially when the scale of missing labels is large.

In addition, Granular Computing refers to a new computational paradigm which focuses on structuralized knowledge organization and reasoning with information granules [19]. In early works, Zadeh [44] investigated information granulation using fuzzy set theory. In this granulation, information granules are formulated which can be regarded as a collection of examples drawn together by their closeness (resemblance, proximity, functionality, etc.) articulated in terms of some useful spatial, temporal, or functional relationships. Yu and Pedrycz [27] regarded fuzzy sets as information granules and employed the relationships among them to find solutions to various problems. Akram and Zafar [2–4] proposed the techniques of formation of granular structures using fuzzy soft graphs. They provided a valid way to compute the parameterized family of granules using fuzzy sets and fuzzy graphs. Although Granular Computing has played a visible role in the underlying technologies of machine learning and data mining, few works have been done on adopting Granular Computing for handling multi-label data sets [6,28,30]. The idea of information granulation can utilize the structuralized instance information and the label discriminant information to some extent, leaving room for further performance improvement of label prediction. To this end, in this paper, information granulation is introduced to capture the sample distributions of data and sample granules are leveraged to enhance the label discrimination.

3. Problem statement and notations

Given a training data set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ and a label set $L = \{l_1, l_2 \cdots, l_q\}$. The ith instance $\mathbf{x}_i$ is represented by a d-dimensional real-valued vector, which is associated with a q-dimensional binary label vector $\mathbf{y}_i$. Formally, denote $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ by the feature matrix (T is the transposition), and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n\times q}$ by the label matrix.

In weakly labeled data, the ground-truth label matrix $\mathbf{Y}$ is often unavailable and only an incomplete label matrix $\widetilde{\mathbf{Y}}$ is observed due to the missing annotations. In this case, each entry would be $\widetilde{\mathbf{Y}}_{ij} \in \{1, 0, -1\}$, where $\widetilde{\mathbf{Y}}_{ij} = 1$ means the ith instance has the jth label, $\widetilde{\mathbf{Y}}_{ij} = -1$ means the ith instance does not has the jth label, and $\widetilde{\mathbf{Y}}_{ij} = 0$ means the annotation is not available whose real value may be 1 or -1. The notations related in this paper are summarized in Table 1.

Our goal is to generate the coefficient matrix $\mathbf{W} \in R^{d\times q}$ which performs precisely on multi-label prediction and to construct the approximately accurate label matrix $\mathbf{Y}$ satisfying the following properties:

(1) Coefficient matrix $\mathbf{W}$ is used to establish the link between the instance matrix and the label matrix. This can help us utilize the feature information to reconstruct the label matrix.
(2) $\mathbf{Y}$ is consistent with the occurrence items in $\widetilde{\mathbf{Y}}$, i.e., $\mathbf{Y}_{ij} = \widetilde{\mathbf{Y}}_{ij}$ for $\widetilde{\mathbf{Y}}_{ij} \neq 0$.
(3) $\mathbf{Y}$ satisfies local manifold smoothness. Similar class labels generally own larger concurrence percentage, and similar instances generally have more relevant labels than dissimilar ones. In detail, each entry of $\mathbf{Y}$ can be derived by the nearest rows and columns in $\widetilde{\mathbf{Y}}$. This can help us utilize the observed label manifold information to estimate the label matrix.
(4) The instance granules extracted from data can appropriately characterize the underlying structure of the data. To reveal the global distribution of data, $\mathbf{W}$ can discriminate the instance granular structure. To be specific, the instance granules with the same labels should preserve large similarity on their predicted outputs, whereas the instance granules with different labels should preserve large diversity on their predicted outputs.
(5) $\ell_{2,1}$ regularization is used which can also boost the influence of discriminative features on reconstructing specific labels.

For a formal description, we introduce an integrated optimization loss function to express the above ideas:

$$\min \mathscr{F}(\cdot) + \lambda_1 \mathscr{S}(\cdot) + \lambda_2 \mathscr{M}(\cdot) + \lambda_3 \mathscr{R}(\cdot) + \lambda_4 \Gamma(\cdot), \tag{1}$$

where $\mathscr{F}(\cdot)$ is used to link the instance matrix with the label matrix, $\mathscr{S}(\cdot)$ is to preserve the consistency between $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$, $\mathscr{M}(\cdot)$ is the regulation term to utilize the local manifold smoothness, $\mathscr{R}(\cdot)$ is to enhance the discrimination of predicted outputs of instance granules, $\Gamma(\cdot)$ is the sparse regulation term of variables, and $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are trade-off parameters to balance the regularization terms.

Table 1
Summary of notations used in this paper.

Notations	Meaning
n	Number of instances
d	Number of features
q	Number of labels
$\circ$	Hadamard product
$\otimes$	Tensor product
$\mathbf{x}_i \in \mathbb{R}^d$	Feature vector of ith instance
$\mathbf{y}_i \in \mathbb{R}^q$	Label vector of ith instance
$\mathbf{I}$	Identity matrix
$\mathbf{1}_u$	Unit vector
$\mathbf{X} \in \mathbb{R}^{n\times d}$	Instance feature matrix
$\widetilde{\mathbf{Y}} \in \{+1, 0, -1\}^{n\times q}$	Observed label matrix
$\mathbf{S} \in \mathbb{R}^{n\times n}$	Instance similarity matrix
$\mathbf{C} \in \mathbb{R}^{q\times q}$	Label similarity matrix
$\mathbf{M}_{i\cdot}$	ith row of matrix $\mathbf{M}$
$\mathbf{M}_{\cdot j}$	jth column of matrix $\mathbf{M}$
$\mathbf{M}_{ij}$	i,jth entry of matrix $\mathbf{M}$
$\mathbf{M}^T$	Transpose of matrix $\mathbf{M}$
$tr(\mathbf{M})$	Trace of matrix $\mathbf{M}$
$\mathbf{vec}(\mathbf{M})$	Vectorization of matrix $\mathbf{M}$
$\mathbf{diag}(\mathbf{x})$	Diagonalization matrix of vector $\mathbf{x}$

4. Proposed method

4.1. Mapping consistency

There is high correlation between the feature space and the label space. The reconstruction for label matrix $\mathbf{Y}$ should take the advantage of the feature information. The reconstruction error can be represented by

$$\mathscr{F}(\mathbf{W}, \mathbf{Y}) = ||(\mathbf{XW} - \mathbf{Y}) \circ \mathbf{J}||_F^2, \tag{2}$$

where $\circ$ is the Hadamard product (entrywise product) and $\mathbf{J}$ is the penalty factor matrix with each $\mathbf{J}_{ij}$ setting as follows: If $\widetilde{\mathbf{Y}}_{ij} = 1$, then $\mathbf{J}_{ij} = -\frac{N^-}{N}$; If $\widetilde{\mathbf{Y}}_{ij} = -1$, then $\mathbf{J}_{ij} = \frac{N^+}{N}$, where N^+ is the number of entries valued $+1$ in matrix $\widetilde{\mathbf{Y}}$, N^- is the number of entries valued -1 in matrix $\widetilde{\mathbf{Y}}$, and $N = N^+ + N^-$. In this way, the erroneous predictions for positive labels as well as negative labels can be treated with equal importance. Moreover, due to the unknown labels of missing entries, the penalty is indiscriminately set $\mathbf{J}_{ij} = \frac{1}{2}$ when $\widetilde{\mathbf{Y}}_{ij} = 0$, which is the neutralization of the penalties of the positive and negative labels.

To select the label-specific features and to reduce the influence of irrelevant features, the sparse regulation term is learnt as $\ell_{2,1}$ norm of $\Gamma(\mathbf{W}) = ||\mathbf{W}||_{2,1}$.

4.2. Label consistency

The predicted label matrix $\mathbf{Y}$ should be consistent with the provided items in $\widetilde{\mathbf{Y}}$, i.e., $\mathbf{Y}_{ij} = \widetilde{\mathbf{Y}}_{ij}$ for $\widetilde{\mathbf{Y}}_{ij} \neq 0$. Then, the loss term is formulated as

$$\mathscr{S}(\mathbf{W}) = ||\left(\mathbf{Y} - \widetilde{\mathbf{Y}}\right) \circ \mathbf{T}||_F^2. \tag{3}$$

The penalty factor matrix $\mathbf{T}$ is set as: If $\widetilde{\mathbf{Y}}_{ij} \neq 0$, then $\mathbf{T}_{ij} = \mathbf{J}_{ij}$; If $\widetilde{\mathbf{Y}}_{ij} = 0$, then $\mathbf{T}_{ij} = 0$. The penalty factor matrix $\mathbf{T}$ can make a balance between positive labels and negative labels. Moreover, the penalty for missing entries is not incurred due to the unknown labels.

4.3. Local manifold smoothness

Missing labels can be derived from the information provided by nearest instances and nearest labels. We incorporate the local manifold smoothness for label propagation, which is formulated as a new regularization term:

$$\mathscr{M}(\mathbf{W}, \mathbf{Y}) = ||\left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C} - \mathbf{Y}\right) \circ \mathbf{J}||_F^2, \tag{4}$$

where $\mathbf{C}$ is the label similarity matrix which is needed to learn, and $\widetilde{\mathbf{S}}$ is the sparse instance similarity matrix such that:

$$\widetilde{\mathbf{S}}_{ij} = \begin{cases} \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right), & \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) or \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j); \\ 0, & otherwise, \end{cases}$$

where σ is the parameter which is set to 1 for simplicity and $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of top-k nearest neighbors of $\mathbf{x}_i$. For simplicity, the Euclidean distance is employed to compute the top-k nearest neighbors of instance.

Denote $\widehat{\mathbf{S}}$ by a diagonal matrix with $\widehat{\mathbf{S}}_{ii} = \sum_{j=1}^{n} \widetilde{\mathbf{S}}_{ij}$ equaling to the sum of the ith row of $\widetilde{\mathbf{S}}$. To make the similarity matrix invariant to different scalings, we normalize the similarity matrix as $\mathbf{S} \leftarrow \widehat{\mathbf{S}}^{-\frac{1}{2}} \widetilde{\mathbf{S}} \widehat{\mathbf{S}}^{-\frac{1}{2}}$.

Note that $\left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C}\right)_{ij} = \sum_p \sum_m \mathbf{S}_{ip} \widetilde{\mathbf{Y}}_{pm} \mathbf{C}_{mj}$, which means the i,jth entry of the label matrix can be derived by instance and label correlations. The larger the similarity between the ith and pth instances (and the larger the similarity between the mth and jth labels) is, the larger weights imposed on $\widetilde{\mathbf{Y}}_{pm}$ for computing $\mathbf{Y}_{ij}$ becomes.

4.4. Instance granular discrimination

Granular discrimination-based regularization can be employed to minimize the inner-class distance and maximize the intraclass distance of positive and negative instances[14,43]. We employ the cluster algorithm to partition the instance sets into p disjoint clusters whose centers are respectively denoted as $\{\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \cdots \widetilde{\mathbf{x}}_p\}$. In view of Granular computing [27,32,33,36,38], the cluster centers appropriately characterize the global granular structure of the data space. The aggregation and divergence degrees of those granules under the mth ($m = 1, \cdots q$) label can be approximately reflected by the following distance:

$$||(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)\mathbf{W}_{\cdot m}||_2^2 = \mathbf{W}_{\cdot m}^T (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)^T (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)\mathbf{W}_{\cdot m}, \tag{5}$$

where $\mathbf{W}_{\cdot m}$ is the mth column of $\mathbf{W}$ and $||\cdot||_2$ is the vector norm. Given the ith and jth cluster centers $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$, denote a matrix $\mathbf{D}_{\{ij\}} = (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)^T (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)$ (T is the transpose operator). Then, Eq. (5) is rewritten as $\mathbf{W}_{\cdot m}^T \mathbf{D}_{\{ij\}} \mathbf{W}_{\cdot m}$. Furthermore, one can see that the mapping distances between $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ under all q labels are stored in the diagonal entries of matrix $\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}$.

Denote a diagonal matrix $\widetilde{\mathbf{T}}_{\{ij\}} = \mathbf{diag}\left(\mathbf{T}_{id(\widetilde{\mathbf{x}}_i)\cdot} \circ \mathbf{T}_{id(\widetilde{\mathbf{x}}_j)\cdot}\right)$, where $id(\widetilde{\mathbf{x}}_i)$ is the index of the row that $\widetilde{\mathbf{x}}_i$ lies and $\mathbf{diag}(\cdot)$ is the diagonalization matrix of vector. Then, the discrimination loss of all cluster center pairs $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ under all labels can be formulated as

$$\mathscr{R}(\mathbf{W}) = \sum_{i,j=1}^{p} tr\left(\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right) \circ \widetilde{\mathbf{T}}_{\{ij\}}\right). \tag{6}$$

Property 1. $tr\left(\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right) \circ \widetilde{\mathbf{J}}_{\{ij\}}\right) = \sum_{m=1}^{q} ||(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j)\mathbf{W}_{\cdot m}||_2^2 \mathbf{T}_{im}\mathbf{T}_{jm}$.

Proof. We have $tr\left(\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right) \circ \widetilde{\mathbf{T}}_{\{ij\}}\right) = \sum_{m=1}^{q} \left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right)_{mm} \left(\widetilde{\mathbf{T}}_{\{ij\}}\right)_{mm}$. Since $\left(\widetilde{\mathbf{T}}_{\{ij\}}\right)_{mm} = \mathbf{T}_{im}\mathbf{T}_{jm}$, the conclusion is then derived. □.

The semantics of Property 1 can be elaborated as follows. Since the sign of each entry in $\mathbf{T}$ is consistent with each corresponding entry in $\widetilde{\mathbf{Y}}$, if the mth label notations of $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ are the same, i.e., both with $+1$ or -1 notations, then $\left(\widetilde{\mathbf{T}}_{\{ij\}}\right)_{mm} = \mathbf{T}_{im}\mathbf{T}_{jm} \geqslant 0$. In this case, $\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right)_{mm} \left(\widetilde{\mathbf{T}}_{\{ij\}}\right)_{mm} \geqslant 0$, which implies a positive penalty is imposed to minimize the diversity between their outputs on mth label. Otherwise, if the mth label notations of $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ are different, i.e., one with $+1$ label and another with -1 label, then a negative penalty is imposed to maximize the diversity between their outputs on mth label. Of course, missing labels involved are not penalized in Eq. (6).

Denote two matrixes by $\widetilde{\mathbf{T}}^+_{\{ij\}} = \max\left(\widetilde{\mathbf{T}}_{\{ij\}}, 0\right)$ and $\widetilde{\mathbf{T}}^-_{\{ij\}} = \max\left(-\widetilde{\mathbf{T}}_{\{ij\}}, 0\right)$, which retain the positive and negative entries in $\widetilde{\mathbf{T}}_{\{ij\}}$, respectively. We have $\widetilde{\mathbf{T}}_{\{ij\}} = \widetilde{\mathbf{T}}^+_{\{ij\}} - \widetilde{\mathbf{T}}^-_{\{ij\}}$. Then Eq. (6) can be written as the difference of two convex parts:

$$\sum_{i,j=1}^{p} tr\left(\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right) \circ \widetilde{\mathbf{T}}^+_{\{ij\}}\right) - \sum_{i,j=1}^{p} tr\left(\left(\mathbf{W}^T \mathbf{D}_{\{ij\}} \mathbf{W}\right) \circ \widetilde{\mathbf{T}}^-_{\{ij\}}\right). \tag{7}$$

4.5. Optimization

Based on the above considerations, the objective function, which is to find $\mathbf{W}, \mathbf{Y}$ and $\mathbf{C}$, can be summarized as follows:

$$\begin{aligned}\mathscr{L}(\mathbf{W},\mathbf{C},\mathbf{Y}) &= ||(\mathbf{XW}-\mathbf{Y})\circ\mathbf{J}||_F^2+\lambda_1||\left(\mathbf{Y}-\widetilde{\mathbf{Y}}\right)\circ\mathbf{T}||_F^2\\ &+\lambda_2||\left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C}-\mathbf{Y}\right)\circ\mathbf{J}||_F^2+\lambda_3\sum_{i,j=1}^{p}tr\left(\left(\mathbf{W}^T\mathbf{D}_{\{i,j\}}\mathbf{W}\right)\circ\widetilde{\mathbf{T}}^{+}_{\{i,j\}}\right)\\ &-\lambda_3\sum_{i,j=1}^{p}tr\left(\left(\mathbf{W}^T\mathbf{D}_{\{i,j\}}\mathbf{W}\right)\circ\widetilde{\mathbf{T}}^{-}_{\{i,j\}}\right)+\lambda_4||\mathbf{W}||_{2,1}.\end{aligned} \tag{8}$$

The proposed optimization model can simultaneously make use of the feature information and label information for label reconstruction. Moreover, the instance discrimination is considered to improve the discrimination power of the prediction model.

4.6. Update $\mathbf{W}$ with fixed $\mathbf{C}$ and $\mathbf{Y}$

When $\mathbf{C}$ and $\mathbf{Y}$ are fixed, the objective function in Eq. (8) w.r.t. $\mathbf{W}$ is simplified as:

$$\mathscr{L}_1(\mathbf{W},\mathbf{C},\mathbf{Y}) = \mathscr{G}_{vex}(\mathbf{W})+\mathscr{G}_{cav}(\mathbf{W}) \tag{9}$$

where $\mathscr{G}_{vex}(\mathbf{W})$ and $\mathscr{G}_{cav}(\mathbf{W})$ are respectively the convex part and the concave part, satisfying:

$$\begin{aligned}\mathscr{G}_{vex}(\mathbf{W}) = \;& ||(\mathbf{XW}-\mathbf{Y})\circ\mathbf{J}||_F^2\\ &+\lambda_3\sum_{i,j=1}^{p}tr\left(\left(\mathbf{W}^T\mathbf{D}_{\{i,j\}}\mathbf{W}\right)\circ\widetilde{\mathbf{T}}^{+}_{\{i,j\}}\right)+\lambda_4||\mathbf{W}||_{2,1},\\ \mathscr{G}_{cav}(\mathbf{W}) = \;& -\lambda_3\sum_{i,j=1}^{p}tr\left(\left(\mathbf{W}^T\mathbf{D}_{\{i,j\}}\mathbf{W}\right)\circ\widetilde{\mathbf{T}}^{-}_{\{i,j\}}\right).\end{aligned} \tag{10}$$

The derivatives of $\mathscr{G}_{vex}(\mathbf{W})$ and $\mathscr{G}_{cav}(\mathbf{W})$ w.r.t. $\mathbf{W}$ are calculated as follows: (denote $\mathbf{J}_1=\mathbf{J}\circ\mathbf{J}$ in the full paper),

$$\begin{aligned}\nabla\mathscr{G}_{vex}(\mathbf{W}) &= 2\mathbf{X}^T((\mathbf{XW}-\mathbf{Y})\circ\mathbf{J}_1)\\ &+2\lambda_3\sum_{i,j=1}^{p}\mathbf{D}_{\{i,j\}}\mathbf{W}\widetilde{\mathbf{T}}^{+}_{\{i,j\}}+2\lambda_4\Theta\mathbf{W},\\ \nabla\mathscr{G}_{cav}(\mathbf{W}) &= -2\lambda_3\sum_{i,j=1}^{p}\mathbf{D}_{\{i,j\}}\mathbf{W}\widetilde{\mathbf{T}}^{-}_{\{i,j\}},\end{aligned} \tag{11}$$

where Θ is the diagonal matrix satisfying $\Theta_{ii}=\frac{1}{2||\mathbf{W}_{i\cdot}||_2}$.

Denote $\mathbf{W}^t$ by the solution of $\mathbf{W}$ at each the tth iteration. According to the CCCP algorithm [40], $\mathbf{W}^{t+1}$ satisfies that

$$\nabla\mathscr{G}_{vex}\left(\mathbf{W}^{t+1}\right) = -\nabla\mathscr{G}_{cav}(\mathbf{W}^t). \tag{12}$$

This leads to the following update procedure:

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}}\left(\mathscr{G}_{vex}(\mathbf{W})+\nabla\mathscr{G}_{cav}(\mathbf{W}^t)\mathbf{W}\right). \tag{13}$$

Consequently, the optimization for $\mathbf{W}$ can be converted to solving the convex objective function. By setting the right-hand side formula in Eq. (13) to zero, we transform the linear equation into the vector form based on the theorem presented in [12]:

$$\begin{aligned}&\left(\mathbf{I}\otimes\mathbf{X}^T\right)\mathbf{diag}(\mathbf{vec}(\mathbf{J}_1))(\mathbf{I}\otimes\mathbf{X})\mathbf{vec}\left(\mathbf{W}^{t+1}\right)\\ &+\lambda_3\sum_{i,j=1}^{p}\left(\widetilde{\mathbf{T}}^{+}_{\{i,j\}}\otimes\mathbf{D}_{\{i,j\}}\right)\mathbf{vec}\left(\mathbf{W}^{t+1}\right)\\ &+\lambda_4\left(\mathbf{I}\otimes\Theta^{t+1}\right)\mathbf{vec}\left(\mathbf{W}^{t+1}\right)\\ &=\mathbf{vec}\left(\mathbf{X}^T(\mathbf{Y}\circ\mathbf{J}_1)\right)+\lambda_3\sum_{i,j=1}^{p}\left(\widetilde{\mathbf{T}}^{-}_{\{i,j\}}\otimes\mathbf{D}_{\{i,j\}}\right)\mathbf{vec}(\mathbf{W}^t),\end{aligned} \tag{14}$$

where $\otimes$ is the tensor product, $\mathbf{vec}(\cdot)$ is the vectorization of a matrix, and $\mathbf{diag}(\cdot)$ is the diagonalization matrix of a vector. It is observed that Θ^{t+1} is determined by $\mathbf{W}^{t+1}$ at each iteration, which is difficult to be directly obtained. Because the iteration is convergent, the tth iteration result Θ^t approximately substitutes Θ^{t+1} when computing $\mathbf{W}^{t+1}$ in Eq. (14).

For simplicity, we denote the following symbols:

$\mathbf{H} = \mathbf{I} \otimes \mathbf{X}$

$\mathbf{Q}^+ = \sum_{i,j=1}^{p} \left(\widetilde{\mathbf{T}}^+_{\{i,j\}} \otimes \mathbf{D}_{\{i,j\}} \right)$

$\mathbf{Q}^- = \sum_{i,j=1}^{p} \left(\widetilde{\mathbf{T}}^-_{\{i,j\}} \otimes \mathbf{D}_{\{i,j\}} \right)$

$\mathbf{U} = \mathbf{X}^T(\mathbf{Y} \circ \mathbf{J}_1)$.

We can rewrite Eq. (14) at the tth iteration as

$$\begin{aligned} &\left(\mathbf{H}^T\mathbf{diag}(\mathbf{vec}(\mathbf{J}_1))\mathbf{H} + \lambda_3\mathbf{Q}^+ + \lambda_4(\mathbf{I} \otimes \Theta^t)\right)\mathbf{vec}\left(\mathbf{W}^{t+1}\right) \\ &= \mathbf{vec}(\mathbf{U}^t) \qquad\qquad +\lambda_3\mathbf{Q}^-\mathbf{vec}\left(\mathbf{W}^t\right). \end{aligned} \tag{15}$$

Eq. (15) is a linear equation w.r.t. $\mathbf{W}^{t+1}$ which can be efficiently solved by many optimization packages.

4.7. Update C with Fixed W and Y

When $\mathbf{W}$ and $\mathbf{Y}$ are fixed, the objective function w.r.t $\mathbf{C}$ can be written as

$$\mathscr{L}_2(\mathbf{C}) = \lambda_2 || \left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C} - \mathbf{Y} \right) \circ \mathbf{T} ||_F^2. \tag{16}$$

Denote that $\mathbf{T}_1 = \mathbf{T} \circ \mathbf{T}$. The derivative of $\mathscr{L}_2(\mathbf{C})$ is calculated as follows:

$$\nabla\mathscr{L}_2 = 2\lambda_2 \left(\mathbf{S}\widetilde{\mathbf{Y}} \right)^T \left(\left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C} - \mathbf{Y} \right) \circ \mathbf{T}_1 \right). \tag{17}$$

By setting Eq. (17) to zero, we transform the linear equation into the vector form:

$$\begin{aligned} &\left(\mathbf{I} \otimes \left(\mathbf{S}\widetilde{\mathbf{Y}} \right)^T \right) \mathbf{diag}(\mathbf{vec}(\mathbf{T}_1)) \left(\mathbf{I} \otimes \left(\mathbf{S}\widetilde{\mathbf{Y}} \right) \right) \mathbf{vec}(\mathbf{C}) \\ &= \left(\mathbf{I} \otimes \left(\mathbf{S}\widetilde{\mathbf{Y}} \right)^T \right) \mathbf{diag}(\mathbf{vec}(\mathbf{T}_1))\mathbf{vec}(\mathbf{Y}). \end{aligned} \tag{18}$$

Denote $\mathbf{N} = \mathbf{I} \otimes \left(\mathbf{S}\widetilde{\mathbf{Y}} \right)$ for simplicity. We can rewrite Eq. (18) at the tth iteration as

$$\mathbf{N}^T\mathbf{diag}(\mathbf{vec}(\mathbf{T}_1))\mathbf{N}\mathbf{vec}\left(\mathbf{C}^{t+1}\right) = \mathbf{N}^T\mathbf{diag}(\mathbf{vec}(\mathbf{T}_1))\mathbf{vec}\left(\mathbf{Y}^t\right). \tag{19}$$

When $\mathbf{W}$ and $\mathbf{Y}$ are fixed, the solution of $\mathbf{C}$ at each iteration can be similarly obtained by solving the linear equation Eq. (19).

4.8. Update Y with fixed W and C

When $\mathbf{W}$ and $\mathbf{C}$ are fixed, the objective function w.r.t. $\mathbf{Y}$ can be written as

$$\begin{aligned} \mathscr{L}_3(\mathbf{Y}) = \; &||(\mathbf{X}\mathbf{W} - \mathbf{Y}) \circ \mathbf{J}||_F^2 + \lambda_1 || \left(\mathbf{Y} - \widetilde{\mathbf{Y}} \right) \circ \mathbf{T} ||_F^2 \\ &+\lambda_2 || \left(\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C} - \mathbf{Y} \right) \circ \mathbf{J} ||_F^2. \end{aligned} \tag{20}$$

Recall that $\mathbf{J}_1 = \mathbf{J} \circ \mathbf{J}$ and $\mathbf{T}_1 = \mathbf{T} \circ \mathbf{T}$. The derivative of $\mathscr{R}(\mathbf{Y})$ w.r.t. $\mathbf{Y}$ is calculated as follows:

$$\begin{aligned} \nabla\mathscr{L}_3 = \; &2(\mathbf{Y} - \mathbf{X}\mathbf{W}) \circ \mathbf{J}_1 + 2\lambda_1 \left(\mathbf{Y} - \widetilde{\mathbf{Y}} \right) \circ \mathbf{T}_1 \\ &+2\lambda_2 \left(\mathbf{Y} - \mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C} \right) \circ \mathbf{J}_1. \end{aligned} \tag{21}$$

By setting Eq. (21) to zero, we can transform the linear equation at the tth iteration into the vector form:

$$\begin{aligned} &((1 + \lambda_2)\mathbf{diag}(\mathbf{vec}(\mathbf{J}_1)) + \lambda_1\mathbf{diag}(\mathbf{vec}(\mathbf{T}_1)))\mathbf{vec}\left(\mathbf{Y}^{t+1}\right) \\ &= \mathbf{vec}\left(\left(\mathbf{X}\mathbf{W}^{t+1} + \lambda_2\mathbf{S}\widetilde{\mathbf{Y}}\mathbf{C}^{t+1} \right) \circ \mathbf{J}_1 + \lambda_1\widetilde{\mathbf{Y}} \circ \mathbf{T}_1 \right). \end{aligned} \tag{22}$$

By solving the linear equation Eq. (22), the solution of $\mathbf{Y}$ can be similarly obtained. As shown by the above analysis, the optimizations with respect to $\mathbf{W}, \mathbf{C}$ and $\mathbf{Y}$ are respectively convex. Hence, the three subroutines can be solved by alternating optimization based on the block coordinate descent algorithm [35] until convergence has occurred.

Algorithm 1: C2ML Algorithm

Input: Data matrix $\mathbf{X}$, observed label matrix $\widetilde{\mathbf{Y}}$, parameters $\lambda_i, (i = 1, 2, 3, 4)$ and the number of nearest neighbors k;
Output: Mapping matrix $\mathbf{W}$, label similarity matrix $\mathbf{C}$ and predicted label matrix $\mathbf{Y}$.

1: Initialize $\mathbf{W}^1 \leftarrow \left(\mathbf{X}^T\mathbf{X} + \eta\mathbf{I}\right)^{-1}\mathbf{X}^T\widetilde{\mathbf{Y}}, \mathbf{Y}^1 \leftarrow \widetilde{\mathbf{Y}}$, and randomly initialize $\mathbf{C}^1$;
2: Compute penalty factor matrices $\mathbf{J}$ and $\mathbf{T}$, instance similarity matrix $\mathbf{S}$;
3: **Repeat**
4: Compute the diagonal matrix Θ^t;
5: Update $\mathbf{W}^{t+1}$ with fixed $\mathbf{W}^t, \mathbf{C}^t, \mathbf{Y}^t$, and Θ^t by solving Eq. (15);
6: Update $\mathbf{C}^{t+1}$ with fixed $\mathbf{Y}^t$ by solving Eq. (19);
7: Update $\mathbf{Y}^{t+1}$ with fixed $\mathbf{W}^{t+1}$ and $\mathbf{C}^{t+1}$ by solving Eq. (22);
8: **Until** convergence;
9: **Return** $\mathbf{W}, \mathbf{C}$ and $\mathbf{Y}$.

5. Experiments

5.1. Experiment preparation

In this section, we perform a sequence of experiments to validate the effectiveness of the proposed method by comparing it with current state-of-art algorithms on data sets with missing labels. The multi-label data sets used for comparison are downloaded from the websites of Mulan[1] and Uco[2], which are outlined in Table 1. Most data sets have no less than 5,000 instances, which are usually regarded as large data in multi-label learning [43]. In Table 1, "Instance", "Features" and "Label" represent the number of instances, the number of features, and the number of labels, respectively. "Card." means the label cardinality, which is the number of labels distributed evenly to all instances, and "Dens." denotes the normalization of label cardinality by the number of labels. The incomplete label ratios vary with different portions of random missing labels, that is, 30%, 50%, and 70%. (See Table 2).

Four widely used rank-based metrics, including *MacroaverageAUC*, *Hammingloss*, *Coverage*, and *MacroF*1 are employed to examine the performances of different multi-label learning algorithms. These metrics measures the predictive accuracy from various aspects, whose detailed definitions can be found in [42]. For *MacroaverageAUC* and *MacroF*1, the greater the values, the better the performance; whereas for *Hammingloss* and *Coverage*, the smaller the values, the better the performance.

5.2. Comparing algorithms

We compare the proposed algorithms against some state-of-the-art algorithms, including MSWL [41], Glocal [47], ESMC [1], LSML [13], MNECM [8], DM2L-l [26], and DM2L-nl [26]. Note that ESMC and LSML were released in 2019 and MSWL, MNECM, DM2L-l, and DM2L-nl were released in 2020, and the advancement and superiority of the proposed algorithms can be guaranteed through comparative verification. All these algorithms were downloaded from open source websites, and their configuration parameters are those suggested in their original sources. Each of the algorithms focuses on weakly supervised learning tasks and is capable of handling missing labels.

C2ML and C2MLE: Two co-training multi-label methods proposed in this work. After obtaining the satisfactory $\mathbf{W}, \mathbf{Y}$ and $\mathbf{C}$, the two types of prediction schemes are implemented as follows. C2ML is based on the direct linear mapping: $\mathbf{Y}_{test} = \mathbf{X}_{test}\mathbf{W}$, where $\mathbf{X}_{test}$ is the matrix of test instances, and $\mathbf{Y}_{test}$ is the predicted label matrix for $\mathbf{X}_{test}$; C2MLE is based on ensemble learning: $\mathbf{Y}_{test} = \mathbf{X}_{test}\mathbf{W} + \lambda_2\mathbf{S}_{test,train}\mathbf{YC}$, where $\mathbf{S}_{test,train}$ is the similarity matrix between the test instances and training instances. The parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are searched in $\left\{10^{-3}, 10^{-2}, \cdots, 1\right\}$. For simplicity, the number of clusters p and the number of neighbors k are both set to 10. The influence when different parameters vary is also roughly examined as shown in Fig. 5.

MSWL [41]: It first fills missing labels by global label correlation with a *one-to-all* style, and then uses feature manifold to build the regularizer. The parameters α and β are tuned in $\left\{10^{-3}, 10^{-1}, \cdots, 10^{3}\right\}$. and γ is tuned in $\left\{10^{-6}, 10^{-5}, \cdots, 10^{6}\right\}$.

Glocal [47]: It applies the low-rank structure of the label matrix, and learns a mapping from the feature space to the latent labels based on global and local label correlation. The parameter λ is set to $1, \lambda_2$ is set to $10^{-3}, \lambda_3$ and λ_4 are searched in $\left\{10^{-4}, 10^{-3}, \cdots, 1\right\}$, and k and g are tuned in $\{5, 10, 15, 20\}$.

ESMC [1]: Non-linearly embedding-based methods is used to represent the label assignments in a low-dimensional space, which can predict the tail labels more accurately. The parameters λ, ρ and σ_z are searched in $\left\{10^{1}, 10^{2}, \cdots, 10^{5}\right\}$.

[1] http://mulan.sourceforge.net/datasets.html
[2] http://www.uco.es/kdis/mllresources/

Table 2
Description of multi-label data sets.

Data sets	Instances	Features	Labels	Card.	Dens.
Birds	645	260	19	1.470	0.074
Business	5,000	438	30	1.588	0.053
CAL500	502	68	174	26.043	0.149
Computers	5,000	681	33	1.509	0.048
Education	5,000	550	33	1.461	0.044
Health	5,000	612	32	1.662	0.052
Recreation	5,000	606	22	1.423	0.065
Reference	5,000	793	33	1.169	0.035
Science	5,000	743	40	1.451	0.036
Social	5,000	636	27	1.283	0.033
Society	5,000	636	27	1.692	0.063
Stackexchess	1,334	585	227	2.411	0.011

LSML [13]: A method to learn label-specific features for multi-label classification with incomplete labels. The parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are searched in $\left\{10^{-5}, 10^{-4}, \cdots, 10^{3}\right\}$.

MNECM [8]: A label confidence matrix is constructed using the positive and negative label density and kernel extreme learning machine in introduced for linear prediction. The kernel and regularization parameters are both set to 1 and the kernel function chooses RBF.

DM2L-l and DM2L-nl [26]: Two models based on the local low-rank and global high-rank structures of the label space. They employ linear and gaussian kernel functions to map the feature matrix into a high dimensional space. The parameter λ_d is searched in $\left\{10^{-5}, 10^{-4}, \cdots, 10^{5}\right\}$ and the threshold value δ is fixed at 0.005.

As suggested by [41], fivefold cross-validation are used to conduct the experiment with all the comparing methods. The experimental process on each data set is repeated five times and the average result of the five results is finally record.

5.3. Result analysis

Tables 3–14 record the experimental results of different methods in terms of the four evaluation metrics when the proportions of missing labels are respectively 30%, 50%, and 70%. The best result among all the algorithms on each data set is highlighted in boldface.As shown in Tables 3–14, each algorithm under comparison exhibits its own advantage in addressing the multi-label learning task with missing labels while the proposed method has statistically superior performance among all in terms of each of the evaluation metrics. Specifically, in Table 3, the proposed C2MLE outperforms all the other Algorithms 7 times among the data sets whereas the proposed C2ML is only inferior to C2MLE on most of the data sets. Similar cases are also reported in other tables with different metrics or different missing label ratios. The advantage becomes more pronounced when 50% of the labels are missing. In most cases, C2MLE constantly obtains the best result, and C2ML constantly obtains the second-best result. Furthermore, in Table 14, C2MLE and C2ML only perform worse than DM2L-l at four cases, and perform the best two at other cases. To summarize, the ensemble classifier C2MLE improves the results to a certain extent compared to the nonensemble version C2ML, and the proposed method performs satisfactorily against other comparison methods at different proportions of missing labels. (See Table 15).

As the results reported in these tables, we arrive at a couple of observations. All the learning algorithms can achieve relatively good performances in terms of the four metrics. In addition, the proposed algorithms performs almost among the best

Table 3
Comparative analysis of prediction performance of different algorithms in terms of *AUC* ↑ while the missing label ratio is 30%, where the best results (the larger the better) are shown in bold.

Methods	*AUC* ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.836	0.834	0.726	0.810	0.814	**0.853**	0.768	0.840	0.840
Business	0.898	0.917	0.904	0.895	0.906	0.913	0.913	**0.921**	0.920
CAL500	0.702	0.753	0.562	0.708	0.720	**0.795**	0.787	0.763	0.760
Computer	0.860	0.851	0.831	0.810	0.841	0.860	0.845	**0.867**	0.867
Education	0.849	0.843	0.843	0.809	0.858	0.848	0.871	**0.879**	0.877
Health	0.901	0.891	0.890	0.868	0.904	0.896	0.892	**0.909**	0.907
Recreation	0.544	0.782	0.752	0.751	0.785	**0.788**	0.739	0.785	0.784
Reference	0.862	0.858	0.855	0.824	0.866	0.870	0.868	**0.892**	0.891
Science	0.862	0.808	0.795	0.776	0.817	0.823	0.807	**0.838**	0.836
Social	0.880	0.863	0.860	0.844	0.882	0.882	0.873	**0.902**	0.900
Society	0.785	0.773	0.776	0.743	0.782	0.789	**0.797**	0.787	0.785
Stackexchess	0.880	**0.884**	0.777	0.825	0.842	0.826	0.816	0.867	0.865

Table 4
Comparative analysis of prediction performance of different algorithms in terms of *Rankingloss* ↓ while the missing label ratio is 30%, where the best results (the smaller the better) are shown in bold.

Methods	*Rankingloss* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.132	0.135	0.238	0.151	0.156	**0.117**	0.189	0.131	0.131
Business	0.250	**0.056**	0.065	0.074	0.060	0.059	0.059	**0.056**	0.059
CAL500	0.296	0.246	0.431	0.291	0.302	0.202	0.213	**0.205**	0.210
Computer	0.124	0.130	0.129	0.165	0.120	0.122	0.117	**0.108**	0.119
Education	0.131	0.135	0.135	0.167	0.118	0.126	0.118	**0.108**	0.110
Health	0.074	0.079	0.081	0.101	**0.067**	0.075	0.088	0.069	0.072
Recreation	0.448	0.185	0.209	0.213	0.173	0.180	0.228	**0.180**	0.187
Reference	0.121	0.124	0.123	0.157	0.112	0.111	0.117	**0.091**	0.098
Science	0.160	0.165	0.179	0.198	0.154	0.151	0.170	**0.140**	0.142
Social	0.091	0.105	0.105	0.121	0.085	0.085	0.095	**0.073**	0.074
Society	**0.164**	0.175	0.178	0.203	0.165	0.159	0.168	0.169	0.172
Stackexchess	0.120	**0.116**	0.220	0.174	0.150	0.170	0.171	0.132	0.135

Table 5
Comparative analysis of prediction performance of different algorithms in terms of *Coverage* ↓ while the missing label ratio is 30%, where the best results (the smaller the better) are shown in bold.

Methods	*Coverage* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	3.656	3.653	5.774	3.910	4.130	0.133	4.604	3.520	3.526
Business	9.164	3.122	3.570	3.956	3.564	0.138	3.308	**2.838**	2.912
CAL500	158.848	149.892	165.669	156.764	138.925	**137.661**	139.087	140.151	141.523
Computer	5.613	5.937	6.135	7.376	5.934	0.093	5.512	**4.911**	5.220
Education	5.810	6.147	6.057	7.385	5.511	0.116	4.929	**4.693**	4.763
Health	4.084	4.512	4.527	5.352	4.069	0.212	4.386	**3.768**	3.836
Recreation	10.902	5.135	5.629	5.833	**4.926**	0.193	5.866	5.019	5.120
Reference	4.759	4.901	4.924	5.964	4.451	0.114	4.450	**3.576**	3.752
Science	8.205	8.513	9.188	10.041	8.162	**0.139**	8.457	**7.226**	7.308
Social	4.816	5.541	5.548	6.245	4.571	**0.157**	4.933	**3.810**	3.866
Society	6.696	7.056	7.017	7.859	6.769	0.125	**6.480**	6.674	6.736
Stackexchess	49.854	**47.824**	85.095	71.239	64.754	**0.060**	70.363	54.787	55.243

Table 6
Comparative analysis of prediction performance of different algorithms in terms of *Macro F*1 ↑ while the missing label ratio is 30%, where the best results (the larger the better) are shown in bold.

Methods	*Macro F*1 ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.135	**0.314**	0.092	0.259	0.038	0.133	0.059	0.294	0.291
Business	**0.221**	0.125	0.144	0.131	0.074	0.138	0.138	0.194	0.154
CAL500	0.626	0.156	0.261	0.177	0.183	0.139	0.143	**0.140**	0.145
Computer	0.086	0.106	0.042	0.111	0.076	0.093	0.021	**0.147**	0.135
Education	0.086	0.115	**0.138**	0.117	0.076	0.116	0.061	0.121	0.119
Health	0.097	0.186	0.185	0.204	0.085	0.212	0.083	**0.232**	0.165
Recreation	0.115	0.188	0.060	0.192	0.117	0.193	0.060	**0.195**	0.189
Reference	0.072	0.110	0.117	0.114	0.060	0.114	0.061	**0.149**	0.103
Science	0.079	0.117	0.045	0.127	0.065	**0.139**	0.137	0.123	0.123
Social	0.067	0.133	0.107	0.130	0.055	**0.157**	0.051	0.144	0.144
Society	0.115	0.134	0.054	0.143	0.106	0.125	0.016	**0.152**	0.151
Stackexchess	0.026	0.014	0.017	0.019	0.020	**0.060**	0.012	0.038	0.037

two on most of the data sets. The success of the proposed algorithms is due to the simultaneous utilization of feature manifold and label manifold for training and the semi-supervised setting for prediction. Hence, the predictive classifier ensembles more compact and sufficient information of features and labels.

To further reflect the superiority of the proposed method, Friedman test [11] and Bonferroni-Dunn test [9] are utilized to evaluate the statistical significance of the compared algorithms. Let r_j be the average rank of algorithm j on all data sets, N the number of multi-label data sets, and K the number of multi-label feature selection algorithms. Under the null-hypothesis, Friedman statistic F_F follows a Fisher distribution: $F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}$, where $\chi_F^2 = \frac{12N}{K(K+1)}\left(\sum_{j=1}^{K} r_j^2 - \frac{K(K+1)^2}{4}\right)$.

Table 7
Comparative analysis of prediction performance of different algorithms in terms of *AUC* ↑ while the missing label ratio is 50%, where the best results (the larger the better) are shown in bold.

Methods	*AUC* ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.812	0.826	0.704	0.794	0.809	0.829	0.758	**0.830**	0.829
Business	0.676	0.918	0.888	0.876	0.891	0.908	0.913	**0.923**	0.922
CAL500	0.6858	0.7573	0.5955	0.6882	0.5121	0.7652	0.7791	**0.7932**	0.7819
Computer	0.848	0.851	0.816	0.785	0.825	0.848	0.845	**0.865**	0.864
Education	0.836	0.854	0.825	0.782	0.831	0.833	0.870	**0.885**	0.884
Health	0.889	0.893	0.878	0.835	0.884	0.881	0.890	**0.910**	0.909
Recreation	0.616	0.779	0.728	0.726	0.759	0.773	0.738	**0.791**	0.790
Reference	0.852	0.866	0.817	0.797	0.844	0.860	0.869	**0.893**	0.892
Science	0.797	0.804	0.758	0.743	0.788	0.806	0.807	**0.841**	0.839
Social	0.865	0.862	0.843	0.812	0.860	0.871	0.871	**0.903**	0.903
Society	0.769	0.768	0.752	0.715	0.756	0.773	0.796	**0.800**	0.799
Stackexchess	0.854	**0.866**	0.736	0.793	0.814	0.784	0.813	0.864	0.861

Table 8
Comparative analysis of prediction performance of different algorithms in terms of *Rankingloss* ↓ while the missing label ratio is 50%, where the best results (the smaller the better) are shown in bold.

Methods	*Rankingloss* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.152	0.136	0.258	0.168	0.151	0.135	0.194	**0.133**	0.133
Business	0.318	0.057	0.077	0.088	0.071	0.064	0.060	**0.055**	0.057
CAL500	0.3124	0.2425	0.4001	0.3110	0.4886	0.2315	0.2181	**0.2053**	0.2170
Computer	0.136	0.130	0.146	0.191	0.137	0.133	0.116	**0.106**	0.118
Education	0.146	0.126	0.156	0.196	0.145	0.143	0.119	**0.102**	0.104
Health	0.087	0.081	0.092	0.133	0.085	0.088	0.090	**0.070**	0.072
Recreation	0.371	0.189	0.238	0.242	0.200	0.193	0.230	**0.174**	0.180
Reference	0.129	0.117	0.162	0.185	0.133	0.121	0.115	**0.092**	0.099
Science	0.177	0.169	0.217	0.230	0.181	0.164	0.171	**0.139**	0.141
Social	0.102	0.104	0.120	0.150	0.105	0.094	0.096	**0.073**	0.073
Society	0.186	0.185	0.203	0.236	0.192	0.177	0.169	**0.162**	0.164
Stackexchess	0.148	0.132	0.257	0.200	0.174	0.208	0.178	**0.128**	0.133

Table 9
Comparative analysis of prediction performance of different algorithms in terms of *Coverage* ↓ while the missing label ratio is 50%, where the best results (the smaller the better) are shown in bold.

Methods	*Coverage* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	4.121	3.842	6.235	4.542	4.235	**3.700**	4.827	3.731	3.737
Business	10.787	3.150	4.137	4.689	4.200	3.620	3.077	**2.830**	2.905
CAL500	165.3944	151.2510	165.6255	157.0757	230.0460	145.8653	160.1598	**138.7689**	146.0916
Computer	6.031	5.858	6.773	8.218	6.510	6.012	5.481	**4.916**	5.246
Education	6.424	5.711	6.832	8.251	6.482	6.504	4.948	**4.430**	4.469
Health	4.658	4.564	5.083	6.572	4.857	5.026	4.450	**3.752**	3.802
Recreation	9.133	5.209	6.308	6.447	5.563	5.373	5.858	**4.847**	4.944
Reference	5.058	4.578	6.204	6.885	5.186	4.823	4.386	**3.568**	3.735
Science	8.988	8.602	10.725	11.316	9.318	8.662	8.486	**7.080**	7.147
Social	5.497	5.557	6.235	7.471	5.529	5.190	5.027	**3.796**	3.813
Society	7.276	7.275	7.832	8.683	7.552	7.117	6.514	**6.338**	6.383
Stackexchess	60.859	55.238	96.792	80.657	73.702	83.375	71.785	**54.287**	55.144

We can further obtain the Friedman statistic F_F on the evaluation metrics and the critical value as shown in Table 14 while the proportions of missing labels are respectively 30%, 50%, and 70%. As seen in these tables, the null hypothesis, that the performance of all methods is equivalent, is clearly rejected on each evaluation metric at significance level $\alpha = 0.05$. Then, Bonferroni-Dunn test is used to further evaluate the performance of the methods as shown in Figs. 1–3. The performance of two methods is regarded as different, if their average ranks exceeds the critical distance $CD_\alpha = q_\alpha \sqrt{\frac{K(K+1)}{6N}}$, where $q_\alpha = 3.102$ is the critical value of the test. We can calculate that $CD_\alpha = 3.4681$ ($N = 12, K = 9$).

Table 10
Comparative analysis of prediction performance of different algorithms in terms of *MacroF*1 ↑ while the missing label ratio is 50%, where the best results (the larger the better) are shown in bold.

Methods	*MacroF*1 ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.145	**0.302**	0.081	0.293	0.125	0.104	0.068	0.298	0.297
Business	0.069	0.123	0.145	0.128	0.074	0.131	0.098	**0.212**	0.141
CAL500	0.2804	**0.3809**	0.2939	0.3448	0.3123	0.3425	0.3657	0.3416	0.3421
Computer	0.087	0.099	0.044	0.107	0.076	0.095	0.021	**0.123**	0.116
Education	0.086	0.102	**0.131**	0.109	0.076	0.115	0.061	0.111	0.110
Health	0.098	0.186	0.183	0.176	0.085	0.211	0.070	**0.215**	0.150
Recreation	0.126	0.183	0.060	0.183	0.118	0.195	0.115	**0.189**	0.184
Reference	0.073	0.111	0.112	0.109	0.060	0.121	0.061	**0.147**	0.083
Science	0.079	0.104	0.046	**0.115**	0.065	0.123	0.056	0.105	0.104
Social	0.068	0.120	0.114	0.105	0.055	**0.172**	0.052	0.151	0.151
Society	0.116	0.122	0.045	0.132	0.106	0.117	0.016	**0.140**	0.139
Stackexchess	0.026	0.013	0.016	0.016	0.020	**0.052**	0.011	0.034	0.033

Table 11
Comparative analysis of prediction performance of different algorithms in terms of *AUC* ↑ while the missing label ratio is 70%, where the best results (the larger the better) are shown in bold.

Methods	*AUC* ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.782	0.779	0.649	0.721	0.757	**0.817**	0.758	0.799	0.799
Business	0.680	**0.917**	0.887	0.851	0.875	0.897	0.915	**0.917**	0.916
CAL500	0.6981	0.7328	0.5787	0.6512	0.5111	0.7851	0.7845	**0.7897**	0.7895
Computer	0.813	0.848	0.772	0.744	0.793	0.820	0.845	**0.858**	**0.858**
Education	0.807	0.852	0.772	0.726	0.776	0.807	0.870	**0.882**	0.881
Health	0.857	0.892	0.847	0.792	0.852	0.865	0.890	**0.913**	0.912
Recreation	0.594	0.766	0.694	0.681	0.717	0.755	0.735	**0.785**	0.784
Reference	0.829	0.863	0.783	0.749	0.805	0.837	0.868	0.889	**0.890**
Science	0.782	0.798	0.726	0.697	0.747	0.781	0.807	**0.837**	0.836
Social	0.841	0.872	0.812	0.770	0.829	0.854	0.873	**0.895**	0.895
Society	0.751	0.777	0.736	0.695	0.741	0.760	**0.796**	0.793	0.792
Stackexchess	0.824	**0.845**	0.668	0.741	0.758	0.742	0.807	0.835	0.833

Table 12
Comparative analysis of prediction performance of different algorithms in terms of *Rankingloss* ↓ while the missing label ratio is 70%, where the best results (the smaller the better) are shown in bold.

Methods	*Rankingloss* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.177	0.181	0.325	0.237	0.210	**0.151**	0.204	0.169	0.169
Business	0.312	**0.056**	0.077	0.111	0.084	0.072	0.058	0.059	0.061
CAL500	0.3004	0.2669	0.4173	0.3480	0.4916	0.2346	0.2315	**0.2093**	0.2138
Computer	0.171	0.132	0.184	0.236	0.170	0.159	0.115	**0.110**	0.120
Education	0.178	0.134	0.215	0.258	0.206	0.173	0.120	**0.106**	0.107
Health	0.115	0.082	0.121	0.178	0.115	0.103	0.089	**0.066**	0.069
Recreation	0.397	0.201	0.275	0.293	0.247	0.211	0.230	**0.180**	0.186
Reference	0.153	0.122	0.195	0.233	0.171	0.143	0.117	**0.096**	0.101
Science	0.196	0.180	0.252	0.282	0.226	0.192	0.171	**0.145**	0.146
Social	0.125	0.098	0.147	0.196	0.134	0.109	0.094	**0.076**	0.076
Society	0.203	0.177	0.222	0.262	0.211	0.191	0.170	**0.167**	0.168
Stackexchess	0.180	**0.147**	0.322	0.254	0.230	0.248	0.183	0.157	0.160

As indicates in Figs. 1–3, the rank of C2MLE exceeds that of LSML, ESMC, Glocal, DM2L-nl, and MSWL over one *CD* on all the four metric with 30% missing label, and C2MLE outperforms LSML, ESMC, MNECM, DM2L-l, and MSWL by one *CD* margin on *AUC*, *Rankingloss* and *Coverage* with 50% missing label. Accordingly, the proposed method has significantly different performance to each of the comparing method. Moreover, in Fig. 1, C2ML can defeat LSML, ESMC, Glocal and MSWL on *AUC* and *Coverage*, and can defeat DM2L-nl and MNECM on *MacroF*1. It is shown even greater advantages when comparing the algorithms in Figs. 2 and 3. The proposed method achieves competitive performances against other well-established multi-label classification approaches on those selected data sets.

Table 13
Comparative analysis of prediction performance of different algorithms in terms of *Coverage* ↓ while the missing label ratio is 70%, where the best results (the smaller the better) are shown in bold.

Methods	*Coverage* ↓								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	4.644	4.789	7.421	5.74	5.285	**4.040**	4.926	4.412	4.409
Business	10.707	3.071	4.051	5.366	4.571	3.943	**2.971**	3.008	3.079
CAL500	164.1315	157.1793	165.0359	162.4382	230.2069	153.1256	165.0128	**143.0637**	143.8367
Computer	7.418	5.929	8.197	9.694	7.631	7.085	5.442	**5.114**	5.416
Education	7.543	5.795	8.832	10.387	8.634	7.498	4.980	**4.535**	4.571
Health	5.921	4.541	6.258	8.231	6.097	5.603	4.414	**3.636**	3.687
Recreation	9.868	5.479	7.103	7.515	6.596	5.846	5.858	**4.987**	5.082
Reference	5.882	4.740	7.338	8.480	6.530	5.574	4.442	**3.669**	3.783
Science	9.938	9.061	12.248	13.415	11.163	9.866	8.445	**7.348**	7.381
Social	6.414	5.157	7.405	9.233	6.757	5.835	4.933	**4.092**	4.128
Society	7.871	7.032	8.466	9.337	8.112	7.558	**6.532**	6.535	6.562
Stackexchess	71.47	**63.065**	114.831	94.342	89.143	96.345	75.072	65.29	64.419

Table 14
Comparative analysis of prediction performance of different algorithms in terms of *MacroF*1 ↑ while the missing label ratio is 70%, where the best results (the larger the better) are shown in bold.

Methods	*Macro F*1 ↑								
	MSWL	Glocal	ESMC	LSML	MNECM	DM2L-l	DM2L-nl	C2MLE	C2ML
Birds	0.135	**0.193**	0.100	0.184	0.133	0.098	0.065	0.184	0.184
Business	0.070	0.115	0.134	0.105	0.075	0.135	0.098	**0.185**	0.117
CAL500	0.2924	0.3610	0.2691	0.3509	**0.3611**	0.2691	0.3209	0.3510	0.3540
Computer	0.087	0.078	0.039	0.092	0.077	0.082	0.021	**0.108**	0.103
Education	0.085	0.082	**0.115**	0.094	0.077	0.094	0.061	0.111	0.109
Health	0.100	0.186	0.156	0.141	0.086	**0.206**	0.079	0.203	0.139
Recreation	0.125	0.162	0.034	0.166	0.120	**0.190**	0.033	0.167	0.163
Reference	0.072	0.089	0.099	0.088	0.060	0.104	0.061	**0.130**	0.068
Science	0.080	0.085	0.042	0.100	0.065	**0.131**	0.041	0.095	0.094
Social	0.067	0.099	0.101	0.086	0.055	0.123	0.051	**0.130**	**0.130**
Society	0.116	0.110	0.045	0.123	0.108	0.120	0.017	**0.140**	**0.140**
Stackexchess	0.025	0.014	0.029	0.027	0.020	**0.039**	0.035	0.025	0.026

Table 15
Summary of the Friedman statistics F_F ($K = 9, N = 12$) and the critical value with different evaluation metrics ($\alpha = 0.05$)

Missing ratio	Metrics	F_F	Critical value
30%	*AUC*	19.1563	2.0578
	Rankingloss	13.7891	
	Coverage	17.2458	
	*MacroF*1	11.8652	
50%	*AUC*	43.7896	2.0578
	Rankingloss	43.9635	
	Coverage	43.8596	
	*MacroF*1	17.1256	
c 70%	*AUC*	53.9452	2.0578
	Rankingloss	55.6158	
	Coverage	50.8756	
	*MacroF*1	11.1456	

5.4. Sensitivity analysis

In this part, we analyze the influence of the parameters involved in the experiments, including regularization parameters $\lambda_i, i = 1, 2, 3, 4$, and the number of clusters p. We vary one or two parameters while keeping the others fixed. Figs. 4(a)–(d) demonstrates the influence of λ_1 and λ_2 on the Birds data set. As shown in the figure, the method performs well in some intermediate regions and gradually performs worse toward the outer regions; however, when λ_1 and λ_2 are too small, the feature manifold and label information are not fully utilized, while λ_1 and λ_2 are too large, the performance deteriorates because the effect of feature mapping is weakened. Figs. 4(e)–(l) suggest that there are suitable bounds of λ_3 and λ_4 that can obtain stable performance. This is because that λ_3 and λ_4 are respectively used to control the discrimination of instance

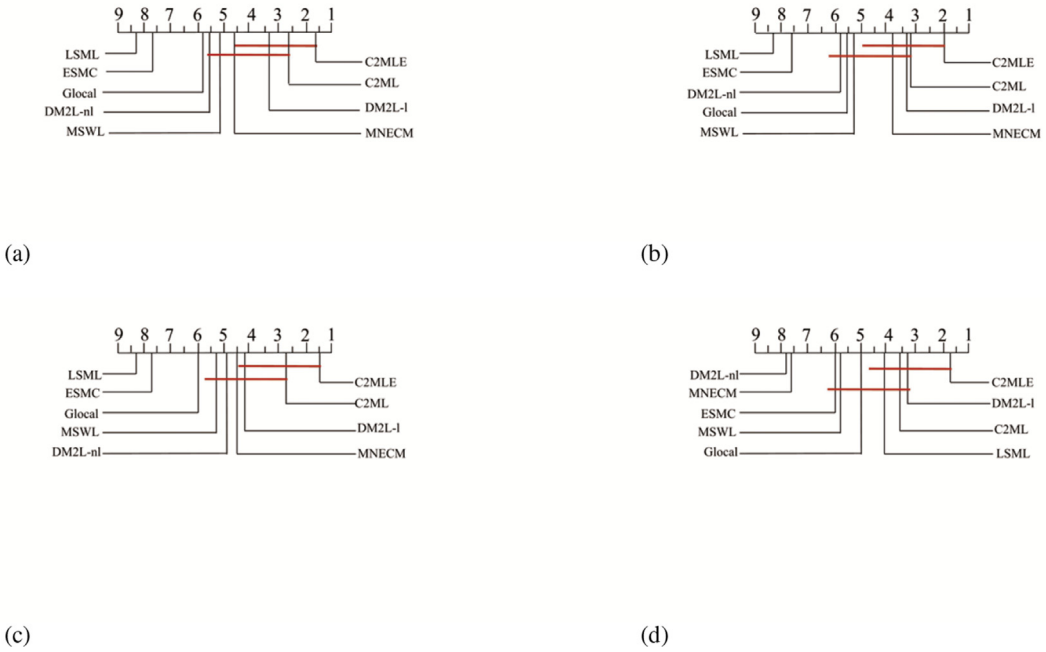

Fig. 1. Comparison of C2MLE and C2ML against algorithms under comparison with the Bonferroni-Dunn test ($CD_\alpha = 3.4681$ at 0.05 significance level) when the missing label ratio is 30% in terms of (a) *AUC*, (b) *Rankingloss*, (c) *Coverage*, (d) *MacroF*1.

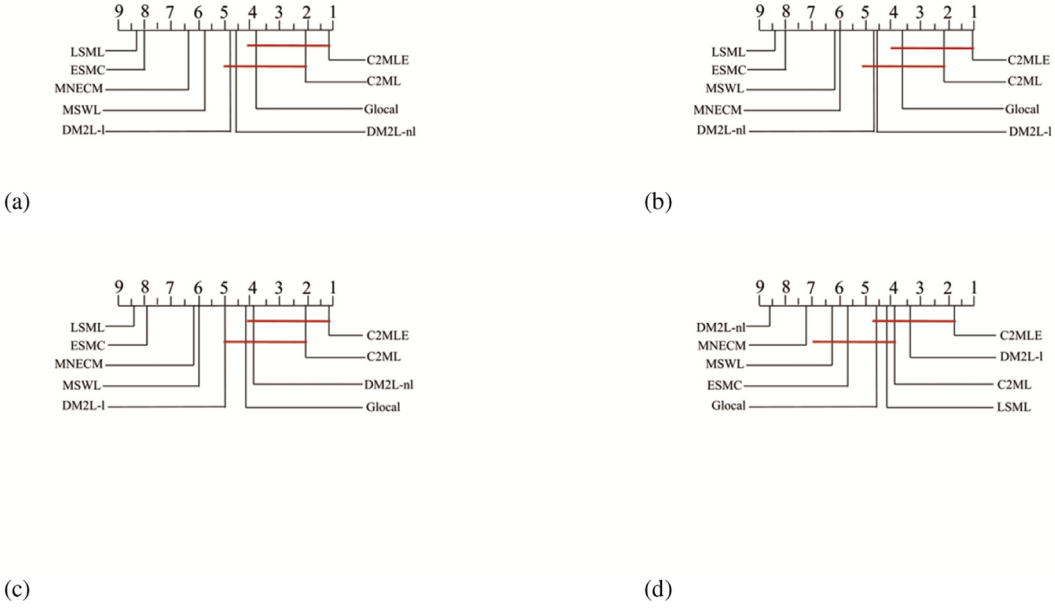

Fig. 2. Comparison of C2MLE and C2ML against algorithms under comparison with the Bonferroni-Dunn test ($CD_\alpha = 3.4681$ at 0.05 significance level) when the missing label ratio is 50% in terms of (a) *AUC*, (b) *Rankingloss*, (c) *Coverage*, (d) *MacroF*1.

granules and the complexity of the mapping, and their values should be limited to a certain range because of their subordinate priority. Fig. 5 shows the influence when the number of clusters varies. Fuzzy c-means cluster algorithm is adopted to partition the instances into p groups. It is suggested that too many or too few clusters cannot achieve satisfactory results.

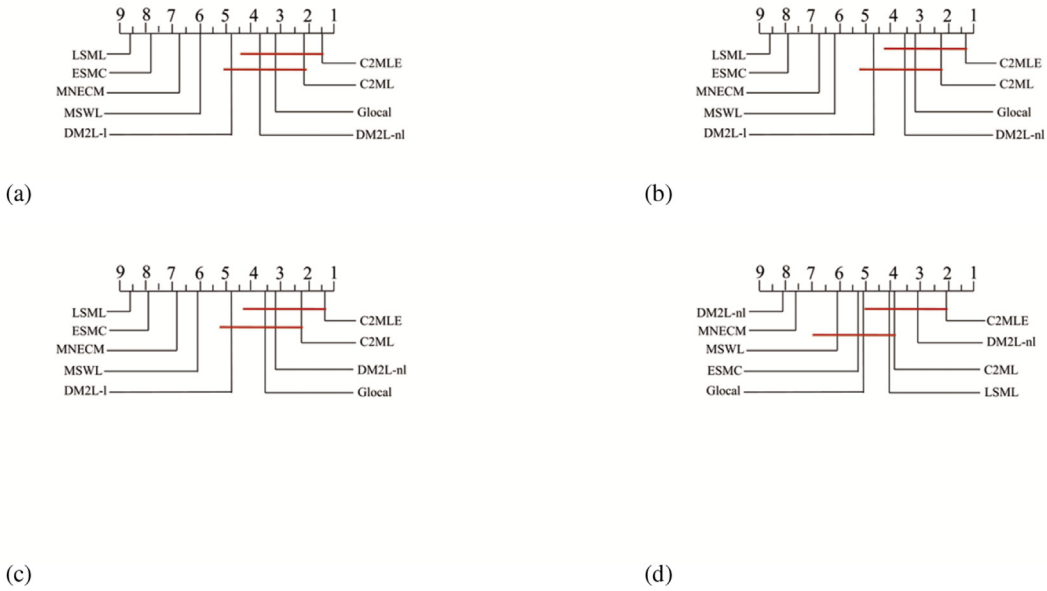

Fig. 3. Comparison of C2MLE and C2ML against algorithms under comparison with the Bonferroni-Dunn test ($CD_\alpha = 3.4681$ at 0.05 significance level) when the missing label ratio is 70% in terms of (a) *AUC*, (b) *Rankingloss*, (c) *Coverage*, (d) *MacroF*1.

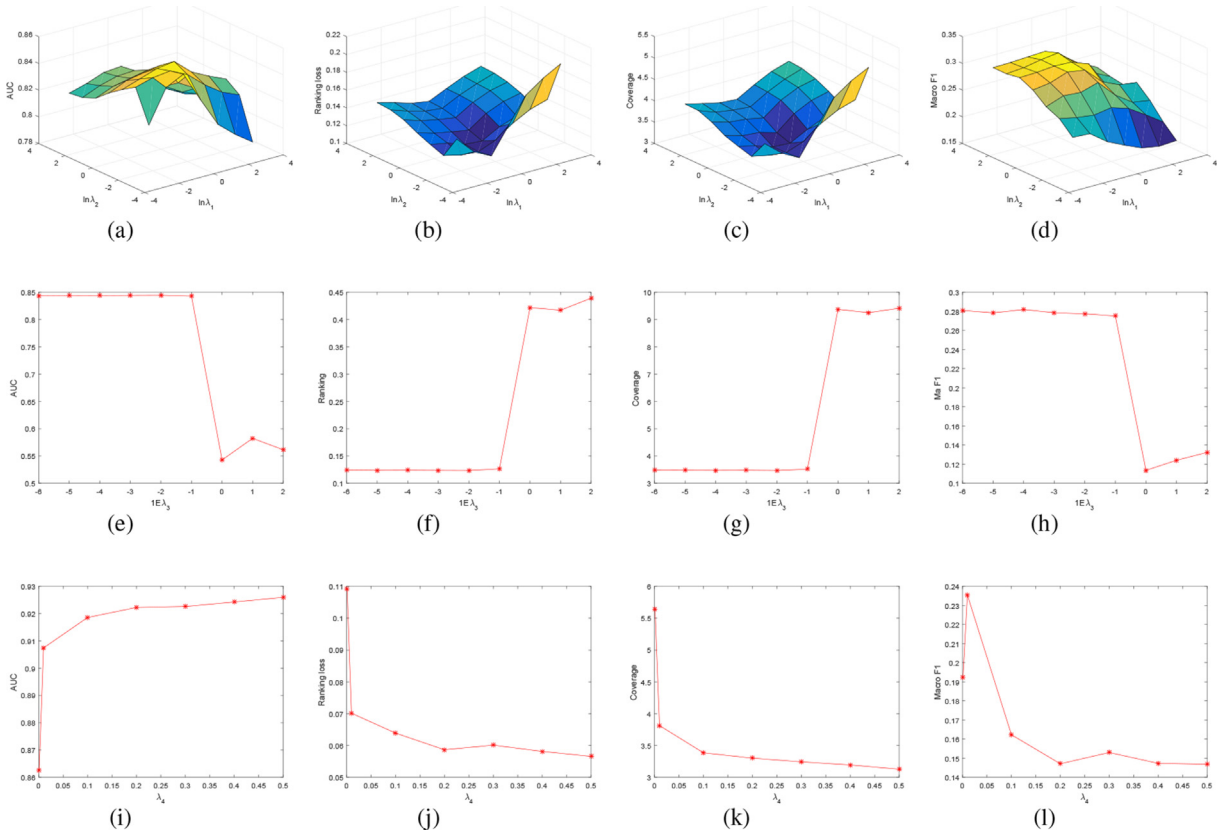

Fig. 4. Parameter sensitivity analysis of C2MLE. (a)–(d) Influence with varying λ_1 and λ_2 on the Birds dataset. (e)–(h) Influence with varying λ_3 on the Birds dataset. (i)–(l) Influence with varying λ_4 on the Health dataset.

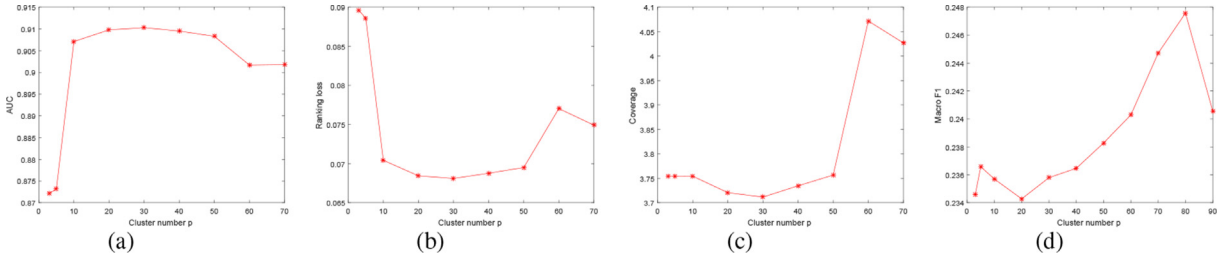

Fig. 5. Influence with varying the number of clusters p on the Health dataset in terms of (a) *AUC*, (b) *Rankingloss*, (c) *Coverage*, (d) *MacroF*1.

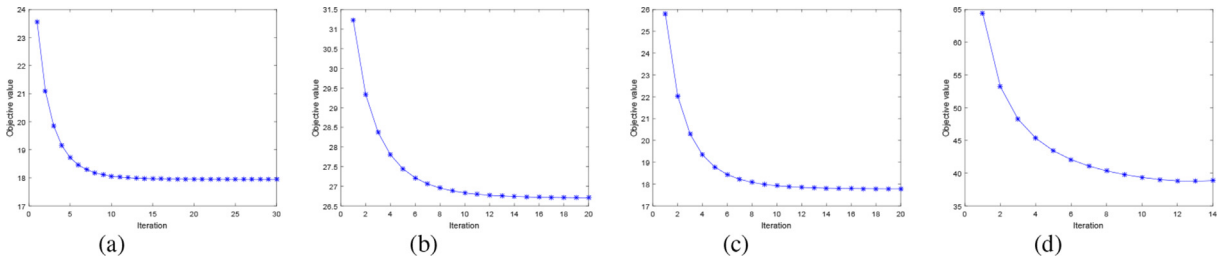

Fig. 6. Convergence analysis of C2MLE on the (a) Education, (b) Recreation, (c) Sciences, (d) Stackexchess.

5.5. Convergence analysis

Below, we analyze the convergence of the proposed algorithm. The alternate iteration and the CCCP we adopted can guarantee the convergence of the proposed algorithm. Furthermore, Figs. 6(a)–(d) outline the change of the objective value with respect to each iteration on four data sets, i.e., Education, Recreation, Sciences, and Stackexchess. It is empirically demonstrated that the curve falls fast within a small number of iterations and then tends towards stability. Hence, the results empirically demonstrate the convergence of the proposed algorithm in practice.

6. Conclusion

Granular computing has provided a structuralized framework of information organization and information reasoning. In this paper, the theory of Granular computing is applied to deal with multi-label data with missing label and the granulation scheme is adopted to partition the data into small sample granules. The label discrimination between different sample granules is formalized to increase the interclass distance and to minimize the innerclass distance of sample granules. The label discrimination of sample granules is further combined with label-specific feature learning and label correlation learning to construct training optimization problem for multi-label learning with missing label. Moreover, an adaptive penalty mechanism is imposed on the labels with different scales to achieve equal importance. A comparative study involving some well-established approaches verifies the competitive performance of the proposed method. In short, one of the important contributions of this work is that a unified multi-label learning framework is proposed to the joint learning of classifier, label correlation, and sample granular discrimination. In future work, we will explore fast convergent algorithms and incorporate the theory of granular computing into weakly supervised learning and hierarchical classification.

CRediT authorship contribution statement

Anhui Tan: Conceptualization, Methodology, Software. **Xiaowan Ji:** Data curation, Writing - original draft. **Jiye Liang:** Visualization, Investigation. **Yuzhi Tao:** Supervision. **Wei-Zhi Wu:** Software, Validation. **Witold Pedrycz:** Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgements

This work was supported by National Key Research and Development Program of China (No. 2020AAA0106100) and National Natural Science Foundation of China (Nos. 62076221 and 61976194).

References

[1] A. Akbarnejad, M. Baghshah, An efficient semi-supervised multi-label classifier capable of handling missing labels, IEEE Trans. Knowl. Data Eng. 31 (2019) 229–242.
[2] M. Akram, K. Anam, Granulation of ecological networks under fuzzy soft environment, Soft Comput. 24 (2020) 11867–11892.
[3] M. Akram, A. Al-Kenani, A. Luqman, Certain models of granular computing based on rough fuzzy approximations, J. Intel. Fuzzy Syst. 39 (2020) 2797–2816.
[4] M. Akram, A. Al-Kenani, A. Luqman, Degree based models of granular computing under fuzzy indiscernibility relations, Math. Biosc. Eng. 18 (2021) 415–8443.
[5] S. Bucak, R. Jin, A. Jain, Multi-label learning with incomplete class assignments, in: Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit., Providence, RI, USA, 2011, pp. 2801–2808..
[6] X. Che, D. Chen, J. Sheng, A novel approach for learning label correlation with application to feature selection of multi-label data, Inf. Sci. 512 (2020) 795–812.
[7] G. Chen, Y. Song, F. Wang, C. Zhang, Semi-supervised multi-label learning by solving a sylvester equation, in: Proc. SIAM Int. Conf. Data Mining, pp. 410–419, 2008..
[8] Y. Cheng, K. Qian, Y. Wang, D. Zhao, Missing multi-label learning with non-equilibrium based on classification margin, Appl. Soft Comput. 86 (2020) 105924.
[9] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2016) 1–30.
[10] H. Dong, Y. Li, Z. Zhou, Learning from semi-supervised weaklabel data, in: Proc. 32nd AAAI Conf. Artif. Intell., New Orleans, LA, USA, 2018, pp. 2926–2933..
[11] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Statist. 11 (1940) 86–92.
[12] R. Horn, C. Johnson, Topics in matrix analysis, Cambridge University Presss, Cambridge, pp. 37:39, 1991..
[13] J. Huang, F. Qin, X. Zheng, Z. Cheng, Z. Yuan, W. Zhang, Q. Huang, Improving multi-label classification with missing labels by learning label-specific features, Inf. Sci. 492 (2019) 124–146.
[14] J. Huang, G. Li, Q. Huang, X. Wu, Joint feature selection and classification for multi-label learning, IEEE Trans. Cybern. 48 (2018) 876–889.
[15] L. Jing, L. Yang, Y. Jian, M.K. Ng, Semi-supervised low-rank mapping learning for multi-label classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1483–1491, Jun. 2015..
[16] I. Katakis, G. Tsoumakas, I. Vlahavas, Multi-label text classification for automated tag suggestion, in Proc. ECML/PKDD 2008..
[17] X. Kong, M.K. Ng, Z. Zhou, Transductive multi-label learning via label set propagation, IEEE Trans. Knowl. Data Eng. 25 (2013) 704–719.
[18] X. Li, B. Shen, B. Liu, Y. Zhang, Ranking-preserving low-rank factorization for image annotation with missing labels, IEEE Trans. Multimedia 20 (2018) 1169–1178.
[19] T. Lin, Granular computing, In: Announcement of the BISC special interest group on granular computing, 1997..
[20] Y. Lin, Q. Hu, J. Liu, X. Zhu, X. Wu, MULFE: Multi-label learning via label-specific feature space ensemble, ACM Trans. Knowl. Discov. Data 16 (2022) 5:1–5:24..
[21] Z. Lin, G. Ding, M. Hu, J. Wang, Multi-label classification via feature-aware implicit label space encoding, in: Proc. 31th Int. Conf. Mach. Learn., un. 2014, pp. 325–333..
[22] W. Liu, X. Shen, H. Wang, I.W. Tsang, The emerging trends of multi-label learning, arXiv:2011.11197, Dec. 2020..
[23] Y. Liu, K. Wen, Q. Gao, X. Gao, F. Nie, SVM based multi-label learning with missing labels for image annotation, Pattern Recognit. 78 (2018) 307–317.
[24] Y. Luo, T. Liu, D. Tao, C. Xu, Multiview matrix completion for multi-label image classification, IEEE Trans. Image Process. 24 (2015) 2355–2368.
[25] J. Ma, Z. Tian, H. Zhang, T.W.S. Chow, Multi-label low-dimensional embedding with missing labels, Knowl.-Based Syst. 137 (2017) 65–82.
[26] Z. Ma, S. Chen, Expand globally, shrink locally: Discriminant multi-label learning with missing labels, Pattern Recognit. 111 (2021) 107675.
[27] W. Pedrycz, A. Bargiela, Granular clustering: a granular signature of data, IEEE Trans. Syst. Man. Cybern. B, Cybern. 32 (2002) 212–224.
[28] W. Qian, Y. Xiong, J. Yang, W. Shu, Feature selection for label distribution learning via feature similarity and label correlation, Inf. Sci. 582 (2022) 38–59.
[29] L. Sun, T. Wang, W. Ding, J. Xu, Y. Lin, Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification, Inf. Sci. 578 (2021) 887–912.
[30] L. Sun, T. Yin, W. Ding, Y. Qian, J. Xu, Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy, IEEE Trans. Fuzzy Syst. (2020), https://doi.org/10.1109/TFUZZ.2021.3053844, in press.
[31] Y. Sun, Y. Zhang, Z. Zhou, multi-label learning with weak label, in: Proc. 24th AAAI Conf. Artificial Intelligence, 2010..
[32] A. Tan, W. Wu, Y. Qian, J. Liang, J. Chen, J. Li, Intuitionistic fuzzy rough set-based granular structures and attribute subset selection, IEEE Trans. Fuzzy Syst. 27 (2019) 527–539.
[33] A. Tan, S. Shi, W. Wu, J. Li, W. Pedrycz, Granularity and entropy of intuitionistic fuzzy information and their applications, IEEE Trans. Cybern. 52 (2022) 192–204.
[34] Q. Tan, G. Yu, C. Domeniconi, J. Wang, Z. Zhang, Multi-view weak-label learning based on matrix completion, in: Proc. SIAM Int. Conf. Data Min., San Diego, CA, USA, 2018, pp. 450–458..
[35] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (2001) 475–494.
[36] C. Wang, Y. Huang, W. Ding, Z. Cao, Attribute reduction with fuzzy rough self-information measures, Inf. Sci. 549 (2021) 68–86.
[37] B. Wu, F. Jia, W. Liu, B. Ghanem, S. Lyu, Multi-label learning with missing labels using mixed dependency graphs, Int. J. Comput. Vis. 126 (2018) 875–896.
[38] S. Xia, Y. Liu, X. Ding, G. Wang, H. Yu, Y. Luo, Granular ball computing classifiers for efficient, scalable and robust learning, Inf. Sci. 483 (2019) 136–152.
[39] H. Yu, P. Jain, P. Kar, I.S. Dhillon, Large-scale multi-label learning with missing labels, in: Proc. Conf. Mach. Learn., Beijing, China, 31th Int., 2014, pp. 593–601.
[40] A. Yuille, A. Rangarajan, The concave-convex procedure, Neural Comput. 15 (2003) 915–936.
[41] J. Zhang, S. Li, M. Jiang, K.C. Tan, Learning from weakly labeled data based on manifold regularized sparse model, IEEE Trans. Cybern. DOI: 10.1109/TCYB.2020.3015269, 2020, in press..
[42] M. Zhang, Z. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (2014) 1819–1837.
[43] M. Zhang, L. Wu, LIFT: Multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 107–120.
[44] L. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (1997) 111–127.
[45] F. Zhao, Y. Guo, Semi-supervised multi-label learning with incomplete labels, in Proc. IJCAI, 2015, pp. 4062–4068..
[46] G. Zhu, S. Yan, Y. Ma, Image tag refinement toward low-rank content-tag prior and error sparsity, in: Proc. ACM Int. Conf. Multimedia, pp. 461–470, 2010..
[47] Y. Zhu, J.T. Kwok, Z. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 30 (2018) 1081–1094.