Multi-label Feature Selection via Global Relevance and Redundancy Optimization

Jia Zhang[1] , **Yidong Lin**[2] , **Min Jiang**[1] , **Shaozi Li**[1,*] , **Yong Tang**[3] and **Kay Chen Tan**[4]
[1]Department of Artificial Intelligence, Xiamen University, China
[2]School of Mathematical Sciences, Xiamen University, China
[3]School of Computer Science, South China Normal University, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong
{j.zhang, linyidong}@stu.xmu.edu.cn, {minjiang, szlig}@xmu.edu.cn, ytang@m.scnu.edu.cn, kaytan@cityu.edu.hk

Abstract

Information theoretical based methods have attracted a great attention in recent years, and gained promising results to deal with multi-label data with high dimensionality. However, most of the existing methods are either directly transformed from heuristic single-label feature selection methods or inefficient in exploiting labeling information. Thus, they may not be able to get an optimal feature selection result shared by multiple labels. In this paper, we propose a general global optimization framework, in which feature relevance, label relevance (i.e., label correlation), and feature redundancy are taken into account, thus facilitating multi-label feature selection. Moreover, the proposed method has an excellent mechanism for utilizing inherent properties of multi-label learning. Specially, we provide a formulation to extend the proposed method with label-specific features. Empirical studies on twenty multi-label data sets reveal the effectiveness and efficiency of the proposed method. Our implementation of the proposed method is available online at: https://jiazhang-ml.pub/GRRO-master.zip.

1 Introduction

Multi-label learning deals with examples which may be associated with multiple labels simultaneously. It has attracted significant interests from the research community and has a wide range of applications. For example, in text categorization, a text needs to be tagged with several topics [Schapire and Singer, 2000]; in image annotation, an image needs to be tagged with multiple scenes [Boutell *et al.*, 2004]; in bioinformatics, one wishes to recognize a gene with multiple functions [Elisseeff and Weston, 2001]. Normally, the aforementioned resources (e.g., text, image, and gene) are represented by feature vectors with high dimensionality. The high dimensionality of multi-label data not only leads to the increasing of the computational cost and memory storage requirement, but also limits the usage of machine learning models in real applications [Jian *et al.*, 2016]. Feature selection is proved to be effective for removing irrelevant and redundant features in the feature representation, thus carrying the most discriminative information for multi-label learning [Lee and Kim, 2017; Lin *et al.*, 2016; Wei and Li, 2019; Zhang *et al.*, 2019].

However, it is unwise to perform feature selection for multi-label learning directly. It means that achieving the purpose in a label-wise effective manner is deemed to be crucial for improving the generalization performance. First, labels in multi-label data are not independent but inherently correlated. For example, an image is likely to be annotated as *sky* if it has label *cloud*. Thus, it is necessary to capture label correlation to guide the feature selection process. Second, labels have their own inherent properties (e.g., the issues of the class-imbalance [Krawczyk, 2016], the relative labeling-importance [Li *et al.*, 2015] and label-specific features [Zhang and Wu, 2015]), and utilizing these inherent properties is also beneficial for multi-label feature selection.

To tackle the learning problem, a large family of existing algorithms is information theoretical based methods [Vergara and Estévez, 2014]. Algorithms in this family mainly focus on exploiting feature evaluation criteria, such as feature relevance maximization and feature redundancy minimization [Peng *et al.*, 2005], to access the importance of features. For implementation, they select candidate features one by one with heuristic search, until obtain a size-specific subset of relevant features. Nevertheless, these methods are easily trapped in local optima. Not surprisingly, they may be in trouble to find an optimal feature subset. Furthermore, such heuristic search is time-consuming while ineffective and repetitive calculations are involved in the criteria function.

In order to address the aforementioned problems, we propose a new multi-label feature selection method via global relevance and redundancy optimization, named GRRO. In particular, we present a global optimization method with the goal of considering feature relevance, label relevance (i.e., label correlation), and feature redundancy for feature evaluation. Allowing for achieving the goal efficiently, the proposed method only needs to go through the relevance and redundancy information one time, and can be easily solved for generating the optimal solution. By analyzing discriminative features for each label, we also give an extension of the proposed method to conduct label-specific feature selection. Extensive experiments on twenty multi-label data sets demonstrate the advantages of the proposed method.

*Shaozi Li is the corresponding author.

2 Related Work

Towards information theoretical based methods for multi-label feature selection, one straightforward way for feature evaluation is to maximize the relevance of candidate feature (e.g., f^+) with all labels on label set L [Li *et al.*, 2018].

$$J_{Max\_Rel}(f^+) = I(f^+, L) \tag{1}$$

where $I(.,.)$ denotes the function for mutual information estimation. In Eq. (1), it describes the decreased uncertainty for f^+ while labeling information L is given, that is, their shared information. Based on this, some other feature evaluation criteria are proposed, which can be roughly categorized into two groups: Feature redundancy minimization and newly classification information maximization.

Methods in the first group is based on the assumption that good features should be strongly correlated with labels, but not be highly correlated with each other. Thus, the general framework to evaluate f^+ can be formulated as follow:

$$J_{Red}(f^+) = I(f^+, L) - I(f^+, \mathcal{S}) \tag{2}$$

where $\mathcal{S}$ denotes the selected feature subset, and $I(f^+, \mathcal{S})$ denotes the redundant information between f^+ and $\mathcal{S}$. Lots of methods in the first group focus on modeling feature redundancy to improve the performance. For example, many methods used multivariate mutual information to measure the conditional redundancy under multiple labels [Lee and Kim, 2013; Lin *et al.*, 2015]. Besides, some authors give a concern for computational efficiency. For example, Lee and Kim [Lee and Kim, 2015] proposed a fast multi-label feature selection method by discarding unnecessary calculations and reusing pre-calculated entropy terms. They also proposed a scalable relevance criterion for large label set [Lee and Kim, 2017], which can estimate feature relevance and feature redundancy efficiently.

For the second group, methods in this group are expected to guarantee that the selected feature subset has a strong predictive ability with the smaller redundancy. Specially, the general framework of the methods can be defined as $I(f^+, L|\mathcal{S})$, which quantifies the new classification information provided by f^+ while $\mathcal{S}$ is given. For multi-label feature selection, some methods extended the criterion to the classification model [Sechidis *et al.*, 2014], or directly adopted this criterion to achieve the purpose [Bermejo *et al.*, 2018].

By employing the aforementioned criteria, heuristic strategy is widely used to rank features. Another strategy regards the feature selection process as an optimization problem, but the research is still lacking. A few methods tried to obtain the optimized feature weight using the aforementioned criteria [Lim and Kim, 2016; Sun *et al.*, 2019]. However, these methods have the poor ability to exploit labeling information. Moreover, they utilize iterative optimization with the gradient descent strategy to approximate the solution, which may cause the inefficiency for large-scale data analysis.

3 The Proposed Method

3.1 Preliminaries

We denote matrices with bold uppercase letters (e.g., $\mathbf{A}$), vectors with bold lowercase letters (e.g., $\mathbf{a}$), the (i,j)-th element of $\mathbf{A}$ as a_{ij}, and the i-th row and column of $\mathbf{A}$ as $\mathbf{a}_{i.}$ and $\mathbf{a}_{.i}$ respectively. The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}^T$, and the trace of $\mathbf{A}$ is denoted by $tr(\mathbf{A})$. Suppose that in the multi-label data set, there are d features, we denote the training data matrix as $\mathbf{F} = [\mathbf{f}_{.1}, \mathbf{f}_{.2}, ..., \mathbf{f}_{.d}]$, where $\mathbf{f}_{.i}$ $(1 \leq i \leq d)$ is the vector which contains the information of feature f_i. Then each instance can be denoted by a d-dimensional feature vector. Additionally, each instance is associated with a finite set of q possible labels $L = \{l_1, l_2, ..., l_q\}$. Let $\mathbf{Y} = [\mathbf{y}_{.1}, \mathbf{y}_{.2}, ..., \mathbf{y}_{.q}]$ be the label matrix, $\mathbf{y}_{.j}$ $(1 \leq j \leq q)$ contains the ground-truth label information of all training data on label l_j. Arbitrary element in $\mathbf{Y}$ whose value is 1 indicates that a label is relevant to an instance, otherwise the value is -1.

3.2 Global Relevance and Redundancy Optimization - GRRO

Aimed at seeking relevant features for multi-label feature selection, leveraging information theory to exploit feature evaluation criteria is a popular and effective way. In general, feature selection in this way is with heuristic search, which easily leads to a suboptimal feature subset. In addition, performing feature selection in a label-wise effective manner contributes to multi-label learning, hence the appropriate usage of labeling information (e.g., label correlation) is a crucial step during the feature selection process.

Considering that many feature evaluation criteria are proposed to maximize feature relevance and minimize feature redundancy [Brown *et al.*, 2012] (Eq. (2) shows the general framework), we take the advantage to measure feature importance. Different to the heuristic search, we propose a global learning framework to achieve the purpose with optimization.

$$\max_{\mathbf{Z}} \quad \sum_{u=1}^{q}\sum_{i=1}^{d}(I(f_i, l_u)z_{iu} - \sum_{j=1}^{d} I(f_i, f_j)z_{iu}z_{ju}) \tag{3}$$

where $\mathbf{Z} \in \mathcal{R}^{d\times c}$ denotes the feature coefficient matrix, $z_{iu} \in \mathbf{Z}$ denotes the importance of feature f_i with respect to label l_u. $I(f_i, l_u)$ and $I(f_i, f_j)$ denote the mutual information of feature f_i with label l_u and feature f_j, respectively. From Eq. (3), we can see that the importance of a feature to a label is positively related to the relevance between the feature and the label (as shown in the first term), which is also limited by the redundancy of the feature with other features (as shown in the second term). Based on this, the importance of all features to each label can be identified by assigning feature weights, i.e., matrix $\mathbf{Z}$. To provide a compact formulation in a quadratic form, Eq. (3) can infer to the following optimization problem.

$$\min_{\mathbf{Z}} \quad ||\mathbf{Z} - \mathbf{C}||_F^2 + \sum_{u=1}^{q} \mathbf{z}_{.u}^T \mathbf{G} \mathbf{z}_{.u} \tag{4}$$

where $\mathbf{C}$ is the matrix which preserves the correlation between features and labels. For arbitrary element $c_{ij} \in \mathbf{C}$, $c_{ij} = I(f_i, l_j)$. $\mathbf{G}$ is the symmetric matrix containing the correlation information of features, whose arbitrary element $g_{ij} = I(f_i, f_j)$.

Next, we learn the global label relevance by exploiting second-order label correlation, thereby improving the generalization performance. According to the feature selection

mechanism in Eq. (4), feature selection result for each label is determined by a feature weight vector of $\mathbf{Z}$, such as vector $\mathbf{z}_{.u}$ for label l_u. For any two labels l_u and l_j, with the case that the two labels are strongly correlated, their feature selection results (i.e., the corresponding coefficients $\mathbf{z}_{.u}$ and $\mathbf{z}_{.j}$) should be similar. Otherwise, the distribution of discriminative features indicated by $\mathbf{z}_{.u}$ and $\mathbf{z}_{.j}$ should be in difference, hence the newly classification information for label l_u and label l_j can be protected. To achieve the purpose, a regularizer for matrix $\mathbf{Z}$ is defined as follow:

$$\sum_{u=1}^{q}\sum_{j=1}^{q} r_{uj}\mathbf{z}_{.u}^T\mathbf{z}_{.j} \tag{5}$$

where $r_{uj} = 1 - s_{uj}$, s_{uj} denotes the correlation between label l_u and label l_j. For simplicity, s_{uj} is calculated by the Cosine similarity.

Integrating Eq. (4) with Eq. (5), we can obtain the optimization objective function of GRRO. In the light of that $\sum_{u=1}^{q}\mathbf{z}_{.u}^T\mathbf{G}\mathbf{z}_{.u} = \sum_{u=1}^{q}(\mathbf{Z}^T\mathbf{G}\mathbf{Z})_{uu} = tr(\mathbf{Z}^T\mathbf{G}\mathbf{Z})$ and $\sum_{u=1}^{q}\sum_{j=1}^{q} r_{uj}\mathbf{z}_{.u}^T\mathbf{z}_{.j} = tr(\mathbf{R}\mathbf{Z}^T\mathbf{Z})$, the optimization objective function can be further written as the following form:

$$\min_{\mathbf{Z}} \quad ||\mathbf{Z}-\mathbf{C}||_F^2 + \alpha tr(\mathbf{Z}^T\mathbf{G}\mathbf{Z}) + \beta tr(\mathbf{R}\mathbf{Z}^T\mathbf{Z}) \tag{6}$$

where α and β are tradeoff parameters. It can be observed in Eq. (6) that feature coefficient matrix $\mathbf{Z}$ is involved in all the terms, which makes the optimized feature selection result affected by the feature relevance, feature redundancy, and label correlation simultaneously.

3.3 Discussions & Practical Issues

Several important issues are discussed to make the proposed GRRO method practical and complete.

We first introduce the solution of GRRO. From Eq. (6), we can see that matrices $\mathbf{G}$ and $\mathbf{R}$ are positive semidefinite, and $\mathbf{Z}^T\mathbf{G}\mathbf{Z} \geq 0$ and $\mathbf{R}\mathbf{Z}^T\mathbf{Z} \geq 0$ for any nonzero $\mathbf{Z}$ hold. Thus, we can get the solution for $\mathbf{Z}$ by setting the derivative of the objective function in Eq. (6) $w.r.t.$ $\mathbf{Z}$ to 0, as follow:

$$2(\mathbf{Z}-\mathbf{C}) + \alpha(\mathbf{G}+\mathbf{G}^T)\mathbf{Z} + \beta\mathbf{Z}(\mathbf{R}+\mathbf{R}^T) = 0 \tag{7}$$

Both $\mathbf{G}$ and $\mathbf{R}$ are symmetric matrices, therefore, we can transform Eq. (7) into the following one:

$$(\mathbf{I} + \alpha\mathbf{G})\mathbf{Z} + \beta\mathbf{Z}\mathbf{R} = \mathbf{C} \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix. Eq. (8) is the matrix equation with the form of **AZ + ZB = C**, where $\mathbf{A} = \mathbf{I} + \alpha\mathbf{G}$ and $\mathbf{B} = \beta\mathbf{R}$. To solve this equation, some existing methods [Wu *et al.*, 2014; Zhang *et al.*, 2019] can be employed to obtain matrix $\mathbf{Z}$. Here, we use the *Lyapunov* function[1] in Matlab to solve the mathematical problem. After that, the importance of each feature can be obtained based on the value of $||\mathbf{z}_{i.}||_2$ $(1 \leq i \leq d)$.

Time complexity. GRRO first calculates the correlation in terms of features and labels, and then uses these calculations to generate the optimal solution for feature selection. In this process, the time cost is dominated by these correlation calculations, which lead to a complexity of $\mathcal{O}(d^2 + q^2 + dq)$. It can be seen that the time complexity is quadratic regarding the number of features d and the number of labels q.

[1]https://www.mathworks.com/help/control/ref/lyap.html

Data set	Training	Test	Features	Labels	Domain
Bibtex	4880	2515	1836	159	Text
Birds	322	323	260	19	Audio
Corel5k	4500	500	499	374	Image
Corel16k001	9241	4525	500	153	Image
Corel16k002	9165	4596	500	164	Image
Emotions	391	202	72	6	Music
Genbase	463	199	1186	27	Biology
Image	1000	1000	294	5	Image
Langlog	978	482	1004	75	Text
Medical	645	333	1449	45	Text
Slashdot	2546	1236	1079	22	Text
Yeast	1499	918	103	14	Biology
Arts	2000	3000	462	26	Text
Business	2000	3000	438	30	Text
Entertainment	2000	3000	640	21	Text
Health	2000	3000	612	32	Text
Recreation	2000	3000	606	22	Text
Reference	2000	3000	793	33	Text
Science	2000	3000	743	40	Text
Social	2000	3000	1047	39	Text

Table 1: Characteristics of multi-label data sets

Scalability. GRRO achieves multi-label feature selection across all labels while feature selection result for each label is available. By virtue of this property, GRRO is easily scalable for multi-label data understanding. For example, the proposed method enables label correlation exploitation with feature selection results of different labels. Moveover, it is flexible to utilize the inherent properties of multi-label data. In this next section, we consider label-specific features to further enhance the generalization performance.

3.4 Extension to Label-specific Feature Selection

For GRRO, the importance of each feature is determined by summing the weights of the feature to all labels. In the light of that different labels have their own inherent characteristics for distinguishing each other [Zhang and Wu, 2015], a more reasonable manner is that selected features should be label-specific. Following the principle, we extend GRRO with label-specific features, and call the extension as GRRO-LS.

Specially, the global optimization result, i.e., matrix $\mathbf{Z}$ generated by Eq. (6), is utilized as the $priori$ knowledge to exploit label-specific feature selection locally. Considering that feature selection result for each label is available by employing matrix $\mathbf{Z}$, label-specific feature learning can be modeled by a search on discriminative features for each label. Formally, for arbitrary label l_u $(1 \leq u \leq q)$, we define $\mathcal{T}_u$ as the index set to preserve the location of its label-specific features, which satisfies:

$$\max_{\mathcal{T}_u} \quad \sum_{h\in\mathcal{T}_u} z_{hu} \quad \text{s.t.} \quad |\mathcal{T}_u| = k \tag{9}$$

From Eq. (9), we can see that index set $\mathcal{T}_u$ indicates the k features with the larger weight based on $\mathbf{z}_{.u}$, and these features are specified as the label-specific features with respect to label l_u. As the index sets of all labels are found out, we construct a new feature coefficient matrix $\mathbf{Z}^{new} \in \mathcal{R}^{d\times c}$. For arbitrary element $z_{iu}^{new} \in \mathbf{Z}^{new}$, it is defined as follow:

$$z_{iu}^{new} = \begin{cases} z_{iu}, & i \in \mathcal{T}_u \\ 0, & otherwise \end{cases} \tag{10}$$

Note that the generated $\mathbf{Z}^{new}$ only contains the weight information of label-specific features. Then we estimate the importance of each feature by the value of $||\mathbf{z}_{i.}^{new}||_2$ $(1 \leq i \leq d)$, thus achieving label-specific feature selection.

Data set	*Macro-F1*↑						
	GRRO	GRRO-LS	PMU	MDMR	FIMF	SCLS	MICO
Bibtex	0.0495±0.0283 (2)	**0.0545±0.0332** (1)	0.0026±0.0017 (7)	0.0225±0.0050 (5)	0.0062±0.0032 (6)	0.0355±0.0160 (4)	0.0362±0.0188 (3)
Birds	**0.1318±0.0236** (1)	0.1222±0.0246 (2)	0.0890±0.0215 (6)	0.1061±0.0146 (4)	0.0769±0.0175 (7)	0.1012±0.0205 (5)	0.1110±0.0286 (3)
Corel5k	0.2979±0.0015 (3)	**0.3010±0.0024** (1)	0.2968±0.0006 (5)	0.2968±0.0006 (5)	0.2968±0.0008 (5)	0.2981±0.0014 (2)	0.2979±0.0013 (3)
Corel16k001	0.0021±0.0013 (3)	**0.0035±0.0016** (1)	0.0001±0.0001 (7)	0.0019±0.0013 (4)	0.0025±0.0018 (2)	0.0016±0.0013 (6)	0.0017±0.0012 (5)
Corel16k002	0.0049±0.0014 (2)	**0.0050±0.0015** (1)	0.0001±0.0001 (6)	0.0035±0.0004 (5)	0.0001±0.0000 (6)	0.0039±0.0006 (4)	0.0044±0.0014 (3)
Emotions	**0.5635±0.0821** (1)	0.5510±0.0808 (2)	0.4908±0.1069 (7)	0.5451±0.0680 (3)	0.5229±0.1138 (5)	0.5094±0.0617 (6)	0.5363±0.0673 (4)
Genbase	0.5858±0.1435 (2)	0.5803±0.1446 (3)	0.5393±0.1225 (7)	0.5533±0.1231 (6)	**0.5903±0.1571** (1)	0.5663±0.1317 (4)	0.5646±0.1425 (5)
Image	0.4270±0.1033 (5)	0.4281±0.0931 (3)	0.4281±0.0689 (3)	0.4417±0.0657 (2)	0.3170±0.1089 (7)	**0.4486±0.0726** (1)	0.3774±0.0974 (6)
Langlog	**0.1199±0.0027** (1)	0.1178±0.0030 (2)	0.1079±0.0015 (6)	0.1103±0.0022 (5)	0.1070±0.0005 (7)	0.1139±0.0029 (4)	0.1150±0.0051 (3)
Medical	0.3516±0.0596 (2)	**0.3520±0.0573** (1)	0.2472±0.0162 (7)	0.3022±0.0342 (6)	0.3122±0.0559 (5)	0.3228±0.0490 (3)	0.3139±0.0555 (4)
Slashdot	0.2193±0.0272 (2)	**0.2254±0.0318** (1)	0.1718±0.0106 (7)	0.1893±0.0155 (6)	0.2024±0.0203 (5)	0.2126±0.0244 (4)	0.2167±0.0271 (3)
Yeast	0.3205±0.0473 (3)	0.3220±0.0472 (2)	0.3049±0.0410 (5)	0.2987±0.0402 (7)	0.3008±0.0500 (6)	**0.3320±0.0522** (1)	0.3196±0.0468 (4)
Arts	**0.0777±0.0191** (1)	0.0759±0.0212 (3)	0.0499±0.0143 (7)	0.0632±0.0107 (5)	0.0529±0.0182 (6)	0.0773±0.0180 (2)	0.0702±0.0152 (4)
Business	0.1556±0.0162 (2)	0.1533±0.0137 (3)	0.0519±0.0095 (7)	0.1404±0.0071 (5)	0.1329±0.0149 (6)	0.1519±0.0121 (4)	**0.1582±0.0179** (1)
Entertainment	0.1102±0.0363 (2)	**0.1227±0.0399** (1)	0.0641±0.0120 (7)	0.0864±0.0121 (5)	0.0642±0.0197 (6)	0.1183±0.0319 (3)	0.0927±0.0342 (4)
Health	0.2237±0.0345 (2)	**0.2280±0.0330** (1)	0.1291±0.0193 (7)	0.2008±0.0190 (6)	0.2048±0.0247 (5)	0.2175±0.0304 (3)	0.2141±0.0257 (4)
Recreation	0.1297±0.0316 (3)	**0.1371±0.0356** (1)	0.0073±0.0077 (7)	0.1231±0.0238 (5)	0.0585±0.0266 (6)	0.1302±0.0270 (2)	0.1284±0.0305 (4)
Reference	**0.1236±0.0181** (1)	0.1220±0.0177 (2)	0.0399±0.0074 (7)	0.1004±0.0068 (6)	0.1040±0.0138 (5)	0.1200±0.0170 (3)	0.1168±0.0186 (4)
Science	0.0534±0.0116 (3)	**0.0604±0.0143** (1)	0.0163±0.0069 (7)	0.0416±0.0074 (5)	0.0387±0.0099 (6)	0.0486±0.0092 (4)	0.0541±0.0136 (2)
Social	0.1190±0.0180 (3)	**0.1317±0.0220** (1)	0.1138±0.0111 (6)	0.1149±0.0125 (5)	0.1099±0.0149 (7)	0.1213±0.0159 (2)	0.1121±0.0170 (4)
Data set	*Micro-F1*↑						
	GRRO	GRRO-LS	PMU	MDMR	FIMF	SCLS	MICO
Bibtex	0.2165±0.0601 (2)	**0.2205±0.0715** (1)	0.0153±0.0107 (7)	0.1610±0.0124 (5)	0.0794±0.0354 (6)	0.1957±0.0382 (4)	0.1988±0.0475 (3)
Birds	**0.5045±0.0344** (1)	0.4890±0.0322 (3)	0.3661±0.0390 (6)	0.4080±0.0383 (5)	0.3607±0.0403 (7)	0.4314±0.0353 (4)	0.4999±0.0316 (2)
Corel5k	0.0064±0.0041 (2)	**0.0170±0.0095** (1)	0.0003±0.0006 (7)	0.0004±0.0009 (6)	0.0010±0.0012 (5)	0.0050±0.0044 (4)	0.0055±0.0042 (3)
Corel16k001	0.0052±0.0038 (2)	**0.0072±0.0045** (1)	0.0001±0.0002 (7)	0.0015±0.0010 (6)	0.0018±0.0012 (5)	0.0033±0.0031 (4)	0.0046±0.0036 (3)
Corel16k002	0.0089±0.0036 (2)	**0.0095±0.0035** (1)	0.0002±0.0002 (6)	0.0038±0.0005 (5)	0.0001±0.0001 (7)	0.0078±0.0027 (4)	0.0086±0.0039 (3)
Emotions	**0.6014±0.0616** (1)	0.5955±0.0561 (2)	0.5178±0.0948 (7)	0.5782±0.0506 (4)	0.5615±0.0969 (5)	0.5527±0.0477 (6)	0.5935±0.0510 (3)
Genbase	0.9065±0.1285 (2)	**0.9066±0.1290** (1)	0.8778±0.1282 (7)	0.9003±0.1258 (4)	0.8948±0.1378 (6)	0.9053±0.1280 (3)	0.8974±0.1285 (5)
Image	0.4394±0.0976 (4)	0.4409±0.0920 (3)	0.4355±0.0660 (5)	0.4495±0.0615 (2)	0.3320±0.1037 (7)	**0.4564±0.0691** (1)	0.3934±0.0904 (6)
Langlog	**0.0973±0.0169** (1)	0.0782±0.0175 (2)	0.0058±0.0080 (6)	0.0306±0.0231 (5)	0.0011±0.0016 (7)	0.0536±0.0236 (4)	0.0638±0.0435 (3)
Medical	0.7022±0.0707 (2)	**0.7117±0.0771** (1)	0.5447±0.0362 (7)	0.6576±0.0710 (6)	0.6587±0.0870 (5)	0.6599±0.0628 (4)	0.6624±0.0674 (3)
Slashdot	0.2276±0.0423 (2)	**0.2330±0.0492** (1)	0.1338±0.0182 (7)	0.1739±0.0243 (6)	0.1990±0.0321 (5)	0.2220±0.0407 (4)	0.2240±0.0419 (3)
Yeast	0.6059±0.0340 (3)	0.6054±0.0349 (4)	0.5949±0.0280 (6)	0.5884±0.0281 (7)	0.5976±0.0341 (5)	**0.6085±0.0340** (1)	0.6065±0.0344 (2)
Arts	0.1967±0.0381 (2)	**0.1996±0.0468** (1)	0.1227±0.0355 (7)	0.1693±0.0274 (5)	0.1276±0.0482 (6)	0.1891±0.0368 (3)	0.1797±0.0339 (4)
Business	0.6905±0.0087 (3)	0.6907±0.0086 (2)	0.6757±0.0039 (7)	0.6825±0.0046 (5)	0.6789±0.0057 (6)	0.6894±0.0065 (4)	**0.6913±0.0098** (1)
Entertainment	0.2866±0.0858 (2)	**0.2931±0.0773** (1)	0.1085±0.0216 (7)	0.2158±0.0316 (5)	0.1217±0.0438 (6)	0.2806±0.0544 (3)	0.2387±0.0838 (4)
Health	**0.5167±0.0242** (1)	0.5123±0.0245 (2)	0.4376±0.0290 (7)	0.4585±0.0290 (6)	0.4682±0.0270 (5)	0.4777±0.0254 (4)	0.4996±0.0216 (3)
Recreation	0.2671±0.0532 (2)	**0.2713±0.0552** (1)	0.0105±0.0117 (7)	0.2302±0.0329 (5)	0.1174±0.0492 (6)	0.2623±0.0449 (4)	0.2636±0.0535 (3)
Reference	0.3796±0.0353 (2)	**0.3830±0.0421** (1)	0.3265±0.0266 (7)	0.3267±0.0254 (6)	0.3289±0.0366 (5)	0.3693±0.0390 (3)	0.3665±0.0390 (4)
Science	0.1659±0.0346 (3)	**0.1830±0.0423** (1)	0.0407±0.0194 (7)	0.1187±0.0223 (5)	0.1148±0.0240 (6)	0.1502±0.0312 (4)	0.1700±0.0379 (2)
Social	0.5040±0.0550 (2)	**0.5152±0.0627** (1)	0.4379±0.0527 (7)	0.4468±0.0595 (6)	0.4740±0.0469 (5)	0.4813±0.0491 (4)	0.4936±0.0531 (3)

Table 2: Comparison results of multi-label feature selection methods (mean±std. deviation) in terms of *macro-F1* and *micro-F1*

4 Experiments

4.1 Experimental Setup

Data sets. A total of twenty benchmark multi-label data sets are employed in the experiment[2]. Table 1 summarizes detailed characteristics of these data sets, which are mainly from the domains including text, multimedia, and biology. We use the same train/test splits of these data sets to report and compare the results.

Evaluation metrics. Six widely used multi-label evaluation metrics are employed for performance evaluation, including two label-based metrics *macro-F1* and *micro-F1*, and four example-based metrics *Hamming loss*, *ranking loss*, *coverage*, and *average precision*. Concrete metric definitions can be found in [Wu and Zhou, 2017]. These metrics can evaluate the performance of multi-label algorithms from various aspects. For *macro-F1*, *micro-F1* and *average precision*, the larger the values the better the performance. For the other metrics, the smaller the values the better the performance.

Comparing algorithms. Five multi-label feature selection methods are selected to compare. All of them are information theoretical based methods, including four heuristic methods PMU [Lee and Kim, 2013], MDMR [Lin *et al.*, 2015], FIMF [Lee and Kim, 2015], and SCLS [Lee and Kim, 2017], and one optimization method MICO [Sun *et al.*, 2019].

[2]Public available at http://www.uco.es/kdis/mllresources/

Hyper-parameters. For the proposed method, both of α and β are searched in $\{10^{-3}, 10^{-2}, ..., 10^{3}\}$, and k is searched in $\{5, 10, ..., 50\}$. The parameter of each comparing method (if any) is set as the corresponding reference suggested. ML-KNN [Zhang and Zhou, 2007] (with the default setting) is used as the classifier for performance evaluation. For the parameter-tuning, we adopt a grid-search strategy to seek the optimal parameter, which is determined by making the average classification result (ACR) on test data smallest. Here, we define the formula for the ACR as follow:

$$ACR(para) = \sum_{i=1}^{30}(HL_i(\mathtt{f},\mathcal{U}) + RL_i(\mathtt{f},\mathcal{U})) \qquad (11)$$

where $para$ denotes the collection of algorithm parameter(s), $\mathcal{U}$ is the test set, and $\mathtt{f}$ denotes the classifier, i.e., ML-KNN. $HL_i(\mathtt{f},\mathcal{U})$ and $RL_i(\mathtt{f},\mathcal{U})$ output the result of *Hamming loss* and *ranking loss* respectively while selecting top-i features.

Computational device. Experiments are performed on a PC with an Intel i7-7700K 4.20GHz CPU and 32GB RAM.

4.2 Performance Evaluation

To evaluate the performance, we focus on top-50 features selected by each method, and the average result with 50 groups of feature subsets[3] is recorded to make a comparison. Due to

[3]The first group is composed of the top-1 feature, the second one is composed of the top-2 features, and so on.

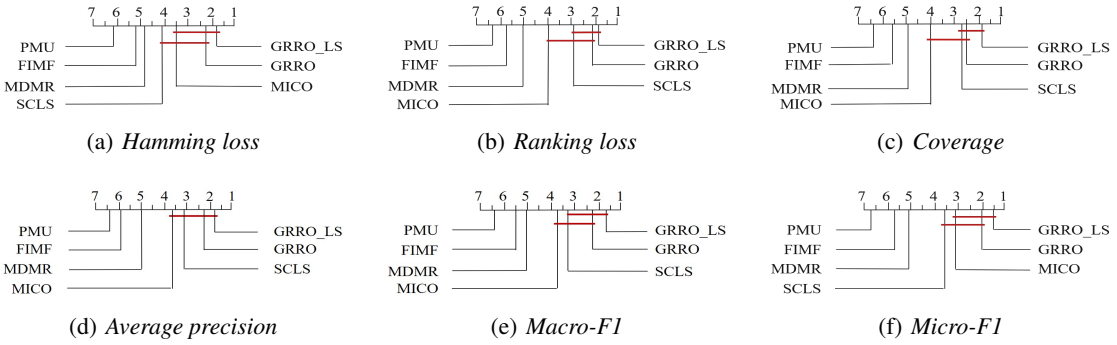


Figure 1: Comparison of the control method against comparing methods with the Nemenyi test ($CD = 2.0146$ at 0.05 significance level)

space limit, we only present the average result of each method on *macro-F1* and *micro-F1* in this paper (as shown in Table 2), and the details on the other metrics are available on the web[4]. From Table 2, we have a couple of observations. (1) Compared with the selected comparing methods, GRRO and GRRO-LS achieve better performance on 11 and 14 out of 20 data sets with respect to *macro-F1* respectively, while on *micro-F1*, both of them can win on 16 out of 20 data sets. (2) On all the 20 data sets, GRRO-LS is superior to GRRO on 13 data sets regarding *macro-F1*, 15 data sets regarding *micro-F1*. This suggests that exploiting label specific features is conducive to the performance improvement. (3) These comparing methods achieve the best performance on up to 2 out of 20 data sets. Thus, we conclude that the proposed method is effective for multi-label feature selection, and has the advantages compared with some other well-established methods.

To further analyze the performance among all the methods, Friedman test [Demsar, 2006] is used as the favorable statistical significance test for the method comparison on the 20 data sets. Table 3 illustrates the Friedman statistic F_F and the corresponding critical value on each metric, and we can see that the null hypothesis, which follows the principle that all the methods have equal performance, is clearly rejected in terms of each metric at significance level $\alpha = 0.05$. Thus, the post-hoc Nemenyi test [Demsar, 2006] is utilized to complete the performance analysis. Here, GRRO or GRRO-LS is regarded as the control method respectively whose average rank difference against the comparing method is calibrated with the critical difference (CD). Accordingly, GRRO or GRRO-LS is deemed to have significantly different performance to one comparing method if their average ranks differ by at least one CD ($CD = 2.0146$ in this paper).

Fig. 1 shows the CD diagrams [Demsar, 2006] *w.r.t.* each metric. Specially, any comparing method whose average rank is within one CD to that of GRRO or GRRO-LS is connected. Otherwise, the method, which is not connected with GRRO or GRRO-LS, is considered to have the significant different performance with the control method. From Fig. 1, we can see that GRRO-LS and GRRO rank 1st and 2nd respectively among all the methods, which have no significant difference on all the metrics, and significantly perform better than PMU, FIMF, and MDMR. Compared with MICO, GRRO-LS has significantly better performance on *ranking loss*, *coverage*, and *macro-F1*, which also significantly outperforms SCLS on *Hamming loss* and *micro-F1*. Thus, the proposed method (namely GRRO-LS) can achieve highly competitive performance against the selected comparing methods.

[4]https://jiazhang-ml.pub/Supplement-GRRO.pdf

Evaluation metric	F_F	Critical value($\alpha = 0.05$)
Hamming loss	21.1661	≈2.17
Ranking loss	36.8824	
Coverage	32.2524	
Average precision	38.4204	
Macro-F1	33.3880	
Micro-F1	77.3768	

Table 3: Friedman statistics F_F and the critical value on evaluation metrics (# comparing algorithms $c = 7$, # data sets $N = 20$)

4.3 Influence of Selected Features

In this section, an experiment is conducted on the Medical date set to learn the influence of selected features.

The experimental result is shown in Fig. 2, in which the performance on each metric is figured out by varying the number of selected features. From Fig. 2, we observe: With the increasing of the number of selected features, the performance of all the methods first has a significant improvement, and then keeps stable or even degrades. It can be seen that feature selection benefits to the performance.

4.4 Running Time Comparison

Table 4 shows the timing results (in second) on multi-label feature selection. According to Table 4, we can observe that the proposed method performs the best in terms of average ranking (*Ave. Rank.*). To be specific, GRRO is obviously superior to PMU and MDMR. FIMF avoids ineffective and repetitive calculations to improve the computation efficiency, hence GRRO is slower than FIMF on some data sets. However, FIMF is less efficient than GRRO on large-scale data analysis, such as Bibtex, Corel16k001, and Corel16k002. SCLS designs feature evaluation criteria to handle large label sets. We can see from Table 4 that GRRO can achieve superior or at least comparable performance on such data sets against SCLS, such as Bibtex and Corel5k, and GRRO also achieves statistically superior to SCLS on most of the other data sets. Similar with GRRO, MICO adopts an optimization strategy to conduct multi-label feature selection, but it has a

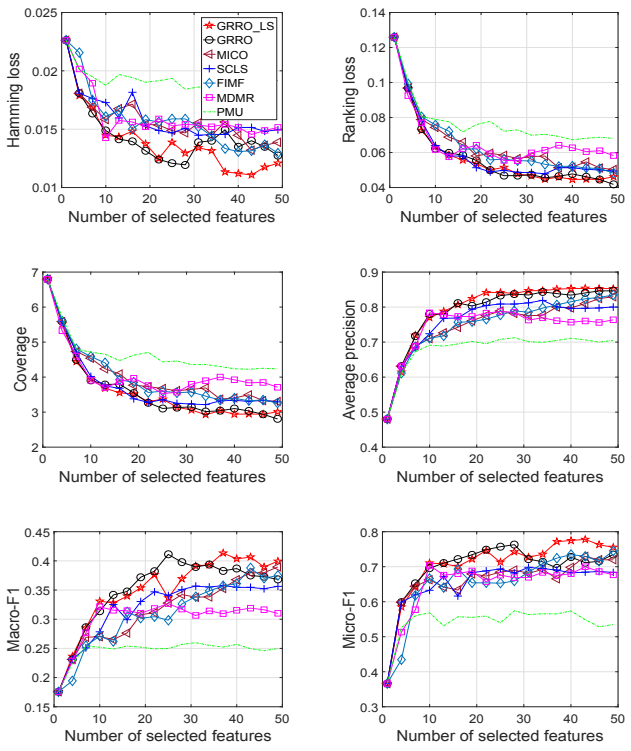
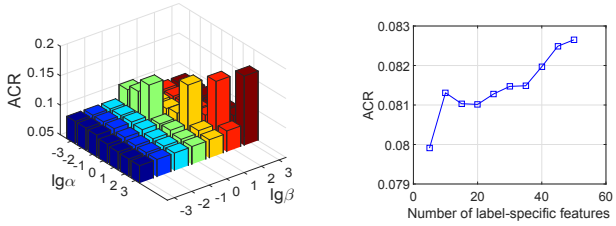


Figure 2: Influence of selected feature number on Medical data set



(a) The influence of α and β (b) The influence of k

Figure 3: Parameter analysis on the Medical data set

low convergence rate with iterative optimization. Thus, GRRO is more computationally efficient than MICO. Remark: GRRO-LS and GRRO have the similar result on running time.

4.5 Parameter Analysis

Our method includes three parameters: α and β are to reflect the influence of feature redundancy and label correlation respectively, and k is the number of label-specific features. We analyze these parameters on the Medical data set, and show the experimental result in Fig. 3. Fig. 3(a) shows the average classification result (ACR), which is calculated by Eq. (11), when different pairs of α and β are employed. From Fig. 3(a), we observe that ACR is going to change dramatically in some cases. Thus, the proposed method is sensitive to α and β. Fig. 3(b) shows the influence of parameter k, and it reflects that the performance is going to deteriorate while k becomes large, and the optimal result is obtained while $k = 5$. A similar phenomenon occurs on most of other data sets, as listed in Table 5. This suggests that generally a small value of k (e.g., $k = 5$) helps to the label-specific feature selection.

Data set	PMU	MDMR	FIMF	SCLS	MICO	GRRO
Bibtex	–	–	–	**194.30**	–	200.61
Birds	40.02	37.65	**0.66**	1.51	1.39	0.89
Corel5k	–	–	–	95.22	42.18	**22.95**
Corel16k001	–	–	743.89	87.20	50.02	**32.46**
Corel16k002	–	–	937.84	92.20	50.04	**33.23**
Emotions	2.83	2.73	0.06	0.32	0.64	**0.05**
Genbase	295.57	263.61	5.95	**3.04**	11.01	10.83
Image	23.26	22.39	**0.09**	2.47	2.80	1.16
Langlog	–	–	49.99	16.65	100.89	**14.15**
Medical	876.15	713.29	21.50	**13.12**	193.50	20.90
Slashdot	808.58	717.25	**8.47**	22.48	56.61	36.07
Yeast	30.21	27.21	0.25	1.22	0.95	**0.23**
Arts	345.59	304.52	**4.05**	8.22	11.39	5.42
Business	387.55	334.92	5.19	8.12	9.72	**4.98**
Entertainment	383.35	344.17	**3.78**	10.81	28.30	10.01
Health	589.30	502.54	**8.23**	12.01	24.68	9.50
Recreation	382.83	339.72	**3.92**	10.23	23.67	9.12
Reference	796.20	671.31	**11.51**	16.01	56.24	15.98
Science	944.48	765.26	15.54	16.03	46.30	**14.28**
Social	–	–	**20.82**	22.70	148.17	27.57
Ave. Rank.	6.00	5.00	2.05	2.65	3.55	**1.75**

Table 4: Running time (sec) of the multi-label feature selection methods. – denotes that time cost is over 1000 seconds

Data set	Optimal k value
Bibtex, Birds, Corel5k, Corel16k001, Emotions, Entertainments, Health, Medical, Recreation, Science, Slashdot, Social	5
Business,Langlog, Reference	10
Education, Image	15
Corel16k002, Genbase, Yeast	$\geq$20

Table 5: Optimal value distribution of k on the 20 data sets

5 Conclusion

In this paper, we developed information theoretical based methods for multi-label feature selection. Our main contribution is to propose a general global optimization framework incorporating feature relevance, feature redundancy, and label correlation. In addition to label correlation exploitation, the proposed method is capable for exploiting the other properties of multi-label learning to further improve the performance, such as label-specific features. Experiments on twenty benchmark data sets in terms of six evaluation metrics showed that the proposed method can significantly improve the performance with feature selection.

In future work, we have interest in further study of labeling information exploitation considering the issues of the class-imbalance and the relative labeling-importance, and will also pay attention to the analysis of genetic data with high dimensionality, such as the application on autism spectrum disorder.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61876159, No. 61806172, No. 61572409, No. U1705286, No. 61571188, No. 61772211 & No. U1811263), the National Key Research and Development Program of China (No.2018YFC0831402), Fujian Province 2011 Collaborative Innovation Center of TCM Health Management, Collaborative Innovation Center of Chinese Oolong Tea Industry-Collaborative Innovation Center (2011) of Fujian Province.

References

[Bermejo *et al.*, 2018] Pablo Bermejo, José A. Gámez, and José Miguel Puerta. Adapting the CMIM algorithm for multilabel feature selection. A comparison with existing methods. *Expert Systems*, 35(1), 2018.

[Boutell *et al.*, 2004] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[Brown *et al.*, 2012] Gavin Brown, Adam Craig Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.

[Demsar, 2006] Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[Elisseeff and Weston, 2001] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14, Vancouver, Canada*, pages 681–687, 2001.

[Jian *et al.*, 2016] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, NY*, pages 1627–1633, 2016.

[Krawczyk, 2016] Bartosz Krawczyk. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[Lee and Kim, 2013] Jae-Sung Lee and Dae-Won Kim. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters*, 34(3):349–357, 2013.

[Lee and Kim, 2015] Jae-Sung Lee and Dae-Won Kim. Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognition*, 48(9):2761–2771, 2015.

[Lee and Kim, 2017] Jae-Sung Lee and Dae-Won Kim. SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, 66:342–352, 2017.

[Li *et al.*, 2015] Yu-Kun Li, Min-Ling Zhang, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic City, NJ*, pages 251–260, 2015.

[Li *et al.*, 2018] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys*, 50(6):94:1–94:45, 2018.

[Lim and Kim, 2016] Hyunki Lim and Dae-Won Kim. Convex optimization approach for multi-label feature selection based on mutual information. In *Proceedings of the 23rd International Conference on Pattern Recognition, Cancún, Mexico*, pages 1512–1517, 2016.

[Lin *et al.*, 2015] Yaojin Lin, Qinghua Hu, Jinghua Liu, and Jie Duan. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, 168:92–103, 2015.

[Lin *et al.*, 2016] Yaojin Lin, Qinghua Hu, Jia Zhang, and Xindong Wu. Multi-label feature selection with streaming labels. *Information Sciences*, 372:256–275, 2016.

[Peng *et al.*, 2005] Hanchuan Peng, Fuhui Long, and Chris H. Q. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[Schapire and Singer, 2000] Robert E. Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[Sechidis *et al.*, 2014] Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown. Information theoretic feature selection in multi-label data through composite likelihood. In *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, Joensuu, Finland*, pages 143–152, 2014.

[Sun *et al.*, 2019] Zhenqiang Sun, Jia Zhang, Liang Dai, Candong Li, Changen Zhou, Jiliang Xin, and Shaozi Li. Mutual information based multi-label feature selection via constrained convex optimization. *Neurocomputing*, 329:447–456, 2019.

[Vergara and Estévez, 2014] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014.

[Wei and Li, 2019] Tong Wei and Yu-Feng Li. Learning compact model for large-scale multi-label data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, Hawaii*, pages 5385–5392, 2019.

[Wu and Zhou, 2017] Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia*, pages 3780–3788, 2017.

[Wu *et al.*, 2014] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels. In *Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, Sweden*, pages 1964–1968, 2014.

[Zhang and Wu, 2015] Min-Ling Zhang and Lei Wu. LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2015.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang *et al.*, 2019] Jia Zhang, Zhiming Luo, Candong Li, Changen Zhou, and Shaozi Li. Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, 95:136–150, 2019.