Neurocomputing 321 (2018) 92–102

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-label learning of non-equilibrium labels completion with mean shift

Cheng Yusheng[a,b,*], Zhao Dawei[a], Zhan Wenfa[a], Wang Yibin[a]

[a] School of Computer and Information, Anqing Normal University, Anhui, Anqing 246011, China
[b] Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Fujian, Zhangzhou 363000, China

ARTICLE INFO

Article history:
Received 26 April 2018
Revised 21 July 2018
Accepted 10 September 2018
Available online 20 September 2018

Communicated by Dr. Chenping Hou

Keywords:
Multi-label classification
Label correlation
Information entropy
Label completion
Mean shift

ABSTRACT

In multi-label learning, the use of labels correlation is crucial for the improvement of multi-label learning performance. Most of the existing methods for studying labels correlation usually do not consider the study of feature-space information. Further study is deserved about how to synchronize rich information contained in features-space and labels-space. In this paper, a multi-label learning algorithm of **N**on-**E**quilibrium **L**abels **C**ompletion with **M**ean **S**hift (i.e. NeLC-MS) was proposed. The aim of this research was to mine the feature hidden information by reconstructing the features space, and introduce non-equilibrium label correlation information so as to better improve the robustness of multi-label learning classification. First, the mean shift clustering method was used to reconstruct the information between features in the feature space to obtain the hidden information between features. Then, the new information entropy was used to measure the correlation between labels which gets the basic labels confidence matrix. Then the basic labels confidence matrix was improved to construct a Non-equilibrium labels completion matrix by the non-equilibrium parameters. Finally, the new training set was constructed by using the reconstructed features space and the Non-equilibrium Labels Completion matrix, and the existing linear classifier was used for predicting the new training set. The experimental results of the proposed algorithm in the opening benchmark multi-label datasets showed that the NeLC-MS algorithm would have some advantages over other comparative multi-label learning algorithms, and the effectiveness of the proposed method was further illustrated by the use of statistical hypothesis test and stability analysis.

1. Introduction

The multi-label learning [1] is one of the important learning frameworks for dealing with real-world objects with rich semantics. At present, most of the multi-label learning methods, such as binary relevance (BR), label power set (LP) [2], back-propagation for multi-label (BP-MLL) [3] and multi-label lazy learning methods(such as ML-KNN) [4], are usually considered to be independent individuals. In fact, in the real world, they're not independent from each other among the labels, and there is a certain correlation between them. For example, if a document contains "*Sports*" and "*physical education*", the possibility of marking "*Olympics*" will be larger, and the possibility of marking "*politics*" will be smaller. In order to build strong generalization performance, it is an important issue in the research of multi-label classification learning algorithm about how to make full use of the correlation information between labels.

The relevant algorithms targeted to the correlation between labels have been proposed in great numbers and achieved good results. For example, by transforming the multi-label learning problems into BR-based classifier chains, the Classifier Chains (CC) algorithm [5] achieves accurate predictions, while it considers the correlation between labels. However, the chain is randomly arranged and only considers the correlation between labels. For the Calibrated Label Ranking (CLR) [6], it tells correlated labels from those uncorrelated ones based on artificially-calibrated labels, but yields algorithmically-complicated, yet less accurate predictions at the cost of the number of sub-classifiers, if encountered with huge data sets and large amounts of classified labels. The Random *k*-Label sets (RAkEL) [7], based on an integrated algorithm of LP classifiers and rated as an improved LP version, consider the correlation between labels, but it leads to the complexity of the algorithm. By employing the maximum interval criterion strategy

* Corresponding author at: School of Computer and Information, Anqing Normal University, Anhui, Anqing 246011, China.
E-mail addresses: chengyshaq@163.com (C. Yusheng), like854@qq.com (Z. Dawei), zhanwf@aqnu.edu.cn (Z. Wenfa), wangyb07@mail.ustc.edu.cn (W. Yibin).

https://doi.org/10.1016/j.neucom.2018.09.033
0925-2312/

to adapt to the multi-label learning, the Support Vector Machine Ranking (RankSVM) [8] constructs SVM classifiers in its modeling to the ranking loss of correlated and uncorrelated labels of corresponding samples, and it is time-consuming with regard to huge parameter calculation. On the basis of the ML-KNN algorithm, Younes et al. [9] add the domain relation on the basis of the maximum posterior probability (MAP), considering the label correlation; Gweon et al. [10] propose a novel multi-label learning method, which uses a dual distance nearest neighbor marker set (NLDD), but NLDD only implicitly considers the correlation among the labels. Zhang and Yeung [11] intend to describe the correlation between paired labels (positively or negatively correlated, or uncorrelated) based on covariance matrices between labels, but it solves only the correlation between paired labels.

At the same time, information entropy [12] as an effective measure of uncertainty has been widely applied to the study of labels correlation. Due to consideration to the importance of the correlation between labels, Zhang et al. [13] propose a multi-label classification algorithm based on the correlation information entropy to measure weak and strong correlations between labels on the basis of the RAkEL algorithm. Lee et al. [14] propose a new multi-label learning method based on CC algorithm. The correlation of labels is modeled by a directed acyclic graph, which maximizes the correlation between labels by using conditional entropy, and good results have been achieved. Park et al. [15] propose a multi-label learning system with probability distribution, which uses normalized entropy as a standard for the system to measure the accuracy of the whole classification.

It is not difficult to find that it is feasible to use the information entropy to represent the correlation among labels. However, these methods often measure the mutual influence between the marked labels, which neglects the influence of annotation of unknown labels to the quality of label sets and the influence of known annotations to unknown ones. In literature [16], it is found that there is an asymmetric relationship between labels through reusable weight calculation, which indicates that there are some problems using the traditional mutual information or cosine similarity method. Moreover, these methods often measure the information entropy or mutual information between the marked labels. It can be seen that using this relationship to measure the related information between labels only takes into account the interaction between the marked labels, but neglects the influence of the unmarked labels. It is undeniable that a lot of valuable information may be included in unmarked labels, such as an "*apple*" and no "*mobile*" in a document, but a "*cell phone*" often determines the tendencies of the labels set as a whole. Moreover, the unmarked labels in the labels space may contain a lot of effective information.

In fact, in the multi-label learning data set, the number of labels is generally more, but the average number and the density of labels for each object are not high. This phenomenon is also consistent with the common sense that the known labels of an object should not be greater than the unknown labels, otherwise the multi-labels of the object will lose its meaning. At the same time, it is undeniable that a lot of valuable information may be contained in unknown labels, which is common in the real world.

Based on this consideration, we introduce the non-equilibrium parameter and propose a non-equilibrium labeling completion algorithm. First, the strength of the relationship between labels is measured by the amount of conditional information between labels, and the basic confidence matrix of labels is obtained. Then a more accurate mark confidence matrix of the data set is obtained by using the proposed non-equilibrium labeling confidence matrix calculation method. Finally, the initial incomplete standard is used to reinforce by the label confidence matrix. It can be seen that it will undoubtedly lead to a more accurate classification model by modeling with non-equilibrium confidence matrix.

Traditional information entropy theory has been applied to measure the correlation between labels and achieved good results. But the traditional entropy has a high complexity of computation because it has no nature of complement. Therefore, a new definition about the rough entropy will be introduced in this paper.

Besides, the use of labels correlation to reconstruct information in feature space has also been widely applied. For example, Zhang [17] proposes an improved algorithm IMMLA on the basis of ML-KNN algorithm. The algorithm takes the labels correlation to improve the performance of the classifier, but it does not accurately reflect the complex relationship between the labels. The LIFT [18] method first uses the *K*-means clustering algorithm to cluster the positive and negative examples of each label, and calculates the distance between the sample and the cluster center to generate each labels. Zhang et al. [19] propose a multi-label information increase algorithm (Multi-label Learning with Feature-induced Labeling Information Enrichment, MLFE), which helps to change the structure information in the feature space by enriching the labels information, and the classification effect of the algorithm has some advantages. A new semi-supervised and multi-label active learning method is proposed by Wu [20], which combines automatic annotation and manual annotation to reduce the amount of annotation related to the active learning process.

In addition, the mean shift clustering [21] algorithm does not need the prior knowledge of cluster number and the shape of the cluster. At the same time, the Gauss kernel function and weight value are added to the mean shift algorithm, which makes the information effectively preserved in the process of the feature space reconstruction and can effectively extract the fuzzy information between the feature space features. After adding the Gauss kernel function, the reconstructed feature space is stable. Besides, non-equilibrium labels completion is introduced to add labels correlation information. It can make the training centralization feature and label space rich in information to improve the generalization performance of the classifier, which makes NeLC-MS more stable performance with the combination of mean shift clustering and non-equilibrium labeling complement.

Based on the idea of the above reconstruction feature space information, this paper proposes an unbalanced parameter algorithm (Multi-label learning algorithm of Non-equilibrium Labels Completion with Mean Shift, NeLC-MS). Firstly, it uses mean shift clustering [20] algorithm to extract the ambiguity between features in the feature space. In this way, the feature space is reconstructed, and a new training set is obtained by introducing the labels correlation information into the Non-equilibrium labels completion method, which enhances the training set information and improves the generalization performance of the classifier. Both experimental results and statistical hypothesis tests of the NeLC-MS shows that the algorithm has a certain validity and stability. At the same time, it also confirms the rationality of the combination of feature space reconfiguration and the correlation between labels to improve the performance of the algorithm.

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 modeling of the non-equilibrium matrix and neighboring labels space for the labels matrix completion. Our proposed method for the multi-label classification of NeLC-MS is proposed in Section 4. In Section 5, experimental results of the NeLC-MS in opening multi-label data sets shows that our algorithm is effective. Statistical hypothesis tests further prove our method in Section 6. In the last section, we sum up what has been discussed and put forward further research.

2. The multi-label learning and rough entropy

2.1. The multi-label learning and traditional entropy

Definition 1 [1]. Suppose the matrix of sample feature $X = [x_1, \ldots, x_N]^T \in R^{N\times d}$, where N and $d$denote the number of samples and the dimension of features of the training data, respectively; $x_i \in R^d$, the feature vector corresponding to the ith sample; $Y = [y_1, \ldots, y_N]^T \in R^{N\times k}$, the label matrix corresponding to the sample, where k the number of labels in the data; $y_i = \{1, -1\}^k$, the binary label indicator vector corresponding to the ith sample. Therefore, the multi-label training data set containing N samples is:

$$S = \{(x_i, Y_i)|1 \le i \le N\} \subset R^d \times \{+1, -1\}^k \quad (1)$$

Definition 2 [12,22]. Suppose the set $A = \{a_1, \ldots, a_m\}$, and $p(a_i)$denotes the prior probability of the element a_i,

$$H(A) = -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i) \quad (2)$$

then $H(A)$ is the information entropy of the set $\boldsymbol{A}$, and the larger value of it, the more uncertainty of the set.

Definition 3 [12,22]. Suppose the set $A = \{a_1, \ldots, a_m\}$ and the set $B = \{b_1, \ldots, b_n\}$, then the conditional entropy of the set $\boldsymbol{B}$ under the given constraints of the set $\boldsymbol{A}$ is:

$$H(B|A) = -\sum_{i=1}^{m}\sum_{j=1}^{n} H(b_j|a_i) \quad (3)$$

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:

$$H(b_j|a_i) = -p(a_i b_j)\log_2 p(b_j|a_i) \quad (4)$$

The conditional entropy is thus employed to describe the uncertainty of the set B with the appearing set $\boldsymbol{A}$.

Meanwhile, the traditional entropy is often used in the multi-label learning algorithms and it has a high complexity of computation because it has no nature of complement. Therefore, a new definition about the rough entropy will be introduced in this paper.

2.2. New definition about the rough entropy

An information system is usually denoted as triplet S = (U, A, f), which is called a decision table, where $U$is the universe which consists of a finite set of objects, $A$is the set of attributes. With every attribute a ∈ A, set of its values V_a is associated. Each attribute a determines an information function f: $U \to V_a$ such that for any a ∈ A and x ∈ U, f(x) ∈ V_a. Each non-empty subset P⊆A determines an indiscernible relation

$$R_P = \{(x, y) : \forall a \in P, f_a(x) = f_a(y), x, y \in U\}$$

R_P is called a equivalence relation and partitions U into a family of a disjoint subsets; U/R_P is called a quotient set of U:

$$U/R_P = \{X_1, X_2, X_3, \ldots, X_n\}$$

In the traditional entropy definition, $log_2 \frac{1}{p(X_i)}$ is used to measure the information quantity of the equivalence classes X_i. Similarly, we construct the definition of information quantity expressed by equivalence classes based on rough set theory as follows:

$$I(X_i) = 1 - \frac{|X_i|}{|U|} \quad (5)$$

|.| represents the cardinality of the set element and $0 \le I(x_i) < 1 - \frac{1}{|U|}$.

Definition 4 [22]. For an information system S = (U, A, f), P ⊆ A, $U/R_P = \{X_1, X_2, X_3, \ldots, X_n\}$, the information entropy of attributes P is defined as follows,

$$E(P) = E(X) = \sum_{i=1}^{n} \frac{|X_i|}{|U|} I(X_i) = \sum_{i=1}^{n} \frac{|X_i|}{|U|}\left(1 - \frac{|X_i|}{|U|}\right) = \sum_{i=1}^{n} \frac{|X_i|}{|U|}\frac{|X_i|^C}{|U|} \quad (6)$$

In which C represents the complement. It is easy for E(X) to be a rough entropy and $0 \le E(X) < 1 - \frac{1}{|U|}$ [27].

Similarly, if a partition of the feature space is defined as X = $\{X_1, X_2, X_3, \ldots, X_n\}$, and a partition of the labels space is marked as Y = $\{Y_1, Y_2, Y_3, \ldots, Y_m\}$. According to the definition of I(X_i), we can construct the conditional information I($x_i|y_j$) in multi-label learning as follows:

$$I(X_i|Y_j) = \frac{|X_i^c - Y_j^c|}{|U|} \quad (7)$$

Correspondingly, the space composed of (X, Y) is recorded as (X, Y) = $\{X_iY_j : X_i \in X, Y_j \in Y, i = 1 \cdots n, j = 1 \ldots m\}$ in multi-label learning, then each element (X_i, Y_j) on the (X, Y) is the average value of the joint probability weighted statistics from the amount of information, therefore, a new definition about the conditional entropy on the set (X, Y) about the multi-label system, can be defined as follows:

$$\begin{aligned} E(X|Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|(X_i \cap Y_j)|}{|U|} I\left(\left(X_i|Y_j\right)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{|(X_i \cap Y_j)|}{|U|}\frac{|X_i^c - Y_j^c|}{|U|} \end{aligned} \quad (8)$$

3. The modeling of the non-equilibrium label completion matrix

The number of unannotated items of a sample in the real world is much larger than that of annotated ones, as seen in an example that a picture with known labels including *green mountains* and *clear water* is more probable to contain unannotated *forests*, rather than unannotated *deserts* or *sea*. We have found in many cases that researchers calculate the conditional information between annotated and unannotated elements in each label set of the sample by applying Eq.(4), to obtain the basic label confidence matrix. Suppose the matrix of training samples$Y = [y_1, \cdots, y_N]^T \in R^{N\times k}$ and $y_i = \{1, -1\}^k$, and according to Eq. (4) of traditional entropy, we have:

$$a_{ij} = \frac{1}{H(\bar{l}_j|l_i)}, b_{ij} = \frac{1}{H(l_j|\bar{l}_i)}$$

where l_i and $\bar{l}_i$ denote that the value of y_i is "1", and $\bar{l}_i$ "−1"; $i = 1, \cdots, k$, $j = 1, \cdots, k$ and $i \ne j$.

According to Eq. (7) of new rough entropy, the new basic label confidence matrix can be redefined as follows:

$$newa_{ij} = \frac{1}{I(l_j^C|l_i)}, newb_{ij} = \frac{1}{I(l_j|l_i^C)}$$

Therefore, $newa_{ij}$, the new basic label confidence matrix, focuses on the confidence of known labels to unknown ones, while $newb_{ij}$ the confidence of unknown labels to known ones, and it directly affects the quality of label sets. Since most multi-label data sets are currently artificially annotated, annotating an unknown sample may directly affect the quality of multi-label data sets. The paper therefore introduces α, the unbalanced parameter and proposes the algorithm of the non-equilibrium label confidence matrix (NeLCM) based on weighted calculation of decreasing the basic label confidence matrix (BCLM) of $newa_{ij}$ and increasing that of $newb_{ij}$:

$$Conf_{ij} = -\alpha \times newa_{ij} + (1 - \alpha) \times newb_{ij} \quad (9)$$

Algorithm 1 Non-equilibrium label confidence matrix (NeLCM).

Input: $\boldsymbol{Y}$, the matrix of training samples, and α, the non-equilibrium parameter;
Output: $\hat{Y}$, the NeLCM
1) $Y = \{Y_i | i = 1, \ldots, k\}$/*The label set of the training set*/
2) for each l_i, l_j
3) While $i \neq j$
4) $newa_{ij} = \frac{1}{I(l_j^r | l_i)}$, $newb_{ij} = \frac{1}{I(l_j | l_i^r)}$/* Calculate $newa_{ij}$ and $newb_{ij}$ by employing Eq. (7).*/
5) elseif $i = j$
6) $newa_{ij} = newb_{ij} = 0$; /*Set the diagonal element as 0.*/
7) end
8) Normalize the matrix $\boldsymbol{a}$, $\boldsymbol{b}$ by row and obtain the corresponding matrix $\boldsymbol{a}$, $\boldsymbol{b}$
9) While $i = j$
10) $newa_{ij} = newb_{ij} = 1$;
11) end/*Set the diagonal element as 1. */
12) $Conf_{ij} = -\alpha \times newa_{ij} + (1 - \alpha) \times newb_{ij}$;/*Obtain the confidence matrix by employing Eq. (9).*/
13) end
14) $\hat{Y} = Conf \times Y$ /* non-equilibrium label confidence matrix */
15) return $\hat{Y}$

We suggest the range of the unbalanced parameter $0 \leq \alpha \leq 0.5$.

Inspired by the idea of labels propagation dependency [23], the non-equilibrium label completion matrix is defined as follows:

$$\hat{Y} = Conf \times Y \tag{10}$$

Introduced non-equilibrium parameters, the algorithm of non-equilibrium label confidence matrix is calculated as follows: Algorithm 1.

4. The modeling of non-equilibrium label completion matrix combined with mean shift

4.1. Combined with mean shift of Gauss kernel function

Mean shift clustering algorithm is a non-parametric clustering technology [24], It does not need to determine the number of clusters, nor does it limit the shape of clusters. In this paper, the most widely used Gauss kernel function is added to the mean shift algorithm.

$$K_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x\|^2}{2}} \tag{11}$$

The definition of multi-variate kernel density estimation is as follows:

Definition 5. There are n data points $X_i, i = 1,2,3\ldots N$, in the D-dimensional space R^d. The kernel density estimates of kernel function $K(x)$ and window radius h is shown as Eq. (12).

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{12}$$

The gradient of kernel density estimation is obtained with Eq. (13) as follows:

$$\nabla f(x) = \frac{2\sum_{i=1}^{n} (x - x_i) k'\left(\left\|\frac{x_i - x}{h}\right\|^2\right)}{nh^{d+2}} \tag{13}$$

Suppose $g(x) = -k'(x), G(x) = g(\|x\|^2)$, then Eq. (13) is replaced by Eq. (14) as follows:

$$\begin{aligned} \nabla f(x) &= \frac{2\sum_{i=1}^{n} (x_i - x) G\left(\frac{x_i - x}{h}\right)}{nh^{d+2}} \\ &= \frac{2}{h^2} \frac{\sum_{i=1}^{n} (x_i - x) G\left(\frac{x_i - x}{h}\right)}{n\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right)} \frac{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right)}{nh^d} \end{aligned} \tag{14}$$

If we decompose Eq. (14), we can get the offset vector of mean shift as Eq. (15) follows:

$$M(x) = \frac{\sum_{i=1}^{n} x_i G\left(\frac{x_i - x}{h}\right)}{n\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right)} - x \tag{15}$$

The $M(x)$ vector always points to the maximum direction of the density gradient, so that Eq. (15) can be decomposed to Eq. (16) as follows.

$$m(x) = \frac{\sum_{i=1}^{n} x_i G\left(\frac{x_i - x}{h}\right)}{n\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right)} \tag{16}$$

Finally, the iterative Eq. (17) can be obtained as follows.

$$M(x) = m(x) - x \tag{17}$$

4.2. Multi-label classifier modeling combined with mean shift

Given a multi label training set $S = \{(x_1, Y_1), \ldots (x_N, Y_N)\}$, where $x_i \in X$ is a single instance, $Y_i \in y$ is a set x_i of associated labels, and the goal of the multi label learning system is to learn a function h: $X \to 2^y$ from S, which is used to predict a label set for an unknown instance.

In this paper, traditional Euclidean distance is used for similarity computation to measure the similarity between two instances x_i and x_j features, and the similarity matrix D_E is defined as follows:

$$D_E(x_i, x_j) = \left(\sum_{h=1}^{d} (x_i^h - x_j^h)^2\right) \tag{18}$$

Among them, x_i^h and x_j^h represent the h-dimension of instance x_i and x_j respectively, and the Euclidean distance $d = D_E$ is introduced into the Eq. (15) as follows:

$$M(d) = \frac{\sum_{i=1}^{n} d_i G\left(\frac{d_i - d}{h}\right)}{\sum_{i=1}^{n} G\left(\frac{d_i - d}{h}\right)} - x \tag{19}$$

Then the iterative Eq. (17) of mean shift algorithm is rewritten as:

$$M(d) = m(d) - d \tag{20}$$

Therefore, the mean shift algorithm combined with the Gauss kernel function is as follows Algorithm 2.

Algorithm 2 Gaussian kernel function with mean shift (GMS).

Input: Multi-label data $S = \{(x_1, Y_1), \ldots (x_N, Y_N)\}$ and the size of Gauss kernel window h;
Output: Cluster M_t and cluster center $C_j (t = 1, 2, ..., k, j = 1, 2, ..., k)$
The description about the algorithm: GMS(S, h)
1) for each x_i, $x_j \in S$
2) $d_i = d_E(x_i, x_j)$; /*Calculating the Euclidean distance between examples by Eq. (18)*/
3) end for
4) repeat iterative Eq. (20)
5) until M_t doesn't change /*Iteration stop */
6) return M_t, C_j

In the training set, we measure the similarity matrix D_E between samples according to Eq. (17), and then by GMS algorithm. The similarity matrix D_E is divided into t intersecting clusters $\{M_1, M_2, \cdots, M_t\}$, and then converted to a D-dimension vector $M_{x_i}(l) = [\varphi_1(x_i), \varphi_2(x_i), \cdots, \varphi_d(x_i)]^T$. $\varphi_j(x_i) = D_E(x_i, C_j)$ is used to calculate the Euclidean distance C_j of the instance i and j. Cluster centers $C_j = \{C_1, C_2, \cdots, C_j\}$, $t, j = 1, 2, ..., k$. C_j is defined as follows:

$$C_j = \arg\min_{x_i \in D_E} \sum_{x_j \in D_E} D_E(x_i, x_j) \tag{21}$$

The weight W is obtained by minimizing the sum of squares error functions:

$$f(i, l) = w_l^T \cdot M_{x_i} \tag{22}$$

Considering the improvement of generalization ability of multi-label classifiers, the NeLC algorithm is used to complement the original Y set.

$$\hat{Y} = Conf \times Y \tag{23}$$

$$E = \frac{1}{2} \sum_{i=1}^{m} \sum_{l \in y} \left(f(i, l) - \hat{Y}(i, l)\right)^2 \tag{24}$$

By using the reconstruction error of the non-equilibrium labels on the training set according to the minimization Eq.(24), the weights $w_l (l \in y)$ are used to predict the labels of unknown instances. The objective function of Eq. (24) is differentiated into w_l and the derivative is zero. Then the normal equation of the least squares problem is defined as follows.

$$\begin{aligned} &(\Phi^T \Phi) \cdot W = \Phi^T \hat{Y} \\ &\Phi = [\varphi_{il}]_{m \times N}, \varphi_{il} = M_{x_i}(l) \end{aligned} \tag{25}$$

After obtaining the best fitting model, the training of algorithm is finished. The labels of the new sample h are predicted as follows:

$$Y^* = \left\{ l | f(h, l) = w_l^T \cdot M_{x_i} > 0, l \in Y, x_i \in h \right\} \tag{26}$$

Therefore, the algorithm NeLC-MS based on the Non-equilibrium Labels Completion model is presented. The corresponding algorithm is described as follows: Algorithm 3.

5. NeLC-MS experiment and its results

5.1. Description of the experimental data sets

In order to illustrate the effectiveness of the algorithm NeLC-MS, we choose 14 sets of data sets such as *Birds, Emotions* and 6 *Mulan* datasets and 7 sets of *Yahoo Web Page s*and *Image*. The *Mulan* dataset is from http://mulan.sourceforge.net/datasets-mlc.html. The *Yahoo Web Pages* dataset is from http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar. The Image dataset is from http://cse.seu.edu.cn/PersonalPage/zhangml/. The specific description is shown in Table 1.

5.2. The experimental environment and evaluation indicators

The experiment is conducted on a computer equipped with Windows 7 Operation System, Intel®Core$_{(TM)}$ i5-2380p, and 3.10 GHz CPU, and in Matlab2016a for the operation of experimental codes. We choose 5 commonly-applied evaluation criteria, namely, Average Precision, Coverage, Hamming Loss, One-Error, and Ranking Loss [25] to evaluate the MLLA performance. The criteria are abbreviated as AP↑, CV↓, HL↓, OE↓, and RL↓ for convenience, where ↑ indicates the higher value, the better, and ↓ indicates the lower, the better. Suppose $h(\cdot)$, the multi-label classifier; $f(\cdot, \cdot)$, the prediction function; $rank_f$, the ranking function; $D = \{(x_i, Y_i | 1 \le i \le n)\}$, the MLD. The formal methods of these criteria are defined as follows:

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

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

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

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

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

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

(4) One-Error (OE): Evaluating the occurrence number of labels when top-ranking labels are not correct:

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

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

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

5.3. Choice of algorithms and the configuration of related parameters

In order to verify the performance of the proposed algorithm, the NeLC-MS algorithm is compared with 5 multi-label mainstream classification algorithms, which are ML-KNN, IMMLA, RankSVM, MLFE and LIFT, respectively. In the NeLC-MS algorithm, the Non-equilibrium parameter is set to be [0,0.5]. In the ML-KNN algorithm, the nearest neighbor K and the smoothing parameters

Algorithm 3 Multi-label learning algorithm of non-equilibrium labels completion with mean shift (NeLCMS).

Input: $S = \{(x_i, Y_i)|1 \le i \le N\}$, the train data set; $S^* = \{(x_i, Y_i)|1 \le i \le M\}$, the test data set; Clusters M_t, Clusters centers C_j
Output: Y^*, the prediction label.
1) $Y = \{Y_i|i = 1, \cdots, k\}$/* The training label set */
2) for each $x_i \in S$
3) $\hat{Y}$/* according to the Algorithm 1*/
4) end
5) for each $x_i \in R^d$
6) $x_i \to M_{x_i}(l) = [\phi_1(x_i), \phi_2(x_i), \cdots, \phi_t(x_i)]^T$/*The feature instance is converted to D-dimensional vector */
7) $\phi_i(x_i) = D_E(x_i, C_j)$;/*The Euclidean distance of the example x_i with clusters center C_j*/
8) $E = \frac{1}{2}\sum_{i=1}^{m}\sum_{l\in y}\left(f(i,l) - \hat{Y}(i,l)\right)^2$;
9) end for
10) for each $x_i \in S^*$
11) $\overset{*}{Y} = \{l|f(h,l) = w_l^T \cdot M_{x_i} > 0, l \in Y, x_i \in h\}$;/* the label exists when the prediction value $f_l(x_i^*) > 0$*/
12) end for
13) return $\overset{*}{Y}$

Table 1
Detailed descriptions of multi-labels data sets.

Data set	Training sets	Test sets	No. of labels	No. of features	Average no. of labels	Label density	Fields
Birds	322	323	20	260	1.470	0.074	Audio
Emotions	391	202	6	72	1.868	0.311	Music
Enron	1123	579	53	1001	3.378	0.064	Text
Natural scene	1000	1000	5	294	1.236	0.247	Images
Image	1000	1000	5	294	1.236	0.245	Images
Yeast	1500	917	14	103	4.237	0.303	Biology
Arts	2000	3000	26	462	1.636	0.063	Text
Business	2000	3000	30	438	1.588	0.053	Text
Recreation	2000	3000	22	606	1.423	0.065	Text
Reference	2000	3000	33	793	1.169	0.035	Text
Science	2000	3000	40	743	1.451	0.036	Text
Social	2000	3000	39	1047	1.283	0.033	Text
Society	2000	3000	26	462	1.692	0.063	Text
Corel5K	4000	1000	374	499	3.522	0.009	Images

Table 2
AP (↑) results of all 14 datasets.

Data set	NeLC-MS	MLKNN	IMLLA	RankSVM	MLFE	LIFT
Birds	**0.7585**$_{(1)}$	0.6750$_{(4)}$	0.6649$_{(5)}$	0.6112$_{(6)}$	0.7508$_{(2)}$	0.7016 ± 0.0056$_{(3)}$
Emotions	**0.7974**$_{(1)}$	0.7878$_{(2.5)}$	0.7878$_{(2.5)}$	0.7503$_{(5)}$	0.7822$_{(4)}$	0.7456 ± 0.0082$_{(6)}$
Enron	0.7028$_{(2)}$	0.6332$_{(5)}$	0.6476$_{(4)}$	0.5225$_{(6)}$	**0.7125**$_{(1)}$	0.6881 ± 0.0033$_{(3)}$
Image	**0.8374**$_{(1)}$	0.7908$_{(5)}$	0.8082$_{(4)}$	0.7894$_{(6)}$	0.8237$_{(2)}$	0.8174 ± 0.0034$_{(3)}$
Natural scene	**0.8367**$_{(1)}$	0.7649$_{(6)}$	0.7987$_{(4)}$	0.7689$_{(5)}$	0.8166$_{(2)}$	0.8063 ± 0.0023$_{(3)}$
Yeast	**0.7637**$_{(1)}$	0.7567$_{(3.5)}$	0.7567$_{(3.5)}$	0.7566$_{(5)}$	0.7545$_{(6)}$	0.7591 ± 0.0015$_{(2)}$
Arts	**0.6095**$_{(1)}$	0.5455$_{(5)}$	0.4817$_{(6)}$	0.5690$_{(4)}$	0.5912$_{(3)}$	0.6072 ± 0.0039$_{(2)}$
Business	0.8807$_{(3)}$	0.8819$_{(2)}$	0.8660$_{(6)}$	0.8711$_{(5)}$	0.8775$_{(4)}$	**0.8827 ± 0.0013**$_{(1)}$
Computers	0.6979$_{(2)}$	0.6286$_{(4)}$	0.6105$_{(6)}$	0.6150$_{(5)}$	0.6848$_{(3)}$	**0.6980 ± 0.0048**$_{(1)}$
Recreation	0.6172$_{(2)}$	0.4482$_{(5)}$	0.4066$_{(6)}$	0.5686$_{(4)}$	0.6104$_{(3)}$	**0.6213 ± 0.0032**$_{(1)}$
Reference	**0.7102**$_{(1)}$	0.6141$_{(5)}$	0.5842$_{(6)}$	0.6250$_{(4)}$	0.6977$_{(3)}$	0.6991 ± 0.0021$_{(2)}$
Science	**0.5910**$_{(1)}$	0.5371$_{(4)}$	0.4282$_{(6)}$	0.4849$_{(5)}$	0.5689$_{(3)}$	0.5882 ± 0.0024$_{(2)}$
Society	**0.6322**$_{(1)}$	0.6137$_{(3)}$	0.5762$_{(6)}$	0.5917$_{(5)}$	0.6095$_{(4)}$	0.6317 ± 0.0019$_{(2)}$
Corel5K	**0.2643**$_{(1)}$	0.2384$_{(4)}$	0.2283$_{(5)}$	N/A	0.2432$_{(3)}$	0.2510 ± 0.0038$_{(2)}$
Average Ranking	**1.38**	4.14	5	5	3.07	2.35

Table 3
CV (↓) results of all 14 datasets.

Data set	NeLC-MS	MLKNN	IMLLA	RankSVM	MLFE	LIFT
Birds	2.8885$_{(2)}$	3.6563$_{(3)}$	3.9102$_{(5)}$	4.2446$_{(6)}$	**2.8824**$_{(1)}$	3.6641 ± 0.1247$_{(4)}$
Emotions	**1.8515**$_{(1)}$	1.8762$_{(3)}$	1.8663$_{(2)}$	2.2426$_{(6)}$	1.9703$_{(4)}$	2.1752 ± 0.0491$_{(5)}$
Enron	13.1675$_{(3)}$	13.3713$_{(4)}$	15.1537$_{(6)}$	14.8411$_{(5)}$	12.5250$_{(2)}$	**12.1149 ± 0.0693**$_{(1)}$
Image	**0.7940**$_{(1)}$	0.9530$_{(5)}$	0.9000$_{(4)}$	0.9760$_{(6)}$	0.8190$_{(2)}$	0.8584 ± 0.0091$_{(3)}$
Natural scene	**0.7980**$_{(1)}$	1.0520$_{(5)}$	0.9430$_{(4)}$	1.0550$_{(6)}$	0.8440$_{(2)}$	0.8957 ± 0.0044$_{(3)}$
Yeast	6.3871$_{(3)}$	6.4318$_{(4)}$	6.2672$_{(2)}$	**6.2475**$_{(1)}$	6.5027$_{(6)}$	6.4689 ± 0.0254$_{(5)}$
Arts	5.6993$_{(5)}$	5.1163$_{(3)}$	6.1877$_{(6)}$	4.7070$_{(2)}$	5.6857$_{(4)}$	**4.6974 ± 0.0733**$_{(1)}$
Business	2.6743$_{(4)}$	2.1693$_{(2)}$	2.9090$_{(6)}$	2.2877$_{(3)}$	2.7190$_{(5)}$	**2.1121 ± 0.0283**$_{(1)}$
Recreation	4.7023$_{(4)}$	5.1720$_{(5)}$	5.5460$_{(6)}$	3.9350$_{(2)}$	4.5443$_{(3)}$	**3.8044 ± 0.0348**$_{(1)}$
Reference	3.9440$_{(4)}$	3.4927$_{(3)}$	4.0067$_{(6)}$	2.9730$_{(2)}$	4.0130$_{(5)}$	**2.7317 ± 0.0353**$_{(1)}$
Science	7.0833$_{(4)}$	5.8567$_{(2)}$	8.1877$_{(6)}$	6.0240$_{(3)}$	7.3880$_{(5)}$	**5.5897 ± 0.0700**$_{(1)}$
Society	6.5010$_{(6)}$	5.3357$_{(2)}$	6.2433$_{(4)}$	**5.3160**$_{(1)}$	6.3313$_{(5)}$	5.3604 ± 0.0541$_{(3)}$
Corel5K	183.9030$_{(3)}$	148.0400$_{(2)}$	190.5390$_{(5)}$	N/A	184.7130$_{(4)}$	**144.8499 ± 1.1037**$_{(1)}$
Average Ranking	3.21	3.36	4.86	3.69	3.77	**2.21**

Table 4
HL (↓) results of all 14 datasets.

Data set	NeLC-MS	MLKNN	IMLLA	RankSVM	MLFE	LIFT
Birds	**0.0466**$_{(1)}$	0.0579$_{(5)}$	0.0577$_{(4)}$	0.0893$_{(6)}$	0.0502$_{(3)}$	0.0499 ± 0.0008$_{(2)}$
Emotions	0.2129$_{(2)}$	0.2195$_{(3)}$	**0.2030**$_{(1)}$	0.2946$_{(6)}$	0.2459$_{(5)}$	0.2372 ± 0.0052$_{(4)}$
Enron	0.0463$_{(2)}$	0.0517$_{(5)}$	0.0511$_{(4)}$	0.0641$_{(6)}$	**0.0454**$_{(1)}$	0.0465 ± 0.0003$_{(3)}$
Image	**0.1440**$_{(1)}$	0.1740$_{(5)}$	0.1640$_{(4)}$	0.1756$_{(6)}$	0.1554$_{(2)}$	0.1598 ± 0.0020$_{(3)}$
Natural scene	**0.1544**$_{(1)}$	0.1866$_{(6)}$	0.1668$_{(4)}$	0.1854$_{(5)}$	0.1624$_{(2)}$	0.1649 ± 0.0018$_{(3)}$
Yeast	**0.1932**$_{(1)}$	0.1987$_{(4)}$	0.1948$_{(2)}$	0.2024$_{(5)}$	0.2038$_{(6)}$	0.1974 ± 0.0010$_{(3)}$
Arts	**0.0545**$_{(1)}$	0.0604$_{(4)}$	0.0624$_{(5)}$	0.0659$_{(6)}$	0.0576$_{(3)}$	0.0546 ± 0.0002$_{(2)}$
Business	0.0256$_{(2)}$	0.0271$_{(4)}$	0.0278$_{(5)}$	0.0291$_{(6)}$	**0.0254**$_{(1)}$	0.0262 ± 0.0001$_{(3)}$
Computers	0.0349$_{(2)}$	0.0415$_{(4)}$	0.0430$_{(6)}$	0.0441$_{(5)}$	0.0358$_{(3)}$	**0.0343 ± 0.0001**$_{(1)}$
Recreation	**0.0546**$_{(1)}$	0.0628$_{(4)}$	0.0641$_{(5)}$	0.0665$_{(6)}$	0.0571$_{(3)}$	0.0548 ± 0.0002$_{(2)}$
Reference	**0.0252**$_{(1)}$	0.0319$_{(4)}$	0.0350$_{(5)}$	0.0356$_{(6)}$	0.0256$_{(2)}$	0.0256 ± 0.0002$_{(3)}$
Science	**0.0305**$_{(1)}$	0.0329$_{(4)}$	0.0351$_{(6)}$	0.0411$_{(5)}$	0.0313$_{(2)}$	0.0316 ± 0.0001$_{(3)}$
Society	**0.0511**$_{(1)}$	0.0536$_{(4)}$	0.0572$_{(5)}$	0.0603$_{(6)}$	0.0531$_{(3)}$	0.0525 ± 0.0002$_{(2)}$
Corel5K	**0.0089**$_{(1)}$	0.0093$_{(2)}$	0.0096$_{(4)}$	N/A	0.0097$_{(5)}$	0.0095 ± 0.0002$_{(3)}$
Average Ranking	**1.29**	4.14	4.29	5.69	3.00	2.64

Table 5
OE (↓) results of all 14 datasets.

Data set	NeLC-MS	MLKNN	IMLLA	RankSVM	MLFE	LIFT
Birds	**0.2817**$_{(1)}$	0.3994$_{(4)}$	0.4272$_{(5)}$	0.5325$_{(6)}$	0.3034$_{(2)}$	0.3458 ± 0.0062$_{(3)}$
Emotions	**0.3069**$_{(1.5)}$	0.3168$_{(3.5)}$	0.3218$_{(5)}$	0.3168$_{(3.5)}$	**0.3069**$_{(1.5)}$	0.3644 ± 0.0110$_{(6)}$
Enron	0.2263$_{(2)}$	0.2936$_{(5)}$	0.2867$_{(4)}$	0.4870$_{(6)}$	**0.2159**$_{(1)}$	0.2511 ± 0.0111$_{(3)}$
Image	**0.2470**$_{(1)}$	0.3230$_{(6)}$	0.2960$_{(4)}$	0.3160$_{(5)}$	0.2790$_{(2)}$	0.2798 ± 0.0071$_{(3)}$
Natural scene	**0.2560**$_{(1)}$	0.3640$_{(6)}$	0.3070$_{(4)}$	0.3500$_{(5)}$	0.2910$_{(2)}$	0.3011 ± 0.0052$_{(3)}$
Yeast	0.2366$_{(3.5)}$	0.2410$_{(5)}$	**0.2312**$_{(1)}$	0.2366$_{(3.5)}$	0.2356$_{(2)}$	0.2412 ± 0.0034$_{(6)}$
Arts	**0.4740**$_{(1)}$	0.5753$_{(5)}$	0.6597$_{(6)}$	0.5627$_{(4)}$	0.4953$_{(3)}$	0.4920 ± 0.0063$_{(2)}$
Business	**0.1147**$_{(1.5)}$	0.1190$_{(4)}$	0.1297$_{(5)}$	0.1367$_{(6)}$	**0.1147**$_{(1.5)}$	0.1222 ± 0.0023$_{(3)}$
Computers	0.3617$_{(2)}$	0.4457$_{(4)}$	0.4623$_{(5)}$	0.4830$_{(6)}$	0.3770$_{(3)}$	**0.3614 ± 0.0069**$_{(1)}$
Recreation	**0.4730**$_{(1)}$	0.7142$_{(5)}$	0.7653$_{(6)}$	0.5737$_{(4)}$	0.4860$_{(3)}$	0.4815 ± 0.0052$_{(2)}$
Reference	**0.3640**$_{(1)}$	0.4837$_{(4)}$	0.5123$_{(5)}$	0.5143$_{(6)}$	0.3820$_{(2)}$	0.3865 ± 0.0020$_{(3)}$
Science	**0.4923**$_{(1)}$	0.5747$_{(4)}$	0.7077$_{(6)}$	0.6533$_{(5)}$	0.5187$_{(3)}$	0.5103 ± 0.0040$_{(2)}$
Social	**0.2800**$_{(1)}$	0.3197$_{(4)}$	0.4070$_{(5)}$	0.4333$_{(6)}$	0.2923$_{(2)}$	0.2941 ± 0.0033$_{(3)}$
Society	**0.3927**$_{(1)}$	0.4347$_{(4)}$	0.4753$_{(5)}$	0.4830$_{(6)}$	0.4273$_{(3)}$	0.4059 ± 0.0019$_{(2)}$
Corel5K	**0.6600**$_{(1)}$	0.7350$_{(5)}$	0.7130$_{(3)}$	N/A	0.6930$_{(2)}$	0.7141 ± 0.0114$_{(4)}$
Average Ranking	**1.39**	4.61	4.5	5.15	2.14	3.07

Table 6
RL (↓) results of all 14 datasets.

Data set	NeLC-MS	MLKNN	IMLLA	RankSVM	MLFE	LIFT
Birds	0.1028$_{(2)}$	0.1358$_{(4)}$	0.1455$_{(5)}$	0.1629$_{(6)}$	**0.1011**$_{(1)}$	0.1327 ± 0.0066$_{(3)}$
Emotions	**0.1625**$_{(1)}$	0.1692$_{(2)}$	0.1693$_{(3)}$	0.2244$_{(5)}$	0.1803$_{(4)}$	0.2271 ± 0.0107$_{(6)}$
Enron	0.0847$_{(3)}$	0.0944$_{(4)}$	0.1039$_{(5)}$	0.1132$_{(6)}$	**0.0790**$_{(1)}$	0.0804 ± 0.0008$_{(2)}$
Image	**0.1326**$_{(1)}$	0.1715$_{(5)}$	0.1593$_{(4)}$	0.1766$_{(6)}$	0.1404$_{(2)}$	0.1485 ± 0.0021$_{(3)}$
Natural scene	**0.1347**$_{(1)}$	0.1952$_{(5)}$	0.1667$_{(4)}$	0.1965$_{(6)}$	0.1452$_{(2)}$	0.1559 ± 0.0012$_{(3)}$
Yeast	0.1696$_{(2.5)}$	0.1733$_{(5)}$	**0.1662**$_{(1)}$	0.1696$_{(2.5)}$	0.1777$_{(6)}$	0.1697 ± 0.0011$_{(4)}$
Arts	0.1474$_{(4)}$	0.1393$_{(3)}$	0.1743$_{(6)}$	0.1243$_{(2)}$	0.1489$_{(5)}$	**0.1208 ± 0.0018**$_{(1)}$
Business	0.0443$_{(4)}$	0.0370$_{(2)}$	0.0562$_{(6)}$	0.0402$_{(3)}$	0.0464$_{(5)}$	**0.0346 ± 0.0001**$_{(1)}$
Recreation	0.1623$_{(4)}$	0.1956$_{(5)}$	0.2123$_{(6)}$	0.1408$_{(2)}$	0.1568$_{(3)}$	**0.1310 ± 0.0015**$_{(1)}$
Reference	0.0933$_{(4)}$	0.0906$_{(3)}$	0.1049$_{(6)}$	0.0743$_{(2)}$	0.0970$_{(5)}$	**0.0661 ± 0.0010**$_{(1)}$
Science	0.1329$_{(4)}$	0.1129$_{(2)}$	0.1666$_{(6)}$	0.1152$_{(3)}$	0.1399$_{(5)}$	**0.1036 ± 0.1008**$_{(1)}$
Social	0.0767$_{(4)}$	0.0550$_{(2)}$	0.0891$_{(6)}$	0.0619$_{(3)}$	0.0779$_{(5)}$	**0.0539 ± 0.0013**$_{(1)}$
Society	0.1507$_{(4)}$	0.1328$_{(3)}$	0.1590$_{(6)}$	**0.1278**$_{(1)}$	0.1547$_{(5)}$	0.1286 ± 0.0012$_{(2)}$
Corel5K	**0.1543**$_{(1)}$	0.1735$_{(3)}$	0.2387$_{(5)}$	N/A	0.2317$_{(4)}$	0.1689 ± 0.0019$_{(2)}$
Average Ranking	2.82	3.5	4.93	3.58	3.79	**2.21**

are set to 15 and 1, respectively. In the IMMLA algorithm, the nearest neighbor number k is set to 15. In RankSvm, the cost parameter is set to 1, and RBF is selected as the kernel function. In the MLFE algorithm, the kernel function selects RBF, and the kernel parameters β_1, β_2 and β_3, are selected from {1,2. 10}, {1,10,15} and {1,10}, which are cross validated on the training set, respectively. In the LIFT algorithm, the parameter $r = 0.1$ and the kernel function is selected to be Linear. Because the result of LIFT algorithm is unstable, in order to improve the accuracy, run LIFT 10 times, and the average (mean) and standard deviation (STD) are given in our experiments.

5.4. Experimental results

The experimental results of the NeLC-MS and other 5 algorithms on 14 data sets are shown in Tables 2–6, where the

Table 7
AP (↑) results of various data sets with different parameters.

Data set	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Birds	0.7587	0.7585	0.7585	0.7596	**0.7604**
Emotions	0.7917	0.7925	0.7931	0.7942	**0.7974**
Natural scene	0.8361	0.8374	0.8373	**0.8388**	0.8367
Arts	0.6081	0.6084	0.6086	0.6095	**0.6099**

Table 8
CV (↓) results of various data sets with different parameters.

Data Set	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Birds	2.8854	2.8824	2.8885	**2.7183**	2.7554
Emotions	1.8515	**1.8416**	1.8515	**1.8416**	1.8515
Natural scene	0.8040	0.7990	0.7990	**0.7920**	0.7980
Arts	5.8267	5.7797	5.7380	5.6993	**5.6603**

Table 9
HL (↓) results of various data sets with different parameters.

Data set	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Birds	0.0676	0.0607	**0.0466**	0.0500	0.0475
Emotions	0.2616	0.2409	0.2285	0.2211	**0.2129**
Natural scene	0.2016	0.1794	0.1644	0.1580	**0.1544**
Arts	0.0607	0.0589	0.0570	**0.0545**	0.0546

Table 10
OE (↓) results of various data sets with different parameters.

Data set	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Birds	**0.2817**	**0.2817**	**0.2817**	0.2879	0.2848
Emotions	0.3267	0.3267	0.3218	0.3218	**0.3069**
Natural scene	0.2570	0.2540	0.2540	**0.2520**	0.2560
Arts	0.4753	0.4757	0.4750	0.4740	**0.4733**

Table 11
RL (↓) results of various data sets with different parameters.

Data set	$\alpha=0.1$	$\alpha=0.2$	$\alpha=0.3$	$\alpha=0.4$	$\alpha=0.5$
Birds	0.1026	0.1024	0.1028	**0.0920**	0.0925
Emotions	0.1632	0.1634	0.1639	0.1631	**0.1625**
Natural scene	0.1361	0.1349	0.1348	**0.1330**	0.1347
Arts	0.1510	0.1496	0.1485	0.1474	**0.1464**

ranking of the experimental results corresponding to each data set is shown in Tables 2–6 in the form of subscripts, and the best one is highlighted in bold. The average ranking of each algorithm in all data sets is given in the last row, where the lower the average ranking, the better the algorithm (Note: No ± value indicates that the algorithm is stable with no change in the total 10 operations; (RankSVM does not yield results in 48 h, so N/A is used.).

It is found in Table 2 that the average ranking of the NeLC-MS algorithm in all 13 data sets is the best. As shown in Table 3, the NeLC-MS algorithm performs poorly on the 7 text data sets of the *Yahoo*, but has a better performance on the 6 data sets of the *Yahoo*, and its CV ranks the second best. In terms of the Hamming loss index, as shown in Table 4, the NeLC-MS algorithm ranks second on *Emotions, Enron, Business*, and *Computers*, and the other data sets are all optimal. In terms of the OE shown in Table 5, in addition to the *Enron, Yeast* and *Computes* data sets, the performance is not the best and the other data sets have the best performance. the NeLC-MS algorithm is shown on the RL of the 14 data sets, as shown in Table 6, it has a better performance on the 6 data sets of the *Mulan*, and the whole performance is the second best.

6. Correlational analyses and statistical hypothesis test

In order to further illustrate the effectiveness of the proposed method, the parameter sensitivity analysis, the stability analysis and hypothesis testing of the algorithm are carried out based on the experimental results.

6.1. Parameter sensitivity analysis

According to the idea of our method, the value of non-equilibrium parameter is selected in the interval [0.1,0.5]. Since the values of non-equilibrium parameters have a certain influence on the algorithm in this paper, Tables 7–11 give the effect of non-equilibrium parameters on 4 data sets, such as Emotions, Natural scene et al. The text highlighted in bold indicates the best results in the experiment.

It is not difficult to observe from Tables 7 to 11 that the values of the non-equilibrium parameters can be obtained with better values in the interval [0.3,0.5], illustrating the importance of non-equilibrium parameters to mining information contained in unknown labels.

6.2. Stability analysis

In order to verify the stability of different multi label learning algorithms, the spider net diagram is used to represent the stability analysis of algorithm [26]. Because the results of the prediction classification are very different in different data sets for different evaluation indicators, we standardize the results between [0.1,0.5] as a general standard. Finally, the stability index is represented through normalized values. Fig. 1 shows the stability of the algorithm under different data sets for each evaluation index (Note: RankSVM did not get results on the Corel5K dataset, so the data set was not considered in stability analysis.).

As shown in Fig. 1, we can observe: (1) for AP, NeLC-MS obtains a fairly stable effect between the stable finger values of the 12 data sets in the [0.45,0.5]. (2) for CV, the stable value of NeLC-MS on 4 datasets is between [0.45,0.5], and the solution is quite stable compared to the MLFE and IMMLA algorithms. (3) for the HL, NeLC-MS can get more stable results on 9 datasets, and the other 4 datasets are also stable in [0.4,0.5], which are more stable than MLKNN, IMLLA, RankSVM, MLFE and LIFT algorithms. (4) for the OE, NeLC-MS can provide a more stable solution on 8 data sets, and the remaining 3 data sets are also in [0.4,0.5]. (5) for RL, NeLC-MS achieves a stable solution on 5 data sets and is more stable than MLFE and IMLLA algorithms. Therefore, the results in Fig. 1 show that NeLC-MS is more stable and has better prediction performance.

6.3. Statistical hypothesis test

We statistically employ the Nemenyi Test [18,27] with significance of 5% to compare the experimental results of the NeLC-MS and other algorithms in all 13data sets (Note: RankSVM did not get results on the Corel5K dataset, so the data set was not considered in stability analysis). We also believe there is no significant difference between any two algorithms when their difference of the average ranking in all data sets is smaller or equal to the critical difference (CD), or there is significant difference. Every two algorithms are compared in terms of different evaluation indicators, as shown in Fig. 2, where the CD on the top line equals 2.0913, and

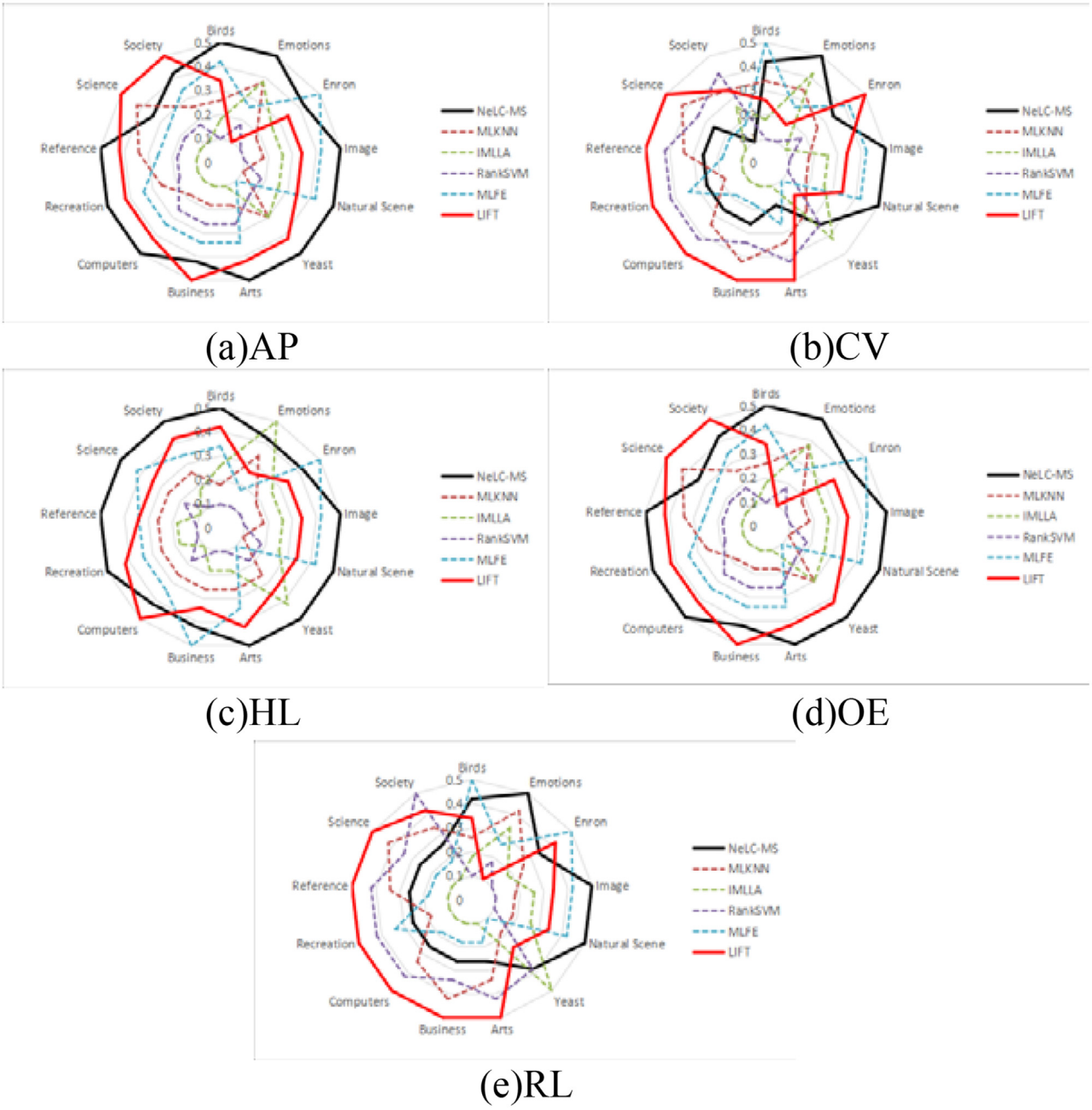

Fig. 1. The stability index values obtained on 13 benchmark multi-label datasets with different evaluation metrics.

the algorithms with no significant difference are connected by colorful lines. The algorithms are ranked in a decreasing order from left to right in each figure.

For each algorithm, there are 25 comparative results (5 comparative algorithms and 5 evaluation criteria). It is found in Fig. 2 that:

- For the NeLC-MS algorithm, there is no statistically significant difference from the other algorithms about 64%. In terms of the AP, as shown in Fig. 2(a), there is no significant difference between the NeLC-MS algorithm and the LIFT and MLFE algorithms. In terms of the CV, as shown in Fig. 2(b), there is no significant difference between the NeLC-MS algorithm and other algorithms. In terms of the HL, as shown in Fig. 2(c), NeLC-MS algorithm and LIFT and MLFE algorithms do not have a significant difference. In terms of OE, as shown in Fig. 2(d), there is no significant difference between the NeLC-MS algorithm and the LIFT and MLFE algorithms. In terms of the RL, as shown in Fig. 2(e), there is no significant difference between the NeLC-MS algorithm and the other algorithms. Therefore, the NeLC-MS is superior to other algorithms in 36% cases.
- For the LIFT algorithm, there is no statistical difference between it and other algorithms in 72% of the conditions, but in 24% cases, it is superior to other algorithms.
- For the MLFE algorithm, there is no statistical difference between it and other algorithms in 84% of the conditions, but in 16% cases, it is superior to other algorithms.

It is not difficult to see that NeLC-MS algorithm is ranked first in AP, CV, HL, OE and RL. It can be concluded that the NeLC-MS is

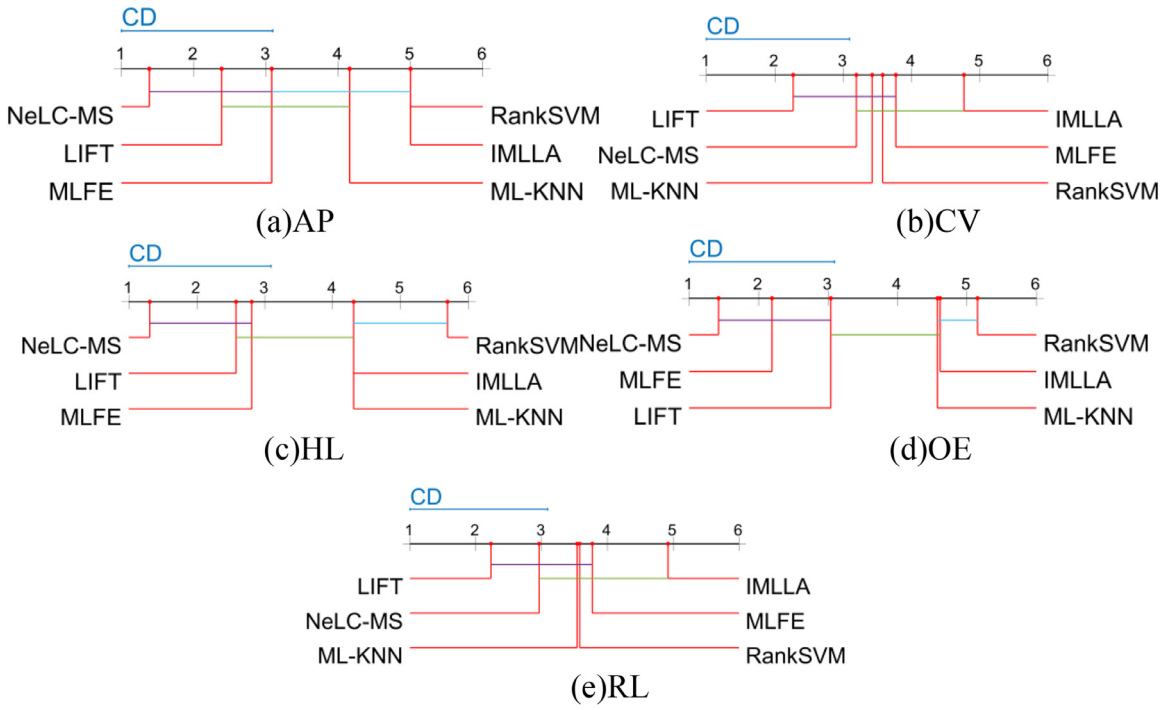

Fig. 2. Comparison of the performance of algorithms.

optimized and statistically better than other algorithms in 36% of the conditions, and it is not worse than other algorithms.

From the above analysis, the NeLC-MS algorithm has the best performance, and the experiment further illustrates the effectiveness of the NeLC-MS algorithm.

7. Conclusion

In multi-label classification learning, it is very important to study the correlation between feature information and labels in multi label learning. In the sake of making full use of the correlation, we introduced the unbalanced parameters, and proposed the NeLC-MS, Non-Equilibrium Label Completion with Mean Shift using a new rough entropy, which attempts to add the fuzzy relation between the features and the correlation between the labels by the reconstruction input space, so that the related information contained in the feature space and the labels space can be fully investigated. Although the new entropy cannot improve the accuracy and performance of the classifier, it has a simple calculation relative to the traditional entropy, and it can be used as an effective measure in the study of multi-label correlation. The combination of the unbalanced label confidence matrix and the nearest neighbor label space improves the quality of the nearest neighbor label space. Experimental results show that NeLC-MS algorithm is better than some common multi-label learning algorithms.

Because the new features cannot be theoretically guaranteed and have strong correlation between labels, the further work is to study the relationship between the feature and label space, fully excavating the effective information contained in the input space, and combining these methods to build a unified multi label learning framework.

Acknowledgments

This research is supported by the Natural Science Foundation of Higher Education of Anhui Province (No. KJ2017A177), and the Fujian Provincial Key Laboratory Fund (NO.D1801).

References

[1] M.L. Zhang, Z.H. Zhou, Multi-label learning, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, Berlin, 2017, pp. 875–881.
[2] M.R. Boutell, J. Luo, X. Shen, et al., Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.
[3] M.L. Zhang, Z.H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization, IEEE Trans. Knowl. Data Eng. 18 (10) (2006) 1338–1351.
[4] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.
[5] J. Read, B. Pfahringer, G. Holmes, et al., Classifier chains for multi-label classification, Mach. Learn. 85 (3) (2011) 333.
[6] K. Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2) (2008) 133–153.
[7] G. Tsoumakas, I. Katakis, I. Vlahavas, Random k-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (7) (2011) 1079–1089.
[8] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, 2002, pp. 681–687.
[9] Z. Younes, F. Abdallah, T. Denoeux, Multi-label classification algorithm derived from K-nearest neighbor rule with label dependencies in: Proceedings of the IEEE Signal Processing Conference , 2015:1–5.
[10] H. Gweon, M. Schonlau, S. Steiner, Nearest labelset using double distances for multi-label classification. 2017, arXiv:1702.04684.
[11] Y. Zhang, D.Y. Yeung, Multilabel relationship learning, ACM Trans. Knowl. Discov. Data 7 (2) (2013) 1–30.
[12] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[13] Z.H. Zhang, S.N. Li, Z.G. Li, A multi-lable classification algorithm using correlation information entropy, J. Northwestern Polytech. Univ. 30 (6) (2012) 968–973 (In Chinese).
[14] J. Lee, H. Kim, N.R. Kim, et al., An approach for multi-label classification by directed acyclic graph with label correlation maximization, Inf. Sci. 351 (C) (2016) 101–114.
[15] L.A. Park, S. Simoff., Using entropy as a measure of acceptance for multi-label classification, in: Proceedings of the International Symposium on Intelligent Data Analysis, Springer, 2016, pp. 217–228.
[16] S.J. Huang, Y. Yu, Z.H. Zhou, Multi-label hypothesis reuse, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 525–533.
[17] M.L. Zhang, An improved multi-label lazy learning approach, J. Comput. Res. Dev. 49 (11) (2012) 2271–2282 (In Chinese).
[18] M.L. Zhang, L. Wu, Lift: Multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 107–120.
[19] Q.W. Zhang, Y. Zhong, M.L. Zhang, Feature-induced labeling information enrichment for multi-label learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, AAAI, New Orleans, 2018 in Press.

[20] J. Wu, C. Ye, V.S. Sheng, et al., Active learning with label correlation exploration for multi-label image classification, Iet Comput. Vis. 11 (7) (2017) 577–584.
[21] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. Inf. Theory 21 (1) (1975) 32–40.
[22] J.Y. Liang, C.Y. Dang, K.S. Chin, A new method for measuring of rough sets and rough relational databases, Inf. Sci. 31 (4) (2002) 331–342.
[23] C. Pizzuti, A multi-objective genetic algorithm for community detection in networks, in: Proceedings of the IEEE International Conference on TOOLS with Artificial Intelligence, IEEE Computer Society (2009) 379–386.
[24] Y.B. Wang, Y.S. Cheng, G.S. Pei, Improved algorithm for multi-instance multi-label learning based on mean shift, J. Nanjing Univ. Nat. Sci. 54 (2) (2018) 422–435 (In Chinese).
[25] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[26] Y.J. Lin, Y.W. Li, C.X. Wang, J.K. Chen, Attribute reduction for multi-label learning with fuzzy rough set, Knowl. Based Syst. (2018), doi:10.1016/j.knosys.2018.04.004.
[27] J. Ar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (1) (2006) 1–30.

Cheng Yusheng he is a professor at Anqing Normal University(AQNU), Anhui, China. He received his Ph.D. in the School of Computer and Information Science of Hefei University of Technology in 2007. His research interests concern the rough set theory and algorithm, semi supervised learning and data mining. He is the author of more than 50 papers in journals and conference proceedings such as *Information Science, Journal of Applied Mathematics, Electronic Journal, System Engineering Theory and Practice, Pattern Recognition and Artificial Intelligence*, PAKDD and so on.

Zhao Dawei he is a graduate student of the computer and Information School of Anqing Normal University. His main research includes machine learning, data mining and statistics.

Zhan Wenfa received his Ph.D. in the School of Computer and Information Science of Hefei University of Technology, Anhui Province in 2009, and BS in VLSI from the School of Electric and Automation of Hefei University of Technology, Anhui Province in 2004. He is a Professor in the School of Computer & Information Science of Anqing Normal University, Anhui Province. His research interests include test data compression, big data, ATPG algorithms, etc. He has published over 60 papers in refereed journals and conference proceedings and hold ten Chinese patents.

Wang Yibin Professor of computer and Information College, Anqing Normal University. The main research directions include multi label learning, machine learning and software security.