Multi-label feature selection method based on dynamic weight

Ping Zhang1,2 · Jiyao Sheng3 · Wanfu Gao1,2,3 · Juncheng Hu1,2 · Yonghao Li1,2

Accepted: 27 November 2021 / Published online: 30 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

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 that are 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. Afterward, a Weighted Feature Relevancy is defined to evaluate the candidate features. Finally, a new multi-label feature selection method based on Weighted Feature Relevancy (WFRFS) is proposed. The experiments obtain encouraging results of WFRFS in comparison with 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
In recent years, multi-label learning has emerged in many areas such as text categorization (Chen et al. 2007; Esuli et al. 2008; Luo et al. 2011), semantic image (Yu et al. 2013) and bioinformatics (Zhang and Zhou 2006; Heider et al. 2013). 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 (Mampaey et al. 2015; Kashef and Nezamabadi-pour 2015). Multi-label feature selection intends to obtain a compact feature subset by selecting relevant features and eliminating the irrelevant and redundant features (Yu and Wang 2014; Khan et al. 2017; Spolaôr et al. 2016; Gu et al. 2011).

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 (Pereira et al. 2018). Filter methods are independent of any learning algorithm, and they use predefined criteria to evaluate the importance of features (Spolaôr et al. 2012; Vergara and Estévez 2014). Wrapper methods depend on the
classification performance of a specific classifier to select the optimal feature subset (Zhang et al. 2009; Gharroudi et al. 2014). Embedded methods implement the classification task and the process of feature selection simultaneously (Cai and Zhu 2018). 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 (Suping et al. 2016; Lim et al. 2017). 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 (Madjarov et al. 2012; Kong and Philip 2012).

Information theory is widely utilized to measure the correlations between features and the label set in algorithm adaption-based methods. Many multi-label feature selection methods based on information theory have been proposed (Lin et al. 2016; Lee and Kim 2015, 2013, 2017), 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 in 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 that are 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 Weighted Feature Relevance (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 Sect. 4, we present the proposed multi-label feature selection method WFRFS. Section 5 conducts the experiments to verify the effectiveness of WFRFS. In Sect. 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 (Shannon 2001; Cover and Thomas 2003). 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:

\[
H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)
\]

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

\[ H(X|Y) = - \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i|y_j) \] (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 follows:

\[ I(X; Y) = H(Y) - H(Y|X). \] (4)

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

\[ I(X; Y|Z) = H(X|Z) - H(X|Y, Z). \] (5)

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

\[ I(X; Y; Z) = I(X; Y) - I(X; Y|Z). \] (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 (Wu and Zhou 2017). 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 follows:

\[ \text{Macro - F1} = \frac{1}{q} \sum_{i=1}^{q} \frac{2TP_i}{2TP_i + FP_i + FN_i} \] (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:

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

\[ \text{HL} = \frac{1}{N} \sum_{i=1}^{N} |L'_i \oplus L_i| \] (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 as follows:

\[ \text{CE} = \frac{1}{N} \sum_{i=1}^{N} \max_{l \in L_i} \{r_i(l)\} - 1 \] (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

In recent years, many multi-label feature selection methods have been proposed. In problem transformation-based feature selection methods, N. Spolaôr et al. SpolaôR et al. (2013) use Binary Relevance (BR) Boutell et al. (2004) and Label Power set (LP) Trohidis et al. (2008) 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. Doquire and Verleysen (2013) propose a multi-label feature selection method based on mutual information using Pruned Problem Transformation (PPT) Read (2008) (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) (Read 2008). 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. Kashef and Nezamabadi-pour (2019) 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) (Jian et al. 2016) 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. Li et al. (2017) 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. Lee and Kim (2013) propose an information-theoretical-based multi-label feature selection method named PMU. Its evaluation criterion is defined as follows:

\[ J(f_k) = \sum_{l_i \in L} I(f_k; l_i) - \sum_{f_j \in S} \sum_{l_i \in L} I(f_k; f_j; l_i) - \sum_{l_i \in L} I(f_k; l_i; l_j) \]  

(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) (Lee and Kim 2015) is proposed to efficiently evaluate the dependency of features in multi-label data. The criterion of D2F is defined by

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

(12)

According to 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 (Lee and Kim 2017) 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:

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

(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 (Lee and Kim 2015, 2013, 2017).

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 in 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 can-
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 4.1 [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 $R\text{-}Ratio(l_i, S)$ for $l_i$ is defined as follows:

$$R\text{-}Ratio(l_i, S) = \sum_{f_j \in S} \left( \frac{H(l_i|f_j)}{H(l_i)} \right)$$

(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 $f_j$. $R\text{-}Ratio(l_i, S)$ 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 in already-selected features. The larger the value of $R\text{-}Ratio(l_i, S)$ 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 4.2 [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:

$$Rel(f_k; L) = \sum_{l_i \in L} [R\text{-}Ratio(l_i, S) \ast I(f_k; l_i)]$$

$$= \sum_{l_i \in L} \left\{ \sum_{f_j \in S} \left( \frac{H(l_i|f_j)}{H(l_i)} \right) \ast I(f_k; l_i) \right\}$$

(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 $f_k$. $R\text{-}Ratio(l_i, S)$ is used as the weight coefficient to evaluate the dynamic change of the uncertainty of label $l_i$.

We use Fig. 1 to illustrate the effect of $Rel(f_k; L)$. $f_{k1}$ 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 $l_1, l_2$ is $R\text{-}Ratio(l_1, S)$ and $R\text{-}Ratio(l_2, S)$, respectively. As shown in 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 $Rel(f_k; L)$ can effectively address this issue. According to Formula (15), For $f_{k1}$, $Rel(f_{k1}; L) = R\text{-}Ratio(l_1, S) \ast I(f_{k1}; l_1) + R\text{-}Ratio(l_2, S) \ast I(f_{k1}; l_2)$. For $f_{k2}$, $Rel(f_{k2}; L) = R\text{-}Ratio(l_1, S) \ast I(f_{k2}; l_1) + R\text{-}Ratio(l_2, S) \ast I(f_{k2}; l_2)$. There exist three cases:
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      | 919   |
| 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  |

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 | 0.3636 ± 0.0566   |
| medical    | 0.3178 ± 0.0734 | 0.2483 ± 0.0464 | 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 ± 0.0977 | 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 ± 0.0332 | 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      |

(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 $I(f_{k1}; l_1) + I(f_{k2}; l_2) = I(f_{k2}; l_1) + I(f_{k2}; l_2)$ and $R_{Ratio}(l_1, S) > R_{Ratio}(l_2, S)$, it holds that $Rel(f_{k1}; L) > Rel(f_{k2}; L)$, which means that using $Rel(f_{k1}; L)$ 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 $Rel(f_{k1}; L) < Rel(f_{k2}; L)$. In this case, $f_{k2}$ is more important than $f_{k1}$.

(3) If $R_{Ratio}(l_1, S) = R_{Ratio}(l_2, S)$, then $Rel(f_{k1}; L) = Rel(f_{k2}; L)$. 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 $I(f_{k1}; l_1) + I(f_{k2}; l_1)$ 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, Formula (15) also holds.
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:

$$J(f_k) = Rel(f_k; L) - Rel(f_k; S)$$

$$= \sum_{l_i \in L} \left( \sum_{f_j \in S} \left( \frac{H(l_i | f_j)}{H(l_i)} \right) * I(f_k; l_i) \right)$$

In Formula (16), $Rel(f_k; L)$ 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 pseudocode of WFRFS is as follows:

In summary, employing the new feature relevancy can more accurately capture the discriminative features.
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 | 3.5857 ± 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 ± 0.1411 | 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 | 5.6863 ± 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 |

Algorithm 1 WFRFS

**Input:**
- A training sample \( D \) with a full feature set \( F = \{ f_1, f_2, ..., f_n \} \) and the label set \( L = \{ l_1, l_2, ..., l_k \} \); User-specified threshold \( K \).

**Output:**
- The already-selected feature subset \( S \).

1. \( S \leftarrow \emptyset \);
2. \( k \leftarrow 0 \);
3. for \( i = 1 \) to \( n \) do
   4. calculate the accumulated mutual information \( \sum_{j \in L} I(f_i; l_j) \);
5. end for
6. while \( k < K \) do
7. if \( k == 0 \) then
8. select the feature \( f_j \) with the largest \( \sum_{l \in L} I(f_i; l_j) \);
9. \( k = k + 1 \);
10. \( S = S \cup \{ f_j \} \);
11. \( F = F - \{ f_j \} \);
12. end if
13. for each candidate feature \( f_j \in F \) do
14. Calculate the Rel \( (f_j; L) \);
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 \( J(f_j) \);
19. \( S = S \cup \{ f_j \} \);
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 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 Sect. 5.1. Then, WFRFS is compared to three information-theoretical-based feature selection methods (D2F (Lee and Kim 2015), PMU (Lee and Kim 2013) and SCLS (Lee and Kim 2017)) and two problem transformation-based methods (PPT+MI (Doquire and Verleysen 2013) and PPT+CHI (Read 2008)) and one embedded-based method (MIFS (Jian et al. 2016)) in terms of four evaluation metrics in Sect. 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 (Tsoumakas et al. 2011). 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 set 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 (Lee and Kim 2015). In addition, the training set and test set have been already separated in Mulan Library (Tsoumakas et al. 2011).

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 (Zhang and Zhou 2007) ($K=10$) 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,
Fig. 4  Classification performance on Science data set: a Macro-F1, b Micro-F1, c Hamming Loss, d Coverage Error

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.

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.

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 are more effective.
To show vividly the classification performance of our method and other compared feature selection methods, Figs. 3, 4 and 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.

Figure 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. 3a, b and c. Figures 4 and 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. The Friedman statistics $F_F$ and critical value in terms of each evaluation criterion are shown in Table 6. The Friedman test is a non-parametric statistical test used to compare multiple groups. The Nemenyi post hoc test is a post hoc test used to determine which groups are different from each other. The experiment results illustrate that our method obtains better classification performance as the size of the already-selected features subset grows larger.

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               |                                  |

Fig. 5 Classification performance on Social data set: a Macro-F1, b Micro-F1, c Hamming Loss, d Coverage Error
performance of the proposed method and other compared methods. Table 6 records the Friedman statistic $F_F$ 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 $\alpha = 0.05$. 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 $\alpha = 0.05$. Fig. 6 shows the CD results on each evaluation criterion, where the average ranks of seven methods are plotted along the axis.

In Fig. 6, the methods whose average rank is within one CD are 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 differ 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.

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 20140204004G]; 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
Multi-label feature selection method...

Education Ministry of China; Fundamental Research Funds for the Central.

Declarations

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

References

Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recogn 37(9):1757–1771

Cai Z, Zhu W (2018) Multi-label feature selection via feature manifold learning and sparsity regularization. Int J Mach Learn Cybern 9(8):1321–1334

Chen W, Yan J, Zhang B, Chen Z, Yang Q (2007) Document transformation for multi-label feature selection in text categorization. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 451–456. IEEE

Cover TM, Thomas JA (2003) Elements of information theory. Tsinghua University Press

Doquire G, Verleysen M (2013) Mutual information-based feature selection for multi-label classification. Neurocomputing 122:148–155

Esuli A, Fagni T, Sebastiani F (2008) Boosting multi-label hierarchical text categorization. Inf Retr 11(4):287–313

Gharroudi O, Elghazel H, Aussem A (2014) A comparison of multi-label feature selection methods using the random forest paradigm. In: Canadian conference on artificial intelligence, pp. 95–106. Springer

Gu Q, Li Z, Han J (2011) Correlated multi-label feature selection. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1087–1096. ACM

Heider D, Senge R, Cheng W, Hüllermeier E (2013) Multi-label classification for exploiting cross-resistance information in hiv-1 drug resistance prediction. Bioinformatics 29(16):1946–1952

Jian L, Li J, Shu K, Liu H (2016) Multi-label informed feature selection. In: IJCAI, pp. 1627–1633

Kashef S, Nezamabadi-pour H (2015) An advanced aco algorithm for feature subset selection. Neurocomputing 147:271–279

Kashef S, Nezamabadi-pour H (2019) A label-specific multi-label feature selection algorithm based on the pareto dominance concept. Pattern Recogn 88:654–667

Khan MA, Ekbal A, Mencía EL, Fürnkranz J (2017) 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

Kong X, Philip SY (2012) glmc: a multi-label feature selection framework for graph classification. Knowl Inf Syst 31(2):281–305

Lee J, Kim DW (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recogn Lett 34(3):349–357

Lee J, Kim DW (2015) Mutual information-based multi-label feature selection using interaction information. Exp Syst Appl 42(4):2013–2025

Lee J, Kim DW (2017) Scs: Multi-label feature selection based on scalable criterion for large label set. Pattern Recogn 66:342–352

Li F, Miao D, Pedrycz W (2017) Granular multi-label feature selection based on mutual information. Pattern Recogn 67:410–423

Lim H, Lee J, Kim DW (2017) Optimization approach for feature selection in multi-label classification. Pattern Recogn Lett 89:25–30

Lin Y, Hu Q, Liu J, Chen J, Duan J (2016) Multi-label feature selection based on neighborhood mutual information. Appl Soft Comput 38:244–256

Luo Q, Chen E, Xiong H (2011) A semantic term weighting scheme for text categorization. Exp Syst Appl 38(10):12708–12716

Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S (2012) An extensive experimental comparison of methods for multi-label learning. Pattern Recogn 45(9):3084–3104

Mampaey M, Nijsse S, Feelders A, Konijn R, Knobbe A (2015) Efficient algorithms for finding optimal binary features in numeric and nominal labeled data. Knowl Inf Syst 42(2):465–492

Pereira RB, Plasticino A, Zadrozny B, Merschmann L.H (2018) Categorizing feature selection methods for multi-label classification. Artif Intell Rev 49(1):57–78

Read J (2008) A pruned problem transformation method for multi-label classification. In: Proc. 2008 New Zealand Computer Science Research Student Conference, pp. 143–150

Shannon CEA (2001) A mathematical theory of communication. at&t tech j. Acm Sigmobile Mobile Comput Commun Rev 5(1):3–55

Spolaor N, Gomiero EA, Brand OK, Lee HD (2012) Filter approach feature selection methods to support multi-label learning based on relief and information gain. In: Brazilian Symposium on Artificial Intelligence, pp. 72–81. Springer

Spolaor N, Gomiero EA, Brand OK, Lee HD (2013) A comparison of multi-label feature selection methods using the problem transformation approach. Electron Notes Theor Comput Sci 292:135–151

Spolaor N, Monard MC, Tsoumakas G, Lee HD (2016) A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing 180:3–15

Suping Xu, Yang X, Yunsong QL (2016) Multi-label learning with label-specific feature reduction. J Computer Appl 104:52–61

Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I (2008) Multi-label classification of music into emotions. Blood 90(9):3438–3443

Tsoumakas G, Spyromitros-Xioufis E, Vlachos V, Vlahavas I (2011) Multilabel classification of music into emotions. Blood 90(9):3438–3440

Wu XZ, Zhou ZH (2017) 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

Yu Y, Pedrycz W, Miao D (2013) Neighborhood rough sets based multi-label classification for automatic image annotation. Int J Approx Reason 54(9):1373–1387

Yu Y, Wang Y (2014) Feature selection for multi-label learning using mutual information and ga. In: International Conference on Rough Sets and Knowledge Technology, pp. 454–463. Springer

Zhang ML, Peña JM, Robles V (2009) Feature selection for multi-label learning. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1075–1080. ACM

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.