Multi-Label Feature Selection Method Based on Dynamic Weight

Ping Zhang
Jilin University
Jiyao Sheng
Jilin University
Wanfu Gao (gaowf@jlu.edu.cn)
Jilin University
Juncheng Hu
Jilin University
Yonghao Li
Jilin University

Research Article

Keywords: Multi-label learning, Multi-label feature selection, Information theory, Weighted Feature Relevancy

Posted Date: July 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-604646/v1

Noname manuscript No.
(will be inserted by the editor)

Multi-label feature selection method based on dynamic weight

Ping Zhang · Jiyao Sheng · Wanfu Gao · Juncheng Hu · Yonghao Li

Received: date / Accepted: date

Abstract Multi-label feature selection attracts considerable attention from multi-label learning. Information-theory based multi-label feature selection methods intend to select the most informative features and reduce the uncertain amount of information of labels. Previous methods regard the uncertain amount of information of labels as constant. In fact, as the classification information of the label set is captured by features, the remaining uncertainty of each label is changing dynamically. In this paper, we categorize labels into two groups: one contains the labels with few remaining uncertainty, which means that most of classification information with respect to the labels has been obtained by the already-selected features; another group contains the labels with extensive remaining uncertainty, which means that the classification information of these labels is neglected by already-selected features. Feature selection aims to select the new features with highly relevant to the labels in the second group. Existing methods do not distinguish the difference between two label groups and ignore the dynamic change amount of information of labels. To this end, a Relevancy Ratio is designed to clarify the dynamic change amount of information of each label under the condition of the already-selected features. Afterwards, a Weighted Feature Relevancy is defined to evaluate the candidate features. Finally, a new multi-label Feature Selection method based

Wanfu Gao✉
College of Computer Science and Technology, JiLin University, Changchun, P.R.China;(Ping Zhang, Wanfu Gao, Juncheng Hu, Yonghao Li)
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, P.R.China;(Ping Zhang, Wanfu Gao, Juncheng Hu, Yonghao Li)
College of Chemistry, Jilin University, Changchun, P.R.China.(Jiyao Sheng and Wanfu Gao)
Ping Zhang: E-mail: zhangping18@mails.jlu.edu.cn
Jiyao Sheng: E-mail: shengjiyao@jlu.edu.cn
Wanfu Gao: E-mail: gaowf@jlu.edu.cn
Juncheng Hu: E-mail: jchu19@mails.jlu.edu.cn
Yonghao Li: E-mail: yonghao17@mails.jlu.edu.cn

on Weighted Feature Relevancy (WFRFS) is proposed. The experiments obtain encouraging results of WFRFS in comparison to six multi-label feature selection methods on thirteen real-world data sets.

Keywords Multi-label learning · Multi-label feature selection · Information theory · Weighted Feature Relevancy

1 Introduction

Recent years, multi-label learning has emerged in many areas such as text categorization [3,6,21], semantic image [35] and bioinformatics [38,9]. In multi-label data, each instance is associated with multiple labels simultaneously. High-dimensional multi-label data sets often contain many irrelevant and redundant features. These irrelevant and redundant features do not only increase the computational burden, but also degrade the classification performance of multi-label learning [23,11]. Multi-label feature selection intends to obtain a compact feature subset by selecting relevant features and eliminating the irrelevant and redundant features [36,13,29,8].

From the perspective with respect to the relationship between learning algorithm and feature selection, multi-label feature selection methods can be divided into three categories: filter methods, wrapper methods and embedded methods [24]. Filter methods are independent of any learning algorithm, and they use predefined criteria to evaluate the importance of features [27,33]. Wrapper methods depend on the classification performance of a specific classifier to select the optimal feature subset [37,7]. Embedded methods implement the classification task and the process of feature selection simultaneously [2]. Filter methods are simple and efficient. In this paper, we focus on the filter-based multi-label feature selection methods.

Different from single-label feature selection that deals with the data sets containing only one label, there are two ways to handle multi-label data in multi-label feature selection: problem transformation and algorithm adaption [30,19]. Problem transformation is a straightforward way that transforms the multi-label data set into single-label data sets (binary or multiclass) and then the feature subset is selected using single-label feature selection methods based on the transformed data sets. However, this way may create too many new labels or lose some label information. Algorithm adaption methods directly select features using the multi-label data set. Many algorithm adaptation-based multi-label feature selection methods have been proposed in recent years [22,14].

Information theory is widely utilized to measure the correlations between features and the label set in algorithm adaptation-based methods. Many multi-label feature selection methods based on information theory have been proposed [20,16,15,17], which have been proved to be effective in reducing high dimension. Mutual information is an effective criterion that measures the reduced uncertainty for one variable when another variable is given. In the process of feature selection, feature relevancy can be regarded as selecting the

features that maximize the reduction of uncertainty in label set to obtain the largest amount of classification information for label set. In fact, the uncertainty of labels is changing dynamically when different features are given. Feature selection should ensure that the selected features can effectively cut down the uncertainty for all labels. However, existing methods regard the uncertainty of labels as constant. To avoid the issue that the much information of some labels is not obtained, it is necessary to study the uncertain changes of labels under the effect of selected features. In this paper, we divide labels into two groups: the first group contains the labels with few remaining uncertainty; the second group contains labels that are contrary to the first group, that is, the remaining uncertainty of these labels is large under the condition of already-selected features. The proposed method aims to select the new features with highly relevant to the labels with extensive remaining uncertainty. The main contributions are as follows:

(1) A Relevancy Ratio is designed to clarify the degree of the contribution of the already-selected features to different labels.
(2) A new feature relevancy term named Weighted Feature Relevancy (WFR) is defined that combines the mutual information with the Relevancy Ratio to evaluate the importance of candidate features.
(3) A novel multi-label feature selection method named multi-label Feature Selection method based on Weighted Feature Relevancy (WFRFS) is proposed, which considers the WFR and the feature redundancy between candidate features and already-selected features.
(4) To evaluate the classification performance of the proposed method, WFRFS is compared to six multi-label feature selection methods on thirteen real-world multi-label data sets. The experimental results show that WFRFS obtains better classification performance in terms of multiple evaluation criteria.

The rest of this paper is organized as follows. Section 2 introduces the preliminaries including some basic concepts of information theory and four evaluation criteria for multi-label classification performance. Section 3 briefly reviews the related work. In Section 4, we present the proposed multi-label feature selection method WFRFS. Section 5 conducts the experiments to verify the effectiveness of WFRFS. In Section 6, we make a conclusion and give the future research direction.

2 Preliminaries

2.1 Some basic concepts of information theory

In this subsection, we introduce some basic information-theoretic concepts for feature selection [26,4]. Information theory provides a way to measure the amount of information for random variables. Let $X = \{x_1, x_2, ..., x_n\}$ be a

discrete random variable. The information entropy $H(X)$ measures the uncertainty of X. It is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{1}$$

where $p(x_i)$ is the probability of x_i and the base of log is 2. Let $Y = \{y_1, y_2, ..., y_m\}$ be another discrete random variable. $H(X,Y)$ is the joint entropy of X and Y. $H(X|Y)$ is the conditional entropy of X given Y, which measures the remaining uncertainty of X under the condition of Y. $H(X,Y)$ and $H(X|Y)$ are defined as follows:

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i, y_j), \tag{2}$$

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i|y_j). \tag{3}$$

where $p(x_i, y_j)$ 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.

Mutual information is a measure of the amount of information shared by two variables. The greater the mutual information, the more relevant the two variables. The mutual information can be defined as:

$$I(X;Y) = H(Y) - H(Y|X). \tag{4}$$

Let Z be a discrete random variable, conditional mutual information between X and Y given Z is defined by:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \tag{5}$$

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

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z). \tag{6}$$

2.2 Multi-label evaluation metrics

In multi-label learning, the evaluation metrics of classification performance are different from the traditional single-label learning, which are more complicated [34]. In this paper, we employ Macro-F1, Micro-F1, Hamming Loss and Coverage Error as the evaluation metrics.

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

$$Macro - F1 = \frac{1}{q}\sum_{i=1}^{q} \frac{2TP_i}{2TP_i + FP_i + FN_i}. \tag{7}$$

where q is the number of labels, TP_i, FP_i and FN_i are the number of true positives, false positives and false negatives in the i-th label, respectively. Micro-F1 is a weighted average of the $F1$ score over all labels. Micro-F1 is defined as follows:

$$Micro - F1 = \frac{\sum_{i=1}^{q} 2TP_i}{\sum_{i=1}^{q}(2TP_i + FP_i + FN_i)}. \tag{8}$$

the larger the value of Macro-F1 and Micro-F1 is , the better the classification performance is.

Suppose that $D = \{(x_i, L_i), 1 \leq i \leq N\}$ is the test set and $L = \{l_1, l_2, ..., l_q\}$ is the label set, where 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) calculates the average fraction of misclassified labels on the test data. HL is defined as:

$$HL = \frac{1}{N}\sum_{i=1}^{N}\frac{|L_i' \oplus L_i|}{q}. \tag{9}$$

where $\oplus$ denotes the symmetric difference between the label set L_i and L_i'. The smaller the value of HL is, the better the classification performance is.

Coverage Error (CE) measures the average search depth in the label ranking list to cover all the correct labels for the instance. The label ranking list is obtained according to the real-valued likelihood between x_i and each label $l \in L_i$ based a multi-label classifier. The definition for CE is:

$$CE = \frac{1}{N}\sum_{i=1}^{N} max_{l \in L_i}\{r_i(l)\} - 1. \tag{10}$$

where $r_i(l)$ is the label rank of $l \in L_i$ corresponding to the instance x_i. The smaller the value of CE is, the better the classification performance is.

3 Related work

Recent years, many multi-label feature selection methods have been proposed. In problem transformation-based feature selection methods, N. Spolaôr et al [28] use Binary Relevance (BR) [1] and Label Power set (LP) [31] to transform the multi-label data sets into single-label data sets. And then, they employ ReliefF and mutual information to evaluate the features. Doquire et al [5] propose a multi-label feature selection method based on mutual information using Pruned Problem Transformation (PPT) [25] (PPT+MI). BR decomposes the label set into independent binary classes. LP transforms instance's label combinations into new single-labels. PPT removes the instance with label combinations that occur too infrequently by defining the minimum occurrence to improve the LP. CHI square statistic is also used to select the effective features (PPT+CHI)[25]. However, the problem transformation methods may create too many new classes or lose the label information.

Algorithm adaptation-based multi-label feature selection methods directly select features from the multi-label data set. S Kashef et al [12] propose a label-specific multi-label feature selection algorithm based on the Pareto dominance concept without data transformation, and the method considers the effects of each feature on each label and transforms the multi-label feature selection problem to a multi-objective optimization problem. Multi-label Informed Feature Selection (MIFS) [10] adopts the latent semantics of the multi-label data to exploit label correlations and alleviates the negative effects of noisy and incomplete labels for feature selection. Li et al [18] propose a granular multi-label feature selection method based on mutual information, the method firstly employs a balanced k-means method to granulate labels into several information granules containing the labels with local dependency and then takes into account the maximal correlation minimal redundancy criterion based on mutual information to evaluate the features on each information granule.

Lee et al [15] propose an information-theoretical-based multi-label feature selection method named PMU. Its evaluation criterion is defined as follows:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in S} \sum_{l_i \in L} I(f_k; f_j; l_i) - \sum_{l_i \in L} \sum_{l_j \in L} I(f_k; l_i; l_j). \quad (11)$$

where L is the label set and S is an already-selected feature subset. f_k is a candidate feature and f_j is an already-selected feature, 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. In addition, Multi-label feature selection method using interaction information (D2F) [16] is proposed to efficiently evaluate the dependency of features in multi-label data. The criterion of D2F is defined by:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in S} \sum_{l_i \in L} I(f_k; f_j; l_i). \quad (12)$$

According to the Formulas (11) and (12), PMU and D2F use the accumulated mutual information between candidate features and each label $\sum_{l_i \in L} I(f_k; l_i)$ to evaluate the feature relevancy. In addition, SCLS [17] is presented to design a new multi-label feature selection method based on scalable relevance evaluation, SCLS evaluates the conditional relevance of features more accurately when a large number of labels are involved. It is denoted as follows:

$$J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in S} \frac{I(f_k; f_j)}{H(f_k)} \sum_{l_i \in L} I(f_k; l_i). \quad (13)$$

Similar to the information-theoretical-based multi-label feature selection methods mentioned above, to the best of our knowledge, these existing feature selection methods do not consider the dynamic change of classification information of labels. In fact, as the classification information in the label set is obtained in the process of feature selection, the remaining uncertainty of labels is changing under the effect on the already-selected features. In this paper, we design a Relevancy Ratio to present the change of uncertainty of each label.

Moreover, a new feature relevancy term named Weighted Feature Relevancy (WFR) is defined based on Relevancy Ratio. Finally, a novel method named multi-label feature selection method based on Weighted Feature Relevancy (WFRFS) is proposed.

4 Proposed multi-label feature selection method

4.1 The definition of Weight Feature Relevancy

In multi-label data sets, there exist different probability distributions for each label in all the instances. The information-theoretical-based multi-label feature selection methods intend to select the feature that is closely related to the probability distribution of labels. Information entropy $H(.)$ can quantify the probability distribution of labels by numerical value, and it is the measure of uncertainty. The larger the value of entropy indicates the greater the uncertainty, and vice versa. An ideal feature f inclines to the same probability distribution with the label l_i, indicating that the uncertainty of l_i is 0 under the condition of f, that is, the conditional entropy $H(l_i|f) = 0$; an irrelevant feature f' is independent of l_i, that is, under the condition of f', the uncertainty of l_i remains unchanged, i.e., $H(l_i|f) = H(l_i)$. Let L be the label set. The task of feature selection is to find a feature subset S, which minimizes the conditional entropy $H(L|S)$, that is, the remaining uncertainty of label set L is the smallest given S. However, according to the definition of the conditional entropy Formula (3), it is difficult to directly calculate $H(L|S)$ due to the high dimensionality of the feature set and the limitation of the number of samples. Therefore, many feature selection methods focus on reducing the uncertainty of each label in the label set to obtain an approximately optimal feature subset [16,15,17].

Previous methods consider that the uncertainty of each label is constant when they evaluate the feature relevancy between candidate features and the label set. In fact, with the increase of already-selected features, the uncertainty of labels is changing dynamically in the process of feature selection. Let S be the already-selected feature subset and L be the label set. For a certain label $l_i \in L$, the initial uncertainty is represented by $H(l_i)$, and the remaining uncertainty of l_i is $H(l_i|f_j)$ under the condition of feature $f_j \in S$. The larger the value of $H(l_i|f_j)$ is, the greater the remaining uncertainty of l_i is, that is, f_j provides less information for l_i. The smaller the value of $H(l_i|f_j)$ is, the more classification information f_j provides for l_i. When the feature relevancy between candidate features and the label set is evaluated, we should not only consider the relationship between candidate features and each label, but also consider the dynamic change of uncertainty of the labels under the condition of already-selected features. If the remaining uncertainty of label l_i is very small under the condition of the already-selected features, it means that the already-selected features have provided enough classification information for the label l_i. In this situation, we should pay less attention to the feature

relevancy between the candidate features and l_i, and vice versa. Moreover, the features in S have different degrees of contribution to different labels. Therefore, in order to select the features that are highly correlated with each label obtained less classification information in the already-selected feature subset, firstly, we propose the definition of Relevancy Ratio to clarify the degree of the contribution of the already-selected features to different labels. Then, a new feature relevancy term is proposed based on Relevancy Ratio.

Definition 41 (Relevancy Ratio) *Let L be the label set and $l_i \in L$, and S is the already-selected feature subset and $f_j \in S$. Then, the Relevancy Ratio $R\_Ratio(l_i, S)$ for l_i is defined as follows:*

$$R\_Ratio(l_i, S) = \sum_{f_j \in S} \left(\frac{H(l_i|f_j)}{H(l_i)} \right). \tag{14}$$

where $H(l_i)$ is the uncertainty of l_i and $H(l_i|f_j)$ is the remaining uncertainty of label l_i given f_j. $R\_Ratio(l_i, S)$ calculates the proportion of the remaining uncertainty of the label l_i given already-selected features in the initial uncertainty of the label l_i. The Relevancy Ratio is changing dynamically with the increase of already-selected features. The larger the value of $R\_Ratio(l_i, S)$ is, the more the remaining information of the label l_i is, as a result, we should pay more attention to the feature relevancy between candidate features and the label l_i, and vice versa. Naturally, we propose a new feature relevancy term based on Relevancy Ratio.

Definition 42 (Weighted Feature Relevancy) *Let F be the original feature set, L is the label set and $l_i \in L$, S is the already-selected feature subset and $f_k \in F - S$, $f_j \in S$. Then, the Weighted Feature Relevancy (WFR) is defined as follows:*

$$\begin{aligned} \boldsymbol{Rel(f_k; L)} &= \sum_{l_i \in L} \{R\_Ratio(l_i, S) * I(f_k; l_i)\} \\ &= \sum_{l_i \in L} \left\{ \sum_{f_j \in S} \left(\frac{H(l_i|f_j)}{H(l_i)} \right) * I(f_k; l_i) \right\}. \end{aligned} \tag{15}$$

where $I(f_k; l_i) = H(l_i) - H(f_k|l_i)$ measures the relationship between the candidate feature f_k and the label l_i which means the reduced uncertainty of l_i given feature f_k. $R\_Ratio(l_i, S)$ is used as the weight coefficient to evaluate the dynamic change of the uncertainty of label l_i.

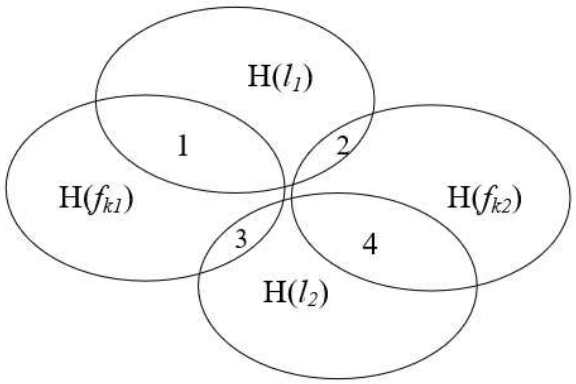

Fig. 1 The relationship between the candidate features f_{k1}, f_{k2} and the labels l_1, l_2.

We use Fig. 1 to illustrate the effect of $\boldsymbol{Rel(f_k; L)}$. f_{k1} and f_{k2} are two candidate features, and $L = \{l_1, l_2\}$ are two labels. Suppose that the Relevancy Ratio between the already-selected feature subset S and l_1, l_2 are $R\_Ratio(l_1, S)$ and $R\_Ratio(l_2, S)$, respectively. As shown in the Fig. 1, the area 1 is $I(f_{k1}; l_1)$, the area 2 is $I(f_{k2}; l_1)$, the area 3 is $I(f_{k1}; l_2)$ and the area 4 is $I(f_{k2}; l_2)$, respectively. Obviously, the area 1 is larger than the area 2 and the area 4 is larger than the area 3, that is $I(f_{k1}; l_1) > I(f_{k2}; l_1)$ and $I(f_{k1}; l_2) < I(f_{k2}; l_2)$. If $I(f_{k1}; l_1) + I(f_{k1}; l_2)$ is equal to $I(f_{k2}; l_1) + I(f_{k2}; l_2)$, that is, $1 + 3 = 2 + 4$. In such situation, traditional feature relevancy term, that is, the accumulated mutual information between candidate feature and each label cannot determine which feature is more important for label set L. Employing $\boldsymbol{Rel(f_k; L)}$ can effectively address this issue. According to the Formula (15), For f_{k1}, $\boldsymbol{Rel(f_{k1}; L)} = R\_Ratio(l_1, S) * I(f_{k1}; l_1) + R\_Ratio(l_2, S) * I(f_{k1}; l_2)$. For f_{k2}, $\boldsymbol{Rel(f_{k2}; L)} = R\_Ratio(l_1, S) * I(f_{k2}; l_1) + R\_Ratio(l_2, S) * I(f_{k2}; l_2)$. There exist three cases:

(1) If $R\_Ratio(l_1, S) > R\_Ratio(l_2, S)$, then the already-selected features provide more information for label l_2 than label l_1. In this case, more classification information for label l_1 should be obtained from the candidate features. Observing Fig. 1, f_{k1} is more informative with respect to l_1 than f_{k2}. According to $I(f_{k1}; l_1) + I(f_{k1}; l_2) = I(f_{k2}; l_1) + I(f_{k2}; l_2)$ and $R\_Ratio(l_1, S) > R\_Ratio(l_2, S)$, it holds that $\boldsymbol{Rel(f_{k1}; L) > Rel(f_{k2}; L)}$, which means that using $\boldsymbol{Rel(f_k; L)}$ can capture accurately the key features for labels.

(2) If $R\_Ratio(l_1, S) < R\_Ratio(l_2, S)$, then the already-selected features provide more classification information for label l_1 than label l_2. Similar to the analysis in the case (1), it holds that $\boldsymbol{Rel(f_{k1}; L) < Rel(f_{k2}; L)}$. In this case, f_{k2} is more important than f_{k1}.

(3) If $R\_Ratio(l_1, S) = R\_Ratio(l_2, S)$, then $\boldsymbol{Rel(f_{k1}; L) = Rel(f_{k2}; L)}$. In this case, f_{k1} and f_{k2} have the same relevancy for the label set L.

As shown in Fig. 2, when the candidate features f_{k1} and f_{k2} are related with respect to the labels l_1 and l_2. In this situation, the union of areas 1 and 5 is $I(f_{k1}; l_1)$, the union of areas 2 and 5 is $I(f_{k2}; l_1)$, the union of areas 3 and 6 is $I(f_{k1}; l_2)$ and the union of areas 4 and 6 is $I(f_{k2}; l_2)$, respectively. If $I(f_{k1}; l_1) + I(f_{k1}; l_2)$ is equal to $I(f_{k2}; l_1) + I(f_{k2}; l_2)$, that is, 1+5+3+6=2+5+4+6. Thus, the areas of 5 and 6 are the common information of two features with labels. Then, the effect of the common information is eliminated in the comparison of the two candidate features. It holds that 1+3=2+4, which is the same issue as Fig. 1. Therefore, the Formula (15) also holds.

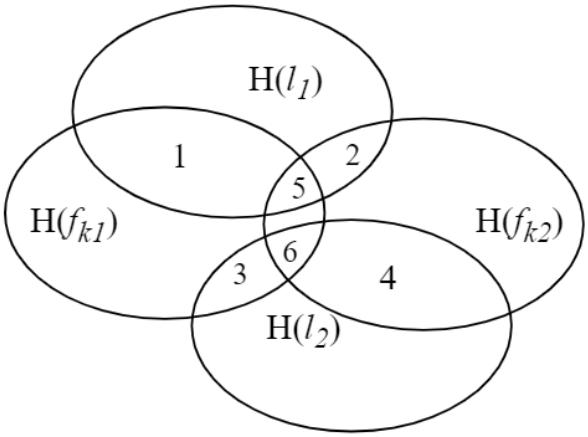

Fig. 2 The relationship between the candidate features f_{k1}, f_{k2} and the labels l_1, l_2.

In summary, employing the new feature relevancy can more accurately capture the discriminative features.

4.2 Proposed method

Based on the definition of WFR, we propose a novel multi-label feature selection method based on Weighted Feature Relevancy (WFRFS). The evaluation function is as follows:

$$\begin{aligned} J(f_k) &= \boldsymbol{Rel(f_k; L)} - \boldsymbol{Red(f_k; S)} \\ &= \sum_{l_i \in L} \left\{ \sum_{f_j \in S} \left(\frac{H(l_i|f_j)}{H(l_i)} \right) * I(f_k; l_i) \right\} - \sum_{f_j \in S} I(f_k; f_j). \end{aligned} \tag{16}$$

In the Formula (16), $\boldsymbol{Red(f_k; S)}$ can be considered as the feature redundancy term that is measured by the accumulated mutual information between candidate features and each already-selected feature. If the candidate features have

the same relevancy for the label set, then the features that have less redundancy with the already-selected features are selected. The sequential search strategy is used in the process of feature selection in our method. We select the feature f_k that achieves the maximal value of $J(f_k)$. The pseudo code of WFRFS is as follows:

Algorithm 1 WFRFS

Input:
A training sample D with a full feature set $F = \{f_1, f_2, ..., f_n\}$ and the label set $L = \{l_1, l_2, ..., l_q\}$; User-specified threshold K.
Output:
The already-selected feature subset S.

1: $S \leftarrow \varnothing$;
2: $k \leftarrow 0$;
3: **for** $i = 1$ to n **do**
4: calculate the accumulated mutual information $\sum_{l_j \in L} I(f_i; l_j)$;
5: **end for**
6: **while** $k < K$ **do**
7: **if** $k == 0$ **then**
8: select the feature f_j with the largest $\sum_{l_j \in L} I(f_i; l_j)$;
9: $k = k + 1$;
10: $S = S \cup \{f_j\}$;
11: $F = F - \{f_j\}$;
12: **end if**
13: **for** each candidate feature $f_i \in F$ **do**
14: Calculate the $\boldsymbol{Rel(f_i; L)}$;
15: Calculate the $I(f_i; f_j)$;
16: According to the Formula (16) update the $J(f_i)$;
17: **end for**
18: Select the feature f_j with the largest $J(f_i)$;
19: $S = S \cup \{X_j\}$;
20: $F = F - \{X_j\}$;
21: $k = k + 1$;
22: **end while**

There are three stages in the WFRFS method. In the first stage (lines 1-5), it initializes the parameters, which includes the already-selected feature subset S and the number of selected features k in lines 1-2, and calculates the accumulated mutual information between each feature and label set in lines 3-5. The second stage (lines 7-12) selects the maximum value of the accumulated mutual information as the first selected feature. The third stage (lines 13-21) calculates the Formula (16) to update the $J(f_i)$ for each feature.

5 Experimental results and analysis

In this section, we verify the effectiveness of the proposed method WFRFS on thirteen real-world multi-label data sets. First, the description of data sets and the experimental settings are introduced in Section 5.1. Then, WFRFS

is compared to three information-theoretical-based feature selection methods (D2F [16], PMU [15] and SCLS [17]) and two problem transformation-based methods (PPT+MI [5] and PPT+CHI [25]) and one embedded-based method (MIFS [10]) in terms of four evaluation metrics in Section 5.2.

5.1 Data sets and experimental settings

To evaluate the classification performance of WFRFS, the experiments are conducted on thirteen real-world multi-label data sets that are from Mulan Library [32]. The description of the data sets is presented in Table 1. These data sets cover five different application areas where the data set birds is used in the audio categorization, the data set emotions is used for the emotional classification in music, the data set scene is collected for image categorization, the data sets yeast and genbase are sampled from biological domain and the remaining data sets are widely applied to text categorization. The continuous features of these data sets are discretized into three bins using the equal-width strategy, as recommend in the literature [16]. In addition, the training set and test set have been already separated in Mulan Library [32].

Table 1 Description of data sets.

No.	Data set	#Instances	#Features	#Labels	#Training	#Test
1	birds	645	260	19	322	323
2	emotions	593	72	6	391	202
3	medical	978	1449	45	333	645
4	scene	2407	294	6	1211	1196
5	yeast	2417	103	14	1500	917
6	Education	5000	550	33	2000	3000
7	Entertain	5000	640	21	2000	3000
8	Health	5000	612	32	2000	3000
9	Science	5000	743	40	2000	3000
10	Social	5000	1047	39	2000	3000
11	Computers	5000	681	33	2000	3000
12	genbase	662	1185	27	463	199
13	Society	5000	636	27	2000	3000

The experimental setting is as follows: first, the number of already-selected features K varies from 1 to M with a step size of 1, where M is the 20% of the total number of features (M=17% in medical data set). Second, the MLKNN [39] (K=10) is employed as the multi-label classifier to evaluate the Hamming Loss and Coverage Error performance for the proposed method WFRFS and other six compared feature selection methods. Finally, the Liblinear-based Support Vector Machine (SVM) is used as the binary classifier to evaluate the Macro-F1 and Micro-F1 performance for seven feature selection methods.

5.2 Experimental comparison and analysis

Tables 2 and 3 show the multi-label classification performance in terms of Macro-F1 and Micro-F1 obtained by WFRFS and six compared methods. Tables 4-5 record the classification performance in terms of Hamming Loss and Coverage Error. These tables record the average classification results and the standard deviations across the M groups of feature subsets selected by each feature selection method. The bold fonts represent the best performance for each evaluation metric on thirteen real-world data sets. In addition, the last row "Avg.rank" presents the average rank of each method over all the multi-label data sets.

As shown in Tables 2 and 3, the proposed method WFRFS obtains better classification in terms of Macro-F1 and Micro-F1 than the compared methods on nine and ten data sets, respectively. As a result, WFRFS ranks the best in terms of the average rank. In particular, WFRFS significantly outperforms three information-theoretical-based feature selection methods D2F, PMU and SCLS on these data sets.

Table 2 Experimental results of seven feature selection methods in terms of Macro-F1 (mean ± std).

Data set	WFRFS	PPT+MI	PPT+CHI	MIFS	D2F	PMU	SCLS
birds	**0.1069±0.0504**	0.1026±0.0495	0.0981±0.0438	0.0754±0.0361	0.0767±0.0405	0.0755±0.0359	0.0386±0.026
emotions	0.385±0.0956	0.3736±0.1056	**0.3945±0.0938**	0.1654±0.1172	0.3149±0.0636	0.2392±0.0985	0.3363±0.0566
medical	**0.3178±0.0734**	0.2483±0.046	0.2609±0.0386	0.2216±0.0484	0.1912±0.0547	0.1882±0.0572	0.0793±0.0128
scene	0.4357±0.0759	0.2211±0.0886	0.2113±0.097	0.2063±0.1453	0.4625±0.0799	**0.4734±0.0887**	0.26±0.0454
yeast	**0.2762±0.0408**	0.2717±0.0329	0.2735±0.0311	0.2188±0.0488	0.2576±0.0345	0.2618±0.032	0.2073±0.0147
Education	**0.0653±0.0131**	0.0633±0.0113	0.0529±0.0148	0.0329±0.0159	0.046±0.0089	0.0267±0.0082	0.0379±0.0062
Entertain	**0.1221±0.0244**	0.1089±0.0255	0.0935±0.0197	0.0589±0.0163	0.0813±0.0061	0.0507±0.0042	0.0673±0.006
Health	**0.1435±0.0269**	0.1259±0.0337	0.1362±0.0285	0.0619±0.028	0.0886±0.0077	0.0781±0.0079	0.0894±0.0098
Science	**0.0546±0.0167**	0.0457±0.0153	0.0481±0.0129	0.0438±0.0184	0.0208±0.0031	0.0091±0.0053	0.0159±0.0036
Social	**0.1038±0.0287**	0.0925±0.0201	0.0936±0.0233	0.0508±0.0309	0.0696±0.0101	0.0518±0.012	0.0521±0.0062
Computers	0.0208±0.0001	**0.0209±0.0001**	0.0207±0.0002	0.0207±0.0001	0.0208±0.0001	0.0207±0.0001	0.0207±0.0001
genbase	**0.7651±0.1274**	0.6643±0.1076	0.6696±0.0939	0.6894±0.1093	0.706±0.1075	0.6277±0.0929	0.2413±0.0221
Society	0.068±0.0131	0.0581±0.0197	**0.0764±0.0122**	0.0342±0.0174	0.0514±0.0088	0.0443±0.0061	0.0267±0.0065
Avg.rank	**1.42**	2.92	2.81	5.73	3.96	5.50	5.65

Table 3 Experimental results of seven feature selection methods in terms of Micro-F1 (mean ± std).

Data set	WFRFS	PPT+MI	PPT+CHI	MIFS	D2F	PMU	SCLS
birds	**0.1984±0.0715**	0.1844±0.0734	0.1882±0.0679	0.116±0.0598	0.1354±0.0759	0.1285±0.056	0.0596±0.0408
emotions	0.4514±0.0751	0.4413±0.0906	**0.4655±0.0659**	0.1995±0.1306	0.3722±0.0395	0.295±0.1024	0.4223±0.0391
medical	**0.7569±0.0547**	0.7282±0.0484	0.736±0.0669	0.7149±0.1061	0.6293±0.0704	0.625±0.075	0.3697±0.0095
scene	0.4637±0.0716	0.2544±0.0982	0.2439±0.1083	0.235±0.164	0.4768±0.0703	**0.485±0.0781**	0.2979±0.0478
yeast	**0.5855±0.031**	0.5806±0.0245	0.5814±0.0227	0.5472±0.0363	0.5652±0.0239	0.5706±0.0216	0.5323±0.0084
Education	**0.2113±0.046**	0.1961±0.0416	0.1276±0.0413	0.1151±0.0641	0.1166±0.0173	0.077±0.0142	0.1376±0.0232
Entertain	**0.2594±0.0575**	0.2306±0.0612	0.1693±0.0491	0.1146±0.0457	0.1633±0.0146	0.0959±0.0128	0.1488±0.0163
Health	**0.4748±0.0383**	0.4536±0.0656	0.4718±0.0292	0.3856±0.0506	0.418±0.0118	0.3914±0.0289	0.4059±0.0038
Science	**0.1312±0.0339**	0.1157±0.029	0.0941±0.0285	0.1133±0.0457	0.0526±0.0105	0.0243±0.0158	0.058±0.0136
Social	**0.432±0.0759**	0.417±0.0741	0.3827±0.1434	0.199±0.1211	0.3955±0.0722	0.3097±0.0697	0.3843±0.0487
Computers	**0.411±0.0004**	0.411±0.0006	0.4115±0.0003	0.4113±0	0.4098±0.0008	0.4097±0.0009	0.4102±0.0002
genbase	**0.979±0.0674**	0.9684±0.0697	0.9648±0.0956	0.9717±0.1012	0.9678±0.0657	0.9458±0.0662	0.541±0.0137
Society	0.21±0.0462	0.2705±0.1101	**0.365±0.0471**	0.1665±0.0702	0.2053±0.0464	0.3305±0.0422	0.1818±0.0429
Avg.rank	**1.65**	2.88	2.92	5.38	4.46	5.38	5.31

Tables 4 and 5 show that the proposed method WFRFS achieves the best average rank than other compared methods in terms of Hamming Loss and Coverage Error. In Table 4, the average rank of WFRFS is 1.38, which is the best Hamming Loss performance compared to six feature selection methods on these data sets. In Table 5, the best average rank on Coverage Error performance is achieved by the proposed method WFRFS, followed by PPT+CHI, D2F, PMU, PPT+MI, SCLS and MIFS. In general, our method outperforms the compared feature selection methods.

Table 4 Experimental results of seven feature selection methods in terms of HL (mean ± std).

Data set	WFRFS	PPT+MI	PPT+CHI	MIFS	D2F	PMU	SCLS
birds	**0.0504±0.0011**	0.0531±0.0017	0.0512±0.0022	0.0518±0.001	0.0527±0.0017	0.0523±0.0017	0.0544±0.0019
emotions	0.2788±0.019	0.2875±0.0124	**0.2729±0.017**	0.337±0.023	0.2941±0.0137	0.3185±0.0095	0.2795±0.0074
medical	0.0174±0.001	0.0178±0.0013	0.0168±0.0015	**0.0165±0.0021**	0.0196±0.001	0.0197±0.0011	0.0233±0.0002
scene	**0.1435±0.0093**	0.167±0.0059	0.1674±0.0069	0.1704±0.0097	0.1492±0.0064	0.1473±0.0066	0.1734±0.003
yeast	0.2265±0.0036	0.2273±0.0043	**0.2258±0.0029**	0.2302±0.0041	0.2278±0.0029	0.2279±0.0037	0.2332±0.0044
Education	**0.0427±0.0007**	0.0431±0.0007	0.0434±0.0005	0.0436±0.0007	0.0443±0.0007	0.0445±0.0008	0.0441±0.001
Entertain	**0.0623±0.0015**	0.0641±0.0011	0.065±0.0008	0.0658±0.0008	0.0657±0.0013	0.0671±0.0011	0.0659±0.0014
Health	**0.045±0.0012**	0.0458±0.0009	0.045±0.002	0.0502±0.001	0.0483±0.0005	0.0493±0.0006	0.0485±0.0011
Science	**0.0352±0.0004**	0.0356±0.0006	0.0356±0.0002	0.0355±0.0003	0.0358±0.0004	0.0363±0.0004	0.0358±0.0004
Social	**0.0274±0.0008**	0.0278±0.0007	0.0296±0.0009	0.0317±0.0013	0.0303±0.0005	0.0309±0.0003	0.0287±0.0007
Computers	**0.0428±0.0008**	0.0432±0.0006	0.0435±0.0006	0.0449±0.0002	0.044±0.0005	0.0441±0.0005	0.0434±0.0005
genbase	**0.0026±0.004**	0.0029±0.0043	0.003±0.0053	0.0026±0.0055	0.0032±0.0039	0.0047±0.0041	0.0309±0.0004
Society	**0.0583±0.0006**	0.0589±0.0006	0.0587±0.0008	0.0596±0.0009	0.0587±0.0004	0.0597±0.0009	0.0594±0.0003
Avg.rank	**1.38**	3.27	2.81	4.81	4.54	5.77	5.42

Table 5 Experimental results of seven feature selection methods in terms of CE (mean ± std).

Data set	WFRFS	PPT+MI	PPT+CHI	MIFS	D2F	PMU	SCLS
birds	4.2896±0.094	4.3258±0.094	4.5997±0.2752	5.1498±0.1735	**4.2836±0.1303**	4.4849±0.1148	4.8103±0.0898
emotions	**3.5325±0.099**	3.6658±0.1542	3.5587±0.1499	4.3055±0.1765	3.7182±0.1272	3.9332±0.1439	3.6407±0.147
medical	5.7876±0.2897	5.9455±0.2748	**5.7447±0.2885**	6.1604±0.4141	6.3598±0.4012	6.4201±0.4025	8.3118±0.1098
scene	**2.2289±0.2958**	2.7558±0.3592	2.698±0.2863	2.9801±0.434	2.3015±0.2357	2.3129±0.2443	2.7828±0.1086
yeast	8.8679±0.4212	8.9803±0.3477	8.9062±0.2958	9.0812±0.506	**8.7833±0.2726**	8.9352±0.3673	9.0711±0.3446
Education	**6.3422±0.1765**	6.3911±0.2169	6.3883±0.2465	6.6549±0.3149	6.4315±0.1631	6.52±0.1874	6.5709±0.2155
Entertain	5.5483±0.2744	5.7594±0.249	**5.471±0.2402**	5.9338±0.5407	5.7088±0.2277	5.6683±0.2167	5.7602±0.1751
Health	5.7523±0.1505	5.8222±0.1233	5.9239±0.2105	6.228±0.376	5.7394±0.1555	**5.7229±0.1426**	5.8251±0.1802
Science	**10.2313±0.4803**	10.5728±0.3964	10.3599±0.4058	10.9172±0.5485	10.5342±0.358	10.5707±0.3044	10.8335±0.4279
Social	**5.8965±0.3843**	6.2753±0.2171	6.2043±0.2101	6.955±0.4051	6.1474±0.191	6.2101±0.1976	6.0108±0.3113
Computers	7.1504±0.2396	7.1848±0.28	**7.0589±0.382**	7.5371±0.5373	7.2455±0.2641	7.1926±0.2168	7.1822±0.2216
genbase	1.7429±0.2547	1.7437±0.2892	1.7246±0.2948	1.6516±0.289	**1.5958±0.267**	1.7925±0.2669	3.5393±0.0684
Society	**8.3571±0.3297**	8.4226±0.3732	8.4962±0.4755	8.6349±0.4791	8.4876±0.2665	8.4146±0.2669	8.5074±0.2349
Avg.rank	**1.77**	4.23	3.00	6.38	3.15	4.15	5.31

In Tables 2-3, the experimental results indicate that the Macro-F1 and Micro-F1 performances of the proposed method WFRFS on 12 experimental data sets are better than the other three information-theoretic based feature selection methods D2F, PMU and SCLS. In Table 4, the proposed method has better Hamming Loss performance than D2F, PMU and SCLS methods on all experimental data sets. The results show that it is useful to consider the dynamic change of the uncertainty of labels in design of information-theoretic based methods. In addition, WFRFS outperforms the other compared methods PPT+MI, PPT+CHI and MIFS in terms of Macro-F1, Micro-F1 and Hamming Loss performance on most data sets. In Table 5, PPT+CHI and D2F methods obtain better coverage error performance than WFRFS method on three data sets. WFRFS obtains better coverage error performance than PPT+MI, MIFS and SCLS methods on all experimental data sets. Overall, the selected feature subsets using the proposed feature selection method is more effective.

To show vividly the classification performance of our method and other compared feature selection methods, Figs. 3-5 show the experimental results on three data sets (Entertain, Science and Social). In these figures, the X-axis represents the number of already-selected features, which is varied as {1%, 2%, . . . , 20%} of the total number of features. The Y-axis represents the classification performance in terms of four evaluation criteria. Different colors and shapes represent different multi-label feature selection methods.

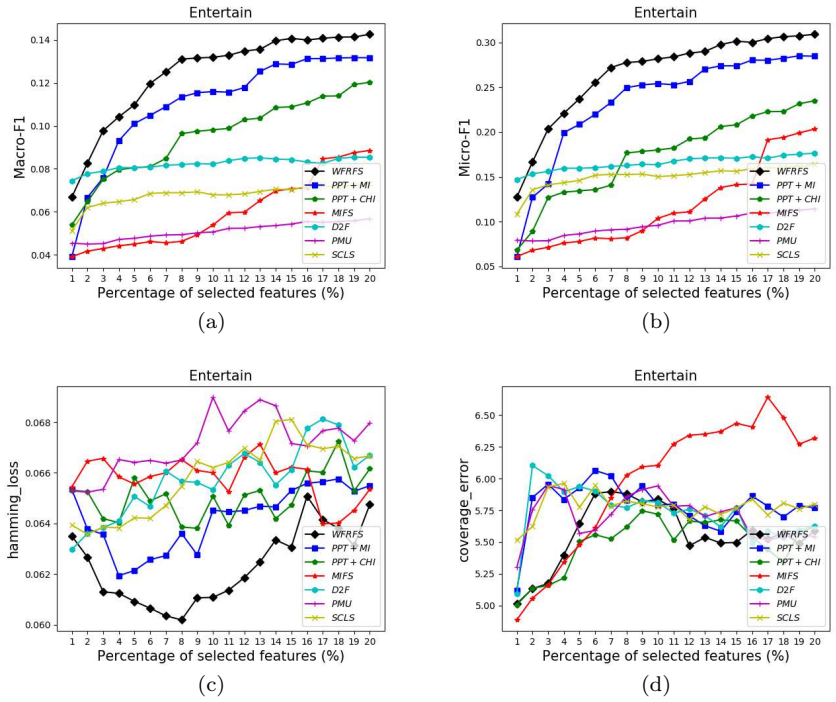

Fig. 3 Classification performance on Entertain data set: (a) Macro-F1, (b) Micro-F1, (c) Hamming Loss, (d) Coverage Error.

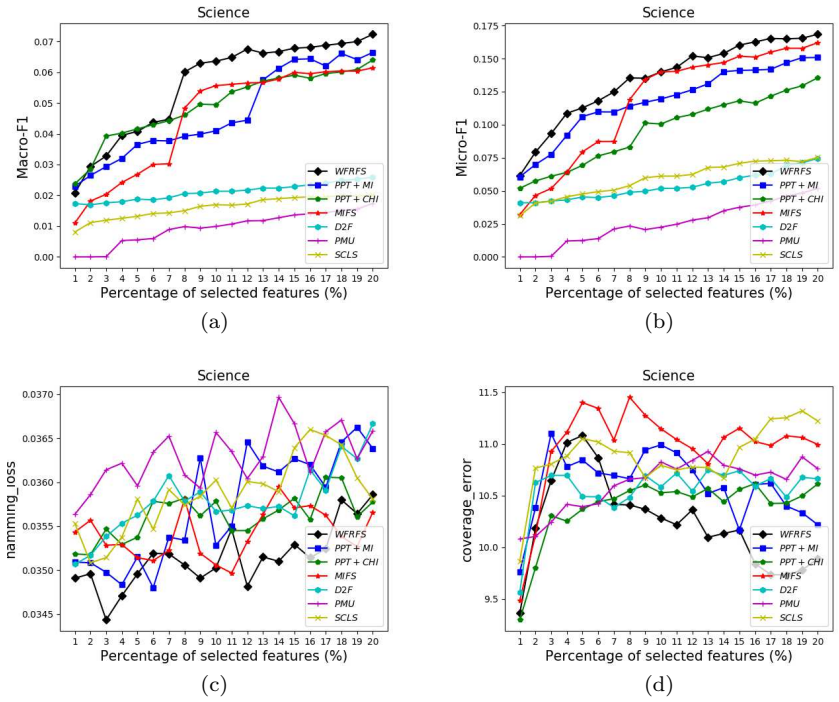

Fig. 4 Classification performance on Science data set: (a) Macro-F1, (b) Micro-F1, (c) Hamming Loss, (d) Coverage Error.

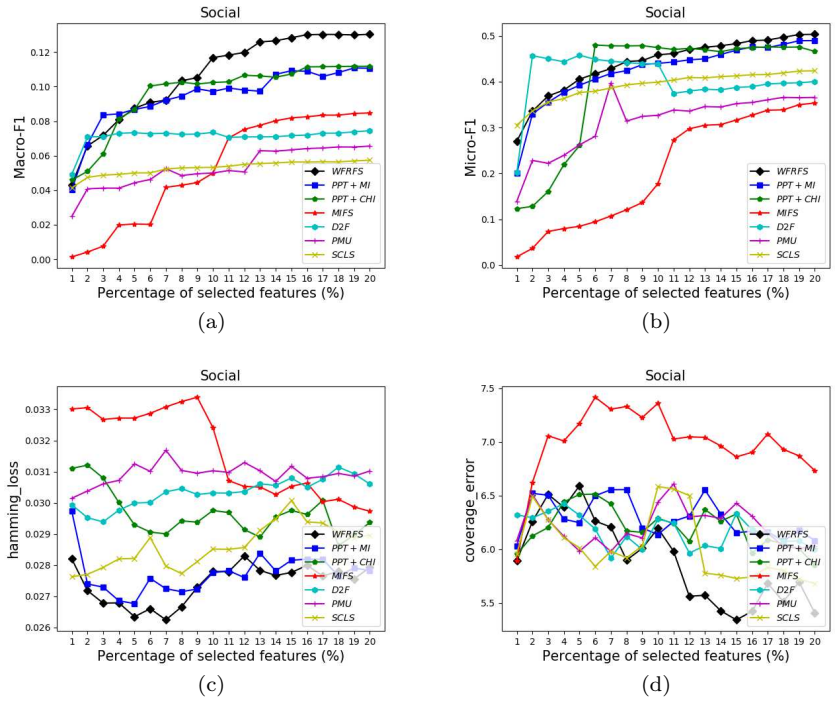

Fig. 5 Classification performance on Social data set: (a) Macro-F1, (b) Micro-F1, (c) Hamming Loss, (d) Coverage Error.

Fig. 3 shows the classification performance of seven feature selection methods on the Entertain data set. The proposed method WFRFS outperforms the compared methods. In particular, WFRFS significantly outperforms the compared methods in terms of the Macro-F1, Micro-F1 and Hamming Loss, as shown in Fig. 3(a), (b) and (c). Fig. 4 and Fig. 5 show the classification performance of each method on Science data set and Social data set respectively. The experiment results illustrate that our method obtains better classification performance as the size of the already-selected features subset grows larger.

We employ the Friedman test and the Nemenyi post-hoc test to present the statistical significance of the classification performance of the proposed method and other compared methods. Table 6 records the Friedman statistic FF of each evaluation criterion and the corresponding critical value. As the results shown in Table 6, the null hypothesis (i.e., all methods have equal classification performance) is clearly rejected in terms of four evaluation criteria at significance level α=0.05. Therefore, the Nemenyi post-hoc test is used to further analyze the relative performance of each pairwise methods. The classification performances of two feature selection methods are significantly different if the distance of the average ranks exceeds the critical distance CD. The value of CD is 2.499 at the significance level α=0.05. Fig. 6 shows the CD

results on each evaluation criterion, where the average ranks of seven methods are plotted along the axis.

Table 6 The Friedman statistics F_F and critical value in terms of each evaluation criterion.

Evaluation metrics	F_F(k=7,N=13)	Critical value ($\alpha = 0.05$)
Macro-F1	18.899	2.227
Micro-F1	11.204	
Hamming Loss	13.669	
Coverage Error	12.123	

In Fig. 6, the methods whose average rank within one CD is interconnected. The results indicate that the proposed method WFRFS ranks No.1 among all the methods. WFRFS method significantly performs better than D2F, PMU, SCLS and MIFS methods in terms of the Macro-F1, Micro-F1 and Hamming Loss evaluation criteria. Additionally, the proposed method does not significantly different with PPT+MI, PPT+CHI, D2F and PMU methods in terms of Coverage Error. As a result, the proposed method obtains highly competitive performance against the compared methods.

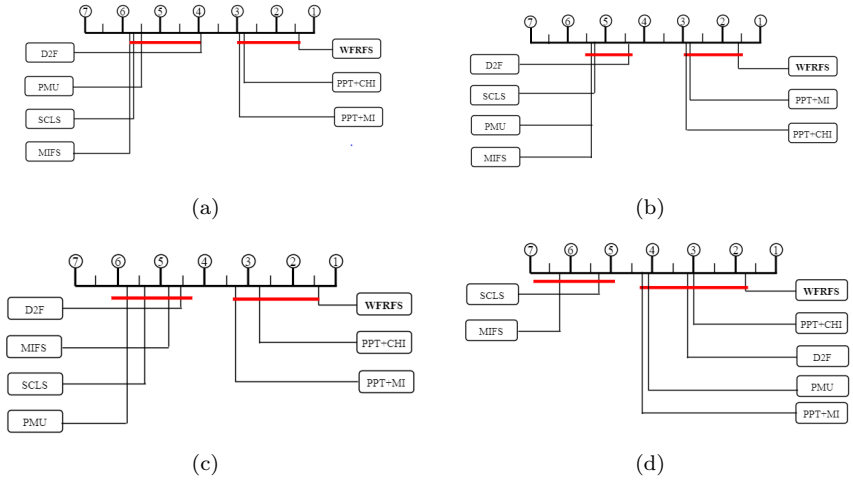

Fig. 6 The CD results: (a) Macro-F1, (b) Micro-F1, (c) Hamming Loss, (d) Coverage Error.

6 Conclusions and future work

Previous information-theoretical-based multi-label feature selection methods ignore the dynamic change of classification information of labels in the process of feature selection. To address this issue, first, we define a Relevancy Ratio to

present the change of uncertainty of each label. Second, we propose a new feature relevancy term named Weighted Feature Relevancy (WFR) that combines mutual information with the Relevancy Ratio to evaluate the relationship between candidate features and labels. Finally, the multi-label feature selection method based on Weighted Feature Relevancy (WFRFS) is proposed.

To verify the effectiveness of our method, WFRFS is compared to six representative multi-label feature selection methods (PPT+MI, PPT+CHI, MIFS, D2F, PMU and SCLS) using SVM classifier and MLKNN classifier on thirteen benchmark multi-label data sets in terms of Macro-F1, Micro-F1, Hamming Loss and Coverage Error. Compared to the six multi-label feature selection methods, WFRFS obtains better classification performance. As a result, we can conclude that WFRFS can effectively select the compact feature subset for multi-label data.

In our future work, we intend to explore the effect of label correlations in the process of feature selection.

Acknowledgements This work is supported by Postdoctoral Innovative Talents Support Program under Grant No. BX20190137; National Nature Science Foundation of China [grant number 61772226,61373051,61502343]; Science and Technology Development Program of Jilin Province [grant number 20140204004GX]; Science Research Funds for the Guangxi Universities [grant number KY2015ZD122]; Science Research Funds for the Wuzhou University [grant number 2014A002]; Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering); Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China; Fundamental Research Funds for the Central.

Compliance with ethical standards

Conflict of interest All the authors declare that there is no conflict of interest.

References

1. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9), 1757–1771 (2004)
2. Cai, Z., Zhu, W.: Multi-label feature selection via feature manifold learning and sparsity regularization. International Journal of Machine Learning and Cybernetics **9**(8), 1321–1334 (2018)
3. Chen, W., Yan, J., Zhang, B., Chen, Z., Yang, Q.: Document transformation for multi-label feature selection in text categorization. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 451–456. IEEE (2007)
4. Cover, T.M., Thomas, J.A.: Elements of information theory. Tsinghua University Pres (2003)
5. Doquire, G., Verleysen, M.: Mutual information-based feature selection for multilabel classification. Neurocomputing **122**, 148–155 (2013)
6. Esuli, A., Fagni, T., Sebastiani, F.: Boosting multi-label hierarchical text categorization. Information Retrieval **11**(4), 287–313 (2008)
7. Gharroudi, O., Elghazel, H., Aussem, A.: A comparison of multi-label feature selection methods using the random forest paradigm. In: Canadian conference on artificial intelligence, pp. 95–106. Springer (2014)

8. Gu, Q., Li, Z., Han, J.: Correlated multi-label feature selection. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1087–1096. ACM (2011)
9. Heider, D., Senge, R., Cheng, W., Hüllermeier, E.: Multilabel classification for exploiting cross-resistance information in hiv-1 drug resistance prediction. Bioinformatics **29**(16), 1946–1952 (2013)
10. Jian, L., Li, J., Shu, K., Liu, H.: Multi-label informed feature selection. In: IJCAI, pp. 1627–1633 (2016)
11. Kashef, S., Nezamabadi-pour, H.: An advanced aco algorithm for feature subset selection. Neurocomputing **147**, 271–279 (2015)
12. Kashef, S., Nezamabadi-pour, H.: A label-specific multi-label feature selection algorithm based on the pareto dominance concept. Pattern Recognition **88**, 654–667 (2019)
13. Khan, M.A., Ekbal, A., Mencía, E.L., Fürnkranz, J.: Multi-objective optimisation-based feature selection for multi-label classification. In: International Conference on Applications of Natural Language to Information Systems, pp. 38–41. Springer (2017)
14. Kong, X., Philip, S.Y.: gmlc: a multi-label feature selection framework for graph classification. Knowledge and Information Systems **31**(2), 281–305 (2012)
15. Lee, J., Kim, D.W.: Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters **34**(3), 349–357 (2013)
16. Lee, J., Kim, D.W.: Mutual information-based multi-label feature selection using interaction information. Expert Systems with Applications **42**(4), 2013–2025 (2015)
17. Lee, J., Kim, D.W.: Scls: Multi-label feature selection based on scalable criterion for large label set. Pattern Recognition **66**, 342–352 (2017)
18. Li, F., Miao, D., Pedrycz, W.: Granular multi-label feature selection based on mutual information. Pattern Recognition **67**, 410–423 (2017)
19. Lim, H., Lee, J., Kim, D.W.: Optimization approach for feature selection in multi-label classification. Pattern Recognition Letters **89**, 25–30 (2017)
20. Lin, Y., Hu, Q., Liu, J., Chen, J., Duan, J.: Multi-label feature selection based on neighborhood mutual information. Applied Soft Computing **38**, 244–256 (2016)
21. Luo, Q., Chen, E., Xiong, H.: A semantic term weighting scheme for text categorization. Expert Systems with Applications **38**(10), 12,708–12,716 (2011)
22. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern recognition **45**(9), 3084–3104 (2012)
23. Mampaey, M., Nijssen, S., Feelders, A., Konijn, R., Knobbe, A.: Efficient algorithms for finding optimal binary features in numeric and nominal labeled data. Knowledge and Information Systems **42**(2), 465–492 (2015)
24. Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.: Categorizing feature selection methods for multi-label classification. Artificial Intelligence Review **49**(1), 57–78 (2018)
25. Read, J.: A pruned problem transformation method for multi-label classification. In: Proc. 2008 New Zealand Computer Science Research Student Conference, pp. 143–150 (2008)
26. Shannon, C.E.A.: A mathematical theory of communication. at&t tech j. Acm Sigmobile Mobile Computing & Communications Review **5**(1), 3–55 (2001)
27. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: Filter approach feature selection methods to support multi-label learning based on relieff and information gain. In: Brazilian Symposium on Artificial Intelligence, pp. 72–81. Springer (2012)
28. SpolaôR, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science **292**, 135–151 (2013)
29. Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D.: A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing **180**, 3–15 (2016)
30. Suping, X.U., Yang, X., Yunsong, Q.I.: Multi-label learning with label-specific feature reduction. Journal of Computer Applications **104**, 52–61 (2016)
31. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. Blood **90**(9), 3438–3443 (2008)
32. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research **12**(7), 2411–2414 (2011)

33. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural computing and applications **24**(1), 175–186 (2014)
34. Wu, X.Z., Zhou, Z.H.: A unified view of multi-label performance measures. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 3780–3788. JMLR. org (2017)
35. Yu, Y., Pedrycz, W., Miao, D.: Neighborhood rough sets based multi-label classification for automatic image annotation. International Journal of Approximate Reasoning **54**(9), 1373–1387 (2013)
36. Yu, Y., Wang, Y.: Feature selection for multi-label learning using mutual information and ga. In: International Conference on Rough Sets and Knowledge Technology, pp. 454–463. Springer (2014)
37. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences **179**(19), 3218–3229 (2009)
38. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE transactions on Knowledge and Data Engineering **18**(10), 1338–1351 (2006)
39. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition **40**(7), 2038–2048 (2007)