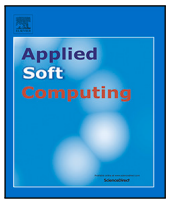
Applied Soft Computing Journal 175 (2025) 113028

Contents lists available at ScienceDirect

Applied Soft Computing

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



Feature selection via label enhancement and neighborhood rough set for multi-label data with unbalanced distribution

Wenbin Qian [a,b,*], Wenyong Ruan [a], Xiwen Lu [a], Wenji Yang [b], Jintao Huang [c]

[a] School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China
[b] School of Software, Jiangxi Agricultural University, Nanchang 330045, China
[c] Department of Computer Science, Hong Kong Baptist University, 999077, Hong Kong, China

ARTICLE INFO

Keywords:
Feature selection
Neighborhood rough set
Multi-label learning
Label enhancement
Label distribution

ABSTRACT

Multi-label learning has gained significant attention in classification tasks, but challenges remain in handling high-dimensional data. Although feature selection techniques can alleviate these issues, neglecting the unbalanced data distribution problem severely undermines the models' accuracy. Furthermore, existing methods fail to account for the importance and correlation of labels. In this paper, we present a novel multi-label feature selection algorithm that addresses these issues through three innovations: (1) using k-nearest neighbors to capture local similarities in unbalanced data, (2) enhancing labels by converting them into distributions to enrich semantic information, and (3) introducing a new evaluation function to assess label correlations. A multi-criteria strategy is established to maximize feature-label relevance, minimize redundancy, and strengthen label correlations. Experimental results on fifteen multi-label datasets demonstrate the algorithm's superiority over five state-of-the-art methods.

1. Introduction

Recently, multi-label learning [1] has attracted increasing interest, leading to wide applications in text categorization, video annotation, gene function, and other fields. However, similar to single-label data, some multi-label data may be unevenly distributed [2], and homogeneous sample spaces rarely exist. As an example of data with unbalanced distributed sample space (Fig. 1), samples may be sparsely or densely distributed in the sample space. There are two different types of samples: class 1 and class 2, which are marked as red stars and blue dots, respectively. The blue circles represent the δ-neighborhood, and the black circles indicate the k-nearest neighbors of the sample. Under this unevenly distributed sample space, local similarity samples cannot be well obtained through the traditional neighborhood or kNN theory and may weaken the mining of local sample information. Furthermore, the information from label supervision obtained by local similar instances is also incomplete. Moreover, in many practical applications, it is always the case that numerous irrelevant and redundant features are stored in multi-label data, which usually leads to high computational complexity and low multi-label classification performance. Therefore, how to reduce redundant or irrelevant features from high-dimensional multi-label data to avoid curse of dimensionality and improve classification performance is an urgent problem to resolve.

According to the study and exploration of researchers, many multi-label feature selection methods have been developed in recent years. For instance, Qu et al. introduced information gain for performing preliminary dimensionality reduction on datasets with high dimensionality [3]. M.A.N.D.Sewwandi et al. designed a novel class-specific feature selection algorithm using k-nearest neighbor and neighborhood rough set theories [4]. The research mentioned above typically regards each class label as having equal significance. However, the semantic information carried by different labels varies. This limitation may lead to a weakening of the label-supervised information. To deal with this issue, Geng et al. introduced some label enhancement methods to convert labels in logical form into a label distribution form to obtain more semantic information [5,6]. Yin et al. developed a method via label enhancement and β-precision for feature selection [7]. Liu et al. designed a novel label distribution feature selection algorithm using instance information distribution [8]. Although these studies have demonstrated the effectiveness of exploring inter-sample information for label enhancement to enrich multi-label information, multi-label datasets are often unevenly distributed. This unbalanced distribution will seriously affect the exploration of similar samples and then the acquisition of local labeling supervision information. To tackle the problem, the k-nearest neighborhood is used to explore local similarity samples based on the advantages of kNN and δ-neighborhood. Then,

* Corresponding author at: School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang 330045, China.
E-mail address: qianwenbin1027@126.com (W. Qian).

https://doi.org/10.1016/j.asoc.2025.113028
Received 22 April 2024; Received in revised form 17 December 2024; Accepted 14 March 2025
Available online 27 March 2025

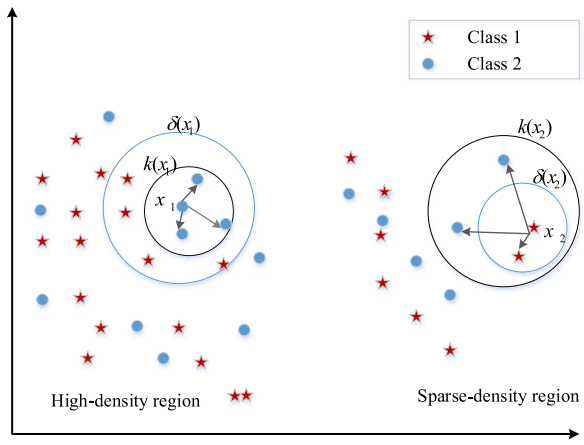


Fig. 1. Illustration of sample neighborhood under high-density and sparse-density regions.

a k-nearest neighborhood-based label enhancement method is defined as converting the labels in logical form into a label distribution form. Additionally, the relationships between features and between features and labels are often considered in feature selection methods. Since mutual information is a filter method based on feature selection criteria and has been used in many articles, it is preferable to use different descriptions to express this problem.

Specifically, this paper transforms multi-label learning into label distribution learning through label enhancement. In particular, sample similarity and label correlation are analyzed to convert label-supervised information from the logical form into a label distribution form. Then, based on the label distribution decision system, the relationship between different labels is approximated by the degree of overlap between their corresponding membership functions. Finally, a multi-criteria strategy is proposed based on maximum relevance, minimum redundancy, and label correlation metrics. In summary, the main contributions of this paper are summarized as follows.

- To tackle the challenges posed by unbalanced distribution in multi-label data and the heavy reliance of existing neighborhood models on sample data distribution, the k-nearest neighborhood is used to explore local similarity samples via neighborhood granularity. Then, the local similarity samples set can be obtained, and more sample information can be explored.
- A k-nearest neighborhood-based label enhancement algorithm using local similarity samples and label correlation is proposed, which can recover the original logical label into label distribution. Meanwhile, the local prior probability and centroid distance measurement are proposed to judge the possibility of unrelated labels being enhanced.
- To study the uncertainty measures of enhanced multi-label data, neighborhood entropy combined with neighborhood credibility degree is further explored. Furthermore, label correlation is characterized by rough approximations. Finally, a novel feature evaluation criterion is presented using a multi-criteria strategy.
- Extensive experiments are carried out to analyze and compare the performance of the proposed method with five multi-label feature selection methods on fifteen publicly available multi-label datasets. The effectiveness of the proposed method is verified by experimental results on six multi-label evaluation metrics.

The remainder of this paper is organized as follows. Section 2 introduces the related work on feature selection for multi-label learning and label distribution learning. Some basic concepts about multi-label learning, label distribution learning, and neighborhood rough set are reviewed in Section 3. In Section 4, a k-nearest neighborhood-based label enhancement method is presented by associating label correlation and sample similarity. Then, a feature selection algorithm using the multi-criteria measurement strategy is proposed. Section 5 reports the experimental results. Lastly, we summarize the main conclusions of this paper in Section 6 and discuss future research.

2. Related work

2.1. Multi-label feature selection

Over the last few years, although multi-label learning has been successfully applied in a variety of research fields, one challenge with existing research is the difficulty in processing high-dimensional feature spaces for multi-label learning. To overcome this challenge, feature selection is a significant preprocessing step that can reduce feature dimensions and improve classification performance [9]. At present, many feature selection methods have been studied using various feature evaluation criteria, such as dependency degree [10,11], distance metrics [12,13], and mutual information [14,15].

Mutual information, one of the most commonly used evaluation criteria for measuring the correlation between variables, has been extensively utilized in multi-label feature selection [16,17]. Sun et al. [18] proposed a method based on neighborhood mutual information and ML-ReliefF. Qian et al. [19] designed a feature selection method using label distribution and mutual information for multi-label learning. Furthermore, Gao et al. [20] introduced a feature redundancy term in the relevancy between the class, given that each already-selected feature and candidate feature are considered. In [21,22], fuzzy mutual information is generalized to adapt to multi-label learning, and then the fuzzy neighborhood mutual information-based method is utilized to measure the performance of the features. Additionally, multi-label learning with streaming label problems using mutual information has also been a concern. Liu et al. [23] introduced the circumstances of streaming labels and proposed the corresponding feature selection method.

In recent years, label correlation-based methods have been concerned with multi-label feature selection. For example, Che et al. [24] investigated a local attribute reduction with fuzzy rough sets by exploring label correlations from local and global viewpoints. Fan et al. [25] established a method via feature redundancy and label correlations. Dai et al. [26] introduced a modified redundancy-removal feature selection method to remove the redundant features. Besides, some label correlation-based methods have also been studied in online feature selection and missing label cases. For instance, You et al. [27] introduced a feature selection method to handle streaming features, where the weight of each label is obtained by considering the correlation between labels. What is more, a novel joint learning framework, which includes multi-label classification, feature selection, and label correlations, was introduced by He et al. [28]. In summary, label correlations have received much attention from many researchers [29,30].

Furthermore, it should be noted that the above methods not only show that mutual information is a useful feature evaluation criterion but also indicate that taking label correlation into consideration in feature selection can obtain superior results. Motivated by this finding, we try to consider the feature dependency, feature redundancy, and label correlation simultaneously. Meanwhile, a feature selection method via a multi-criteria strategy is explored, which is also the focus of this study.

2.2. Label distribution learning

Label distribution learning is proposed as a framework for dealing with label ambiguity problems by taking more label information into consideration, which is regarded as an extension of multi-label learning. Based on such advantages, label distribution learning has been extensively explored by many researchers in various studies in

recent years [31,32]. Jia et al. [33] designed a label distribution learning algorithm by exploiting the label ranking relation, which implies more valuable semantic information. Zhang et al. [34] introduced a method based on learning non-negative components. In addition, label distribution learning is also applied in facial expression recognition, which is a hot topic in practical applications [35]. Feature selection has received extensive attention in label distribution learning. For instance, Deng et al. [36] used a dual similarity measure based on neighborhood fuzzy entropy to solve the label distribution feature selection problem. In [37], Qian et al. presented a label distribution-based feature selection method, which explores feature similarity using neighborhood granularity and generates label correlations through a correlation coefficient. To address the problem of the number of possible label sets that may dynamically increase [38], the correlations between different labels were exploited.

According to the above studies, it is worth noting that increasing attention is being attracted to label distribution learning because it can solve the more common case of label ambiguity and obtain superior performance via the label distribution learning method. However, in the real world, most instances only involve logical values rather than label distribution values due to the significant challenge of obtaining label distribution values directly. To solve this issue, label enhancement is proposed, which enhances labels in logical form into a label distribution form [39,40]. Recently, label enhancement has been widely studied by researchers, and it can provide more supervised information about labels. Meanwhile, multi-label feature selection problems can be more effectively solved by mining more label information.

For the aforementioned descriptions, it would be beneficial for us to study label distribution learning and design a label enhancement method for recovering label distributions from logical labels to obtain more label-supervised information. Therefore, this paper attempts to explicitly express label ambiguity in raw multi-label data using label distribution.

3. Preliminaries

3.1. Multi-label learning and label distribution learning

Let $X = \{x_1, x_2, \ldots, x_n\} \in R^{n\times d}$ represent the sample space, where $x_i = \{x_i^1, x_i^2, \ldots, x_i^d\}$ represents the ith sample, n denotes the number of samples, and d denotes the dimension of the features. Let $Y = \{y_1, y_2, \ldots, y_n\} \in R^{n\times q}$ denote the label space, where $y_i = \{y_i^1, y_i^2, \ldots, y_i^q\}$ is the corresponding label set associated with sample x_i. Let $L = \{l_1, l_2, \ldots, l_q\}$ represent the label set, where q is the cardinality of the labels. It is worth noting that $y_i^j = 1$ if the jth class label is allocated to sample x_i, and $y_i^j = 0$ otherwise. The degree of each sample's membership in the corresponding label is expressed by the logical value (0 or 1).

For multi-label data, the relative significance of each label is different and cannot be simply expressed as 0 or 1. Therefore, label distribution learning is introduced, which points out that each sample should be expressed as the degree of association with the corresponding labels rather than just 0 and 1. A more detailed description is provided as follows.

In label distribution learning, $X = R^{n\times|L|}$ and $L = \{l_1, l_2, \ldots, l_q\}$ describe the sample space and the label set, respectively. Assume $LD = \{d_1, d_2, \ldots, d_n\}$ denotes the label distribution set; then, the label description degree of each sample x_i is expressed as $d_{x_i} = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \ldots, d_{x_i}^{y_q}\}$. Let $d_{x_i}^{y_j}$ denote the label significance of the jth class label l_j for the ith sample x_i. Additionally, the corresponding description degrees $d_{x_i}^{y_j} (1 \le j \le q)$ should satisfy the following two conditions: (1) $d_{x_i}^{y_j} \ge 0$, and (2) $\sum_{j=1}^{q} d_{x_i}^{y_j} = 1$.

3.2. Neighborhood rough set

To introduce a neighborhood rough set, some basic concepts and notations are defined as follows. Let $NDS = \langle U, C, D, f, \triangle, \delta\rangle$ represent the neighborhood decision system, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty sample set; $C = \{c_1, c_2, \ldots, c_m\}$ represents the set of conditional attributes, and $D = \{d_1, d_2, \ldots, d_q\}$ denotes the set of decision attributes; Let $V = \cup_{a\in\{C\cup D\}} V_a$, where the value set of attribute a is V_a; $f : U \times \{C \cup D\} \to V$ denotes a mapping function; $\triangle$ represents the distance function; and δ represents a neighborhood radius with $0 \le \delta \le 1$. For convenience, $NDS = \langle U, C, D, \delta\rangle$ is the neighborhood decision system in this article. For any feature subset $B \subseteq C$ and $x \in U$, the neighborhood relation can be described as [41]

$$NR_\delta(B) = \{(x, y) \in U \times U \mid \triangle(x, y) \le \delta\}, \tag{3.1}$$

and the neighborhood class of sample x under the feature subset B is expressed as

$$\delta_B(x) = \{y | x, y \in U, \triangle(x, y) \le \delta\}, \tag{3.2}$$

where $\triangle(x, y) = \sqrt{\sum_{t=1}^{|B|}(f(x, c_t) - f(y, c_t))^2}$ denotes the distance function, and $\delta_B(x)$ is the neighborhood information granularity produced by x. Generally speaking, the neighborhood granularity produced by x relies on the size of the threshold δ to a certain extent. For a conditional feature set $C = \{c_1, c_2, \ldots, c_m\}$, with $c_t \in C$, the threshold is defined as [42]

$$\delta = \frac{1}{\omega} Avg\left[\sum_{t=1}^{m} \frac{\sigma(c_t)}{\omega}\right], \tag{3.3}$$

where ω is a user-defined parameter, and $\sigma(c_t)$ denotes the standard deviation of the tth column feature. Then, the mean of $\sigma(c_t)$ for the entire conditional feature set can be denoted by Avg.

Let $NDS = \langle U, C, D, \delta\rangle$ represent the neighborhood decision system. With $B \subseteq C$ and any $x_i \in X \subseteq U$, the neighborhood lower approximation set $\underline{N_B}(X)$ and the neighborhood upper approximation set $\overline{N_B}(X)$ of X under feature subset B can be expressed, respectively, [41], as

$$\underline{N_B}(X) = \{x_i | \delta_B(x_i) \subseteq X\}, \tag{3.4}$$

$$\overline{N_B}(X) = \{x_i | \delta_B(x_i) \cap X \ne \emptyset\}. \tag{3.5}$$

Furthermore, for any $B \subseteq C$, the neighborhood uncertainty of x_i can be expressed as [41]

$$NH_\delta^{x_i}(B) = -log\frac{|\delta_B(x_i)|}{|U|}. \tag{3.6}$$

4. The proposed approach

The whole framework is shown in Fig. 2. There are two main mechanisms in our proposed method: label enhancement and feature selection. First, the original multi-label data are inputted. Then, the logical sample similarity and label correlation are explored based on the proposed k-nearest neighborhood and label relevance. Based on a smoothness assumption that similar examples have the same label distribution and combining label correlation. The k-nearest neighborhood-based label enhancement method is proposed to enhance labels in logical form into a label distribution form to obtain richer label-supervised information. Second, in feature selection methods, the maximum relevance minimum redundancy criterion has been successfully applied. However, most existing research methods on multi-label feature selection may not take the label correlation into account well. Based on this, a feature selection method for enhanced multi-label data is designed using the multi-criteria measurement strategy, which not only considers the correlation between features and labels and the correlation among features but also takes the correlation among labels into account. Finally, a ranked feature subset is outputted. This section will provide more details of the proposed approach.

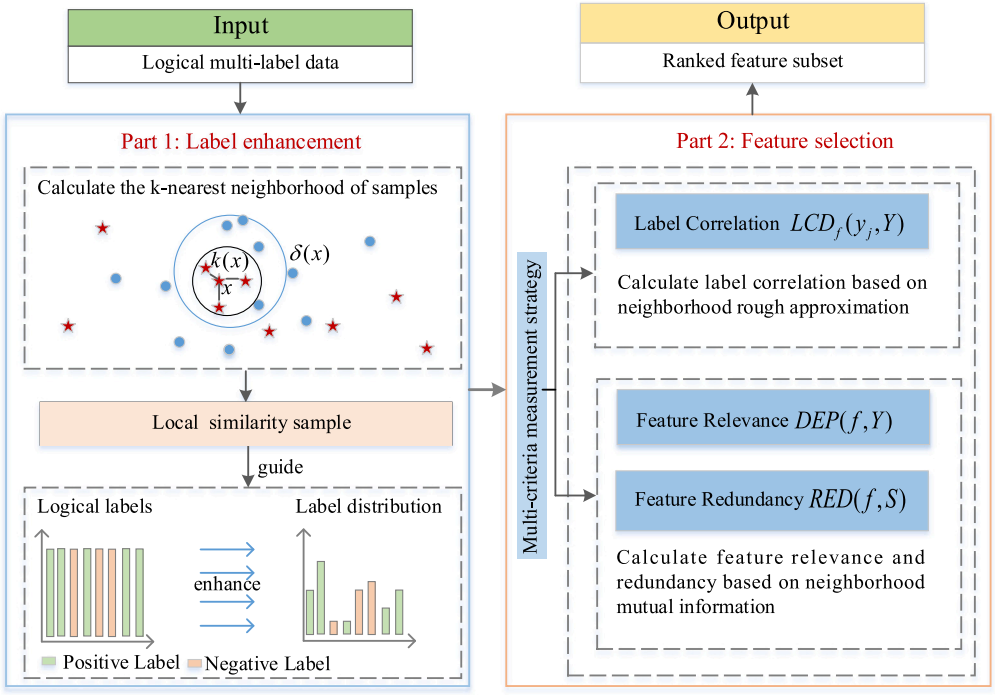


Fig. 2. The whole framework of the proposed method.

4.1. Label enhancement for multi-label data with unbalanced distribution

Based on the smoothness assumption, similar feature spaces can be viewed as having the same label distribution [42]. Therefore, the key idea is to explore the similarities between samples. However, for multi-label datasets, most data are characterized by uneven distribution, with either dense or sparse distributions. Finding a suitable set of similar examples is difficult based on traditional neighborhood granularity or kNN theory. To solve this problem, the k-nearest neighborhood is used to handle unevenly distributed data. Thus, a label enhancement method is presented to acquire the label description degree of multi-label data via k-nearest neighborhood-based local sample similarity and label relevance. The specific details are introduced as follows.

Definition 1. Given a neighborhood decision system $NDS = \langle U, C, D, \delta\rangle$ with $L \subseteq D$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a sample set, $C = \{c_1, c_2, \ldots, c_m\}$ represents a set of conditional attributes, and $L = \{l_1, l_2, \ldots, l_q\}$ is a set of decision attributes . δ is a neighborhood parameter such that $0 \le \delta \le 1$. For convenience, let KN_B^δ be a k-nearest neighborhood with $B \subseteq C$; k is a parameter that denotes taking the k samples closest to sample x_i. Then, the k-nearest neighborhood of $x_i \in U$ with respect to B is defined as

$$KN_B^\delta(x_i) = \{x_j \in U | x_j \in \delta_B(x_i) \cap K_B(x_i)\}, \tag{4.1}$$

where $K_B(x_i) = \{x_j \in U |\ \triangle_B(x_i, x_j) > \triangle_B(x_i, x_j^h)\}$ represents k-nearest-neighbor information granules. x_j^h denotes the hth sample in ascending order of distance values between x_i and other samples. Specifically, it means the k samples closest to x_i. $\delta_B(x_i) = \{x_j \in U |\ \triangle_B(x_i, x_j) \le \delta\}$ denotes the neighborhood granularity of sample x_i with respect to B.

For any sample x_i, the process begins by identifying its neighborhood $\delta_B(x_i)$ under the feature set B. Next, the closest k samples $K_B(x_i)$ within this neighborhood are determined, resulting in the k-nearest neighborhood $KN_B^\delta(x_i)$. This k-nearest neighborhood approach not only overcomes the limitations of traditional δ-neighborhoods and k-nearest neighbor methods for finding locally similar examples but is also better suited for handling uneven data distributions. In summary, a set of locally similar samples is generated through the construction of k-nearest neighborhood information granules.

Based on similar local samples, for any label $l_j \in L$ with respect to x_i, if $l_j = 1$, it can be viewed as the relevant label for sample x_i; if $l_j = 0$, it can be viewed as the irrelevant label for sample x_i. However, in most existing cases, even though these labels are marked as 0, they still contain implicit supervisory information. Specifically, the significance of the possible relevance of the label $l_j = 0$ to sample x_i can also be explored.

Definition 2. Given a neighborhood decision system $NDS = \langle U, C, D, \delta\rangle$ with $L \subseteq D$ and $L = \{l_1, l_2, \ldots, l_q\}$, for all $B \subseteq C$, the k-nearest neighborhood of $x_i \in U$ under feature subset B is expressed as $KN_B^\delta(x_i)$. For any $x_i \in U$, the corresponding label vector can be expressed as $y_i = \{y_{i1}, y_{i2}, \ldots, y_{iq}\} \subset \{0,1\}^q$. Then, the local prior probability of label l_j can be defined as

$$LP(x_i) = \begin{cases} LP_{x_i}^+ = \frac{|N_{l_j}^+|}{|N_{l_j}^+| + |N_{l_j}^-|} & if\ l_j = 1 \\ LP_{x_i}^- = \frac{|N_{l_j}^-|}{|N_{l_j}^+| + |N_{l_j}^-|} & else, \end{cases} \tag{4.2}$$

where $N_{l_j}^+ = \{x_i | x_i \in KN_B^\delta(x_i) \wedge y_{ij} = 1\}$ represents the cardinality of the local associated sample set of label l_j with respect to x_i and $N_{l_j}^- = \{x_i | x_i \in KN_B^\delta(x_i) \wedge y_{ij} = 0\}$. For simplicity, we use $LP_{x_i}^+$ and $LP_{x_i}^-$ to represent the local positive prior probability and the local negative prior probability with respect to x_i, respectively.

Based on the local prior probability LP of label l_j, one can infer that the possible related label $l_j = 0$ for sample x_i should be converted to a label distribution if $LP_{x_i}^+$ is larger than $LP_{x_i}^-$. In other words, for a sample x_i with respect to $l_j = 0$, there are more samples with the label marked as 1 rather than 0 in its k-nearest neighborhood.

To better determine whether the possible related labels should be enhanced, the distance of sample x_i to the class-positive center and the class-negative center should be calculated. Let $l_j^+ = \{x_i | x_i \in U \wedge y_{ij} = 1\}$

and $l_j^- = \{x_i | x_i \in U \wedge y_{ij} = 0\}$ represent the class-positive sample set of label l_j and the class-negative sample set of label l_j, respectively. Then, let $C_{l_j}^+$ and $C_{l_j}^-$ represent the class-positive center and class-negative center of label l_j, respectively.

Based on the definition above, the distances of sample x_i to the class-positive center and class-negative center are proposed to measure the relative distance, which can be expressed as follows:

$$\triangle(x_i, C_{l_j}^+) = \left(\sum_{t=1}^{|B|} |v(x_i, c_t) - v(C_{l_j}^+, c_t)|^p \right)^{\frac{1}{p}}, \tag{4.3}$$

$$\triangle(x_i, C_{l_j}^-) = \left(\sum_{t=1}^{|B|} |v(x_i, c_t) - v(C_{l_j}^-, c_t)|^p \right)^{\frac{1}{p}}, \tag{4.4}$$

where $p = 2$ in this paper and denotes the Euclidean distance. Euclidean distance is a simple and effective metric that has been widely applied in various domains. Therefore, it is adopted in this paper as the fundamental distance measure.

Similarly, the $r_{l_j}^+$ and $r_{l_j}^-$ are the radii of class-positive samples and class-negative samples for label l_j, respectively, which are used to represent the size of the class cluster. They are defined as follows:

$$r_{l_j}^+ = \frac{1}{|l_j^+|} \sum_{i=1}^{|l_j^+|} \triangle(x_i, C_{l_j}^+), \tag{4.5}$$

$$r_{l_j}^- = \frac{1}{|l_j^-|} \sum_{i=1}^{|l_j^-|} \triangle(x_i, C_{l_j}^-). \tag{4.6}$$

Combining the definition above, we introduce the difference value (DV) to assess whether the possible related label l_j should be enhanced. The DV is calculated as the difference between the distance from sample x_i to the class center and the radius of the class cluster. The detailed formula description for DV is given as follows:

$$DV_{x_i}^+ = |\triangle(x_i, C_{l_j}^+) - r_{l_j}^+|, \tag{4.7}$$

$$DV_{x_i}^- = |\triangle(x_i, C_{l_j}^-) - r_{l_j}^-|, \tag{4.8}$$

where $DV_{x_i}^+$ represents the difference value between the distance of sample x_i to the class-positive center and the radius of class-positive samples for label l_j, and $DV_{x_i}^-$ denotes the difference value between the distance of sample x_i to the class-negative center and the radius of class-negative samples for label l_j. Based on this, it can be inferred that the possible relevance of label $l_j = 0$ to sample x_i should be enhanced if $DV_{x_i}^+$ is smaller than $DV_{x_i}^-$. For simplicity, the possible related label l_j is closer to the related class-positive cluster and should be enhanced.

Based on the definitions above, we assess whether to enhance labels marked as 0 from both local and global perspectives. From the local perspective, we calculate the proportions of labels marked as 1 and 0 within each k-nearest neighborhood. From the global perspective, we measure the distances from an instance to the centers of positive and negative class instances, taking these factors into account to determine whether to enhance the labels marked as 0.

Label correlation is a key factor in obtaining better classification performance in multi-label learning. To preserve the original label structure of multi-label data after label enhancement, label correlation is considered when evaluating labels, and the degree of relevance between labels is calculated to explore the information about label significance.

Definition 3. Given a multi-label decision table, the label set is represented by D, $\forall L \subseteq D$ and $L = \{l_1, l_2, \ldots, l_q\}$. Then, the label relevance matrix $LM(LL)$ can be denoted as

$$LM(LL) = \begin{bmatrix} LR(l_1; l_1) & LR(l_1; l_2) & \ldots & LR(l_1; l_q) \\ LR(l_2; l_1) & LR(l_2; l_2) & \ldots & LR(l_2; l_q) \\ \vdots & \vdots & \vdots & \vdots \\ LR(l_q; l_1) & LR(l_q; l_2) & \ldots & LR(l_q; l_q) \end{bmatrix}, \tag{4.9}$$

where $LR(l_i, l_j) = \frac{|l_i \cap l_j|}{|l_i \cup l_j|}$ measures the relative relationship between the two labels l_i and l_j, LR can be measured by the Jaccard similarity coefficient.

From this definition, it can be inferred that for any label l_i, the label correlation between l_i and the entire label space L can be calculated by $LR(l_i; L) = \sum_{j=1}^{q} LR(l_i; l_j)$. Then, the label weight of each label $l_i \in L$ can be defined as $w(l_i) = \frac{\sum_{j=1}^{q} LR(l_i; l_j)}{\sum_{i=1}^{q} LR(l_i; L)}$.

Based on the above analysis, and combined with the local similarity of samples and label correlation, the label enhancement method is designed to convert the labels from a logical form into a label distribution. It is worth noting that all sample points in the k-nearest neighborhood can be considered as similar samples. According to the smoothness assumption and the probabilistic model, the detailed design of the proposed label enhancement method is as follows.

Definition 4. Given a neighborhood decision system $NDS = \langle U, C, D, \delta \rangle$ with $B \subseteq C$ and $L \subseteq D$, where $L = \{l_1, l_2, \ldots, l_q\}$. For $x_i \in U$, the label set $Y_i \subseteq D$ is defined as $Y_i = \{y_1, y_2, \ldots, y_q\}$. Then, the label distribution can be calculated via

$$d_{x_i}^{y_j} = \frac{count[\sum_{x_a \in KN_B^\delta(x_i)} (x_a^{y_j} = 1)]}{count[\sum_{x_a \in KN_B^\delta(x_i)} \sum_{k=1}^{q} (x_a^{y_k} = 1)]} . w(y_j), \quad \forall x_i \in U, y_j \in Y_i, \tag{4.10}$$

where $x_a \in KN_B^\delta(x_i)$ denotes the ath similar sample of sample point x_i, and $count[\cdot]$ denotes the number of labels with a value of 1 for sample point x_a. $x_a^{y_j}$ represents the value of label y_j for sample point x_a. Here, the corresponding description degree should meet the condition: if $\sum_{j=1}^{q} d_{x_i}^{y_j} < 1$, then we have $d_{x_i}^{y_j} = \frac{d_{x_i}^{y_j}}{\sum_{k=1}^{q} d_{x_i}^{y_k}}$.

Subsequently, the label distribution space $LD = \{d_1, d_2, \ldots, d_n\}$ can be obtained in the sample space $U = \{x_1, x_2, \ldots, x_n\}$. For each sample $x_i \in U$, its corresponding label distribution set is defined as $d_{x_i} = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \ldots, d_{x_i}^{y_q}\}$. Here, Definition 4 meets the following two constraint conditions: (1) non-negativity ($d_{x_i}^{y_j} \geq 0$) and (2) standardization ($\sum_{j=1}^{q} d_{x_i}^{y_j} = 1$).

Example 1. Given a neighborhood decision system $NDS = \langle U, C, D, \delta \rangle$ with $L \subseteq D$, $L = \{l_1, l_2, l_3, l_4\}$. Given an instance x_1 with labels $L(x_1) = \{1, 1, 0, 1\}$. Assuming that $KN_B^\delta(x_1) = \{x_1, x_6\}$ and $L(x_6) = \{0, 1, 1, 0\}$. On this basis, we can compute the label weight of each label: $w(l_1) = \frac{5}{9}$, $w(l_2) = \frac{1}{3}$, $w(l_3) = \frac{5}{9}$, $w(l_4) = \frac{5}{9}$. Subsequently, through Definition 4 and Algorithm 1, we can obtain $d_{x_1}^{y_j}$ for each label $d_{x_1}^{y_1} = \frac{1}{5} * \frac{5}{9} = \frac{1}{9}$, $d_{x_1}^{y_2} = \frac{2}{5} * \frac{1}{3} = \frac{2}{15}$, $d_{x_1}^{y_3} = 0 * \frac{5}{9} = 0$, $d_{x_1}^{y_4} = \frac{1}{5} * \frac{5}{9} = \frac{1}{9}$. However, $Sum(d_{x_1}^{Y}) = \frac{16}{45} \neq 1$, and we need to transform $d_{x_1}^{y_j}$ for each label. Therefore $d_{x_1}^{y_1} = \frac{1}{9} / \frac{16}{45} = \frac{5}{16}$, $d_{x_1}^{y_2} = \frac{2}{15} / \frac{16}{45} = \frac{3}{8}$, $d_{x_1}^{y_3} = 0$, $d_{x_1}^{y_4} = \frac{1}{9} / \frac{16}{45} = \frac{5}{16}$.

To describe the whole process of the algorithm intuitively, a detailed introduction to the label enhancement algorithm is provided, as shown in Algorithm 1.

The details of the algorithm are introduced as follows. In the first stage, the multi-label data are normalized. Then, in step 2, for each sample $x_i \in U$, its k-nearest neighborhood is calculated. Furthermore, the local prior probability LP and the difference value DV are calculated to measure whether the possible related labels should be enhanced in steps 3–4. Subsequently, LM is constructed to represent the correlations among labels in step 5. With the aid of the label relevance matrix, the label weight can be obtained in steps 6–8. Based on sample similarity and label weight, the description degree $d_{x_i}^{y_j}$ of the label for each sample is calculated in step 9. In the end, the label distribution space LD is obtained in step 10.

Compared to traditional multi-label data with logical labels, label distribution forms provide more supervisory information. Moreover,

Algorithm 1 K-nearest Neighborhood-based Label Enhancement algorithm combined label correlation(KNLE)

Input:MLDT=$\{U, C \cup D\}$: a multi-label decision table; neighborhood adjustment parameter k

Output:LDDT=$\{U, C \cup D, LD\}$: a label distribution decision table

Begin

1. Normalize the column data for each conditional feature of original features C;
2. Calculate the k-nearest neighborhood $KN_C^{\delta}(x_i)$ of sample x_i with respect to C ;
3. Calculate local prior probability LP of labels for each k-nearest neighborhood ;
4. Calculate the difference value DV of unrelated labels for each sample by Definition 2;
5. Construct the label relevance matrix LM;
6. For each label $l_j \in D$:
7. Calculate label weight $w(l_j)$;
8. End for
9. Calculate the description degree $d_{x_i}^{y_j}$ of label for each sample by Eq.(4.10);
10. Obtain the label distribution space LD based on sample similarity and label weight;
11. Return a label distribution decision table $LDDT =< U, C \cup D, LD >$.

End

the label correlation is also worth exploring for label distribution vectors. Thus, a novel label correlation evaluation method for enhanced multi-label data needs to be developed.

4.2. Feature selection on enhanced multi-label data

Recently, multi-label feature selection has become a concern in multi-label learning. Thus, a number of feature selection methods have been proposed. However, the existing methods mainly regard the importance of labels related to each sample as being the same and ignore label ambiguity. Different from applying the 0/1 to model the relative importance of labels, label distribution data can explicitly describe label ambiguity. Moreover, the maximum relevance minimum redundancy is a typical method that can achieve good results in feature selection but ignores the correlation among labels. Most existing studies on feature selection methods use this criterion strategy, which only takes into account the correlation between features and the correlation between features and labels. To overcome this drawback, in Section 4.1, a label enhancement method is designed. Subsequently, inspired by the multi-criteria measurement strategy and considering the correlation among labels, this section presents a new feature selection method.

Drawing on the set-theoretic model of three-way decision, we try to divide the labels into three different sets: the related, the possible relevance, and the unrelated sets of class-samples via enhanced multi-label data. In traditional binary classification of multi-label data, the label partitions the sample set U into two decision classes $U/D = \{D_1, D_2\}$. To learn from the methods of classical rough set theory in the label distribution decision table, let $X_{(y)}$ denote a set of instances $X = \{x_1, x_2, \ldots, x_n\}$ for label y. Assuming that the label weight parameter h denotes the importance of the label y, which is significant when $d_{x_i}^{y} \geq h$, the $X_{(y)}$ only consists of x where $d_{x_i}^{y} \geq h$. The label weight parameter e represents the significance of the label y, which is unimportant when $d_{x_i}^{y} = e$. A more detailed introduction is provided as follows.

Definition 5. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD\rangle$ with $B \subseteq C$, for a sample point $x_i \in U$, the corresponding label set $Y_i \subseteq D$ is defined as $Y_i = \{y_1, y_2, \ldots, y_q\}$. Let $X_{(y_j)}$ represent the set of all samples x that belong to label y_j. The detailed definition of $X_{(y_j)}$ is given as

$$X_{(y_j)} = \begin{cases} \{x_i | \forall x_i \in U : d_{x_i}^{y_j} \geq h\} \\ \{x_i | \forall x_i \in U : e < d_{x_i}^{y_j} < h\}, \\ \{x_i | \forall x_i \in U : d_{x_i}^{y_j} = e\} \end{cases} \tag{4.11}$$

where $d_{x_i}^{y_j}$ is the label description degree, and each label description degree $d_{x_i}^{y_j} \in [0, 1]$. The parameter $h = \frac{\sum_{i=1}^{n} d_{x_i}^{y_j}}{n}$ represents the related label where $d_{x_i}^{y_j} \geq h$, $e = 0$ is a threshold parameter that indicates the unrelated label where $d_{x_i}^{y_j} = e$, and the label can be viewed as a possible relevant label when $e < d_{x_i}^{y_j} < h$. Based on this definition, the set of samples is classified into three categories according to the label space.

Based on the concept of neighborhood, we define neighborhood granules on the universe $\{\delta_B(x_i) \cup x_i | x_i \in U, i = 1, 2, \ldots, n\}$, where $\cup_{i=1}^{n}(\delta_B(x_i) \cup x_i) = U$. Then, from the conceptual point of view of rough set theory, the lower and upper approximations can be defined as follows.

Definition 6. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD\rangle$, $\forall B \subseteq C$, $Y \subseteq D$ and $Y = \{y_1, y_2, \ldots, y_q\}$, given any $X_{y_j} \subseteq U$ as subsets of the sample with classes: the related, the possibly related and unrelated, and a family of δ neighborhood granules $\{\delta_B(x_i) \cup x_i | x_i \in U\}$. Then, the upper approximation $\overline{R_B}(y_j)(x_i)$ and the lower approximation $\underline{R_B}(y_j)(x_i)$ of label y_j under attributes B are defined, respectively, as

$$\overline{R_B}(y_j)(x_i) = \{x_i | x_i \in U, \{\delta_B(x_i) \cup x_i\} \cap X_{(y_j)} \neq \emptyset\}, \tag{4.12}$$

$$\underline{R_B}(y_j)(x_i) = \{x_i | x_i \in U, \{\delta_B(x_i) \cup x_i\} \subseteq X_{(y_j)}\}. \tag{4.13}$$

Then, the positive region $POS_B(Y)$ and the dependency function $\gamma_B(Y)$ of Y related to attributes B can be defined, respectively, as

$$POS_B(Y) = U_{j=1}^{q} \underline{R_B}(y_j), \tag{4.14}$$

$$\gamma_B(Y) = \frac{\sum_{x_i \in U} POS_B(Y)(x_i)}{|U|}. \tag{4.15}$$

According to the definitions of rough approximations, let $Y = \{y_1, y_2, \ldots, y_q\}$ and $B \subseteq C$. The dependency function of label y related to B can be used as a criterion for the membership information of samples in positive regions. It is easy to observe that it can employ membership information to analyze the correlation among labels. In the following, a new evaluation metric function will be defined: the label correlation degree function, which reflects the correlation among labels based on the overlap degree of rough approximations.

Definition 7. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD\rangle$, $\forall B \subseteq C$, $Y = \{y_1, y_2, \ldots, y_q\} \subseteq D$. The label correlation dependency of y_j and y_k related to B is defined as

$$LCD_B(y_j; y_k) = \frac{1}{|U|} \sum_{x_i \in U} \frac{|\underline{R_B}(y_j)(x_i) \cap \underline{R_B}(y_k)(x_i)|}{|U|}, \tag{4.16}$$

where $|\underline{R_B}(y_j)(x_i) \cap \underline{R_B}(y_k)(x_i)|$ is the cardinality of the intersection of two lower approximations. In other words, it shows the degree of overlap between different sets. The larger the overlap degree, the greater the label correlation dependency. Based on the observation that the overlap degree of different low approximations related to different labels is close to the correlation among labels, a new label correlation degree function can be constructed.

Furthermore, the label correlation dependency of the jth label with respect to the whole label space Y based on B can be written as

$$LCD_B(y_j; Y) = \sum_{k=1}^{q} LCD_B(y_j; y_k). \tag{4.17}$$

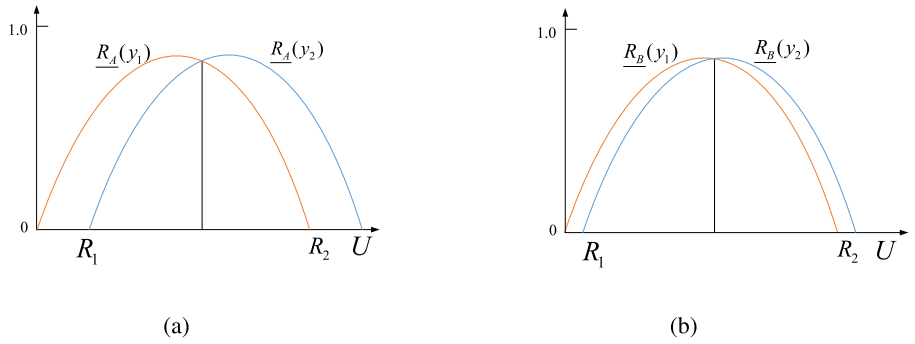


Fig. 3. Membership curves in two different situations. (a) Membership curves of two labels in feature subspace A. (b) Membership curves of two labels in feature subspace B.

In Fig. 3(a) and (b), $\underline{R_A}(y_1)$ ($\underline{R_B}(y_1)$) and $\underline{R_A}(y_2)$ ($\underline{R_B}(y_2)$) represent the membership curves of label y_1 and label y_2 in the feature subspace $A(B)$, respectively. We assume that $\underline{R_A}(y_1) \cup \underline{R_A}(y_2) = \underline{R_B}(y_1) \cup \underline{R_B}(y_2)$; it is not difficult to infer that the positive regions in the feature subspaces A and B are equal, and both feature subspaces possess the same classification abilities based on classical rough set theory. However, we can easily observe that the overlap degree of two labels in feature subspaces $A(B)$ is different, which raises the question: will the overlap degree influence the corresponding result of feature selection? Based on the above analysis and findings, we aim to define a new label correlation dependency function.

To the best of our knowledge, the redundancy between features and correlations between labels and features plays a crucial role in constructing evaluation metrics for feature selection. Therefore, incorporating label correlation into a multi-criteria measurement strategy is still worth further exploration. To this end, we have proposed a series of uncertainty measures to evaluate the candidate features by combining algebraic and informational perspectives.

Definition 8. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD \rangle$, for all $B \subseteq C$, $Y = \{y_1, y_2, \ldots, y_q\} \subseteq D$, where $\delta_B(x_i)$ represents a neighborhood class of sample $x_i \in U$, and $X_{(y_j)}$ is the set of all samples x that belong to label y_j. A neighborhood credibility degree $NC_i^B(Y)$ for sample $x_i \in U$ related to Y is defined as

$$NC_i^B(Y) = \sum_{j=1}^{q} \frac{|\delta_B(x_i) \cap X_{(y_j)}|}{|\delta_B(x_i)|}, \tag{4.18}$$

where $NC_B(Y)$ is used to describe the classification accuracy of Y relative to B.

Subsequently, the neighborhood entropy of x_i related to B is defined as follows:

$$NH^{\delta}(B) = -\frac{\sum_{i=1}^{|U|} NC_i^B(Y)}{|U|} \sum_{i=1}^{|U|} log \frac{|\delta_B(x_i)|}{|U|}. \tag{4.19}$$

In the neighborhood entropy $NH^{\delta}(B)$, $\sum_{i=1}^{|U|} NC_i^B(Y)$ is the neighborhood credibility degree of Y related to B based on an algebraic perspective, and $-\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|\delta_B(x_i)|}{|U|}$ represents the neighborhood entropy of B based on an informational perspective. Furthermore, the neighborhood joint entropy $NH^{\delta}(B,Y)$ and neighborhood conditional entropy $NH^{\delta}(Y|B)$ can be further constructed, which are described as follows:

$$NH^{\delta}(B,Y) = -\frac{\sum_{i=1}^{|U|} NC_i^B(D)}{|U|} \sum_{i=1}^{|U|} \sum_{j=1}^{q} log \frac{|\delta_B(x_i) \cap X_{(y_j)}|}{|U|}, \tag{4.20}$$

$$NH^{\delta}(Y|B) = -\frac{\sum_{i=1}^{|U|} NC_i^B(D)}{|U|} \sum_{i=1}^{|U|} \sum_{j=1}^{q} log \frac{|\delta_B(x_i) \cap X_{(y_j)}|}{|\delta_B(x_i)|}, \tag{4.21}$$

where $\delta_B(x_i)$ and $X_{(y_j)}$ represent the neighborhood granules relative to the feature subset B and the similar distribution set related to label y_j, respectively. $\delta_B(x_i) \cap X_{(y_j)}$ represents the intersection between $\delta_B(x_i)$ and $X_{(y_j)}$, where $\cap$ denotes the intersection operator.

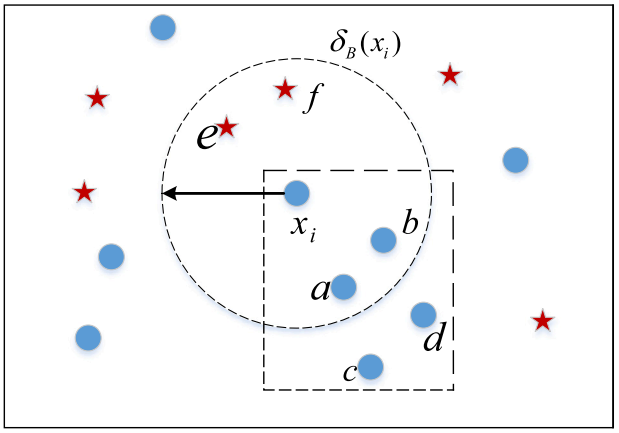


Fig. 4. The information granule of sample x_i in two-dimensional space.

Example 2. Fig. 4 provides an intuitive explanation of $\delta_B(x_i)$ in two-dimensional space. The distribution of samples in the decision class is often adjacent to the feature space. Therefore, samples c and d in Fig. 4, which are in the decision class of sample x_i, have the possibility of being included in the similar class of sample x_i. It is better to include samples a, b, c, and d in $\delta_B(x_i)$ than to include samples a and b. However, traditional information measures pay more attention to the similar class of sample x_i, thus ignoring the valid samples in the neighborhood information granules. Obviously, the joint information entropy measures proposed from both an algebraic perspective and an informational perspective can overcome the shortcomings of traditional information measures.

Definition 9. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD \rangle$, for $\forall B \subseteq C$, $Y = \{y_1, y_2, \ldots, y_q\} \subseteq D$, the neighborhood mutual information $NMI(B; y_j)$ is defined to measure the corresponding shared information between candidate feature subset B and jth label. The formula for this can be described as follows:

$$NMI(B; y_j) = -\frac{1}{|U|} \sum_{i=1}^{|U|} log \frac{|\delta_B(x_i)|^{NC_i^B(Y)}.|X_{(y_j)}|}{|\delta_B(x_i) \cap X_{(y_j)}|^{NC_i^B(Y)}.|U|}. \tag{4.22}$$

Furthermore, due to the fact that different labels have different label significance in enhanced multi-label data, the corresponding NMI is also influenced to some extent by the varying importance of the labels. Therefore, we take the significance of the labels into consideration. The

neighborhood mutual information between feature subset B and the whole label space Y can be computed as follows:

$$NMI(B;Y) = \sum_{j=1}^{q} sig(y_j) * NMI(B;y_j), \tag{4.23}$$

where $sig(y_j) = \frac{\sum_{i=1}^{n} d_{x_i}^{y_j}}{n}$ is the significance of the jth label in the enhanced multi-label data, which satisfies the condition $\sum_{j=1}^{q} sig(y_j) = 1$.

During the multi-criteria measurement strategy, redundancy and dependency are considered. For the enhanced multi-label data, we exploit mutual information and consider label significance to measure the dependency between feature f and the label space Y. Nevertheless, feature redundancy may also exist. In contrast to most existing methods for calculating label correlation, the correlation among labels is defined via the overlap degree of rough approximations.

Definition 10. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD\rangle$ and a candidate feature f, with $\forall B \subseteq C$ and $Y = \{y_1, y_2, \ldots, y_q\} \subseteq D$, the dependency between the candidate feature f and label set Y is defined as follows:

$$DEP(f,Y) = \sum_{y_j \in Y} sig(y_j) * NMI(f;y_j), \tag{4.24}$$

where $NMI(f;y_j)$ represents the neighborhood mutual information between candidate f and label y_j.

In addition, suppose $S = \{f_1, f_2 \ldots, f_k\}$ is the set of selected features. Then, the redundancy between candidate feature f and the selected feature S can be denoted as follows:

$$RED(f,S) = \frac{1}{k} \sum_{f_s \in S} NMI(f;f_s), \tag{4.25}$$

where k is the size of the selected feature subset S. The neighborhood mutual information between the candidate feature f and the selected feature f_s can be expressed as $NMI(f;f_s) = NH^{\delta}(f) - NH^{\delta}(f|f_s)$.

Definition 10 is used to measure a feature f that is highly associated with the label set Y and weakly related to the selected feature subset S. $DEP(f,Y)$ is utilized to measure the feature dependency, while $RED(f,S)$ denotes the feature redundancy. However, during feature selection, it is necessary to explore the correlation among labels. This can be expressed as follows.

Definition 11. Given a label distribution decision table $LDDT = \langle U, C \cup D, LD\rangle$ with $\forall B \subseteq C$ and $Y = \{y_1, y_2, \ldots, y_q\} \subseteq D$, the label correlation degree can be defined as follows:

$$LCD_f(y_j,Y) = \sum_{k=1}^{q} LCD_f(y_j;y_k), \tag{4.26}$$

where $\sum_{k=1}^{q} LCD_f(y_j;y_k)$ represents the label correlation degree of the jth label to the whole label space Y with respect to f.

Definition 11 can be used to measure the correlation between labels with respect to the candidate feature f. In the feature selection process, it is necessary to explore useful information from both the feature space and the label space, considering feature dependency, feature redundancy, and label correlation simultaneously. Therefore, we introduce a scoring function for feature selection as follows.

Definition 12. Given a candidate feature f, the label set Y, and the selected feature subset S, the score function for candidate feature f is formulated as follows:

$$J(DEP,RED,LCD) = \quad DEP(f,Y) - RED(f,S) + LCD_f(y_j,Y). \tag{4.27}$$

Based on the definition above, the score function is constructed using feature dependency, feature redundancy, and label correlation. To balance the weight between feature correlation and label correlation, a threshold λ is introduced to combine the information through a linear combination. Then, the score function $J(DEP,RED,LCD)$ can be reconstructed to measure the significance of the candidate feature f, which is defined as follows:

$$\begin{aligned} J(DEP,RED,LCD) &= \lambda * (DEP(f,Y) - RED(f,S)) \\ &\quad + (1-\lambda) * LCD_f(y_j,Y) \\ &= \lambda * \Big(\sum_{y_j \in Y} sig(y_j) * NMI(f;y_j) \\ &\quad - \tfrac{1}{k} \sum_{f_s \in S} NMI(f;f_s)\Big) \\ &\quad + (1-\lambda) * \sum_{j=1}^{q} LCD_f(y_j;y_k). \end{aligned} \tag{4.28}$$

Consequently, via Eq. (4.24), the first feature can be selected and added to an empty feature set S. Then, the importance of each candidate feature $f \in C-S$ can be measured using Eq. (4.29), in which each candidate feature satisfies the following condition:

$$f = arg\,max_{f \in C-S}\{J(DEP,RED,LCD)\}. \tag{4.29}$$

To describe the entire process of the algorithm intuitively, we provide a detailed introduction to the FSMRL algorithm, as shown in Algorithm 2.

Algorithm 2 Feature Selection algorithm using Multi-label neighborhood Rough set based on Label distribution (FSMRL)

Input:MLDT=$\{U, C \cup D\}$: a multi-label decision table, threshold λ
Output:S: the ranking set of selected features
Begin
1. Initialize the selected feature subset $S \leftarrow \{\emptyset\}$;
2. Construct the label distribution decision table $LDDT = \{U, C \cup D, LD\}$based on $MLDT$ by Algorithm 1 ;
3. For each $f \in C$:
4. Calculate the label dependency $DEP(f,Y)$ based on Eq.(4.24);
5. Calculate the feature redundancy $RED(f,S)$ based on Eq.(4.25);
6. Calculate the label correlation $LCD_f(y_j,Y)$ based on Eq.(4.27);
7. End for
8. **Repeated**
9. Select the feature $f \in C$ which satisfies Eq.(4.29);
10. Set $S = S \cup \{f\}$, $C = C - \{f\}$;
11. **Until** $|S| = |C|$;
12. Return the ranking set of selected features S.
End

Based on Algorithm 2, it is possible to understand the complete procedure of the feature selection algorithm using a multi-criteria measurement strategy based on enhanced multi-label data. The algorithm can be divided into the following three main steps. We initialize the selected feature subset and enhance the relevant labels for each sample in the original multi-label dataset into label distribution forms according to the algorithm in the first stage (steps 1–2). In the second phase, the label dependency, the feature redundancy, and label correlation are calculated (steps 3–7). In the final step (steps 8–11), the candidate feature with maximum feature dependency, minimum feature redundancy, and maximum label correlation is selected via a heuristic search process.

5. Experiments

In this section, to demonstrate the superiority of the proposed method, we compare the proposed algorithm with five feature selection algorithms on fifteen multi-label datasets. All the experimental results of these datasets are uniformly evaluated according to six evaluation metrics. The details of the experiments are presented as follows: First, Section 5.1 introduces some basic characteristics of the fifteen multi-label datasets, and the experimental settings are described in Section 5.2. Then, Section 5.3 analyzes the experimental results based

Table 1
Descriptions of experimental datasets.

Datasets	Samples	Features	Labels	LC	LD	Domains
Flags	194	19	7	3.392	0.485	Image
Virus	207	440	6	2.217	0.203	Biology
Emotions	593	72	6	1.869	0.311	Music
Plant	978	440	12	1.079	0.09	Biology
Birds	645	260	19	1.014	0.053	Audio
Yeast	2417	103	14	1.185	0.085	Biology
Gpositive	519	440	4	1.008	0.252	Biology
Guardian	352	1000	6	1.126	0.188	Text
BBC	302	1000	6	1.126	0.188	Text
Eukaryote	7766	440	22	1.146	0.052	Biology
Image	2000	294	5	1.236	0.247	Image
Human	3106	440	14	1.185	0.085	Biology
3sources	294	1000	6	1.126	0.188	Text
Stackex	9270	635	274	2.556	0.009	Text
Yelp	10 806	671	5	1.638	0.328	Text

on the experimental settings. After that, the stability of the proposed algorithm is assessed using statistical tests in Section 5.4. Furthermore, to study the corresponding impact of label correlation on FSMRL, an ablation study is conducted in Section 5.5, and the computational complexity is described in Section 5.6. Lastly, the sensitivity of the parameters is analyzed in Section 5.7.

5.1. Datasets

Experiments are conducted on fifteen multi-label datasets selected from the Mulan Library[1] and MLL Resources,[2] namely Flags, Virus, Emotions, Plant, Birds, Yeast, Gpositive, Guardian, BBC, Eukaryote, Image, Human, 3sources (reuters1000), Stackex (cs), Yelp, respectively. It is worth noting that these datasets are either sparse or dense due to their feature distribution characteristics. The datasets can be divided into five domains: image, biology, music, audio, and text. Table 1 displays more specific details and useful descriptions of these datasets, such as name (Datasets), number of samples (Samples), number of labels (Labels), label cardinality (LC), label density (LD), and application domains (Domains).

5.2. Experiment design and settings

In this study, we compare FSMRL with several other outstanding multi-label feature selection methods, including SSFS [43], PMFS [44], MDFS [45], MIFS [44], and PMU [46]. The details of all the algorithms are as follows:

- SSFS: It considers the influence of potential feature structure on label correlation in the process of feature selection. As suggested in a previous study, the corresponding parameters α, β, and γ are searched in $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$.
- PMFS: It models the multi-label feature selection problem as a two-objective optimization problem focused on the relevance and redundancy of features. According to previous research, the parameter λ is set to 10.
- MDFS : It uses local and global label correlation, and popular regularization discriminant feature selection is carried out for multi-label learning. The parameters α, β, and γ are set as 1, 1, and 100, respectively; the parameter k is set to $q - 1$, where q denotes the size of the label set.
- MIFS: It proposes a multi-label informed feature selection framework that exploits label correlations to select discriminative features across multiple labels. The parameters α, β and γ of MIFS are searched in $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 10\}$.
- PMU : It proposes a multi-label classification feature selection method based on mutual information applied to several multi-label classification problems. As suggested in previous research, the equal-width strategy is utilized to discretize the continuous features into four bins.

Several parameters need to be set in advance in our algorithm, including the neighborhood threshold ω, neighborhood adjustment parameter k, and balance parameter λ. We obtain the neighborhood threshold ω from 0.5 to 0.9 for each multi-label dataset with a step size of 0.02. The parameter k ranges from 0.01 N to 0.1 N in steps of 0.01 N, where N represents the size of each multi-label dataset. Here, k is set to 0.02 N for the later comparative experiment. In addition, we obtain the best balance parameter λ within the range of $[0, 1]$ with a step of 0.1. It is worth noting that, for fairness, MLKNN (k = 10) [47] is chosen as the basic multi-label classifier to assess the corresponding performance of all comparison methods. Furthermore, the 10-fold cross-validation strategy is used to maintain the consistency of all experimental results. In order to rationalize the comparison in the experiments, a reasonable feature ratio $r\%$ is adopted. As recommended in [48], the feature ratio $r\%$ is defined as follows:

- if $d < 100$, the top 40% of candidate features from the ranking set were selected.
- if $100 < d < 500$, the top 30% of candidate features from the ranking set were selected.
- if $500 < d < 1000$, the top 20% of candidate features from the ranking set were selected.
- if $d > 1000$, the top 10% from candidate features of the ranking set were selected.

where d indicates the dimensionality of the original features, and the top $r\%$ of candidate features from the ranking set are selected for experiments. Additionally, six multi-label evaluation metrics are adopted to assess the corresponding performance of algorithms from various perspectives, including instance-based evaluation metrics (Hamming Loss (HL), Coverage (CV), Average Precision (AP), Ranking Loss (RL), and One Error (OE)) and label-based evaluation metrics (Micro-F1 (F1)).

5.3. Experimental results

In this subsection, we compare the FSMRL algorithm with five existing multi-label feature selection algorithms. These experiments aim to evaluate predictive performance in terms of HL, CV, AP, RL, OE, and Micro-F1. The experiments are performed on fifteen multi-label datasets, which are selected from Table 1. Furthermore, a more detailed description of the experimental results is displayed in Tables 2–7. The best predictive performance result is marked in bold. For convenience, symbols "↓(↑)" represent "the smaller (larger) result being better". To count the corresponding number of results achieved by the FSMRL algorithm that is better than, equal to, or worse than those of other comparison algorithms on all comparative datasets. In addition, a row named "Rank" is added to each table to display the average performance across all multi-label datasets. Based on the experimental results, the corresponding analysis and summary can be drawn as follows.

By viewing the corresponding experimental results, it is easy to find that the FSMRL algorithm obtains a better result than all comparison algorithms. The average ranks are 1.40, 1.40, 1.67, 1.20, 1.27, and 1.60 for AP, CV, HL, OE, RL, and F1, respectively. Furthermore, the average performance of FSMRL on six evaluation metrics is better than that of the five comparison algorithms. We can analyze the experimental results as follows.

[1] http://mulan.sourceforge.net.
[2] http://www.uco.es/kdis/mllresources.

Table 2
Average precision(↑) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	0.8163 (2)	0.8019 (5)	0.8062 (3)	0.8036 (4)	0.7850 (6)	**0.8220 (1)**
Virus	0.6610 (4)	0.6691 (2)	0.6567 (5)	0.6471 (6)	0.6667 (3)	**0.6861 (1)**
Emotions	0.7854 (3)	0.7777 (5)	0.7799 (4)	0.7901 (2)	0.7706 (6)	**0.7959 (1)**
Plant	0.5393 (2)	0.5279 (3)	0.5260 (4)	0.5133 (6)	0.5174 (5)	**0.5429 (1)**
Birds	0.5850 (3)	0.5919 (2)	0.5586 (5)	0.5685 (4)	0.5523 (6)	**0.6160 (1)**
Yeast	0.7468 (3)	**0.7645 (1)**	0.7295 (6)	0.7454 (4)	0.7450 (5)	0.7544 (2)
Gpositive	0.8334 (2)	0.8217 (3)	0.8076 (4)	0.7818 (6)	0.7968 (5)	**0.8367 (1)**
Guardian	0.4764 (6)	0.4993 (4)	0.5772 (2)	0.4796 (5)	0.5001 (3)	**0.5791 (1)**
BBC	0.5061 (3)	0.5040 (4)	**0.5955 (1)**	0.4997 (6)	0.5010 (5)	0.5737 (2)
Eukaryote	0.5521 (2)	0.5268 (4)	0.5508 (3)	0.5160 (6)	0.5186 (5)	**0.5676 (1)**
Image	0.7649 (5)	**0.7897 (1)**	0.7586 (6)	0.7719 (3)	0.7693 (4)	0.7815 (2)
Huamn	0.5632 (2)	0.5355 (5)	0.5496 (3)	0.5369 (4)	0.5341 (6)	**0.5650 (1)**
3sources	0.4876 (5)	0.4789 (6)	0.5721 (2)	0.4894 (4)	0.4955 (3)	**0.5928 (1)**
Stackex	0.2381 (6)	0.2484 (2)	0.2382 (5)	0.2467 (3)	**0.2520 (1)**	0.2426 (4)
Yelp	0.8347 (2)	0.8314 (3)	0.8016 (6)	0.8206 (4)	0.8193 (5)	**0.8363 (1)**
Rank(↓)	3.33	3.33	3.93	4.47	4.53	**1.40**
Win/Tie/Loss	15/0/0	13/0/2	14/0/1	15/0/0	14/0/1	–

Table 3
Coverage(↓) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	**3.6992 (1)**	3.9074 (5)	3.7311 (2)	3.8505 (4)	3.9111 (6)	3.7476 (3)
Virus	1.3607 (5)	1.2645 (2)	1.3414 (4)	1.3838 (6)	1.3095 (3)	**1.2100 (1)**
Emotions	1.8254 (2)	1.8655 (3)	1.8896 (5)	1.8689 (4)	1.9095 (6)	**1.7880 (1)**
Plant	2.4618 (2)	2.5907 (5)	2.5271 (3)	2.6818 (6)	2.5783 (4)	**2.3823 (1)**
Birds	2.4858 (2)	2.8118 (4)	2.7882 (3)	2.8503 (5)	2.8861 (6)	**2.2653 (1)**
Yeast	6.4571 (3)	**6.3045 (1)**	6.6731 (6)	6.5104 (5)	6.4728 (4)	6.4434 (2)
Gpositive	0.4491 (2)	0.4704 (3)	0.5127 (4)	0.5953 (6)	0.5377 (5)	**0.4374 (1)**
Guardian	2.3794 (6)	2.2889 (4)	1.8276 (2)	2.3167 (5)	2.2211 (3)	**1.8204 (1)**
BBC	2.1452 (3)	2.1598 (4)	**1.6931 (1)**	2.1913 (5)	2.2143 (6)	1.8047 (2)
Eukaryote	2.9060 (2)	3.0972 (5)	2.9291 (3)	3.1450 (6)	3.0914 (4)	**2.8415 (1)**
Image	1.0415 (3)	**0.9610 (1)**	1.0770 (6)	1.0540 (5)	1.0420 (4)	1.0065 (2)
Huamn	2.5339 (2)	2.6821 (6)	2.5983 (3)	2.6569 (4)	2.6676 (5)	**2.5076 (1)**
3sources	2.2929 (5)	2.2468 (3)	1.7841 (2)	2.2993 (6)	2.2800 (4)	**1.7749 (1)**
Stackex	84.4309 (6)	82.7974 (4)	83.9263 (5)	82.5252 (3)	**80.885 (1)**	82.3297 (2)
Yelp	1.3132 (3)	1.3119 (2)	1.4425 (6)	1.3753 (4)	1.3796 (5)	**1.2909 (1)**
Rank(↓)	3.13	3.47	3.67	4.93	4.40	**1.40**
Win/Tie/Loss	14/0/1	14/0/2	14/0/1	15/0/0	14/0/1	–

Table 4
Hamming loss(↓) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	0.2895 (3)	0.2930 (4)	0.2825 (2)	0.3010 (5)	0.3105 (6)	**0.2762 (1)**
Virus	0.1986 (5)	0.1987 (6)	**0.1937 (1)**	0.1964 (4)	0.1940 (2)	0.1945 (3)
Emotions	0.2061 (4)	0.2119 (5)	0.2032 (2)	0.2047 (3)	0.2227 (6)	**0.2001 (1)**
Plant	0.0901 (5.5)	0.0901 (5.5)	0.0897 (2.5)	0.0897 (2.5)	0.0900 (4)	**0.0893 (1)**
Birds	**0.0472 (1)**	0.0493 (3)	0.0502 (5)	0.0505 (6)	0.0498 (4)	0.0476 (2)
Yeast	0.2030 (4)	**0.1947 (1)**	0.2144 (6)	0.2036 (5)	0.2028 (3)	0.2010 (2)
Gpositive	0.1508 (2)	0.1556 (3)	0.1600 (4)	0.1686 (5)	0.1759 (6)	**0.1465 (1)**
Guardian	0.1887 (3)	0.1921 (5)	**0.1822 (1)**	0.1904 (4)	0.1926 (6)	0.1882 (2)
BBC	0.1909 (6)	0.1894 (5)	**0.1795 (1)**	0.1890 (4)	0.1866 (3)	0.1857 (2)
Eukaryote	0.0513 (3.5)	0.0513 (3.5)	0.0512 (2)	0.0517 (5.5)	0.0517 (5.5)	**0.0511 (1)**
Image	0.1880 (6)	**0.1690 (1)**	0.1870 (5)	0.1822 (3)	0.1824 (4)	0.1810 (2)
Huamn	0.0831 (2)	0.0839 (5)	0.0834 (3)	0.0836 (4)	0.0841 (6)	**0.0830 (1)**
3sources	0.1882 (3)	0.1894 (5)	0.1842 (2)	0.1911 (6)	0.1888 (4)	**0.1820 (1)**
Stackex	0.0093 (4)	0.0093 (4)	0.0093 (4)	0.0093 (4)	0.0093 (4)	**0.0093 (1)**
Yelp	**0.2309 (1)**	0.2417 (3)	0.2457 (5)	0.2337 (2)	0.2583 (6)	0.2454 (4)
Rank(↓)	3.53	3.93	3.03	4.20	4.63	**1.67**
Win/Tie/Loss	13/0/2	13/0/2	12/0/3	15/0/0	15/0/0	–

- For the AP index, FSMRL achieves superior performance on eleven datasets, except for the Yeast, BBC, Image, and Stackex datasets. For these four datasets, the AP of FSMRL ranks second or fourth and is close to the best performance. Taking the Virus dataset as an example, FSMRL is 3.79%, 2.54%, 4.48%, 6.03%, and 2.91% better than SSFS, PMFS, MDFS, MIFS, and PMU, respectively. Although FSMRL does not achieve optimal performance on all datasets, it outperforms MIFS and SSFS on all datasets, which shows the effectiveness of FSMRL.
- As for the CV index in Table 3, FSMRL achieves the best performance values on the Virus, Emotions, Plant, Birds, Gpositive, Guardian, Eukaryote, Human, 3sources, and Yelp datasets. However, the performance of FSMRL on Flags, Yeast, BBC, Image, and Stackex datasets is slightly lower than the values achieved by MDFS, PMFS, SSFS, and PMU. In addition, on the Flags dataset, FSMRL is outperformed by MIFS and PMU by 2.67% and 4.18%, respectively. FSMRL shows the best performance when compared with MIFS, according to the result of "Win/Tie/Loss." As can be

Table 5
One-Error(↓) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	0.2582 (5)	0.2316 (3)	0.2839 (6)	0.2211 (2)	0.2526 (4)	**0.1968 (1)**
Virus	0.5493 (4)	0.5452 (3)	0.5593 (5)	0.5690 (6)	0.5314 (2)	**0.5157 (1)**
Emotions	0.2988 (4)	0.3222 (6)	0.2952 (3)	**0.2802 (1)**	0.3221 (5)	0.2834 (2)
Plant	0.6616 (2)	0.6697 (3)	0.6789 (4)	0.6902 (6)	0.6881 (5)	**0.6576 (1)**
Birds	0.7271 (3)	0.7085 (2)	0.7458 (6)	0.7300 (4)	0.7380 (5)	**0.6914 (1)**
Yeast	0.2437 (3)	**0.2209 (1)**	0.2528 (6)	0.2503 (4)	0.2524 (5)	0.2358 (2)
Gpositive	0.2853 (2)	0.3065 (3)	0.3315 (4)	0.3680 (6)	0.3508 (5)	**0.2795 (1)**
Guardian	0.7378 (5)	0.7116 (4)	0.6360 (2)	0.7449 (6)	0.7018 (3)	**0.6253 (1)**
BBC	0.7132 (5)	0.7106 (4)	0.6560 (2)	0.7160 (6)	0.7102 (3)	**0.6306 (1)**
Eukaryote	0.6285 (3)	0.6564 (4)	0.6281 (2)	0.6724 (6)	0.6716 (5)	**0.6056 (1)**
Image	0.3660 (5)	**0.3260 (1)**	0.3700 (6)	0.3490 (3)	0.3540 (4)	0.3365 (2)
Huamn	0.6217 (2)	0.6546 (5)	0.6407 (3)	0.6542 (4)	0.6584 (6)	**0.6201 (1)**
3sources	0.7314 (5)	0.7646 (6)	0.6332 (2)	0.7245 (4)	0.7038 (3)	**0.5847 (1)**
Stackex	0.7310 (6)	0.7139 (2)	0.7233 (5)	0.7216 (4)	0.7179 (3)	**0.7132 (1)**
Yelp	0.3032 (2)	0.3081 (3)	0.3373 (6)	0.3093 (4)	0.3186 (5)	**0.3003 (1)**
Rank(↓)	3.73	3.33	4.13	4.40	4.20	**1.20**
Win/Tie/Loss	15/0/0	14/0/2	15/0/0	14/0/1	15/0/0	–

Table 6
Ranking loss(↓) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	0.2111 (2)	0.2230 (4)	0.2119 (3)	0.2296 (5)	0.2390 (6)	**0.2052 (1)**
Virus	0.2283 (5)	0.2128 (2)	0.2221 (4)	0.2335 (6)	0.2198 (3)	**0.2007 (1)**
Emotions	0.1757 (2)	0.1805 (4)	0.1845 (5)	0.1767 (3)	0.1951 (6)	**0.1667 (1)**
Plant	0.2120 (2)	0.2228 (5)	0.2186 (3)	0.2327 (6)	0.2214(4)	**0.2044 (1)**
Birds	0.0878 (2)	0.0985 (3)	0.1014 (4)	0.1027 (5)	0.1034 (6)	**0.0758 (1)**
Yeast	0.1788 (3)	**0.1696 (1)**	0.1929 (6)	0.1821 (5)	0.1792 (4)	0.1774 (2)
Gpositive	0.1463 (2)	0.1546 (3)	0.1685 (4)	0.1959 (6)	0.1772 (5)	**0.1434 (1)**
Guardian	0.4442 (6)	0.4213 (4)	0.3403 (2)	0.4287 (5)	0.4097 (3)	**0.3339 (1)**
BBC	0.3985 (4)	0.3970 (3)	**0.3073 (1)**	0.4039 (5)	0.4088 (6)	0.3303 (2)
Eukaryote	0.1245 (2)	0.1337 (5)	0.1258 (3)	0.1359 (6)	0.1333 (4)	**0.1213 (1)**
Image	0.1927 (3)	**0.1730 (1)**	0.2020 (6)	0.1961 (5)	0.1936 (4)	0.1839 (2)
Huamn	0.1705 (2)	0.1821 (6)	0.1759 (3)	0.1802 (4)	0.1813 (5)	**0.1686 (1)**
3sources	0.4238 (5)	0.4139 (3)	0.3248 (2)	0.4239 (6)	0.4217 (4)	**0.3225 (1)**
Stackex	0.1693 (6)	0.1654 (4)	0.1683 (5)	0.1647 (3)	**0.1611 (1)**	0.1646 (2)
Yelp	0.1469 (3)	0.1465 (2)	0.1765 (6)	0.1608 (4)	0.1625 (5)	**0.1420 (1)**
Rank(↓)	3.13	3.33	3.80	4.93	4.40	**1.27**
Win/Tie/Loss	15/0/0	13/0/2	14/0/1	15/0/0	14/0/1	–

Table 7
Micro-F1(↑) of comparing algorithms on fifteen datasets.

Datasets	SSFS	PMFS	MDFS	MIFS	PMU	FSMRL
Flags	0.7108 (3)	0.7007(4)	0.7134 (2)	0.6873 (5)	0.6786 (6)	**0.7138 (1)**
Virus	0.2606 (6)	0.2734 (5)	0.2808 (3)	0.2803 (4)	0.2946 (2)	**0.3102 (1)**
Emotions	0.6440 (2)	0.6395 (3)	0.6374 (4)	0.6373 (5)	0.6197 (6)	**0.6610 (1)**
Plant	**0.0802 (1)**	0.0582 (3)	0.0467 (4)	0.0146 (6)	0.0463 (5)	0.0780 (2)
Birds	0.3126 (3)	0.3141 (2)	0.2265 (4)	0.2094 (5)	0.2051 (6)	**0.3171 (1)**
Yeast	0.6193 (5)	0.6212 (2)	0.5827 (6)	0.6195 (4)	0.6202 (3)	**0.6216 (1)**
Gpositive	0.6744 (2)	0.6503 (3)	0.6346 (4)	0.6199 (5)	0.5883 (6)	**0.6753 (1)**
Guardian	0.0054 (4.5)	0.0000 (6)	**0.0877 (1)**	0.0277 (3)	0.0054(4.5)	0.0428 (2)
BBC	0.0045 (6)	0.0099 (5)	0.1039 (2)	0.0148 (4)	0.0325 (3)	**0.1078 (1)**
Eukaryote	0.1047 (3)	0.0549 (4)	0.1163 (2)	0.0336 (6)	0.0409 (5)	**0.1261 (1)**
Image	0.5233 (5)	**0.5755 (1)**	0.5077 (6)	0.5420 (3)	0.5377 (4)	0.5473 (2)
Huamn	0.1153 (2)	0.0814 (5)	0.0913 (3)	0.0867 (4)	0.0738 (6)	**0.1242 (1)**
3sources	0.0169 (5)	0.0221 (4)	0.1436 (2)	0.0335 (3)	0.0059 (6)	**0.1729 (1)**
Stackex	0.0408 (2)	0.0376 (3)	0.0278 (5)	**0.0452 (1)**	0.0371 (4)	0.0257 (6)
Yelp	**0.5954 (1)**	0.5681 (3)	0.5331 (5)	0.5680 (4)	0.5318 (6)	0.5721 (2)
Rank(↓)	3.37	3.53	3.53	4.13	4.83	**1.60**
Win/Tie/Loss	13/0/2	14/0/1	14/0/1	14/0/1	15/0/0	–

seen from Table 3, FSMRL achieves better performance results in most cases.

- For the HL index, FSMRL obtains remarkable performance on Flags, Emotions, Plant, Gpositive, Eukaryote, Human, 3sources, and Stackex datasets. Furthermore, for Virus, Birds, Yeast, Guardian, BBC, and Image datasets, the HL values of FSMRL are close to the best performance values. As shown in Table 4, it can be noted that MDFS, PMFS, SSFS, PMU, and FSMRL outperform other comparison algorithms on 3, 2, 2, 1, and 8 datasets, respectively. As for FSMRL, although its value of HL is 0.0004 inferior to SSFS for the Birds dataset, 0.0008, 0.006, and 0.0062 inferior to MDFS for the Virus, Guardian, and BBC datasets, respectively, it achieves the best average performance.
- Regarding the OE index in Table 5, FSMRL achieves the highest values across all 12 experimental datasets, including the Stackex and Yelp datasets. FSMRL ranks second in the Emotions, Yeast,

and Image datasets. Specifically, on the Emotions dataset, the MIFS method obtains the lowest value of OE. On the Yeast dataset, the PMFS algorithm attains the lowest performance value for OE. In most cases, FSMRL outperforms the other methods. According to the row "Rank", the outperform ratios of the proposed FSMRL algorithm are 70.9% compared to MDFS, 72.7% compared to MIFS,64% compared to PMFS, 71.4% compared to PMU, and 67.8% compared to SSFS.

- Concerning the RL index, we can observe that the proposed FSMRL algorithm ranks first in eleven datasets and second in four datasets. In other words, the FSMRL algorithm is superior to the other five algorithms in most cases. For example, on the Emotions dataset, the Ranking Loss value of FSMRL is 0.1667. In addition, in the row "Rank", we can obtain the average rank of all algorithms as follows: SSFS (3.13), PMFS (3.33), MDFS (3.80), MIFS (4.93), PMU (4.4), FSMRL (1.27).
- As indicated in Table 7, FSMRL obtains better performance on ten datasets. For the remaining datasets, FSMRL ranks second on the Plant, Guardian, Image, and Yelp datasets. We can observe that MDFS, MIFS, PMFS, PMU, SSFS, and FSMRL achieve optimal performance on one, one, one, zero, two, and ten datasets, respectively. Although FSMRL performs slightly worse on the Stackex dataset, it has achieved good performance in the average ranks. For the Plant dataset, the performance result of FSMRL is slightly inferior to SSFS but shows significant advantages over the remaining comparison algorithms. According to the above analysis, it still demonstrates the effectiveness of the proposed algorithm.

Overall, it can be observed that FSMRL shows outstanding predictive performance on fifteen multi-label datasets. Meanwhile, the effect of the proposed label enhancement method on these datasets can also be verified. The most important reason for FSMRL being superior to other comparison methods is that the label significance achieved by the proposed label enhancement method is beneficial for obtaining more label supervision information. In addition, the correlation between labels has been explored in constructing feature scoring functions, providing a comprehensive measurement perspective. Another key factor is the use and expansion of traditional MRMR measurement strategies. In conclusion, the proposed FSMRL algorithm is feasible for enhanced feature selection in multi-label datasets.

To further conduct a detailed comparison between FSMRL and other comparison methods, a series of experiments are carried out to show the performance changes trends with the feature dimension, as displayed in Figs. 5–6. Taking the Emotions and Virus datasets as representatives, we illustrate the performance changes with six different evaluation metrics. In each subfigure, the x-coordinate represents the feature dimension, while the y-coordinate indicates the value of the corresponding evaluation measure. In addition, the trend of the curves in different colors reflects the changes in classification performance with feature subsets under different feature selection algorithms.

For the Emotions dataset, it can be observed from Fig. 5 that FSMRL can show obvious advantages overall compared to the other algorithms across all evaluation measures as the size of selected features increases. In addition, PMU exhibits the worst performance among the six evaluation algorithms, varying with the feature dimension. For the Emotions dataset in Fig. 5, the trend of the curve gradually becomes stable with the change in the number of features, which further confirms the efficiency of the FSMRL algorithm. Additionally, FSMRL consistently achieves optimal performance on the Emotions dataset. Based on the above analysis, we can further illustrate the effectiveness of the FSMRL algorithm.

As can be observed in Fig. 6, the size of the selected candidate features changes from 10 to 130 in step 10. It can be seen that the trend of each method's variation with the number of selected features under various evaluation metrics is generally similar. The trend of the curve does not monotonically decrease or increase with changes in the feature dimension, which may be caused by irrelevant, redundant features. From Fig. 6(a) to (f), FSMRL can obtain the optimal performance in all the evaluation metrics as the size of the feature subset changes. According to the above analysis, the better performance of FSMRL indicates the efficiency of enhancing the labels from a logical form into a label distribution form to explore the label information and label correlation is explored in enhanced multi-label data.

Furthermore, to display the effectiveness of FSMRL in more detail,'single measure' and 'overall measure' are used to evaluate the overall performance of six comparison algorithms in Fig. 7, respectively. The average ranking for each metric is based on a single measure perspective, while the overall average rank for each method is derived from an overall measure perspective. According to (a) and (b) of Fig. 7, we can see that FSMRL demonstrates better performance than the other five comparison algorithms.

Table 8
The Friedman statistics F_F and the Critical value ($\alpha = 0.05$) on different metrics ($K = 6, N = 13$).

Evaluation metrics	F_F(MLKNN)	Critical value
Hamming loss	6.3362	2.346
One error	9.3304	
Coverage	10.3336	
Ranking loss	9.5658	
Average precision	8.4451	
Micro_F1	6.8059	

5.4. Statistical test

To further analyze the experimental effect of the FSMRL algorithm through statistical tests, we select the commonly used Friedman test [49] and Bonferroni–Dunn test [50] for statistical analysis. Among them, the Friedman test relies on a null hypothesis, which states that all compared algorithms perform equivalently, and the formula is as follows:

$$\chi_F^2 = \frac{12N}{K(K+1)}\left(\sum_{j=1}^{K} R_j^2 - \frac{K(K+1)^2}{4}\right),$$

where N is the number of experimental datasets, and K represents the number of compared algorithms; $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$ indicates the average rank value of the jth algorithm on all experimental datasets. Then, the Friedman statistic F_F can be computed as

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}.$$

Note that the F_F follows the F-distribution with $K-1$ and $(K-1)(N-1)$ degrees of freedom. At a significance level of $a = 0.05$, the critical value (2.368) is found from the corresponding inquiry form for $K = 6$ and $N = 15$. Taking the Ranking Loss metric as an example, the mean ranks $R_1 = 3.13$, $R_2 = 3.33$, $R_3 = 3.80$, $R_4 = 4.93$, $R_5 = 4.40$, and $R_6 = 1.27$, corresponding to the $SSFS$, $PMFS$, $MDFS$, $MIFS$, PMU, and $FSMRL$ algorithms, can be obtained, respectively (as displayed in Table 6). Based on the average rank $R_j (j = 1, 2, \ldots, 6)$ and the formula for the Friedman test, the $\chi_F^2 = 30.44$ and $F_F = 9.5658$. Table 10 displays the calculation results of the F_F.

In Table 8, it is not difficult to find that all F_F values on the six evaluation metrics are far greater than the critical value. Therefore, according to the Friedman test, the null hypothesis can be rejected on all six evaluation metrics at $\alpha = 0.05$. That is to say, our proposed method in this paper significantly outperforms the other five representative comparison algorithms across fifteen datasets and six evaluation metrics. In order to compare the relative performance between algorithms, the Bonferroni–Dunn test is also used to analyze the statistical results. According to this test, the relative distance between any two

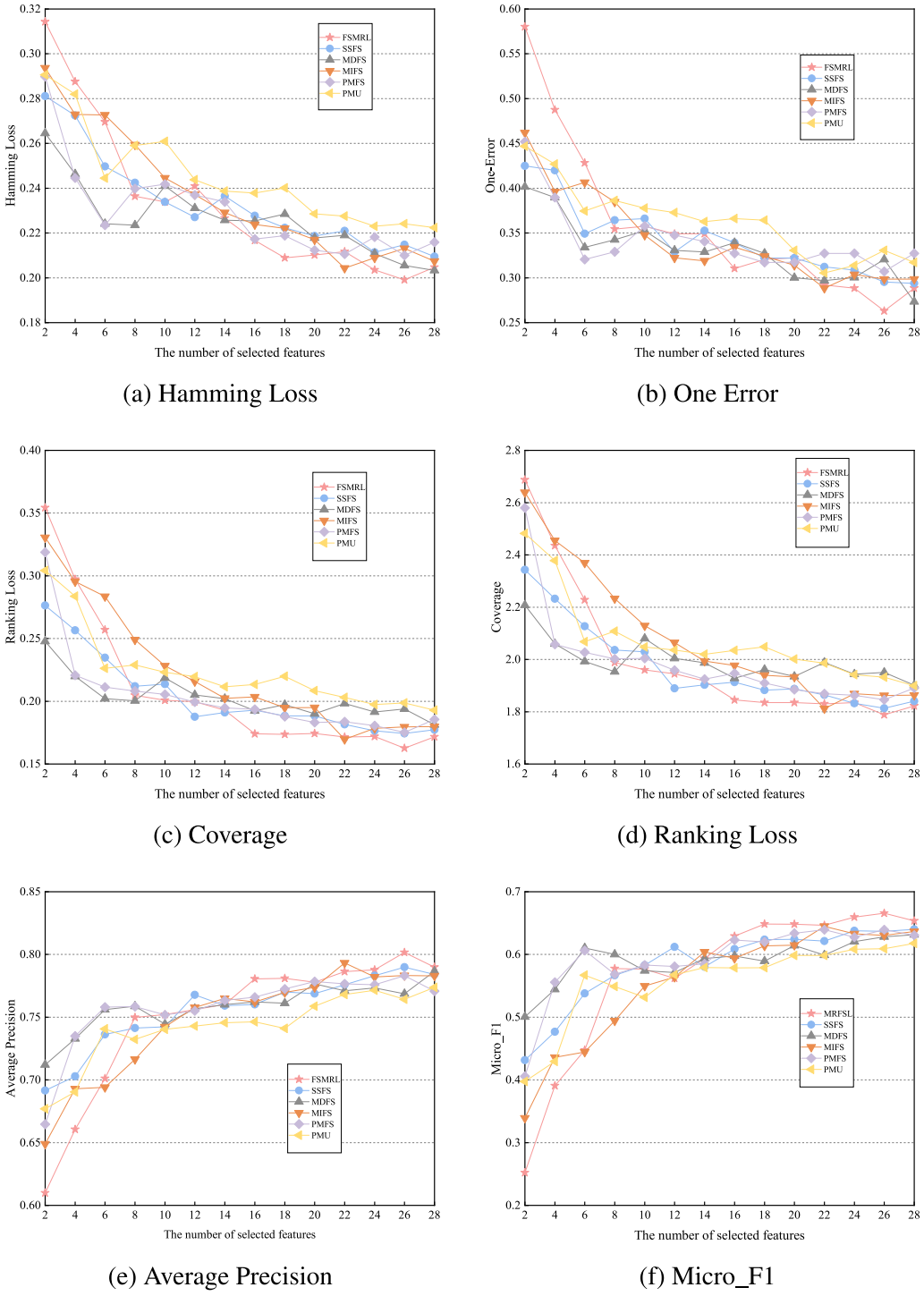


Fig. 5. The performance variation of Emotions dataset.

algorithms is calculated, and if the algorithms are considered to have a significant difference, the critical difference (CD) is formulated as

$$CD_\alpha = q_\alpha \sqrt{\frac{K(K+1)}{6N}},$$

where the definitions of K and N are given as introduced previously. The q_α represents the critical value for the Bonferroni–Dunn test, and the value of $q_\alpha = 2.576$ can be obtained when setting the significance level α to 0.05.

As shown in Fig. 8, if the interval between the average ranking of an algorithm and the proposed algorithm(FSMRL's) exceeds a CD length, then it can be observed that the algorithm has a significant difference from FSMRL. On the contrary, the performance of an algorithm and FSMRL is considered to be comparable. To further display the performance comparison of FSMRL and the other five comparison algorithms, the CD diagrams on six multi-label evaluation metrics are shown, and the specific results are shown in Fig. 8(a)–(f). In each sub-graph, there is a red equal-width line, which is utilized to represent the rank of the algorithm. If an algorithm is close to the proposed algorithm, it has a similar performance to the proposed algorithm. In other words, if the interval between an algorithm and the proposed algorithm exceeds

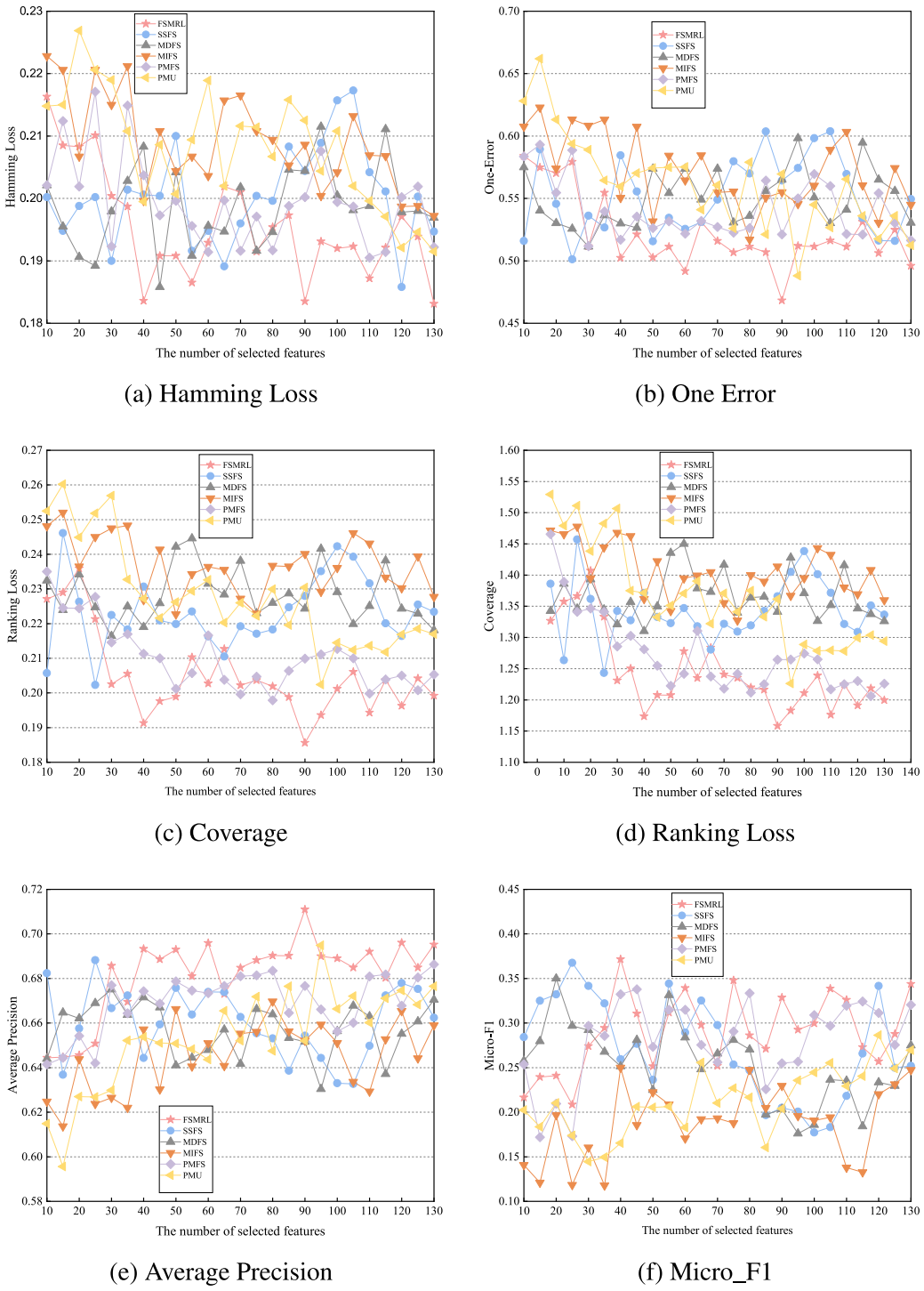


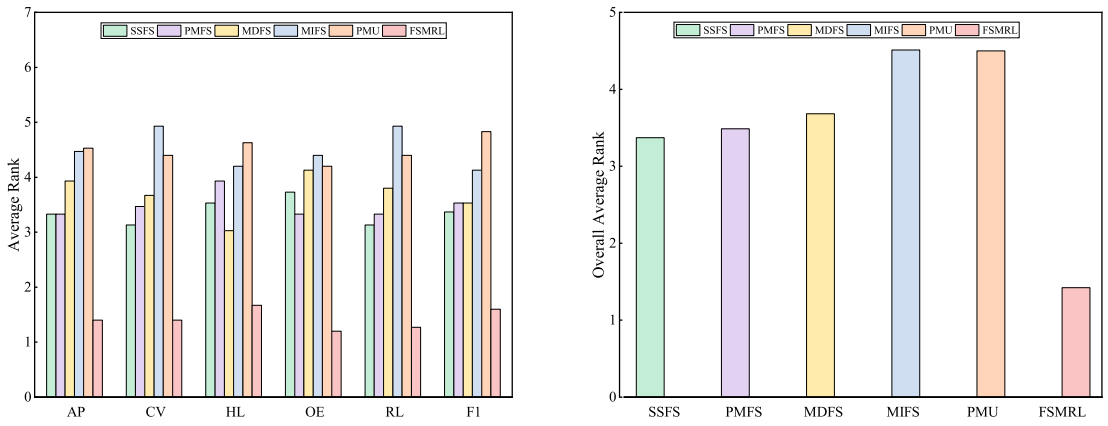
Fig. 6. The performance variation of Virus dataset.

the red CD line, then this algorithm is significantly different from the proposed algorithm.

In Fig. 8, we can come to the following conclusions:(1) The results for six evaluation metrics show that FSMRL is better than the other comparison algorithms, as the mean rank value of the FSMRL is the smallest and is located to the far right of the coordinate axis relative to other algorithms. (2) For the Average Precision metrics and One Error, there is no connection between the FSMRL algorithm and other comparison algorithms. In other words, FSMRL has achieved significant differences. The main reason for this may be influenced by exploring richer label semantic information through label enhancement. (3) Although for Hamming Loss, Ranking Loss, Coverage, and Micro_F1, FSMRL has a slight connection with some algorithms, there is no obvious evidence to reflect a difference among all the comparison algorithms.

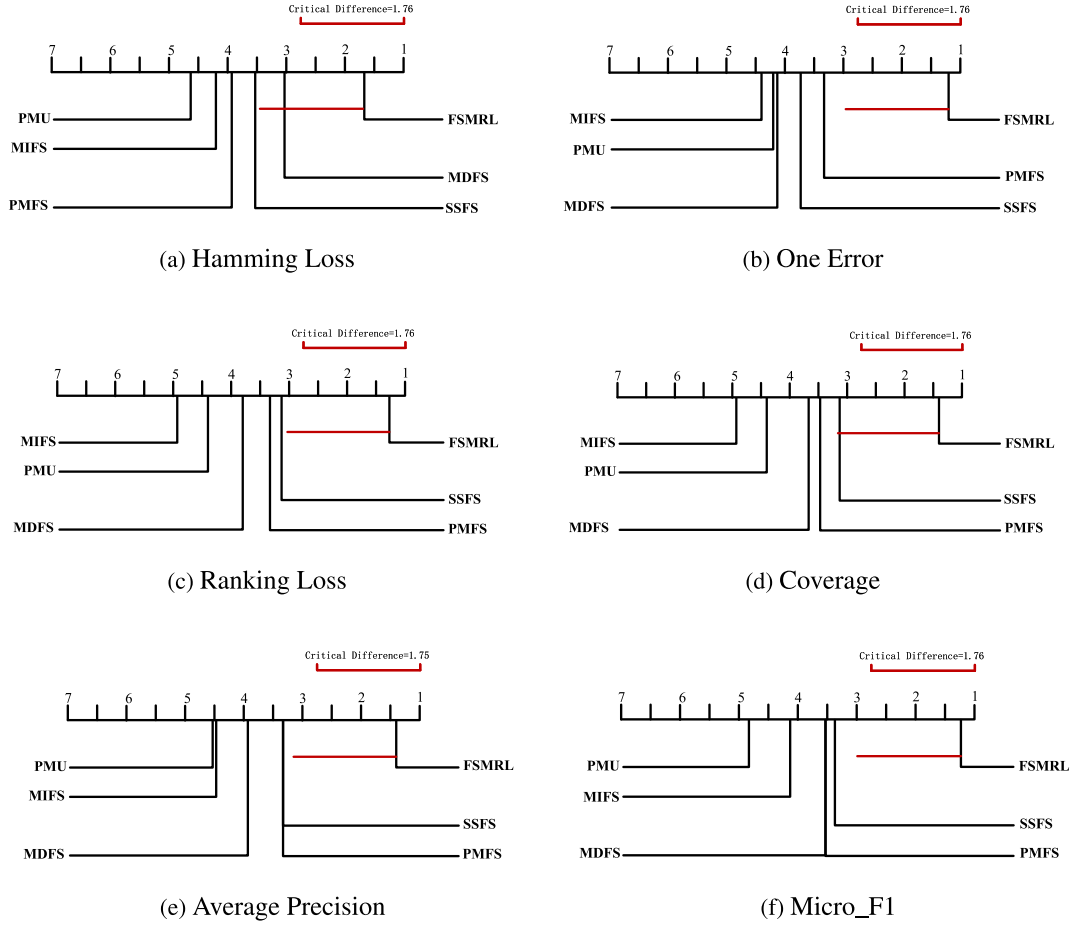
5.5. Ablation study

In this section, to explore the specific impact of label correlation terms on FSMRL, an ablation study is conducted on four representative multi-label datasets, such as Virus, Emotions, Birds, and Plant. The experimental results are compared and analyzed using six multi-label



(a) Average Rank (b) Overall Rank

Fig. 7. The results of two ranks on seven comparison methods.



(a) Hamming Loss (b) One Error

(c) Ranking Loss (d) Coverage

(e) Average Precision (f) Micro_F1

Fig. 8. Comparison of proposed algorithm against others with Bonferroni–Dunn test.

evaluation metrics. Here, we combine the feature dependency and feature redundancy as one part and view the label correlation as the other part. As for FSMRL, the necessity of each part can be analyzed and compared based on the classification performance of the two parts alone and the classification performance after they are combined. The results of the experiments are shown as follows:

As discussed in the aforementioned analysis, DEP+RED represents the evaluation function that only considers feature dependency and feature redundancy, LCD means that only label correlation is considered, DEP+RED+LCD indicates that it not only describes the significance of features via the feature perspective but also from a label perspective. The performance results are shown in Fig. 9. From the sub-figs. (a)–(f), it can be observed that the performance of DE+RED+LCD is significantly superior to the result of DEP+RED and LCD. In addition, when measuring the significance of features through either the feature-based perspective or label-based perspective alone, DEP+RED is better than LCD on most datasets. The most important reason is that the former takes into consideration the correlation between features and

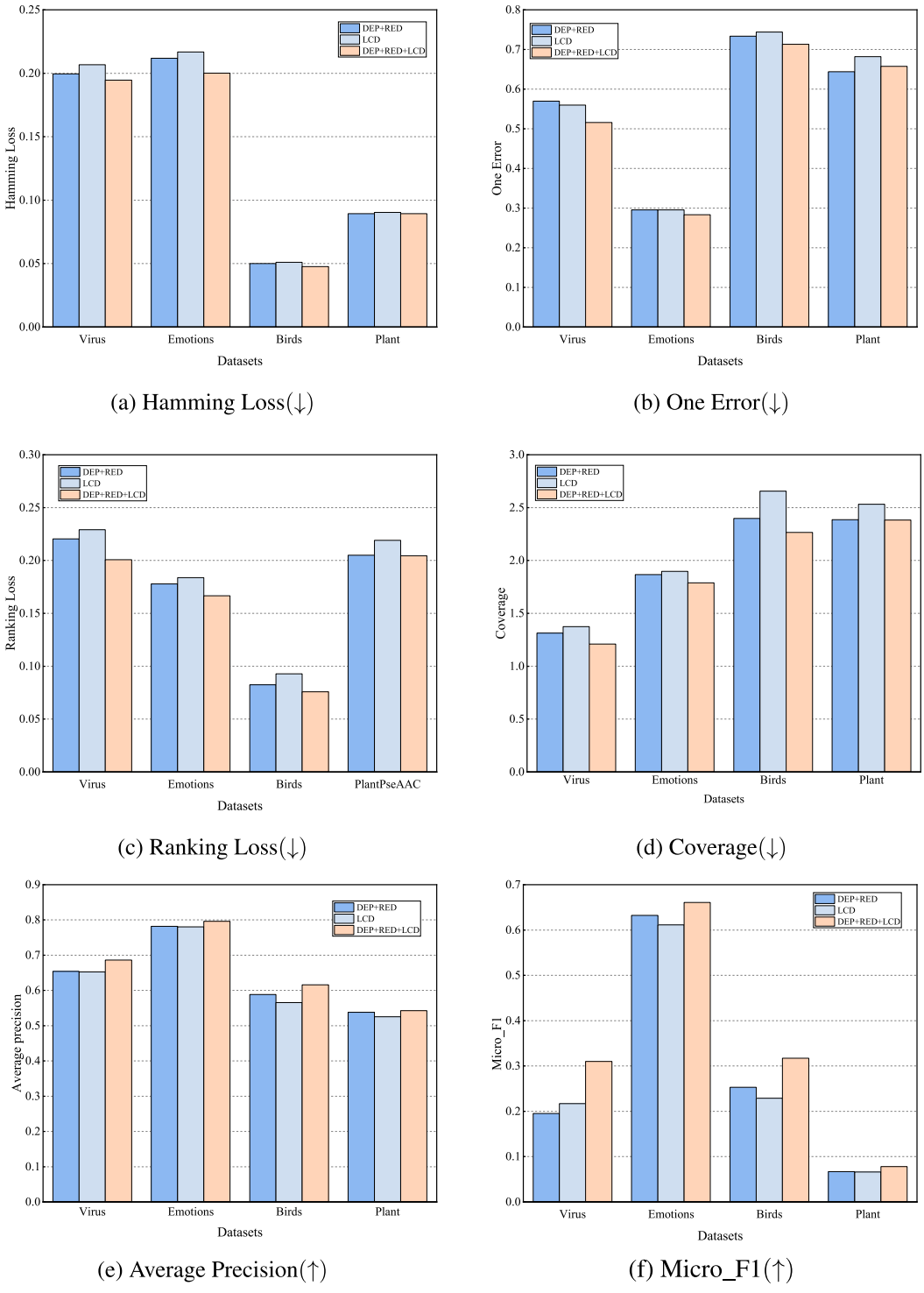


Fig. 9. Ablation study of the proposed method FSMRL on four datasets.

labels simultaneously. In summary, the proposed method to evaluate the significance of features by combining the feature perspective and label perspective can improve the classification performance.

5.6. Computational complexity

The computational complexity of the proposed algorithm FSMRL consists of two main parts: constructing the label distribution decision table and obtaining the ranking set of selected features. The first part includes normalizing the data, which typically takes $O(n)$, calculating k-nearest neighbors usually takes $O(n^2)$, and constructing the label relevance matrix, which is generally $O(k)$ for each of the k neighbors, leading to a complexity of $O(nk)$. The label weight calculation can also take $O(d)$. Therefore, the overall complexity is roughly $O(n^2+nk+d)$. To obtain the ranking set of selected features, assuming each of the calculations (DEP, RED, LCD) has a rough complexity of $O(n)$, the total complexity would be $O(n+n^2+dn+d^2)$, simplifying to $O(n^2+dn)$. To combine the two-time complexities $O(n^2+nk+d)$ and $O(n^2+dn)$, the

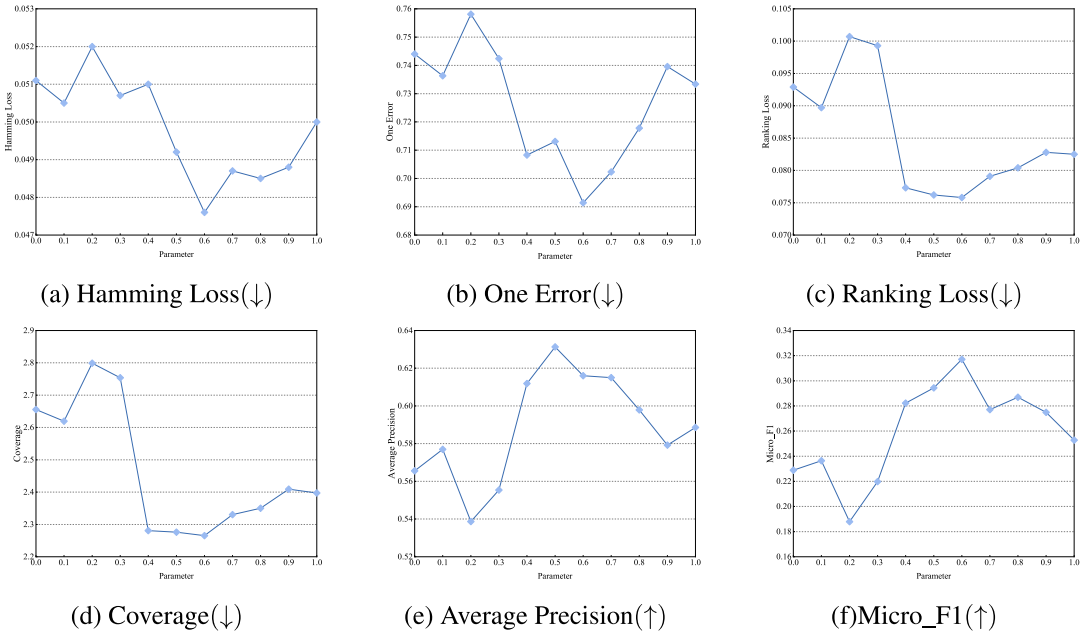


Fig. 10. Classification performance on six evaluation metrics of Birds dataset with parameter λ.

Table 9
Experimental feature selection methods with complexity.

Method	SFFS	PMFS	MDFS	MIFS	PMU	FSMRL
Time complexity	$O(Tn(d+c))$	$O\left(d^2n+d^2\right)$	$O(nd^3+d^3)$	$O\left(ndc+n^2\right)$	$O(nd^2)$	$O(n^2+nk+dn)$
Ref.	[43]	[44]	[45]	[44]	[46]	Proposed

overall time complexity can be expressed as $O(n^2+nk+dn)$, where k indicates the neighborhood adjustment parameter, n indicates the number of instances, d represents the number of features, and c represents the number of labels. The complexity of the comparison algorithm is shown in Table 9, Where T indicates the number of iterations.

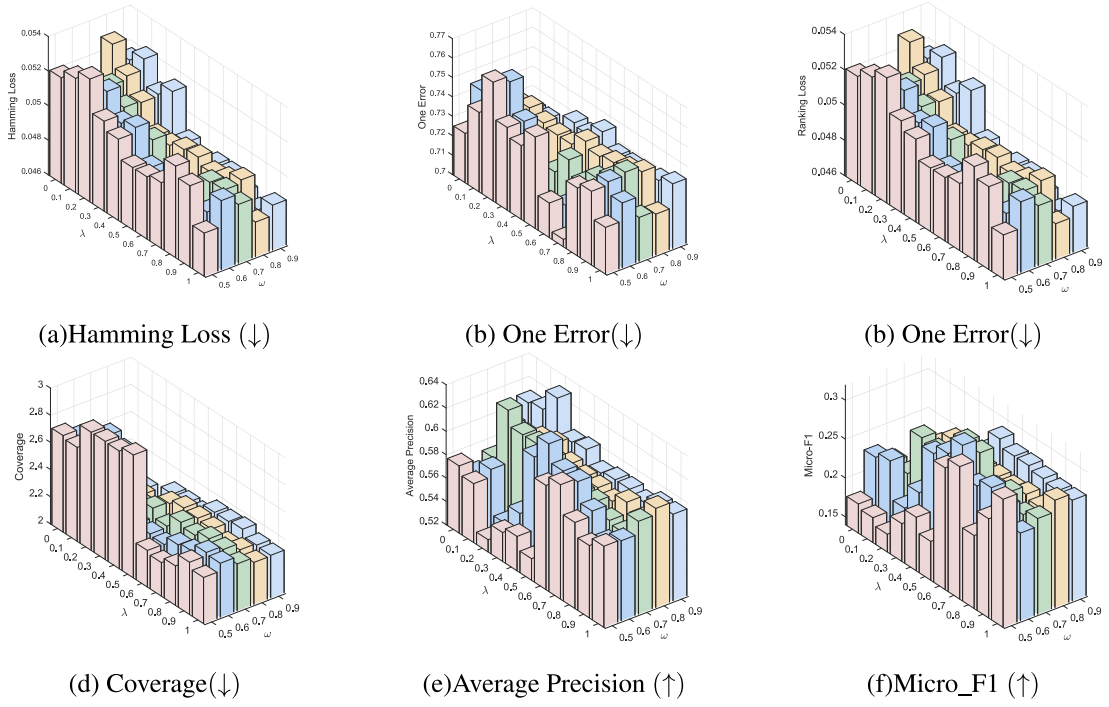
5.7. Parameter analysis

In the FSMRL algorithm, the trade-off between feature dependency, feature redundancy, and label relevance is controlled by the balance parameter λ. To verify the availability of λ in constructing the evaluation function, six evaluation metrics for different λ values are compared using the FSMRL algorithm. In the experiments, the parameter λ is varied in the range $[0,1]$ in steps of 0.1. Take the dataset Birds as an example. Fig. 10 displays the performance of the six evaluation metrics as the parameter λ changes. It is observed that when the value of λ is equal to 0.6, it ensures the selection of highly informative features. Additionally, the method requires different balance parameters λ for different datasets. Therefore, to select highly informative features and achieve better performance, the specific parameter λ can be obtained by analyzing the corresponding dataset.

Meanwhile, in this subsection, we also conduct a series of comparative studies on the effects of different parameters on classification performance under different values. Due to limitations in the length of the paper, we take the effects of ω and λ on the Birds dataset under different values as an example. As shown in Fig. 11, the six evaluation metrics are displayed. The value of ω is set in the range $\{0.5,0.6,0.7,0.8,0.9\}$, and the value of λ is set in the range $\{0,0.1,\ldots,1\}$. The three sub-figures show that when ω and λ are set to 0.6, the best performance is achieved. In other words, the optimal feature subset can be selected under these parameters. As the values of λ increase, the Average Precision and Micro_F1 values increase, while the Hamming Loss values, One Error values, Coverage values, and Ranking Loss values decrease. In short, the classification performance becomes better, and the parameter λ can affect classification performance. We can also see that as the parameter ω changes, there is no significant increase or decrease in classification performance. This is because the k-nearest neighborhood used in this paper may reduce the impact of changes in parameter ω. From Fig. 11, the six evaluation metric values are going to change dramatically with the change of two parameters. Thus, it can be concluded that the proposed FSMRL method is sensitive to variations in its parameters.

6. Conclusions

Multi-label feature selection has gained increasing attention, yet challenges remain in addressing high dimensionality in multi-label learning, specifically: (a) uneven data distribution, (b) unequal label significance, and (c) neglect of label correlations. This paper presents a feature selection algorithm that simultaneously considers feature dependency, redundancy, and label correlation. First, local similarity samples are identified using k-nearest neighbors. Second, a label enhancement method leveraging samples and combined label correlations is introduced to enrich label-supervised information. Additionally, label correlations are derived by analyzing the overlap of different low approximations related to various labels in the enhanced multi-label data. Finally, we present a novel algorithm, FSMRL, which employs a multi-criterion evaluation strategy for multi-label feature selection. Comparative experiments with five representative feature selection algorithms across fifteen diverse datasets and six evaluation metrics demonstrate the superiority of our method. Despite these advancements, practical applications face challenges in acquiring labels due to limitations in human and material resources. Future work will focus on developing feature selection algorithms tailored for incomplete data scenarios.



(a)Hamming Loss (↓) (b) One Error(↓) (b) One Error(↓)

(d) Coverage(↓) (e)Average Precision (↑) (f)Micro_F1 (↑)

Fig. 11. Classification performance on six evaluation metrics of Birds dataset with different $\omega = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\lambda = \{0, 0.1, \ldots, 1\}$.

CRediT authorship contribution statement

Wenbin Qian: Methodology, Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Wenyong Ruan:** Methodology, Data curation, Software, Writing – original draft. **Xiwen Lu:** Validation, Writing – review & editing. **Wenji Yang:** Formal analysis, Writing – review & editing. **Jintao Huang:** Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62366019, No. 62366018, and No. 61966016), and the Natural Science Foundation of Jiangxi Province, China (No. 20242BAB23014, No. 20224BAB202020 and No. 20224BAB202015).

Data availability

Data will be made available on request.

References

[1] Q. Lai, J. Zhou, Y. Gan, C.M. Vong, C.L.P. Chen, Single-stage broad multi-instance multi-label learning (BMIML) with diverse inter-correlations and its application to medical image classification, IEEE Trans. Emerg. Top. Comput. Intell. (2023) 1–12.

[2] W. Xu, Z. Yuan, Z. Liu, Feature selection for unbalanced distribution hybrid data based on k -nearest neighborhood rough set, IEEE Trans. Artif. Intell. (2023) 1–15.

[3] K. Qu, J. Xu, Q. Hou, K. Qu, Y. Sun, Feature selection using Information Gain and decision information in neighborhood decision system, Appl. Soft Comput. 136 (2023) 110100.

[4] M.A.N.D. Sewwandi, Y. Li, J. Zhang, A class-specific feature selection and classification approach using neighborhood rough set and K-nearest neighbor theories, Appl. Soft Comput. 143 (2023) 110366.

[5] K. Wang, N. Xu, M. Ling, X. Geng, Fast label enhancement for label distribution learning, IEEE Trans. Knowl. Data Eng. 35 (2023) 1502–1514.

[6] C. Tan, S. Chen, G. Ji, X. Geng, Algorithm for multi-label distribution learning, IEEE Trans. Knowlegdge Data Eng. 34 (2022) 5098–5113.

[7] T. Yin, H. Chen, T. Li, Z. Yuan, C. Luo, Robust feature selection using label enhancement and β-precision fuzzy rough sets for multilabel fuzzy decision system, Fuzzy Sets and Systems 461 (2023) 1–34.

[8] J. Liu, Y. Lin, W. Ding, H. Zhang, C. Wang, J. Du, Multi-label feature selection based on label distribution and neighborhood rough set, Neurocomputing 524 (2023) 142–157.

[9] Y. Kou, G. Lin, Y. Qian, S. Liao, A novel multi-label feature selection method with association rules and rough set, Inf. Sci. 624 (2023) 299–323.

[10] J. Jiang, X. Zhang, Feature selection based on self-information combining double-quantitative class weights and three-order approximation accuracies in neighborhood rough sets, Inf. Sci. 657 (2024) 119945.

[11] W. Shu, Q. Xia, W. Qian, Neighborhood multigranulation rough sets for cost-sensitive feature selection on hybrid data, Neurocomputing 565 (2024).

[12] D. Xia, G. Wang, Q. Zhang, J. Yang, S. Li, M. Gao, Incremental approximation feature selection with accelerator for rough fuzzy sets by knowledge distance, IEEE Trans. Fuzzy Syst. 31 (2023) 3959–3973.

[13] X. Yang, H. Chen, T. Li, J. Wan, B. Sang, Neighborhood rough sets with distance metric learning for feature selection, Knowl.- Based Syst. 224 (2021) 107076.

[14] J. Xu, X. Meng, K. Qu, Y. Sun, Q. Hou, Feature selection using relative dependency complement mutual information in fitting fuzzy rough set model, Appl. Intell. 53 (2023) 18239–18262.

[15] D. Huang, Y. Chen, F. Liu, Z. Li, Feature selection for multiset-valued data based on fuzzy conditional information entropy using iterative model and matrix operation, Appl. Soft Comput. 142 (2023) 110345.

[16] P. Zhang, G. Liu, J. Song, MFSJMI: Multi-label feature selection considering join mutual information and interaction weight, Pattern Recognit. 138 (2023) 109378.

[17] K. Liu, T. Li, X. Yang, H. Ju, X. Yang, D. Liu, Feature selection in threes: Neighborhood relevancy, redundancy, and granularity interactivity, Appl. Soft Comput. 146 (2023) 110679.

[18] L. Sun, T. Yin, W. Ding, Y. Qian, J. Xu, Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems, Inf. Sci. 537 (2020) 401–424.

[19] W. Qian, J. Huang, Y. Wang, W. Shu, Mutual information-based label distribution feature selection for multi-label learning, Knowl.- Based Syst. 195 (2020) 105684.

[20] W. Gao, L. Hu, P. Zhang, Feature redundancy term variation for mutual information-based feature selection, Appl. Intell. 50 (2020) 1272–1288.

[21] J. Liu, Y. Lin, W. Ding, H. Zhang, J. Du, Fuzzy mutual information-based multi-label feature selection with label dependency and streaming labels, IEEE Trans. Fuzzy Syst. 31 (2022) 77–91.
[22] Z. Yuan, H. Chen, P. Zhang, J. Wan, T. Li, A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information, IEEE Trans. Fuzzy Syst. 30 (2022) 3395–3409.
[23] J. Liu, Y. Li, W. Weng, J. Zhang, B. Chen, S. Wu, Feature selection for multi-label learning with streaming label, Neurocomputing 387 (2020) 268–278.
[24] X. Che, D. Chen, J. Mi, Label correlation in multi-label classification using local attribute reductions with fuzzy rough sets, Fuzzy Sets and Systems 426 (2022) 121–144.
[25] Y. Fan, B. Chen, W. Huang, J. Liu, W. Weng, W. Lan, Multi-label feature selection based on label correlations and feature redundancy, Knowl.- Based Syst. 241 (2022) 108256.
[26] J. Dai, J. Chen, Y. Liu, H. Hu, Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation, Knowl.- Based Syst. 207 (2020) 106342.
[27] D. You, Y. Wang, J. Xiao, Y. Lin, M. Pan, Z. Chen, L. Shen, X. Wu, Online multi-label streaming feature selection with label correlation, IEEE Trans. Knowl. Data Eng. 35 (2023) 2901–2915.
[28] Z.F. He, M. Yang, Y. Gao, H.D. Liu, Y. Yin, Joint multi-label classification and label correlations with missing labels and feature selection, Knowl.- Based Syst. 163 (2019) 145–158.
[29] J. Liu, W. Wei, Y. Lin, L. Yang, H. Zhang, Learning implicit labeling-importance and label correlation for multi-label feature selection with streaming labels, Pattern Recognit. 147 (2024) 110081.
[30] L. Sun, Y. Ma, W. Ding, J. Xu, Sparse feature selection via local feature and high-order label correlation, Appl. Intell. 54 (2023) 565–591.
[31] G.L. Li, H.R. Zhang, F. Min, Y.N. Lu, Two-stage label distribution learning with label-independent prediction based on label-specific features, Knowl.- Based Syst. 267 (2023) 110426.
[32] J. Wang, X. Geng, Label distribution learning by exploiting label distribution manifold, IEEE Trans. Neural Netw. Learn. Syst. 34 (2023) 839–852.
[33] X. Jia, X. Shen, W. Li, Y. Lu, J. Zhu, Label distribution learning by maintaining label ranking relation, IEEE Trans. Knowl. Data Eng. 35 (2021) 1695–1707.
[34] T. Zhang, Y. Mao, F. Shen, J. Zhao, Label distribution learning through exploring nonnegative components, Neurocomputing 501 (2022) 212–221.
[35] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui, Label distribution learning on auxiliary label space graphs for facial expression recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13981–13990.
[36] Z. Deng, T. Li, D. Deng, K. Liu, P. Zhang, S. Zhang, Z. Luo, Feature selection for label distribution learning using dual-similarity based neighborhood fuzzy entropy, Inf. Sci. 615 (2022) 385–404.
[37] W. Qian, Y. Xiong, J. Yang, W. Shu, Feature selection for label distribution learning via feature similarity and label correlation, Inf. Sci. 582 (2022) 38–59.
[38] X. Jia, Z. Li, X. Zheng, W. Li, S.J. Huang, Label distribution learning with label correlations on local samples, IEEE Trans. Knowl. Data Eng. 33 (2021) 1619–1631.
[39] W. Qian, J. Huang, Y. Wang, Y. Xie, Label distribution feature selection for multi-label classification with rough set, Internat. J. Approx. Reason. 128 (2021) 32–55.
[40] Q. Zhang, S.Liu, J.Wang, Z.Li, C.Wen, Feature selection for multi-labeled data based on label enhancement technique and mutual information, Inf. Sci. 679 (2024) 121113.
[41] L. Sun, X. Zhang, Y. Qian, J. Xu, S. Zhang, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Inf. Sci. 502 (2019) 18–41.
[42] W. Qian, X. Long, Y. Wang, Y. Xie, Multi-label feature selection based on label distribution and feature complementarity, Appl. Soft Comput. J. 90 (2020) 106167.
[43] W. Gao, Y. Li, L. Hu, Multilabel feature selection with constrained latent structure shared term, IEEE Trans. Neural Netw. Learn. Syst. 34 (2023) 1253–1262.
[44] A. Hashemi, M. Bagher Dowlatshahi, H. Nezamabadi-pour, An efficient Pareto-based feature selection algorithm for multi-label classification, Inf. Sci. 581 (2021) 428–447.
[45] J. Zhang, Z. Luo, C. Li, C. Zhou, S. Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recognit. 95 (2019) 136–150.
[46] X. Wang, Y. Zhou, Multi-label feature selection with conditional mutual information, Comput. Intell. Neurosci. 2022 (2022).
[47] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007) 2038–2048.
[48] W. Qian, C. Xiong, Y. Wang, A ranking-based feature selection for multi-label classification with fuzzy relative discernibility, Appl. Soft Comput. 102 (2021) 106995.
[49] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[50] O.J. Dunn, Multiple comparisons among means, J. Amer. Statist. Assoc. 56 (1961) 52–64.