Neurocomputing 465 (2021) 128–140

Contents lists available at ScienceDirect

Neurocomputing

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

Fuzzy rough discrimination and label weighting for multi-label feature selection

Anhui Tan [a,b], Jiye Liang [a,*], Wei-Zhi Wu [b,c], Jia Zhang [d], Lin Sun [e], Chao Chen [f]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, PR China
[b] School of Information Engineering, Zhejiang Ocean University, Zhoushan, Zhejiang 316022, PR China
[c] Key Laboratory of Oceanographic Big Data Mining and Application of Zhejiang Province, Zhoushan, Zhejiang 316022, PR China
[d] College of information science and technology, Jinan University, Guangzhou 510632, PR China
[e] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, PR China
[f] Marine Science and Technology College, Zhejiang Ocean University, Zhoushan, Zhejiang 316022, PR China

ARTICLE INFO

Article history:
Received 30 March 2021
Revised 10 August 2021
Accepted 2 September 2021
Available online 08 September 2021
Communicated by Zidong Wang

Keywords:
Fuzzy rough set
Discernibility matrix
Feature selection
Feature weighting
Fuzzy relation
Multi-label data

ABSTRACT

Fuzzy rough set is a theoretical framework of fuzzy uncertainty management, and discernibility matrix offers a mathematical foundation for algorithm construction of feature learning. The approaches of fuzzy rough set and discernibility matrix have been successfully applied in single-label learning. However, few works have been done on investigating the foundation of fuzzy rough discernibility matrix on multi-label data. There will be two pivotal problems to be addressed when using fuzzy rough discernibility matrix for multi-label data analysis. One is how to extract sample-level and label-level correlations; and the other is how to utilize the discernibility matrix for algorithm construction. For this reason, in this paper the fuzzy rough discrimination matrix is introduced to deal with the problem of multi-label feature selection. First, the significance of labels in the label space is captured based on the label correlation. Labels with different significances contribute to different weights for measuring the similarity between samples. Hence, a sample similarity matrix in the label space is computed based on the label weighting strategy. Then, a framework of a fuzzy decision system is formalized, in which the discernibility matrix of fuzzy rough sets is introduced as a foundation to evaluate the sample discrimination ability of features. Under the discernibility matrix criterion, a multi-label learning algorithm is developed to select discriminative features from multi-label data. A series of experimental analysis verifies the effectiveness and efficiency of the proposed method.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, multi-label learning has rapidly expanded in various real-word domains such as image annotation, text categorization and medical diagnosis [1,2]. In multi-label learning, a sample is usually associated with multiple class labels simultaneously rather than a single one, and the main purpose of learning is to construct a classifier that can accurately predict the possible label set for test samples.

As we know, the performance of multi-label classification is strongly influenced by the input features. It is overextended for a machine when it directly handles high-dimensional data sets with a huge number of features. In fact, most of features are redundant and offer repetitive or even no semantic information which may decrease the classification performance of multi-label learning. Therefore, feature selection which aims to eliminate irrelevant features is an essential pre-processing step to alleviate the curse of dimensionality and to improve the learning performance. Up to now, various multi-label feature selection algorithms have been developed from different viewpoints, such as information metric [3–6], large margin [7,8], and sparse learning [9–11]. In many related works, label correlation is an important concern for constructing learning models, which can influence the accuracy of label annotation. According to the order of correlation, the strategies of multi-label learning are roughly divided into three categories: first-order, second-order and high-order strategies. First-order strategy treats a multi-label task as a set of single-label ones and deals with multi-label data in a label-by-label manner. Second-order strategy takes the pairwise relationship between labels into account [12–14], which may overlook the interactions

* Corresponding author.
E-mail addresses: tananhui86@163.com (A. Tan), ljy@sxu.edu.cn (J. Liang), wuwz@zjou.edu.cn (W.-Z. Wu), zhangjia_gl@163.com (J. Zhang), sunlin@htu.edu.cn (L. Sun), chenchao@zjou.edu.cn (C. Chen).

https://doi.org/10.1016/j.neucom.2021.09.007
0925-2312/© 2021 Elsevier B.V. All rights reserved.

among the class label subsets. Therefore, high-order strategy is developed to explore the correlation between one label and the subsets of class labels [15]. However, high-order approaches are usually more complex and less effective due to time complexity and algorithmic adaptability. Furthermore, label correlation can also be explored from the viewpoints of globality and locality simultaneously [16], and the local correlation can be captured via clustering and low-rank embedding.

Most of existing methods related to multi-label learning are based on the mapping or interaction between the feature space and the label space [9], while they don't consider the direct interaction of sample distributions between the two spaces. This leads to the fact that discriminative information hidden in the set of samples cannot be discovered in some degree. To address this deficiency, in this paper, we propose a sample-oriented multi-label feature selection method via label weighting and fuzzy rough discrimination. First, the pairwise label correlation is explored, based on which the weights are assigned for measuring the significance of each label in the label space. Second, the similarities between samples in the label space are generated by combing the set of weighted labels. Third, fuzzy rough discernibility matrix is introduced to discriminate inter-class samples in the framework of the fuzzy decision system. Finally, a multi-label feature selection algorithm based on fuzzy rough discrimination is designed, and experimental results demonstrate that the proposed method is superior to some state-of-the-art multi-label learning algorithms for multi-label feature selection.

The rest of this paper is organized as follows. We recall related notations and construct the sample similarity matrix in the label space in Section 2. In Section 3, we develop a framework of a fuzzy decision system, introduce the fuzzy rough discernibility matrix, and present our algorithm in detail. In Section 4, we analyze the effectiveness of the proposed method by a series of experiments. Finally, we summarize the study in Section 5.

1.1. Related work

Feature selection is an essential pre-processing step for many machine leaning tasks, which aims to delete irrelevant or redundant features so as to speed up the computational progress and improve the learning accuracy. In recent years, various types of multi-label feature selection algorithms have been developed. The algorithms based on information theory [3–5,17] provide scalable evaluation for selecting relative features. Lee et al. [3–5] proposed mutual information based multi-label feature selection methods by maximizing the dependency between selected features and labels. Li et al. [19] utilized information theory to evaluate the discrimination information provided by the selected feature subset. Lin et al. [20] employed mutual information to construct metrical scales by integrating max-dependency and min-redundancy. Qian et al. [21] transformed multi-label data into multi-label systems and constructed an information fusion-based feature selection method.

Sparse learning provides a well-established perspective for multi-label feature selection. Huang et al. [12] introduced a label-specific feature selection method, in which each class label was only determined by a sparse subset of relevant features. Huang et al. [13] proposed a feature selection method, namely manifold-based constraint laplacian score. The method first transformed the original logical labels into numerical ones via manifold learning, and then calculated the similarity between numerical labels based on an affinity matrix between corresponding samples. Zhang and Wu [9] developed a method that projected the original feature space into a lower-dimensional one by maximizing the dependence between features and associated class labels. Zhang et al. [22] constructed a sparse optimization model for multi-label feature selection based on label manifold and feature manifold. In the model, the local label-level and sample-level correlations were captured by embedding the label space and feature space into the unified lower-dimensional matrix via Laplacian mappings.

Fuzzy rough set [23] and discernibility matrix [24], as two important tools for uncertainty measurement in granular computing, have been widely applied in single-label feature selection. Since they are closely related to our study, we will briefly introduce them in the following. In fuzzy rough sets, a pair of lower and upper approximations are defined in terms of fuzzy membership functions to handling real-valued features. The model encapsulates both the concepts of fuzziness and indiscernibility for characterizing the uncertainty hidden in knowledge. Jensen and Shen [25,26] first proposed the feature selection methods with fuzzy rough sets, where the approximation qualities of decision classes were measured based on the fuzzy similarity of samples induced from features. Following the work of Jensen and Shen, a series of evolutionary versions were put forward to improve the algorithmic efficiency [27–29]. On the other hand, the concept of discernibility matrix clearly reveals the basic structures of optimal feature subsets under fuzzy rough sets. In the discernibility matrix, each entry stores a subset of features that can discriminate corresponding sample pairs. Tsang et al. [30] constructed the discernibility matrix of fuzzy rough sets and proposed an algorithm using discernibility matrix to compute all optimal feature subsets. Chen et al. [31] pointed out that the minimal elements in the discernibility matrix were sufficient and necessary for constructing optimal feature subsets. Hu et al. [32,33] introduced kernelized fuzzy rough sets to evaluate the approximation quality and approximation abilities of features in hybrid and multi-modality data. Zhang et al. [34] employed the discernibility matrix of fuzzy rough sets to introduce a type of information entropy for single-label feature selection. By considering the relevance and diversity between samples simultaneously, the first author of this paper [35,36] investigated the discernibility matrix and feature selection with a generalized intuitionistic fuzzy rough set. In recent years, the approaches of fuzzy rough sets have been gradually introduced and improved to deal with multi-label data. Li et al. [37] developed a multi-label kernelized fuzzy rough sets for multi-label feature selection. Lin [38] adopted the mean distance between samples to introduce a new fuzzy rough set model. The above-mentioned algorithms mainly utilized the approximations in fuzzy rough set models to design feature evaluation functions, while they pay less attention to the discrimination information of samples. Hence, Che et al. [39] employed the sample discrimination to formulate the correlation between label pairs and developed a discernibility matrix on special sample pairs. Noticeably, this method treated the labels equally and divided the samples into two disjoint parts in a label-by-label manner. In this paper, we introduce a sample discrimination method by using the discernibility matrix of fuzzy rough sets. To be specific, the similarities between samples are characterized by weighting the labels according to the diversities by comparing each label with label subsets. The inter-class discrimination and intra-class aggregation are recognized as criteria of feature selection based on the discernibility matrix of fuzzy rough sets (Table 1).

Table 1
Summary of notations used in this paper.

Notations	Meaning
n	number of samples
p	number of features
m	number of labels
1_m	m-dimensional all one row vector
$\|\|\cdot\|\|_1$	ℓ_1-norm
$\|\|\cdot\|\|_F$	Frobenius norm
Y	label matrix
$Y(i)$	ith row vector
$Y(i,j)$	(i,j)th entry
Y^T	transpose matrix of matrix Y
$\widetilde{Y}$	reverse matrix satisfying $\widetilde{Y}(i,j) = 1 - Y(i,j)$
ø	Hadamard division (entry-wise division)
C_L	label correlation matrix
S_S	sample similarity matrix
w_L	label weighting vector
$w(i)$	ith entry of vector w
Σ_w	diagonal matrix of vector w satisfying $\Sigma_w(i,i) = w(i)$

2. The sample similarity matrix of multi-label data

In this section, we first review basic notations in multi-label learning. Then, the weighting method of labels is proposed based on pairwise label correlation, and the sample similarity matrix in the label space is produced by combing the weights of related class labels. Notations used are summarized in Table 1

A multi-label data set is denoted by $MLD =< U, F, L >$, where $U = \{x_1, x_2, \cdots, x_n\}$ is the universe of n samples, $F = \{f_1, f_2, \cdots, f_p\}$ is the set of p features, and $L = \{l_1, l_2, \cdots, l_m\}$ is the set of m class labels. Assume that $Y \in \{0,1\}^{n\times m}$ is the ground-truth label matrix, and $Y(i)$ is the label vector of sample x_i. If $Y(i,j) = 1$, then x_i is associated with label l_j; otherwise $Y(i,j) = 0$.

Suppose $C_L \in \mathbb{R}^{m\times m}$ is the label correlation matrix. According to the assumption of label manifold, each label theoretically can be recovered by its neighboring labels. This assumption suggests us to induce the following optimization problem:

$$\begin{aligned} &\min_{C_L, Y_0} ||Y_0||_1 \\ &s.t. \quad Y = YC_L + Y_0 \end{aligned} \tag{1}$$

Here, $Y_0 \in \mathbb{R}^{n\times m}$ retains the noise labeling information which is constrained to be sparse by using the ℓ_1-norm. Each column of C_L contains the label correlation w.r.t. one label and each entry of C_L measures the similarity between a pair of labels. The larger the value is, the stronger the correlation between two labels is.

The above optimization problem is with a smooth convex loss function involving ℓ_1-norm regularization. To solve the above problem, the augmented Lagrange multiplier method [40] is adopted to induce the following Lagrange function:

$$\min_{C_L, Y_0} ||Y_0||_1 + < A, YC_L + Y_0 - Y > + \frac{\mu}{2}||YC_L + Y_0 - Y||_F^2 \tag{2}$$

where $A \in \mathbb{R}^{n\times m}$ is a Lagrange multiplier matrix, μ is a penalty parameter, and $||\cdot||_F$ is the Frobenius norm of a matrix.

According to the LADMAP method [41], Eq. (2) can be rewritten as

$$\min_{C_L, Y_0} ||Y_0||_1 + \frac{\mu}{2}||YC_L + Y_0 - Y + \frac{A}{\mu}||_F^2 \tag{3}$$

Specially, when Y_0 is fixed, the optimal solution of C_L is calculated as

$$C_L = \left(Y^T Y\right)^{-1}\left(Y^T Y_0 - Y^T Y + Y^T \frac{A}{\mu}\right) \tag{4}$$

The optimizing rule for Y_0 can be calculated based on the element-wise shrinkage operator [42] such that:

$$Y_0 = \mathscr{S}_{\frac{1}{\mu}}\left(Y - YC_L - \frac{A}{\mu}\right) \tag{5}$$

where $\mathscr{S}$ is the element-wise shrinkage operator [42], which is defined as $\mathscr{S}_w(a) = (a - w)_+ - (-a - w)_+$.

The Lagrange multiplier matrix A and the parameter μ at the t-th iteration are updated as:

$$\begin{aligned} A^{t+1} &= A^t + \mu^{t+1}(Y - YC_L - Y_0) \\ \mu^{t+1} &= \min(\mu_{\max}, \rho\mu^t) \end{aligned} \tag{6}$$

where ρ is a positive scalar.

The label correlation matrix records the pairwise correlation of labels, based on which the weights for measuring the significance of labels are defined as follows.

Definition 1. Let $MLD =< U, F, L >$ be a multi-label data set and C_L the label correlation matrix. Denote the label weighting vector $w_L \in \mathbb{R}^{1\times m}$ as

$$w_L = \left(1_m \widetilde{C_L}\right) / \left(1_m \widetilde{C_L} 1_m^T\right) \tag{7}$$

where $\widetilde{C_L} \in \mathbb{R}^{1\times m}$ is the reverse matrix of C_L satisfying that $\widetilde{C_L}(i,j) = 1 - C_L(i,j)$.

In Definition 1, the semantics of the numerator and denominator for calculating w_L can be explained as follows.

Property 1. (1) $\left(1_m \widetilde{C_L}\right)(i) = \sum_{k=1}^m \widetilde{C_L}(k,i)$;
(2) $1_m \widetilde{C_L} 1_m^T = \sum_{i=1}^m \sum_{k=1}^m \widetilde{C_L}(k,i)$;
(3) $\sum_{i=1}^m w_L(i) = 1$.

Proof. (1) It holds that $\left(1_m \widetilde{C_L}\right)(i) = \sum_{k=1}^m 1 \times \widetilde{C_L}(k,i) = \sum_{k=1}^m \widetilde{C_L}(k,i)$.
(2) and (3) are not hard to be induced from (1). □

We see from Property 1 that $\left(1_m \widetilde{C_L}\right)(i)$ is the total diversity of label l_i compared with all the other labels. The larger the value of $\left(1_m \widetilde{C_L}\right)(i)$ is, the more information the label l_i takes when deleting it from the label space. In other words, $w_L(i)$ is the weight assigned to label l_i, which is in accordance with the specific of the label in the label space.

Labels with different weights have different contributions to measuring the similarity between samples. The labels assigned with greater weights would contribute to greater significance when calculating the similarity of samples. Based on this idea, the sample similarity matrix is introduced as follows.

Definition 2. Let $MLD =< U, F, L >$ be a multi-label data set and w_L the label weighting vector. We define the sample similarity matrix $S_S \in \mathbb{R}^{n\times n}$ in the label space as

$$S_S = \left(Y\Sigma_{w_L} Y^T\right) ø \left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right) \tag{8}$$

where Σ_w is the diagonal matrix of vector w satisfying $\Sigma_w(i,i) = w(i)$.

In Definition 2, each entry $S_S(i,j)$ represents the similarity degree between samples x_i and x_j. To clearly illustrate the semantics of the sample similarity matrix, we state the following conclusions.

Theorem 1. *The following statements hold.*

(1) $\left(Y\Sigma_{w_L}Y^T\right)(i,j) = \sum\{w_L(k)|Y(i,k)=1 \wedge Y(j,k)=1, 1\leqslant k\leqslant m\}$;

(2) $\widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}(i,j) = \sum\{w_L(k)|Y(i,k)=0 \wedge Y(j,k)=0, 1\leqslant k\leqslant m\}$;

(3) $\left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right)(i,j) =$
$\sum\{w_L(k)|Y(i,k)=1 \vee Y(j,k)=1, 1\leqslant k\leqslant m\}$;

(4) $S_S(i,j) = \frac{\sum\{w_L(k)|Y(i,k)=1\wedge Y(j,k)=1\}}{\sum\{w_L(k)|Y(i,k)=1\vee Y(j,k)=1\}}$.

Proof. (1) It is not hard to prove that $\left(Y\Sigma_{w_L}Y^T\right)(i,j) = \sum_{k=1}^m w_L(k)Y(i,k)Y(j,k)$.

With the fact that Y is a 0–1 matrix, it hold $\left(Y\Sigma_{w_L}Y^T\right)(i,j) = \sum\{w_L(k)|Y(i,k)=1\wedge Y(j,k)=1\}$.

(2) From (1), we have $\left(\widetilde{Y}\Sigma_{w_L}\widetilde{Y}^T\right)(i,j) = \sum\left\{w_L(k)|\widetilde{Y}(i,k)=1\wedge\widetilde{Y}(j,k)=1\right\}$.

It follows that $\widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}(i,j) = \sum\{w_L(k)|Y(i,k)=0\wedge Y(j,k)=0\}$.

(3) Since $1_n^T 1_n(i,j)=1$, it holds from (2) that

$$\left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right)(i,j) = 1 - \sum\{w_L(k)|Y(i,k)=0\wedge Y(j,k)=0\}.$$

Combining with $\sum_{k=1}^m w_L(k)=1$, we have.

$$\left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right)(i,j) = \sum\{w_L(k)|Y(i,k)=1\vee Y(j,k)=1\}.$$

(4) It is direct from (1) and (3). □

In Item (1), $\left(Y\Sigma_{w_L}Y^T\right)(i,j)$ is the total weights of the labels jointly owned by samples x_i and x_j, whereas in Item (3), $\left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right)(i,j)$ is the total weights of the labels owned by at least one of samples x_i and x_j. Item (4) characterizes the semantic of each entry $S_S(i,j)$ in the matrix: If the weights of all labels are equal, $S_S(i,j)$ is just the ratio between the labels jointly owned by samples x_i and x_j and the labels owned by either x_i or x_j. As an extension, $S_S(i,j)$ is the similarity between samples x_i and x_j by combining the weights of all related labels.

According to the above analysis, we present a detailed algorithm for calculating the sample similarity matrix in the label space as shown in Algorithm 1.

Algorithm 1: An algorithm for computing the sample similarity matrix in the label space of a multi-label data set

Input: A multi-label data set $MLD =< U, F, L >$ with the label matrix Y.
Output: The sample similarity matrix S_S.
1: Compute the label correlation matrix by solving the optimization problem (1);
2: Compute the label weighting vector $w_L = \left(1_m\widetilde{C_L}\right)/\left(1_m\widetilde{C_L}1_m^T\right)$ by Eq. (7);
3: Construct the diagonal matrix Σ_{w_L} of w_L;
4: Compute the matrix $S_S = \left(Y\Sigma_{w_L}Y^T\right)\emptyset\left(1_n^T 1_n - \widetilde{Y}\Sigma_{w_L}\widetilde{Y^T}\right)$ by Eq. (8);
5: **Output** S_S.
6: **End**

3. A fuzzy decision information system induced from multi-label data

In the section above, the sample similarity matrix S_S is produced from the label space by weighting the pairwise correlation of labels. Each entry of the matrix reflects the similarity between the corresponding pair of samples in the label space. In the same sense, in the feature space, each feature f_k can induce a fuzzy relation $R_k = [R_k(x_i,x_j)] \in \mathbb{R}^{n\times n}$ by using various methods, e.g., similarity methods or kernel methods, which characterizes the pairwise similarity of samples on feature f_k. In particularly, denote $R_k(x_i,x_j) = \exp\left(-\frac{\|(x_i-x_j)_k\|^2}{2\sigma^2}\right)$, where σ is the parameter which is simply set by $\sigma=1$ for graph construction. Moreover, $(x_i-x_j)_k$ is the difference between x_i and x_j at the k-th feature. Consequently, a family of fuzzy relations $\mathscr{R} = \{R_1, R_2, \cdots, R_p\}$ can be obtained from the whole feature space.

Given a multi-label data set $MLD =< U, F, L >$, we below utilize the framework of fuzzy relations to represent the sample information in the data. Consequently, a fuzzy decision information system is formalized as $FS = (U, \mathscr{R}, S_S)$, where $\mathscr{R} = \{R_1, R_2, \cdots, R_p\}$ is a set of fuzzy relations induced from the the feature space and S_S is the sample similarity matrix in the label space. Under the framework of a fuzzy decision information system, we will investigate the multi-label feature evaluation and feature selection.

Given a subset of fuzzy relations $\mathscr{B}\subseteq\mathscr{R}$, the interaction fuzzy relation $\mathbf{B} = \cap\mathscr{B}$ can be obtained. The approximation ability of a subset of fuzzy relations is defined based on fuzzy rough set as follows.

Definition 3. [23] Let $FS = (U, \mathscr{R}, S_S)$ a fuzzy decision information system with $\mathscr{B}\subseteq\mathscr{R}$. The lower and upper approximations of each $x_i \in U$ w.r.t. $\mathscr{B}$ are respectively defined as:

$$\begin{aligned}\underline{\mathscr{B}}(x_i) &= \inf_{x_j\in U}\max(1-\mathbf{B}(x_i,x_j), S_S(x_i,x_j)),\\ \overline{\mathscr{B}}(x_i) &= \sup_{x_j\in U}\min(\mathbf{B}(x_i,x_j), S_S(x_i,x_j)).\end{aligned} \tag{9}$$

$\underline{\mathscr{B}}$ and $\overline{\mathscr{B}}$ are fuzzy sets induced by $\mathbf{B}$, which satisfy $\underline{\mathscr{B}}(x_i)\subseteq\mathbf{B}(x_i)\subseteq\overline{\mathscr{B}}(x_i)$ for each $x_i\in U$. $\underline{\mathscr{B}}(x_i)$ is the certainty degree of x_i associating to the fuzzy set and $\overline{\mathscr{B}}(x_i)$ is the plausibility degree of x_i associating to the fuzzy set. The larger the $\underline{\mathscr{B}}(x_i)$ or/and the smaller the $\overline{\mathscr{B}}(x_i)$, the stronger the performance of subset $\mathscr{B}$.

Moreover, the lower and upper approximations satisfy monotonicity, i.e., the lower (upper) approximation monotonically increases (decreases) with the increasing of fuzzy relations. This can raise the definition of reduct, which is exactly a subset of fuzzy relations that achieves the same approximation ability as the whole set.

Definition 4. Let $FS = (U, \mathscr{R}, S_S)$ a fuzzy decision information system with $\mathscr{B}\subseteq\mathscr{R}$. $\mathscr{B}$ is referred to as a consistent set iff $\underline{\mathscr{B}}(x_i) = \underline{\mathscr{R}}(x_i)$ for each $x_i\in U$. Furthermore, if $\mathscr{B}$ is a consistent set and any $\mathscr{B}\prime\subseteq\mathscr{B}$ is not any more, then $\mathscr{B}$ is a reduct of S.

In Definition 4, due to the monotonicity of lower approximations, there always exist minimal subsets of fuzzy relations that constitute the reducts of a fuzzy decision information system. For simplicity, denote $\lambda_i = \underline{\mathscr{R}}(x_i)$ in all the following contents. We can see that λ_i is the fuzzy degree of the i-th sample belonging to the lower approximation w.r.t. the whole set of fuzzy relations.

We next present the basic structures of reducts by introducing the notion of discernibility set.

Definition 5. Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system. Define the discernibility set of each $(x_i,x_j)\in U\times U$ as

$$c_{ij}=\begin{cases}\{R\in\mathscr{R}|1-R(x_i,x_j)\geqslant\lambda_i\}, & S_S(x_i,x_j)<\lambda_i;\\ \varnothing, & else.\end{cases}$$

Definition 5 explicitly indicates that, if the label-level similarity of two samples is less than the confidence degree induced by the fuzzy rough lower approximation, then some related features need to be selected to distinguish the two samples at the feature level.

In previous literatures [28,30,31], the fuzzy rough discernibility set is constructed when the decision space takes categorical values; while in Definition 6, the items in the decision space and the sample space both take fuzzy values, which is a generalized case and can deal with multi-label data.

The discernibility set defined in Definition 6 can be used to elaborate the reduction principle and to construct the reducts as follows.

Theorem 2. *Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system. For $\mathscr{B}\subseteq\mathscr{R}$, the following statements are equivalent:*

(1) *$\mathscr{B}$ is a consistent set of S;*
(2) *$\mathscr{B}\cap c_{ij}\neq\varnothing$ for any $c_{ij}\neq\varnothing$;*

Proof. (1)⇒(2): Since $\mathscr{B}$ is a consistent set, it implies $\underline{\mathscr{B}}(x_i)=\lambda_i$. We now assume $c_{ij}\neq\varnothing$.

We have $S_S(x_i,x_j)<\lambda_i$, and there is at least one $R\in\mathscr{R}$ such that $1-R(x_i,x_j)\geqslant\lambda_i$.

Suppose by contradiction that $\mathscr{B}\cap c_{ij}=\varnothing$. It follows that $1-R(x_i,x_j)<\lambda_i$ for any $R\in\mathscr{B}$. Hence, $1-(\cap\mathscr{B})(x_i,x_j)<\lambda_i$, which equals to $1-\mathbf{B}(x_i,x_j)<\lambda_i$. It must hold that $S_S(x_i,x_j)\geqslant\lambda_i$. It is a contradiction to $S_S(x_i,x_j)<\lambda_i$.

(2)⇒(1): We need to prove $\underline{\mathscr{B}}(x_i)=\lambda_i$. It is obvious that $\underline{\mathscr{B}}(x_i)\leqslant\lambda_i$ for $\mathscr{B}\subseteq\mathscr{R}$. We next prove $\underline{\mathscr{B}}(x_i)\geqslant\lambda_i$.

Suppose by contradiction that $\underline{\mathscr{B}}(x_i)<\lambda_i$. With the fact of $\underline{\mathscr{B}}(x_i)=\inf\limits_{x_j\in U}\max(1-\mathbf{B}(x_i,x_j),S_S(x_i,x_j))$, there is some $x_j\in U$ such that $1-\mathbf{B}(x_i,x_j)<\lambda_i$ and $S_S(x_i,x_j)<\lambda_i$. It can be verify that $1-\mathbf{R}(x_i,x_j)\geqslant\lambda_i$. Hence, there is some $R\in\mathscr{R}$ such that $1-R(x_i,x_j)\geqslant\lambda_i$. This implies $c_{ij}\neq\varnothing$. Combining with $1-\mathbf{B}(x_i,x_j)<\lambda_i$, we have $\mathscr{B}\cap c_{ij}=\varnothing$. This is a contradiction.

We complete the proof. □.

Corollary 1. *Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system. Then, $\mathscr{B}\subseteq\mathscr{R}$ is a reduct iff $\mathscr{B}$ is a minimal subset satisfying $\mathscr{B}\cap c_{ij}\neq\varnothing$ for any $c_{ij}\neq\varnothing$.*

Proof. The proof is direct from Theorem 2. □.

After obtaining the discernibility sets by Definition 6, a discernibility function can be defined as $f(\mathscr{R})=\wedge\{\vee(c_{ij})|c_{ij}\neq\varnothing\}$, where $\vee(c_{ij})$ is the disjunction of all elements in c_{ij}, and $\wedge\{\vee(c_{ij})\}$ is the conjunction of all $\vee(c_{ij})$. By transforming the discernibility function into a reduced disjunctive form, the reducts can be constructed as shown in following theorem.

Theorem 3. *Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system and $f(\mathscr{R})$ the discernibility function of FS. Then, $\mathscr{B}=\{R_1,R_2,\cdots,R_p\}\subseteq\mathscr{R}$ is a reduct iff the conjunction $R_1\wedge R_2\cdots\wedge R_p$ is a prime implicant of $f(\mathscr{R})$.*

Proof. The proof follows from Corollary 1. □.

Corollary 1 indicates that a reduct shares at least one element in common with each of the nonempty discernibility sets. In this sense, the fuzzy relations with larger coverage on the discernibility sets are relative significant. Moreover, it is expected to select a subset of fuzzy relations that can minimize the intra-class diversity and maximize the inter-class diversity. Based on these ideas, we employ all the lower approximation values λ_i as confidence levels to intercept indiscernible sample pairs. Consequently, we define

$$IDS(\mathscr{R})=\{(x_i,x_j)\in U\times U|S_S(i,j)\geqslant\lambda_i\}.$$

We can see that $IDS(\mathscr{R})$ contains the pairs of samples those are near and cannot be discerned at the level of λ_i. For a given sample x_i, denote $IDS(\mathscr{R})(x_i)=\{x_j\in U|(x_i,x_j)\in IDS(\mathscr{R})\}$. We further obtain that $IDS(\mathscr{R})(x_i)=\{x_j\in U|S_S(i,j)\geqslant\lambda_i\}$. Intuitively speaking, the samples in $IDS(\mathscr{R})(x_i)$ have large similarity with x_i and can be seen as the near neighbors of x_i in the label space.

Definition 6. Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system and $M=(c_{ij})$ be the discernibility matrix with $R\in\mathscr{R}$. A discernibility relation induced by $R\in\mathscr{R}$ is defined as $DS(R)=\{(x_i,x_j)\in U\times U|R\in c_{ij}\}$.

In Definition 6, if $(x_i,x_j)\in DS(R)$, then x_j can be discerned with x_i by relation R. Given any $\mathscr{B}\subseteq\mathscr{R}$, denote $DS(\mathscr{B})=\cup_{R\in\mathscr{B}}DS(R)$. We arrive at the following property.

Property 2. Let $FS=(U,\mathscr{R},S_S)$ be a fuzzy decision information system with $\mathscr{B}\subseteq\mathscr{R}$. Then,

(1) $IDS(\mathscr{R})=U\times U-DS(\mathscr{R})$;
(2) $\mathscr{B}$ is a consistent set iff $DS(\mathscr{B})=DS(\mathscr{R})$.

Proof. It is not hard to be proved from Theorem 2. □.

Algorithm 2: An algorithm for computing all the discernibility relations

Input: A fuzzy decision information system $FS=(U,\mathscr{R},S_S)$.
Output: The discernibility relations $DS(R)$ for all $R\in\mathscr{R}$.
For Each $R\in\mathscr{R}$
  Initialize $DS(R)\leftarrow\varnothing$;
End For
For Each $x_i\in U$
  Set $\mathbf{R}=\cap\mathscr{R}$ in Definition 3 and compute $\lambda_i=\underline{\mathscr{R}}([x_i]_d)(x_i)$;
  For Each $x_j\in U$ satisfying $S_S(i,j)<\lambda_i$
    For Each $R\in\mathscr{R}$
      If $1-R(x_i,x_j)\geqslant\lambda_i$
        Let $DS(R)\leftarrow DS(R)\cup\{(x_i,x_j)\}$;// compute the discernibility sample pairs;
      End If
    End For
  EndFor
End For
Output All $DS(R)$.
End

Algorithm 2 provides a detailed process for calculating all the discernibility relations induced by the fuzzy relations. A fuzzy relation is considered to be significant if it can discriminate as many

discernibility sample pairs as possible. Moreover, the margin between intra-class similarity and inter-class similarity is an important index for evaluating the discrimination ability of a fuzzy relation. Following these ideas, Algorithm 3 is introduced to select a discriminative subset of fuzzy relations in a fuzzy decision information system.

Algorithm 3: An algorithm for finding a reduct of a fuzzy decision information system

Input: A fuzzy decision information system $FS = (U, \mathscr{R}, S_S)$.
Output: One reduct $\mathscr{B}$.
1: Initialize $\mathscr{B} \leftarrow \varnothing$;
2: Compute $DS(R)$ for all $R \in \mathscr{R}$ by Algorithm 2;
3: Compute $DS(\mathscr{R}) = \cup_{R\in\mathscr{R}} DS(R)$ and $IDS(\mathscr{R}) = U \times U - DS(\mathscr{R})$;
4: **For** Each $R_k \in \mathscr{R} - \mathscr{B}$
5: Initialize its significance as $Sig(R_k) \leftarrow 0$;
6: Let $\mathscr{B}_k \leftarrow \mathscr{B} \cup \{R\}$;
7: **For** Each $x_i \in U$
8: Compute the average similarity between x_i and its near neighbors: $\mathrm{Aver}(x_i) = \frac{\sum\{\mathbf{B_k}(x_i,x_j)|x_j\in IDS(\mathscr{R})(x_i)\}}{|IDS(\mathscr{R})(x_i)|}$;
9: **For** Each x_j satisfying $(x_i, x_j) \in DS(\mathscr{R})$
10: **If** $\mathbf{B_k}(x_i, x_j) < \mathrm{Aver}(x_i)$
11: Update $Sig(R) \leftarrow Sig(R) + 1$;
12: **End If**
13: **End For**
14: **End For**
15: **End For**
16: Find some $R_0 \in \mathscr{R} - \mathscr{B}$ maximizing Sig and update $\mathscr{B} \leftarrow \mathscr{B} \cup \{R_0\}$;
17: If $DS(\mathscr{B}) = DS(\mathscr{R})$, terminate the algorithm; otherwise go to Step 4;
18: **Output** $\mathscr{B}$.
19: **End**

In Algorithm 3, Step 8 computes the average similarity between each sample and its near neighbors. Steps 9–13 compute the significance of each remaining fuzzy relation one by one by adding it to the current reduct. Step 9 evaluates whether or not the fuzzy relation can discriminate the discernible sample pairs, whereas Step 10 evaluates each candidate relation that whether or not the fuzzy relation can separate the inter-class samples from the intra-class samples. In the end, Step 17 is the termination condition, which guarantees that the selected subset can cover all the discernible sample pairs.

In Algorithm 2, The time complexity of computing the fuzzy rough lower approximations in Step 5 is $O\left(|U|^2\right)$, and the time complexity of computing the discernibility sample pairs in Step 9 is $O\left(|U|^2|\mathscr{R}|\right)$. In Algorithm 3, the time complexity of Step 3 is $O(|\mathscr{R}|)$ and Step 8 can be done within $O(|U|)$ in each loop. Moreover, the computation of the significance of the fuzzy relations in Step 11 can be done within $O\left(|U|^2|\mathscr{R}|\right)$. In summarize, the total time complexity of Algorithms 2 and 3 is $O\left(|U|^2|\mathscr{R}|\right)$.

4. Experiments

4.1. Experiment preparation

In this section, we perform a sequence of experiments to demonstrate the effectiveness of the proposed algorithm (FRD) by comparing it with current state-of-art algorithms on some public data sets. The multi-label data sets used for comparison are downloaded from (http://mulan.sourceforge.net/datasets.html [44]), which are outlined in Table 2. The features in each data set have a variety of characteristics-some binary/discrete and some continuous. The continuous features are normalized into [0,1] and discrete features are converted to distinct values.

The sample similarity matrix S_S evaluates the pairwise similarity of samples. For the sake of simplicity, we cut the matrix S_S by using top-k nearest neighbors as follows:

$$S_S(i,j) = \begin{cases} S_S(i,j), & x_j \in \mathscr{N}_k(x_i) \text{ or } S_S(i,j) = 1; \\ 0, & \text{otherwise}. \end{cases}$$

where $\mathscr{N}_k(x_i)$ is the set of the top-k nearest neighbors of x_i in the label space according to the sample similarity matrix S_S. The parameter k is set to 10 in the substantial implementation in the following. After these preparations, Algorithms 2 and 3 can be implemented to select of a subset of features.

4.2. Comparison methods

We compare the proposed algorithm against some multi-label feature selection algorithms, including MDDMspc [46], MDDMproj [46], MLNB [47], MCLS [13], MIFS [49], and PMU [5]. The ideas of the six algorithms for comparison are listed below.

MDDM: It contains two substantial algorithms, i.e., MDDMspc and MDDMproj, which adopt different types of strategies to project the original data into a lower-dimensional feature space by maximizing the dependence between the original feature space and the

Table 2
Description of multi-label data sets.

No.s	Data sets	Instances	Training	Test	Features	Labels	Domain
1	Birds	645	322	323	260	19	Audio
2	Business	5000	2000	3000	438	30	Text
3	Computers	5000	2000	3000	681	33	Text
4	Education	5000	2000	3000	550	33	Text
5	Emotions	593	391	202	72	6	Music
6	Entertainment	5000	2000	3000	640	21	Text
7	Health	5000	2000	3000	612	32	Text
8	Recreation	5000	2000	3000	606	22	Text
9	Reference	5000	2000	3000	793	33	Text
10	Scene	2407	1211	1196	294	6	Image
11	Science	5000	2000	3000	743	40	Text
12	Society	5000	2000	3000	636	27	Text
13	Yeast	2417	1499	918	103	14	Biology
14	Bibtex	7395	4880	2515	1836	159	Text
15	Slashdot	3782	2546	1236	1079	22	Text

class labels. MLNB: It adapts the traditional naive Bayes classifiers to deal with multi-label samples.

MCLS: It employs manifold learning to transform logical label space to Euclidean label space.

MIFS: It exploits label correlations to select discriminative features across multiple labels.

PMU: It evaluates features by maximizing mutual information between selected features and labels.

The configuration parameters of the six algorithms for comparison are suggested by their original literature, which are also listed in [22] in detail. ML-KNN [50] ($K = 10$) is employed as the classifier to evaluate the performance of the feature subset searched by the algorithms. To fully reveal the effectiveness of the proposed method, we randomly divide each data set into training and test parts for five times. As the validation is iterated five times, we obtain five results and record the average result on each data set.

4.3. Experimental results

4.3.1. Evaluation metrics

We get the predictive classification performances of all comparison algorithms on the sequentially selected top-50%. Tables 3–8 record the experimental results of different multi-label feature selection methods in terms of the six evaluation metrics. Six

Table 3
Comparison results of multi-label feature selection methods in terms of *Hamming Loss* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	**0.0550**	0.0647	0.0657	0.0662	0.0607	0.0551	0.0687
Business	**0.0272**	0.0277	0.0276	0.0287	0.0278	0.0276	0.0275
Computers	0.0413	0.0435	0.0436	0.0434	0.0430	0.0422	**0.0412**
Education	0.0420	0.0432	0.0426	0.0429	0.0442	0.0423	**0.0412**
Emotions	**0.2336**	0.2402	0.2635	0.2384	0.3284	0.2452	0.2676
Entertainment	**0.0605**	0.0642	0.0629	0.0648	0.0650	0.0614	0.0623
Health	0.0437	0.0475	0.0452	0.0473	0.0487	0.0440	**0.0433**
Recreation	**0.0611**	0.0649	0.0651	0.0648	0.0650	0.0636	0.0650
Reference	**0.0300**	0.0334	0.0322	0.0347	0.0344	0.0312	0.0303
Scene	**0.1208**	0.1549	0.1753	0.1606	0.1552	0.1452	0.1238
Science	**0.0340**	0.0356	0.0354	0.0353	0.0356	0.0346	0.0352
Society	**0.0565**	0.0592	0.0594	0.0589	0.0601	0.0593	0.0571
Yeast	**0.2064**	0.2233	0.2226	0.2226	0.2119	0.2122	0.2117
Bibtex	**0.0537**	0.0545	0.0545	0.0547	0.0545	0.0584	0.0547
Slashdot	**0.0540**	0.0542	0.0542	0.0543	0.0546	0.0547	0.0545

Table 4
Comparison results of multi-label feature selection methods in terms of *Ranking Loss* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	0.1381	0.1420	0.1430	0.1465	0.1480	**0.1382**	0.1364
Business	**0.0412**	0.0419	0.0432	0.0445	0.0433	0.0415	0.0414
Computers	0.0981	0.1029	0.1039	0.1028	**0.0967**	0.1015	0.0994
Education	0.0960	0.1008	0.0971	0.0992	0.1069	0.0972	**0.0933**
Emotions	**0.1908**	0.2090	0.2310	0.2048	0.4147	0.2053	0.2563
Entertainment	**0.1213**	0.1327	0.1315	0.1367	0.1407	0.1249	0.1287
Health	**0.0632**	0.0735	0.0681	0.0718	0.0773	0.0657	0.0637
Recreation	**0.1912**	0.2119	0.2159	0.2121	0.2114	0.2057	0.2151
Reference	**0.0896**	0.0981	0.0973	0.0983	0.0981	0.0939	0.0900
Scene	**0.1304**	0.2494	0.3058	0.2583	0.2115	0.1836	0.1444
Science	**0.1407**	0.1497	0.1515	0.1492	0.1512	0.1381	0.1411
Society	0.1500	0.1563	0.1574	0.1550	0.1601	0.1581	**0.1476**
Yeast	**0.1822**	0.2023	0.2003	0.1996	0.1853	0.1867	0.1843
Bibtex	**0.1041**	0.1052	0.1047	0.1053	0.1063	0.1059	0.1045
Slashdot	0.2022	0.2117	0.2035	0.2152	0.2171	0.2084	**0.2020**

Table 5
Comparison results of multi-label feature selection methods in terms of *Coverage* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	3.6900	3.8669	3.5944	**3.4396**	3.4768	3.6749	3.6625
Business	**2.3015**	2.3407	2.3567	2.4110	2.3907	2.3213	2.3183
Computers	4.6556	4.8351	4.8650	4.8318	**4.5720**	4.7916	4.7008
Education	4.0864	4.2092	4.0884	4.1657	4.4467	4.1000	**3.9655**
Emotions	**2.0177**	2.1289	2.2365	2.0873	3.0691	2.1152	2.3276
Entertainment	**3.3290**	3.5070	3.4768	3.5852	3.6684	3.3379	3.4275
Health	**3.4112**	3.7745	3.5883	3.7161	3.9084	3.4789	3.4141
Recreation	**5.0805**	5.4935	5.5735	5.5013	5.4738	5.3753	5.5653
Reference	**3.4553**	3.7404	3.7079	3.7472	3.7427	3.6051	3.4679
Scene	**0.7576**	1.3505	1.6367	1.3952	1.1655	1.0264	0.8238
Science	**6.9032**	7.4565	7.4985	7.4315	7.4831	6.9418	7.0459
Society	**5.8026**	6.0466	6.0869	6.0044	6.1678	6.0986	5.8058
Yeast	**6.5610**	6.7942	6.6983	6.8061	6.6082	6.6346	6.5941
Bibtex	13.9636	14.2850	14.2435	14.2936	14.2971	13.9931	**13.2457**
Slashdot	**4.7618**	4.9296	4.7621	5.0162	5.0526	4.8568	4.7623

Table 6
Comparison results of multi-label feature selection methods in terms of *AveragePrecision* (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	**0.7014**	0.6439	0.6378	0.6303	0.6541	0.6805	0.6244
Business	**0.8789**	0.8743	0.8729	0.8654	0.8714	0.8736	0.8737
Computers	**0.6142**	0.6010	0.6001	0.6011	0.6112	0.6057	0.6113
Education	0.5414	0.5056	0.5256	0.5176	0.4859	0.5285	**0.5448**
Emotions	**0.7710**	0.7564	0.7332	0.7585	0.6022	0.7562	0.7185
Entertainment	**0.5678**	0.5326	0.5403	0.5232	0.5129	0.5624	0.5482
Health	0.6753	0.6407	0.6603	0.6450	0.6304	0.6753	**0.6756**
Recreation	**0.4670**	0.3946	0.3854	0.3934	0.3923	0.4197	0.3881
Reference	**0.6150**	0.5835	0.5854	0.5825	0.5850	0.6084	0.6097
Scene	**0.7921**	0.6593	0.5932	0.6462	0.6874	0.7165	0.7791
Science	**0.4571**	0.4175	0.4143	0.4262	0.4073	0.4568	0.4455
Society	**0.5815**	0.5579	0.5546	0.5594	0.5495	0.5589	0.5810
Yeast	**0.7463**	0.7199	0.7214	0.7211	0.7400	0.7402	0.7417
Bibtex	**0.6118**	0.5839	0.6038	0.5835	0.5894	0.5603	0.5800
Slashdot	**0.3890**	0.3875	0.3885	0.3724	0.3707	0.3781	0.3888

Table 7
Comparison results of multi-label feature selection methods in terms of $Macro-F1$ (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	**0.0964**	0.0749	0.0808	0.0740	0.0858	0.0954	0.0910
Business	0.1453	0.1358	**0.1498**	0.1115	0.1279	0.1581	0.0859
Computers	0.0458	0.0224	0.0208	0.0238	**0.0521**	0.0312	0.0439
Education	**0.1061**	0.0933	0.1062	0.1005	0.0726	0.1052	0.0659
Emotions	**0.5601**	0.5477	0.4644	0.5338	0.1412	0.5130	0.4908
Entertainment	**0.1178**	0.0591	0.0704	0.0486	0.0346	0.1160	0.0975
Health	**0.1967**	0.1408	0.1771	0.1502	0.1052	0.1824	0.1362
Recreation	**0.0776**	0.0167	0.0051	0.0179	0.0132	0.0375	0.0073
Reference	**0.1174**	0.0869	0.0893	0.0872	0.0900 0.1161	0.0442	
Scene	**0.5878**	0.3103	0.1516	0.2826	0.3611	0.4119	0.5778
Science	0.0413	0.0104	0.0122	0.0136	0.0085	**0.0437**	0.0274
Society	**0.0524**	0.0270	0.0215	0.0270	0.0172	0.0260	0.0514
Yeast	**0.3258**	0.2410	0.2339	0.2386	0.3007	0.3014	0.3049
Bibtex	**0.754**	0.0608	0.0619	0.0602	0.0521	0.0706	0.0751
Slashdot	0.1288	0.1029	**0.1297**	0.1008	0.0970	0.0937	0.1128

Table 8
Comparison results of multi-label feature selection methods in terms of $Micro-F1$ (mean).

Data sets	FRD	MDDMproj	MLNB	MDDMspc	MCLS	MIFS	PMU
Birds	**0.4638**	0.3718	0.3614	0.3580	0.3753	0.4580	0.3537
Business	**0.6983**	0.6844	0.6909	0.6691	0.6799	0.6927	0.6849
Computers	0.3559	0.3348	0.3421	0.3367	0.3495	0.3517	**0.3597**
Education	0.1543	0.0759	0.1186	0.1086	0.0140	0.1357	**0.1951**
Emotions	**0.5937**	0.5838	0.5126	0.5782	0.1967	0.5722	0.5178
Entertainment	**0.2305**	0.1366	0.1660	0.1091	0.0983	0.2234	0.1878
Health	0.4162	0.3371	0.3963	0.3533	0.3021	0.4159	**0.4260**
Recreation	**0.1441**	0.0228	0.0060	0.0251	0.0185	0.0666	0.0105
Reference	**0.3306**	0.2466	0.2744	0.2873	0.2707	0.3138	0.3235
Scene	**0.5980**	0.3260	0.1657	0.2980	0.3829	0.4472	0.5833
Science	**0.1085**	0.0289	0.0518	0.0547	0.0217	0.1077	0.0665
Society	**0.2677**	0.1990	0.2181	0.1968	0.1749	0.1927	0.2667
Yeast	**0.6116**	0.5631	0.5556	0.5588	0.5932	0.5900	0.5949
Bibtex	**0.4635**	0.4220	0.4489	0.4210	0.3896	0.3829	0.4623
Slashdot	0.0389	0.0212	**0.0416**	0.0186	0.0159	0.0067	0.0302

evaluation metrics, including *Hamming Loss*, *Ranking Loss*, *Coverage*, *Average Precision*, $Macro-F1$, and $Micro-F1$, are employed to examine the performances of the selected feature subsets obtained by different multi-label feature selection algorithms. For *Hamming Loss*, *Ranking Loss*, and *Coverage*, the smaller the value, the better the performance; whereas for *Average Precision*, $Macro-F1$, and $Micro-F1$, the larger the value is, the better the performance is. To fully reveal the effectiveness of the proposed method,

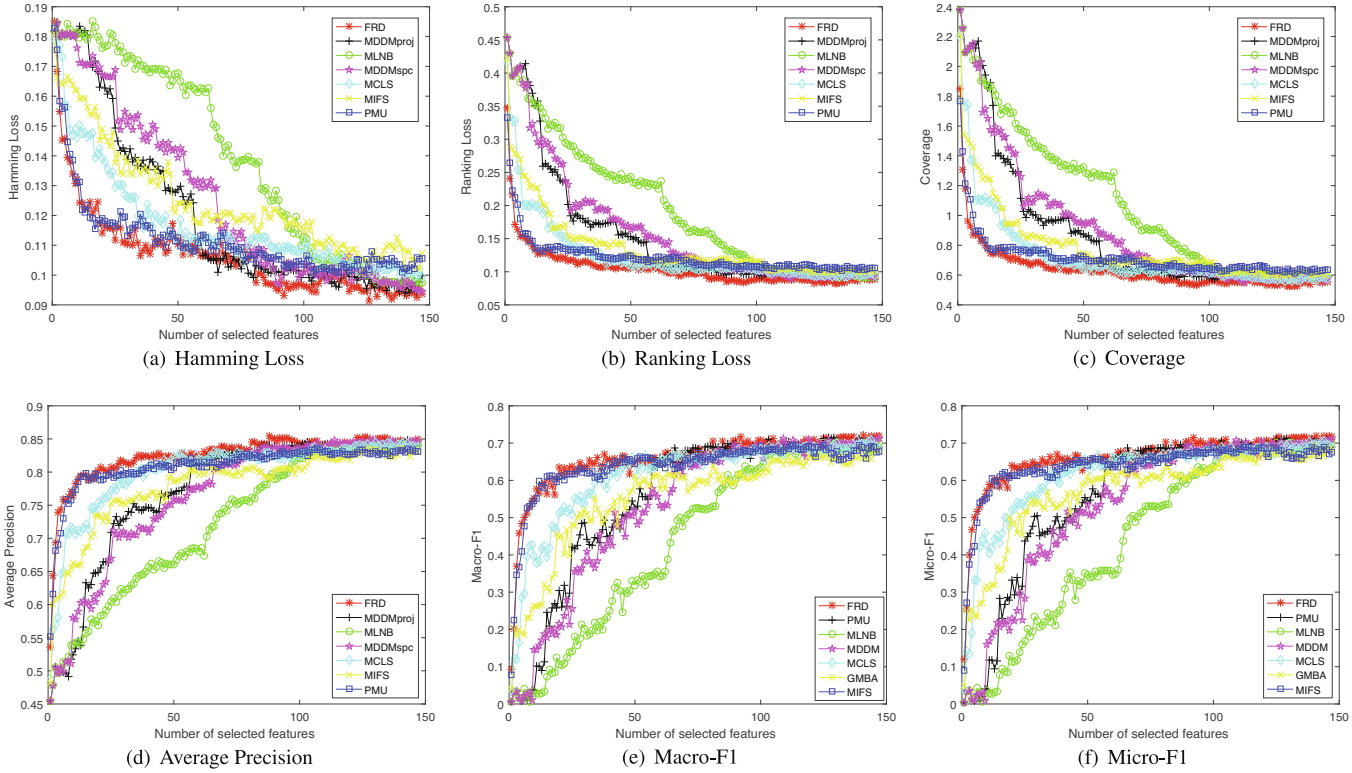
4.3.2. Overall observation of experimental results

Based on the results reported in these tables, we have a couple of observations: (1) Algorithm FRD performs almost the best one on most of the data sets. (2) For *Hamming Loss* and *Average Precision*, FRD obtains the best performance at least on twelve multi-label data sets, and is only worse than PMU in some cases. (3) For $Macro-F1$, and $Micro-F1$, the predictive classification performance of FRD is much better than MDDMspc, MDDMproj, MLNB, MCLS, and MIFS at least on ten multi-label data

sets. Meanwhile, the predictive classification performance of FRD is also extremely close to the performance of PMU on the other multi-label data sets.

4.3.3. Detailed experimental results

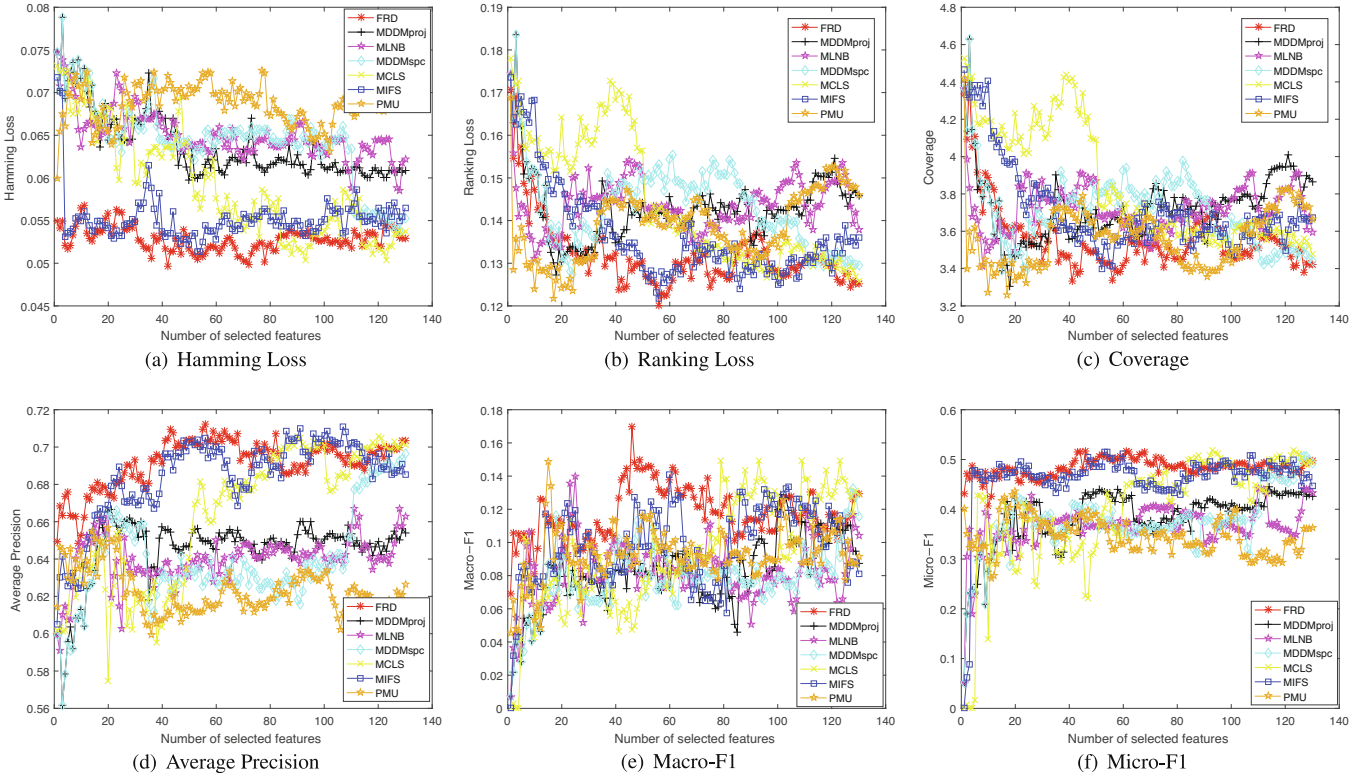
We below compare the performance of the selection algorithms on each data set in detail. Table 3 reveals the experimental results



(a) Hamming Loss (b) Ranking Loss (c) Coverage

(d) Average Precision (e) Macro-F1 (f) Micro-F1

Fig. 1. Comparison results of different comparing algorithms on Scene.



(a) Hamming Loss (b) Ranking Loss (c) Coverage

(d) Average Precision (e) Macro-F1 (f) Micro-F1

Fig. 2. Comparison results of different comparing algorithms on Birds.

Table 9
Summary of the Friedman statistics F_F $(K = 7, N = 15)$ and the critical value with different evaluation metrics

Metrics	F_F	Critical value ($\alpha = 0.05$)
Hamming loss	15.23	2.5836
Ranking Loss	12.62	
Coverage	13.23	
Average precision	14.35	
Macro-F1	12.69	
Micro-F1	15.98	

of multi-label feature selection methods in terms of *HammingLoss*. In this table, Algorithm FRD achieves superior results against the comparing methods on 12 out of 15 dada sets. Meanwhile, on the other 3 data sets Computers, Education and Health, FRD is inferior but approximate to PMU and outperforms the other algorithms.

Table 4 reveals the selection results in terms of *RankingLoss*. As shown in the table, FRD achieves the best performance 10 times over the 15 dada sets compared with all the other algorithms. On the other 3 data sets, FRD is inferior to PMU for 3 times and is inferior to MCLS and MIFS for 2 times. The average loss indicates that FRD is significantly superior to all the other algorithms with a large margin.

We can see the similar results on the remaining evaluation metrics. With each of the metrics, FRD offers the most evident improvement on nearly 75% of the data sets when compared with the other algorithms. Furthermore, the average performance recorded in each table shows that FRD constantly ranks 1 with a larger margin compared with other methods.

For a more intuitive display, Figs. 1 and 2 also depict the variation diagram of the performance w.r.t. the selected features on two data sets with different scales which are collected from different domains i.e., Scene and Birds. As shown Fig. 2, FRD significantly outperforms the comparing algorithms in terms of each of the evaluation metrics *Ranking Loss*, *Coverage*, $Macro - F1$, and $Micro - F1$, but it is inferior to the algorithms PMU, MDDMproj, and MDDMspc in terms of *Hamming Los*.

4.3.4. Performance analysis

The above phenomenons indicate that the proposed method has the overall advantages and can achieve good performances in most cases, whereas some comparing algorithms are also capable to handle some special tasks from different domains. The success of the proposed algorithm relies on the ability to disclose the instance-level compact and sufficient information and to preserve the consistency between the information from the feature space and the label space.

4.3.5. Performance analysis in terms of hypothesis tests

For further illustration, Friedman test [51] and Bonferroni-Dunn test [52] are utilized to evaluate the statistical significance of the comparing algorithms. let r_j be the average rank of algorithm j

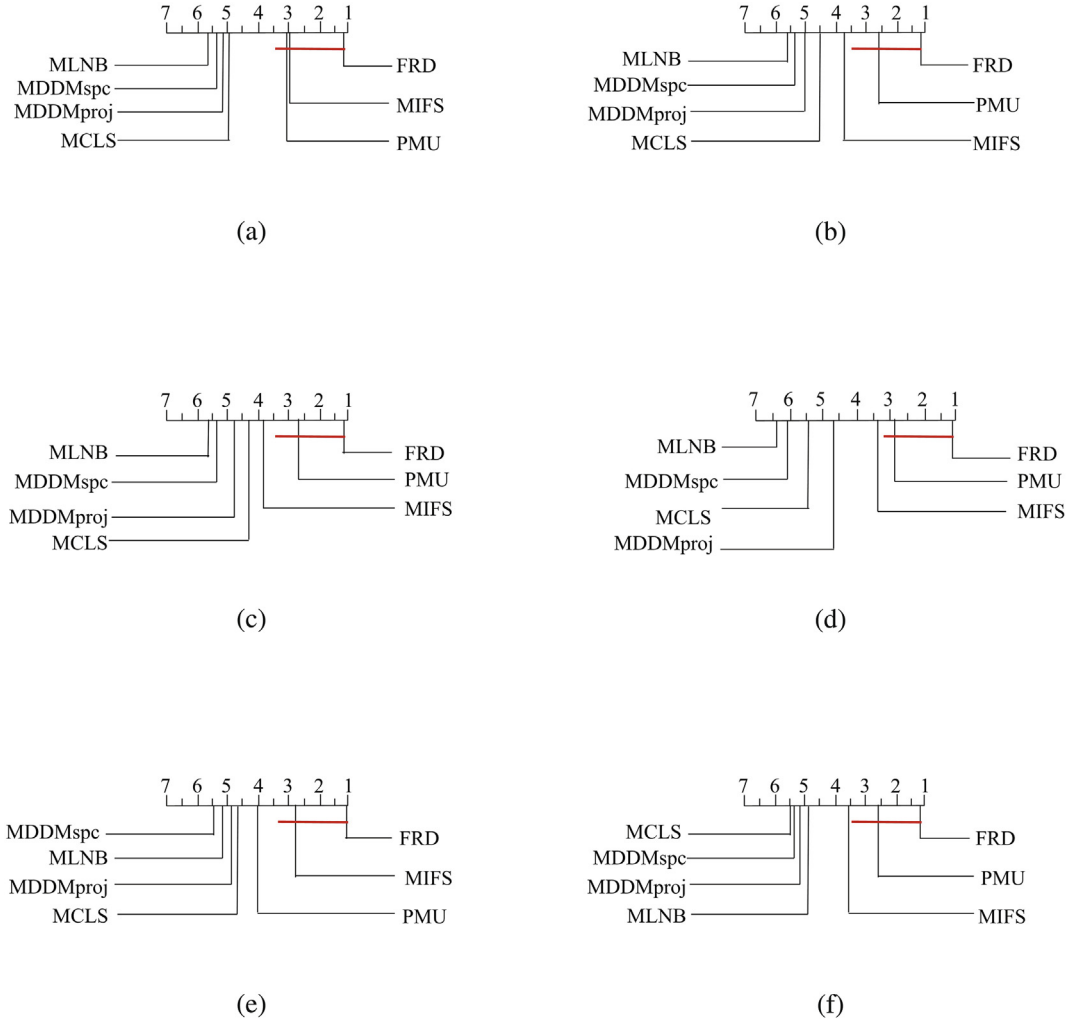
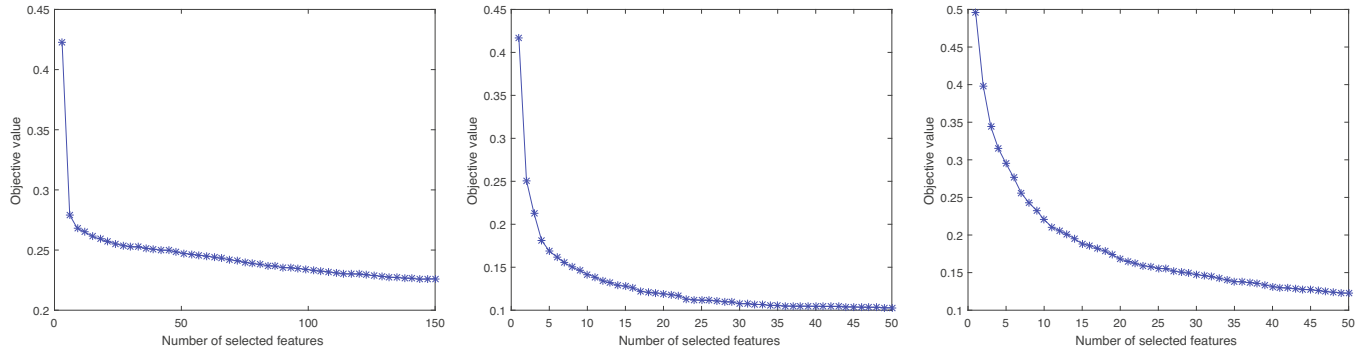


Fig. 3. Comparison of FRD against algorithms under comparison with the Bonferroni-Dunn test in terms of (a) Hamming Loss, (b) Ranking Loss, (c) Coverage, (d) Average Precision, (e) Macro-F1, (f) Micro-F1.



(a) Birds (b) Emotions (c) Yeast

Fig. 4. Convergence analysis. The objective value for misclassifying inter-class sample pairs with Algorithm FRD w.r.t. the number of iteration.

on all data sets, N the number of multi-label data sets, and K the number of multi-label feature selection algorithms. Under the null-hypothesis, Friedman statistic F_F follows a Fisher distribution:

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

where

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

We can further obtain the Friedman statistic F_F on the six evaluation metrics and the critical value shown in Table 9. As seen in the table, the null hypothesis, that the performance of all methods is equivalent, is clearly rejected on each evaluation metric at significance level $\alpha = 0.05$. Then, Bonferroni-Dunn test [52] is used to further evaluate the performance of the methods. The performance of two methods is regarded as different, if their average ranks exceeds the critical distance:

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

where $q_\alpha = 2.638$ is the critical value of the test. We can calculate that $CD_\alpha = 2.0809$ ($\#$ datasets $N = 15, \#$ comparing algorithms $K = 7$). Fig. 3 presents the CD diagrams on each metric, where the red line of one CD is represented by the red line. Two comparing methods are considered to have no significant difference if their average ranks are connected within one CD. We can see from Fig. 3 that FRD is significantly superior to MDDMproj, MLNB, MDDMspc, and MCLS on *Hamming Loss*, *Ranking Loss*, *Coverage*, *Macro* $-$ $F1$, and *Micro* $-$ $F1$. Furthermore, FRD outperforms MIFS on *Micro* $-$ $F1$, and outperforms PMU on *Macro* $-$ $F1$. Hence, the proposed algorithm can achieve competitive performance and is advisable to be as an alternative version of feature selection for multi-label data sets than some existing methods.

4.3.6. Convergence analysis

We below analysis the convergence of the proposed algorithm. As shown in in Definition 3, the fuzzy rough approximations are monotonous with the size of the subset of features. More features selected indicates more inter-class sample pairs discriminated with the algorithm. The objective value of misclassifying inter-class sample pairs decreases monotonously with the size of selected features. Consequently, the convergence of the proposed method can be guaranteed. Fig. 4 (a)-(c) illustrate the change of the objective value w.r.t. the size of selected features on three data sets, including Birds, Emotions, and Yeast. In the figures, it is demonstrated that FRD falls fast within a small number of selected features and then tends towards stability. Hence, the results empirically verify the convergence of the proposed algorithm in practice.

5. Conclusion

The approaches of discernibility matrix provide a mathematical foundation for fuzzy rough set-based data analysis. However, a way how to effectively realize the superiority for dealing with multi-label data sets, which generate more than one groups of class label partitions, has not been systematically studied so far. In this paper, we introduced a multi-label feature selection algorithm based on fuzzy rough discrimination. Note that different labels may provide different/repetitive sample distribution information. To make proper use of the label information, a weighting method of label significance was proposed based on label correlation. Then, a sample similarity matrix in the label space was computed via label weighting. The discernibility matrix of fuzzy rough set for discriminating related sample pairs was constructed. Based on the fuzzy rough discrimination, the procedure was implemented to evaluate and select discriminative features in multi-label data. Moreover, the margin preservation for discriminating different classes' samples was also considered in our algorithm construction.

There are still some problems that deserve further consideration. For example, 1) The relevance between candidate features and the selected feature subset was calculated by the proposed forward searching algorithm, which was the main step of time consuming. The selection process will be improved for dealing with large-scale data in our future study. 2) The intersection operation was adopted to generate the fuzzy relation w.r.t. a given subset of features. This method may cause a deviation when characterizing sample similarity over a subset of features. Hence, a generalized framework of fuzzy information system will be introduced to overcome this limitation. 3) The idea of fuzzy rough discrimination will be applied to handle the multi-label classification learning with noise labels.

CRediT authorship contribution statement

Anhui Tan: Conceptualization, Software, Writing - original draft. **Jiye Liang:** Methodology, Supervision. **Wei-Zhi Wu:** Formal analysis. **Lin Sun:** Software, Validation. **Chao Chen:** 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.

Acknowledgements

This work was supported by National Key Research and Development Program of China (No. 2020AAA0106100), National Natural Science Foundation of China (Nos. 61876103, 62076221 and 61976194), and the Natural Science Foundation of Zhejiang Province (No. LY18F030017).

References

[1] K. Barnard, P. Duygulu, D.A. Forsyth, N. de Freitas, D.M. Blei, M.I. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135.
[2] I. Katakis, G. Tsoumakas, and I. Vlahavas, Multilabel text classification for automated tag suggestion, in Proc. ECML/PKDD 2008 Discover. Challenge, Antwerp, Belgium, 2008, pp. 75-83..
[3] J. Lee, D.W. Kim, Mutual information-based multi-label feature selection using interaction information, Expert Syst. Appl. 42 (2015) 2013–2025.
[4] J. Lee, D.W. Kim, SCLS: Multi-label feature selection based on scalable criterion for large label set, Pattern Recognit. 66 (2017) 2989–3004.
[5] J. Lee, D.W. Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (2013) 349–357.
[6] J.H. Dai, J.L. Chen, J. Liu, H. Hu, Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation, Knowl.-Based Syst. 207 (2020) 106342.
[7] D.G. Kong, C. Ding, H. Huang, H.F. Zhao, Multi-label ReliefF and F-statistic feature selections for image annotation, Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 2352–2359..
[8] L. Sun, T.Y. Yin, W.P. Ding, Y.H. Qian, J.C. Xu, Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems, Inf. Sci. 537 (2020) 401–424.
[9] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1609–1614.
[10] J. Zhang, S.Z. Li, M. Zhang, K.C. Tan, Learning from weakly labeled data based on manifold regularized sparse model, IEEE Trans. Cybern. DOI: 10.1109/TCYB.2020.3015269, 2020 (In press)..
[11] C.Q. Zhang, Z.W. Yu, H.Z. Fu, P.F. Zhu, L. Chen, Q.H. Hu, Hybrid noise-oriented multilabel learning, IEEE Trans. Cybern. 50 (2020) 2837–2850.
[12] J. Huang, G.R. Li, Q.M. Huang, X.D. Wu, Learning label-specific features and class-dependent labels for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (2016) 3309–3323.
[13] R. Huang, W.D. Jiang, G.L. Sun, Manifold-based constraint laplacian score for multi-label feature selection, Pattern Recognit. Lett. 112 (2018) 346–352.
[14] J. Fürnkranz, E. Hllermeier, E.L. Mencła, K. Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2008) 133–153.
[15] Y. Yu, W. Pedrycz, D. Miao, Multi-label classification by exploiting label correlations, Expert Syst. Appl. 41 (2014) 2989–3004.
[16] Y. Zhu, J. Kwok, Z.H. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 30 (2017) 1081–1109.
[17] P. Zhang, G.X. Liu, W.F. Gao, Distinguishing two types of labels for multi-label feature selection, Pattern Recognit. 95 (2019) 72–82.
[19] F. Li, D.Q. Miao, W. Pedrycz, Granular multi-label feature selection based on multual information, Pattern Recognit. 67 (2017) 410–423.
[20] Y.J. Lin, Q.H. Hu, J.H. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.
[21] W.B. Qian, S.D. Yu, J. Yang, Y.L. Wang, J.H. Zhang, Multi-label feature selection based on information entropy fusion in multi-source decision system, Evol. Intell. 13 (2020) 255–268.
[22] J. Zhang, Z.M. Luo, C.D. Li, C.G. Zhou, S.Z. Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recogn. 95 (2019) 136–150.
[23] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.
[24] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, Intelligent Decision Support Theory and Decision Library 11 (1992) 331–362.
[25] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. Data Eng. 16 (2004) 1457–1471.
[26] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Trans. Fuzzy Syst. 15 (2007) 73–89.
[27] J.H. Dai, H. Hu, W.-Z. Wu, Y.H. Qian, D.B. Huang, Maximal discernibility pair based approach to attribute reduction in fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (2018) 2174–2187.
[28] C.Z. Wang, Y. Wang, M.W. Shao, Y.H. Qian, D.G. Chen, Fuzzy rough attribute reduction for categorical data, IEEE Trans. Fuzzy Syst. 28 (2020) 818–830.
[29] C.Z. Wang, Y. Huang, M.W. Shao, Q.H. Hu, D.G. Chen, Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (2020) 4031–4042.
[30] C.C. Tsang, D.G. Chen, S.D. Yueng, W.T. Lee, X.Z. Wang, Attribute reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (2008) 1130–1141.
[31] D.G. Chen, L. Zhang, S.Y. Zhao, Q.H. Hu, P.F. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2012) 385–389.
[32] Q.H. Hu, D.R. Yu, W. Pedrycz, D.G. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (2011) 1649–1667.
[33] Q.H. Hu, L.J. Zhang, Y.C. Zhou, W. Pedrycz, Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (2018) 226–238.
[34] X. Zhang, C.L. Mei, D.G. Chen, J.H. Li, Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[35] A.H. Tan, W.-Z. Wu, Y.H. Qian, J.Y. Liang, J.K. Chen, J.J. Li, Intuitionistic fuzzy rough set-based granular structures and attribute subset selection, IEEE Trans. Fuzzy Syst. 27 (2019) 527–539.
[36] A.H. Tan, S.W. Shi, W.-Z. Wu, J.J. Li, and W. Pedrycz, Granularity and entropy of intuitionistic fuzzy information and their applications, IEEE Trans. Cybern., DOI: 10.1109/TCYB.2020.2973379, 2020 (In press)..
[37] Y.W. Li, Y.J. Lin, J.H. Liu, W. Weng, Z.K. Shi, S.X. Wu, Feature selection for multi-label learning based on kernelized fuzzy rough sets, Neurocomputing 318 (2018) 271–286.
[38] Y.J. Lin, Y.W. Li, C.X. Wang, J.K. Chen, Attribute reduction for multi-label learning with fuzzy rough set, Knowl.-Based Syst. 152 (2018) 51–61.
[39] X.Y. Che, D.G. Chen, J.S. Mi, A novel approach for learning label correlation with application to feature selection of multi-label data, Inf. Sci. 512 (2020) 795–812.
[40] Y. Zhang, D. Shi, J. Gao, D. Cheng, Low-rank-sparse subspace representation for robust regression, in: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 7445–7454.
[41] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, Adv. Neural Inform. Processing Systems (2011) 612–620.
[42] X. Zhang, Y. Ma, Z. Lin, H. Gao, L. Zhuang, N. Yu, Non-negative low rank and sparse graph for semi-supervised learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2328–2335.
[44] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, Mulan: A java library for multi-label learning, J. Mach. Learn. Res. 12 (2011) 2411–2414.
[46] Y. Zhang, Z.H. Zhou, Multilabel dimensionality reduction via dependence maximization, ACM Trans. Knowl. Discov. Data 4 (2010) 1–21.
[47] M.L. Zhang, J.M. Peña, V. Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (2009) 3218-3229..
[49] L. Jian, J.D. Li, K. Shu, H. Liu, Multi-label informed feature selection, in Proc. IJCAI, 2016, pp. 1627–1633..
[50] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recogn. 40 (2007) 2038–2048.
[51] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Statist. 11 (1940) 86–92.
[52] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

Chen Chao received his M.S. degree from the School of Mathematics at Shandong University, Jinan, China, in 2012, and the Ph.D. degree from the School of Mathematics at Xiamen University, Xiamen, China, in 2015. He is currently an associate professor with the School of Information Engineering, Zhejiang Ocean University, China. He is actively pursuing research in granular computing and machine learning.

Jia Zhang received the PhD degree from Xi'an Jiaotong University, Xian, China. He is currently a professor in Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, the School of Computer and Information Technology, Shanxi University, Taiyuan, China. His research interests include artificial intelligence, granular computing, and machine learning. He has published more than 120 papers in his research fields, including TPAMI, TKDE, KDD, and Artificial Intelligence.

Jiye Liang received the M.Sc. degree in Mathematics from East China Normal University, Shanghai, China, in 1992, and the Ph.D. degree in Applied Mathematics from Xi'an Jiaotong University, Xi'an, China, in 2002. He is currently a Professor with the School of Information Engineering, Zhejiang Ocean University. He has published 3 monographs and more than 150 articles in international journals and book chapters. His current research interests include granular computing, approximate reasoning, and data mining. Dr. Wu also serves in the editorial boards of several international journals.

Lin Sun received the Ph.D. degree in artificial intelligence from Xiamen University, Xiamen, China, in 2020. He is currently lecturer with the College of information science and technology, Jinan University. He is currently working on multi-label learning, data fusion, feature selection, and weakly-supervised learning.

Wei-Zhi Wu received the M.S. degree in Computer Science and Technology from Henan Normal University in 2007 and the Ph.D. degree in pattern recognition and intelligent systems from Beijing University of Technology in 2015. He is currently an associate professor at the college of Computer and Information Engineering with Henan Normal University, China. He was a visiting scholar at University of Regina, Canada, in 2019. His main research interests include rough sets, granular computing and big data mining. He has served as a reviewer for several prestigious peer-reviewed international journals.

Anhui Tan received the B.E. degree in communication engineering and the M.S. degree in photogrammetry and remote sensing from the Shandong University of Science and Technology, Qingdao, China, in 2005 and 2009, respectively, and the Ph.D. degree in cartography and geography information system from Peking University, Beijing, China, in 2013. He is currently an Associate Professor with Zhejiang Ocean University, Zhoushan, China. His research interests include the data mining and image analysis of remote sensing