Feature Selection with Missing Labels Using Multilabel Fuzzy Neighborhood Rough Sets and Maximum Relevance Minimum Redundancy

Lin Sun, Member, IEEE, Tengyu Yin, Weiping Ding, Senior Member, IEEE, Yuhua Qian, Member, IEEE, and Jiucheng Xu

Abstract—Recently, multilabel classification has generated considerable research interest. However, the high dimensionality of multilabel data incurs high costs; moreover, in many real applications, a number of labels of training samples are randomly missed. Thus, multilabel classification can have great complexity and ambiguity, which means some feature selection methods exhibit poor robustness and yield low prediction accuracy. To solve these issues, this paper presents a novel feature selection method based on multilabel fuzzy neighborhood rough sets (MFNRS) and maximum relevance minimum redundancy (MRMR) that can be used on multilabel data with missing labels. First, to handle multilabel data with missing labels, a relation coefficient of samples, label complement matrix, and label-specific feature matrix are constructed and implemented in a linear regression model to recover missing labels. Second, the margin-based fuzzy neighborhood radius, fuzzy neighborhood similarity relationship, and fuzzy neighborhood information granule are developed. The MFNRS model is built based on multilabel neighborhood rough sets combined with fuzzy neighborhood rough sets. Based on algebra and information views, certain fuzzy neighborhood entropy-based uncertainty measures are proposed for MFNRS. The fuzzy neighborhood mutual information-based MRMR model with label correlation is improved to evaluate the performance of candidate features. Finally, a feature selection algorithm is designed to improve the performance for multilabel data with missing labels. Experiments on twenty datasets verify that our method is effective not only for recovering missing labels but also for selecting significant features with better classification performance.

Index Terms—Feature selection, fuzzy neighborhood entropy, multilabel fuzzy neighborhood rough sets, MRMR.

I. INTRODUCTION

In recent years, multilabel classification has attracted increasing interest from scholars in various fields [1]. Feature selection is a crucial pre-processing step that aims to eliminate redundant features, find an optimal feature subset, and improve the performance of multilabel classification. However, it is difficult to obtain all the proper labels in real applications [2]. Typically, a few labels will be missing, which presents a significant challenge for multilabel feature selection. Currently, feature selection models can be roughly categorized as filter, wrapper, or embedded methods [3], [4]. It is excellent for filter methods to effectively evaluate candidate features [5]. The embedded and wrapper approaches are time-consuming, and in some cases, their selected features can be dependent on specific classifier [6], [7]. Therefore, we focus on feature selection using filter to deal with multilabel data with missing labels.

In many real-world applications, due to the unavailability of all labels, there exist numerous instances with missing labels [8], [9]. Namely, only partial labels are available in label-related applications. These missing labels result in inaccurate measurements between candidate features and label sets, which leads to the loss of valuable features in feature selection [10]. This limits the practical applications of multilabel classification. Zhu et al. [11] proposed a feature selection algorithm for multilabel data with missing labels under l_{2, 1} norm loss. Ma et al. [12] combined input and updated labels in unlabelled space for multilabel classification with missing labels. In general, the aforementioned methods employed all the features available to distinguish all labels, which may be inaccurate. For multilabel classification, each label is affected by its own specific features. Huang et al. [13] used sparsity regularisation and manifold regularisation induced by local feature correlation to select related features. Zhang et al. [14] employed label-specific features to represent samples to predict corresponding labels. Therefore, we focus on feature selection using filter to deal with multilabel data with missing labels.

In recent years, multilabel neighborhood rough sets (MNRS)
and fuzzy neighborhood rough sets (FNRS) have been favoured as two efficient tools for feature selection [5], [16]–[19]. MNRS can deal with continuous and numerical data. Duan et al. [20] presented an MNRS-based multilabel feature selection algorithm. Sun et al. [5] proposed a multilabel feature selection model based on BPSO and MNRS. Liu et al. [16] designed an MNRS-based online multilabel feature selection method. However, because these models use neighborhood similarity classes to approximately describe decision equivalence classes, they cannot represent the fuzziness of instances under a fuzzy background [17], [18]. To overcome this drawback, using fuzzy information granules to describe instance decisions, FNRS can construct a robust distance and thereby reduce error rate of data classification [17]. Chen et al. [19] studied a variable-precision FNRS-based multilabel feature selection method. Vluymans et al. [21] investigated multilabel classification using fuzzy rough neighborhood consensus. However, these FNRS-based models manually select the neighborhood radius, which causes high computational cost, ignores the correlation among labels, and leads to randomness and uncertainty in multilabel classification. To address these issues, fuzzy neighborhood radius based on margin [5] is proposed, using all similar and heterogeneous instances under each label, which will automatically set a neighborhood radius for each dataset to reduce time cost and interference from noisy and improve accuracy. To date, there have been few reports of combining MNRS with FNRS for multilabel feature selection. Therefore, it would be beneficial for us to study multilabel fuzzy neighborhood rough sets (MFNRS) and design an MFNRS-based feature selection algorithm for multilabel data with missing labels.

Mutual information is an effective metric for evaluating uncertainty in random variables [22]–[27]. Iricio et al. [22] designed a mutual information-based filter feature selection model. To date, feature selection based on mutual information has been developed for multilabel data. Gonzalez-Lopez et al. [23] studied mutual information and proposed a continuous feature selection method for multilabel classification. Qian et al. [24] presented a feature selection method using label distribution and mutual information for multilabel learning. However, these studies did not obtain probability and joint distributions of the variables, and the discretization of features easily led to loss of key information. Moreover, mutual information in a fuzzy scenario cannot describe the correlation and redundancy of features [25]. Zhang et al. [26] designed a fuzzy mutual information-based multilabel feature selection for continuous data. Wang et al. [27] proposed a label distribution-based multilabel feature selection method using fuzzy mutual information. Thus, the fuzzy mutual information measure can handle multilabel data with continuous probability distribution well in label space. However, few scholars have focused on multilabel feature selection for missing labels to deal with the probability distribution of data. Based on this observation, a novel fuzzy neighborhood radius based on margin is defined to reflect the diversity and differences of samples, and a fuzzy similarity relationship is developed for label set to represent the inner correlation between labels. However, few algebra- and information-based measures for multilabel feature selection have been reported for multilabel fuzzy neighborhood decision systems. Thus, to study fuzziness from the perspectives of algebra and information, fuzzy neighborhood entropy-based uncertainty measures are proposed for multilabel fuzzy neighborhood decision systems. Furthermore, the maximum relevance minimum redundancy (MRMR) [28] criterion is employed to study fuzzy neighborhood entropy. To solve the problem that MRMR ignores the correlation among labels, label correlation based on fuzzy similarity relationship within the label set is developed and implemented in MRMR. Finally, the improved MRMR with label correlation is presented to evaluate the performance of candidate feature subsets.

Our main contributions can be summarized as follows:

1) To handle the problem of missing labels in multilabel data, a relation coefficient between samples is investigated to discover topological information, and a label complement matrix is defined to obtain label semantic information and learn high-order label correlation. Furthermore, a label-specific feature matrix is implemented in the linear regression model to learn the relations among labels with specific features. Based on the aforementioned approaches, a multilabel learning method based on linear regression is constructed to obtain the complete label matrix as a pre-processing step for feature selection in multilabel data with missing labels.

2) To solve the issue that the neighborhood radius for each dataset is manually set, the margin combining all similar and heterogeneous samples under each label is introduced, and then a novel margin-based fuzzy neighborhood radius is set to granulate all instances using fuzzy neighborhood information granules automatically. Furthermore, the MFNRS model is constructed by combining MNRS with FNRS. To integrate the advantages of MNRS and FNRS, the fuzzy neighborhood lower and upper approximations and fuzzy neighborhood approximate accuracy are provided in MFNRS. Thus, the robust performance of multilabel classification can be significantly improved for the MFNRS model.

3) To study the uncertainty measures of multilabel data with missing labels, fuzzy neighborhood entropy combined with fuzzy neighborhood approximate accuracy is studied from both algebra and information viewpoints, and subsequent entropy measures are proposed. Then, based on the MRMR strategy, fuzzy neighborhood mutual information is proposed to evaluate the redundancy among features and correlation between features and labels. Furthermore, the label correlation based on fuzzy similarity relationship within the label set is implemented in MRMR. Thus, a new MRMR approach is developed to evaluate candidate features. Finally, a feature selection algorithm for multilabel data with missing labels is designed for multilabel fuzzy neighborhood decision systems.

The remainder of this paper is organised as follows. In Section II, related concepts are reviewed. Section III presents a multilabel learning model with missing labels and the MFNRS model; moreover, an improved MRMR approach is proposed. Section IV describes the design of the multilabel feature selection algorithm for missing labels. Experiments are reported in Section V, and Section VI summarizes the findings and contributions of this study.
II. PRELIMINARIES

A. Multilabel neighborhood rough sets

Let \( \text{NDS} = \{U, C, D, V, F, \Lambda, \delta\} \) represent a neighborhood decision system, where \( U = \{x_1, x_2, \cdots, x_m\} \); \( C \) is a set of conditional attributes; \( D \) is a set of decision attributes; \( V = \bigcup_{a \in A} V_a \), where \( V_a \) is a value set of attribute \( a \); \( F: U \times \{C \cup D\} \rightarrow V \) is a map function; \( \Lambda \) denotes a distance function; and \( 0 \leq \delta \leq 1 \) is a neighborhood radius. Let \( \text{MNDS} = \{U, C, D, V, F, \Lambda, \delta\} \) be a multilabel neighborhood decision system, which can be abbreviated to \( \text{MNDS} = \{U, C, D, \Lambda, \delta\} \), where \( D = \{d_1, d_2, \cdots, d_l\} \) is a label set. For any \( B \subseteq C \), the neighborhood relationship is denoted [5] as

\[
\text{NR}_B(B) = \{x, y \in U \mid \Lambda(x, y) \leq \delta, \delta \geq 0\},
\]

and the neighborhood class of \( x \) in \( B \) is expressed [5] as

\[
\delta^*_B(x) = \{y \in x, y \in U, \Lambda(x, y) \leq \delta, \delta \geq 0\},
\]

where \( \Lambda(x, y) \) denotes the Euclidean distance function, and \( \delta^*_B(x) \) is also referred to as the neighborhood granularity of \( x \).

Given \( \text{MNDS} = \{U, C, D, \Lambda, \delta\} \) with \( B \subseteq C \), \( L = \{l_1, l_2, \cdots, l_M\} \) and \( L \subseteq D \), \( D' \) represents a set with label \( l_i \), and \( D \) denotes a set of labels associated with \( x \). The lower and upper approximations of \( D \) to \( B \) are respectively described as [5]

\[
N^_(x) = \{x \mid \forall l \in D, \delta(x) \leq \delta\} \subset \text{U} = \{x \mid \forall l \in D, \delta(x) \geq \delta\} \subset \text{U},
\]

\[
N^y(x) = \{x \mid \forall l \in D, \delta(x) \geq \delta\} \subset \text{U} = \{x \mid \forall l \in D, \delta(x) \leq \delta\} \subset \text{U}.
\]

The neighborhood entropy of \( x \in U \) is denoted [5] as

\[
H(B) = -\log \left( \frac{\delta^*_B(x)}{|U|} \right).
\]

B. Fuzzy neighborhood system

Suppose that there exists a fuzzy neighborhood decision system \( \text{FNDS} = \{U, C, D, V, F, \Lambda, \delta\} \) with the fuzzy neighborhood parameter \( \delta^* \), or in short, \( \text{FNDS} = \{U, C, D, \delta^*\} \). For any \( a \in B \subseteq C \), the fuzzy similarity relation \( R_B \) can be induced on \( U \) if \( R_B \) satisfies the following [27]

(1) Reflexivity: \( R_B(x, y) = 1, \forall x, y \in U \).

(2) Symmetry: \( R_B(x, y) = R_B(y, x), \forall x, y \in U \).

(3) Transitivity: \( R_B(x, y) \geq \min\{R_B(x, z), R_B(z, y)\}, \forall x, y, z \in U \).

Then, the fuzzy neighborhood similarity can be expressed as

\[
[x, y] = R_B(x, y) \text{ and } [\bar{x}, \bar{y}] = \min_{x \in X, y \in Y} \{R_B(x, y)\}.
\]

Given \( \text{FNDS} = \{U, C, D, \delta^*\} \) with \( B \subseteq C \), \( U/D = \{X_1, X_2, \cdots, X_l\} \), for any \( x, y \in U \), the fuzzy neighborhood information granule of \( x \) with respect to \( B \) is expressed [18, 29] as

\[
\alpha^B_d(x) = \left[ \begin{array}{c}
R_B(x, y) \\
0
\end{array} \right] \begin{array}{c}
R_B(x, y) < 1 - \delta^* \\
R_B(x, y) \geq 1 - \delta^*
\end{array}.
\]

The fuzzy neighborhood lower and upper approximations of \( X \) with respect to \( B \) are respectively expressed as [18]

\[
\text{FNL}_B^y(X) = \{x \in U \mid \alpha^B_d(x) \subseteq X\},
\]

\[
\text{FNU}_B^y(X) = \{x \in U \mid \alpha^B_d(x) \cap X \neq \emptyset\}.
\]

For any \( B \subseteq C \), the approximate accuracy of \( D \) with respect to \( B \) is expressed [18] as

\[
\text{AP}_B^y = \frac{\text{FNU}_B^y(X)}{|FNU}_B^y(X)|
\]

III. FEATURE SELECTION IN MULTILABEL DATA WITH MISSING LABELS

A. Multilabel learning with missing labels

Missing labels significantly interfere with classification performance on multilabel data. To overcome this drawback, the relation coefficient of instances is defined to discover topological information between two instances, and a label complement matrix is designed for integration with a linear regression model to obtain more semantic information of labels. Thus, a relation matrix of label-specific features is introduced to enhance the robustness and prediction accuracy of the linear regression model with \( l_i \) norm regularisation.

Definition 1: Suppose that there exists a sample set \( U \) with \( X \subseteq C \). Let \( X \) be a training data matrix. Then, for any \( x \in U \), the relation coefficient \( L \) of \( x_i \) and \( x_j \) is defined as

\[
L = 1 - \frac{CV_y}{\max(CV_y) - \min(CV_y)},
\]

where \( CV_y \) is a transpose of training data matrix, \( CV_y \in R^{m \times m} \) is the correlation matrix of samples, \( \max(CV_y) \) is the maximal value of \( CV_y \), and \( \min(CV_y) \) is the minimal value of \( CV_y \).

Definition 2: Suppose that \( Y \) is a training label matrix and \( C \) is a label correlation matrix. To describe the dependence degree between a data sample and its labels of data when there are missing labels, a minimum function is defined as follows to solve the multilabel problem:

\[
\min_{x \in \mathbb{R}^c} \frac{1}{2} \| Y - XW\|_F^2 + \lambda_1 \text{Tr}(C^T Y^TYC) + \lambda_2 \|C\|_1.
\]

where \( \lambda_1, \lambda_2, \lambda_3 \) are weighting parameters, \( \text{Tr}(C^T Y^TYC) \) is the trace of matrix \( C^T Y^TYC \), and \( C \) represents the label correlation. If two samples are highly similar, they may have similar labels. The topological structure of the data can be extracted by \( CV \). Furthermore, by minimizing the correlation matrix trace \( \text{Tr}(C^T Y^TYC) \) of samples labels, sufficient structural information of the original data can be obtained. To ensure that a label of a sample is only determined by a subset of specific features in the original dataset, the regression coefficient \( W \) is employed to indicate the label-specific feature matrix. Then, \( l_i \) regularisation is implemented in the linear regression model to induce sparsity.

Definition 3: Suppose that \( W \) is a label-specific feature matrix. Then, the optimization problem of multilabel data with missing labels can be expressed as

\[
\min_{w \in \mathbb{R}^c} \frac{1}{2} \| Y - XW\|_F^2 + \frac{1}{2} \| Y - Y\|_F^2 + \lambda_1 \text{Tr}(C^T Y^TYC) + \lambda_2 \| W \|_1 + \lambda_3 \| C \|_1,
\]

where \( \lambda_1, \lambda_2, \lambda_3 \) represent the weighting parameters. Note that the optimization problem in Eq. (12) is convex, but it is not smooth because the objective function contains the \( l_i \) regularisation and trace terms. Then, the accelerated proximal gradient method in [9] is used to solve this non-smooth objective function, where \( \Phi \) is a combined variable of \( W \) and \( C \).

The optimization problem can be transformed as

\[
\min_{w \in \mathbb{R}^c} \| g(\Phi) = f(\Phi) + g(\Phi) \|
\]

where \( f(\Phi) = \frac{1}{2} \| Y - XW\|_F^2 + \frac{1}{2} \| Y - Y\|_F^2 + \lambda_1 \text{Tr}(C^T Y^TYC) \)

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
and \(g(\Phi) = \lambda_3||W||_1 + \lambda_4||\mathbb{C}||_1\). Note that \(f(\Phi)\) is convex and differentiable, and \(g(\Phi)\) is not differentiable. Therefore, for any \(L > 0\), \(\Omega_t(\Phi, \Phi^0) = f(\Phi^0) + \langle \nabla f(\Phi^0), \Phi - \Phi^0 \rangle + \frac{L}{2}||\Phi - \Phi^0||_2^2 + g(\Phi)\).

For any \(L \geq L_f\), the given iteration over \(t\), \(\Omega_t(\Phi, \Phi^0) \geq G(\Phi)\) holds, where \(L_f\) is the Lipschitz constant. Then, the quadriatic model is employed to approximate \(G(\Phi)\). \(H_t = \Phi^0 - \frac{1}{L}\nabla f(\Phi^0)\), where \(\Phi\) can be obtained by minimizing \(\Omega_t(\Phi, \Phi^0)\) using

\[
p_L(\Phi) = \arg\min_{\Phi \in \mathbb{R}^{D \times D}} \Omega_t(\Phi, \Phi^0),
\]

where \(\Phi\) can be obtained by minimizing \(\Omega_t(\Phi, \Phi^0)\) using

\[
p_L(\Phi) = \arg\min_{\Phi \in \mathbb{R}^{D \times D}} \Omega_t(\Phi, \Phi^0),
\]

where \(\Phi\) is defined as prox\(_{\Phi}^t(\Phi^0)\) using

\[
p_L(\Phi) = \arg\min_{\Phi \in \mathbb{R}^{D \times D}} \Omega_t(\Phi, \Phi^0),
\]

where \(\Phi\) is defined as prox\(_{\Phi}^t(\Phi^0)\) for the \(t\)th iteration.

Thus, using one variable each time and fixing the other using its previous value, the parameters \(W\) and \(\mathbb{C}\) can be minimized alternately through the following steps:

**Step 1.** By fixing \(\mathbb{C}\), the derivation of \(f(\Phi)\) with respect to \(\mathbb{C}\) is obtained by \(\nabla f(\Phi) = X^T*X - X^T*\mathbb{C}\). If \(\epsilon\) represents the step size and the \(\lambda_1\) localisation can be solved by the soft-thresholding operator \(\text{prox}_\alpha(W_0) = (|w_{ij}| - \epsilon)*\text{sign}(w_{ij})\), then the accelerated proximal gradient for \(W\) is given as

\[
W^{(t+1)} = W^t + \frac{\alpha_{t+1} - 1}{\alpha_t}(W^t - W_{t-1}); W_{t+1} = \text{prox}_\alpha(W^{(t+1)}) - \frac{1}{L} \nabla f(W^{(t+1)}),
\]

**Step 2.** By fixing \(W\), the derivation of \(f(\Phi)\) with respect to \(\mathbb{C}\) is obtained by \(\nabla f(\Phi) = (1 + \lambda_1)Y^T\mathbb{C} - Y^T*X + \lambda_2Y^T(L + L^T)Y\). When the soft-thresholding operator can be defined as \(\text{prox}_\lambda(C_{y}) = (|C_{y}| - \epsilon)*\text{sign}(C_{y})\), the accelerated proximal gradient for \(\mathbb{C}\) is given as

\[
C^{(t+1)} = C^t + \frac{\alpha_{t+1} - 1}{\alpha_t}(C^t - C_{t-1}); C_{t+1} = \text{prox}_\lambda(C^{(t+1)}) - \frac{1}{L} \nabla f(W^{(t+1)}).
\]

**Theorem 1:** Given \(X \subseteq U\), the optimization problem of multilabel data with missing labels in Eq. (12) is Lipschitz continuous, and the Lipschitz constant can be denoted as \(L_f = \sqrt{2[||X^T*X||_2^2 + ||X^T*Y||_2^2 + ||Y^T*X||_2^2 + ||Y^T*Y||_2^2] + \lambda_2y(L + L^T)||\mathbb{C}||_2^2}\).

**Proof.** It follows immediately from Steps 1 and 2 that

\[
||\nabla f(\Phi_t) - \nabla f(\Phi_t)||_2 \leq \frac{1}{L_f} \nabla f(W^{(t+1)}),
\]

Therefore, the fuzzy neighborhood radius \(\delta\) is defined as

\[
\delta^F = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\Delta_i(x, NS_i(x))}{||NS_i(x)||} = \frac{\Delta_i(x, NT_i(x))}{||NT_i(x)||},
\]

where \(NS_i(x)\) and \(NT_i(x)\) represent the heterogeneous and similar samples of \(x\) with respect to label \(l_i\) respectively, and \(\Delta_i(x, NS_i(x))\) and \(\Delta_i(x, NT_i(x))\) denote the distance from \(x\) with respect to \(NS_i(x)\) and \(NT_i(x)\) under \(l_i\), respectively.

**Definition 5:** Suppose that there exists \(MFNDS = \langle U, C, D, \delta^F \rangle\) with label set \(L = \{l_1, l_2, \ldots, l_M\}\) and \(L \subseteq D\). For any \(x, y \in U\) and \(f \in B\), the fuzzy neighborhood similarity matrix based on \(B\) can be expressed as

\[
R_f(x, y) = \begin{cases} \frac{|f(x) - f(y)|}{\delta^F}, & \text{if } f(x) \neq f(y) \\ 0, & \text{otherwise} \end{cases}
\]

Then, the fuzzy neighborhood similarity matrix \(R_f(x, y)\) is defined as

\[
R_f(x, y) = \begin{cases} \frac{|f(x) - f(y)|}{\delta^F}, & \text{if } f(x) \neq f(y) \\ 0, & \text{otherwise} \end{cases}
\]

**Definition 6:** Suppose that there exists \(MFNDS = \langle U, C, D, \delta^F \rangle\) with \(B = \{f_1, f_2, \ldots, f_f\} \subseteq B\). For any \(x, y \in U\), the fuzzy neighborhood information granule of \(x\) related to \(B\) is defined as

\[
FN^B_f[x](y) = \begin{cases} R_f(x, y) \leq \delta^F, & \text{if } R_f(x, y) \leq \delta^F \\ 0, & \text{otherwise} \end{cases}
\]

**Definition 7:** Suppose that there exists \(MFNDS = \langle U, C, D, \delta^F \rangle\) with \(L = \{l_1, l_2, \ldots, l_M\}\) and \(L \subseteq D\). For any \(x, y \in U\), the fuzzy similarity relationship under label set \(L\) is defined as

\[
\eta^L_i = \begin{cases} \frac{d(x, y)}{\max(d(x, x_i), d(x, y))}, & \text{if } d(x, y) < \delta^F \\ \frac{d(x, y)}{\max(d(x, x_i), d(x, y))}, & \text{otherwise} \end{cases}
\]

where \(d(x, y) = \sqrt{\sum_{i=1}^{M} (c_i - c_{i,y})^2}\), \(c_i = \frac{L(x_i \cup L(x))}{L(x_i)}\) is the Jaccard similarity coefficient, and \(L(x)\) indicates the label set of \(x \in U\). Note that \(c_i\) maps a sample label to Euclidean space \([30]\), which is used to compare the similarities and differences between finite samples. However, in multilabel datasets, there are much fewer positive labels for each sample than there are negative. Inspired by the label correlation presented in Section III.B of Lin’s paper [31], a finer-grained measure of similarity between samples in label space based on \(c_i\) is given in Eq. (20), from which the fuzzy relationship matrix of the label set can be obtained, reflecting the intrinsic correlation among labels.
Definition 8: Suppose that there exists MFNDS = <U, C, D, δ > with B ⊆ C, L = {l₁, l₂, ..., lₖ}, and L ⊆ D. D’ is a set with label lᵢ, and D₀ is the label set associated with xᵢ. Then, the fuzzy neighborhood lower and upper approximations of D with respect to B are respectively defined as

\[ \text{FN}_B^{-}(D) = \{ x_i \mid \forall l_i \in D, \text{FN}_{B_{l_i}}^{+}(x_i) \subseteq D', \ x_i \in U \}, \]

\[ \text{FN}_B^{+}(D) = \{ x_i \mid \exists l_i \in D, \text{FN}_{B_{l_i}}^{-}(x_i) \cap D' \neq \emptyset, \ x_i \in U \}. \]

Then, the fuzzy neighborhood approximate accuracy of D with respect to B is defined as

\[ \text{FAP}_B(D) = \frac{\text{FN}_B^{-}(D)}{\text{FN}_B^{+}(D)}. \]

Definition 9: Suppose that there exists MFNDS = <U, C, D, δ > with B ⊆ C, L = {l₁, l₂, ..., lₖ}, and L ⊆ D. The fuzzy neighborhood entropy of B is defined as

\[ \text{FAP}_B(D) = \frac{\text{FN}_B^{-}(D)}{\text{FN}_B^{+}(D)}. \]

Remark 1: Definition 9 shows that FAP_B(D) is the fuzzy neighborhood approximate accuracy of D relative to B from an algebra perspective and \(-\frac{1}{|U|} \sum_{i=1}^{l} \log |\text{FN}_{B_{l_i}}^{+}(x_i)| \) is the fuzzy neighborhood entropy of B from an information perspective. Then, new fuzzy neighborhood entropy compensates for the defects of information entropy in multilabel classification.

Property 1: Let MFNDS = <U, C, D, δ > with B ⊆ C. Then, \( 0 \leq \text{FAP}_B(D) \leq \log |U| \).

Definition 10: Suppose that there exists MFNDS = <U, C, D, δ > with L = {l₁, l₂, ..., lₖ} and L ⊆ D. For any B₁, B₂ ⊆ C, the fuzzy neighborhood joint entropy of B₁ and B₂ is defined as

\[ \text{FNH}(B_1, B_2) = -\frac{\text{FAP}_{B_1 \cup B_2}(D)}{|U|} \sum_{i=1}^{l} \log \frac{|\text{FN}_{B_{l_i}}^{+}(x_i) \cap \text{FN}_{B_{l_i}}^{-}(x_i)|}{|U|}. \]

Definition 11: Suppose that there exists MFNDS = <U, C, D, δ >. For any B₁, B₂ ⊆ C and xᵢ ∈ U, the fuzzy neighborhood conditional entropy of B₁ with respect to B₂ is defined as

\[ \text{FNH}(B_1 | B_2) = -\frac{\text{FAP}_{B_1 \cap B_2}(D)}{|U|} \sum_{i=1}^{l} \log \frac{|\text{FN}_{B_{l_i}}^{+}(x_i) \cap \text{FN}_{B_{l_i}}^{-}(x_i) \cap \text{FN}_{B_{l_i}}^{+}(x_i) \cap \text{FN}_{B_{l_i}}^{-}(x_i)|}{|U|}. \]

Definition 12: Suppose that there exists MFNDS = <U, C, D, δ >. For any B₁, B₂ ⊆ C and xᵢ ∈ U, the fuzzy neighborhood mutual information of B₁ and B₂ is defined as

\[ \text{FMI}(B_1, B_2) = -\frac{\text{FAP}_{B_1 \cap B_2}(D)}{|U|} \sum_{i=1}^{l} \log \frac{|\text{FN}_{B_{l_i}}^{+}(x_i) \cap \text{FN}_{B_{l_i}}^{-}(x_i)|}{|U|}. \]

Remark 2: Definitions 10–12 combine the fuzzy neighborhood approximate accuracy from algebra perspective and the fuzzy information entropy from information perspective, which allows the uncertainty of multilabel fuzzy neighborhood decision systems with missing labels to be measured accurately.

Property 2: Suppose that there exists MFNDS = <U, C, D, δ > with L = {l₁, l₂, ..., lₖ} and L ⊆ D. For any B₁, B₂ ⊆ C, the following properties hold:

1. \text{FMI}(B_1; B_2) ≥ 0;
2. \text{FMI}(B_1; B_2) = \text{FMI}(B_2; B_1);
3. \text{FMI}(B_1; B_2) = \text{FMI}(B_1 + B_2) - \text{FMI}(B_1, B_2);
4. \text{FMI}(B_1; B_2) = \text{FMI}(B_1; B_2) - \text{FMI}(B_1, B_2).

Proof. (1) The proof is straightforward.
(2) It follows immediately from Definitions 9–11 that \( \text{FMI}(B_1; B_2) = \text{FMI}(B_1 \cup B_2) - \text{FMI}(B_1, B_2) \).
C. MRMR based on fuzzy neighborhood mutual information

MRMR only considers the redundancy among features and correlation between labels and features, while it ignores the impact of label correlation on features [27], [32], [33]. Moreover, MRMR-based methods cannot sufficiently eliminate redundant features or evaluate the integrity of knowledge. To overcome these drawbacks, the correlation among labels is defined and implemented in MRMR with the fuzzy neighborhood similarity relationship on the label set, which can measure the significance of features and reflect the inner correlation between labels to improve classification performance.

Definition 14: Suppose that there exists \( MFNDS = \langle U, C, D, \delta^D \rangle \) with \( S \subseteq C \), \( S \) is the selected features, \( L = \{ l_1, l_2, \ldots, l_m \} \), and \( L \subseteq D \). The maximum relevance is formulated as

\[
\max_{f \neq \phi} REL(f, L), \quad REL(f, L) = \frac{1}{|L|} \sum_{i,j \in L} FNMI(f; l_i, l_j). \tag{28}
\]

Definition 15: Suppose that there exists \( MFNDS = \langle U, C, D, \delta^D \rangle \) with \( S \subseteq C \), \( S \) is the selected features, \( L = \{ l_1, l_2, \ldots, l_m \} \), and \( L \subseteq D \). The minimum redundancy is formulated as

\[
\min_{S \subseteq C} RED(f, S), \quad RED(f, S) = \frac{1}{|S|} \sum_{i,j \in S} FNMI(f; l_i, l_j). \tag{29}
\]

Definition 16: Suppose that there exists \( MFNDS = \langle U, C, D, \delta^D \rangle \) with \( L = \{ l_1, l_2, \ldots, l_m \} \), and \( L \subseteq D \). The label correlation with the fuzzy similarity relationship on label set \( L \) is defined as

\[
\max_{l \in L} LCD(l, L), \quad LCD(l, L) = \sum_{i \in L} t_{ij}^L \tag{30}
\]

where \( t_{ij}^L \) represents the fuzzy similarity relationship for label \( l_i \) and \( l_j \), and \( t_{ij}^L \) is the fuzzy similarity relation for label set \( L \).

Definition 17: Suppose that there exists \( MFNDS = \langle U, C, D, \delta^D \rangle \) with \( S \subseteq C \), \( S \) is the selected features, \( L = \{ l_1, l_2, \ldots, l_m \} \), and \( L \subseteq D \). MRMR with the label fuzzy similarity relationship is defined as

\[
\max_{S \subseteq C} J(REL, RED, LCD), \quad J(REL, RED, LCD) = \frac{REL(f, L) + LCD(l, L)}{RED(f, S)} \tag{31}
\]

\[
= \frac{FNMI(f; l_i) + \sum_{i,j \in L} t_{ij}^L}{|S| \sum_{i,j \in S} FNMI(f; l_i, l_j)}. \]

Remark 3: Definition 17 shows that \( REL(f, L) \) analyses the relevance between \( f \) and \( L \), \( LCD(l, L) \) reflects the inner correlation between labels, and \( RED(f, S) \) focuses on the redundancy between \( f \) and \( S \). \( J(REL, RED, LCD) \) evaluates the significance of each feature one by one and obtain an optimal feature subset for multilabel datasets with missing labels.

D. Multilabel feature selection algorithm for missing labels

To recover missing labels, the multilabel learning algorithm using accelerated proximal gradient optimization (MLAPG) is summarised in Algorithm 1. Suppose that \( m, n \) and \( l \) describe the numbers of samples, features, and labels, respectively. The time complexity of Algorithm 1 is mostly from Steps 3, 4, and 6. Step 3 calculates the Lipschitz constant, and its complexity is approximately \( O(n^3 + \tilde{F}) \). The complexity of calculating the gradient of \( f(\Phi) \) with respect to \( W \) in Step 4 is \( O(n^2 m + n^2 l + n ml + n^2 \tilde{F}) \). Similarly, the complexity of Step 6 is \( O(ml^2 + l^2 + nl + ml) \). Because \( m > n > l \) in most cases, the worst total time complexity of Algorithm 1 is \( O(n^2 (m + n) + l (n + m) + nl) \).

Algorithm 1. MLAPG

**Input:** Training data set \( X \in \mathbb{R}^{n \times l} \), training label set \( Y \in \mathbb{R}^{n \times m} \), parameters \( \lambda_1, \lambda_2, \lambda_3 \) and \( \lambda_4 \).

**Output:** Optimal solution \( W^* \) and \( C^* \).

1. Initialise \( W_0, W_1 = \text{rand} (n, l) \). \( C_0, C_1 = \text{zeros} (l, l) \). \( \Phi = \{ W, C \}; t = 1 \).
2. WHILE not converged
3. Calculate Lipchitz constant \( L_t \) according to Theorem 1.
4. Update \( W^{(t)} \) and \( W_{ct} \) with Eq. (15).
5. Let \( W^{(t+1)} = W^{(t)} \).
6. Update \( C^{(t)} \) and \( C_{ct} \) with Eq. (16).
7. Let \( C^{(t+1)} = C^{(t)} \).
8. Let \( \alpha_t = (1 + \sqrt{4\alpha_t^2 + 1})/2 \).
9. Let \( t = t + 1 \).
10. END WHILE
11. \( W^* = W_t \).
12. \( C^* = C_t \).

The aforementioned MLAPG algorithm is a pre-processing step of the multilabel feature selection used to recover the missing labels. Then, multilabel feature selection for missing labels using MRMR (MFSMR) is described by Algorithm 2.

Algorithm 2: MFSMR

**Input:** Multilabel fuzzy neighborhood decision system \( MFNDS = \langle U, C, D, \delta^D, \text{fuzzy neighborhood parameter } \delta^D \rangle \).

**Output:** Optimal feature subset \( S \).

1. Initialise \( S = \emptyset, k = 1 \).
2. Use MLAPG to obtain the complete multilabel dataset.
3. Calculate the fuzzy neighborhood granule \( \text{FN}_S^\delta \) with Eq. (19).
4. FOR \( f_i \in C \)
5. Calculate \( REL(f_i, L) \) with Eq. (28), where \( l \in D \).
6. Calculate \( RED(f_i, S) \) with Eq. (29), where \( l \in D \).
7. Calculate \( LCD(l, L) \) with Eq. (30), where \( l_i, l_j \in D \).
8. Find \( f_i \) satisfying Eq. (31).
9. Let \( S = S \cup \{ f_i \} \) and \( C = C - \{ f_i \} \).
10. Let \( k = k + 1 \).
11. END FOR
12. RETURN Reduced feature subset \( S \).

In Algorithm 2, based on Algorithm 1, the time complexity of Step 2 is \( O(n^2 (m + n) + \tilde{F} (n + m) + nl) \). Step 3 calculates the fuzzy neighborhood with complexity \( O(ml) \). The main time cost of MFSMR is from Steps 4-11. The time complexity of Step 5 is \( O(ml + n) \), and that of Step 6 is \( O(ml \log n + n) \), where the complexity of calculating the label correlation is \( O(n ml \log n) \). In addition, there exists a loop in Step 4. Therefore, in the worst case, the total time complexity of MFSMR is \( O(n^2 (m + n) + l (n + m) + nl + n^2 ml \log n) \).

IV. EXPERIMENTAL ANALYSIS

A. Experiment preparation

To demonstrate the performance of our MLAPG and MFSMR algorithms, several experiments were performed on twenty multilabel datasets from various fields, which were downloaded from http://mulan.sourceforge.net/datasets.html, http://meka.sourceforge.net/#datasets and http://computer.njnu.edu.cn/Lab/LABIC/LABIC_Software.html, respectively. The characteristics of these datasets are described in Table I.
As in [34], [35], a multi-label k-nearest neighbours (MLKNN) algorithm evaluates the classification performance of all feature selection methods; its smoothing parameter is 1 and \( K = 10 \). Then, MLKNN is employed to describe the processing results for the original dataset. Nine evaluation metrics are used to demonstrate the classification performance of feature selection: number of selected features (\( N \)), average precision (\( AP \)), coverage (\( CV \)), one error (\( OE \)), ranking loss (\( RL \)), Hamming loss (\( HL \)), macro-averaging F1 (MacF1), micro-averaging F1 (MicF1), and macro-AUC (AUC) [9], [10], [36]. For \( AP \), MacF1, and MicF1, the larger the values, the better the performance is; for \( CV \), \( OE \), \( RL \), and \( HL \), the lower the values, the better the performance is. Then, the experimental results for the selected features are obtained using five-fold cross-validation with all the test datasets. For convenience, “↑” represents a larger result being better, and “↓” represents the contrary. The optimal value for each index is given in bold font.

### B. MLAPG compared with other multilabel classification methods with missing labels

These experiments aimed to evaluate MLAPG under different missing percentages in terms of \( AP \), \( CV \), \( OE \), \( RL \), \( HL \), and AUC. Five state-of-the-art multilabel classification algorithms were selected for comparison: MLMF [37], MLNB [38], CDN-LR [39], sCDN-LR [40], and GLOCAL [41]. Following the approaches to setting parameters in [9], [42], [43], the four parameters \( \lambda_1 \), \( \lambda_2 \), \( \lambda_3 \), and \( \lambda_4 \) of MLAPG for the training samples of each dataset were set to values \( \{10^{-3}, 10^{-4}, \ldots, 10^3\} \). The parameters of other algorithms can be found in [37]–[41]. Following the experiments presented in [37], Arts, Business, Computer, Education, Entertainment, Health, Medical, Recreation, Reference, Scene, Science, Social, and Society were selected from Table I for comparison, and the classification results of seven methods on thirteen datasets with different missing percentages (\( p \)) of labels are provided in six tables. From Table II, the \( AP \) of MLAPG is the highest under different missing percentages of labels on most datasets, except for the Education, Entertainment, and Medical datasets. For the Education and Entertainment datasets, MLAPG is second to MLMF in metrics of high \( p \). For the Medical dataset, there is no obvious difference between MLAPG and MLMF, and MLAPG outperforms the other algorithms. More comparison results on the thirteen datasets in terms of \( CV \), \( OE \), \( RL \), \( HL \), and AUC can be found in the supplementary file. It follows from the experimental results in all tables that MLAPG is clearly the best performing algorithm in terms of the six considered metrics for multilabel datasets with missing labels; the classification performances of other algorithms show a downward trend as the missing percentage of labels increases.

### C. MFSMR compared with other multilabel feature selection algorithms with missing labels

The first part of this subsection demonstrates the efficiency of MFSMR on eight multilabel datasets in terms of \( AP \), \( CV \), \( OE \), and \( RL \). MFSMR was compared with seven state-of-the-art multilabel feature selection algorithms with missing labels: MDDM proj [44], MDDM_spc [44], MLNB [38], MDMR [45], MLFRS [46], PMU [47], and MFML [10]. Following the experimental strategies and results in [10], the \( N \) value of eight datasets determined by MLNB was adopted in this experiment. Figs. 1–4 show the classification variation tendency of eight algorithms under various missing percentages, where the horizontal and vertical axes denote the missing percentage of labels and classification results of each metric, respectively.

Fig. 1 shows that in terms of \( AP \), MFSMR is the best performing algorithm on the Arts, Computer, Enron, Entertainment, Recreation, Reference, and Science datasets. For the Health dataset, MFSMR performs as well as MFML and is better than the other six algorithms. From Fig. 2, MFSMR performs better on six datasets: Arts, Computer, Entertainment, Health, Reference, and Science; however, its \( CV \) on the Recreation dataset is volatile and reaches its optimal values when the missing percentages are 0% and 40%. On the Enron dataset, the performances of MFSMR and three algorithms (MDMR, MLFRS, and MFML) are all similar. Fig. 3 indicates that in terms of \( OE \), MFSMR performs better than the other algorithms for almost all missing percentages on five datasets: Arts, Computer, Enron, Entertainment, and Health. On the Recreation and Reference datasets, the difference between the performance of MFSMR and that of MFML is insignificant, whereas MFSMR is superior to the other six algorithms. For the Science dataset, MFSMR outperforms other methods when the missing percentage of labels is less than or equal to 70%. Fig. 4 illustrates that in terms of \( RL \), MFSMR achieves better results than the other algorithms on the Arts, Computer, Health, Reference, and Science datasets. For the Enron and Entertainment datasets, PMU, MDMR, MLFRS, MFML, and MFSMR cannot clearly distinguish pros and cons when the missing percentage is more than 20%. For the Recreation dataset, the classification performance is unstable, and there is no obvious advantage for any algorithm. According to the aforementioned values of all evaluated metrics, it can be concluded that MFSMR obtains better results than the other seven algorithms and indeed improves the classification performance for multilabel datasets with different missing percentages of labels.

**TABLE I**

<table>
<thead>
<tr>
<th>NO.</th>
<th>Datasets</th>
<th>Instance</th>
<th>Feature</th>
<th>Label</th>
<th>LC</th>
<th>LD</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arts</td>
<td>5000</td>
<td>462</td>
<td>26</td>
<td>1.636</td>
<td>0.063</td>
<td>Text</td>
</tr>
<tr>
<td>2</td>
<td>Bibtex</td>
<td>7395</td>
<td>1836</td>
<td>159</td>
<td>2.402</td>
<td>0.015</td>
<td>Text</td>
</tr>
<tr>
<td>3</td>
<td>Birds</td>
<td>645</td>
<td>260</td>
<td>19</td>
<td>1.470</td>
<td>0.074</td>
<td>Audio</td>
</tr>
<tr>
<td>4</td>
<td>bookmarks</td>
<td>7395</td>
<td>1836</td>
<td>159</td>
<td>2.402</td>
<td>0.015</td>
<td>Text</td>
</tr>
<tr>
<td>5</td>
<td>Business</td>
<td>5000</td>
<td>438</td>
<td>30</td>
<td>1.588</td>
<td>0.053</td>
<td>Text</td>
</tr>
<tr>
<td>6</td>
<td>Computer</td>
<td>5000</td>
<td>681</td>
<td>33</td>
<td>1.508</td>
<td>0.046</td>
<td>Text</td>
</tr>
<tr>
<td>7</td>
<td>Corel5k</td>
<td>500</td>
<td>499</td>
<td>374</td>
<td>3.522</td>
<td>0.009</td>
<td>Image</td>
</tr>
<tr>
<td>8</td>
<td>Delicious</td>
<td>16105</td>
<td>500</td>
<td>983</td>
<td>19.02</td>
<td>0.002</td>
<td>Text</td>
</tr>
<tr>
<td>9</td>
<td>Education</td>
<td>5000</td>
<td>550</td>
<td>33</td>
<td>1.461</td>
<td>0.044</td>
<td>Text</td>
</tr>
<tr>
<td>10</td>
<td>Enron</td>
<td>1702</td>
<td>1001</td>
<td>53</td>
<td>3.378</td>
<td>0.064</td>
<td>Text</td>
</tr>
<tr>
<td>11</td>
<td>Entertainment</td>
<td>5000</td>
<td>640</td>
<td>21</td>
<td>1.42</td>
<td>0.068</td>
<td>Text</td>
</tr>
<tr>
<td>12</td>
<td>Health</td>
<td>5000</td>
<td>612</td>
<td>32</td>
<td>1.662</td>
<td>0.052</td>
<td>Text</td>
</tr>
<tr>
<td>13</td>
<td>Medical</td>
<td>978</td>
<td>1449</td>
<td>45</td>
<td>1.245</td>
<td>0.028</td>
<td>Text</td>
</tr>
<tr>
<td>14</td>
<td>Recreation</td>
<td>5000</td>
<td>606</td>
<td>22</td>
<td>1.423</td>
<td>0.065</td>
<td>Text</td>
</tr>
<tr>
<td>15</td>
<td>Reference</td>
<td>5000</td>
<td>793</td>
<td>33</td>
<td>1.169</td>
<td>0.035</td>
<td>Text</td>
</tr>
<tr>
<td>16</td>
<td>Scene</td>
<td>2407</td>
<td>294</td>
<td>6</td>
<td>1.074</td>
<td>0.179</td>
<td>Image</td>
</tr>
<tr>
<td>17</td>
<td>Science</td>
<td>5000</td>
<td>743</td>
<td>40</td>
<td>1.451</td>
<td>0.036</td>
<td>Text</td>
</tr>
<tr>
<td>18</td>
<td>Social</td>
<td>5000</td>
<td>1047</td>
<td>39</td>
<td>1.283</td>
<td>0.033</td>
<td>Text</td>
</tr>
<tr>
<td>19</td>
<td>Society</td>
<td>5000</td>
<td>636</td>
<td>27</td>
<td>1.692</td>
<td>0.063</td>
<td>Text</td>
</tr>
<tr>
<td>20</td>
<td>Yeast</td>
<td>2417</td>
<td>103</td>
<td>14</td>
<td>4.237</td>
<td>0.303</td>
<td>Biology</td>
</tr>
</tbody>
</table>

*LC: label cardinality; LD: label density [33].

In summary, MFSMR is superior to the other algorithms in almost all missing percentages on five datasets: Arts, Computer, Enron, Entertainment, and Health. For the Recreation and Reference datasets, MFSMR achieves better results than the other algorithms in almost all missing percentages. For the Science dataset, MFSMR outperforms other methods when the missing percentage of labels is less than or equal to 70%. However, its \( CV \) on the Recreation dataset is volatile and reaches its optimal values when the missing percentages are 0% and 40%. On the Enron dataset, the performances of MFSMR and the other three algorithms (MDMR, MLFRS, and MFML) are all similar. For the Recreation, and Reference datasets, the difference between the performance of MFSMR and that of MFML is insignificant, whereas MFSMR is superior to the other six algorithms. For the Science dataset, MFSMR outperforms other methods when the missing percentage of labels is less than or equal to 70%. For the Recreation dataset, the classification performance is unstable, and there is no obvious advantage for any algorithm. According to the aforementioned values of all evaluated metrics, it can be concluded that MFSMR obtains better results than the other seven algorithms and indeed improves the classification performance for multilabel datasets with different missing percentages of labels.
TABLE II

<table>
<thead>
<tr>
<th>Datasets</th>
<th>ML.KNN</th>
<th>MLMF</th>
<th>MLNB</th>
<th>CDN-LR</th>
<th>sCDN-LR</th>
<th>GLOCAL</th>
<th>MLAPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.5214</td>
<td>0.6156</td>
<td>0.2347</td>
<td>0.3345</td>
<td>0.5881</td>
<td>0.6023</td>
<td>0.6259</td>
</tr>
<tr>
<td>Business</td>
<td>0.6978</td>
<td>0.7025</td>
<td>0.1826</td>
<td>0.2697</td>
<td>0.6782</td>
<td>0.6812</td>
<td>0.7163</td>
</tr>
<tr>
<td>Computer</td>
<td>0.6229</td>
<td>0.6944</td>
<td>0.1696</td>
<td>0.2233</td>
<td>0.6721</td>
<td>0.6746</td>
<td>0.7175</td>
</tr>
<tr>
<td>Education</td>
<td>0.5671</td>
<td>0.5997</td>
<td>0.1923</td>
<td>0.1401</td>
<td>0.5197</td>
<td>0.5776</td>
<td>0.5999</td>
</tr>
<tr>
<td>Entertainment</td>
<td>0.6042</td>
<td>0.6853</td>
<td>0.2701</td>
<td>0.6021</td>
<td>0.6670</td>
<td>0.6999</td>
<td>0.7092</td>
</tr>
<tr>
<td>Health</td>
<td>0.7080</td>
<td>0.7862</td>
<td>0.1037</td>
<td>0.1454</td>
<td>0.7616</td>
<td>0.7685</td>
<td>0.7929</td>
</tr>
<tr>
<td>Medical</td>
<td>0.7617</td>
<td>0.8879</td>
<td>0.0592</td>
<td>0.1453</td>
<td>0.8698</td>
<td>0.8399</td>
<td>0.8819</td>
</tr>
<tr>
<td>Recreation</td>
<td>0.4493</td>
<td>0.6248</td>
<td>0.5077</td>
<td>0.5824</td>
<td>0.5973</td>
<td>0.6119</td>
<td>0.6422</td>
</tr>
<tr>
<td>Reference</td>
<td>0.6141</td>
<td>0.7118</td>
<td>0.1093</td>
<td>0.2530</td>
<td>0.6972</td>
<td>0.6935</td>
<td>0.7217</td>
</tr>
<tr>
<td>Scene</td>
<td>0.8475</td>
<td>0.8545</td>
<td>0.8174</td>
<td>0.8298</td>
<td>0.8186</td>
<td>0.8266</td>
<td>0.8586</td>
</tr>
<tr>
<td>Science</td>
<td>0.5298</td>
<td>0.5891</td>
<td>0.2082</td>
<td>0.3135</td>
<td>0.5546</td>
<td>0.5791</td>
<td>0.6067</td>
</tr>
<tr>
<td>Social</td>
<td>0.7483</td>
<td>0.7734</td>
<td>0.1436</td>
<td>0.2546</td>
<td>0.7568</td>
<td>0.7576</td>
<td>0.7849</td>
</tr>
<tr>
<td>Society</td>
<td>0.6131</td>
<td>0.6357</td>
<td>0.3236</td>
<td>0.3595</td>
<td>0.6175</td>
<td>0.6239</td>
<td>0.6462</td>
</tr>
</tbody>
</table>

Fig. 1. AP index of the eight algorithms on eight multilabel datasets with different missing percentages.
The following evaluates the effectiveness of MFSMR on six datasets in terms of $AP$, $OE$, $HL$, and MacF1. The eight state-of-the-art multilabel feature selection algorithms for comparison included MDDM [44], PMU [47], SFUS [48], MDMR [45], and MLMLFS ($p = 0.4, p = 0.6, p = 0.8$, and $p = 1$) [11]. Following the strategies of feature selection with missing labels designed by Zhu et al. [11], optimal scores under certain features are provided for all compared algorithms, and the experimental results for $80\%$, $50\%$, $25\%$, and $0\%$ missing labels are provided. As the missing percentage increases, the structure information of labels is degraded to a greater extent, resulting in worse classification performance. As shown in Table III, when $p = 80\%$, the structure information of labels has been completely destroyed, and all other compared algorithms obtain worse performances compared to MFSMR, which has remarkable performance for the Arts, bookmarks, Reference, and Social datasets in terms of $AP$ and MacF1. In terms of $OE$ and $HL$, the performance of MFSMR is not significantly prominent compared with that of MLMLFS; however, MFSMR is far superior to the other algorithms. More comparison results under $50\%$, $25\%$, and $0\%$ missing labels in terms of $AP$, $OE$, $HL$, and MacF1 can be found in the supplementary file. In general, MFSMR achieves superior classification performance on six datasets with missing labels.
The final part further demonstrates the performance of MFSMR on four datasets in terms of AP, CV, OE, RL, and MacF1. The five state-of-the-art multilabel feature selection algorithms were compared with MFSMR: CMFS [49], MSSL [50], CSFS [51], MLMLFS [11], and FSLCLC [13]. Following the experimental technologies in [13], the number of missing labels is set as $m$; for example, $m = 3$ denotes that three labels of all training samples are randomly masked. The significance of features is sorted through MFSMR, and the features are selected when the values of $\lambda$ are varied while the other parameters are fixed at their best setting. The experimental results are shown in Fig. 5. From Fig. 5, it can be observed that MFSMR is relatively insensitive to the parameters with wide ranges, and the classification performance decreases when the values of $\lambda_3$ and $\lambda_4$ are increased. The reason for this is that with the increase of $\lambda_3$ and $\lambda_4$, the discriminative features are lost and the correlated labels are filtered out, which indicates the significant contribution of adding the new label correlation matrix and label-specific feature matrix in the training phase.

D. Parameter analysis

Here, the parameter sensitivity of MFSMR for the four parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ is analysed in detail, where $\lambda_1$ controls the difference between the recovered label matrix manifold and original label matrix, $\lambda_3$ controls the new label matrix manifold, $\lambda_3$ controls the sparsity of the feature matrix, and $\lambda_4$ controls the sparsity of the label matrix. These parameters were tuned using fivefold cross-validation from $10^{-5}$ to $10^3$ with a step size of $10^1$ for each dataset. Following the strategies of parameter analysis provided in [8], [9], the results on the Bitex dataset with 60% missing labels are given in terms of the AP, CV, OE, RL, HL, and AUC indices; one parameter is varied while the other parameters are fixed at their best setting. The experimental results are shown in Fig. 5. From Fig. 5, it can be observed that MFSMR is relatively insensitive to the parameters with wide ranges, and the classification performance decreases when the values of $\lambda_3$ and $\lambda_4$ are increased. The reason for this is that with the increase of $\lambda_3$ and $\lambda_4$, the discriminative features are lost and the correlated labels are filtered out, which indicates the significant contribution of adding the new label correlation matrix and label-specific feature matrix in the training phase.

![Parameter sensitivity analysis on dataset Bitex under 60% missing label percentage.](image-url)
To analyze the statistical performance among all the metrics, the Friedman test and Nemenyi test [52] were employed for performance analysis. The Friedman statistic is expressed as follows:

\[ \chi^2 = \frac{12T}{k(k+1)} \sum_{i=1}^{k} R_i^2 - s(s+1) \]  

where \( T \) and \( s \) are the numbers of datasets and methods, respectively; \( R_i(T-1) \) represents the mean rank of the \( i \)-th method on all datasets. At significance level \( \alpha = 0.1 \), the null hypothesis that all the compared methods perform equivalently is rejected in terms of each evaluation index. The
critical difference among these methods is described as
\[ CD = q_a \sqrt{\frac{(s+1)}{6T}} \]
where \( q_a \) denotes the critical tabulated value.

Following the statistical tests in [5], [18], CD diagrams are employed to visually display the correlation among all the methods. If the average rank of the any compared algorithm and our proposed methods is within one CD, they will be connected. Otherwise, any algorithm not connected with the proposed methods is deemed to be significantly different.

From the aforementioned Table II and Tables I–V in the supplementary file, the \( F_F \) and \( \chi^2_F \) results are list in Table V. Comparison of MFMSR against the other algorithms with the Nemenyi test is displayed in Fig. 6, where \( q_a = 2.693 \) at a significance level \( \alpha = 0.1 \), and the \( CD = 2.2818 \) (ix = 7, T = 13).

As show in Fig. 6, MLAPG performs significantly better than the other algorithms on each index. For all evaluation metrics, MLAPG outperforms MLKNN, MLNB, CDN-LR, sCDN-LR, and GLOCAL, and obtains comparable results against to MLMF. In general, all tested results show that MLAPG provides a competitive performance in all compared methods. Based on Figs. 1–4, Table VI shows \( \chi^2_F \) and \( F_F \) in terms of four metrics and the null hypothesis at \( \alpha = 0.1 \), \( q_a = 2.780 \), and \( CD = 2.3131 \) (ix = 8, T = 8). From Fig. 7, the results for AP, CV, and OE show that MFMSR is better than the other methods; for RL, MFMSR is statistically better than the other methods, and there is no consistent evidence to indicate a statistical equivalence among MFMSR, MFML, MDDMspc, MDDMproj, and MLFRS.

According to the results in Table III and in Tables VI–VIII in the supplementary file, for the Nemenyi test, \( q_a = 2.855 \) when \( \alpha = 0.1 \), and \( CD = 4.5142 \) (ix = 9, T = 6). As shown in Fig. 8, MFMSR clearly outperforms MLMLFS (\( \alpha = 0.1 \)), MLMLFS (\( \alpha = 0.6 \)), MLMLFS (\( \alpha = 0.8 \)), SFUS, MDMR, PMU, and MDDM in metrics of all evaluation indices. Thus, there is no significant difference among MLMLFS (\( \alpha = 0.1 \)), MLMLFS (\( \alpha = 0.4 \)), MLMLFS (\( \alpha = 0.6 \)), and MLMLFS (\( \alpha = 0.8 \)) based on the statistical test. Based on Table IV, \( F_F \) for the five evaluation indices is given in Table VIII, and for \( \alpha = 0.1 \), the null hypothesis of equal performance among the seven methods is rejected under the Friedman test. \( q_a = 2.693 \) when \( \alpha = 0.1 \), and thus, \( CD = 4.114 \) (ix = 7, T = 4). The Nemenyi test results are shown in Fig. 9. For AP, OE, RL, and MacF1, MFMSR achieves statistically superior performance compared to MSSL, MLKNN, MLMLFS, CMFS, and CSFS. There is no consistent evidence for statistical differences between MFMSR and FSLCL. Overall, MFMSR obtains excellent performance when compared with other six methods.

Table V

<table>
<thead>
<tr>
<th align="center">( \chi^2_F )</th>
<th align="center">( F_F )</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">70.58</td>
<td align="center">114.21</td>
</tr>
<tr>
<td align="center">69.68</td>
<td align="center">100.50</td>
</tr>
<tr>
<td align="center">69.68</td>
<td align="center">100.50</td>
</tr>
<tr>
<td align="center">53.27</td>
<td align="center">25.84</td>
</tr>
<tr>
<td align="center">60.59</td>
<td align="center">41.78</td>
</tr>
<tr>
<td align="center">62.44</td>
<td align="center">48.15</td>
</tr>
</tbody>
</table>

Table VI

<table>
<thead>
<tr>
<th align="center">( \chi^2_F )</th>
<th align="center">( F_F )</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">29.91</td>
<td align="center">8.02</td>
</tr>
<tr>
<td align="center">23.10</td>
<td align="center">4.92</td>
</tr>
<tr>
<td align="center">27.53</td>
<td align="center">6.77</td>
</tr>
<tr>
<td align="center">5.90</td>
<td align="center">0.82</td>
</tr>
</tbody>
</table>

Table VII

<table>
<thead>
<tr>
<th align="center">( \chi^2_F )</th>
<th align="center">( F_F )</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">38.67</td>
<td align="center">20.73</td>
</tr>
<tr>
<td align="center">37.34</td>
<td align="center">17.51</td>
</tr>
<tr>
<td align="center">37.26</td>
<td align="center">17.34</td>
</tr>
<tr>
<td align="center">38.49</td>
<td align="center">20.24</td>
</tr>
</tbody>
</table>

Table VIII

<table>
<thead>
<tr>
<th align="center">( \chi^2_F )</th>
<th align="center">( F_F )</th>
</tr>
</thead>
<tbody>
<tr>
<td align="center">14.49</td>
<td align="center">4.57</td>
</tr>
<tr>
<td align="center">16.21</td>
<td align="center">6.24</td>
</tr>
<tr>
<td align="center">17.63</td>
<td align="center">8.31</td>
</tr>
<tr>
<td align="center">14.84</td>
<td align="center">4.86</td>
</tr>
<tr>
<td align="center">16.19</td>
<td align="center">6.22</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, a multilabel feature selection method using MFNRS and MRMR was proposed to improve classification performance of multilabel data with missing labels. First, in combination with the relation coefficient between samples, the label complement matrix and label-specific feature matrix based on linear regression model were studied, and then the
multilabel learning model was presented to recover missing labels. Second, a margin-based fuzzy neighborhood radius was presented, and the MNRS model was constructed by combining MNRSs with FNRSs. By integrating algebra and information viewpoints, fuzzy neighborhood entropy-based uncertainty measures were investigated. Third, the label correlation based on the fuzzy similarity within the label set was defined, and the new MRMR model was developed to evaluate the performance of candidate feature subsets. Finally, the multilabel feature selection algorithm with missing labels was designed to efficiently eliminate redundant features and optimize classification performance on multilabel data. Extensive experiments showed that our method can achieve competitive and promising results. However, because the accelerated proximal gradient strategy is used to solve the model optimization of MFSMR and the solution process for the Lipschitz constant requires a large number of matrix operations, high time cost easily appears. In addition, MFSMR cannot achieve better classification performance when the missing percentage is very high. To improve classification performance and decrease computational cost of our model for multilabel data with missing labels, more efficient optimal search strategies and uncertainty measures based on multilabel fuzzy neighborhood rough sets should be explored in future work.

ACKNOWLEDGEMENT

The authors would like to express their sincere appreciation to the anonymous reviewers for their insightful comments, which greatly improved the quality of this paper.

REFERENCES


[33] A. Agarwal and A. Gupta, “A maximum relevancy and minimum
redundancy feature selection approach for median filtering forensics,” 


Jiucheng Xu is currently a Professor at the College of Computer and Information Engineering, Henan Normal University. He received the M.S. degree and the Ph.D. degree in Computer Science and Technology from Xian Jiaotong University in 1995 and 2004, respectively. He has received funding from grants from the National Natural Science Foundation of China, the Key Scientific Research Project of Higher Education of Henan Province, and the Key Scientific and Technological Project of Henan Province. He has published over 100 articles. His research interests include granular computing, data mining, and pattern recognition. He has received the title of Henan's Distinguished High Profile Professional, and has served as a reviewer in several prestigious peer-reviewed international journals.