Knowledge-Based Systems 159 (2018) 148–157

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-label learning with label-specific features by resolving label correlations

Jia Zhang[a,b], Candong Li[c], Donglin Cao[a,b], Yaojin Lin[d], Songzhi Su[a,b], Liang Dai[a,b], Shaozi Li[*,a,b]

[a] Department of Cognitive Science, Xiamen University, Xiamen 361005, PR China
[b] Fujian Key Laboratory of Brain-inspired Computing Technique and Applications, Xiamen University, Xiamen 361005, PR China
[c] College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, PR China
[d] School of Computer Science, Minnan Normal University, Zhangzhou 363000, PR China

ARTICLE INFO

Keywords:
Multi-label learning
Optimization framework
Label-specific features
Label correlations
Traditional Chinese medicine

ABSTRACT

In multi-label learning, different labels may have their own inherent characteristics for distinguishing each other, in the meanwhile, exploiting the correlations among labels is another practical yet challenging task to improve the performance. In this work, we present a new method for the joint learning of label-specific features and label correlations. The key is the design of an optimization framework to learn the weight assignment scheme of features, and the correlations among labels are taken into account by constructing additional features at the same time. Through iteratively optimizing the two sets of unknown variables, which are referred to feature weights and label correlations-based features, label-specific features of each label are available to achieve multi-label classification. Comprehensive experiments on various multi-label data sets including two collected traditional Chinese medicine data sets reveal the advantages of our proposed algorithm.

1. Introduction

Multi-label learning deals with objects associated with multiple class labels simultaneously [1,2], which stems from the research of text categorization where a news document could be related to several topics, such as *politics, reform*, and *economics* [3,4]. Recently, multi-label learning has attracted increasing interests from the research community, and has a wide range of applications in reality [5–8], such as bioinformatics [9], web mining [10], and automatic image annotation [11]. As an example in detail, syndrome ('Zheng' in Chinese) diagnosis is based on all clinical manifestations with systematic description [12,13], which can be regarded as a typical multi-label learning problem for diagnosing disease locations and natures of patients, and syndrome differentiation for the identification of health-state in traditional Chinese medicine (TCM) [14,15].

Since multi-label learning has great potential for the purpose of practical applications, many multi-label learning approaches have been witnessed during the past years [16–19]. For example, Boutell et al. [20] presented a simple and straightforward binary relevance method, which transformed multi-label learning into a series of binary classification subproblems, and trained a classifier for each label independently. Zhang and Zhou [21] developed a multi-label version of traditional k-nearest neighbor method. By searching for k-nearest neighbors of an unseen instance, the authors utilized maximum a posteriori (MAP) principle to predict the relevant labels. The aforementioned methods are state-of-the-art methods for multi-label learning, but the effectiveness of these methods may be suboptimal due to the ignorance of label correlations. As we know, it is essential to improve the system performance by resolving the correlations among labels [2]. For instance, Wu et al. [22] constructed a set of hierarchical trees to exploit label correlations, and made multi-label prediction via combining these constructing trees as an ensemble.

Except for the challenge of label correlations, the high-dimensional problem is another key challenge for multi-label learning [23,24]. In general, feature selection and dimension reduction are two widely used methods to reduce feature dimensionality. For instance, Zhang et al. [25] introduced a naive Bayes-based multi-label learning method as basic model. Through adopting the principal component analysis and genetic algorithms for dimension reduction, the proposed algorithm achieved a higher performance compared with the basic model. Based on the min-redundancy and max-relevance (mRMR) criterion in single-label feature selection [26], Lin et al. [27] achieved multi-label feature selection based on max-dependency and min-redundancy. Specifically, a kind of mutual information-based measure method was employed to maximize the dependency between the candidate feature and all labels, and minimize the conditional redundancy between the candidate

* Corresponding author at: Department of Cognitive Science, Xiamen University, Xiamen 361005, PR China.
E-mail addresses: zhangjia_gl@163.com (J. Zhang), fjzylcd@126.com (C. Li), another@xmu.edu.cn (D. Cao), yjlin@mnnu.edu.cn (Y. Lin), ssz@xmu.edu.cn (S. Su), dust_on_ice@126.com (L. Dai), szlig@xmu.edu.cn (S. Li).

https://doi.org/10.1016/j.knosys.2018.07.003
Received 5 September 2017; Received in revised form 30 June 2018; Accepted 3 July 2018
Available online 04 July 2018
0950-7051/

feature and the selected feature subset simultaneously. It is noteworthy that these methods obtain the same feature ranking or subspace for multi-label classification, nevertheless, each label possesses its own label-specific features [28], which are preferred in discriminating the label without losing class-discriminative information.

Inspired by the previous work, we propose a novel multi-label learning method, which attempts to learn label-specific features of each label by exploiting the correlations among labels. In detail, we first design an optimization framework to model the label-specific feature learning problem by incorporating label correlations estimation. Under this framework, the assignment scheme of feature weights and label-producing features involving the label correlations information are iteratively updated to describe the characteristics of each label, and then the label-specific features of the corresponding label can be represented by these optimized characteristics to distinguish with other labels. Furthermore, considering the label vector of an unseen instance is unknown in advance, we design a KNN-like method to generate additional features for this unseen instance. Finally, the linear model is utilized to obtain the relevant labels, by conducting extensive experiments, we can conclude that the proposed algorithm is superior on various evaluation metrics compared with some state-of-the-art multi-label methods, and major contributions of our work are summarized as follows:

- We propose to learn label-specific features using sparsity regularized optimization in multi-label setting, which cover the information of label correlations.
- We model this multi-label learning problem by an optimization framework in which the weights of features and label correlations-based features are defined as two sets of unknown variables, and introduce a iterative optimization method to update these unknown variables.
- Label correlations are represented by additional features generated in the optimization process, and a KNN-like method is designed to obtain label correlations-based features of test data.
- Extensive experiments demonstrate the advantages of our proposed algorithm. In addition, two real-world data sets on TCM are collected, and our proposed algorithm is further validated on these two data sets in terms of the identification of health-state in TCM.

The rest of this paper is organized as follows. Section 2 briefly reviews existing methods to multi-label learning, and then we give a brief review of related studies for l_1-regularization and accelerated proximal gradient in Section 3. In Section 4, we introduce the proposed learning framework, and explain the experimental results in Section 5. Finally, Section 6 concludes and discusses several issues for future work.

2. Related work

Towards a emerging machine learning paradigm, multi-label learning has a broad background for practical applications in many fields, such as the prediction of the classes of anatomical therapeutic chemicals [9], automatic image annotation [11], and the diagnosis for coronary heart disease in traditional Chinese medicine [14], however, which is also confronted with arduous challenges. In particular, label correlations [2], label-specific features [28], class-imbalance [29], and streaming data [30] are hot research topics proposed for multi-label learning in recent years. In this paper, we mainly focus on the research of label correlations and label-specific feature learning.

For the first problem, the literature [2] has theoretically summarized the common strategies, such as second-order and high-order label correlations. Second-order methods explore pairwise relationships between labels in a label-by-label style, such as CLR [6] and BPMLL [31]. High-order methods tackle the multi-label learning problem by mining relations among all labels or subsets of labels. For example, Read et al. [17] proposed a chain method, namely CC, to exploit high-order label correlations by using the label vector as additional features, and received a chain of binary classification for multi-label learning. Tsoumakas and Vlahavas [18] transformed multi-label problem into an ensemble of multi-class classification problems, and presented a random k-labelsets method based on random projections of label space. In addition to these methods, many researchers are also devoted to the research of label correlations. Huang and Zhou [32] first designed an optimization objective function to optimize the LOC code via considering label correlations locally, and then generated new classification features including the LOC code for the support vector machine (SVM) classification. Mencía and Janssen [33] introduced a bootstrapped stacking method to learn a separate ruleset for each label, which considered the correlations among labels by using the remaining labels as additional features for training samples.

For label-specific feature learning, Zhang and Wu [28] pioneered the idea in multi-label learning, and assumed that different labels ought to possess specific characteristics themselves. To model this problem, the authors proposed a multi-label learning algorithm called LIFT. First, a feature mapping method was proposed by applying clustering techniques to positive and negative training samples of each label, and then utilized the SVM method to modeling the multi-label classification. LIFT is a reasonable and effective learning approach, nevertheless, this method neglected the influence of label correlations. Moreover, Huang et al. [7] developed a new method called JFSC via the joint learning of feature selection and multi-label classification. To be specific, JFSC designed an optimization framework to learn the low-dimensional data representation of each label, and considered the shared features to exploit pairwise label correlations simultaneously. Xu et al. [19] presented a multi-label learning approach based on fuzzy rough set, which utilized the approximation quality to evaluate the significance of features. Through applying the forward greedy search strategy, the proposed method achieved label-specific feature reduction.

3. Preliminaries

3.1. Optimization methods for l_1-regularization

Sparse representation with l_1-norm minimization is one of hot topics in the research community, which is with great success in both theoretical researches and practical applications [34,35]. For solving l_1-regularization problems, many optimization approaches have been brought forward based on different functional forms, mainly including the logistic regression negative log-likelihood and the squared error in a linear regression model [36]. Compared with the method of logistic regression, the l_1-regularization with least squares loss is more general. In a significant amount of the early work [35,37], the stable version of squared error based l_1-regularization has been widely studied, which is proved to be effective in regularizing highly underdetermined linear regression and consistent in some noisy overdetermined settings [38]. In contrast, the l_1-regularized logistic regression is usually taken as an effective method to classification problems, but is limited by some severe challenges. For example, the method is usually very slow in practice to achieve a higher learning performance [39,40], and general purpose solvers for l_1-norm regularization may be inflexible in large-scale logistic regression problems [41]. Thus, we focus on the optimization method based on least squares loss for l_1-regularization.

3.2. Accelerated proximal gradient

Owing to the non-smoothness of l_1-regularization term, the optimization solution based on standard proxy function is not an antidote to the scenario. In this paper, the accelerated proximal gradient method [34,37] is introduced to solve the non-smooth convex optimization problem, which can be represented by a general optimization framework as follow:

$$\min_{W\in\mathcal{H}} \quad F(W) = f(W) + \mu\|W\|_1, \tag{1}$$

where $\mathcal{H}$ is a real Hilbert space, and μ is the optimized parameter. To handle this task, a sequence of separable quadratic approximations are minimized instead of directly optimizing W of Eq. (1), denoted as:

$$Q_{L_f}(W, W^{(t)}) = f(W^{(t)}) + <\nabla f(W^{(t)}), W - W^{(t)}> + \frac{L_f}{2}\|W - W^{(t)}\|_F^2 + \mu\|W\|_1, \tag{2}$$

where L_f is the *Lipschitz* constant of ∇f, which satisfies . In addition, the literature [42] has set:

$$W^{(t)} = W_t + \frac{b_{t-1} - 1}{b_t}(W_t - W_{t-1}), \tag{3}$$

which is helpful to improve the convergence rate to $O(t^{-2})$ for a sequence b_t satisfying $b_{t+1}^2 - b_{t+1} \le b_t^2$, and W_t is the value of W at the tth iteration.

Let $G^{(t)} = W^{(t)} - \frac{1}{L_f}\nabla f(W^{(t)})$. Then, we minimize $Q_{L_f}(W, W^{(t)})$ to obtain the optimal solution of W:

$$\arg\min_{W} \quad Q_{L_f}(W, W^{(t)}) = \frac{L_f}{2}\left\|W - G^{(t)}\right\|_F^2 + \mu\|W\|_1, \tag{4}$$

In other words, the value of W can be solved by minimizing the following optimization problem in each iteration:

$$W^{t+1} = \arg\min_{W} \quad \frac{1}{2}\left\|W - G^{(t)}\right\|_F^2 + \frac{\mu}{L_f}\|W\|_1. \tag{5}$$

4. The proposed algorithm

In this section, we describe the design of our proposed algorithm. First additional features are defined to exploit the correlations among labels, which are incorporated with the original features in form of a new feature space. Then we formulate the multi-label learning problem as an optimization problem for multi-label classification, which proposes to update label correlations-based features and the weights of features from multi-label data. By launching a iterative procedure for optimization, we develop a multi-label learning framework with label-specific features by resolving label correlations.

4.1. Problem formulation

Formally, let $X = \Re^{n\times d}$ be a d-dimensional feature space associated with a finite set of q possible class labels denoted by $L = \{l_1, l_2, \cdots, l_q\}$, and $\mathcal{D} = \{(\mathbf{x}_i, \mathcal{Y}_i)|1 \le i \le n\}$ be a multi-label training data set, where $\mathbf{x}_i \in X$ is a feature vector denoted by $\{x_{i1}, x_{i2}, ..., x_{id}\}$, and $Y_i = \{y_{i1}, y_{i2}, ..., y_{iq}\}$ is the set of the ground truth labels of $\mathbf{x}_i$. For arbitrarily element $y_{ij} \in Y_i$, $y_{ij} = 1$ in case that l_j is relevant to $\mathbf{x}_i$, otherwise $y_{ij} = 0$. Then the purpose of multi-label learning is to derive a real-valued function $h: X \to 2^L$, such that some loss functions or specific evaluation are satisfied. For each instance $\mathbf{x}_i \in X$, the function outputs a set of relevant labels $h(\mathbf{x}_i) \subseteq L$.

4.2. Optimization framework

The key insight behind the proposed method is that each label essentially possesses specific characteristics of its own, and the correlations among labels are supposed to be fully excavated to promote the model generalization ability, so the joint learning of label-specific features and label correlations should be a feasible solution to construct a high-performance learning system. Following this principle, we present an optimization objective function.

$$\min_{W,C} \quad f(W, C) = \frac{1}{2}\left\|Y' - Y\right\|_F^2 + \frac{\alpha}{2}\sum_{i=1}^{n}\left\|\mathbf{c}_i - Y_i\right\|^2 + \beta\|W\|_1, \tag{6}$$

where W is the weight matrix of features, and label-specific features of each label are indicated by non-zero components of the corresponding matrix row. $C \in \Re^{n\times q}$ is the label-producing feature space, and each feature vector $\mathbf{c}_i \in C$ is constructed to fit the label correlations between $\mathbf{x}_i$ and other instances. Y' returns the forecast matrix which corresponds to label space Y, and we implement it as:

$$Y' = X(W^x)^T + C(W^c)^T, \tag{7}$$

where $W^x \in \Re^{q\times d}$, $W^c \in \Re^{q\times q}$, and $W = [W^x, W^c]$. The second term of Eq. (6) is defined to enforce that similar label correlations-based features are shared by similar samples in regard to label space Y. Through calculating the similarity of each feature vector $\mathbf{c}_i$ with the corresponding Y_i, additional feature space C is constructed and empowered to depict the correlations relationships among labels. In addition, an l_1-regularization has been included for W as shown in the third term, which could remove irrelevant features of each label for the label-specific feature learning, and α and β are parameters to trade off the three terms.

In short, the collections of feature weights and additional features are learned together to obtain the values of W and C by minimizing the objective function, hence label-specific features can be derived via resolving label correlations. However, in light of the two unknown variables involved in the optimization procedure, it is natural to update the value of one unknown variable while maintaining the value of another set. Thus, we iteratively optimize the objective function in terms of W and C until convergence, and illuminate two elementary steps as follow:

Step 1: update W. With an estimation of C, we obtain the value of W. Since the second term of Eq. (6) is constant, the objective function of Eq. (6) can be transformed into the following formula to calculate W:

$$W = \arg\min_{W} \quad \frac{1}{2}\left\|Y' - Y\right\|_F^2 + \beta\|W\|_1, \tag{8}$$

Step 2: update C. At this step, the value of W is fixed, similarly, we reduce the optimization objective function of Eq. (6), and C is updated by solving the following optimization problem:

$$C = \arg\min_{C} \quad \left\|Y' - Y\right\|_F^2 + \alpha\sum_{i=1}^{n}\|\mathbf{c}_i - Y_i\|^2, \tag{9}$$

Now that the optimization can be achieved through alternatively updating W and C. Nevertheless, the optimization execution is based on the value of one set which is known a *priori*. To address this problem, we initialize the value of C for starting the optimization process. As we know, a label correlation is possible to be shared by a subset of samples [5,32], hence we discover groups which share the same correlation using clustering techniques. Specifically, all samples are clustered according to the information of relevant labels themselves, and the clustering center of each group is viewed as addition features shared by the samples in the group.

4.2.1. Weight learning for label-specific features

First, we discuss the calculation of W by optimizing Eq. (8). Suppose $X' = [X, C]$, $f(W)$ is defined as a convex function:

$$f(W) = \frac{1}{2}\left\|X'W^T - Y\right\|_F^2, \tag{10}$$

Then, the following objective function is defined to solve the non-smooth convex optimization problem with respect to W:

$$W^{t+1} = \arg\min_{W} \quad \frac{1}{2}\left\|W - G^{(t)}\right\|_F^2 + \epsilon\|W\|_1, \tag{11}$$

where $\epsilon = \frac{\beta}{L_f}$, and $\nabla f(W^{(t)})$ is the gradient of Eq. (10) w.r.t. $W^{(t)}$. Then the optimization problem can be solved using soft-thresholding operation $W^{t+1} = O_\epsilon[\mathcal{G}^{(t)}]$. For arbitrary element $w \in \mathbb{R}$, $O_\epsilon[w] = sign(w)(|w| - \epsilon)_+$.

Theorem 1. *Given W_1 and W_2, in consideration of that L_f satisfies $\|\nabla f(W_1) - \nabla f(W_2)\| \leq L_f \|W_1 - W_2\|$, we can calculate:*

$$L_f = \sqrt{2\sigma_{\max}^2\left((X^{'})^T X^{'}\right)}, \tag{12}$$

Proof. According to Eq. (10), $\nabla f(W) = (X^{'})^T X^{'} W^T - (X^{'})^T Y$. Then the following inequality can be derived:

$$\|\nabla f(W_1) - \nabla f(W_2)\|_F^2 \leq 2\|(X^{'})^T X^{'}\|_2^2 \|(W_1^T - W_2^T)\|_F^2, \tag{13}$$

Owing to $\left\|(X^{'})^T X^{'}\right\|_2^2 = \sigma_{\max}^2\left((X^{'})^T X^{'}\right)$, in which $\sigma_{max}(\cdot)$ denotes the maximum singular value of the matrix, then we get the value of L_f. □

4.2.2. Label correlations-based feature construction

For updating C, we estimate each row $\mathbf{c}_i$ of C independently, and the optimization problem of Eq. (9) is transformed into the following optimization problem with respect to $\mathbf{c}_i$:

$$\min_{\mathbf{c}_i} \quad \sum_{l=1}^{q} (\mathbf{x}_i(W_l^x)^T + \mathbf{c}_i(W_l^c)^T - y_{il})^2 + \alpha(\mathbf{c}_i - Y_i)^2, \tag{14}$$

Setting the gradient of Eq. (14) w.r.t. $\mathbf{c}_i$ to zero, we optimize feature vector $\mathbf{c}_i$ corresponding to $\mathbf{x}_i$:

$$\mathbf{c}_i = \left[\sum_{l=1}^{q}(y_{il} - \mathbf{x}_i(W_l^x)^T)W_l^c + \alpha Y_i\right]\left[\sum_{l=1}^{q}(W_l^c)^T W_l^c + \alpha I\right]^{-1}, \tag{15}$$

In brief, the features of $\mathbf{c}_i$ are discriminative to the corresponding sample with other samples via considering the sample correlations in terms of labels. Since the value of W and label space Y are utilized to calculate C, each feature vector $\mathbf{c}_i \in C$ can be independently updated with effect.

4.3. Classification model induction

The optimization procedure is executed for seeking the global optimal solution. Based on the optimization results, a new multi-label training set can be created, as follow:

$$T = \{([\mathbf{x}_i, \mathbf{c}_i], Y_i) | (\mathbf{x}_i, Y_i) \in \mathcal{D}, \mathbf{c}_i \in C, 1 \leq i \leq n\}, \tag{16}$$

Given an unseen instance $\mathbf{x}_t$, however, the feature vector $\mathbf{c}_t$ is unknown, hence a specifically producing mechanism for $\mathbf{c}_t$ is imperative. Thus, we calculate $\mathbf{c}_t$ by searching for k-nearest neighbors from the new training set T. Explanatorily, the similarity between $\mathbf{x}_t$ and other instances is calculated by the Cosine similarity, and then top-k similar neighbors of the test sample are selected in form of its neighbor set NN. Finally, we can obtain the value of $\mathbf{c}_t$ based on additional features of the samples in NN. Besides, concerning the squared loss function of the designed optimization function, for simplicity, a linear model is induced to generate the predicted label vector Y_t:

$$Y_t = sign(P_t - \tau), \tag{17}$$

where τ is the given threshold set to be 0.5, $P_t = [\mathbf{x}_t, \mathbf{c}_t]W^T$, and $\mathbf{c}_t = \frac{1}{k}\sum_{i \in NN}\{c_{ij} | 1 \leq j \leq q\}$. Since k has little effect as the value is small, we simply set the value of k as 10 for reducing the computation complexity.

As stated, the pseudo code of the algorithm is summarized in Algorithm 1. In which we start with the initializations of C and the parameters for optimizing W. Then the proposed learning framework is constructed by iteratively updating W and C to obtain the predicted label set. For updating W, the time cost is dominated by step 7, which is $O(d^2q + dq^2)$. For updating C, the calculation leads to a complexity of $O(nq)$. Thus, the total complexity is $O(t(d^2q + dq^2 + nq))$, where t is the number of iterations. In addition, the proposed algorithm needs memory of $O(d^2 + dq + nd + nq)$.

5. Experiments

5.1. Data sets

We experiment with thirteen multi-label data sets covering various domains including music, image, biology, text, and traditional Chinese medicine (TCM). For these data sets, emotion, genbase, medical, yeast,

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | 1 \leq i \leq n\}$, parameters α, β, and γ, number of clusters m.
Output: Predicted result for unseen instance $\mathbf{x}_t$.

1: Initialize the value of C by k-means;
2: $b_0 = b_1 = 1$, $W_0 = W_1 = [(X^{'})^T X^{'} + \gamma I]^{-1}(X^{'})^T Y$;
3: **repeat**
4: **while** *not converged* **do**
5: $W^{(t)} = W_t + \frac{b_{t-1}-1}{b_t}(W_t - W_{t-1})$;
6: Compute $L_f = \sqrt{2\sigma_{\max}^2((X^{'})^T X^{'})}$;
7: Compute $W^{t+1} = O_\epsilon[G^{(t)}]$;
8: $b_{t+1} = \frac{1}{2}(1 + \sqrt{4b_t^2 + 1})$, $t = t + 1$;
9: **end while**
10: **for** $i = 1$ to n **do**
11: Update $\mathbf{c}_i$ using Eq. (15);
12: **end for**
13: **return** C;
14: **until** *convergence criterion is satisfied;*
15: Create the new training set T using Eq. (16);
16: $P_t = [\mathbf{x}_t, \mathbf{c}_t]W^T$, $Y_t = sign(P_t - \tau)$;

Algorithm 1. Multi-label learning with label-specific features by resolving label correlations (MLFC).

Table 1
Characteristics of multi-label data sets.

Number	Data set	Domains	Samples	Features	Labels	Cardinality
1	emotions	music	593	72	6	1.869
2	genbase	biology	645	1186	27	1.252
3	medical	text	978	1449	45	1.245
4	TCM1	TCM	1146	461	43	4.189
5	TCM2	TCM	1146	504	360	1.994
6	yeast	biology	2417	103	14	4.237
7	arts	text	5000	462	26	1.636
8	computers	text	5000	681	33	1.508
9	corel5k	image	5000	499	374	3.522
10	education	text	5000	550	33	1.461
11	science	text	5000	743	40	1.451
12	social	text	5000	1047	39	1.283
13	society	text	5000	636	27	1.692

arts, computer, corel5k, education, science, social, and society are public available and widely used for the verification of multi-label methods, which can be downloaded free from Mulan Library (http://mulan.sourceforge.net/datasets.html). Besides, TCM1 and TCM2 are collected from the Second People's Hospital Health Management of Fujian Province, which is for the purpose of the multi-label learning task about the identification of health-state in TCM. Concretely, TCM1 is created with 461 symptoms from 1146 patients and has 43 class labels indicating disease locations and natures, and TCM2 has 504 features by incorporating the 43 class labels of TCM1 into 461 symptoms for identifying 360 kinds of syndromes, namely 'Zheng' in Chinese. The detailed data description of all data sets is shown in Table 1, where the cardinality measures the average number of labels in each sample.

5.2. Experimental settings

To evaluate the performance of our proposed algorithm, several state-of-the-art multi-label methods are selected as comparing algorithms, including Rank-SVM [43], ML-KNN [21], MLNB [25], LLSF [44], and LIFT [28]. In addition, we construct another method to further validate the proposed algorithm called MLFC-SVM, which uses the SVM model for multi-label classification instead of the linear model of MLFC, and LibSVM (with linear kernel) [45] is utilized to implement the SVM models for both MLFC-SVM and LIFT. For the compared methods, the parameter values of each algorithm are used as default settings according to the corresponding literature. For MLFC, α, β, and γ are set as 1, 0.1, and 0.1 for all the data sets respectively, and the parameter of m has little effect on prediction performance. This is because the parameter m is set for initializing label correlations-based feature space. However, the initialization does not affect the final results in case that the optimization objective function is convex [46].

Besides, we use five widely used multi-label evaluation metrics to compare the proposed algorithm with the above multi-label methods from different aspects [2,21]. Given test set $T' = \{(\mathbf{x}_i, Y_i)|1 \le i \le t\}$ and the family of q learned functions $\{f_1, f_2, \cdots, f_q\}$, the algorithms predict a relevant label set Y_i' for unseen instance $\mathbf{x}_i$. Then evaluation metrics are defined as follow:

Hamming loss (HL): suppose Δ corresponds to the symmetric difference between two sets, the metric evaluates the fraction of instance-label pairs which have been misclassified.

$$HL = \frac{1}{tq}\sum_{i=1}^{t}|Y_i' \Delta Y_i|, \tag{18}$$

One-error (OE): the metric evaluates the fraction of examples whose top-ranked label is not in the relevant label set.

$$OE = \frac{1}{t}\sum_{i=1}^{t}\left[\left[\left[\arg\max_{l_k \in L} f_k(\mathbf{x}_i)\right] \notin Y_i\right]\right], \tag{19}$$

Coverage (CV): suppose $rank(\mathbf{x}_i, l_k) = \sum_{j=1}^{q}[\![f_j(\mathbf{x}_i) \ge f_k(\mathbf{x}_i)]\!]$ returns the rank of l_k when all labels in L are sorted in descending order based on the q learned functions, the metric evaluates how many steps are needed, on average, to go down the label ranking list so as to cover all the ground-truth labels of the instance.

$$CV = \frac{1}{q}\left(\frac{1}{t}\sum_{i=1}^{t}\max_{l_k \in Y_i} rank(\mathbf{x}_i, l_k) - 1\right), \tag{20}$$

Ranking loss (RL): suppose $D_i = \{(l_j, l_k)|f_j(\mathbf{x}_i) \ge f_k(\mathbf{x}_i), (l_j, l_k) \in Y_i \times \overline{Y_i}\}$, and $\overline{Y_i}$ is the complementary set of Y_i in L, the metric evaluates the fraction of reversely ordered label pairs.

$$RL = \frac{1}{t}\sum_{i=1}^{t}\frac{|D_i|}{|Y_i||\overline{Y_i}|}, \tag{21}$$

Average precision (AP): suppose $L_i = \{l_j|rank(\mathbf{x}_i, l_j) \le rank(\mathbf{x}_i, l_k)\}$, the metric evaluates the average fraction of relevant labels ranked higher than a particular label $l_k \in Y_i$.

$$AP = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{|Y_i|}\sum_{l_j, l_k \in Y_i}\frac{|L_i|}{rank(\mathbf{x}_i, l_k)}, \tag{22}$$

In short, the performance of multi-label algorithms can obtain the objective evaluation data by these metrics. In general, few algorithms can perform better than other algorithms on all evaluation metrics. Specially, for hamming loss, one-error, coverage and ranking loss, the smaller the value, the better the system performance, and the larger value of average precision indicates the better performance.

5.3. Results of multi-label classification

In experiments, 10-fold cross validation is used for evaluating the performance systematically. In detail, all samples are randomly divided into 10 equal subsets, each subset is held-out in turn for test, while the remaining data is merged to training. As the validation is iterated 10 times, the averaged metric values out of ten runs are calculated for the algorithms.

Tables 2–6 demonstrate the experimental results, in which the best performance among all the algorithms is highlighted in boldface. Based on these experimental results, some observations can be made:

(1) Compared with the selected state-of-the-art approaches, we can see that MLFC outperforms Rank-SVM, MLNB, and LLSF on all the data sets, and performs better on 6 out of the 13 data sets in terms of all evaluation metrics. For all the data sets, the proposed algorithm is superior to these multi-label approaches on hamming loss, one-error, and average precision.

(2) MLFC-SVM obtains the more balanced results than MLFC, which is worse than MLFC on some data sets, such as emotions and yeast, but performs better than the other comparing algorithms in most cases on all the data sets. Besides, by comparing MLFC-SVM with LIFT, we can see that the proposed label-specific feature learning method gets the better results.

(3) Some special issues are elaborated. For example, MLFC-SVM obtains the best results on TCM1 among all the algorithms, and MLFC and LIFT have their advantages in terms of syndrome differentiation on TCM2. Since the optimization target of MLNB is constructed with hamming loss and ranking loss, MLNB can receive the best results on genbase, TCM2, and corel5k regarding ranking loss, but has unsatisfying results on the other metrics.

In addition, for demonstration purposes, the experimental results from Tables 2–6 on each evaluation metric are also presented in Fig. 1. In these subgraphs, the horizontal and vertical axes denote the sequence number of data sets and the prediction performance respectively. According to Fig. 1, we can obtain the same conclusions with the above observations, and the proposed algorithms achieve a competitive performance against the state-of-the-art multi-label methods.

Table 2
Prediction performance of each comparing algorithm (mean ± std. deviation) in terms of Hamming Loss.

Data Set	MLFC	MLFC-SVM	Rank-SVM	ML-KNN	MLNB	LLSF	LIFT
emotion	**0.184** ± 0.019	0.195 ± 0.020	0.390 ± 0.019	0.194 ± 0.014	0.198 ± 0.015	0.202 ± 0.019	0.197 ± 0.019
genbase	**0.001** ± 0.001	**0.001** ± 0.001	0.064 ± 0.006	0.004 ± 0.003	0.046 ± 0.003	**0.001** ± 0.000	0.002 ± 0.001
medical	**0.010** ± 0.001	**0.010** ± 0.002	0.038 ± 0.003	0.015 ± 0.002	0.024 ± 0.001	0.012 ± 0.001	0.013 ± 0.002
TCM1	**0.050** ± 0.005	**0.050** ± 0.003	0.111 ± 0.005	0.083 ± 0.008	0.078 ± 0.004	0.060 ± 0.008	**0.050** ± 0.005
TCM2	**0.005** ± 0.000	**0.005** ± 0.000	0.007 ± 0.001	**0.005** ± 0.000	0.006 ± 0.001	0.007 ± 0.001	**0.005** ± 0.000
yeast	**0.194** ± 0.008	0.213 ± 0.010	0.232 ± 0.005	0.195 ± 0.010	0.209 ± 0.010	0.201 ± 0.009	0.202 ± 0.008
arts	**0.053** ± 0.003	**0.053** ± 0.001	0.075 ± 0.019	0.057 ± 0.002	0.068 ± 0.002	0.054 ± 0.002	**0.053** ± 0.003
computer	0.033 ± 0.002	**0.032** ± 0.002	0.044 ± 0.002	0.035 ± 0.002	0.048 ± 0.003	0.034 ± 0.002	0.033 ± 0.002
corel5k	**0.009** ± 0.000	**0.009** ± 0.000	0.012 ± 0.001	**0.009** ± 0.000	0.014 ± 0.001	**0.009** ± 0.000	**0.009** ± 0.000
education	**0.037** ± 0.001	**0.037** ± 0.001	0.056 ± 0.002	0.038 ± 0.001	0.051 ± 0.002	**0.037** ± 0.001	**0.037** ± 0.001
science	0.031 ± 0.001	**0.030** ± 0.001	0.050 ± 0.002	0.032 ± 0.001	0.047 ± 0.002	0.031 ± 0.001	0.031 ± 0.001
social	0.020 ± 0.001	**0.019** ± 0.001	0.037 ± 0.001	0.021 ± 0.001	0.040 ± 0.002	0.021 ± 0.001	**0.019** ± 0.001
society	**0.051** ± 0.002	**0.051** ± 0.002	0.063 ± 0.002	0.053 ± 0.002	0.064 ± 0.003	0.052 ± 0.002	**0.051** ± 0.002

Table 3
Prediction performance of each comparing algorithm (mean ± std. deviation) in terms of One-error.

Data Set	MLFC	MLFC-SVM	Rank-SVM	ML-KNN	MLNB	LLSF	LIFT
emotion	**0.244** ± 0.061	0.256 ± 0.046	0.345 ± 0.060	0.276 ± 0.027	0.271 ± 0.033	0.286 ± 0.029	0.261 ± 0.066
genbase	0.003 ± 0.006	0.002 ± 0.005	0.743 ± 0.052	**0.000** ± 0.000	1.000 ± 0.000	0.003 ± 0.006	**0.000** ± 0.000
medical	**0.121** ± 0.029	0.130 ± 0.033	0.728 ± 0.052	0.246 ± 0.037	0.318 ± 0.033	0.170 ± 0.028	0.155 ± 0.039
TCM1	**0.104** ± 0.029	0.092 ± 0.031	0.590 ± 0.060	0.339 ± 0.024	0.259 ± 0.037	0.147 ± 0.032	0.165 ± 0.031
TCM2	0.560 ± 0.066	0.577 ± 0.051	0.900 ± 0.028	0.600 ± 0.059	0.568 ± 0.045	0.641 ± 0.041	**0.542** ± 0.048
yeast	**0.213** ± 0.031	0.232 ± 0.015	0.252 ± 0.024	0.236 ± 0.030	0.245 ± 0.024	0.229 ± 0.024	0.229 ± 0.029
arts	0.450 ± 0.024	**0.448** ± 0.016	0.770 ± 0.026	0.542 ± 0.029	0.595 ± 0.020	0.456 ± 0.017	0.458 ± 0.019
computer	0.343 ± 0.019	**0.337** ± 0.013	0.476 ± 0.028	0.386 ± 0.023	0.475 ± 0.016	0.357 ± 0.020	0.349 ± 0.029
corel5k	**0.640** ± 0.029	0.655 ± 0.023	0.977 ± 0.018	0.663 ± 0.023	0.803 ± 0.016	0.644 ± 0.013	0.679 ± 0.022
education	**0.456** ± 0.026	**0.456** ± 0.020	0.685 ± 0.022	0.492 ± 0.023	0.589 ± 0.019	0.465 ± 0.020	0.471 ± 0.022
science	0.480 ± 0.022	**0.473** ± 0.030	0.765 ± 0.024	0.551 ± 0.018	0.617 ± 0.018	0.487 ± 0.015	0.480 ± 0.021
social	0.270 ± 0.019	**0.259** ± 0.021	0.575 ± 0.017	0.310 ± 0.012	0.406 ± 0.026	0.285 ± 0.022	0.268 ± 0.016
society	0.390 ± 0.018	**0.379** ± 0.021	0.503 ± 0.023	0.419 ± 0.018	0.485 ± 0.021	0.394 ± 0.019	0.387 ± 0.018

Table 4
Prediction performance of each comparing algorithm (mean ± std. deviation) in terms of Coverage.

Data Set	MLFC	MLFC-SVM	Rank-SVM	ML-KNN	MLNB	LLSF	LIFT
emotion	**0.286** ± 0.022	0.290 ± 0.023	0.411 ± 0.033	0.292 ± 0.024	1.794 ± 0.157	0.310 ± 0.019	0.293 ± 0.029
genbase	**0.013** ± 0.008	0.015 ± 0.008	0.184 ± 0.028	0.022 ± 0.017	21.066 ± 0.400	0.014 ± 0.005	0.016 ± 0.012
medical	**0.025** ± 0.010	0.035 ± 0.017	0.170 ± 0.014	0.055 ± 0.011	13.431 ± 0.483	0.048 ± 0.015	0.040 ± 0.011
TCM1	**0.136** ± 0.022	**0.136** ± 0.010	0.294 ± 0.016	0.237 ± 0.026	14.800 ± 0.576	0.177 ± 0.025	0.171 ± 0.024
TCM2	**0.160** ± 0.017	0.190 ± 0.023	0.314 ± 0.023	0.249 ± 0.021	179.817 ± 6.394	0.206 ± 0.026	0.193 ± 0.028
yeast	0.452 ± 0.014	0.496 ± 0.017	0.535 ± 0.020	**0.447** ± 0.012	6.499 ± 0.244	0.463 ± 0.020	0.464 ± 0.012
arts	0.217 ± 0.012	**0.168** ± 0.009	0.254 ± 0.013	0.187 ± 0.007	6.515 ± 0.233	0.219 ± 0.012	0.172 ± 0.010
computer	0.138 ± 0.012	**0.101** ± 0.006	0.152 ± 0.009	0.111 ± 0.009	5.275 ± 0.514	0.148 ± 0.008	0.104 ± 0.009
corel5k	0.424 ± 0.011	0.292 ± 0.016	0.735 ± 0.036	**0.283** ± 0.011	128.444 ± 3.251	0.447 ± 0.008	0.294 ± 0.015
education	0.165 ± 0.016	**0.099** ± 0.006	0.146 ± 0.007	0.103 ± 0.007	5.374 ± 0.566	0.170 ± 0.015	0.102 ± 0.004
science	0.181 ± 0.015	**0.128** ± 0.011	0.213 ± 0.012	0.145 ± 0.008	6.371 ± 0.332	0.185 ± 0.017	0.130 ± 0.008
social	0.106 ± 0.008	**0.067** ± 0.008	0.113 ± 0.007	0.073 ± 0.004	6.701 ± 0.495	0.121 ± 0.014	0.068 ± 0.005
society	0.241 ± 0.013	**0.183** ± 0.008	0.253 ± 0.011	0.192 ± 0.009	6.940 ± 0.187	0.244 ± 0.012	0.187 ± 0.010

To further analyze the performance among all the algorithms, Friedman test [47] is applied as the favorable statistical significance test for the method comparison on various data sets. Table 7 illustrates the Friedman statistic F_F and the corresponding critical values on each evaluation metric, and we can see from Table 7 that the null hypothesis, which follows the principle that all the algorithms have equal performance, is clearly rejected in terms of each evaluation metric at significance level $\alpha = 0.05$. Thus, we proceed with certain post-hoc test [47] to complete the performance analysis, and the Nemenyi test [47] is utilized to serve this purpose where MLFC or MLFC-SVM is regarded as the control algorithm respectively. Further, the difference between two algorithms is distinguished with the critical difference (CD), as follow:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{23}$$

where $q_\alpha = 2.949$ at significance level $\alpha = 0.05$, and then we can calculate $CD = 2.4987$ ($k = 7$, $N = 13$).

Fig. 2 shows the CD diagrams [47] regarding each evaluation metric. In each sub-figure, any comparing algorithm whose average rank is within one CD to that of MLFC or MLFC-SVM is connected. Otherwise, the algorithm, which is not connected with MLFC or MLFC-SVM, is considered to have the significant different performance with the control algorithm. From Fig. 1, we can observe that the proposed algorithm significantly outperforms Rank-SVM and MLNB, and obtains the better performance than ML-KNN in terms of the metrics of one-error and average precision, while on the metrics of coverage and ranking loss, MLFC-SVM performs better compared with LLSF. In addition, the proposed algorithm achieves statistically superior performance against LIFT regarding label-specific feature learning. The better performance of the proposed algorithm against LIFT indicates the effectiveness of

Table 5
Prediction performance of each comparing algorithm (mean ± std. deviation) in terms of Ranking Loss.

Data Set	MLFC	MLFC-SVM	Rank-SVM	ML-KNN	MLNB	LLSF	LIFT
emotion	**0.145** ± 0.025	0.152 ± 0.021	0.290 ± 0.030	0.157 ± 0.017	0.162 ± 0.025	0.174 ± 0.020	0.155 ± 0.029
genbase	0.002 ± 0.003	0.003 ± 0.004	0.163 ± 0.022	0.007 ± 0.009	**0.000** ± 0.000	0.003 ± 0.003	0.004 ± 0.006
medical	**0.016** ± 0.007	0.020 ± 0.012	0.144 ± 0.012	0.037 ± 0.008	0.040 ± 0.008	0.035 ± 0.013	0.025 ± 0.007
TCM1	0.023 ± 0.004	**0.022** ± 0.003	0.123 ± 0.010	0.080 ± 0.009	0.056 ± 0.006	0.041 ± 0.006	0.040 ± 0.006
TCM2	0.086 ± 0.009	0.097 ± 0.013	0.185 ± 0.008	0.135 ± 0.013	**0.048** ± 0.008	0.115 ± 0.016	0.099 ± 0.017
yeast	**0.166** ± 0.011	0.202 ± 0.010	0.235 ± 0.017	0.167 ± 0.014	0.177 ± 0.014	0.174 ± 0.011	0.173 ± 0.009
arts	0.142 ± 0.009	**0.110** ± 0.007	0.191 ± 0.011	0.130 ± 0.006	0.153 ± 0.007	0.144 ± 0.008	0.113 ± 0.007
computer	0.094 ± 0.008	**0.066** ± 0.003	0.104 ± 0.006	0.074 ± 0.005	0.090 ± 0.007	0.103 ± 0.007	0.067 ± 0.006
corel5k	0.182 ± 0.007	0.124 ± 0.007	0.408 ± 0.035	0.121 ± 0.005	**0.118** ± 0.005	0.195 ± 0.004	0.124 ± 0.007
education	0.114 ± 0.011	**0.071** ± 0.004	0.117 ± 0.004	0.076 ± 0.005	0.089 ± 0.005	0.117 ± 0.011	0.073 ± 0.004
science	0.132 ± 0.013	**0.093** ± 0.009	0.172 ± 0.010	0.110 ± 0.006	0.123 ± 0.007	0.136 ± 0.011	0.094 ± 0.005
social	0.074 ± 0.005	**0.046** ± 0.005	0.088 ± 0.004	0.053 ± 0.004	0.068 ± 0.006	0.086 ± 0.012	0.047 ± 0.004
society	0.152 ± 0.006	**0.117** ± 0.007	0.176 ± 0.009	0.127 ± 0.006	0.152 ± 0.007	0.156 ± 0.008	0.119 ± 0.005

Table 6
Prediction performance of each comparing algorithm (mean ± std. deviation) in terms of Average Precision.

Data set	MLFC	MLFC-SVM	Rank-SVM	ML-KNN	MLNB	LLSF	LIFT
emotion	**0.818** ± 0.035	0.813 ± 0.028	0.684 ± 0.032	0.802 ± 0.017	0.798 ± 0.025	0.790 ± 0.023	0.805 ± 0.035
genbase	**0.997** ± 0.004	0.995 ± 0.004	0.426 ± 0.046	0.992 ± 0.009	0.060 ± 0.004	0.996 ± 0.004	0.995 ± 0.007
medical	**0.916** ± 0.020	0.905 ± 0.026	0.384 ± 0.036	0.816 ± 0.024	0.085 ± 0.002	0.875 ± 0.017	0.881 ± 0.029
TCM1	0.873 ± 0.019	**0.877** ± 0.019	0.497 ± 0.036	0.655 ± 0.024	0.218 ± 0.013	0.828 ± 0.022	0.803 ± 0.022
TCM2	0.468 ± 0.052	0.486 ± 0.034	0.169 ± 0.019	0.446 ± 0.046	0.012 ± 0.001	0.400 ± 0.033	**0.504** ± 0.040
yeast	**0.771** ± 0.017	0.735 ± 0.009	0.686 ± 0.019	0.753 ± 0.019	0.748 ± 0.015	0.759 ± 0.012	0.757 ± 0.013
arts	0.625 ± 0.016	**0.636** ± 0.013	0.407 ± 0.025	0.571 ± 0.020	0.321 ± 0.007	0.619 ± 0.010	0.627 ± 0.015
computer	0.710 ± 0.015	**0.721** ± 0.009	0.596 ± 0.018	0.686 ± 0.016	0.352 ± 0.031	0.702 ± 0.014	0.712 ± 0.020
corel5k	**0.302** ± 0.012	0.290 ± 0.015	0.067 ± 0.007	0.297 ± 0.011	0.039 ± 0.001	0.298 ± 0.008	0.285 ± 0.013
education	0.639 ± 0.017	**0.646** ± 0.013	0.468 ± 0.013	0.618 ± 0.015	0.293 ± 0.045	0.628 ± 0.018	0.639 ± 0.015
science	0.604 ± 0.016	**0.615** ± 0.019	0.373 ± 0.014	0.557 ± 0.013	0.502 ± 0.013	0.597 ± 0.015	0.610 ± 0.016
social	0.777 ± 0.013	**0.793** ± 0.018	0.566 ± 0.016	0.759 ± 0.009	0.203 ± 0.014	0.764 ± 0.017	0.788 ± 0.012
society	0.635 ± 0.011	**0.653** ± 0.015	0.524 ± 0.020	0.626 ± 0.015	0.336 ± 0.009	0.633 ± 0.015	0.646 ± 0.013

utilizing a robust optimization function and modeling intraclass and innerclass distances. Futhermore, LIFT loses sight in label correlations exploitation, whereas the proposed algorithm has designed an effective manner to solve this key problem by generating label correlations-based features, which can be incorporated with the original features freely.

5.4. Properties of MLFC

We further study the properties of MLFC. In light of the influence of label correlations, we conduct experiments to analyze label correlations-based features, and the variation of the results with iterative optimization.

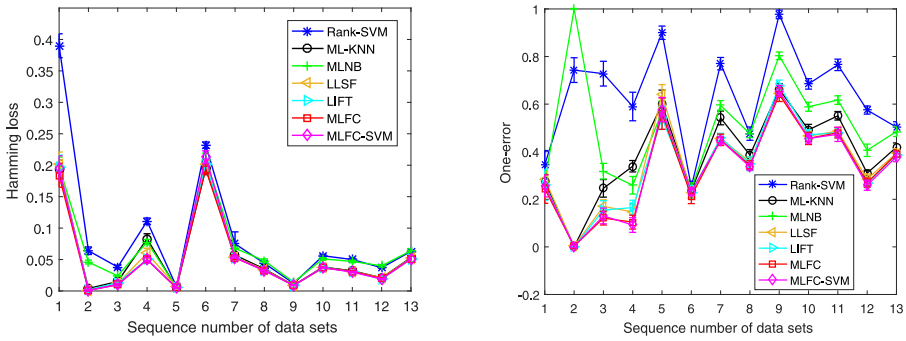
5.4.1. Influence of label correlations

To explain the influence of label correlations, we compare MLFC with the label-specific feature learning method labeled as MLLF, which can be viewed as a degenerated version of the proposed algorithm without considering label correlations-based features. For simplicity, we report the results on four data sets including emotions, TCM1, yeast, and social in regard to each evaluation metric, as shown in Table 8. From Table 8, we can make the conclusion that exploiting the correlations among labels is helpful to improve the performance. Explanatorily, compared with MLLF learning from the original feature space, the proposed algorithm can obtain the better results on these four data sets via incorporating label correlations-based features. Intuitively, it's worth noting that the influenced degree of the label correlations-based features has difference, which are constructed based on different multi-label data sets, and the experiments reveal the performance has a marked improvement in terms of emotions, yeast, and social, whereas the improvement on TCM1 is relatively less.

5.4.2. Result analysis with iterative optimization

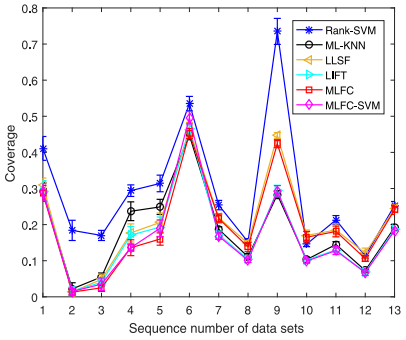
As described in Section 4, the proposed learning framework is constructed with an iterative updating scheme, and the estimation of unknown variable sets, namely W and C, is the crucial link to study the convergence rate. Once one of unknown variable sets converges, the function value reaches a stable stage. Detailedly, the optimization problem is convex w.r.t. C, and can be easily solved to obtain the optimal solution. Nevertheless, the solution for W is based on accelerated proximal gradient, which is designed by iterative shrinkage-thresholding and not a monotone method [37]. For revealing the impact on the convergence of the proposed alternating optimization, as the literatures suggested [18,48], we record the change of the prediction performance with respect to each iteration.

We demonstrate the results of MLFC on hamming loss in terms of emotions and yeast in Fig. 3, in which the horizontal and vertical axes represent the number of iterations and hamming loss respectively, and experiments on other data sets obtain the similar results. Fig. 3(a) shows the optimization results of the proposed algorithm on emotions, and the results of MLNB, ML-KNN, LLSF, and LIFT on hamming loss are also plotted in the sub-figure, while the results of Rank-SVM are not plotted because they are inconvenient to exhibit properly. Concretely, we can see from Fig. 3(a) that the proposed algorithm tends to achieve a better performance with the increasing of the number of iterations, and outperforms other multi-label methods at the second iteration. In addition, hamming loss decreases dramatically, and becomes stable after 7 iterations. Similarly, Fig. 3(b) shows the results on yeast, in which we can observe that the variation trends on hamming loss are similar to emotions, and the optimization result is in minor change within 13 iterations. Thus, we can summarize that the proposed algorithm is feasible and effective with iterative optimization.

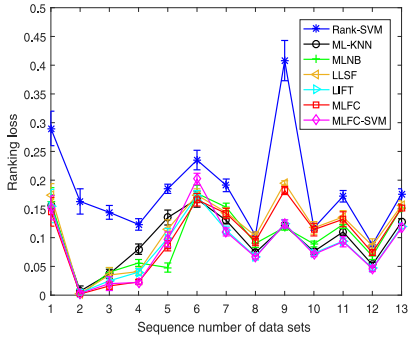


(a) Hamming loss

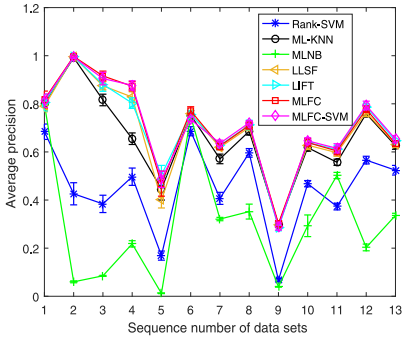
(b) One-error



(c) Coverage



(d) Ranking loss



(e) Average precision

Fig. 1. Performance comparison of MLFC and MLFC-SVM with other comparing algorithms on all the data sets.

Table 7
Summary of the Friedman statistics F_F ($k = 7$, $N = 13$) and the critical value on evaluation metrics(k: comparing algorithms; N: data sets).

Evaluation metric	F_F	Critical value($\alpha = 0.05$)
Hamming Loss	24.2721	
One-error	35.7176	
Coverage	33.6647	2.25
Ranking Loss	16.0138	
Average Precision	35.6576	

6. Discussion and conclusion

This work introduced a novel multi-label algorithm, which combined two independent multi-label learning problems, namely label correlations and label-specific feature learning, into one problem and addressed via the optimization method. Specifically, an optimization framework was designed to learn the most class-discriminative features of each label, and the influence of label correlations was considered in this optimization procedure by generating additional features for matching the correlations among labels. Based on the hypothesis, feature weights and label correlation-based features were defined as two unknown variables, and a iterative optimization method was proposed to obtain the optimal solution, in which non-zero components of the learned weight matrix indicated label-specific features of all class labels. Then we constructed label correlations-based features for test samples, and achieved the multi-label classification by using the linear and SVM models respectively. Finally in experiments, the results on various multi-label data sets including two collected TCM data sets demonstrated that the proposed method is feasible and effective. Compared with some other state-of-the-art multi-label learning approaches, the proposed optimization framework had the advantage with respect to the prediction performance, and it is conformed that the performance of the proposed algorithm had a general improvement via resolving label correlations.

Our proposed algorithm can be further improved by using other loss functions for label-specific feature learning or other measure methods to fit the correlations among labels. Besides, the initialization of the

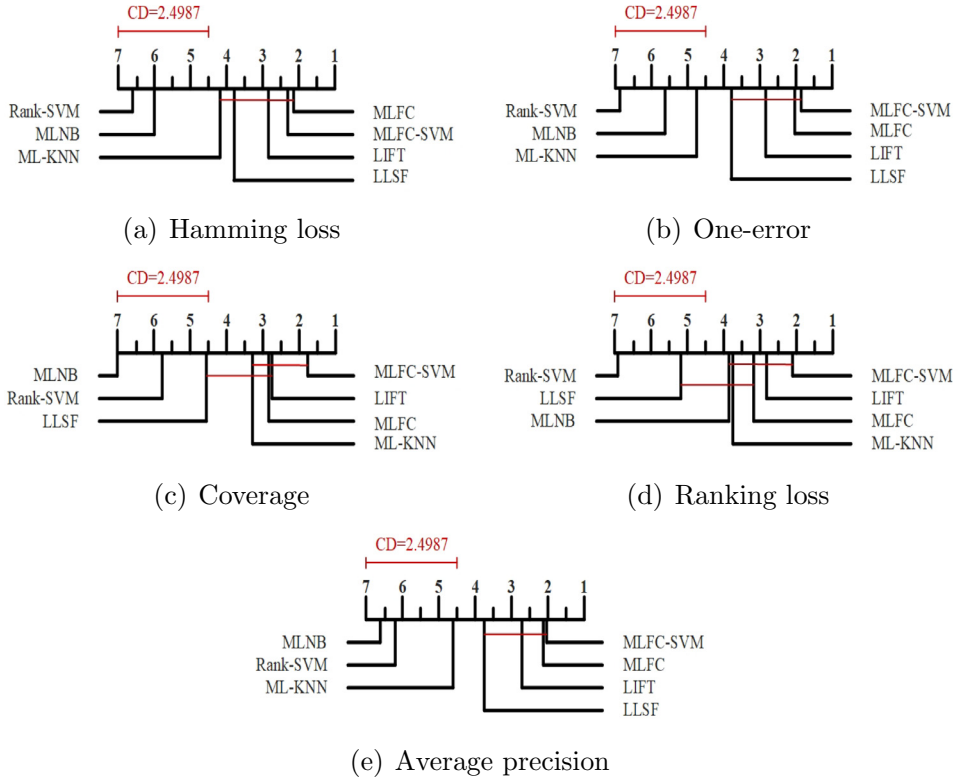


Fig. 2. Comparison of MLFC or MLFC-SVM (control algorithm) against other comparing algorithms with the Nemenyi test.

Table 8
Influence of label correlations.

Data set	Method	Hamming loss	One-error	Coverage	Ranking loss	Average precision
emotion	MLFC	**0.184** ± 0.019	**0.244** ± 0.061	**0.286** ± 0.022	**0.145** ± 0.025	**0.818** ± 0.035
	MLLF	0.199 ± 0.026	0.265 ± 0.058	0.298 ± 0.034	0.160 ± 0.027	0.805 ± 0.027
TCM1	MLFC	**0.050** ± 0.005	**0.104** ± 0.029	**0.136** ± 0.022	**0.023** ± 0.004	**0.873** ± 0.019
	MLLF	0.052 ± 0.003	0.107 ± 0.030	0.137 ± 0.012	**0.023** ± 0.004	0.868 ± 0.015
yeast	MLFC	**0.194** ± 0.008	**0.213** ± 0.031	**0.452** ± 0.014	**0.166** ± 0.011	**0.771** ± 0.017
	MLLF	0.200 ± 0.005	0.224 ± 0.024	0.460 ± 0.015	0.173 ± 0.011	0.761 ± 0.015
social	MLFC	**0.020** ± 0.001	**0.270** ± 0.019	**0.106** ± 0.008	**0.074** ± 0.005	**0.777** ± 0.013
	MLLF	0.021 ± 0.001	0.281 ± 0.019	0.115 ± 0.012	0.082 ± 0.009	0.768 ± 0.015

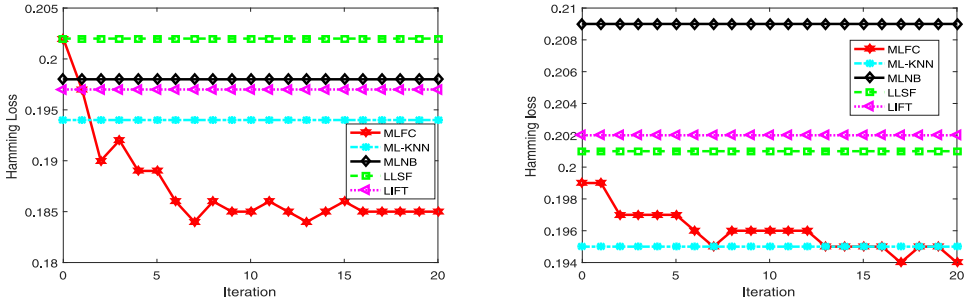


Fig. 3. Hamming loss w.r.t. # iterations on emotions and yeast data sets.

proposed optimization scheme from clustering techniques is typically a good start for iterative optimization, which can be replaceable to improve the rate of convergence. Furthermore, there may be better methods to predict label correlations-based features of test data, such as large margin method. Based on the above analysis, we have interest in further study of these variants. Furthermore, since the TCM syndrome diagnosis is based on comprehensive signals from the multi-model data, including inspection, listening and smelling, inquiry, and palpation, we will pay close attention to the research of multi-label learning via fusing multiple feature modalities.

Acknowledgments

This work is supported by grants from the National Natural Science

Foundation of China (nos. 61572409, U1705286, 61571188 & 61672272), Fujian Province 2011 Collaborative Innovation Center of TCM Health Management and Collaborative Innovation Center of Chinese Oolong Tea Industry-Collaborative Innovation Center (2011) of Fujian Province, Fund for Integration of Cloud Computing and Big Data, Innovation of Science and Education.

References

[1] G. Tsoumakas, E.S. Xioufis, J. Vilcek, I.P. Vlahavas, MULAN: a java library for multi-label learning, J. Mach. Learn. Res. 12 (2011) 2411–2414.
[2] M. Zhang, Z. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[3] A. McCallum, Multi-label text classification with a mixture model trained by EM, Working Notes of the AAAI'99 Workshop on Text Learning, Orlando, FL, (1999).
[4] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2/3) (2000) 135–168.
[5] Z. Fang, Z. Zhang, Simultaneously combining multi-view multi-label learning with maximum margin classification, 12th IEEE International Conference on Data Mining, Brussels, Belgium, (2012), pp. 864–869.
[6] J. Fürnkranz, E. Hüllermeier, E.L. Mencía, K. Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2) (2008) 133–153.
[7] J. Huang, G. Li, Q. Huang, X. Wu, Joint feature selection and classification for multilabel learning, IEEE Trans. Cybern. 48 (3) (2018) 876–889.
[8] F. Li, D. Miao, W. Pedrycz, Granular multi-label feature selection based on mutual information, Pattern Recognit. 67 (2017) 410–423.
[9] X. Cheng, S. Zhao, X. Xiao, K. Chou, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (3) (2017) 341–346.
[10] H. Kazawa, T. Izumitani, H. Taira, E. Maeda, Maximal margin labeling for multi-topic text categorization, Advances in Neural Information Processing Systems 17, Vancouver, British Columbia, Canada, 2004, pp. 649–656.
[11] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, H. Zhang, Correlative multi-label video annotation, Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, (2007), pp. 17–26.
[12] P. Gu, H. Chen, Modern bioinformatics meets traditional Chinese medicine, Brief. Bioinf. 15 (6) (2014) 984–1003.
[13] J. Zhang, C. Li, Y. Lin, Y. Shao, S. Li, Computational drug repositioning using collaborative filtering via multi-source fusion, Expert Syst. Appl. 84 (2017) 281–289.
[14] G. Liu, G. Li, Y. Wang, Y. Wang, Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning, BMC Complem. Altern. Med. 10 (2010) 37.
[15] W. Wu, J. Liu, H. Chang, Latent class model based diagnostic system utilizing traditional chinese medicine for patients with systemic lupus erythematosus, Expert Syst. Appl. 38 (1) (2011) 281–287.
[16] B. Qian, X. Wang, J. Ye, I. Davidson, A reconstruction error based framework for multi-label and multi-view learning, IEEE Trans. Knowl. Data Eng. 27 (3) (2015) 594–607.
[17] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (3) (2011) 333–359.
[18] G. Tsoumakas, I.P. Vlahavas, Random *k*-labelsets: an ensemble method for multi-label classification, Proceedings of the 18th European Conference on Machine Learning, Warsaw, Poland, (2007), pp. 406–417.
[19] S. Xu, X. Yang, H. Yu, D. Yu, J. Yang, E.C.C. Tsang, Multi-label learning with label-specific feature reduction, Knowl.-Based Syst. 104 (2016) 52–61.
[20] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.
[21] M. Zhang, Z. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.
[22] Q. Wu, M. Tan, H. Song, J. Chen, M.K. Ng, ML-FOREST: a multi-label tree ensemble method for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (10) (2016) 2665–2680.
[23] J. Lee, D. Kim, Memetic feature selection algorithm for multi-label classification, Inf. Sci. 293 (2015) 80–96.
[24] J. Liu, Y. Lin, M. Lin, S. Wu, J. Zhang, Feature selection based on quality of information, Neurocomputing 225 (2017) 11–22.
[25] M. Zhang, J.M.P. Sánchez, V. Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19) (2009) 3218–3229.
[26] H. Peng, F. Long, C.H.Q. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[27] Y. Lin, Q. Hu, J. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[28] M. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 107–120.
[29] M. Zhang, Y. Li, X. Liu, Towards class-imbalance aware multi-label learning, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, (2015), pp. 4041–4047.
[30] Y. Lin, Q. Hu, J. Zhang, X. Wu, Multi-label feature selection with streaming labels, Inf. Sci. 372 (2016) 256–275.
[31] M. Zhang, Z. Zhou, Multi-label neural networks with applications to functional genomics and text categorization, IEEE Trans. Knowl. Data Eng. 18 (10) (2006) 1338–1351.
[32] S. Huang, Z. Zhou, Multi-label learning by exploiting label correlations locally, Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada, (2012), pp. 945–955.
[33] E.L. Mencía, F. Janssen, Learning rules for multi-label classification: a stacking and a separate-and-conquer approach, Mach. Learn. 105 (1) (2016) 77–126.
[34] Y. Nesterov, Introductory lectures on convex optimization: a basic course, Springer, 2004.
[35] Z. Zhang, Y. Xu, J. Yang, X. Li, D. Zhang, A survey of sparse representation: algorithms and applications, IEEE Access 3 (2015) 490–530.
[36] M. Schmidt, G. Fung, R. Rosaless, Optimization methods for L1-regularization, Tech. Rep. TR-2009-19, Univ. British Columbia, Vancouver, BC, Canada, 2009.
[37] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE Trans. Image Process. 18 (11) (2009) 2419–2434.
[38] P. Zhao, B. Yu, On model selection consistency of lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
[39] S. Lee, H. Lee, P. Abbeel, A.Y. Ng, Efficient L1 regularized logistic regression, Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Massachusetts, USA, (2006), pp. 401–408.
[40] V. Roth, The generalized LASSO, IEEE Trans. Neural Networks 15 (1) (2004) 16–28.
[41] K. Koh, S. Kim, S.P. Boyd, An interior-point method for large-scale l_1-regularized logistic regression, J. Mach. Learn. Res. 8 (2007) 1519–1555.
[42] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imag. Sci. 2 (1) (2009) 183–202.
[43] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, Advances in Neural Information Processing Systems 14, Vancouver, British Columbia, Canada, (2001), pp. 681–687.
[44] J. Huang, G. Li, Q. Huang, X. Wu, Learning Label Specific Features for Multi-label Classification, 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA, (2015), pp. 181–190.
[45] C. Chang, C. Lin, LIBSVM: A library for support vector machines, ACM trans, Intell. Syst. Technol. 2 (3) (2011) 1–27. 27:27
[46] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, J. Han, Conflicts to harmony: a framework for resolving conflicts in heterogeneous data by truth discovery, IEEE Trans. Knowl. Data Eng. 28 (8) (2016) 1986–1999.
[47] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[48] J. Wu, S. Zhao, V.S. Sheng, J. Zhang, C. Ye, P. Zhao, Z. Cui, Weak labeled active learning with conditional label dependence for multi-label image classification, IEEE Trans. Multimedia 19 (6) (2017) 1156–1169.