Applied Soft Computing Journal 86 (2020) 105924

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

Missing multi-label learning with non-equilibrium based on classification margin

Yusheng Cheng [a,b,*], Kun Qian [a], Yibin Wang [a,b], Dawei Zhao [a]

[a] School of Computer and Information, Anqing Normal University, Anqing 246011, China
[b] The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing 246011, China

ARTICLE INFO

Article history:
Received 13 April 2019
Received in revised form 24 September 2019
Accepted 5 November 2019
Available online 7 November 2019

Keywords:
Multi-label learning
Label completion
Information entropy
Classification margin
Non-equilibrium

ABSTRACT

Multi-labels are more suitable for the ambiguity of the real world. However, missing labels are common in multi-label learning datasets; this results in unbalanced labeling and label diversity, which directly affect the performance of multi-label learning. Therefore, the classification and modeling of imbalanced data in missing multi-label learning are problems that need to be urgently solved Current methods mostly focus on combining sampling techniques with cost-sensitive learning and incorporating label correlation to improve the performance of the classifier, but generally they do not consider label loss caused by label cost. In fact, labeling unknown instances is often affected by the threshold of the discriminant function, especially for the label types near the threshold. Based on our previous research, we believe that information such as data distribution density and label density can be integrated into the label correlation, and that the classification margin can be expanded to effectively solve the labeling quality of labels near the threshold. Therefore, in this paper we propose a non-equilibrium multi-label learning algorithm based on the classification margin and aimed at completing the missing labels. First, the classification margin is proposed, and the label space is expanded by the label density. Then, the information entropy is used to measure the correlation between labels, and the label confidence matrix is constructed. The label confidence matrix is then unbalanced using the positive and negative label density, and the non-equilibrium label confidence matrix is used for label completion to obtain an informative label completion matrix. Finally, the kernel extreme learning machine and the label completion matrix are used for linear prediction. The experimental results show that the proposed algorithm has some advantages over other multi-label learning algorithms.

1. Introduction

A multi-label learning framework can effectively solve complex semantic problems in the real world by associating more objects with related labels [1,2]. Current multi-label learning methods have achieved certain classification effects. The classical multi-label *k*-nearest-neighbor (MLKNN) [3] method used maximum a posteriori probability to predict labels one by one. The multi-label kernel extreme learning machine (MLKELM) [4] method improved the performance of the classifier by dynamically adjusting the threshold function, introducing label correlations, and reducing the impact of tagged label results near the threshold. However, as data complexity increases rapidly, class imbalance issues become more apparent. Zhang et al. [5] proposed a cross-coupling aggregation (COCOA) method, which is a typical multi-label learning method that considers class imbalance. The goal is to use label correlations to study unbalances between classes. Sun et al. [6] proposed a two-stage multi-label hypernetwork (TSMLHN), which uses label correlations to solve the class imbalance problem in multi-label learning and exploits the correlation between common labels and unbalanced labels to improve learning performance. Xiao [7] proposed an unbalanced data fitting method based on two-stage hypergraph reduction. The combination of traditional sampling methods and classifiers is also common in multi-label class imbalance learning. Tsai et al. [8] proposed a cluster-based instance selection (CBIS) method that combines cluster analysis and instance selection. The cluster analysis module groups similar data samples of most class datasets into subclasses, while the instance selection component filters out non-representative data samples from each subclass [9].

The increase in data complexity will also elevate the labeling cost of the labels, and the labels will inevitably be missing [10,11]. Most current practices treat the indeterminate missing labels as unlabeled. Although these methods reduce the impact of the missing label, it is not arbitrary and not conducive to improving the accuracy of the algorithm. Therefore, it is necessary to perform an effective label completion for missing labels. Based on

* Corresponding author at: School of Computer and Information, Anqing Normal University, Anqing 246011, China.
E-mail address: chengyshaq@163.com (Y. Cheng).

https://doi.org/10.1016/j.asoc.2019.105924
1568-4946/

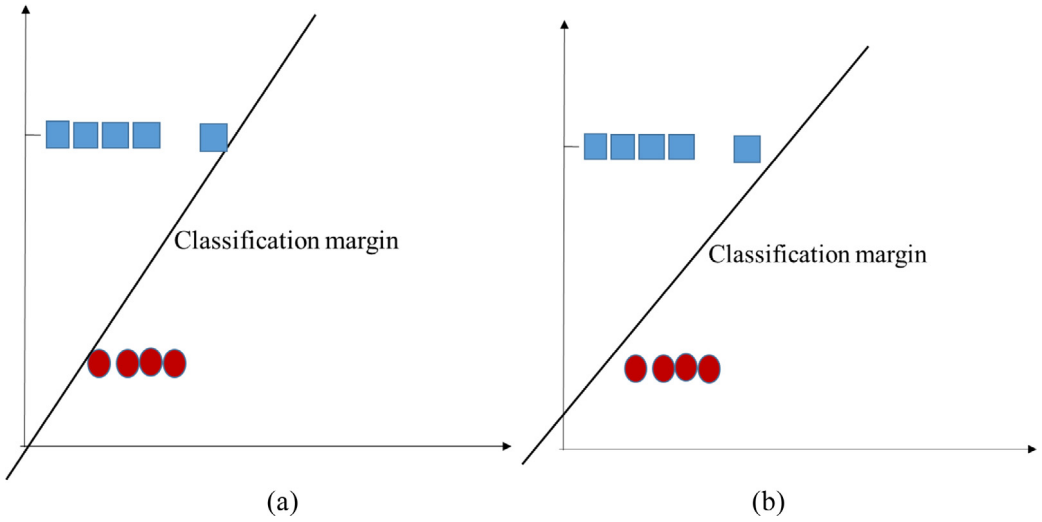

Fig. 1. Classification margin.

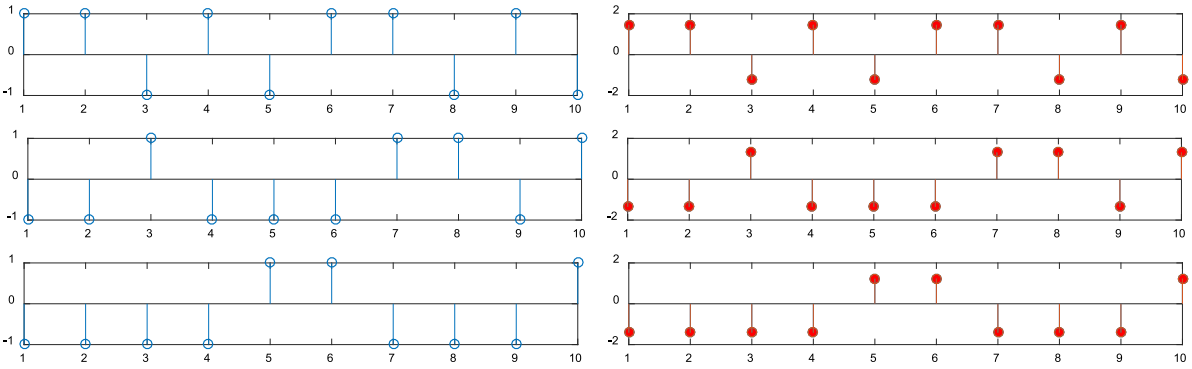

Fig. 2. Label interval.

this, Zhu et al. [12] proposed the global and local label correlation (Glocal) method to learn label correlations from the global and local parts to reduce the impact of missing labels. He et al. [13] proposed the joint multi-label classification and label correlations with missing labels and feature selection (MLMF) method to solve the missing label problem by combining label correlations, missing labels, and feature selection. Obviously, label correlations not only help improve the performance of the method but also improve the performance of label completion algorithms based on the correlation among labels.

As an effective method to measure uncertainty, information entropy is widely used in the study of label correlations [14–19]. Lee et al. [20] proposed a new multi-label learning method based on the classifier chains algorithm, and used the directed acyclic graph to model the correlation of labels; they also used conditional entropy to design a multi-label learning method and maximize the correlation between labels. Meanwhile, Xie et al. [21] proposed ordering methods based on the conditional entropy of labels. They generated a single order instead of multiple orders. Unlike in existing ordering methods, there is no need to train more classifiers than are in the classifier chain (CC) [22].

Obviously, using information entropy to measure label correlations has achieved certain effects. However, either the correlation among unlabeled labels is not considered, or the missing labels are all regarded as unlabeled labels, that is, the correlation between labeled and unlabeled labels is increased [23].

In fact, when we label a sample, the results are often affected by the threshold of the discriminant function, especially for labels near the threshold. Based on this, we integrate data distribution density, label density, and other information into the label correlations, and expand the classification margin, to solve the labeling quality of labels near the threshold. Therefore, this paper proposes a Missing **M**ulti-label Learning with **N**on-**E**quilibrium Based on **C**lassification **M**argin algorithm (MNECM). Considering the class imbalance and missing labels issues [24,25], we first introduce the classification margin and expand the original labels space by expanding the margin among labels by labels density. Then the labels space uses the non-equilibrium labels completion matrix algorithm. Finally, the kernel extreme learning machine is used for classification. The experimental results of the proposed algorithm on open benchmark multi-label datasets show that the MNECM algorithm has some advantages over other comparable multi-label learning algorithms.

The rest of the paper is organized as follows. Section 2 gives some basic notions related to multi-label learning and the rough entropy. Section 3 introduces the classification margin based on label density. The modeling of the non-equilibrium matrix is introduced and our proposed method for the missing multi-label classification of MNECM is proposed in Section 4. Section 5 introduces some experimental designs of MNECM. In Section 6, experimental results of MNECM on open multi-label datasets show that our algorithm is effective. A statistical hypothesis test

further proves our method. In Section 7, we sum up what has been discussed and point out further research directions.

2. Related work

2.1. Multi-label learning

Multi-label learning is a learning framework proposed for ambiguous objects that are prevalent in real life. Under this framework, a sample is composed of multiple features and multiple labels. The goal of learning is to map unknown instances to more correct labels. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}^{\mathrm{T}} \in \mathbf{R}^{N*d}$ denote that there are $\boldsymbol{N}$ samples and the number of features of each sample is d, $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\} \in \{-1, +1\}^{N*L}$ denotes a set of labels, where L denotes the number of labels corresponding to the samples. $\boldsymbol{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i)\}(\boldsymbol{x}_i \in \boldsymbol{X}, \boldsymbol{y}_i \in \boldsymbol{Y})$ denotes a training set. The goal of multi-label learning is to obtain a mapping relationship $f: \boldsymbol{X} \rightarrow \{-1, 1\}^L$ and perform label prediction on samples whose labels are unknown and whose characteristics are known.

2.2. Information entropy

In information theory, information entropy is an important tool to measure the uncertainty of random variables [26,27]. Suppose the sets $A = \{a_1, a_2, \ldots, a_m\}$, $B = \{b_1, b_2, \ldots, b_n\}$. Here $p(a), p(b)$ represent the prior probabilities of the elements in A and B. Then the information entropy of set A is expressed as follows:

$$H(A) = -\sum_{i=1}^{m} p(a_i) \log_2 p(a_i). \tag{1}$$

$H(A)$ is the information entropy of set A; the larger its value, the greater the uncertainty of the set. Then the conditional entropy of set B under the given constraints of set A is expressed as follows:

$$H(B|A) = -\sum_{i=1}^{m}\sum_{j=1}^{n} H(b_j|a_i). \tag{2}$$

where $H(b_j|a_i)$, the conditional information, is employed to describe the uncertainty of the element b_j with the appearing element a_i. The larger the value, the more uncertainty between a_i and b_j, and vice versa:

$$H(b_j|a_i) = -p(a_i b_j) \log_2 p(b_j|a_i). \tag{3}$$

$p(a_i b_j)$ represents the joint probability of element a_i with b_j, and $p(b_j|a_i)$ represents the conditional probability of b_j under the condition of a_i.

2.3. Extreme learning machine

The extreme learning machine (ELM) [28–30] algorithm is an effective single hidden layer feed-forward neural network learning algorithm. The learning parameters of the hidden layer in the ELM algorithm network structure are randomly selected so that it is only necessary to set the number of hidden layer network neurons. Finally, the output weight of the hidden layer is obtained by the least squares method, and no iteration is required for the network weight and offset. Therefore, compared with the traditional neural network algorithm, the ELM algorithm has the advantages of fast training speed and strong generalization ability.

Suppose there are N random samples $\boldsymbol{D} = \left\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)_{i=1}^{N}\right\}$, $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{in}\}^{\mathrm{T}}$, and $\boldsymbol{Y}_i = \left\{\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{im}\right\}^{\mathrm{T}}$. The number of hidden layer network neurons is L, and the activation function is $g(x)$. Then:

$$f_L(\boldsymbol{X}_j) = \sum_{i=1}^{L} \boldsymbol{\beta}_i g_i(\boldsymbol{X}_j) = \sum_{i=1}^{L} \boldsymbol{\beta}_i g_i(\boldsymbol{\omega}_i \cdot \boldsymbol{X}_j + b_i). \tag{4}$$

In Eq. (4), $\boldsymbol{\beta}_i = \left\{\boldsymbol{\beta}_{i1}, \boldsymbol{\beta}_{i2}, \ldots, \boldsymbol{\beta}_{im}\right\}^{\mathrm{T}}$ represents the ith layer output weight, $\boldsymbol{\omega}_i = \{\omega_{i1}, \omega_{i2}, \ldots, \omega_{im},\}$ represents the ith layer input weight, $\cdot$ represents the offset of the ith layer, and M represents the dot product. If $f_L(\boldsymbol{X}_j) - \boldsymbol{y}_j = 0$, there is no error between the output of the single hidden layer neural network and the real label. Then, there are:

$$\sum_{i=1}^{L} \boldsymbol{\beta}_i g_i(\boldsymbol{\omega}_i \cdot \boldsymbol{X}_j + b_i) = \boldsymbol{y}_j. \tag{5}$$

Eq. (5) can be simplified in matrix form as:

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{Y}. \tag{6}$$

3. Classification margin based on label density

3.1. Label density and classification margin

In the multi-label learning framework, $\boldsymbol{y} = +1$ is usually used to indicate that the instance has a label, and $\boldsymbol{y} = -1$ means that the instance does not have a label. To consider the missing label completion problem, we extended the original multi-label learning framework mode. Let $\hat{\boldsymbol{Y}} = \left\{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_n\right\} \in \{-1, 0, +1\}^{N*L}$, where -1 indicates that the instance does not have a label; this is called a negative label. Furthermore, 0 indicates that the instance label is missing, and this is called a 0 label; meanwhile, $+1$ indicates that the instance has a label, and this is called a positive label. For an imbalanced label density distribution, the original positive and negative labels are modified by the statistical label density, which makes the original positive and negative labels become continuous numerical labels. This transformation increases the discrimination between positive, negative, and 0 labels. It is easier for the classifier to distinguish the labels. The positive and negative label densities of a corresponding label are defined as follows:

$$\hat{Y}P_{pos}(j) = \frac{\sum_{i=1}^{N} \left|\hat{y}_{ij} = 1\right|}{\sum_{i=1}^{N}\sum_{l=1}^{L} \left|\hat{y}_{il} = 1\right|}, j = 1, 2, \ldots, L.$$
$$\hat{Y}P_{neg}(j) = \frac{\sum_{i=1}^{N} \left|\hat{y}_{ij} = -1\right|}{\sum_{i=1}^{N}\sum_{l=1}^{L} \left|\hat{y}_{il} = -1\right|}, j = 1, 2, \ldots, L. \tag{7}$$

$\hat{Y}P_{pos}(j)$ denotes the jth positive label density and $\hat{Y}P_{neg}(j)$ denotes the jth negative label density. At the same time, in multi-label learning, in order to acquire the mapping relationship $f: \boldsymbol{X} \rightarrow \{-1, 1\}^L$, the following error function is usually minimized as follows:

$$\min_{w} \|XW - Y\|_{\mathrm{F}}^2. \tag{8}$$

According to Eq. (8), we have $xw - y = 0$. In other words, we have $xw = y$. Owing to $\boldsymbol{Y} \in \{-1, +1\}$, the classification interval is 1. For multi-learning, different label has different classification interval. As shown in Fig. 1(a), we assume the classification interval is 1. For samples near the classification margin, it is difficult to distinguish whether it is a positive or negative label. As shown in Fig. 1(b), we could find that the classification interval is larger than 1 and we could easily distinguish positive and negative labels.

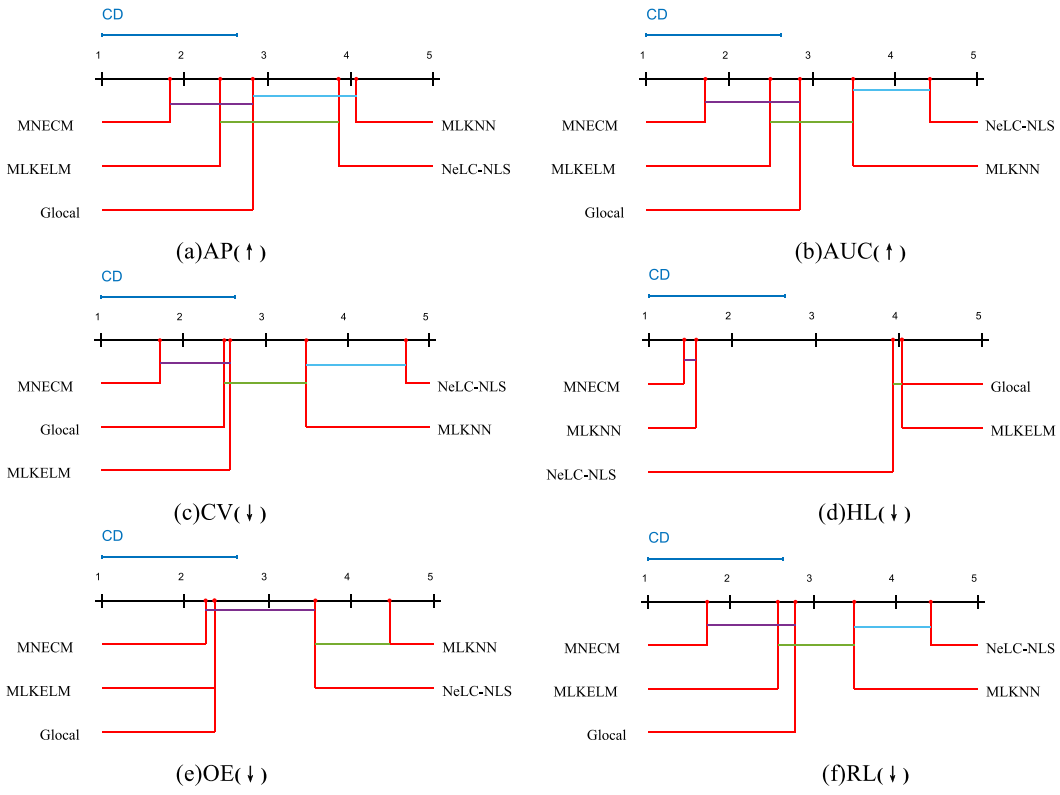

Fig. 3. Performance comparison of algorithms.

Table 1
Virtual label space data.

Number	Original label			Density label		
	y_1	y_2	y_3	y_1	y_2	y_3
1	+1	−1	−1	+1.4615	−1.3529	−1.4118
2	+1	−1	−1	+1.4615	−1.3529	−1.4118
3	−1	+1	−1	−1.2353	+1.3077	−1.4118
4	+1	−1	−1	+1.4615	−1.3529	−1.4118
5	−1	−1	+1	−1.2353	−1.3529	+1.2308
6	+1	−1	+1	+1.4615	−1.3529	+1.2308
7	+1	+1	−1	+1.4615	+1.3077	−1.4118
8	−1	+1	−1	−1.2353	+1.3077	−1.4118
9	+1	−1	−1	+1.4615	−1.3529	−1.4118
10	−1	+1	+1	−1.2353	+1.3077	+1.2308

3.2. Classification margin based on label density

In order to reduce the influence of the labeling results near the threshold, the idea of expanding the classification margin is further proposed. The original label space is transformed into the label density space after the introduction of the classification margin, which is defined as follows:

$$\hat{YP}(j) = \begin{cases} \hat{YP}_{pos}(j) + 1, \hat{y}_j = 1 \\ -\hat{YP}_{neg}(j) - 1, \hat{y}_j = -1, \end{cases} \quad j = 1, 2, \ldots, L. \tag{9}$$

$\hat{\boldsymbol{YP}}$ is the final label density matrix. It can be further found by Eq. (9) that the distance between each label and the margin 0 is enlarged by the label density and the labels are distinguished by this. To make this process easier to understand, we present virtual label space data in Table 1.

As shown in Fig. 2, the label interval shows the interval between the labels in Table 1, where 1–10 are the numbers of instances, blue indicates the original label, and red indicates the label after the change. It can be found that the distance between each label and the margin in the original label space is 1. After the label density information is added to the original label space, the distance between each label and the margin is greater than 1, which thus better distinguishes various labels and simplifies the problem of imbalanced label distribution density.

4. Missing multi-label learning with non-equilibrium

4.1. Modeling of non-equilibrium labels completion matrix

With the rapid increase in the amount of data, the dimensions of data are also increasing, resulting in an increase in the cost of labeling. Therefore, in multi-label datasets, a large number of missing labels are inevitable. The number of unlabeled items of a sample in the real world is much larger than that of annotated ones. Although most algorithms consider the label corrections and the meaning of the unknown labels, missing labels are ignored. Undoubtedly, there may be a lot of valuable information in the unlabeled labels in the sample labels set. For example, a document may include the "*Apple*" label instead of the "*Mobile*" label, but the missing label "*Mobile*" often determines the overall labels tendencies. Based on this, this paper uses information entropy to study the relationship among positive, negative, and 0 labels. The corresponding labels confidence matrix can be expressed as follows:

$$\boldsymbol{a}_{ij} = \frac{1}{H\left(\bar{l}_j|l_i\right)}, \boldsymbol{b}_{ij} = \frac{1}{H\left(l_j|\bar{l}_i\right)}, \boldsymbol{c}_{ij}^{+} = \frac{1}{H\left(l_j^0|l_i\right)}, \boldsymbol{c}_{ij}^{-} = \frac{1}{H\left(l_j^0|\bar{l}_i\right)}. \tag{10}$$

where $\bar{l}_i$, l_j^0 and l_i denote $\hat{\boldsymbol{y}} = -1, \hat{\boldsymbol{y}} = 0, \hat{\boldsymbol{y}} = +1$, $i = 1, 2, \ldots, L$, $j = 1, 2, \ldots, L$, $i \neq j$. $\boldsymbol{a}_{ij}$ is the effect of the positive labels on the negative labels, $\boldsymbol{b}_{ij}$ is the effect of the negative labels on the positive labels, $\boldsymbol{c}_{ij}^{+}$ is the effect of the positive labels on the 0 labels, and $\boldsymbol{c}_{ij}^{-}$ is the effect of the negative labels on the 0 labels.

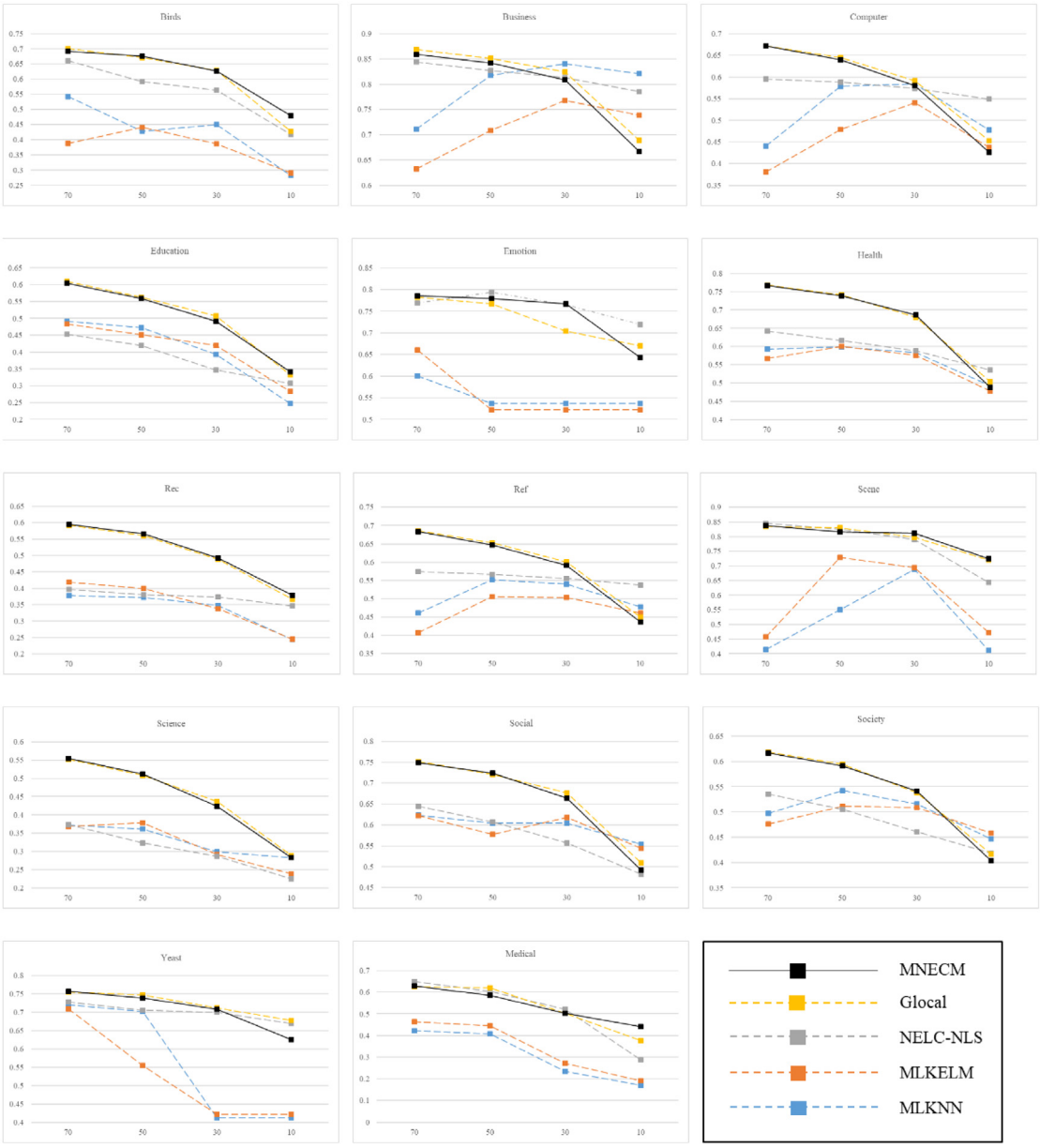

Fig. 4. AP(↑) results on all 14 missing labels datasets.

Researchers have found that unlabeled labels often determine the overall label orientation of the labels set, and therefore, increasing the weight of such labels facilitates the completion of missing labels. We previously proposed the related non-equilibrium parameter method [31], but this method only considers the correlation between positive and negative labels. In addition, this algorithm is very sensitive to non-equilibrium parameters and needs to be constantly adjusted to achieve the desired result. Based on this, the present paper uses the labels density to reduce the dependence of the algorithm on the non-equilibrium parameters and improve the robustness of the model. According to Eq. (7), the positive and negative labels densities can be expressed as follows:

$$\hat{Y}P_{pos} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{L}\left|\hat{y}_{il}=1\right|}{N\times L}, j=1,2,\ldots,L.$$
$$\hat{Y}P_{neg} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{L}\left|\hat{y}_{il}=-1\right|}{N\times L}, j=1,2,\ldots,L. \quad (11)$$

where YP_{pos} represents the positive labels density and YP_{neg} represents the negative labels density. We combine Eqs. (10) and (11) to construct a non-equilibrium labels confidence matrix:

$$Conf_{ij} = -\hat{Y}P_{neg}(\boldsymbol{a}_{ij}+\boldsymbol{c}_{ij}^{+}) + \hat{Y}P_{pos}(\boldsymbol{b}_{ij}+\boldsymbol{c}_{ij}^{-}). \quad (12)$$

Finally, the missing labels are complemented by the non-equilibrium labels confidence matrix:

$$\boldsymbol{Y}^{*} = \hat{\boldsymbol{Y}}\boldsymbol{P}\times\boldsymbol{Conf}. \quad (13)$$

4.2. Missing multi-label learning with non-equilibrium based on classification margin

The information entropy method is used to measure the strength of the relationship among labels. At the same time, the class imbalance is considered, and the label density information is introduced. When the missing label is complemented, the non-equilibrium labels confidence matrix is obtained according to the positive and negative labels density and the labels confidence

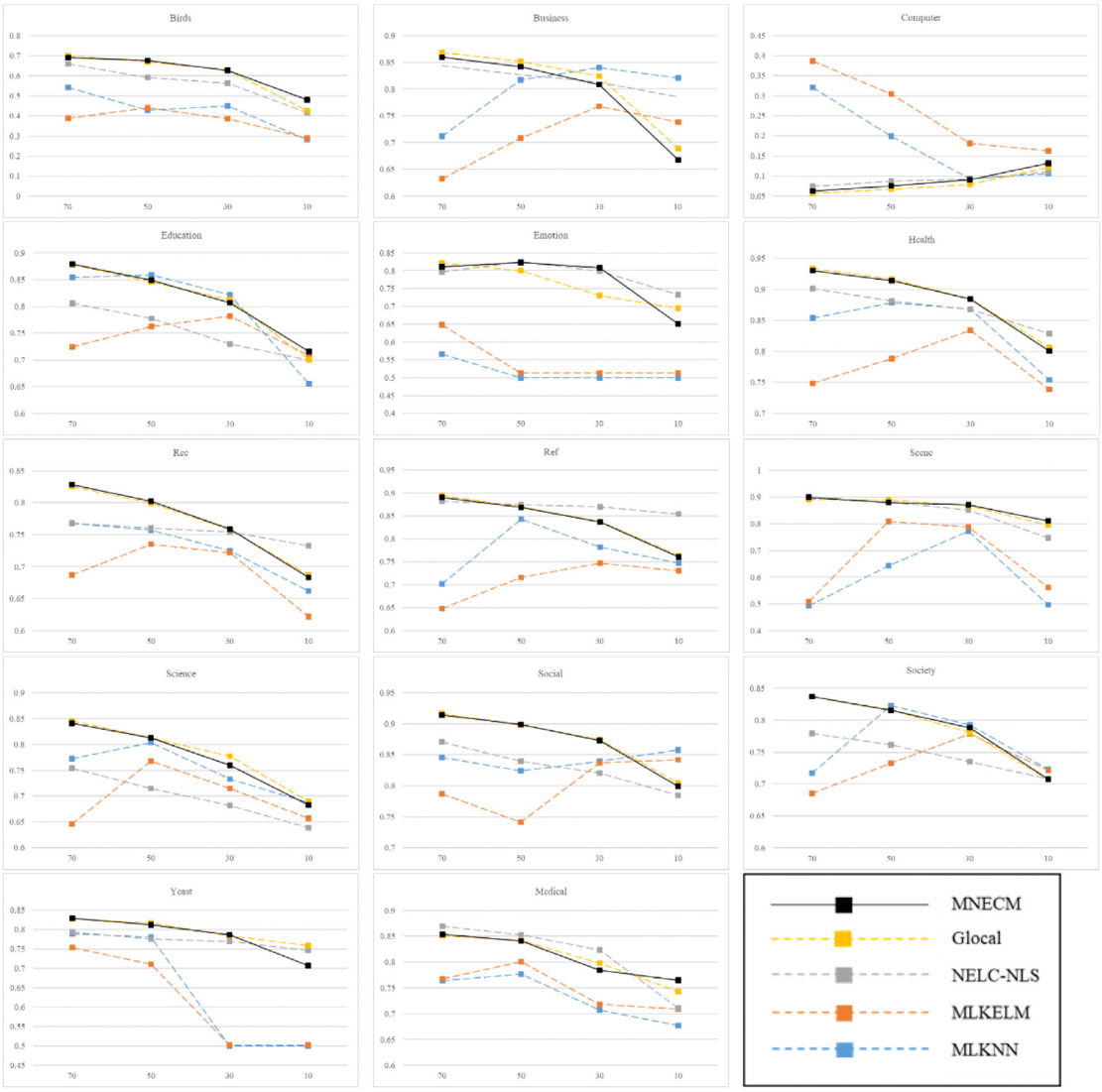

Fig. 5. AUC(↑) results on all 14 missing labels datasets.

matrix, and then the missing labels are complemented. Finally, the labels completion matrix is input into the kernel extreme learning machine [32,33] (KELM, https://www.ntu.edu.sg/home/egbhuang/elm_codes.html).

According to Eqs. (4) and (6), the output function $f_l(x)$ of the ELM can be expressed as:

$$f_l(x) = \boldsymbol{H}\beta = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} \begin{bmatrix} \beta_1^{\mathrm{T}} \\ \vdots \\ \beta_L^{\mathrm{T}} \end{bmatrix}. \tag{14}$$

According to Eq. (6), we can get:

$$\beta = \boldsymbol{H}^{\dagger}\boldsymbol{Y}^{*}$$
$$s.t: \boldsymbol{H}^{\dagger} = \begin{cases} \left(\boldsymbol{H}^{\dagger}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}} \\ \boldsymbol{H}^{T}\left(\boldsymbol{H}^{\dagger}\boldsymbol{H}\right). \end{cases} \tag{15}$$

If $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ is a nonsingular matrix, $\boldsymbol{H}^{\dagger} = \left(\boldsymbol{H}^{\dagger}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}$; or, if $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ is a nonsingular matrix, $\boldsymbol{H}^{\dagger} = \boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}^{\dagger}\boldsymbol{H}\right)^{-1}$, where $\boldsymbol{H}^{\dagger}$ is the Moore–Penrose generalized inverse matrix of $\boldsymbol{H}$. According to the Ridge Regression Theory [34], it is recommended that we add the regularization term $\boldsymbol{C}$ to the diagonal line of $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$ or $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$, which will improve the stability and generalization ability of the algorithm. Then, the minimum goal of Eq. (14) is

$$minL_f = \|\beta\|^2 + \boldsymbol{C}\sum_{i=1}^{N}\|\xi_i\|^2$$
$$s.t: \xi_i = Y_i - f_l(x_i), i = 1, 2, \ldots, N. \tag{16}$$

According to Karush–Kuhn–Tucker, the hidden layer output weight $\boldsymbol{\beta}$ is expressed as follows:

$$\boldsymbol{\beta} = \boldsymbol{H}^{\mathrm{T}}\left(\frac{I}{\boldsymbol{C}} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{Y}^{*}. \tag{17}$$

At this point, the multi-label input function can be expressed as follows:

$$f_l(x) = \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\left(\frac{\boldsymbol{I}}{\boldsymbol{C}} + \boldsymbol{H}\boldsymbol{H}^{T}\right)^{-1}\boldsymbol{Y}^{*}. \tag{18}$$

In the traditional ELM algorithm, the calculation results are easily affected by the random set value. To solve this problem, the kernel matrix is introduced and the kernel extreme learning machine is used for classification:

$$\boldsymbol{\Omega}_{\mathrm{ELM}} = \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}: \boldsymbol{\Omega}_{\mathrm{ELM}(i,j)} = K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right).$$

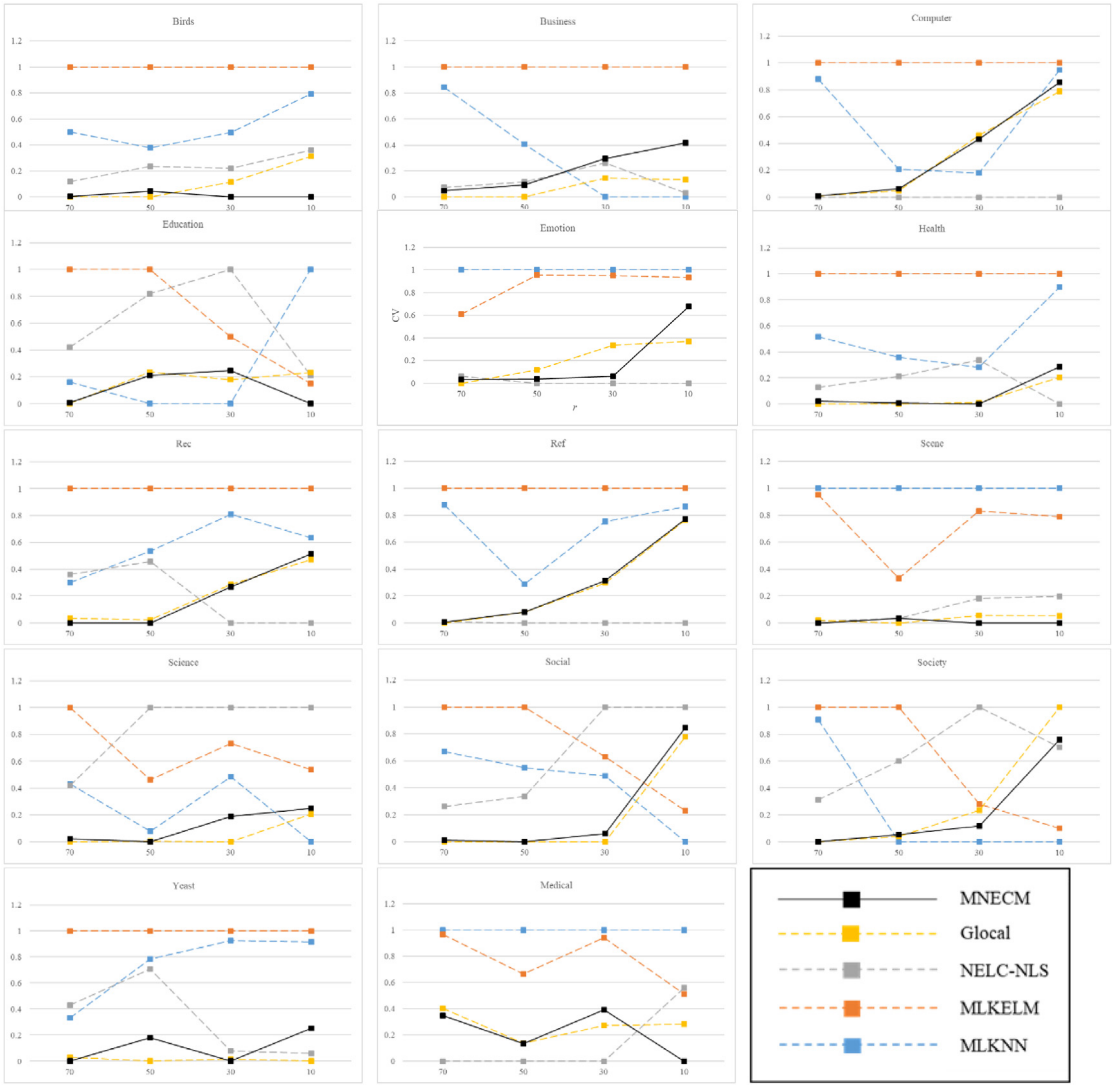

Fig. 6. CV(↓) results on all 14 missing labels datasets.

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = exp\left(-\gamma \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|\right). \tag{19}$$

According to Eq. (19), we can get $\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}} = \begin{bmatrix} K\left(x, x_2\right) \\ \vdots \\ K\left(x, x_N\right) \end{bmatrix}^{\mathrm{T}}$. By combining Eqs. (18) and (19), the objective function of the multi-label kernel extreme learning machine is expressed as

$$\begin{aligned} f_L\left(\boldsymbol{x}\right) &= h\left(\boldsymbol{x}\right)\boldsymbol{H}^{\mathrm{T}}\left(\frac{I}{\boldsymbol{C}} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{Y}^{*} \\ &= \begin{bmatrix} K\left(\boldsymbol{x}, \boldsymbol{x}_2\right) \\ \vdots \\ K\left(\boldsymbol{x}, \boldsymbol{x}_N\right) \end{bmatrix}\left(\frac{I}{\boldsymbol{C}} + \Omega_{ELM}\right)^{-1}\boldsymbol{Y}^{*}. \end{aligned} \tag{20}$$

where $\boldsymbol{C}$ represents the regularization term, and I represents the unit matrix.

Therefore, our proposed method is described in Algorithm 1:

Algorithm 1: Missing Multi-label Learning with Non-equilibrium Based on Classification Margin

Input: training set: $\boldsymbol{D} = \left\{\left(\boldsymbol{X}_i, \bar{\boldsymbol{Y}}_i\right)_{i=1}^{N}\right\}$, testing set: $\boldsymbol{D}^{*} = \left\{\left(\boldsymbol{X}_j^{*}\right)_{j=1}^{N^{*}}\right\}$, regularization term: $\boldsymbol{C}$, kernel parameter: γ
Output: prediction labels: $\boldsymbol{Y}_{D^{*}}$

```
for j in L
    identify Ŷ_pos(j), Ŷ_neg(j) according to Eq.(7)
end
compute Ŷ according to Eq.(9)
compute Ŷ_pos, Ŷ_neg according to Eq.(11)
for y_i, y_j in Ŷ
    if i ≠ j
        identify a_ij, b_ij, c_ij^+, c_ij^- according to Eq. (10)
    else if i = j
        a_ij = 1, b_ij = 1, c_ij^+ = 1, c_ij^- = 1
    end
end
compute Conf according to Eq.(12)
compute Y* according to Eq.(13)
put C, γ, D*, X, Y* into KELM, compute prediction labels f_L(x) according to Eq. (20)
if f_L(x) > 0
    Y_D* = 1
else
    Y_D* = −1
```

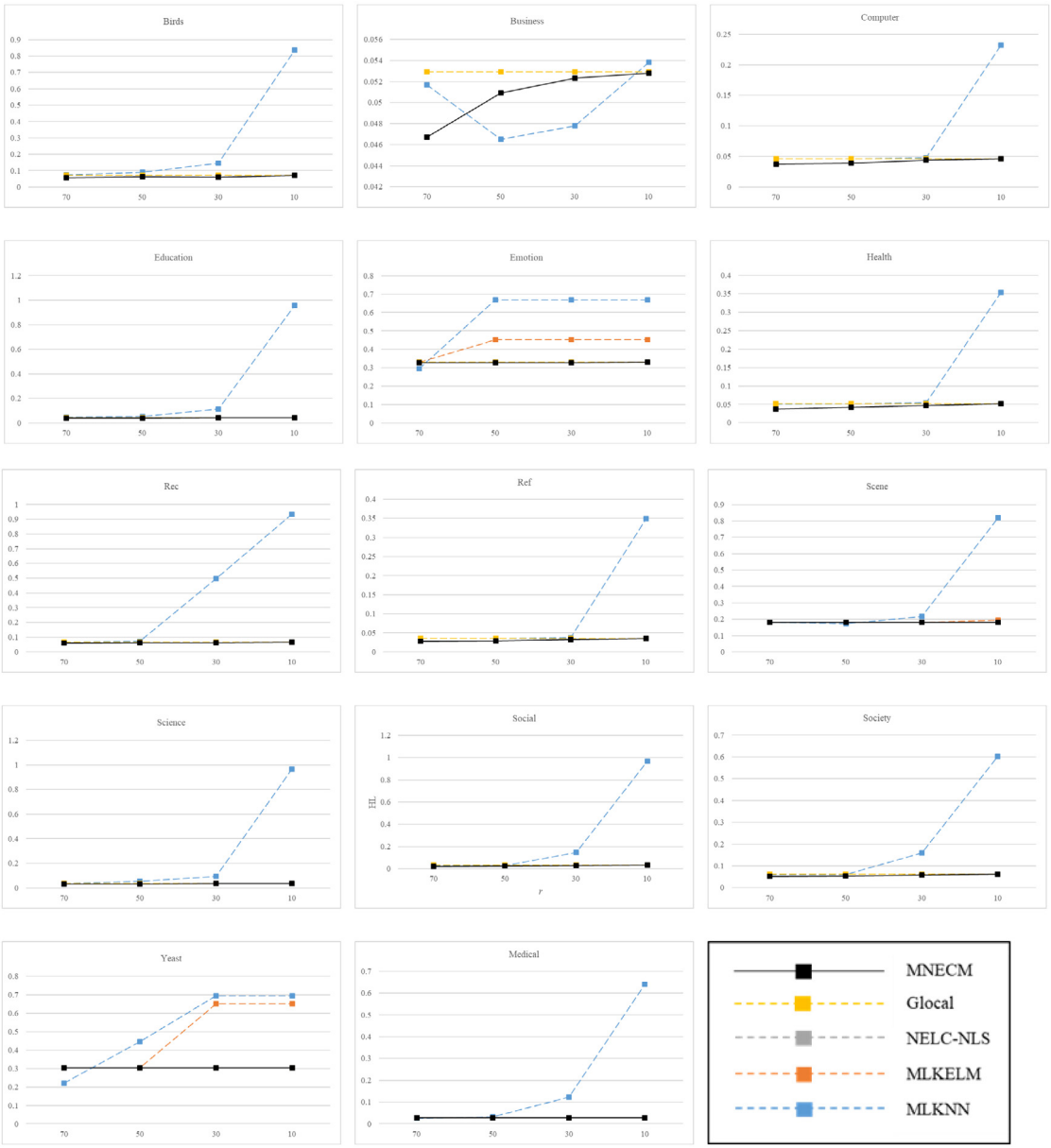

Fig. 7. HL(↓) results on all 14 missing labels datasets.

5. Experimental design

5.1. Datasets

In order to illustrate the effectiveness of the MNECM algorithm, we choose 14 multi-label datasets from *Yahoo Web Pages* (http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar) and *Mulan* (http://mulan.sourceforge.net/datasets-mlc.html). The datasets contain multiple fields such as *text*, *audio*, and *music*. See Table 2 for details. Since the 14 datasets do not contain missing labels, we use the random missing method to obtain the missing labels datasets.

5.2. Environment and evaluation metrics

The experiment is conducted on a computer equipped with a Windows 7 Operation System, Intel®Core$_{(TM)}$ i5-2525M 2.50 GHz CPU and uses Matlab2016a for the operation of experimental codes. We choose six commonly applied evaluation criteria,

Table 2
Detailed descriptions of multi-label datasets.

Datasets	Training instance	Testing instance	Labels	Features	Fields
Birds	322	323	20	260	Audio
Business	2000	3000	30	438	Text
Computer	2000	3000	33	681	Text
Education	2000	3000	33	550	Text
Emotion	391	202	6	72	Music
Health	2000	3000	32	612	Text
Rec	2000	3000	22	606	Text
Ref	2000	3000	33	793	Text
Scene	1211	1196	6	294	Text
Science	2000	3000	40	743	Text
Social	2000	3000	39	1047	Text
Society	2000	3000	27	636	Text
Yeast	1499	918	14	103	Biology
Medical	333	645	45	1449	Text

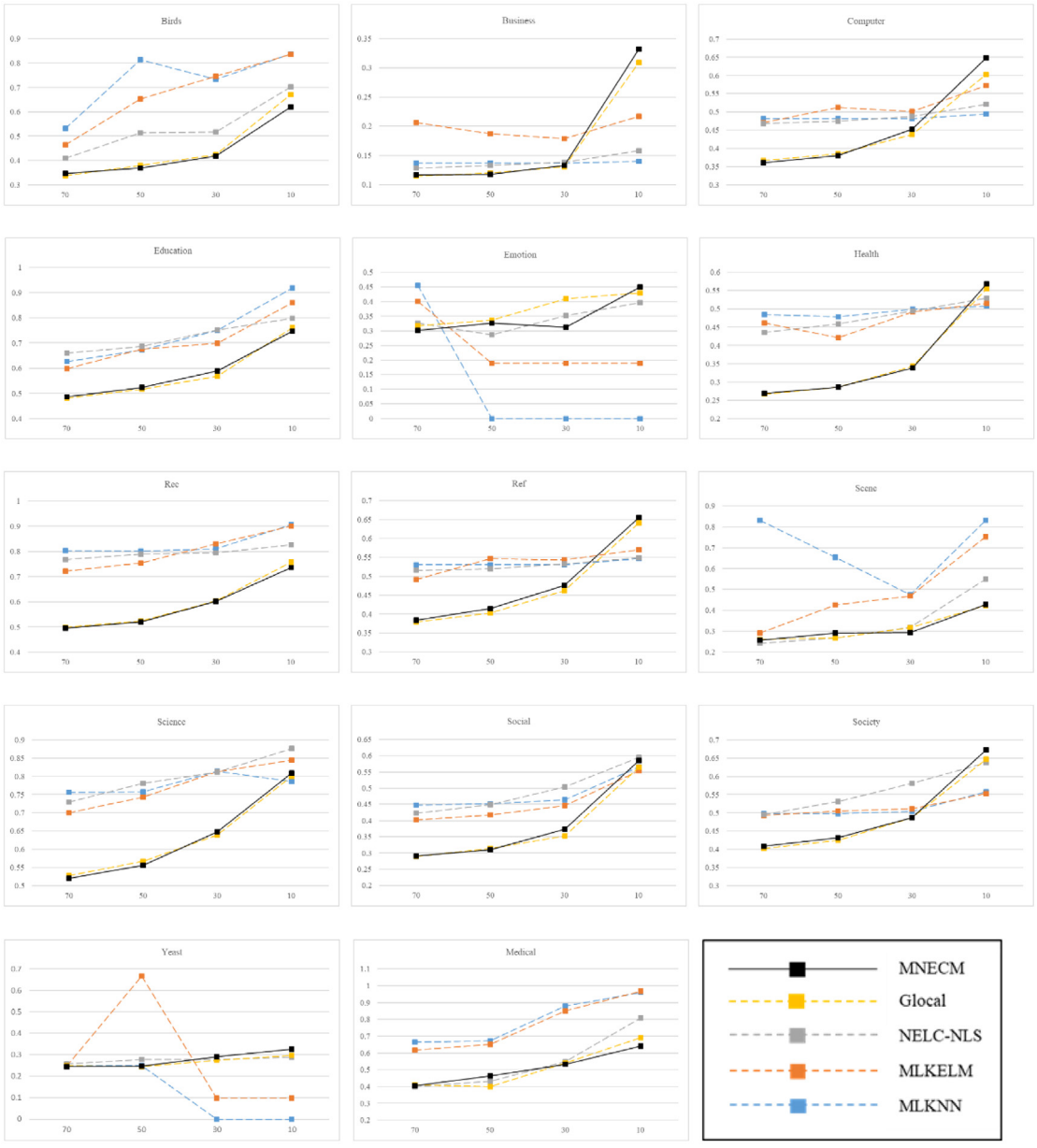

Fig. 8. OE(↓) results on all 14 missing labels datasets.

namely, average precision, average area under the receiver operating characteristic curve, coverage, hamming loss, one-error, and ranking loss [35] to evaluate the MLLA performance. The criteria are abbreviated as AP↑, AUC↑, CV↓, HL↓, OE↓, and RL↓, for convenience, where ↑ indicates that the higher the value, the better, and ↓ indicates that the lower the value, the better. Suppose $h(\cdot)$ is the multi-label classifier; $f(\cdot,\cdot)$ is the prediction function; $rank_f$ is the ranking function; and $D = \{(x_{it}, Y_{il}|1 \le t \le d, 1 \le i \le n, 1 \le l \le L)\}$ is the MLD. The formal equations of these criteria are defined as follows:

(1) AP: Evaluates the average score of correct labels ranked in the specific label $y \in Y_i$:

$$AP_D(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\sum_{y\in Y}\frac{\left|\left\{rank_f\left(x_i, y'\right) \le rank_f\left(x_i, y\right), y' \in Y_i\right\}\right|}{rank_f\left(x_i, y\right)}. \tag{21}$$

(2) AUC: Calculates the score of the positive labels ranking higher than the negative labels, and takes the average of all the labels:

$$AUC_D(f) = \frac{1}{L}\sum_{j=1}^{L}\frac{\left\{\left|f_j\left(x_{i'}\right) \ge f_j\left(x_{i''}\right)\right|, \left(x_{i'}, x_{i''}\right) \in Y_j \times \overline{Y}_j\right\}}{\left|Y_j\right|\left|\overline{Y}_j\right|}. \tag{22}$$

(3) CV: An indicator to measure the average step number for traversing all related labels of the given sample:

$$CV_D(f) = \frac{1}{n}\sum_{i=1}^{n} max_{y\in Y_i} rank_f\left(x_i, y\right) - 1. \tag{23}$$

(4) HL: An indicator to measure real labels in a single label and wrong matches of prediction labels of the given sample:

$$HL_D(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y}\left|h\left(x_i\right) \ne Y_i\right|. \tag{24}$$

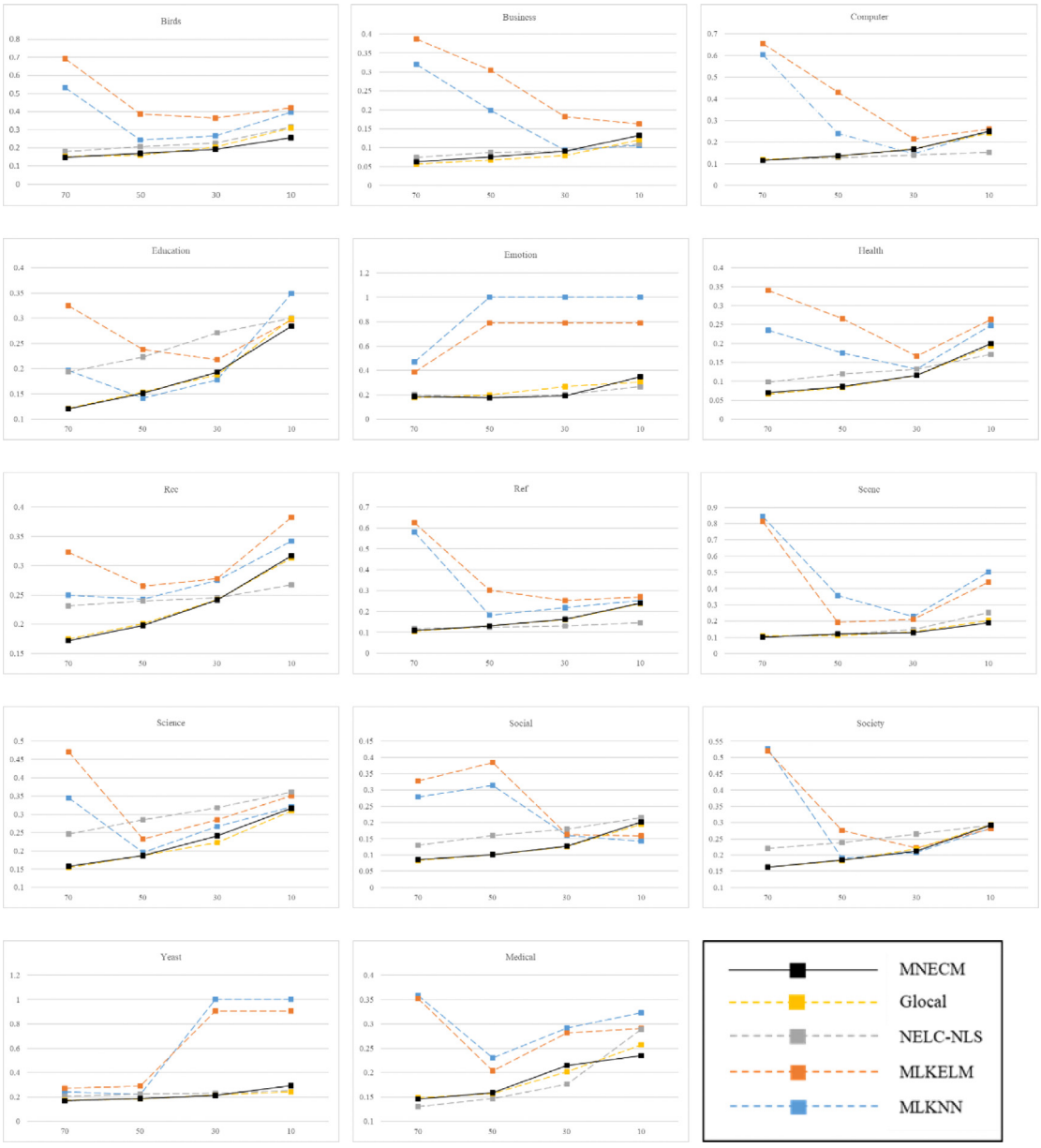

Fig. 9. RL(↓) results on all 14 missing labels datasets.

(5) OE: Evaluates the occurrence of labels when top-ranking labels are not correct:

$$OE_D(f) = \frac{1}{n}\sum_{i=1}^{n}\left[\left[\arg max_{y\in Y} f(x_i, y)\right] \notin Y_i\right]. \tag{25}$$

(6) RL: An indicator to evaluate the circumstances where the ranking of uncorrelated labels of a given sample is lower than that of correlated labels:

$$RL_D(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|\left|\overline{Y_i}\right|} \times \left|\left\{(y_1, y_2)\,|f\left((x_i, y_1), (y_1, y_2) \in Y_i \times \overline{Y_i}\right)\right\}\right|. \tag{26}$$

5.3. Choice of algorithms and the configuration of related parameters

In order to verify the performance of the proposed algorithm, the MNECM algorithm is compared with 4 multi-label mainstream classification algorithms: MLKNN, MLKELM, NeLC-NLS and Glocal. Since MLKNN and MLKELM cannot directly handle missing labels, the missing labels must be complemented first. According to the literature [36], the missing label is completed using the MAXIDE algorithm, that is, the 0 label is complemented by positive and negative labels.

(1) MLKNN [3]: The nearest neighbor k is set to 10 and the smoothing parameter s is set to 1.
(2) MLKELM [4]: The kernel parameter is set to 1, the regularization parameter is set to 1 and the kernel function selects RBF.
(3) NeLC-NLS [23]: The kernel parameter is set to 1, the regularization parameter is set to 1, the kernel function selects RBF, the non-equilibrium parameter is set to 1, the nearest neighbor k is set to 10 and the smoothing parameter s is set to 1.
(4) Glocal [12]: The number of clusters is set to 3, the regularization parameters λ_2, λ_3, andλ_4 are set to 10^{-3}, 0.125, and 0.125, respectively, and the latent matrix dimension is set to 20.

Table 3
AP(↑) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$0.6684_{(5)}$	$0.7425_{(2)}$	$0.6894_{(4)}$	$\mathbf{0.7443}_{(\mathbf{1})}$	$0.7413_{(3)}$
Business	$0.8683_{(5)}$	$0.8866_{(3)}$	$0.8685_{(4)}$	$0.8870_{(2)}$	$\mathbf{0.8871}_{(\mathbf{1})}$
Computer	$0.6111_{(5)}$	$0.7065_{(3)}$	$0.6123_{(4)}$	$0.7064_{(2)}$	$\mathbf{0.7073}_{(\mathbf{1})}$
Education	$0.5181_{(5)}$	$0.6468_{(2)}$	$0.5191_{(4)}$	$0.6467_{(3)}$	$\mathbf{0.6486}_{(\mathbf{1})}$
Emotion	$0.8050_{(2)}$	$0.7867_{(5)}$	$0.7921_{(4)}$	$0.8031_{(3)}$	$\mathbf{0.8077}_{(\mathbf{1})}$
Health	$0.6429_{(5)}$	$0.7913_{(2)}$	$0.6485_{(4)}$	$0.7906_{(3)}$	$\mathbf{0.7924}_{(\mathbf{1})}$
Rec	$0.3966_{(5)}$	$\mathbf{0.6351}_{(\mathbf{1})}$	$0.4096_{(4)}$	$0.6321_{(3)}$	$0.6349_{(2)}$
Ref	$0.5855_{(4)}$	$\mathbf{0.7177}_{(\mathbf{1})}$	$0.5847_{(5)}$	$0.7173_{(3)}$	$0.7174_{(2)}$
Scene	$0.8591_{(2)}$	$0.8482_{(3)}$	$\mathbf{0.8706}_{(\mathbf{1})}$	$0.8481_{(4)}$	$0.8475_{(5)}$
Science	$0.4191_{(5)}$	$0.6035_{(2)}$	$0.4294_{(4)}$	$0.6017_{(3)}$	$\mathbf{0.6040}_{(\mathbf{1})}$
Social	$0.6821_{(4)}$	$0.7717_{(2)}$	$0.6810_{(5)}$	$0.7715_{(3)}$	$\mathbf{0.7726}_{(\mathbf{1})}$
Society	$0.5778_{(4)}$	$0.6433_{(2)}$	$0.5719_{(5)}$	$0.6431_{(3)}$	$\mathbf{0.6441}_{(\mathbf{1})}$
Yeast	$0.7499_{(4)}$	$\mathbf{0.7660}_{(\mathbf{1})}$	$0.7395_{(5)}$	$0.7598_{(2.5)}$	$0.7598_{(2.5)}$
Medical	$0.7256_{(2)}$	$0.7163_{(5)}$	$\mathbf{0.7600}_{(\mathbf{1})}$	$0.7177_{(4)}$	$0.7182_{(3)}$
Average ranking	4.0714	2.4286	3.8571	2.8214	**1.8214**

Table 4
AUC(↑) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$0.8730_{(4)}$	$0.8895_{(2)}$	$0.8684_{(5)}$	$\mathbf{0.8926}_{(\mathbf{1})}$	$0.8874_{(3)}$
Business	$0.9560_{(4)}$	$0.9605_{(2)}$	$0.9484_{(5)}$	$0.9604_{(3)}$	$\mathbf{0.9638}_{(\mathbf{1})}$
Computer	$0.8999_{(4)}$	$0.9184_{(2)}$	$0.8961_{(5)}$	$0.9182_{(3)}$	$\mathbf{0.9243}_{(\mathbf{1})}$
Education	$0.9011_{(4)}$	$0.9138_{(3)}$	$0.8856_{(5)}$	$0.9142_{(2)}$	$\mathbf{0.9235}_{(\mathbf{1})}$
Emotion	$\mathbf{0.8472}_{(\mathbf{1})}$	$0.8290_{(4)}$	$0.8245_{(5)}$	$0.8392_{(3)}$	$0.8432_{(2)}$
Health	$0.9296_{(4)}$	$0.9509_{(3)}$	$0.9278_{(5)}$	$0.9514_{(2)}$	$\mathbf{0.9565}_{(\mathbf{1})}$
Rec	$0.7870_{(4)}$	$0.8617_{(2)}$	$0.7860_{(5)}$	$0.8602_{(3)}$	$\mathbf{0.8665}_{(\mathbf{1})}$
Ref	$0.9006_{(4)}$	$0.9270_{(3)}$	$0.8980_{(5)}$	$0.9271_{(2)}$	$\mathbf{0.9341}_{(\mathbf{1})}$
Scene	$0.9142_{(2)}$	$0.9125_{(3)}$	$\mathbf{0.9195}_{(1)}$	$0.9116_{(4)}$	$0.9084_{(5)}$
Science	$0.8528_{(4)}$	$0.8897_{(2)}$	$0.8285_{(5)}$	$0.8893_{(3)}$	$\mathbf{0.8959}_{(\mathbf{1})}$
Social	$0.9282_{(4)}$	$0.9398_{(3)}$	$0.9094_{(5)}$	$0.9399_{(2)}$	$\mathbf{0.9444}_{(\mathbf{1})}$
Society	$0.8524_{(4)}$	$0.8697_{(2)}$	$0.8328_{(5)}$	$0.8695_{(3)}$	$\mathbf{0.8756}_{(\mathbf{1})}$
Yeast	$0.8232_{(4)}$	$\mathbf{0.8403}_{(\mathbf{1})}$	$0.8055_{(5)}$	$0.8364_{(3)}$	$0.8384_{(2)}$
Medical	$0.9423_{(2)}$	$0.9355_{(4)}$	$\mathbf{0.9517}_{(\mathbf{1})}$	$0.9320_{(5)}$	$0.9374_{(3)}$
Average ranking	3.5000	2.5000	4.4286	2.8571	**1.7143**

Table 5
CV(↓) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$3.3406_{(4)}$	$3.1146_{(2)}$	$3.6161_{(5)}$	$\mathbf{3.0650}_{(\mathbf{1})}$	$3.1517_{(3)}$
Business	$2.4150_{(4)}$	$2.3963_{(2)}$	$2.7317_{(5)}$	$2.3987_{(3)}$	$\mathbf{2.2153}_{(\mathbf{1})}$
Computer	$4.6993_{(4)}$	$4.0440_{(2)}$	$4.9340_{(5)}$	$4.0540_{(3)}$	$\mathbf{3.7633}_{(\mathbf{1})}$
Education	$4.1530_{(4)}$	$4.1497_{(3)}$	$4.8823_{(5)}$	$4.1257_{(2)}$	$\mathbf{3.6330}_{(\mathbf{1})}$
Emotion	$\mathbf{1.7822}_{(\mathbf{1})}$	$1.9257_{(4)}$	$1.9307_{(5)}$	$1.8614_{(3)}$	$1.8315_{(2)}$
Health	$3.6413_{(4)}$	$3.2567_{(3)}$	$3.7797_{(5)}$	$3.2303_{(2)}$	$\mathbf{2.8773}_{(\mathbf{1})}$
Rec	$5.4993_{(4)}$	$4.0963_{(2)}$	$5.5797_{(5)}$	$4.1420_{(3)}$	$\mathbf{3.9477}_{(\mathbf{1})}$
Ref	$3.7863_{(4)}$	$3.1160_{(3)}$	$3.9047_{(5)}$	$3.1087_{(2)}$	$\mathbf{2.8057}_{(\mathbf{1})}$
Scene	$0.5343_{(2)}$	$0.5385_{(3)}$	$\mathbf{0.5092}_{(\mathbf{1})}$	$0.5393_{(4)}$	$0.5543_{(5)}$
Science	$7.2867_{(4)}$	$6.0403_{(2)}$	$8.4523_{(5)}$	$6.0583_{(3)}$	$\mathbf{5.6737}_{(\mathbf{1})}$
Social	$3.7820_{(4)}$	$3.4460_{(3)}$	$4.6997_{(5)}$	$3.4410_{(2)}$	$\mathbf{3.1403}_{(\mathbf{1})}$
Society	$5.7750_{(4)}$	$5.6203_{(2)}$	$6.6383_{(5)}$	$5.6253_{(3)}$	$\mathbf{5.3030}_{(\mathbf{1})}$
Yeast	$6.4423_{(4)}$	$\mathbf{6.1176}_{(\mathbf{1})}$	$6.9074_{(5)}$	$6.2756_{(3)}$	$6.2092_{(2)}$
Medical	$3.5039_{(2)}$	$3.9209_{(4)}$	$\mathbf{3.0078}_{(\mathbf{1})}$	$4.0496_{(5)}$	$3.7442_{(3)}$
Average ranking	3.5000	2.5714	4.7143	2.5000	**1.7143**

Table 6
HL(↓) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$0.0605_{(2)}$	$0.0718_{(4)}$	$0.0718_{(4)}$	$0.0718_{(4)}$	$\mathbf{0.0570}_{\mathbf{1}}$
Business	$\mathbf{0.0276}_{(\mathbf{1})}$	$0.0529_{(4.5)}$	$0.0528_{(3)}$	$0.0529_{(4.5)}$	$0.0396_{(2)}$
Computer	$0.0433_{(2)}$	$0.0461_{(4)}$	$0.0461_{(4)}$	$0.0461_{(4)}$	$\mathbf{0.0347}_{(\mathbf{1})}$
Education	$0.0430_{(2)}$	$0.0442_{(4)}$	$0.0442_{(4)}$	$0.0442_{(4)}$	$\mathbf{0.0373}_{(\mathbf{1})}$
Emotion	$\mathbf{0.2063}_{(\mathbf{1})}$	$0.3292_{(4)}$	$0.3292_{(4)}$	$0.3292_{(4)}$	$0.3284_{(2)}$
Health	$0.0487_{(2)}$	$0.0518_{(4)}$	$0.0518_{(4)}$	$0.0518_{(4)}$	$\mathbf{0.0345}_{(\mathbf{1})}$
Rec	$0.0643_{(2)}$	$0.0650_{(4)}$	$0.0650_{(4)}$	$0.0650_{(4)}$	$\mathbf{0.0568}_{(\mathbf{1})}$
Ref	$0.0350_{(2)}$	$0.0357_{(4)}$	$0.0357_{(4)}$	$0.0357_{(4)}$	$\mathbf{0.0259}_{(\mathbf{1})}$
Scene	$\mathbf{0.0874}_{(\mathbf{1})}$	$0.1810_{(4)}$	$0.1810_{(4)}$	$0.1810_{(4)}$	$0.1788_{(2)}$
Science	$0.0354_{(2)}$	$0.0356_{(4)}$	$0.0356_{(4)}$	$0.0356_{(4)}$	$\mathbf{0.0312}_{(\mathbf{1})}$
Social	$0.0267_{(2)}$	$0.0331_{(4)}$	$0.0331_{(4)}$	$0.0331_{(4)}$	$\mathbf{0.0205}_{(\mathbf{1})}$
Society	$\mathbf{0.0577}_{(\mathbf{1})}$	$0.0624_{(4)}$	$0.0624_{(4)}$	$0.0624_{(4)}$	$0.0580_{(2)}$
Yeast	$\mathbf{0.1995}_{(\mathbf{1})}$	$0.3038_{(4)}$	$0.3038_{(4)}$	$0.3038_{(4)}$	$0.3031_{(2)}$
Medical	$\mathbf{0.0187}_{(\mathbf{1})}$	$0.0276_{(4)}$	$0.0276_{(4)}$	$0.0276_{(4)}$	$0.0275_{(2)}$
Average ranking	1.5714	4.0357	3.9286	4.0357	**1.4286**

Table 7
OE(↓) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$0.4396_{(5)}$	$\mathbf{0.3127}_{(\mathbf{2})}$	$0.4025_{(4)}$	$\mathbf{0.3127}_{(\mathbf{2})}$	$\mathbf{0.3127}_{(\mathbf{2})}$
Business	$0.1337_{(5)}$	$0.1160_{(2.5)}$	$0.1263_{(4)}$	$\mathbf{0.1130}_{(\mathbf{1})}$	$0.1160_{(2.5)}$
Computer	$0.4667_{(5)}$	$0.3523_{(2.5)}$	$0.4623_{(4)}$	$\mathbf{0.3520}_{(\mathbf{1})}$	$0.3523_{(2.5)}$
Education	$0.6307_{(5)}$	$\mathbf{0.4533}_{(\mathbf{1})}$	$0.6187_{(4)}$	$0.4537_{(2.5)}$	$0.4537_{(2.5)}$
Emotion	$0.3020_{(4.5)}$	$0.3020_{(4.5)}$	$0.2871_{(3)}$	$0.2822_{(2)}$	$\mathbf{0.2723}_{(\mathbf{1})}$
Health	$0.4790_{(5)}$	$\mathbf{0.2537}_{(\mathbf{1})}$	$0.4530_{(4)}$	$0.2543_{(3)}$	$0.2540_{(2)}$
Rec	$0.7783_{(5)}$	$\mathbf{0.4603}_{(\mathbf{1})}$	$0.7603_{(4)}$	$0.4610_{(3)}$	$0.4607_{(2)}$
Ref	$0.5133_{(5)}$	$\mathbf{0.3667}_{(\mathbf{1})}$	$0.5123_{(4)}$	$0.3677_{(2)}$	$0.3693_{(3)}$
Scene	$0.2299_{(2)}$	$0.2525_{(4.5)}$	$\mathbf{0.2107}_{(\mathbf{1})}$	$0.2525_{(4.5)}$	$0.2500_{(3)}$
Science	$0.7323_{(5)}$	$\mathbf{0.4907}_{(\mathbf{1})}$	$0.7013_{(4)}$	$0.4930_{(3)}$	$0.4910_{(2)}$
Social	$0.4217_{(5)}$	$\mathbf{0.2887}_{(\mathbf{1})}$	$0.4147_{(4)}$	$0.2890_{(2)}$	$0.2897_{(3)}$
Society	$0.4700_{(4)}$	$0.3940_{(3)}$	$0.4727_{(5)}$	$\mathbf{0.3933}_{(\mathbf{1.5})}$	$\mathbf{0.3933}_{(\mathbf{1.5})}$
Yeast	$0.2560_{(5)}$	$0.2429_{(3)}$	$0.2516_{(4)}$	$0.2386_{(2)}$	$\mathbf{0.2364}_{(\mathbf{1})}$
Medical	$0.3504_{(2)}$	$0.3643_{(5)}$	$\mathbf{0.2977}_{(\mathbf{1})}$	$0.3628_{(3.5)}$	$0.3628_{(3.5)}$
Average ranking	4.4643	2.3571	3.5714	2.3571	**2.2500**

(5) MNECM: The kernel parameter is set to 1, the regularization parameter is set to 1 and the kernel function selects RBF.

6. Experimental results and analysis

6.1. Full-labels experimental results and analysis

The experimental results of the MNECM and other four algorithms on 14 full-labels multi-label datasets are shown in Tables 3–8, where the ranking of the experimental results corresponding to each dataset is shown in Tables 3–8 in the form of subscripts. The average ranking of each algorithm on all datasets is given in the last row; the lower the average ranking, the better the algorithm.

According to the experimental results on the 14 full-label multi-label datasets (Tables 3–8), we can draw the following conclusions:

(1) Table 3 shows that MNECM has the best AP evaluation metric on most datasets, and the average ranking is optimal.
(2) In Table 4, the average ranking of MNECM is lower than that of the AP evaluation metric. It can be seen that MNECM is dominant under the AUC evaluation metric.
(3) In Table 5, MNECM has the lowest CV evaluation metric and is optimal on more than half of the datasets.
(4) In Table 6, the MNECM algorithm has a lower average ranking and a higher-ranking indicator than comparison algorithms.
(5) In Table 7, the average ranking of OE evaluation metric of MNECM is slightly better than that of MLKELM and Glocal algorithms, but it is optimal on four datasets, including Birds and Emotion.
(6) In Table 8, MNECM performs poorly on the Scene dataset, but the average ranking on the RL evaluation metric is optimal and the metric on the upper half of the dataset is optimal.

According to the above analysis, in addition to the OE evaluation metric, MNECM is dominant on most datasets, and the average ranking of the six metrics is optimal. The rationality of the algorithm is further illustrated by statistical hypothesis testing and stability analysis.

Statistical hypothesis test: We employed the Nemenyi test [37] with a significance of 5% to compare the experimental results

Table 8
RL(↓) results on all 14 datasets.

Datasets	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
Birds	$0.1270_{(4)}$	$0.1105_{(2)}$	$0.1316_{(5)}$	$\mathbf{0.1074}_{(1)}$	$0.1126_{(3)}$
Business	$0.0440_{(4)}$	$0.0395_{(2)}$	$0.0516_{(5)}$	$0.0396_{(3)}$	$\mathbf{0.0362}_{(1)}$
Computer	$0.1001_{(4)}$	$0.0816_{(2)}$	$0.1040_{(5)}$	$0.0818_{(3)}$	$\mathbf{0.0757}_{(1)}$
Education	$0.0990_{(4)}$	$0.0862_{(3)}$	$0.1144_{(5)}$	$0.0858_{(2)}$	$\mathbf{0.0765}_{(1)}$
Emotion	$\mathbf{0.1528}_{(1)}$	$0.1710_{(4)}$	$0.1755_{(5)}$	$0.1608_{(3)}$	$0.1568_{(2)}$
Health	$0.0706_{(4)}$	$0.0492_{(3)}$	$0.0723_{(5)}$	$0.0486_{(2)}$	$\mathbf{0.0436}_{(1)}$
Rec	$0.2130_{(4)}$	$0.1383_{(2)}$	$0.2140_{(5)}$	$0.1398_{(3)}$	$\mathbf{0.1335}_{(1)}$
Ref	$0.0995_{(4)}$	$0.0730_{(3)}$	$0.1020_{(5)}$	$0.0729_{(2)}$	$\mathbf{0.0659}_{(1)}$
Scene	$0.0858_{(2)}$	$0.0875_{(3)}$	$\mathbf{0.0805}_{(1)}$	$0.0884_{(4)}$	$0.0916_{(5)}$
Science	$0.1472_{(4)}$	$0.1103_{(2)}$	$0.1715_{(5)}$	$0.1107_{(3)}$	$\mathbf{0.1041}_{(1)}$
Social	$0.0718_{(4)}$	$0.0602_{(3)}$	$0.0906_{(5)}$	$0.0601_{(2)}$	$\mathbf{0.0556}_{(1)}$
Society	$0.1476_{(4)}$	$0.1303_{(2)}$	$0.1672_{(5)}$	$0.1305_{(3)}$	$\mathbf{0.1244}_{(1)}$
Yeast	$0.1768_{(4)}$	$\mathbf{0.1597}_{(1)}$	$0.1945_{(5)}$	$0.1636_{(3)}$	$0.1616_{(2)}$
Medical	$0.0585_{(2)}$	$0.0653_{(4)}$	$\mathbf{0.0490}_{(1)}$	$0.0683_{(5)}$	$0.0630_{(3)}$
Average ranking	3.5000	2.5714	4.4286	2.7857	**1.7143**

of the MNECM and other algorithms on all 14 datasets. We believe there is no significant difference between any two algorithms when their difference in average ranking on all datasets is smaller or equal to the critical difference (CD), otherwise, we believe there is a significant difference. Every two algorithms are compared in terms of different evaluation metrics, as shown in Fig. 3 The CD on the top line equals 1.6303, and the algorithms with no significant difference are connected by colored lines. The algorithms are ranked in decreasing order from left to right in each figure.

As shown in Fig. 3, the Nemenyi test shows that MNECM is significantly different from the other algorithms in 50% of cases. Moreover, for HL, MNECM is slightly better than MLKNN. For OE, MNECM is slightly better than MLKELM and Glocal. For AUC, CV, and RL, MNECM are significantly better than the other algorithms.

(1) As shown in Fig. 3(a), (b), (c), and (f), there is no significant difference among MNECM, MLKELM and Glocal, but there is a significant difference among MNECM, MLKNN and NeLC-NLS.
(2) As shown in Fig. 3(d), there is no significant difference between MNECM and MLKNN. There is a significant difference between MNECM and NeLC-NLS, and between MLKELM and Glocal. There is no significant difference among NeLC-NLS, MLKELM and Glocal.
(3) As shown in Fig. 3(e), there is no significant difference among MNECM, MLKELM, Glocal and NeLC-NLS, but there is a significant difference between MNECM and MLKNN.

The above analysis shows that MNECM is not significantly different from other algorithms in 50% of cases, and the performance is optimal on the six-evaluation metrics. MNECM is superior to NeLC-NLS in all six-evaluation metrics, and there are significant differences between the two algorithms. It can be seen that the integration of labels density improves the performance of the algorithm. Except for with HL, MNECM and Glocal have no significant difference, and MNECM is better than Glocal on six-evaluation metrics. It can be seen that considering the relationship between 0 labels and positive and negative labels can improve the quality of label completion.

6.2. Missing labels experimental results and analysis

Figs. 4–9 show the experimental results of the algorithm and the other four algorithms on 14 missing labels multi-label datasets. The *Y*-axis represents the evaluation metric sizes, and the "70,50,30,10" of the *X*-axis represents the percentage of the label. When $r = 100$, it represents the full-label case.

As shown in Figs. 4–9, MLKNN and MLKELM dominate on a few datasets. NeLC-NLS dominates on some datasets. The Glocal algorithm gets smaller and smaller when r is smaller, and the evaluation metric changes more. MNECM is slightly better than Glocal in most datasets. As shown in Fig. 7, The HL is almost the same. The dominant number of MNECM does not change much when r gets smaller and smaller, and is thus less affected by r. It can be seen that MNECM still maintains good performance as the missing labels density decreases; this is because the correlation between positive, negative and 0 labels and the influence of labels density are taken into account.

In addition, Table 9 lists the *mean* and *std* of the optimal number of each algorithm on the six-evaluation metrics. The best result is in bold, and the suboptimal result is in italics. As shown in Table 9, when r is equal to 70 or 30, MNECM is optimal, and Glocal is suboptimal; when r is equal to 50, Glocal is optimal, and MNECM is suboptimal; and when r is equal to 10, MNECM is optimal, and NeLC-NLS is suboptimal.

As shown in the box plot of Fig. 10, the performance of each algorithm under the six-evaluation metrics is clearly shown when $r = 70$. The *Y*-axis represents the evaluation metric size, and the *X*-axis represents MLKNN, MLKELM, NeLC-NLS, Glocal and MNECM, labeled as 1–5 respectively. The red "+" indicates abnormal data. As shown in Fig. 10(a) and (b), MNECM's medians of the AP and AUC evaluation metrics are higher than other comparison algorithms, and the upper and lower edges are ideal. As shown in Fig. 10(c) and (e), the MNECM graphics edge is ideal. As shown in Fig. 10(d), there are more abnormal points in the data under the HL metric, but MNECM's median is significantly lower than that of the other algorithm. As shown in Fig. 10(f), the RL of MNECM is smaller than that of the other algorithms. Through the above analysis, it is not difficult to find that the proposed algorithm has good robustness and stable performance.

6.3. Model decomposition experimental results and analysis

In order to verify the relationship between label density and classification margin, we decompose the model, and combine KLEM with the classification margin based on label density (CM-KLEM) and the non-equilibrium labels confidence matrix (NE-KELM), respectively. As shown in Table 10, we selected four metrics of AP, AUC, OE, and RL to show the results of KELM, CM-KELM, NE-KELM, and MNECM on 14 datasets.

According to Table 10, we found:

(1) Compared with KELM, the CM-KELM results have improved, which indicates that the classification margin based on the label density can improve the multi-label learning performance, and it is demonstrated that it is effective to combine the label density with the classification margin.
(2) By comparing KELM with NE-KELM, we found that the non-equilibrium method can improve the classification accuracy of the model.
(3) By comparing MNECM with CM-KELM and NE-KELM, we found that the results have improved, which further proves that the non-equilibrium method can improve the classification accuracy of the model and combine the label density with the classification margin is of effect.

7. Conclusion

In this paper, we propose a new multi-label correlation learning algorithm called MNECM, which considers the class imbalance in multi-label learning and the problem of missing labels combined with the classification margin and non-equilibrium. A label confidence matrix is constructed by means of information entropy, which not only takes into account positive, negative, and

Table 9
The max average optimal number on all 14 missing datasets.

r	MLKNN	MLKELM	NeLC-NLS	Glocal	MNECM
70	0.5000 ± 1.2247	0.0000 ± 0.0000	1.3333 ± 0. 8165	*5.8333±2.9944*	**6.3333±2.6583**
50	1.3333 ± 0.8165	0.1667 ± 0.4082	2.5000 ± 1.6432	**5.3333±2.5820**	*5.1667±3.2506*
30	2.1667 ± 0.7528	0.0000 ± 0.0000	2.3333 ± 1.7512	*3.6667±2.6583*	**6.3333±2.8752**
10	3.0000 ± 2.3664	2.0000 ± 3.0332	*5.0000±3.5214*	3.1667 ± 3.8687	**5.5000±3.2094**

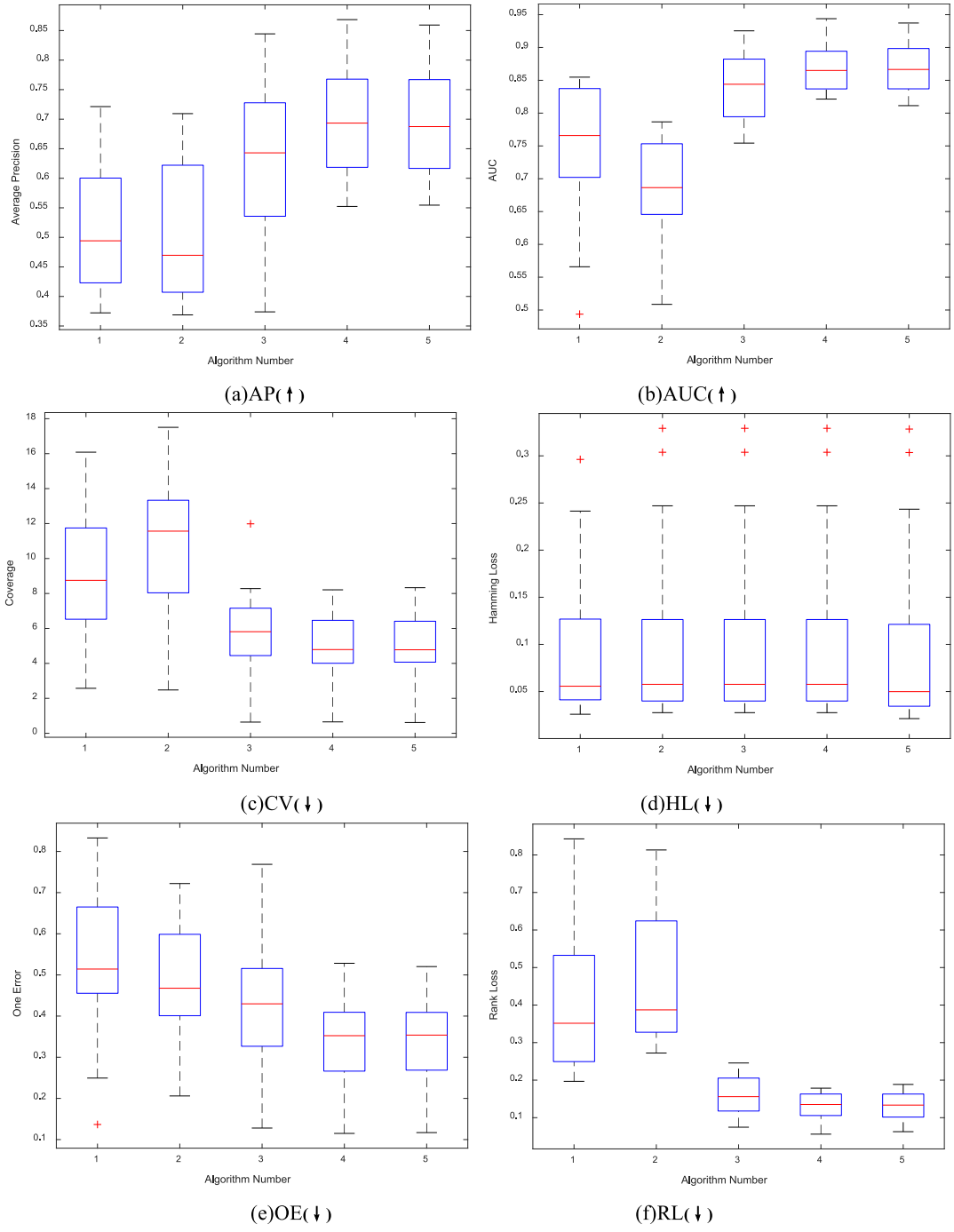

Fig. 10. Box plot.

missing labels but also considers the correlations among them. At the same time, the prior knowledge of labels density is used to increase the discrimination among labels and avoid the influence of class imbalance. Then, the positive and negative labels density is used to control the influence of the labels confidence matrix and avoid the influence of artificial parameters. Finally, MNECM provides a solution for label completion by reducing the influence of missing labels. Our method performs well on many performance evaluation indices, and multiple experiments and analyses prove that our method has certain advantages over existing methods.

Table 10
Model decomposition results on all 14 datasets.

Dataset	AP(↑)				Dataset	AUC(↑)			
	KELM	CM-KELM	NE-KELM	MNECM		KELM	CM-KELM	NE-KELM	MNECM
Birds	0.7078	0.7386	0.7331	**0.7413**	Birds	0.8838	**0.8959**	0.8885	0.8874
Business	0.8615	0.8829	0.8834	**0.8871**	Business	0.9570	0.9630	**0.9651**	0.9638
Computer	0.6247	0.6687	0.6687	**0.7073**	Computer	0.9069	0.9205	0.9234	**0.9243**
Education	0.5551	0.6198	0.6205	**0.6486**	Education	0.9042	0.9213	**0.9241**	0.9235
Emotion	0.7868	0.7972	0.7995	**0.8077**	Emotion	0.8298	0.8369	0.8351	**0.8432**
Health	0.6915	0.7511	0.7512	**0.7924**	Health	0.9369	0.9522	0.9539	**0.9565**
Rec	0.4919	0.5796	0.5798	**0.6349**	Rec	0.8163	0.8553	0.8559	**0.8665**
Ref	0.6403	0.6809	0.6804	**0.7174**	Ref	0.9187	0.9309	**0.9347**	0.9341
Scene	0.8188	**0.8477**	0.8458	0.8475	Scene	0.8938	0.9119	0.9066	**0.9084**
Science	0.4786	0.5568	0.5559	**0.6040**	Science	0.8626	0.8922	0.8936	**0.8959**
Social	0.6838	0.7377	0.7372	**0.7726**	Social	0.9287	0.9418	0.9427	**0.9444**
Society	0.5759	0.6221	0.6219	**0.6441**	Society	0.8483	0.8720	0.8742	**0.8756**
Yeast	0.7389	**0.7654**	0.7549	0.7598	Yeast	0.8118	0.8378	0.8338	**0.8384**
Medical	0.7170	0.7170	**0.7183**	0.7182	Medical	0.9350	0.9355	**0.9375**	0.9374
Dataset	OE(↓)				Dataset	RL(↓)			
	KELM	CM-KELM	NE-KELM	MNECM		KELM	CM-KELM	NE-KELM	MNECM
Birds	0.3653	0.3189	0.3251	**0.3127**	Birds	0.1192	**0.1041**	0.1115	0.1126
Business	0.1400	0.1227	0.1227	**0.1160**	Business	0.0432	0.0370	**0.0349**	0.0362
Computer	0.4627	0.4120	0.4117	**0.3523**	Computer	0.0933	0.0795	0.0766	**0.0757**
Education	0.5750	0.4960	0.4967	**0.4537**	Education	0.0958	0.0787	**0.0759**	0.0765
Emotion	0.3069	0.2871	0.2772	**0.2723**	Emotion	0.1702	0.1631	0.1649	**0.1568**
Health	0.3793	0.3200	0.3203	**0.2540**	Health	0.0633	0.0479	0.0461	**0.0436**
Rec	0.6403	0.5430	0.5437	**0.4610**	Rec	0.1840	0.1447	0.1441	**0.1335**
Ref	0.4647	0.4210	0.4230	**0.3693**	Ref	0.0813	0.0692	**0.0653**	0.0659
Scene	0.3018	0.2525	0.2533	**0.2500**	Scene	0.1065	**0.0881**	0.0934	0.0916
Science	0.6487	0.5563	0.5560	**0.4910**	Science	0.1375	0.1078	0.1064	**0.1041**
Social	0.4087	0.3453	0.3470	**0.2897**	Social	0.0713	0.0582	0.0573	**0.0556**
Society	0.4817	0.4300	0.4310	**0.3933**	Society	0.1518	0.1280	0.1258	**0.1244**
Yeast	0.2658	**0.2342**	0.2375	0.2364	Yeast	0.1882	0.1622	0.1662	**0.1616**
Medical	**0.3628**	**0.3628**	**0.3628**	**0.3628**	Medical	0.0728	0.0651	**0.0627**	0.0630

However, there are still some problems in our method. For example, the more serious the situation of missing labels, the less satisfactory the effect of label completion. Therefore, how to accurately complete labels in the case of less label information will be the focus of our next study.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.105924.

Acknowledgments

This work is supported by the Program for Innovative Research Team in Anqing Normal University, China, Key University Natural Science Research Funds of Anhui Province, China (No. KJ2017A352). We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this revised manuscript.

References

[1] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837, http://dx.doi.org/10.1109/TKDE.2013.39.
[2] X. Zhu, X. Li, S. Zhang, Block-row sparse multiview multilabel learning for image classification, IEEE Trans. Cybern. 46 (2) (2016) 450–461, http://dx.doi.org/10.1109/TCYB.2015.2403356.
[3] M.L. Zhang, Z.H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048, http://dx.doi.org/10.1016/j.patcog.2006.12.019.
[4] F.F. Luo, W.Z. Guo, Y.L. Yu, G.L. Chen, A multi-label classification algorithm based on kernel extreme learning machine, Neurocomputing 260 (2017) 313–320, http://dx.doi.org/10.1016/j.neucom.2017.04.052.
[5] M.L. Zhang, Y.K. Li, X.Y. Liu, Towards class-imbalance aware multi-label learning, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015, pp. 4041-4047. http://dl.acm.org/citation.cfm?id=2832812.
[6] K.W. Sun, C.H. Lee, Addressing class-imbalance in multi-label learning via two-stage multi-label hypernetwork, Neurocomputing 266 (2017) 375–389, http://dx.doi.org/10.1016/j.neucom.2017.05.049.
[7] G.B. Xiao, X. Zhou, Y. Yan, H.Z. Wang, A two-step hypergraph reduction based fitting method for unbalanced data, Pattern Recognit. Lett. http://dx.doi.org/10.1016/j.patrec.2018.11.003.
[8] C.F. Tsai, W.C. Lin, Y.H. Hu, G.T. Yao, Under-sampling class imbalanced datasets by combining clustering analysis and instance selection, Inform. Sci. 477 (2019) 47–54, http://dx.doi.org/10.1016/j.ins.2018.10.029.
[9] W. Feng, W.J. Huang, J.C. Ren, Class imbalance ensemble learning based on the margin theory, Appl. Sci. 8 (5) (2018) 815–843, http://dx.doi.org/10.3390/app8050815.
[10] Y. Liu, K. Wen, Q. Gao, X.B. Gao, F.P. Nie, SVM Based multi-label learning with missing labels for image annotation, Pattern Recognit. 78 (2018) 307–317, http://dx.doi.org/10.1016/j.patcog.2018.01.022.
[11] B. Wu, F. Jia, W. Liu, B. Ghanem, S.W. Lyu, Multi-label learning with missing labels using mixed dependency graphs, Int. J. Comput. Vis. 126 (8) (2018) 875–896, http://dx.doi.org/10.1007/s11263-018-1085-3.
[12] Y. Zhu, J.T. Kwok, Z.H. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 30 (6) (2018) 1081–1094, http://dx.doi.org/10.1109/tkde.2017.2785795.
[13] Z.F. He, M. Yang, Y. Gao, H.D. Liu, Y.L. Yin, Joint multi-label classification and label correlations with missing labels and feature selection, Knowl.-Based Syst. 163 (2019) 145–158, http://dx.doi.org/10.1016/j.knosys.2018.08.018.
[14] S. Hou, S. Zhou, L. Chen, Y. Feng, K. Awudu, Multi-label learning with label relevance in advertising video, Neurocomputing 171 (2016) 932–948, http://dx.doi.org/10.1016/j.neucom.2015.07.022.
[15] W.F. Gao, L. Hu, P. Zhang, J.L. H.E., Feature selection considering the composition of feature relevancy, Pattern Recognit. Lett. 112 (2018) 70–74, http://dx.doi.org/10.1016/j.patrec.2018.06.005.
[16] J. Cai, J.W. Luo, S.L. Wang, S. Yang, Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79, http://dx.doi.org/10.1016/j.neucom.2017.11.077.
[17] D. Panday, R.C. Amorim, P. Lane, Feature weighting as a tool for unsupervised feature selection, Inform. Process. Lett. 129 (2018) 44–52, http://dx.doi.org/10.1016/j.ipl.2017.09.005.
[18] Y.J. Lin, Q.H. Hu, J.H. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103, http://dx.doi.org/10.1016/j.neucom.2015.06.010.
[19] F. Li, D.Q. Miao, W. Pedrycz, Granular multi-label feature selection based on mutual information, Pattern Recognit. 67 (2017) 410–423, http://dx.doi.org/10.1016/j.patcog.2017.02.025.

[20] J. Lee, H. Kim, N.R. Kim, J.H. Lee, An approach for multi-label classification by directed acyclic graph with label correlation maximization, Inform. Sci. 351 (C) (2016) 101–114, http://dx.doi.org/10.1016/j.ins.2016.02.037.
[21] J. Xie, L. Yu, L. Zhu, G.L. Duan, Conditional entropy based classifier chains for multi-label classification, Neurocomputing 335 (2019) 185–194, http://dx.doi.org/10.1016/j.neucom.2019.01.039.
[22] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (3) (2011) 333, http://dx.doi.org/10.1007/978-3-642-04174-7_17.
[23] Y.S. Cheng, D.W. Zhao, K. Qian, Multi-label learning for non-equilibrium labels completion in neighborhood labels space, Pattern Recognit. Artif. Intell. 31 (8) (2018) 740–749, http://dx.chinadoi.cn/10.16451%2fj.cnki.issn1003-6059.201808006.
[24] C.S. Zhang, J.J. Bi, S.X. Xu, E. Ramentol, G.J. Fan, B.J. Qiao, H. Fujita, Multi-imbalance: an open-source software for multi-class imbalance learning, Knowl.-Based Syst. 174 (2019) 137–143, http://dx.doi.org/10.1016/j.knosys.2019.03.001.
[25] M. Sahu, A. Mukhopadhyay, A. Szengel, S. Zachow, Addressing multi-label imbalance problem of surgical tool detection using CNN, Int. J. Comput. Assisted Radiol. Surgery 12 (6) (2017) 1013–1020, http://dx.doi.org/10.1007/s11548-017-1565-x.
[26] N.Y. Chen, Y. Liu, H.Q. Chen, J.J. Cheng, Detecting communities in social networks using label propagation with information entropy, Physica A 471 (2017) 788–798, http://dx.doi.org/10.1016/j.physa.2016.12.047.
[27] D.D. Li, Z. Wang, C. Cao, Y. Liu, Information entropy based sample reduction for support vector data description, Appl. Soft Comput. 71 (2018) 1153–1160, http://dx.doi.org/10.1016/j.asoc.2018.02.053.
[28] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501, http://dx.doi.org/10.1016/j.neucom.2005.12.126.
[29] G. Huang, G.B. Huang, S.J. Song, K.Y. You, Trends in extreme learning machines: a review, Neural Netw. 61 (2015) 32–48, http://dx.doi.org/10.1016/j.neunet.2014.10.001.
[30] Y. Lei, D. Zhao, H.B. Cai, Prediction of length-of-day using extreme learning machine, Geod. Geodyn. 6 (2) (2015) 151–159, http://dx.doi.org/10.1016/j.geog.2014.12.007.
[31] Y.S. Cheng, D.W. Zhao, W.F. Zhan, Y.B. Wang, Multi-label learning of non-equilibrium labels completion with mean shift, Neurocomputing 321 (2018) 92–102, http://dx.doi.org/10.1016/j.neucom.2018.09.033.
[32] W.M. Huang, N. Li, Z.P. Lin, G.B. Huang, W.W. Zong, J.Y. Zhou, Y.P. Duan, Liver tumor detection and segmentation using kernel-based extreme learning machine, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2013, pp. 3662–3665, http://dx.doi.org/10.1109/EMBC.2013.6610337.
[33] R. Ye, Q. Dai, Multitl-KELM: A multi-task learning algorithm for multi-step-ahead time series prediction, Appl. Soft Comput. 79 (2019) 227–253, http://dx.doi.org/10.1016/j.asoc.2019.03.039.
[34] W.Y. Deng, Q.H. Zheng, L. Chen, X.B. Xu, Research on extreme learning of neural networks, Chinese J. Comput. 33 (2) (2010) 279–287, http://dx.doi.org/10.3724/SP.J.1016.2010.00279.
[35] Z.H. Zhou, M.L. Zhang, Multi-label learning, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, Berlin, 2017, pp. 875–881, http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/EncyMLDM2017.pdf.
[36] M. Xu, R. Jin, Z.H. Zhou, Speedup matrix completion with side information: application to multi-label learning, Adv. Neural Inf. Process. Syst. 26 (2013) 2301–2309, http://papers.nips.cc/paper/4999-speedup-matrix-completion-with-side-information-application-to-multi-label-learning.
[37] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30, http://dx.doi.org/10.1007/s10846-005-9016-2.