Multi-Label Feature Selection Based on Min-relevance Label

GAO, Wanfu; PAN, Hanlin

Published in:
IEEE Access

Published: 01/01/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1109/ACCESS.2022.3231871

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher’s copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 23/03/2024
Multi-Label Feature Selection Based on Min-Relevance Label

WANFU GAO1 AND HANLIN PAN1,2
1College of Computer Science and Technology, Jilin University, Changchun 130012, China
2College of Business, City University of Hong Kong, Hong Kong 999077, China

Corresponding author: Hanlin Pan (panhanlin16@126.com)

This work was supported in part by the National Key Research and Development Plan of China under Grant 2017YFA0604500, in part by the National Sci-Tech Support Plan of China under Grant 2014BAH02F00, in part by the National Natural Science Foundation of China under Grant 61701790, in part by the Youth Science Foundation of Jilin Province of China under Grant 201605201JH and Grant 20180201103GX, and in part by the China Postdoctoral Science Foundation under Grant 2018M631873.

ABSTRACT Multi-label feature selection has been widely adopted to address multi-label data with high-dimension features. It is critical to calculate label correlations for multi-label feature selection. Existing methods adopt different schemes to calculate label correlations, which obtain different importance of labels. However, there exist two issues regarding these schemes calculating label importance: first, previous schemes cannot predict the whole labels well because they only focus on the most important labels; second, most of important labels have similar classification information corresponding to redundant features. To this end, we use the mutual information metric to obtain different cores regarding label set rather than calculating the importance of labels. Afterwards, we capture features with respect to each core label, finally, obtaining an optimal feature subset. To verify the effectiveness, our method is compared to state-of-the-art multi-label methods on 16 real-world data sets with several evaluating metrics. The results of experiment proves that the proposed method has achieves the best classification performance among all multi-label feature selection methods.

INDEX TERMS Multi-label learning, feature selection, information theory, label importance.

I. INTRODUCTION

Multi-label learning has extensive applications in many research fields [1], [2], [3] such as document classification [4], image recognition [5], [6] and gene prediction [7], [8].Therefore it attracts significant attentions. In multi-label data sets, each instance is usually associated with multiple labels and high-dimension features simultaneously. High-dimension multi-label data have numerous irrelevant and redundant features, which not only degrade classification performance but also increases calculating cost. To address the problem regarding curse of dimension, feature selection technique becomes the focus [9], [10], [11], [12].

Feature selection methods can be divided into three categories [13], [14], [15], [16]: filter models, wrapper models and embedded models. Filter models obtain a feature subset from the original feature set independent of the subsequent classifier [17]. Wrapper models use the result of the classifier to determine the optimal feature subset [18]. Embedded models choose the optimal feature subset in the training process [19]. Wrapper models and embedded models both involve the process of training models. Compared to them, filter models are more flexible and faster [20], thus this paper will focus on this type of models.

Different from the single-label learning that one instance is only related to one label, in multi-label learning, one instance is related to multiple labels. Therefore, it is critical to consider the label correlations. However, there two issues exist in previous methods. First, previous methods cannot...
predict the whole labels well because they only focus on the most important labels that are defined by the cumulative summation approximation scheme; second, most of important labels have similar classification information corresponding to redundant features. For example, label “sports” has high probability alone with label “baseball”, features are selected around similar classification information, causing that numerous redundant features are selected. In fact, there exist high-order label correlations in multi-label data. Previous methods employ mutual information based on cumulative summation approximation scheme to calculate the relevance between features and labels [21], [22], [23], [24], [25], [26]. The common limitation of these methods is that cumulative summation approximation scheme may overestimate the importance of some candidate features related to groups with more labels, but ignore the importance of groups with fewer labels. In other words, the features selected by these methods may only have a good classification effect for those groups that contain more labels. We believe that a good model should have considerable performance on classification for each label set. Specially, the highlights are concluded in the following:

1) A new label importance evaluation scheme is designed, which consider predict precisely for the whole label set.
2) The redundant features are eliminated by selecting features around different classification information of labels
3) The new feature selection method avoids to overestimate the significance of some features.

The rest of this paper is arranged as follows. Section II introduces some basic concepts of information theory and evaluation criteria of multi-label classification. Section III reviews some related work. In Section IV, we present the proposed multi-label feature selection method. Section V introduces some experimental results to verify the effectiveness of the proposed method. In Section VI, we make a conclusion and give our opinion in future research direction.

II. PRELIMINARIES
A. THE BASIC CONCEPTS OF INFORMATION THEORY

In this subsection, we will introduce some basic concepts of information theory which are used to measure the correlations among random variables [27], [28]. Let \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_m\} \) be two discrete random variables. Information entropy \( H(X) \) is used to measure the uncertainty of \( X \). It is defined as follows:

\[
H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)
\]

\( p(x_i) \) is the probability of \( x_i \) and the base of log is 2. Then we introduce joint entropy and conditional entropy. The joint entropy of \( X \) and \( Y \) can be expressed as \( H(X, Y) \). \( H(X|Y) \) is the conditional entropy of \( X \) considering \( Y \), which is used to measure the remaining uncertainty of \( X \) under the condition of \( Y \). \( H(X, Y) \) and \( H(X|Y) \) are defined as follows:

\[
H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j)
\]

\[
H(X|Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i | y_j) \log p(x_i | y_j)
\]

(2)

(3)

\( p(x_i, y_j) \) is the joint probability of \((x_i, y_j)\) and \( p(x_i | y_j) \) is the conditional probability of \( x_i \) given \( y_j \). With the concept of entropy, mutual information is used to measure the amount of information shared by two variables. The larger the mutual information is, the more relevant the variables are. It is defined as:

\[
I(X; Y) = H(X) - H(X|Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}
\]

(4)

Then we introduce another variable \( Z \). Conditional mutual information means the mutual information between two variables giving another variable, the mutual information of \( X \) and \( Y \) giving \( Z \) can be expressed as:

\[
I(X; Y|Z) = H(X|Z) - H(X|Y, Z)
\]

(5)

Considering \( X, Y \) as a whole. The joint mutual information is defined as:

\[
I(X, Y; Z) = H(Z) - H(Z|X, Y) = I(X; Z|Y) + I(Y; Z).
\]

(6)

Interaction information measures the amount of information shared by three variables, which is defined as:

\[
I(X; Y; Z) = I(X; Z) + I(Y; Z) - I(X, Y; Z).
\]

(7)

B. MULTI-LABEL EVALUATION METRICS

In this section, three evaluation metrics of classification performance used are introduced. They are Hamming Loss, Macro-F1 and Micro-F1 [29], [30], [31].

Let \( D = \{(x_1, L_1), (x_2, L_2), \ldots, (x_n, L_n)\} \) be the multi-label test set where the instance pair is composed of a instance and the real labels, \( L = \{l_1, l_2, \ldots, l_q\} \) is the label set, and \( N \) is the number of instances and \( L_i \subseteq L \) is the label set corresponding to the instance \( x_i \). Let \( L_i' \) be the predicted label set corresponding to the instance \( x_i \). Hamming Loss (HL) is the average fraction of misclassified labels in the test data. HL is defined as:

\[
HL = \frac{1}{n} \sum_{i=1}^{n} \frac{|L_i' \oplus L_i|}{q}.
\]

(8)

In this formula \( \oplus \) is the symmetric difference between the label set \( L_i \) and \( L_i' \). The smaller the value of HL, the better the classification performance.

Macro-F1 and Micro-average Micro-F1 are two widely used evaluation criteria for multi-label learning based on the
F1 score. Macro-F1 is the arithmetic average of the F1 score of all labels. Macro-F1 is defined as follows:

$$\text{Macro - F1} = \frac{1}{q} \sum_{i=1}^{q} \frac{2TP_i}{2TP_i + FP_i + FN_i},$$

(9)

$TP_i$, $FP_i$ and $FN_i$ denote the number of true positives, false positives and false negatives in the $i$th respectively. Micro-F1 is the weighted average of the F1 over all labels:

$$\text{Micro - F1} = \frac{\sum_{i=1}^{q} 2TP_i}{\sum_{i=1}^{q} (2TP_i + FP_i + FN_i)}.$$

(10)

These three multi-label evaluation criteria can be used to measure the classification performance. For the three evaluation criteria, a lower value of HL indicates a better classification performance while the higher Macro-F1 and Micro-F1 are, the better the classification performance is.

### III. RELATED WORK

Traditional multi-label feature selection methods can be divided into two groups: problem transformation and algorithm adaptation [32], [33]. The problem transformation methods involve two steps: (1) transforming the multi-label data sets to multiple groups of single-label data sets; (2) selecting the optimal features from the transformed groups of single-label data sets. Binary Relevance (BR) [6], Label Power set (LP) [34] and Pruned Problem Transformation (PPT) [35] are classic problem transformation methods. BR transforms the multi-label data set into several groups of independent binary classification data sets. LP divides each instance’s label set to a single new class. Using ReliefF (RF) [37] and Information Gain (IG) [38] as the feature evaluation criteria to measure the transformed data, N. Spolaö et al. [36] propose four multi-label feature selection methods (RF-BR, RF-LP, IG-BR and IG-LP) based on BR and LP. However, BR ignores the correlations between labels and LP may cause over-fitting and imbalance problems because it creates too many classes. PPT removes the instances related to rarely occurring labels by a predefined minimal number of occurrences of the label set to improve the effectiveness of LP. Based on this PPT method and mutual information, Doquire and Verleysen [39] propose a multi-label feature selection method (PPT + MI). Additionally, CHI square statistic are used to select the effective features (PPT + CHI) [35]. However, the problem of this type of methods still remain, they usually ignore the correlation between labels.

In recent years, many multi-label feature selection methods based on algorithm adaptation have been proposed, which directly select features from multi-label data sets. S Kashef and H Nezamabadi-pour propose a multi-label feature selection algorithm based on Pareto dominance concept in [40] for label-specific feature selection in multi-objective optimization problem. Sun et al. [25] propose a new feature selection method based on mutual information, which obtains discriminant features considering label correlation through constrained convex optimization (MICO).

Multi-label information feature selection (MIFS) [41] is an integrated feature selection method. It uses latent semantic index (LSI) to decompose multi-label information into low dimensional label space, and then uses the reduced label space to guide the feature selection process through regression model. Gao et al. [53] propose a new feature selection method based on a unified feature selection framework including three low-order information-theoretic terms for multi-label learning named Selected Terms of Feature Selection (STFS) regarding high-order variable correlations. Zhang et al. [54] propose a new feature selection method named Weighted Feature Relevancy Feature Selection (WFRFS) assuming that the remaining uncertainty of candidate features changed dynamically according to the selected features, thus Relevancy Ratio is designed to clarify the dynamic change amount of information and Weighted Feature Relevancy is defined to evaluate the candidate features. Fan et al. [55] propose manifold learning with structured subspace for multi-label feature selection to take the potential structural information of the data into consideration. Wang et al. [56] propose a new feature selection based on neighborhood discrimination index which is used to characterize the distinguishing information of a neighborhood relation. Fan et al. [57] propose a new feature selection method with local discriminant model and label correlations. It uses local discriminant model to obtain cluster assignment matrix and then explores the label correlations matrix based on clusters by taking advantage of a coefficient matrix between feature space and label space. Lee and Kim [21] propose a multi-label feature selection method based on information theory, which is named Pairwise Multi-label Utility (PMU). Its evaluation function is defined as follows:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{j \in S} \sum_{l_i \in L} I(f_k; f_j; l_i)$$

(11)

where $L$ is the label set and $S$ is the feature subset that has been selected. $f_k$ is a candidate feature and $f_j$ is a feature in $S$, $l_i$ and $l_j$ are two labels. The larger the value of $J(f_k)$ is, the more important the candidate feature $f_k$ is.

Multi-label feature selection method using interaction information (D2F) [22] is also proposed to efficiently measure the correlation between features in multi-label data. It can be presented as:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{j \in S} \sum_{l_i \in L} I(f_k; f_j; l_i).$$

(12)

Scalable Criterion for a Large Label Set (SCLS) [23] is proposed to design a new multi-label feature selection method using a scale from 0% to 100% to assess the correlation. It is denoted as follows:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{j \in S} \frac{I(f_k; f_j)}{H(f_k)} \sum_{l_i \in L} I(f_k; l_i).$$

(13)
Lin et al. [42] propose a multi-label feature selection method based on Max-Dependency and Min-Redundancy (MDMR). The general idea of MDMR is to maximize the feature dependency between candidate features and each label using mutual information and minimize the feature redundancy between the candidate feature and all the already-selected features. The criterion of MDMR is defined as

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \frac{1}{|S|} \sum_{f_j \in S} \left\{ I(f_k; f_j) - \sum_{l_i \in L} I(f_k; l_i|f_j) \right\}. \quad (14)$$

|S| is the number of features that have already been selected. Multi-label Feature Selection based on Label Redundancy (LRFS) [2] is also proposed, and LRFS adopts the conditional mutual information between candidate features and each label giving other labels to assess feature relevance. It is defined as follows:

$$J(f_k) = \sum_{l_i \in L} \left( \frac{1}{|S|} \sum_{f_j \in S} I(f_k; l_i|f_j) \right) - \frac{1}{|S|} \sum_{f_j \in S} I(f_k; f_j). \quad (15)$$

Gonzalez-Lopez et al. [43] propose Geometric Mean Maximization (GMM) which selects the optimal features by calculating the geometric mean of the instance’s mutual information. GMM is defined as follows:

$$J(f_k) = \left[ \prod_{i=1}^{q} I(f_k; l_i) \right]^{\frac{1}{2}}. \quad (16)$$

Through the above introduction, we can find that most of previous multi-label feature selection methods based on information theory use cumulative summation approximation to consider label correlation. But the problem is that in real-world multi-label data sets, there is often high-order label correlation in multi-label data, that is, labels tend to be clustered into several groups with high correlation. The common limitation of these methods is that cumulative summation may overestimate the importance of some candidate features related to groups with more labels, but ignore the importance of groups with fewer labels. In other words, the features selected by these methods may only have a good classification effect for those groups that contain more labels. We believe that a good model should have considerable performance on classification for each label in the data set, therefore we propose a method named multi-label Feature Selection method based on Min-relevance Label (MRLFS) that first finds the label set that has minimized relevance and then select features based on the label set.

IV. MRLFS: MIN-RELEVANCE LABEL FEATURE SELECTION

A. METHOD PROPOSED

Observing Formulas (11)-(15) that obtain evaluation functions for multi-label feature selection methods using cumulative summation approximation, it is hard to deal with high-order correlation relation. Labels in real-world multi-label data set are likely to cluster to several groups that share similar topics, thus the number of labels in the group will influence the classification performance of feature selection models. On the one hand, they tend to select more redundant features that are all related to the same group with more labels; on the other hand they neglect some important features that are related to the groups with less labels. For example, if there is a label set with 100 labels that can be divided into 5 groups (g1, . . . , g5) while there are 80 labels belong to g1 while only 20 labels belong to the other four groups, then these models may select extra features that are related to g1 and ignore some features related to the other four groups. Supposing five group are important equally, then these methods may not be satisfying. An effective and comprehensive feature subset should choose features that are related to different semantic groups, which is proved to be effective [44].

To solve these issues, we propose the Min-Relevance Label Feature Selection method (MRLFS). We have mainly taken two stages:

1) By selecting the most irrelevant labels through mutual information to form a new label set, we can avoid the influence of the number of labels in the original label set.

2) The maximum value of mutual information between features and labels is used as the calculation function, which is also to avoid the deviation of experimental results caused by the large difference in the number of different groups of labels.

Let the training sample D with a full feature set $F = \{f_1, f_2, \ldots, f_q\}$ and the label set $L = \{l_1, l_2, \ldots, l_q\}$. First we calculate the mutual information of all pairs of $f_i$ and $l_j$ and get the pair of feature with the largest value, we call this pair as initial features $f_0$ and initial labels $l_0$. And we take this $f_0$ to our feature subset $S$ and $l_0$ to the label subset $L'$. Then we calculate the mutual information of label not in $L'$ with all the labels in $L'$, then we find the label with minimal value and select it into the $L'$, we call this evaluate function as label relevance (LR) and it is denoted as:

$$LR(l_i) = \sum_{l_j \in L'} I(l_i; l_j) \quad (17)$$

Afterwards, the minimal label relevance (MLR) is denoted as:

$$MLR(l_i) = \arg \min_{l_j \in L'} \left( \sum_{l_i \in L'} I(l_i; l_j) \right) \quad (18)$$

this function returns the label with the minimal relevance with the labels in the $L'$. We choose one label at one time and then add the feature having maximal mutual information with this label into subset $F'$, the function is presented as follows:

$$J(f_k) = I(f_k; l_{\text{now}}) - \frac{1}{|S|} \sum_{f_j \in S} I(f_k; f_j) \quad (19)$$

the latter item is used to minimize the feature redundancy in the already-selected feature subset. S is the number of labels in $F'$. Next we repeat this operation until the number
of selected features has arrived the termination condition or all labels have been added into $L'$. If all the labels have been added and the number of feature we selected is still not enough, we use this function to select feature until the termination condition:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in F'} I(f_k; f_j)$$  (20)

the pseudo code of MLRFS is presented in Algorithm 1. There are four stages in the whole algorithm. Lines 1-9 are the first stage where the model selects the first feature and the first label. This stage can be considered as the initialization of model. The second stage is lines 10-11, 14-17 where the model first judges if all the labels are traversed, if not it will select the minimal relevance label. Lines 12-13, 18-20 is stage three where the model selects the feature based on the label subset.

### B. COMPLEXITY ANALYSIS

We provide the complexity analysis for the MRLFS method and other information-theoretical-based feature selection

---

**Algorithm 1 MRLFS**

**Input:**
A training sample $D$ with a full feature set $F = \{f_1, f_2, \ldots, f_d\}$ and the label set $L = \{l_1, l_2, \ldots, l_q\}$; The number of selected features $t$.

**Output:**
The already-selected feature subset $S$.

1: $S \leftarrow \emptyset; L' \leftarrow \emptyset$
2: for $i = 1$ to $d$ do
3: for $j = 1$ to $q$ do
4: calculate the $I(f_i; l_j)$ and get the pair $f_0, l_0$ with the largest value;
5: end for
6: end for
7: Put $f_0$ in $S$, $l_0$ in $L'$ and set label_now=$l_0$;
8: $S = S \cup \{f_0\}$;
9: $L' = L' \cup \{l_0\}$;
10: label_now=$l_0$;
11: while $|S| < t$ do
12: if $|L'| = |L|$ then
13: select feature $f_i$ based on equation (20) ;
14: $S = S \cup \{f_i\}$;
15: else
16: select label $l_m$ based on equation (18) ;
17: $L' = L' \cup \{l_m\}$;
18: label_now=$l_m$;
19: Select feature $f_k$ based on equation (19) ;
20: $S = S \cup \{f_k\}$;
21: end if
22: end while
methods (D2F, PMU and SCLS). For ease of readability and comparison, we provided the result of in a tabular form in Table 1. Suppose that the number of instances is n, the number of features is d and the number of labels is q. The time complexity of mutual information, conditional mutual information and interaction information is $O(n)$ as all the instances need to be traveled for probability estimation. Suppose that the number of selected features is t, the time complexity of MRLFS is $O(ndq + tn(d + q))$. The time complexity of SCLS is $O(ndq + ndq^2)$. PMU and LRFS design the evaluation criteria to consider the second-order label correlations. The time complexity of PMU is $O(ndq + ndq + ndq^2)$.

V. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the classification performance of the proposed method, our method is compared to four state-of-the-art multi-label feature selection methods (MIFS, D2F, PMU and SCLS). Three evaluating metrics (Hamming Loss, Micro-F1 and Marco-F1) are used. And the classifiers used in the experiment are MLKNN classifier and SVM classifier. In this section, we first introduce the experimental data sets.

**TABLE 2. Description of data sets.**

<table>
<thead>
<tr>
<th>No.</th>
<th>Data sets</th>
<th>#Instances</th>
<th>#Features</th>
<th>#Labels</th>
<th>#Training</th>
<th>#Test</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arts</td>
<td>5000</td>
<td>462</td>
<td>26</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>2</td>
<td>Birds</td>
<td>645</td>
<td>260</td>
<td>19</td>
<td>322</td>
<td>323</td>
<td>Audio</td>
</tr>
<tr>
<td>3</td>
<td>Business</td>
<td>5000</td>
<td>438</td>
<td>30</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>4</td>
<td>Computers</td>
<td>5000</td>
<td>681</td>
<td>33</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>5</td>
<td>Education</td>
<td>5000</td>
<td>550</td>
<td>33</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>6</td>
<td>Emotions</td>
<td>593</td>
<td>72</td>
<td>6</td>
<td>391</td>
<td>202</td>
<td>Music</td>
</tr>
<tr>
<td>7</td>
<td>Enron</td>
<td>1702</td>
<td>1001</td>
<td>53</td>
<td>1123</td>
<td>579</td>
<td>Text</td>
</tr>
<tr>
<td>8</td>
<td>Entertain</td>
<td>5000</td>
<td>640</td>
<td>21</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>9</td>
<td>Genbase</td>
<td>662</td>
<td>1185</td>
<td>27</td>
<td>463</td>
<td>199</td>
<td>Biology</td>
</tr>
<tr>
<td>10</td>
<td>Health</td>
<td>5000</td>
<td>612</td>
<td>32</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>11</td>
<td>Medical</td>
<td>978</td>
<td>1449</td>
<td>45</td>
<td>333</td>
<td>645</td>
<td>Text</td>
</tr>
<tr>
<td>12</td>
<td>Recreation</td>
<td>5000</td>
<td>606</td>
<td>22</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>13</td>
<td>Scene</td>
<td>2407</td>
<td>294</td>
<td>6</td>
<td>1211</td>
<td>1196</td>
<td>Images</td>
</tr>
<tr>
<td>14</td>
<td>Science</td>
<td>5000</td>
<td>743</td>
<td>40</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>15</td>
<td>Social</td>
<td>5000</td>
<td>1047</td>
<td>39</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
<tr>
<td>16</td>
<td>Society</td>
<td>5000</td>
<td>636</td>
<td>27</td>
<td>2000</td>
<td>3000</td>
<td>Text</td>
</tr>
</tbody>
</table>

**TABLE 3. Experimental results of multi-label feature selection methods in terms of Marco-F1 score using SVM.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MRLFS</th>
<th>MIFS</th>
<th>D2F</th>
<th>PMU</th>
<th>SCLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.077±0.034</td>
<td>0.074±0.024</td>
<td>0.031±0.005</td>
<td>0.014±0.007</td>
<td>0.027±0.005</td>
</tr>
<tr>
<td>Birds</td>
<td>0.115±0.067</td>
<td>0.075±0.036</td>
<td>0.077±0.041</td>
<td>0.075±0.036</td>
<td>0.039±0.026</td>
</tr>
<tr>
<td>Business</td>
<td>0.048±0.006</td>
<td>0.045±0.005</td>
<td>0.046±0.002</td>
<td>0.032±0.003</td>
<td>0.041±0.001</td>
</tr>
<tr>
<td>Computers</td>
<td>0.21±0</td>
<td>0.02±0</td>
<td>0.42±0</td>
<td>0.02±0</td>
<td>0.42±0</td>
</tr>
<tr>
<td>Education</td>
<td>0.059±0.024</td>
<td>0.033±0.016</td>
<td>0.046±0.009</td>
<td>0.027±0.008</td>
<td>0.038±0.006</td>
</tr>
<tr>
<td>Emotions</td>
<td>0.337±0.094</td>
<td>0.165±0.117</td>
<td>0.315±0.064</td>
<td>0.239±0.099</td>
<td>0.336±0.057</td>
</tr>
<tr>
<td>Enron</td>
<td>0.126±0.037</td>
<td>0.074±0.017</td>
<td>0.132±0.042</td>
<td>0.127±0.045</td>
<td>0.119±0.031</td>
</tr>
<tr>
<td>Entertain</td>
<td>0.117±0.03</td>
<td>0.059±0.016</td>
<td>0.081±0.006</td>
<td>0.051±0.004</td>
<td>0.067±0.006</td>
</tr>
<tr>
<td>Genbase</td>
<td>0.772±0.153</td>
<td>0.689±0.109</td>
<td>0.706±0.107</td>
<td>0.628±0.093</td>
<td>0.241±0.022</td>
</tr>
<tr>
<td>Health</td>
<td>0.134±0.039</td>
<td>0.062±0.028</td>
<td>0.089±0.008</td>
<td>0.078±0.008</td>
<td>0.089±0.01</td>
</tr>
<tr>
<td>Medical</td>
<td>0.316±0.103</td>
<td>0.222±0.048</td>
<td>0.191±0.055</td>
<td>0.188±0.057</td>
<td>0.079±0.013</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.103±0.031</td>
<td>0.095±0.025</td>
<td>0.077±0.009</td>
<td>0.026±0.002</td>
<td>0.044±0.004</td>
</tr>
<tr>
<td>Scene</td>
<td>0.468±0.085</td>
<td>0.206±0.145</td>
<td>0.462±0.08</td>
<td>0.473±0.089</td>
<td>0.26±0.045</td>
</tr>
<tr>
<td>Science</td>
<td>0.057±0.02</td>
<td>0.044±0.018</td>
<td>0.021±0.003</td>
<td>0.009±0.005</td>
<td>0.016±0.004</td>
</tr>
<tr>
<td>Social</td>
<td>0.1±0.034</td>
<td>0.051±0.031</td>
<td>0.07±0.01</td>
<td>0.052±0.012</td>
<td>0.052±0.006</td>
</tr>
<tr>
<td>Society</td>
<td>0.065±0.022</td>
<td>0.034±0.017</td>
<td>0.051±0.009</td>
<td>0.044±0.006</td>
<td>0.027±0.006</td>
</tr>
<tr>
<td>Average values</td>
<td>0.1821</td>
<td>0.1218</td>
<td>0.1510</td>
<td>0.1303</td>
<td>0.0935</td>
</tr>
</tbody>
</table>
and parameter setting of the experiments. Second, extensive experimental results are used to analyze the performance of our method. The experimental framework is presented in Figure 1.

### A. DATASET AND EXPERIMENTAL SETTING

Sixteen real-world data sets are used in our experiments, which are collected from Mulan Library [45]. They are from five different application areas. Birds data set is relevant to Text. Scene data set is relevant to images. Emotions is a multi-label data set that is relevant to music [34]. Genbase data set is relevant to Biology. The remaining data sets are widely applied to text categorization [46]. Table 1 displays the details of the data sets. The training set and test set have been already separated in Mulan Library and the number of instances of training and test data set are shown in the last two columns of Table 2.

Experimental settings are introduced as follows: first, continuous features are discretized using an equal-width strategy into three bins, as recommend in the literature [14], [47]. Second, the number of already-selected features $b$ varies from 1 to $M$ with a step size of 1, where $M$ is 20% of the total

### TABLE 4. Experimental results of multi-label feature selection methods in terms of Mirco-F1 score using SVM.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MRLFS</th>
<th>MIIFS</th>
<th>D2F</th>
<th>PMU</th>
<th>SCLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.15±0.073</td>
<td>0.165±0.054</td>
<td>0.075±0.013</td>
<td>0.033±0.016</td>
<td>0.072±0.015</td>
</tr>
<tr>
<td>Birds</td>
<td>0.17±0.099</td>
<td>0.116±0.06</td>
<td>0.135±0.076</td>
<td>0.129±0.056</td>
<td>0.06±0.041</td>
</tr>
<tr>
<td>Business</td>
<td>0.67±0.005</td>
<td>0.675±0.002</td>
<td>0.673±0.003</td>
<td>0.669±0.002</td>
<td>0.671±0.001</td>
</tr>
<tr>
<td>Computers</td>
<td>0.411±0</td>
<td>0.411±0</td>
<td>0.411±0001</td>
<td>0.411±0001</td>
<td>0.410</td>
</tr>
<tr>
<td>Education</td>
<td>0.183±0.075</td>
<td>0.115±0.064</td>
<td>0.117±0.017</td>
<td>0.077±0.014</td>
<td>0.138±0.023</td>
</tr>
<tr>
<td>Emotions</td>
<td>0.408±0.085</td>
<td>0.199±0.131</td>
<td>0.372±0.039</td>
<td>0.295±0.102</td>
<td>0.422±0.039</td>
</tr>
<tr>
<td>Enron</td>
<td>0.514±0.059</td>
<td>0.372±0.027</td>
<td>0.509±0.033</td>
<td>0.498±0.038</td>
<td>0.488±0.031</td>
</tr>
<tr>
<td>Entertain</td>
<td>0.239±0.079</td>
<td>0.115±0.046</td>
<td>0.163±0.015</td>
<td>0.096±0.013</td>
<td>0.149±0.016</td>
</tr>
<tr>
<td>Genbase</td>
<td>0.695±0.143</td>
<td>0.972±0.101</td>
<td>0.968±0.066</td>
<td>0.946±0.066</td>
<td>0.541±0.014</td>
</tr>
<tr>
<td>Health</td>
<td>0.469±0.035</td>
<td>0.386±0.051</td>
<td>0.418±0.012</td>
<td>0.391±0.029</td>
<td>0.406±0.004</td>
</tr>
<tr>
<td>Medical</td>
<td>0.714±0.142</td>
<td>0.715±0.106</td>
<td>0.629±0.07</td>
<td>0.625±0.075</td>
<td>0.37±0.009</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.186±0.063</td>
<td>0.175±0.051</td>
<td>0.138±0.016</td>
<td>0.038±0.003</td>
<td>0.07±0.007</td>
</tr>
<tr>
<td>Scene</td>
<td>0.483±0.075</td>
<td>0.235±0.164</td>
<td>0.477±0.07</td>
<td>0.485±0.07</td>
<td>0.298±0.048</td>
</tr>
<tr>
<td>Science</td>
<td>0.119±0.049</td>
<td>0.113±0.046</td>
<td>0.053±0.01</td>
<td>0.024±0.016</td>
<td>0.058±0.014</td>
</tr>
<tr>
<td>Social</td>
<td>0.387±0.134</td>
<td>0.199±0.121</td>
<td>0.396±0.072</td>
<td>0.31±0.07</td>
<td>0.384±0.049</td>
</tr>
<tr>
<td>Society</td>
<td>0.203±0.06</td>
<td>0.166±0.07</td>
<td>0.205±0.046</td>
<td>0.331±0.042</td>
<td>0.182±0.043</td>
</tr>
<tr>
<td>Average values</td>
<td>0.3915</td>
<td>0.3206</td>
<td>0.3586</td>
<td>0.3348</td>
<td>0.2949</td>
</tr>
</tbody>
</table>

### TABLE 5. Experimental results of multi-label feature selection methods in terms of hamming loss.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>MRLFS</th>
<th>MIIFS</th>
<th>D2F</th>
<th>PMU</th>
<th>SCLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.062±0.001</td>
<td>0.061±0.001</td>
<td>0.063±0.001</td>
<td>0.064±0.001</td>
<td>0.063±0.001</td>
</tr>
<tr>
<td>Birds</td>
<td>0.051±0.001</td>
<td>0.052±0.001</td>
<td>0.053±0.002</td>
<td>0.052±0.002</td>
<td>0.054±0.002</td>
</tr>
<tr>
<td>Business</td>
<td>0.028±0</td>
<td>0.028±0</td>
<td>0.029±0</td>
<td>0.029±0</td>
<td>0.029±0</td>
</tr>
<tr>
<td>Computers</td>
<td>0.042±0.001</td>
<td>0.045±0</td>
<td>0.044±0</td>
<td>0.044±0</td>
<td>0.043±0.001</td>
</tr>
<tr>
<td>Education</td>
<td>0.043±0.001</td>
<td>0.044±0.001</td>
<td>0.044±0.001</td>
<td>0.045±0.001</td>
<td>0.044±0.001</td>
</tr>
<tr>
<td>Emotions</td>
<td>0.296±0.014</td>
<td>0.337±0.023</td>
<td>0.294±0.014</td>
<td>0.319±0.009</td>
<td>0.279±0.007</td>
</tr>
<tr>
<td>Enron</td>
<td>0.051±0.002</td>
<td>0.057±0.001</td>
<td>0.052±0.001</td>
<td>0.052±0.001</td>
<td>0.053±0.001</td>
</tr>
<tr>
<td>Entertain</td>
<td>0.063±0.001</td>
<td>0.066±0.001</td>
<td>0.066±0.001</td>
<td>0.067±0.001</td>
<td>0.066±0.001</td>
</tr>
<tr>
<td>Genbase</td>
<td>0.004±0.008</td>
<td>0.003±0.005</td>
<td>0.003±0.004</td>
<td>0.005±0.004</td>
<td>0.031±0</td>
</tr>
<tr>
<td>Health</td>
<td>0.046±0.002</td>
<td>0.05±0.001</td>
<td>0.048±0.001</td>
<td>0.049±0.001</td>
<td>0.048±0.001</td>
</tr>
<tr>
<td>Medical</td>
<td>0.018±0.002</td>
<td>0.016±0.002</td>
<td>0.02±0.001</td>
<td>0.02±0.001</td>
<td>0.023±0</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.062±0.001</td>
<td>0.062±0.001</td>
<td>0.062±0.001</td>
<td>0.065±0.001</td>
<td>0.064±0.001</td>
</tr>
<tr>
<td>Scene</td>
<td>0.143±0.009</td>
<td>0.17±0.01</td>
<td>0.149±0.006</td>
<td>0.147±0.007</td>
<td>0.173±0.003</td>
</tr>
<tr>
<td>Science</td>
<td>0.035±0</td>
<td>0.035±0</td>
<td>0.036±0</td>
<td>0.036±0</td>
<td>0.036±0</td>
</tr>
<tr>
<td>Social</td>
<td>0.028±0.002</td>
<td>0.032±0.001</td>
<td>0.03±0</td>
<td>0.03±0</td>
<td>0.029±0.001</td>
</tr>
<tr>
<td>Society</td>
<td>0.058±0.001</td>
<td>0.06±0.001</td>
<td>0.059±0</td>
<td>0.06±0.001</td>
<td>0.059±0</td>
</tr>
<tr>
<td>Average values</td>
<td>0.0644</td>
<td>0.0699</td>
<td>0.0658</td>
<td>0.0678</td>
<td>0.0684</td>
</tr>
</tbody>
</table>
FIGURE 2. Classification performance on Entertain data set.

FIGURE 3. Classification performance on Health data set.

number of features (M = 17% in medical data set used in experiments). Third, MLKNN [44] is used as the multi-label classifier to evaluate the classification performance of the MRLFS method and other compared feature selection methods in terms of Hamming Loss. We set the number of nearest neighbors K as 10. Finally, linear-based Support Vector Machine (SVM) is implemented to evaluate the classification performance in terms of Macro-F1 and Micro-F1. MLkNN and SVM are two popular classifiers, they are widely applied in various literature [48], [49], [50], [51], [52]. We use the package scikit-learn and in Python 2.7 to implement the classifiers.

B. EXPERIMENTAL RESULT AND ANALYSIS

Tables 3–5 show the classification performance of each feature selection method in terms of hamming loss, Marco-F1
These tables present the details of the experimental results of each compared method providing the average classification results and standard deviations of the M groups of feature subsets selected on each data set. The best classification performance for each data set is shown in bold fonts and second best ones are in underline. In addition, the last rows record the average results of each feature selection method over all benchmark data sets.

Tables 3 and 4 record the evaluating results in terms of macro-F1 and micro-F1. From Table 3, we can see that MRLFS achieves the best classification performance on all data sets except data sets Enron and Scene, but MRLFS...
obtains sub-optimal results on the data sets Scene. Similarly, the proposed method obtains the best or sub-optimal results on all the data sets in terms of micro-F1 in Table 4. Both Macro-F1 and Micro-F1 metrics are label-based metrics, the larger the values are, the better classification performance the method achieves. Furthermore, Hamming Loss is a example-based metric adopted for evaluating MRLFS and other compared methods. The results in Table 5 present HL that the proposed method obtains the best or second best performance on all data sets. At last, we calculate the average results of each method on all data sets in terms of the three evaluation metrics. The proposed method achieves the best classification performance on the average values in terms of three metrics. For example, in Table 3, MRLFS obtains 0.1821 compared to MIFS with 0.1218, D2F with 0.1510, PMU with 0.1303 and SCLS with 0.0935. Hamming loss is used to indicate the fraction that the labels are mis-classified. These results prove the conclusions in the previous paragraph again: our model is effective in classify all types of labels other than certain categories with majority.

To further prove our conclusion and clearly display the classification performance, Figures 2–5 show the classification results of MRLFS and other feature selection methods on four multi-label data sets (Entertain, Health, Scene and Science). X-axis of the figures indicates the percentage of the selected features while Y-axis represents the classification performance considering in terms of different evaluation metrics. The curves of different color represent different feature selection methods.

As shown in Figures 2–5, the proposed method outperforms the other compared methods significantly. On Entertain data set, our method is better than all the methods. As for other three data sets (Health, Scene and Science), MRLFS outperforms all the other methods. Overall, our method is better than MIFS, D2F, PMU and SCLS on these four data sets in terms of these three metrics.

VI. CONCLUSION

In this paper, a novel multi-label feature selection method is proposed named Multi-label Feature Selection based on the Min-relevance Label. Our general idea is to find a subset of labels where the labels are as different as possible to represent the whole data set and then select features based on it. In this process, mutual information is used to find the labels with minimal relevance and optimal features. The combination of mutual information between labels and features, and the feature redundancy term constitutes a function to select the features. To demonstrate the effectiveness of our method, MRLFS is compared to four state-of-the-art multi-label feature selection methods (MIFS, D2F, SCLS and PMU) using ML-KNN on 16 real-world multi-label data sets with the evaluating metrics of Hamming Loss. Additionally, the SVM classifier is used to evaluate the classification performance among the five feature selection methods in terms of Micro-F1 and Macro-F1. The experimental results proves that MRLFS is effective and reach our target. In future work, we have two main directions: first, we should further improve the performance of the model and optimize the existing problems with high-order label correlation; second, we could find a term to evaluate the diversity of data sets and adjust the structure of the datasets.

REFERENCES