This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

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 measures 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. Jiang *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. Huang *et al.* [15] learned label-specific features and class-dependent labels using a sparse stacking approach. Although these methods consider the relationship between labels and specific features, they ignore relevant information among labels. Furthermore, because some partial labels are missing for multilabel data, it is important to restore missing labels. To solve these issues, a correlation coefficient between any two samples and a label correlation complement matrix are proposed and implemented in a linear regression model; then a relation matrix between labels and specific features is employed to improve prediction accuracy of label. A novel linear regression model with label correlation and label-specific features is constructed to recover missing labels.

In recent years, multilabel neighborhood rough sets (MNRS)

This work was supported in part by the NSFC Grants 62076089, 61772176, 61976120, 61976082 and 61672332, the Natural Science Foundation of Jiangsu Province Grant BK20191445, and the Six Talent Peaks Project of Jiangsu Province Grant XYDXXJS-048. (*Corresponding author: Weiping Ding*)
L. Sun, T. Yin, and J. Xu are with the College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China (e-mail: linsunok@gmail.com, tengyuhtu@163.com, jiuchxu@gmail. com).
W. Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).
Y. Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China (e-mail: jinchengqyh@126.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

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]. Ircio *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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

II. Preliminaries

A. Multilabel neighborhood rough sets

Let $NDS = <U, C, D, V, F, \Delta, \delta>$ 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: $U \times \{C \bigcup D\} \to V$ is a map function; Δ denotes a distance function; and $0 \le \delta \le 1$ is a neighborhood radius. Let $MNDS = <U, C, D, V, F, \Delta, \delta>$ be a multilabel neighborhood decision system, which can be abbreviated to $MNDS = <U, C, D, \delta>$, 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

$$NR_\delta(B) = \{(x, y) \in U \mid \Delta(x, y) \le \delta, \delta \ge 0\}, \tag{1}$$

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

$$\delta_B(x) = \{y \mid x, y \in U, \Delta(x, y) \le \delta, \delta \ge 0\}, \tag{2}$$

where $\Delta(x, y)$ denotes the Euclidean distance function, and $\delta_B(x)$ is also referred to as the neighborhood granularity of x.

Given $MNDS = <U, C, D, \delta>$ with $B \subseteq C$, $L = \{l_1, l_2, \cdots, l_M\}$ and $L \subseteq D$, D^j represents a set with label l_j, and D_i denotes a set of labels associated with x_i. The lower and upper approximations of D to B are respectively described as [5]

$$\underline{N_B}D = \{x_i \mid \forall l_j \in D_i, \delta_B(x_i) \subseteq D^j, x_i \in U\}, \tag{3}$$

$$\overline{N_B}D = \{x_i \mid \forall l_j \in D_i, \delta_B(x_i) \bigcap D^j \ne \varnothing, x_i \in U\}. \tag{4}$$

The neighborhood entropy of $x_i \in U$ is denoted [5] as

$$H(B) = -\log \frac{|\delta_B(x_i)|}{|U|}. \tag{5}$$

B. Fuzzy neighborhood rough sets

Suppose that there exists a fuzzy neighborhood decision system $FNDS = <U, C, D, V, F, \Delta, \delta^F>$ with the fuzzy neighborhood parameter δ^F, or in short, $FNDS = <U, C, D, \delta^F>$. 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: $R_B(x, x) = 1$, $\forall\, x, y \in U$.

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

(3) Transitivity: $R_B(x, z) \ge \min(R_B(x, y), R_B(y, z))$, $\forall\, x, y \in U$.

Then, the fuzzy neighborhood similarity can be expressed as $[x]_B(y) = R_B(x, y)$ and $[x]^\alpha_B(y) = \min_{a\in B}([x]_a(y))$.

Given $FNDS = <U, C, D, \delta^F>$ 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

$$\alpha_B(x) = [x]^\alpha_B(y) = \begin{cases} 0 & R_B(x, y) < 1-\delta^F \\ R_B(x, y) & R_B(x, y) \ge 1-\delta^F \end{cases}. \tag{6}$$

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

$$\underline{FN^\alpha_B}(X) = \{x \in U \mid \alpha_B(x) \subseteq X\}, \tag{7}$$

$$\overline{FN^\alpha_B}(X) = \{x \in U \mid \alpha_B(x) \bigcap X \ne \varnothing\}. \tag{8}$$

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

$$AP^\alpha_B = \frac{|\underline{FN^\alpha_B}(X)|}{|\overline{FN^\alpha_B}(X)|}. \tag{9}$$

III. Feature Slection 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_1 norm regularisation.

Definition 1: Suppose that there exists a sample set U with $X \subseteq U$. 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

$$\mathrm{L} = 1 - \frac{CV_{ij}}{\max(CV) - \min(CV)}, \tag{10}$$

where $CV = XX^T$, X^T is the transpose of training data matrix, $CV \in R^{m\times m}$ is the correlation matrix of samples, $\max(CV)$ is the maximal value of CV, and $\min(CV)$ is the minimal value of CV.

Definition 2: Suppose that Y is a training label matrix and $\mathbb{C}$ 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:

$$\min_{\mathbb{C}} \frac{\lambda_1}{2}\|\mathrm{Y}\mathbb{C} - \mathrm{Y}\|_F^2 + \lambda_2 \mathrm{Tr}(\mathbb{C}^T \mathrm{Y}^T \mathrm{L} \mathrm{Y} \mathbb{C}) + \lambda_3 \|\mathbb{C}\|_1, \tag{11}$$

where λ_1, λ_2 and λ_3 are weighting parameters, $\mathrm{Tr}(\mathbb{C}^T \mathrm{Y}^T \mathrm{L} \mathrm{Y} \mathbb{C})$ is the trace of matrix $\mathbb{C}^T \mathrm{Y}^T \mathrm{L} \mathrm{Y} \mathbb{C}$, and $\mathbb{C}$ 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 $\mathrm{Tr}(\mathbb{C}^T \mathrm{Y}^T \mathrm{L} \mathrm{Y} \mathbb{C})$ 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_1 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

$$\begin{aligned} \min_{\mathrm{W},\mathrm{C}} \; & \frac{1}{2}\|\mathrm{Y}\mathbb{C} - \mathrm{XW}\|_F^2 + \frac{\lambda_1}{2}\|\mathrm{Y}\mathbb{C} - \mathrm{Y}\|_F^2 + \\ & \lambda_2 \mathrm{Tr}(\mathbb{C}^T \mathrm{Y}^T \mathrm{L} \mathrm{Y} \mathbb{C}) + \lambda_3 \|\mathrm{W}\|_1 + \lambda_4 \|\mathbb{C}\|_1, \end{aligned} \tag{12}$$

where λ_1, λ_2, λ_3, and λ_4 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_1 regularisation and trace terms. Then, the accelerated proximal gradient method in [9] is used to solve this non-smooth objective function, where Φ is a combined variable of W and $\mathbb{C}$. The optimization problem can be transformed as

$$\min_{\Phi} \{G(\Phi) = f(\Phi) + g(\Phi)\}, \tag{13}$$

where $f(\Phi) = \frac{1}{2}\|\mathrm{Y}\mathbb{C} - \mathrm{XW}\|_F^2 + \frac{\lambda_1}{2}\|\mathrm{Y}\mathbb{C} - \mathrm{Y}\|_F^2 + \lambda_2 \mathrm{Tr}(\mathbb{C}^T \mathrm{Y}^T \mathrm{L}\mathrm{Y}\mathbb{C})$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

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 $L > 0$, $\Omega_L(\Phi,\Phi^{(t)}) = f(\Phi^{(t)}) + \langle \nabla f(\Phi^{(t)}), \Phi - \Phi^{(t)}\rangle + \frac{L}{2}\|\Phi - \Phi^{(t)}\|_F^2 + g(\Phi)$. For any $L \geq L_f$, the given iteration over t, $\Omega_L(\Phi, \Phi^{(t)}) \geq G(\Phi)$ holds, where L_f is the Lipschitz constant. Then, the quadratic model is employed to approximate $G(\Phi)$. $H^{(t)} = \Phi^{(t)} - \frac{1}{L}\nabla f(\Phi^{(t)})$, where Φ can be obtained by minimizing $\Omega_L(\Phi, \Phi^{(t)})$ using

$$pL(\Phi) = \underset{\Phi}{\operatorname{argmin}}\left\{\Omega_L(\Phi, \Phi^{(t)})\right\} = \underset{\Phi}{\operatorname{argmin}}\ g(\Phi) + \frac{L}{2}\|\Phi - H(t)\|_F^2 . \quad (14)$$

Huang *et al.* [9] demonstrated that if a sequence α_t satisfies $\alpha_{t+1}^2 - \alpha_{t+1} \leq \alpha_t^2$, the convergence rate can be improved to $O(\frac{1}{t^2})$, when setting $\Phi^{(t)} = \Phi_t + \frac{\alpha_{t-1}-1}{\alpha_t}(\Phi_{t-1})$ for the tth 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:

Step1. By fixing $\mathbb{C}$, the derivation of $f(\Phi)$ with respect to W is obtained by $\nabla_W f(\Phi) = X^TXW - X^TY\mathbb{C}$. If ε represents the step size and the l_1 regularisation can be solved by the soft-thresholding operator $\operatorname{prox}_\varepsilon(w_{ij}) = (|w_{ij}| - \varepsilon)_+\operatorname{sign}(w_{ij})$, then the accelerated proximal gradient for W is given as

$$W^{(t)} = W_t + \frac{\alpha_{t-1}-1}{\alpha_t}(W_t - W_{t-1});\ W_{t+1} = \operatorname{prox}_\varepsilon(W^{(t)} - \frac{1}{L}\nabla_W f(W^{(t)},\mathbb{C})). \quad (15)$$

Step2. By fixing W, the derivation of $f(\Phi)$ with respect to $\mathbb{C}$ is obtained by $\nabla_{\mathbb{C}} f(\Phi) = (1 + \lambda_1)Y^TY\mathbb{C} - Y^TXW - \lambda_1 Y^TY + \lambda_2 Y^T(L+L^T)Y\mathbb{C}$. When the soft-thresholding operator can be defined as $\operatorname{prox}_\varepsilon(\mathbb{C}_{ij}) = (|\mathbb{C}_{ij}| - \varepsilon)_+$, the accelerated proximal gradient for $\mathbb{C}$ is given as

$$\mathbb{C}^{(t)} = \mathbb{C}_t + \frac{\alpha_{t-1}-1}{\alpha_t}(\mathbb{C}_t - \mathbb{C}_{t-1});\ \mathbb{C}_{t+1} = \operatorname{prox}_\varepsilon(\mathbb{C}^{(t)} - \frac{1}{L}\nabla_{\mathbb{C}} f(W,\mathbb{C}^{(t)})). \quad (16)$$

Theorem 1: Given $X \subseteq U$, the optimization problem of multilabel data with missing labels in Eq. (12) is Lipchitz continuous, and the Lipchitz constant L_f can be denoted as

$$L_f = \sqrt{2(\|X^TX\|_2^2 + \|X^TY\|_2^2 + \|(1+\lambda_1)Y^TY\|_2^2 + \|Y^TX\|_2^2 + \|\lambda_2 Y^T(L+L^T)Y\|_2^2)} .$$

Proof. It follows immediately from Steps 1 and 2 that

$$\begin{aligned}
&\left\|\nabla f(\Phi_1) - \nabla f(\Phi_2)\right\|_F^2 \\
&= \|X^TX\Delta W - X^TY\Delta\mathbb{C}\|_F^2 + \|(1+\lambda_1)Y^TY\Delta\mathbb{C} - Y^TX\Delta W \\
&\quad + \lambda_2 Y^T(L+L^T)Y\Delta\mathbb{C}\|_F^2 \\
&\leq 2\|X^TX\|_2^2\|\Delta W\|_F^2 + 2\|X^TY\|_2^2\|\Delta\mathbb{C}\|_F^2 + 2\|(1+\lambda_1)Y^TY\|_2^2\|\Delta\mathbb{C}\|_F^2 \\
&\quad -\|Y^TX\|_2^2\|\Delta W\|_F^2 + 2\|\lambda_2 Y^T(L+L^T)Y\|_2^2\|\Delta\mathbb{C}\|_F^2 .
\end{aligned}$$

Namely, $\left\|\nabla f(\Phi_1) - \nabla f(\Phi_2)\right\|_F^2 \leq 2(\|X^TX\|_2^2 + \|X^TY\|_2^2 + \|(1+\lambda_1)Y^TY\|_2^2 + \|Y^TX\|_2^2\|\Delta W\|_F^2 + \|\lambda_2 Y^T(L+L^T)Y\|_2^2)\left\|\begin{matrix}\Delta W \\ \Delta\mathbb{C}\end{matrix}\right\|_F^2$. Thus, the Lipschitz constant of the objective function is obtained as

$$L_f = \sqrt{2(\|X^TX\|_2^2 + \|X^TY\|_2^2 + \|(1+\lambda_1)Y^TY\|_2^2 + \|Y^TX\|_2^2 + \|\lambda_2 Y^T(L+L^T)Y\|_2^2)} .$$

B. Multilabel fuzzy neighborhood rough sets

Because the fuzzy neighborhood radius is set manually to achieve optimal accuracy in almost all approaches [18], [29], the time cost is high. Moreover, the integrity and information diversity of multilabel data are easily ignored. To address these drawbacks, a new margin-based fuzzy neighborhood radius is presented by combining features and labels of samples; this approach reduces noise caused by weak correlation between samples. By integrating label correlation, fuzzy similarity within label set is defined to explore inner correlation between labels. Based on algebra and information viewpoints, by integrating the MNRS and FNRS models, MFNRS is presented, and various uncertainty measures are proposed to evaluate the performance of candidate features for multilabel classification.

Definition 4: Suppose that there exists a multilabel fuzzy neighborhood decision system $MFNDS = \langle U, C, D, \delta^F\rangle$ with label set $L = \{l_1, l_2, \cdots, l_M\}$ and $L \subseteq D$. For any $x \in U$, the fuzzy neighborhood radius δ^F is defined as

$$\delta^F = \frac{\sum_{j=1}^{|U|}\sum_{i=1}^{|L|}(\frac{\Delta_{l_i}(x, NS_{l_i}(x))}{|NS_{l_i}(x)|} - \frac{\Delta_{l_i}(x, NT_{l_i}(x))}{|NT_{l_i}(x)|})}{|U||L|}, \quad (17)$$

where $NS_{li}(x)$ and $NT_{li}(x)$ represent the heterogeneous and similar samples of x with respect to label l_i, respectively, and $\Delta_{li}(x, NS_{li}(x))$ and $\Delta_{li}(x, NT_{li}(x))$ denote the distance from x with respect to $NS_{li}(x)$ and $NT_{li}(x)$ under l_i, respectively.

Definition 5: Suppose that there exists $MFNDS = \langle U, C, D, \delta^F\rangle$ with $B \subseteq C$, $B = \{f_1, f_2, \cdots, f_n\}$, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. For any $x, y \in U$ and $f \in B$, the fuzzy neighborhood similarity relationship between x and y with respect to f is defined as

$$R_f(x,y) = \begin{cases} 0, & |F(x,f) - F(y,f)| > \delta^F \\ 1 - |F(x,f) - F(y,f)|, & |F(x,f) - F(y,f)| \leq \delta^F \end{cases}. \quad (18)$$

Then, the fuzzy neighborhood similarity matrix $[x]_f(y) = R_f(x, y)$. Therefore, the fuzzy neighborhood similarity matrix based on B can be expressed as $[x]_B(y) = \min([x]_f(y))$.

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

$$FN_B^\delta = [x]_B(y) = \begin{cases} 0, & R_B(x,y) < 1 - \delta^F \\ R_B(x,y), & R_B(x,y) \geq 1 - \delta^F \end{cases}. \quad (19)$$

Definition 7: Suppose that there exists $MFNDS = \langle U, C, D, \delta^F\rangle$ with $L = \{l_1, l_2, \cdots, l_M\}$ and $L \subseteq D$. For any $x_i, x_j, x_s, x_t \in U$, the fuzzy similarity relationship under label set L is defined as

$$r_{ij}^L = \begin{cases} 1 - 4 \times \frac{d(x_i,x_j)}{\max(d(x_s,x_t)) - \min(d(x_s,x_t))}, & \frac{d(x_i,x_j)}{\max(d(x_s,x_t)) - \min(d(x_s,x_t))} \leq 0.25 \\ 0 & otherwise \end{cases}, \quad (20)$$

where $d(x_i, x_j) = \sqrt{(\sum_{r=1}^m (c_{ir} - c_{jr})^2)}$, $c_{ij} = \frac{|L(x_i) \cap L(x_j)|}{|L(x_i) \cup L(x_j)|}$ is the Jaccard similarity coefficient, and $L(x)$ indicates the label set of $x \in U$. Note that c_{ij} 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_{ij} is given in Eq. (20), from which the fuzzy relationship matrix of the label set can be obtained, reflecting the intrinsic correlation among labels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

Definition 8: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $B \subseteq C$, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. D^j is a set with label l_j, and D_i is the label set associated with x_i. Then, the fuzzy neighborhood lower and upper approximations of D with respect to B are respectively defined as

$$\underline{FN_B}D = \{x_i \mid \forall l_j \in D_i, FN_B^\delta(x_i) \subseteq D^j, x_i \in U\}, \quad (21)$$

$$\overline{FN_B}D = \{x_i \mid \forall l_j \in D_i, FN_B^\delta(x_i) \cap D^j \neq \varnothing, x_i \in U\}. \quad (22)$$

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

$$FAP_B(D) = \frac{\underline{FN_B}D}{\overline{FN_B}D}. \quad (23)$$

Definition 9: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $B \subseteq C$, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. The fuzzy neighborhood entropy of B is defined as

$$FNH(B) = -\frac{FAP_B(D)}{|U|}\sum_{i=1}^{|U|}\log\frac{|FN_B^\delta(x_i)|}{|U|}. \quad (24)$$

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 $-\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|FN_B^\delta(x_i)|}{|U|}$ 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 $MFNDS = <U, C, D, \delta^F>$ with $B \subseteq C$. Then, $0 \le FNH(B) \le \log|U|$.

Definition 10: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $L = \{l_1, l_2, \cdots, l_M\}$ and $L \subseteq D$. For any $B_1, B_2 \subseteq C$, the fuzzy neighborhood joint entropy of B_1 and B_2 is defined as

$$FNH(B_1, B_2) = -\frac{FAP_{B_1 \cup B_2}(D)}{|U|}\sum_{i=1}^{|U|}\log\frac{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|}{|U|}. \quad (25)$$

Definition 11: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$. For any $B_1, B_2 \subseteq C$ and $x_i \in U$, the fuzzy neighborhood conditional entropy of B_1 with respect to B_2 is defined as

$$FNH(B_1 \mid B_2) = -\frac{\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}|U|^{FAP_{B_2}(D)}}{|U|^{FAP_{B_1 \cup B_2}(D)}|FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}}\right)}{|U|}. \quad (26)$$

Definition 12: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$. For any $B_1, B_2 \subseteq C$ and $x_i \in U$, the fuzzy neighborhood mutual information of B_1 and B_2 is defined as

$$FNMI(B_1; B_2) = -\frac{\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_1}(D)}|FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}|U|^{FAP_{B_1 \cup B_2}(D)}}{|U|^{FAP_{B_1}(D)+FAP_{B_2}(D)}|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}}\right)}{|U|}. \quad (27)$$

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 $MFNDS = <U, C, D, \delta^F>$ with $L = \{l_1, l_2, \cdots, l_M\}$ and $L \subseteq D$. For any $B_1, B_2 \subseteq C$, the following properties hold:

(1) $FNMI(B_1; B_2) \ge 0$;
(2) $FNH(B_1|B_2) = FNH(B_1, B_2) - FNH(B_2)$;
(3) $FNMI(B_1; B_2) = FNH(B_1) + FNH(B_2) - FNH(B_1, B_2)$;
(4) $FNMI(B_1; B_2) = FNH(B_1) - FNH(B_1|B_2)$.

Proof. (1) The proof is straightforward.

(2) It follows immediately from Definitions 9-11 that

$$\begin{aligned}&FNH(B_1, B_2) - FNH(B_2)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}}{|U|^{FAP_{B_1 \cup B_2}(D)}} \cdot \frac{|U|^{FAP_{B_2}(D)}}{|FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}|U|^{FAP_{B_2}(D)}}{|U|^{FAP_{B_1 \cup B_2}(D)}|FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}}\right).\end{aligned}$$

Then, $FNH(B_1|B_2) = FNH(B_1, B_2) - FNH(B_2)$ holds.

(3) It follows immediately from Definitions 9, 10 and 12 that

$$\begin{aligned}&FNH(B_1) + FNH(B_2) - FNH(B_1, B_2)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_1}(D)}}{|U|^{FAP_{B_1}(D)}} \cdot \frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_2}(D)}}{|U|^{FAP_{B_2}(D)}} \cdot \frac{|U|^{FAP_{B_1 \cup B_2}(D)}}{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_1}(D)}|FN_{B_1}^\delta(x_i)|^{FAP_{B_2}(D)}|U|^{FAP_{B_1 \cup B_2}(D)}}{|U|^{FAP_{B_1}(D)+FAP_{B_2}(D)}|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}}\right).\end{aligned}$$

Thus, $FNMI(B_1; B_2) = FNH(B_1) + FNH(B_2) - FNH(B_1, B_2)$.

(4) It follows immediately from Definitions 9, 11 and 12 that

$$\begin{aligned}&FNH(B_1) - FNH(B_1 \mid B_2)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_1}(D)}}{|U|^{FAP_{B_1}(D)}} \cdot \frac{|U|^{FAP_{B_1 \cup B_2}(D)} \cdot |FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}}{|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)} \cdot |U|^{FAP_{B_2}(D)}}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_{B_1}^\delta(x_i)|^{FAP_{B_1}(D)}|FN_{B_2}^\delta(x_i)|^{FAP_{B_2}(D)}|U|^{FAP_{B_1 \cup B_2}(D)}}{|U|^{FAP_{B_1}(D)+FAP_{B_2}(D)}|FN_{B_1}^\delta(x_i) \cap FN_{B_2}^\delta(x_i)|^{FAP_{B_1 \cup B_2}(D)}}\right).\end{aligned}$$

Hence, obviously $FNMI(B_1; B_2) = FNH(B_1) - FNH(B_1|B_2)$.

Definition 13: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $B \subseteq C$, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. l_{xi} denotes a sample set with the same label as x_i. If $FN_B^\delta(x_i) \subseteq l_{x_i}$, the fuzzy decision of x_i is consistent.

Property 3: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $B \subseteq C$, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. Then,

$$FNMI(B; l) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_B^\delta(x_i)|^{FAP_B(D)}|l_{x_i}|}{|FN_B^\delta(x_i) \cap l_{x_i}|^{FAP_B(D)}|U|}\right).$$

Proof. Suppose that any $x_i \in U$ ($i = 1, 2, 3 \cdots, m$) is consistent. It follows from the proof of Property 3 in [5] that $FN_{B\cup l}^\delta(x_i) = FN_B^\delta(x_i) \cap l_{x_i}$. Then, we have $FN_B^\delta(x_i) \subseteq l_{x_i}$, and $FN_{B\cup l}^\delta(x_i) = FN_B^\delta(x_i)$ can be obtained clearly. Furthermore, from Definition 8, it follows that

$$\begin{aligned}\underline{FN_{B\cup l}}D &= \{x_i | \forall l_j \in D_i, FN_{B\cup l}^\delta(x_i) \subseteq D^j, x_i \in U\}\\ &= \{x_i | \forall l_j \in D_i, FN_B^\delta(x_i) \subseteq D^j, x_i \in U\}\end{aligned}$$

and

$$\begin{aligned}\overline{FN_{B\cup l}}D &= \{x_i \mid \forall l_j \in D_i, FN_{B\cup l}^\delta(x_i) \cap D^j \neq \varnothing, x_i \in U\}\\ &= \{x_i \mid \forall l_j \in D_i, FN_B^\delta(x_i) \cap D^j \neq \varnothing, x_i \in U\}\end{aligned}.$$

From Definition 8, $FAP_{B\cup l}(D) = FAP_B(D)$ holds. Therefore, we clearly have $FNMI(B; l) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|FN_B^\delta(x_i)|^{FAP_B(D)} \cdot |l_{x_i}|}{|FN_B^\delta(x_i) \cap l_{x_i}|^{FAP_B(D)} \cdot |U|}\right)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

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 $MFNDS = <U, C, D, \delta^F>$ with $S \subseteq C$, S is the selected features, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. The maximum relevance is formulated as

$$\max REL(f_i, L), \quad REL(f_i, L) = \frac{1}{|L|} \sum_{f_i \in S, l_i \in L} FNMI(f_i; l_i). \tag{28}$$

Definition 15: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $S \subseteq C$, S is the selected features, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. The minimum redundancy is formulated as

$$\min RED(f_i, S), \quad RED(f_i, S) = \frac{1}{|S|} \sum_{f_i, f_j \in S} FNMI(f_i; f_j). \tag{29}$$

Definition 16: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. The label correlation with the fuzzy similarity relationship on label set L is defined as

$$\max LCD(l_i, L), \quad LCD(l_i, L) = \sum_{l_i \in L} \frac{r_{ij}^{l_i}}{r_{ij}^{L}}, \tag{30}$$

where $r_{ij}^{l_i}$ represents the fuzzy similarity relationship for label l_i, and r_{ij}^{L} is the fuzzy similarity relation for label set L.

Definition 17: Suppose that there exists $MFNDS = <U, C, D, \delta^F>$ with $S \subseteq C$, S is the selected features, $L = \{l_1, l_2, \cdots, l_M\}$, and $L \subseteq D$. MRMR with the label fuzzy similarity relationship is defined as

$$\begin{aligned} &\max J(REL, RED, LCD), \\ &J(REL, RED, LCD) = \frac{REL(f_i, L) + LCD(l_i, L)}{RED(f_i, S)} \\ &= \max\{\frac{FNMI(f_i; l_i) + \sum_{l_i \in L} \frac{r_{ij}^{l_i}}{r_{ij}^{L}}}{\frac{1}{|S|} \sum_{f_i, f_j \in S} FNMI(f_i; f_j)}\}. \end{aligned} \tag{31}$$

Remark 3: Definition 17 shows that $REL(f_i, L)$ analyses the relevance between f_i and L, $LCD(l_i, L)$ reflects the inner correlation between labels, and $RED(f_i, S)$ focuses on the redundancy between f_i 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 $O(n^3 + l^3)$. The complexity of calculating the gradient of $f(\Phi)$ with respect to W in Step 4 is $O(n^2m + n^2l + nml + nl^2)$. Similarly, the complexity of Step 6 is $O(ml^2 + l^3 + nml + n^2l)$. Because $m > n > l$ in most cases, the worst total time complexity of Algorithm 1 is $O(n^2(n + m) + l^2 (n + m) + nml)$.

Algorithm 1. MLAPG

Input: Training data set $X \in R^{m \times n}$; training label set $Y \in R^{n \times l}$; parameters λ_1, λ_2, λ_3 and λ_4.

Output: Optimal solution W^* and $\mathbb{C}^*$.

1. Initialise W_0, W_1 = rand (n, l); $\mathbb{C}_0, \mathbb{C}_1$ = zeros (l, l); $\Phi = \{W, \mathbb{C}\}$; $t = 1$.
2. **WHILE** not converged
3. Calculate Lipchitz constant L_f according to Theorem 1.
4. Update $W^{(t)}$ and W_{t+1} with Eq. (15).
5. Let $W^{(t+1)} = W^{(t)}$.
6. Update $\mathbb{C}^{(t)}$ and $\mathbb{C}_{t+1}$ with Eq. (16).
7. Let $\mathbb{C}^{(t+1)} = \mathbb{C}^{(t)}$.
8. Let $\alpha_{t+1} = (1 + \sqrt{4\alpha_t^2 + 1}) / 2$.
9. Let $t = t + 1$.
10. **END WHILE**
11. $W^* = W_t$.
12. $\mathbb{C}^* = \mathbb{C}_t$.

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 $MFNDS = <U, C, D, \delta^F>$, fuzzy neighborhood parameter δ^F.

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 FN_B^δ with Eq. (19).
4. **FOR** $f_k \in C$
5. Calculate $REL(f_k, L)$ with Eq. (28), where $l_i \in D$.
6. Calculate $RED(f_k, S)$ with Eq. (29), where $l_i \in D$.
7. Calculate $LCD(l_i, L)$ with Eq. (30), where $l_i, L \in D$.
8. Find f_k satisfying Eq. (31).
9. Let $S = S \cup \{f_k\}$ and $C = C - \{f_k\}$.
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 $O(n^2(n + m) + l^2 (n + m) + nml)$. 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 $O(m\log m + n)$, where the complexity of calculating the label correlation is $O(nm\log m)$. In addition, there exists a loop in Step 4. Therefore, in the worst case, the total time complexity of MFSMR is $O(n^2(n + m) + l^2 (n + m) + nml + n^2 m\log m)$.

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.

TABLE I
DESCRIPTION OF THE TWELVE MULTILABEL DATASETS

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

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

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], [35], [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 λ_1, λ_2, λ_3, and λ_4 of MLAPG for the training samples of each dataset were set to values $\{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$. 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

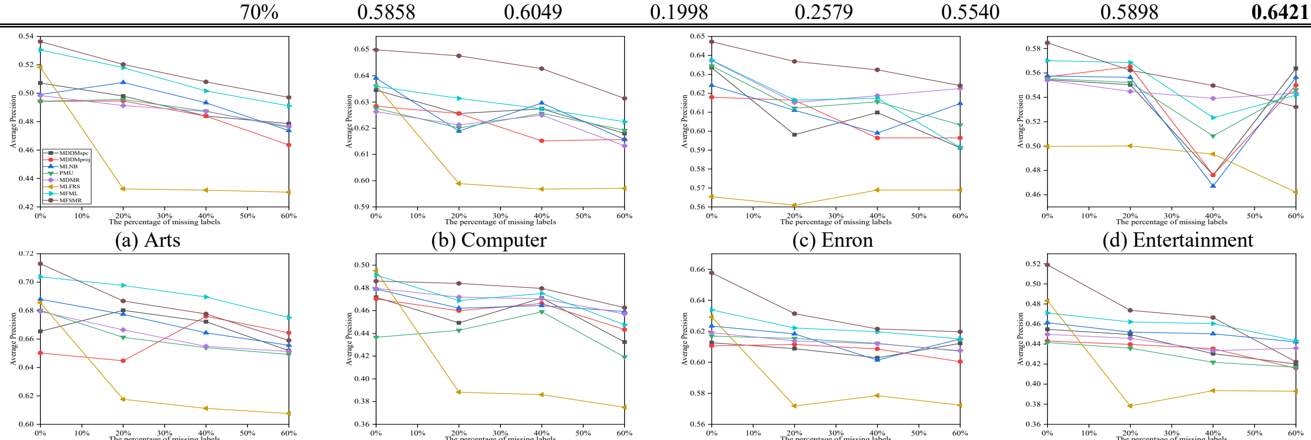
The first part of this subsection demonstrates the efficiency of MFSMR on eight multilabel datasets in terms of AP, CV, OE, and RL. MFSMR was compared with seven state-of-the-art multilabel feature selection algorithms with missing labels: MDDM_proj [44], MDDM_spc [44], MLNB [38], MDMR [45], MLFRS [46], 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

TABLE II
AP (↑) VALUES OF THE SEVEN METHODS ON THE THIRTEEN DATASETS WITH DIFFERENT MISSING LABELS PERCENTAGES

Datasets	*p*	MLKNN	MLMF	MLNB	CDN-LR	sCDN-LR	GLOCAL	MLAPG
Arts	10%	0.5214	0.6156	0.2347	0.3345	0.5881	0.6023	**0.6259**
	30%	0.5330	0.6107	0.2348	0.2901	0.5814	0.5972	**0.6201**
	50%	0.5312	0.5975	0.2321	0.2808	0.5626	0.5838	**0.6091**
	70%	0.5213	0.5770	0.2210	0.2066	0.4785	0.5578	**0.5898**
Business	10%	0.8786	0.8888	0.2470	0.3086	0.8679	0.8728	**0.8898**
	30%	0.8776	0.8862	0.2234	0.2997	0.8685	0.8731	**0.8879**
	50%	0.8754	0.8825	0.1925	0.2840	0.8630	0.8721	**0.8871**
	70%	0.8715	0.8769	0.1471	0.2014	0.8544	0.8686	**0.8865**
Computer	10%	0.6278	0.7025	0.1826	0.2697	0.6782	0.6812	**0.7163**
	30%	0.6229	0.6944	0.1696	0.2233	0.6721	0.6746	**0.7175**
	50%	0.6153	0.6799	0.1440	0.2022	0.6519	0.6621	**0.7139**
	70%	0.6204	0.6546	0.1205	0.1489	0.5965	0.6382	**0.7115**
Education	10%	0.5942	0.6387	0.1846	0.5779	0.6119	0.6235	**0.6427**
	30%	0.5871	0.6302	0.1698	0.3657	0.6087	0.6176	**0.6358**
	50%	0.5780	**0.6212**	0.1204	0.1830	0.5863	0.6064	0.6196
	70%	0.5671	**0.5997**	0.0923	0.1401	0.5197	0.5776	0.5909
Entertainment	10%	0.6042	0.6853	0.2701	0.6021	0.6670	0.6699	**0.6922**
	30%	0.5979	0.6777	0.2752	0.5952	0.6551	0.6626	**0.6797**
	50%	0.5977	**0.6677**	0.1996	0.3028	0.6339	0.6535	0.6642
	70%	0.5865	**0.6491**	0.1671	0.2353	0.5761	0.6314	0.6349
Health	10%	0.7080	0.7862	0.1037	0.1454	0.7616	0.7685	**0.7929**
	30%	0.7218	0.7806	0.1020	0.1342	0.7581	0.7624	**0.7907**
	50%	0.7239	0.7707	0.0978	0.1187	0.7417	0.7624	**0.7902**
	70%	0.7156	0.7502	0.0911	0.0978	0.6812	0.7284	**0.7880**
Medical	10%	0.7857	0.8931	0.0672	0.2084	0.8846	0.8675	**0.9006**
	30%	0.7617	**0.8879**	0.0592	0.1453	0.8698	0.8399	0.8819
	50%	0.7337	0.8759	0.0523	0.0958	0.8401	0.7845	**0.8772**
	70%	0.6878	0.8471	0.0461	0.0641	0.7848	0.6903	**0.8550**
Recreation	10%	0.4493	0.6248	0.5077	0.5824	0.5973	0.6119	**0.6422**
	30%	0.4436	0.6142	0.4847	0.5540	0.5796	0.6022	**0.6329**
	50%	0.4681	0.6025	0.4571	0.4977	0.5602	0.5858	**0.6163**
	70%	0.4950	0.5733	0.3646	0.3622	0.4625	0.5476	**0.5942**
Reference	10%	0.6141	0.7118	0.1093	0.2530	0.6972	0.6935	**0.7217**
	30%	0.6134	0.7020	0.0938	0.1839	0.6876	0.6849	**0.7108**
	50%	0.6324	0.6881	0.0782	0.1299	0.6475	0.6849	**0.6882**
	70%	0.6361	0.6579	0.0656	0.0876	0.6034	0.6390	**0.6695**
Scene	10%	0.8475	0.8545	0.8174	0.8298	0.8186	0.8266	**0.8586**
	30%	0.8422	0.8513	0.8061	0.8219	0.8152	0.8208	**0.8520**
	50%	0.8379	0.8487	0.7893	0.8151	0.8092	0.8087	**0.8492**
	70%	0.7994	0.8245	0.6234	0.8058	0.8072	0.5573	**0.8273**
Science	10%	0.5298	0.5891	0.2082	0.3135	0.5546	0.5791	**0.6067**
	30%	0.5212	0.5799	0.1715	0.2595	0.5554	0.5691	**0.6001**
	50%	0.5106	0.5661	0.1455	0.2302	0.5276	0.5538	**0.5894**
	70%	0.4898	0.5385	0.1051	0.1435	0.4351	0.5237	**0.5475**
Social	10%	0.7483	0.7734	0.1436	0.2546	0.7568	0.7576	**0.7849**
	30%	0.7438	0.7667	0.1353	0.2433	0.7495	0.7527	**0.7668**
	50%	0.7393	0.7574	0.1184	0.1745	0.7353	0.7434	**0.7629**
	70%	0.7263	0.7374	0.0925	0.1424	0.6990	0.7234	**0.7468**
Society	10%	0.6131	0.6357	0.3236	0.3595	0.6175	0.6239	**0.6462**
	30%	0.6062	0.6290	0.3200	0.3409	0.6135	0.6161	**0.6448**
	50%	0.5978	0.6193	0.2664	0.2958	0.5993	0.6065	**0.6422**
	70%	0.5858	0.6049	0.1998	0.2579	0.5540	0.5898	**0.6421**



(a) Arts (b) Computer (c) Enron (d) Entertainment

(e) Health (f) Recreation (g) Reference (h) Science

Fig. 1. *AP* index of the eight algorithms on eight multilabel datasets with different missing percentages.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

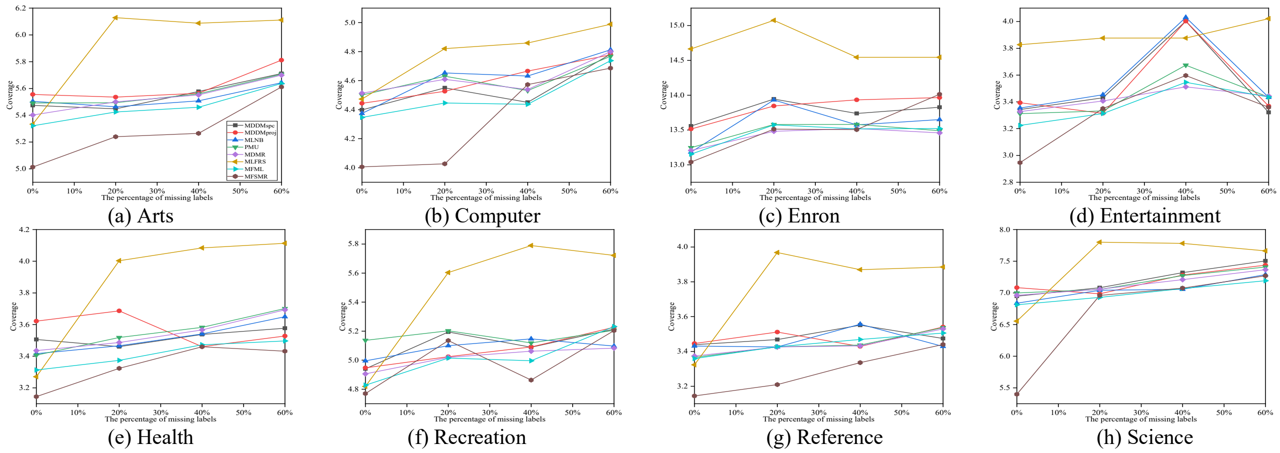


Fig. 2. *CV* index of the eight algorithms on eight multilabel datasets with different missing percentages.

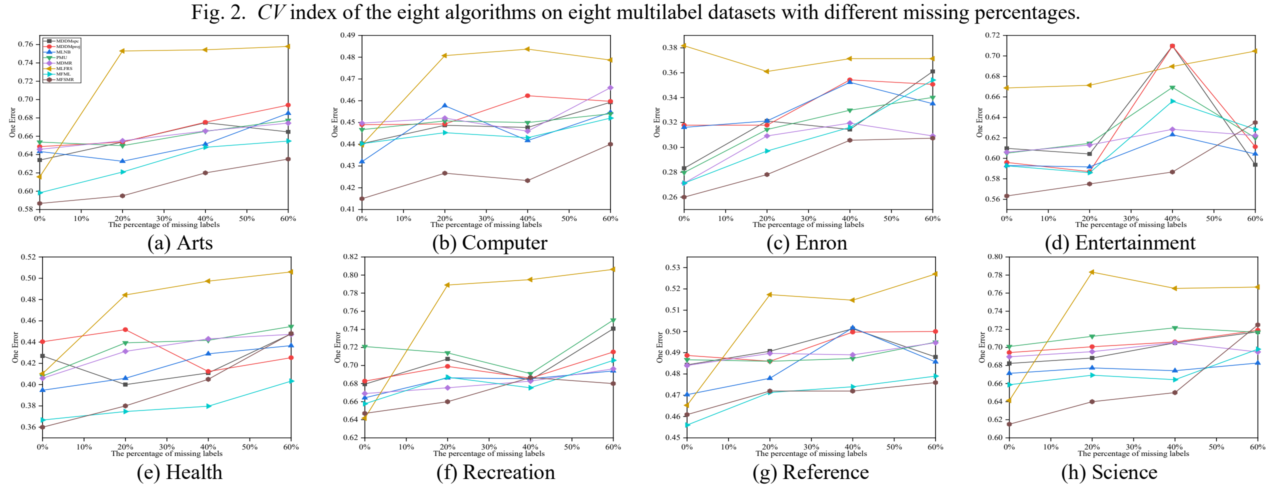


Fig. 3. *OE* index of the eight algorithms on eight multilabel datasets with different missing percentages.

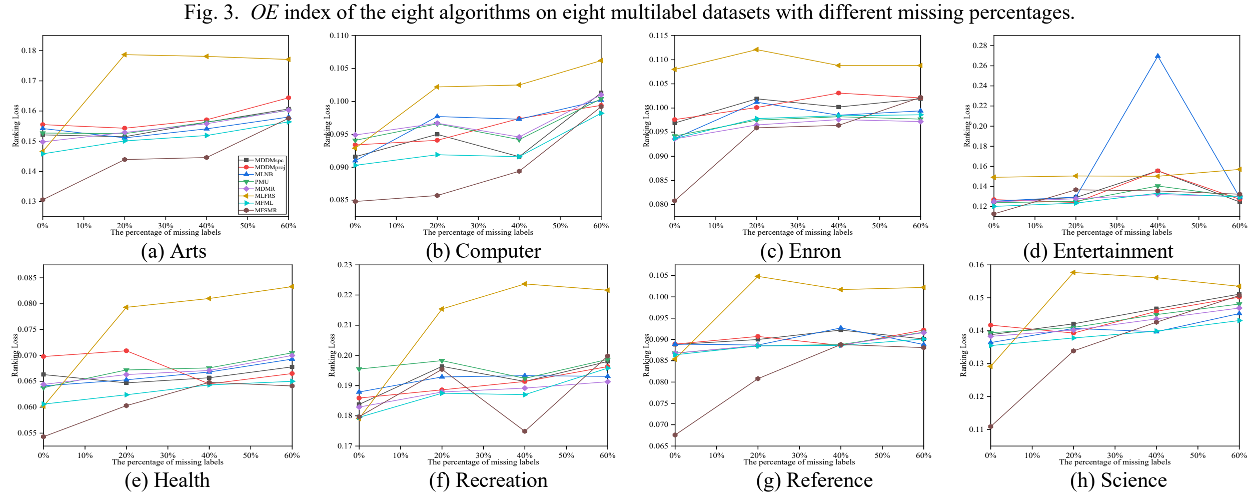


Fig. 4. *RL* index of the eight algorithms on eight multilabel datasets with different missing percentages.

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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

TABLE III
CLASSIFICATION RESULTS OF NINE ALGORITHMS IN TERMS OF FOUR METRICS ON SIX DATASETS WITH 80% MISSING LABELS

Metrics	Datasets	MDDM	PMU	SFUS	MDMR	MLMLFS (p = 0.4)	MLMLFS (p = 0.6)	MLMLFS (p = 0.8)	MLMLFS (p = 1)	MFSMR
	Artificial	0.4410	0.4454	0.4522	0.4531	0.5100	0.5213	0.5118	0.5080	**0.5269**
	Birds	0.6077	0.6087	0.6191	0.6130	0.6451	0.6599	0.6628	**0.6658**	0.6581
AP	Bookmarks	0.2556	0.2488	0.2719	0.2630	0.4117	0.4466	0.4554	0.4521	**0.4565**
(↑)	Reference	0.5728S	0.577	0.5871	0.5722	0.6135	0.6149	0.6164	0.6135	**0.6275**
	Social	0.6535	0.6624	0.6568	0.6691	0.7068	0.7114	0.7113	0.7123	**0.7214**
	Yeast	0.7228	0.7192	0.7244	0.7209	0.7440	**0.7443**	0.7433	**0.7443**	0.7428
	Artificial	0.7347	0.7130	0.7013	0.6960	0.6160	0.5930	0.6083	0.6093	**0.5800**
	Birds	0.5046	0.4551	0.4644	0.4551	0.4087	0.3963	0.3963	**0.3839**	0.4086
OE	Bookmarks	0.6090	0.5890	0.4290	0.5764	0.2063	0.1740	**0.1700**	0.1757	0.1833
(↓)	Reference	0.5220	0.5210	0.5093	0.5200	0.4733	**0.4727**	0.4743	0.4747	0.4800
	Social	0.4607	0.447	0.4583	0.4327	0.3697	0.3650	0.3603	**0.3593**	0.3650
	Yeast	0.2486	0.2465	0.2497	0.2454	0.2388	0.2410	0.2410	0.2410	**0.2050**
	Artificial	0.0630	0.0630	0.0629	0.0628	0.0592	**0.0586**	**0.0586**	**0.0586**	0.0588
	Birds	0.0661	0.0568	0.055	0.0568	0.0542	0.0540	0.0540	**0.0531**	0.0584
HL	Bookmarks	0.0385	0.0433	0.0375	0.0531	0.0324	0.03050	0.0300	0.0301	**0.0253**
(↓)	Reference	0.0351	0.0317	0.0336	0.0317	0.0293	0.0293	**0.0292**	0.0293	0.0305
	Social	0.0298	0.0296	0.0295	0.0291	0.0242	0.0241	0.0239	**0.0235**	0.0241
	Yeast	0.2217	0.2251	0.2224	0.2243	0.2053	0.2053	0.2053	**0.2052**	0.2077
	Artificial	0.0165	0.0292	0.043	0.0435	0.1598	0.1805	0.1623	0.1572	**0.1896**
	Birds	0.4220	0.4396	0.4303	0.4396	0.4752	0.4748	0.4787	**0.4966**	0.4925
MacF1	Bookmarks	0.1237	0.1453	0.1598	0.179	0.3227	0.3736	0.3852	0.3782	**0.4140**
(↑)	Reference	0.2908	0.2092	0.2839	0.2387	0.3987	0.3787	0.3518	0.3541	**0.4040**
	Social	0.266	0.3126	0.2118	0.3341	0.3987	0.4034	0.4209	0.4196	**0.4420**
	Yeast	0.5352	0.5408	0.5416	0.5448	0.6082	0.6082	**0.6139**	0.6094	0.6095

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 from top to bottom gradually, where the ratio of selected features is from 0.1 to 1 with a step size of 0.1. When the ratio is 1, all features are selected. Table IV shows that the five indices vary with m on datasets Bibtex, Corel5k, Enron, and Delicious selected from Table I, from which, MFSMR is better than the other five algorithms on dataset Bibtex in terms of *CV* and *RL*. For indices *AP* and MacF1, when $m = 1$, MFSMR and FSLCL exhibit little difference in performance. MFSMR is slightly inferior to FSLCL in terms of *OE*. For the Corel5k dataset, MFSMR exhibits great performance on most indices when $m = 2$ and 3; it is only second to FSLCL when $m = 1$, whereas it is better than the other algorithms in terms of *AP* and MacF1. As can be seen for dataset Delicious, when $m = 1$ and 2, MFSMR performs better than the other algorithms. When $m = 3$, the performance of MFSMR declines for most indices. For the *CV* index, there are no algorithms that have significant advantages over MLKNN. On the Enron dataset, MFSMR yields the best results in terms of the four metrics when compared with the other six methods. Overall, MFSMR outperforms the six other compared methods on these four multilabel datasets when different labels are masked. In general, MFSMR can select the most relevant features and realise excellent classification performance for multilabel datasets with missing labels.

D. Parameter analysis

Here, the parameter sensitivity of MFSMR for the four parameters λ_1, λ_2, λ_3, and λ_4 is analysed in detail, where λ_1 controls the difference between the recovered label matrix manifold and original label matrix, λ_2 controls the new label matrix manifold, λ_3 controls the sparsity of the feature matrix, and λ_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 Bibtex 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 λ_3 and λ_4 are increased. The reason for this is that with the increase of λ_3 and λ_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.

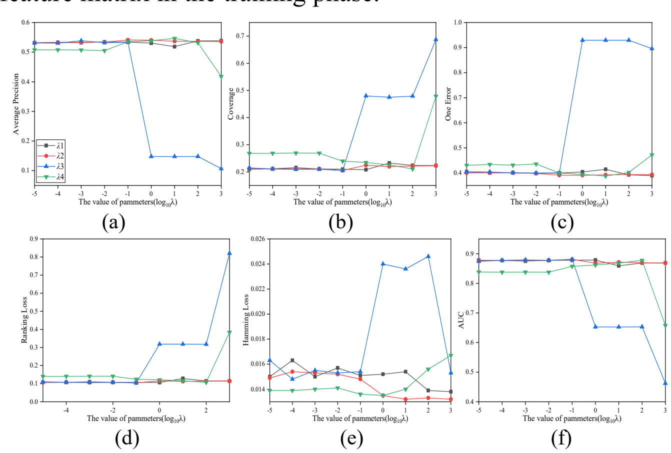


Fig. 5. Parameter sensitivity analysis on dataset Bibtex under 60% missing label percentage.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

TABLE IV
EVALUATION RESULTS OF SEVEN METHODS ON THE Bibtex DATASET WHILE MASKING DIFFERENT LABELS

Datasets	Metrics	*m*	MLKNN	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	MFSMR
Bibtex	*AP* (↑)	1	0.2321 ± 0.0051	0.2235 ± 0.0063	0.3215 ± 0.0216	0.2716 ± 0.0125	0.2196 ± 0.0135	**0.3593 ± 0.0095**	0.3480 ± 0.0205
		2	0.2151 ± 0.0064	0.2095 ± 0.0052	0.2814 ± 0.0149	0.2354 ± 0.0104	0.2120 ± 0.0132	0.3275 ± 0.0082	**0.3341 ± 0.0195**
		3	0.2095 ± 0.0059	0.2039 ± 0.0051	0.2689 ± 0.0166	0.2111 ± 0.0208	0.2079 ± 0.0134	0.3180 ± 0.0072	**0.3231 ± .01126**
	CV (↓)	1	80.386 ± 1.2707	77.975 ± 1.5388	63.300 ± 1.0405	68.650 ± 1.2134	75.848 ± 1.6193	61.365 ± 1.2954	**51.637 ± 1.8632**
		2	77.222 ± 1.4155	81.791 ± 1.5353	71.085 ± 1.5648	75.192 ± 1.1917	79.338 ± 1.0064	67.884 ± 1.3866	**55.876 ± 1.1327**
		3	80.741 ± 1.7322	82.365 ± 1.7928	73.213 ± 2.0608	77.696 ± 2.8423	79.706 ± 1.4055	70.627 ± 2.0682	**57.090 ± 2.4136**
	OE (↓)	1	0.6977 ± 0.0092	0.7169 ± 0.0101	0.6093 ± 0.0171	0.6657 ± 0.0210	0.7259 ± 0.0257	0.5736 ± 0.0161	**0.5640 ± 0.0212**
		2	0.7129 ± 0.0155	0.7242 ± 0.0121	0.6409 ± 0.0201	0.7025 ± 0.0181	0.7286 ± 0.0282	0.5973 ± 0.0122	**0.5888 ± 0.0768**
		3	0.7140 ± 0.0167	0.7278 ± 0.0110	0.6476 ± 0.0202	0.7399 ± 0.0161	0.7322 ± 0.0289	**0.5990 ± 0.0144**	0.6330 ± 0.0341
	RL (↓)	1	0.3420 ± 0.0084	0.3453 ± 0.0086	0.2734 ± 0.0087	0.3044 ± 0.0078	0.3366 ± 0.0111	0.2638 ± 0.0100	**0.2130 ± 0.0105**
		2	0.3545 ± 0.0079	0.3593 ± 0.0090	0.3055 ± 0.0097	0.3279 ± 0.0087	0.3495 ± 0.0066	0.2891 ± 0.0068	**0.2224 ± 0.0064**
		3	0.3523 ± 0.0096	0.3584 ± 0.0109	0.3120 ± 0.0135	0.3377 ± 0.0161	0.3475 ± 0.0078	0.2969 ± 0.0118	**0.2232 ± 0.0124**
	MacF1 (↑)	1	0.1221 ± 0.0044	0.0851 ± 0.0046	0.1836 ± 0.0103	0.1437 ± 0.0049	0.0949 ± 0.0097	**0.2158 ± 0.0073**	0.2004 ± 0.0195
		2	0.1094 ± 0.0044	0.0809 ± 0.0038	0.1566 ± 0.0096	0.1189 ± 0.0071	0.0926 ± 0.0100	0.1977 ± 0.0059	**0.2028 ± 0.0134**
		3	0.1047 ± 0.0050	0.0799 ± 0.0031	0.1458 ± 0.0099	0.1058 ± 0.0088	0.0897 ± 0.0083	0.1930 ± 0.0066	**0.1996 ± 0.0117**
Corel5k	*AP* (↑)	1	0.3536 ± 0.0066	0.3463 ± 0.0069	0.3703 ± 0.0015	0.3696 ± 0.0041	0.3680 ± 0.0047	**0.3771 ± 0.006 3**	0.3641 ± 0.0171
		2	0.3212 ± 0.0062	0.3157 ± 0.0075	0.3330 ± 0.0032	0.3289 ± 0.0043	0.3296 ± 0.0029	0.3361 ± 0.0031	**0.3401 ± 0.0223**
		3	0.2927 ± 0.0054	0.2893 ± 0.0047	0.3000 ± 0.0036	0.3010 ± 0.0049	0.3000 ± 0.0050	0.3059 ± 0.0042	**0.3180 ± 0.0190**
	CV (↓)	1	71.502 ± 1.3846	72.041 ± 1.3523	70.673 ± 1.4609	70.674 ± 1.3782	70.921 ± 1.1386	70.041 ± 1.1948	**70.032 ± 4.3070**
		2	79.127 ± 1.2547	79.380 ± 1.2170	78.457 ± 1.1582	78.657 ± 1.2312	78.634 ± 1.1286	**77.914 ± 1.2839**	78.952 ± 1.7185
		3	85.161 ± 1.0413	85.311 ± 1.0639	85.138 ± 0.9200	84.885 ± 1.0812	85.137 ± 1.1048	84.861 ± 1.0026	**84.548 ± 0.0475**
	OE (↓)	1	0.5930 ± 0.0116	0.5981 ± 0.0142	0.5662 ± 0.0121	0.5679 ± 0.0081	0.5713 ± 0.0093	**0.5507 ± 0.0122**	0.5609 ± 0.0349
		2	0.6243 ± 0.0145	0.6268 ± 0.0162	0.6073 ± 0.0111	0.6111 ± 0.0106	0.6083 ± 0.0145	0.6064 ± 0.0064	**0.5920 ± 0.0049**
		3	0.6559 ± 0.0113	0.6531 ± 0.0122	0.6369 ± 0.0135	0.6362 ± 0.0148	0.6382 ± 0.0142	0.6271 ± 0.0092	**0.6248 ± 0.0034**
	RL (↓)	1	0.1476 ± 0.0033	0.1492 ± 0.0031	0.1443 ± 0.0028	0.1442 ± 0.0027	0.1450 ± 0.0021	0.1422 ± 0.0025	**0.1411 ± 0.0077**
		2	0.1647 ± 0.0026	0.1660 ± 0.0026	0.1617 ± 0.0019	0.1626 ± 0.0020	0.1625 ± 0.0017	0.1600 ± 0.0026	**0.1584 ± 0.0034**
		3	0.1775 ± 0.0023	0.1785 ± 0.0021	0.1761 ± 0.0014	0.1757 ± 0.0018	0.1763 ± 0.0020	**0.1747 ± 0.0021**	0.1784 ± 0.0034
	MacF1 (↑)	1	0.1187 ± 0.0062	0.1122 ± 0.0065	0.1306 ± 0.0065	0.1311 ± 0.0079	0.1325 ± 0.0061	**0.1380 ± 0.0057**	0.1362 ± 0.0023
		2	0.1009 ± 0.0064	0.0950 ± 0.0080	0.1099 ± 0.0055	0.1067 ± 0.0072	0.1068 ± 0.0072	0.1134 ± 0.0036	**0.1242 ± 0.0023**
		3	0.0744 ± 0.0067	0.0741 ± 0.0061	0.0820 ± 0.0038	0.0833 ± 0.0057	0.0825 ± 0.0049	0.0884 ± 0.0050	**0.0914 ± 0.0117**
Delicious	*AP* (↑)	1	0.3230 ± 0.0027	0.2746 ± 0.0135	0.3089 ± 0.0043	0.3027 ± 0.0035	0.2547 ± 0.0017	0.3288 ± 0.0026	**0.3300 ± 0.0093**
		2	0.3200 ± 0.0025	0.2724 ± 0.0139	0.3035 ± 0.0033	0.2991 ± 0.0029	0.2590 ± 0.0024	0.3256 ± 0.0025	**0.3264 ± 0.0102**
		3	0.3169 ± 0.0025	0.2704 ± 0.0136	0.3003 ± 0.0031	0.2952 ± 0.0034	0.2525 ± 0.0016	**0.3226 ± 0.0029**	0.3009 ± 0.0130
	CV (↓)	1	**606.01 ± 0.7518**	638.72 ± 0.0479	625.28 ± 3.6584	629.77 ± 3.7979	654.66 ± 2.9751	608.76 ± 3.1202	607.57 ± 2.3586
		2	**610.87 ± 0.5807**	642.05 ± 5.4990	630.01 ± 3.5864	633.48 ± 3.3993	656.10 ± 2.4705	613.15 ± 3.2041	612.28 ± 4.2212
		3	**615.92 ± 0.7199**	645.09 ± 2.4013	634.14 ± 3.4401	637.24 ± 2.63425	657.99 ± 2.9015	618.11 ± 3.1157	616.68 ± 3.5524
	OE (↓)	1	0.3980 ± 0.0050	0.4629 ± 0.0226	0.4242 ± 0.0110	0.4359 ± 0.0048	0.5135 ± 0.0055	0.3868 ± 0.0092	**0.3866 ± 0.0171**
		2	0.4039 ± 0.0077	0.4697 ± 0.0200	0.4349 ± 0.0084	0.4415 ± 0.0097	0.5165 ± 0.0077	**0.3963 ± 0.0067**	0.4067 ± 0.0181
		3	**0.4109 ± 0.0061**	0.4750 ± 0.0205	0.4408 ± 0.0100	0.4480 ± 0.0086	0.5202 ± 0.0072	0.4028 ± 0.0080	0.4097 ± 0.0154
	RL (↓)	1	0.1277 ± 0.0012	0.1424 ± 0.0035	0.1343 ± 0.0013	0.1365 ± 0.0015	0.1497 ± 0.0010	0.1270 ± 0.0009	**0.1268 ±0.0069**
		2	0.1287 ± 0.0011	0.1437 ± 0.0037	0.1361 ± 0.0012	0.1377 ± 0.0014	0.1501 ± 0.0009	**0.1280 ± 0.0008**	0.1300 ± 0.0074
		3	0.1298 ± 0.0011	0.1441 ± 0.0036	0.1372 ± 0.0012	0.1390 ± 0.0014	0.1506 ± 0.0009	**0.1292 ± 0.0009**	0.1306 ± 0.0090
	MacF1 (↑)	1	0.1034 ± 0.0020	0.0585 ± 0.0079	0.0878 ± 0.0022	0.0841 ± 0.0027	0.0511 ± 0.0019	0.1040 ± 0.0015	**0.1339 ± 0.0016**
		2	0.1017 ± 0.0016	0.0570 ± 0.0077	0.0851 ± 0.0025	0.0825 ± 0.0026	0.0504 ± 0.0019	0.1017 ± 0.0012	**0.1097 ± 0.0092**
		3	0.1001 ± 0.0014	0.0561 ± 0.0073	0.0835 ± 0.0018	0.0802 ± 0.0023	**0.0502 ± 0.0016**	0.1004 ± 0.0024	0.0992 ± 0.0133
Enron	*AP* (↑)	1	0.5577 ± 0.0104	0.5634 ± 0.0102	0.5327 ± 0.0155	0.5343 ± 0.0099	0.5368 ± 0.0097	0.5891 ± 0.0110	**0.6270 ± 0.0215**
		2	0.5446 ± 0.0122	0.5466 ± 0.0126	0.5202 ± 0.0165	0.5237 ± 0.0089	0.5250 ± 0.0209	0.5668 ± 0.0099	**0.6008 ± 0.0187**
		3	0.5386 ± 0.0120	0.5372 ± 0.0115	0.5207 ± 0.0105	0.5239 ± 0.0104	0.5206 ± 0.0124	0.5592 ± 0.0124	**0.5884 ± 0.0255**
	CV (↓)	1	14.039 ± 0.3801	14.221 ± 0.3766	14.875 ± 0.3517	14.926 ± 0.3486	14.751 ± 0.2796	13.781 ± 0.4453	**13.334 ± 0.3295**
		2	14.387 ± 0.3980	14.6295 ± 0.382	15.113 ± 0.5018	15.317 ± 0.3789	15.034 ± 0.3942	14.261 ± 0.4565	**13.661 ± 0.5258**
		3	14.678 ± 0.4034	14.930 ± 0.3940	15.215 ± 0.3982	15.350 ± 0.4335	15.227 ± 0.2661	14.684 ± 0.4261	**14.129 ± 0.7115**
	OE (↓)	1	0.3826 ± 0.0202	0.3761 ± 0.0190	0.4168 ± 0.0215	0.4074 ± 0.0192	0.4168 ± 0.0245	0.3585 ± 0.0146	**0.3026 ± 0.0286**
		2	0.4070 ± 0.0305	0.3978 ± 0.0221	0.4419 ± 0.0323	0.4192 ± 0.0233	0.4444 ± 0.0452	0.3810 ± 0.0234	**0.3324 ± 0.0193**
		3	0.4211 ± 0.0301	0.4082 ± 0.0308	0.4554 ± 0.0256	0.4323 ± 0.0425	0.4595 ± 0.0442	0.3781 ± 0.0180	**0.3516 ± 0.0261**
	RL (↓)	1	0.1036 ± 0.0046	0.1047 ± 0.0048	0.1122 ± 0.0048	0.1125 ± 0.0050	0.1106 ± 0.0038	0.0985 ± 0.0065	**0.0963 ± 0.0033**
		2	0.1071 ± 0.0050	0.1090 ± 0.0050	0.1148 ± 0.0066	0.1165 ± 0.0042	0.1140 ± 0.0060	0.1041 ± 0.0054	**0.0993 ± 0.0056**
		3	0.1089 ± 0.0053	0.1118 ± 0.0057	0.1152 ± 0.0052	0.1166 ± 0.0054	0.1152 ± 0.0043	0.1076 ± 0.0059	**0.1044 ± 0.0073**
	MacF1 (↑)	1	0.1095 ± 0.0080	0.1076 ± 0.0091	0.0832 ± 0.0044	0.0753 ± 0.0034	0.0905 ± 0.0084	0.1243 ± 0.0085	**0.1454 ± 0.0106**
		2	0.1097 ± 0.0115	0.0987 ± 0.0060	0.0818 ± 0.0078	0.0709 ± 0.0023	0.0884 ± 0.0092	0.1158 ± 0.0073	**0.1272 ± 0.0136**
		3	0.1030 ± 0.0130	0.0962 ± 0.0110	0.0835 ± 0.0099	0.0717 ± 0.0056	0.0849 ± 0.0051	0.1135 ± 0.0056	**0.1186 ± 0.0147**

E. Statistical analysis

To analyse the statistical performance among all the compared algorithms on each evaluation metrics, the Friedman test and Nemenyi test [52] were employed for performance analysis. The Friedman statistic is expressed as follows:

$$\chi_F^2 = \frac{12T}{s(s+1)}\left(\sum_{i=1}^{s} R_i^2 - \frac{s(s+1)^2}{4}\right) \text{ and } F_F = \frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2}, \quad (32)$$

where T and s are the numbers of datasets and methods, respectively; R_i ($i = 1, 2, \cdots, s$) represents the mean rank of the i-th methods 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

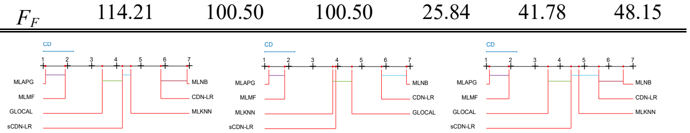
critical difference among these methods is described as $CD_\alpha = q_\alpha \sqrt{\frac{s(s+1)}{6T}}$, where q_α denotes the critical tabulated value.

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.

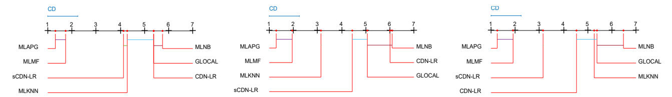
From the aforementioned Table II and Tables I–V in the supplementary file, the F_F and χ_F^2 results are list in Table V. Comparison of MFSMR against the other algorithms with the Nemenyi test is displayed in Fig. 6, where q_α = 2.693 at a significance level α = 0.1, and the CD = 2.2818 (s = 7, T = 13). 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 χ_F^2 and F_F in terms of four metrics and the null hypothesis at α = 0.1, q_α = 2.780, and CD = 2.3131 (s = 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_α = 2.855 when α = 0.1, and CD = 4.5142 (s = 9, T = 6). As shown in Fig. 8, MFMSR clearly outperforms MLMLFS (p = 0.1), MLMLFS (p = 0.4), MLMLFS (p = 0.6), MLMLFS (p = 0.8), SFUS, MDMR, PMU, and MDDM in metrics of all evaluation indices. Thus, there is no significant difference among MLMLFS (p = 0.1), MLMLFS (p = 0.4), MLMLFS (p = 0.6), and MLMLFS (p = 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 α = 0.1, the null hypothesis of equal performance among the seven methods is rejected under the Friedman test. q_α = 2.693 when α = 0.1, and thus, CD = 4.114 (s = 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
STATISTICAL RESULTS OF SEVEN METHODS IN TERMS OF SIX METRICS

	AP	*CV*	*OE*	*RL*	*HL*	AUC
χ_F^2	70.58	69.68	69.68	53.27	60.59	62.44
F_F	114.21	100.50	100.50	25.84	41.78	48.15



(a) *AP* (b) *CV* (c) *OE*

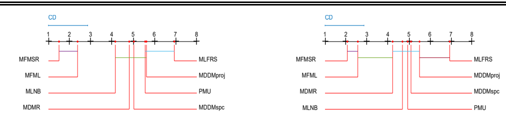


(d) *RL* (e) *HL* (f) AUC

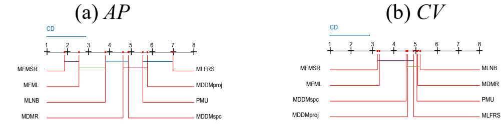
Fig. 6. Comparison between MLAPG and the other six algorithms under the Nemenyi test.

TABLE VI
STATISTICAL RESULTS OF EIGHT METHODS IN TERMS OF FOUR METRICS

	AP	*CV*	*OE*	*RL*
χ_F^2	29.91	23.10	27.53	5.90
F_F	8.02	4.92	6.77	0.82



(a) *AP* (b) *CV*

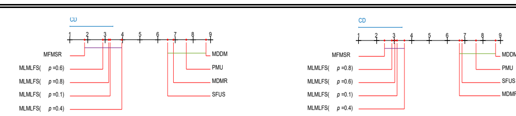


(c) *OE* (d) *RL*

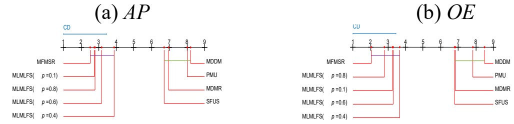
Fig. 7. Comparison between MFMSR and the other eight algorithms under the Nemenyi test.

TABLE VII
STATISTICAL RESULTS OF EIGHT METHODS IN TERMS OF FOUR METRICS

	AP	*OE*	*HL*	MacF1
χ_F^2	38.67	37.34	37.26	38.49
F_F	20.73	17.51	17.34	20.24



(a) *AP* (b) *OE*

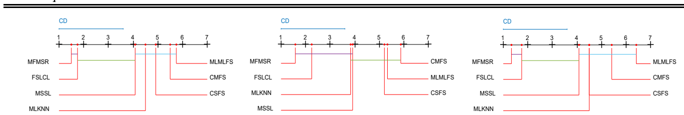


(c) *HL* (d) MacF1

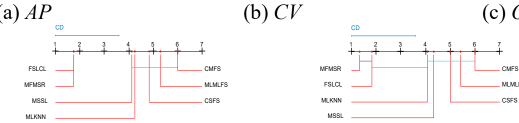
Fig. 8. Comparison between MFMSR and the other nine algorithms under the Nemenyi test.

TABLE VIII
STATISTICAL RESULTS OF EIGHT METHODS IN TERMS OF FIVE METRICS

	AP	*CV*	*OE*	*RL*	MacF1
χ_F^2	14.49	16.21	17.63	14.84	16.19
F_F	4.57	6.24	8.31	4.86	6.22



(a) *AP* (b) *CV* (c) *OE*



(d) *RL* (e) MacF1

Fig. 9. Comparison between MFMSR and the other seven algorithms under the Nemenyi test.

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 MFNRS model was constructed by combining MNRS with FNRS. 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

[1] H. Lim and D. W. Kim, "MFC: initialization method for multi-label feature selection based on conditional mutual information," *Neurocomputing*, vol. 382, pp. 40–51, 2020.

[2] K. M. Ibrahim, E. V. Epure, G. Peeters, and G. Richard, "Confidence-based weighted loss for multi-label classification with missing labels," 2020 *International Conference on Multimedia Retrieval*, Dublin, Ireland, pp. 291–295, 2020.

[3] M. Paniri, M. B. Dowlatshahi, and H. Nezamabadi-Pour, "MLACO: a multi-label feature selection algorithm based on ant colony optimization," *Knowledge Based Systems*, vol. 192, article ID. 105285, 2020.

[4] L. Hu, Y. H. Li, W. F. Gao, P. Zhang, and J. C. Hu, "Multi-label feature selection with shared common mode," *Pattern Recognition*, vol. 104, article ID. 107344, 2020.

[5] L. Sun, T. Y. Yin, W. P. Ding, Y. H. Qian, and J. C. Xu, "Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems," *Information Sciences*, vol. 537, pp. 401–424, 2020.

[6] C. Liu, Q. Ma, and J. H. Xu, "Multi-label feature selection method combining unbiased Hilbert-Schmidt independence criterion with controlled genetic algorithm," *International Conference on Neural Information Processing*, Siem Reap, Cambodia, pp. 3–14, 2018.

[7] H. T. Xu and L. Y. Xu, "Multi-label feature selection algorithm based on label pairwise ranking comparison transformation," 2017 *International Joint Conference on Neural Networks*, Anchorage, Alaska, USA, pp. 1210–1217, 2017.

[8] A. Braytee, W. Liu, A. Anaissi, and P. J. Kennedy, "Correlated multi-label classification with incomplete label space and class imbalance," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, pp. 1–26, 2019.

[9] J. Huang, F. Qin, X. Zheng, Z. K. Cheng, Z. X. Yuan, W. G. Zhang, and Q. M. Huang, "Improving multi-label classification with missing labels by learning label-specific features," *Information Sciences*, vol. 492, pp. 124–146, 2019.

[10] C. X. Wang, Y. J. Lin, and J. H. Liu, "Feature selection for multi-label learning with missing labels," *Applied Intelligence*, vol. 49, no. 8, pp. 3027–3042, 2019.

[11] P. F. Zhu, Q. Xu, Q. H. Hu, C. Q. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognition*, vol. 74, pp. 488–502, 2018.

[12] J. H. Ma and T. W. S. Chow, "Robust non-negative sparse graph for semi-supervised multi-label learning with missing labels," *Information Sciences*, vol. 422, pp. 336–351, 2018.

[13] L. Jiang, G. Yu, M. Guo, and J. Wang, "Feature selection with missing labels based on label compression and local feature correlation," *Neurocomputing*, vol. 395, pp. 95–106, 2020.

[14] M. L. Zhang and L. Wu, "Lift: multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.

[15] J. Huang, G. R. Li, Q. M. Huang, and X. D. Wu, "Learning label specific features for multi-label classification," 15th IEEE *International Conference on Data Mining*, Atlantic, City, NJ, USA, pp. 181–190, 2015.

[16] J. H. Liu, Y. J. Lin, Y. W. Li, W. Weng, and S. X. Wu, "Online multi-label streaming feature selection based on neighborhood rough set," *Pattern Recognition*, vol. 84, pp. 273–287, 2018.

[17] C. Z. Wang, M. M. Shao, Q. He, Y. H. Qiang, and Y. L. Qi, "Feature subset selection based on fuzzy neighborhood rough sets," *Knowledge Based Systems*, vol. 111, pp. 173–179, 2016.

[18] L. Sun, L. Y. Wang, W. P. Ding, Y. H. Qian, and J. C. Xu, "Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 19–33, 2021.

[19] P. Chen, M. Lin, and J. Liu, "Multi-label attribute reduction based on variable precision fuzzy neighborhood rough set," *IEEE Access*, vol. 8, pp. 133565–133576, 2020.

[20] J. Duan, Q. H. Hu, L. J. Zhang, Y. H. Qian, and D. Y. Li, "Feature selection for multi-label classification based on neighborhood rough sets," *Chinese Journal of Computer Research and Development*, vol. 52 no. 1, pp. 56–65, 2015.

[21] S. Vluymans, C. Cornelis, F. Herrera, and Y. Saeys, "Multi-label classification using a fuzzy rough neighborhood consensus," *Information Sciences*, vol. 433–434, pp. 96–114, 2018.

[22] J. Ircio, A. Lojo, U. Mori, and J. A. Lozano, "Mutual information based feature subset selection in multivariate time series classification," *Pattern Recognition*, vol. 108, article ID. 107525, 2020.

[23] J. Gonzalez-Lopez, S. Ventura, and A. Cano, "Distributed selection of continuous features in multilabel classification using mutual information," *IEEE Transactions on Neural Networks*, vol. 31, no. 7, pp. 2280–2293, 2020.

[24] W. B Qian, J. T. Huang, Y. L. Wang, and W. H. Shu, "Mutual information-based label distribution feature selection for multi-label learning," *Knowledge-Based Systems*, vol. 195, article ID. 105684, 2020.

[25] J. H. Dai and J. L Chen, "Feature selection via normative fuzzy information weight with application into tumor classification," *Applied Soft Computing*, vol. 92, article ID. 106299, 2020.

[26] Y. B. Zhang, Y. C. Ma, and X. F. Yang, "Multi-label feature selection based on mutual information," in: 14th International Conference on Natural Computation, *Fuzzy Systems and Knowledge Discovery*, pp. 1379–1386, 2018.

[27] Y. Y. Wang and J. H Dai, "Label distribution feature selection based on mutual information in fuzzy rough set theory," 2019 *International Joint Conference on Neural Network*, Budapest, Hungary, pp. 1–2, 2019.

[28] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2003.

[29] J. C. Xu, Y. Wang, H. Y. Mu, and F. Z. Huang, "Feature genes selection based on fuzzy neighborhood conditional entropy," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 1, pp. 117–126, 2019.

[30] W. P. Ding, C. T. Lin and W. Pedrycz, "Multiple relevant feature ensemble selection based on multilayer co-evolutionary consensus mapreduce," *IEEE Transactions on Cybernetics*, vol. 20, no. 2, pp. 425–439, 2020.

[31] Y. J. Lin, Q. H. Hu, J. H. Liu, J. J. Li, and X. D. Wu, "Streaming feature selection for multilabel learning based on fuzzy mutual information," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1491–1507, 2017.

[32] P. Bugata and P. Drotar, "On some aspects of minimum redundancy maximum relevance feature selection," *Science China Information Sciences*, vol. 63, no. 1, pp. 85–99, 2020.

[33] A. Agarwal and A. Gupta, "A maximum relevancy and minimum

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3053844, IEEE Transactions on Fuzzy Systems

redundancy feature selection approach for median filtering forensics," *Multimedia Tools and Applications*, vol. 79, no. 29–30 pp. 1–28, 2020.

[34] M. L. Zhang and Z. H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[35] S. Kashef and H. Nezamabadi-pour, "A label-specific multi-label feature selection algorithm based on the Pareto dominance concept," *Pattern Recognition*, vol. 88, pp. 654–667, 2019.

[36] P. Zhang, G. X. Liu, and W. F. Gao, "Distinguishing two types of labels for multi-label feature selection," *Pattern Recognition*, vol. 95, pp. 72–82, 2019.

[37] Z. F. He, M. Yang, Y. Gao, H. D. Liu, and Y. L. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowledge-Based Systems*, vol. 163, pp. 145–158, 2019.

[38] M. L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.

[39] Y. H. Guo and S. C. Gu, "Multi-label classification using conditional dependency networks," 2011 *International Joint Conference on Artificial Intelligence*, Barcelona, pp. 1300–1305, 2011.

[40] Y. H. Guo, and W. Xue, "Probabilistic multi-label classification with sparse feature learning," 2013 *International Joint Conference on Artificial Intelligence*, Beijing, China, pp. 1373–1379, 2013.

[41] Y. Zhu, J. T. Kwok, and Z. H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1081–1094, 2018.

[42] L. P. Jing, L. Yang, J. Yu, and K. M. Ng, "Semi-supervised low-rank mapping learning for multi-label classification," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1483–1491, 2015.

[43] Q. Y. Tan, G. X. Yu, C. Domeniconi, J. Wang, and Z. L. Zhang, "Incomplete multi-view weak-label learning," *Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, 2017, pp. 2703–2709.

[44] Y. Zhang and Z. H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 3, article ID.14, 2010.

[45] Y. J. Lin, Q. H. Hu, J. H. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing*, vol. 168, pp. 92–103, 2015.

[46] Y. J. Lin, Y. W. Li, C. X. Wang, and J. K. Chen, "Attribute reduction for multi-label learning with fuzzy rough set," *Knowledge-Based Systems*, vol. 152, pp. 51–61, 2018.

[47] J. Lee and D. W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognition Letters*, vol. 34, no. 3, pp. 349–357, 2013.

[48] Z. G. Ma, F. P. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1021–1030, 2012.

[49] A. Braytee, W. Liu, D. R. Catchpoole, and P. J. Kennedy, "Multi-label feature selection using correlation information," 2017 *ACM Conference on Information and Knowledge Management*, Singapore, pp. 1649–1656, 2017.

[50] Z. L. Cai and W. Zhu, "Multi-label feature selection via feature manifold learning and sparsity regularization," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1321–1334, 2018.

[51] X. J. Chang, F. P. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," 28th *AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, pp. 1171–1177, 2014.

[52] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Lin Sun 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. He has become a Postdoctor with Henan Normal University, China, in 2019. He has received funding from ten grants from the National Natural Science Foundation of China, the China Postdoctoral Science Foundation, etc. His main research interests include rough sets, granular computing and big data mining. He has received the title of Henan's Distinguished Young Scholars for Science and Technology Innovation Talents, and has served as a reviewer for several prestigious peer-reviewed international journals.

Tengyu Yin is currently a postgraduate student in Computer Science and Technology at the College of Computer and Information Engineering with Henan Normal University. She received the B.Sc. degree in Computer Science and Technology from Henan Normal University in 2018. Her main interests include multilabel learning and data mining.

Weiping Ding (M'16-SM'19) received the Ph.D. degree in Computation Application, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013. He was a Visiting Scholar at University of Lethbridge (UL), Alberta, Canada, in 2011. From 2014 to 2015, he is a Postdoctoral Researcher at the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, Taiwan. In 2016, he was a Visiting Scholar at National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney (UTS), Ultimo, NSW, Australia. He is a member of Senior IEEE, IEEE-CIS, and Senior CCF. Dr. Ding is the Chair of IEEE CIS Task Force on Granular Data Mining for Big Data. He is a member of Technical Committee on Soft Computing of IEEE SMCS, a member of Technical Committee on Granular Computing of IEEE SMCS, a member of Technical Committee on Data Mining and Big Data Analytics Technical Committee of IEEE CIS. His main research directions involve deep learning, data mining, evolutionary computing, granular computing and big data analytics. He has co-authored more than 100 peer-reviewed journal and conference papers in these fields. Dr. Ding currently serves on the Editorial Advisory Board of Knowledge-Based Systems, and Editorial Board of Information Fusion and Applied Soft Computing. He serves as an Associate Editor of several prestigious journals, including IEEE Transactions on Fuzzy Systems, Information Sciences, Swarm and Evolutionary Computation, IEEE Access, and Journal of Intelligent & Fuzzy Systems, as well as the leading guest editor in several international journals. He has delivered more than 15 keynote speeches at international conferences and has co-chaired several international conferences and workshops in the area of fuzzy decision-making, data mining, and knowledge engineering.

Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively. He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multi-granulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has published over 80 articles on these topics in international journals. He served on the Editorial Board of the International Journal of Knowledge-Based Organizations and

Artificial Intelligence Research. He has served as the Program Chair or Special Issue Chair of the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and a PC member of many machine learning, data mining, and granular computing conferences.

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 Xi'an 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.