Complex & Intelligent Systems (2022) 8:2105–2129
https://doi.org/10.1007/s40747-021-00636-y

ORIGINAL ARTICLE

Multi-label feature selection based on fuzzy neighborhood rough sets

Jiucheng Xu[1,2] · Kaili Shen[1,2] · Lin Sun[1,2]

Received: 25 July 2021 / Accepted: 23 December 2021 / Published online: 10 January 2022

Abstract

Multi-label feature selection, a crucial preprocessing step for multi-label classification, has been widely applied to data mining, artificial intelligence and other fields. However, most of the existing multi-label feature selection methods for dealing with mixed data have the following problems: (1) These methods rarely consider the importance of features from multiple perspectives, which analyzes features not comprehensive enough. (2) These methods select feature subsets according to the positive region, while ignoring the uncertainty implied by the upper approximation. To address these problems, a multi-label feature selection method based on fuzzy neighborhood rough set is developed in this article. First, the fuzzy neighborhood approximation accuracy and fuzzy decision are defined in the fuzzy neighborhood rough set model, and a new multi-label fuzzy neighborhood conditional entropy is designed. Second, a mixed measure is proposed by combining the fuzzy neighborhood conditional entropy from information view with the approximate accuracy of fuzzy neighborhood from algebra view, to evaluate the importance of features from different views. Finally, a forward multi-label feature selection algorithm is proposed for removing redundant features and decrease the complexity of multi-label classification. The experimental results illustrate the validity and stability of the proposed algorithm in multi-label fuzzy neighborhood decision systems, when compared with related methods on ten multi-label datasets.

Keywords Fuzzy neighborhood rough sets · Multi-label feature selection · Fuzzy neighborhood conditional entropy · Approximation accuracy · Multi-label classification

Introduction

In recent years, multi-label classification occupies a very important position in the fields of artificial intelligence and machine learning, which attracts the attention of more and more scholars and a series of multi-label classification methods are proposed [1–5]. In traditional classification learning, each sample has only one category label, namely single-label learning [6,7]. However, in actual application, most of the samples may belong to multiple category labels at the same time, which named multi-label learning [8–10]. There are a large number of features in multi-label data, but some of which may be irrelevant or redundant information, which will lead to such problems such as high computational cost, overfitting, low classification performance of multi-label learning algorithm and long process of classification learning. Therefore, dimension reduction of multi-label data is the focus of current research. Feature selection is one of the most common dimensionality reduction methods for analyzing high dimensional multi-label data, which aims to eliminate redundant and irrelevant features in classification learning task, and extract useful information [11–13].

With the increasing availability of multi-label data related to multiple labels in an instance, a great quantity of feature selection methods for multi-label learning are developed to reduce dimensions and improve learning performance [14–17]. These methods commonly can be divided into three categories: filter [18–20], wrapper [21,22] and embedded [23] methods, where the filter method is independent of the specific learner, and it has less computation cost and stronger generalization ability. Therefore, our proposed method focuses on the filter strategy. The evaluation criteria commonly based on filter method include information measure [24–32], dependency measure [33–38], distance measure [39–42] and consistency measure [43,44].

✉ Kaili Shen
skl9601@163.com

[1] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China

[2] Engineering Technology Research Center for Computing Intelligence and Data Mining, Xinxiang 453007, China

Rough set theory is a familiar method to deal with uncertain data, which does not need any prior information except data, so it has been widely used in feature selection of data [45]. However, the traditional rough set theory is based on equivalence relation, which is only suitable for discrete data. To solve this problem, some scholars have extended the rough set model. For example, the neighborhood rough sets model (NRS), which is the most common model to deal with numerical data, and the neighborhood relation is used to replace the equivalence relation. Duan et al. [46] defined the lower approximation and dependency of NRS in multi-label learning, and proposed a multi-label feature selection algorithm based on neighborhood rough sets model (MNRS). Unfortunately, NRS cannot deal with the fuzziness of data effectively. So Lin et al. [47] used different fuzzy relations to construct a multi-label fuzzy rough sets model (MFRS), which estimated the similarity between samples under different labels, and directly evaluated the attributes of multi-label data, solved the problem of low separability about fuzzy similarity and defined the dependency function. But FRS is sensitive to noise, these noisy data will affect the calculation of fuzzy lower approximation and limit their practical application [48]. To solve the above problems, the fuzzy neighborhood rough sets model (FNRS) is designed. Wang et al. [49] combined NRS with FRS, proposed a feature selection algorithm based on FNRS via dependency to select feature subset. Chen et al. [48] designed a multi-label attribute reduction method based on variable precision FNRS, which used parameterized fuzzy neighborhood granule to define the fuzzy decision and decision class, and calculated importance of features using dependency measure, but the reduction based on the positive region does not take into account the influence of the uncertain information in the upper approximation on the importance of the attribute. Inspired by these observations, this paper designs a multi-label feature selection method based on FNRS and the approximation accuracy is introduced into our proposed multi-label feature selection method.

In the latest decades, the multi-label feature selection methods are classified into two kinds of views. The first is the algebra view based on approximate accuracy, which considers the effect of some features on the labels with the change of approximation accuracy, while confirms whether these features can be eliminated. For instance, Liang et al. [17] presented the selection of the optimal number of particles in the multi-grain and multi-label decision table, which makes certain positive region reduction more suitable for multi-label datasets. Li et al. [35] designed a robust MFRS by the kernelized information and obtained a lower approximation. The second is the information view based on information entropy, which considers the influence of some features on the decision subset with the information entropy and decides whether these features can be eliminated. For example, Lin et al. [25] designed a multi-label feature selection based on neighborhood mutual information, extended neighborhood information entropy to adapt to multi-label data, and introduced three new measurement methods. Li et al. [29] developed a multi-label feature selection based on information gain, which measured the correlation between features and labels. Xu et al. [24] proposed a fuzzy neighborhood conditional entropy for feature selection. Inspired by these contributions, we design a novel fuzzy neighborhood conditional entropy to judge whether exclude these features on multi-label data. However, these methods cannot provide a more accurate and comprehensive assessment of the importance of features from different perspectives. Therefore, Sun et al. [39] developed a multi-label feature selection which combined neighborhood mutual information with the approximate accuracy in multi-label neighborhood decision systems, and this method of combining two views obtained great the classification performance. Combine the above contributions, this paper proposes a multi-label feature selection method, which combines the fuzzy neighborhood conditional entropy with the approximate accuracy, to evaluate the importance of features from two views. Thus, the major contributions of this article can be briefly described as follows:

– Considering that the similarity of samples is also affected by 0-value label, the average value of decision under different labels is calculated as fuzzy decision. The concepts of fuzzy neighborhood upper approximation, lower approximation and fuzzy neighborhood approximation accuracy are proposed, which improves the integrity of multi-label fuzzy neighborhood decision system.
– This work proposed the definitions of fuzzy neighborhood information entropy, fuzzy neighborhood joint entropy and fuzzy neighborhood conditional entropy for multi-label data, and their related properties and proofs are discussed, by improving the single-label fuzzy neighborhood entropy.
– Combining the approximate accuracy of fuzzy neighborhood under the view of algebra with the fuzzy neighborhood conditional entropy under the view of information theory, a mixed measure method is proposed to evaluate the correlation between feature subset and label set in the multi-label fuzzy neighborhood decision system. Finally, a forward multi-label feature selection algorithm based on fuzzy neighborhood rough sets is designed for multi-label classification.

The remainder of this paper is structured as follows. The next section briefly introduces the related knowledge of NRS, MNRS and FNRS. In the subsequent section, the fuzzy neighborhood rough set model, fuzzy neighborhood conditional entropy and hybrid measure are introduced. The multi-label feature selection algorithm is designed in the next section.

Then the experimental results are provided. Finally, the conclusions of our research are provided in the last section.

Related knowledge

Classical neighborhood rough sets

Suppose there exists a neighborhood decision system which can be simplified as NDS $=< U, A \bigcup D, V, \Delta, \delta >$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty samples set; $A = \{a_1, a_2, \ldots, a_m\}$ is a features set; D is decision class of samples; $V = \bigcup_{a \in A} V_a$, where V_a is the value of feature a; Δ indicates distance function; and $\delta (0 \leq \delta \leq 1)$ is a neighborhood radius. If Δ satisfy the following properties [50] as

(1) $\forall x_1, x_2 \in U, \Delta(x_1, x_2) \geq 0$, where $\Delta(x_1, x_2) = 0$ if and only if $x_1 = x_2$;
(2) $\forall x_1, x_2 \in U, \Delta(x_1, x_2) = \Delta(x_2, x_1)$;
(3) $\forall x_1, x_2, x_3 \in U, \Delta(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$.

Then $\langle U, \Delta \rangle$ is called metric space, in general, the distance in the metric space can be expressed as

$$\Delta(x_i, x_j) = \left(\sum_{a=1}^{m} \left|x_{ia} - x_{ja}\right|^p\right)^{1/\mathrm{p}},$$

when $p = 1$, Δ represents Manhattan distance; when $p = 2$, Δ represents Euclidean distance; when $p \to \infty$, $\Delta(x_i, x_j) = \max_a \left|x_{ia}, x_{ja}\right|$.

Suppose the nonempty metric space $< U, \Delta >$, for $\forall B \subseteq A$, $\delta_B(x) = \{y \,|x, y \in U, \Delta(x, y) \leq \delta,\ \delta \geq 0\}$ [46]. $\Delta(x, y)$ is a function to measure the distance between x and y, $\delta_B(x)$ can also be called the neighborhood granularity of x under B.

Multi-label neighborhood rough sets

Suppose there exists a multi-label neighborhood decision system which can be abbreviated to MNDS $=< U, A \bigcup D, \delta >$, for $\forall B \subseteq A$, $D = \{d_1, d_2, \ldots, d_t\}$, $D_i = \{d_j | d_j(x_i) = 1, d_j \in D\}$ represents the related label set of x_i, and $D^{\mathrm{j}} = \{x_i | d_j(x_i) = 1, x_i \in U\}$ denotes a set of samples with the label d_j. Then the upper approximation and lower approximation of the neighborhood rough sets of D with respect to B are defined [46], respectively, as

$$\overline{N_B}D = \left\{x_i \left|\forall d_j \in D_i, \delta_B(x_i) \bigcap D^j \neq \varnothing, x_i \in U\right.\right\}, \quad (1)$$

$$\underline{N_B}D = \{x_i \left|\forall d_j \in D_i, \delta_B\ (x_i) \subseteq D^j, x_i \in U\right.\}. \quad (2)$$

Then, for $\forall B \subseteq A$, the neighborhood entropy of $x_i \in U$ is expressed [25] as

$$NE(B) = -\log \frac{|\delta_B(x_i)|}{|U|}. \quad (3)$$

Fuzzy neighborhood rough sets

Suppose there exists a fuzzy neighborhood decision system which can be short for FNDS $=< U, A \bigcup D, \delta >$, where $U = \{x_1, x_2, \ldots, x_n\}$ is the nonempty set of samples, and A is the set of features for $\forall B \subseteq A$. The fuzzy binary relation R_B is derived from B [49]. For $\forall x, y \in U$, $R_B(x, y)$ is called fuzzy similarity relation between samples x and y under features set B when it satisfies the following conditions:

(1) Reflexivity: $R_B(x, x) = 1, \forall x \in U$;
(2) Symmetry: $R_B(x, y) = R_B(y, x), \forall x, y \in U$.

Then R_B is also known as the fuzzy similarity relation.
Suppose there exists FNDS $=< U, A \bigcup D, \delta >$ with for $\forall B \subseteq A$, $\forall a \in B$, $\forall x, y \in U$, the fuzzy similarity matrix is $[x]_a(y) = R_{\mathrm{a}}(x, y)$, R_a is a fuzzy similarity relation for $\forall a \in B$, then we can express $R_B = \bigcap_{a \in B} R_a$. Then the fuzzy similarity matrix of x with respect to B over U is defined [24] as

$$[x]_B(y) = \min_{a \in B} ([x]_a(y)), y \in U.$$

Given FNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, $U/D = \{D_1, D_2, \ldots D_{\mathrm{r}}\}$, for $\forall x, y \in U$, the parameterized fuzzy neighborhood information granule is constructed as follows:

$$FN_B(x) = [x]_B^\delta\ (y) = \begin{cases} R_B(x, y), R_B(x, y) \geq \delta \\ 0, \quad R_B(x, y) < \delta \end{cases}, \quad (4)$$

where δ is called the fuzzy neighborhood radius and satisfies $0 \leq \delta \leq 1$. The fuzzy neighborhood of $\forall x \in U$ can be determined by fuzzy similarity relation R_B and neighborhood radius δ.

Let FNDS $=< U, A \bigcup D, \delta >$ be a fuzzy neighborhood decision system, $U/D = \{D_1, D_2, \cdots D_{\mathrm{r}}\}$, for $\forall B \subseteq A$, the upper and lower approximations of D with respect to B are expressed, respectively, as

$$\overline{FN_B^\delta}(D_j) = \left\{x \in U \left|FN_B\ (x) \bigcap D_j \neq \varnothing\right.\right\}, \quad (5)$$

$$\underline{FN_B^\delta}(D_j) = \{x \in U \,|FN_B(x) \subseteq D_j\}. \quad (6)$$

For $\forall B \subseteq C$, the fuzzy neighborhood approximation accuracy of D with respect to B is described as

$$AP_B^{\delta} = \frac{\left|\underline{FN_B^{\delta}}(D_j)\right|}{\left|\overline{FN_B^{\delta}}(D_j)\right|}. \tag{7}$$

Proposed method

In this section, we improve the multi-label fuzzy neighborhood rough set model based on the relevant basic knowledge introduced in the previous section. First, the parameterized fuzzy similarity relation is used to calculate the fuzzy neighborhood granule. Because a sample in multi-label data may belong to multiple labels at the same time, the multi-label fuzzy decision is obtained by averaging values in multiple labels, which is different from the single-label fuzzy decision. Secondly, the fuzzy neighborhood approximation accuracy is introduced to consider the uncertain information of upper approximation. Then the fuzzy neighborhood conditional entropy for multi-label data is proposed. Finally, the fuzzy neighborhood approximation accuracy and fuzzy neighborhood conditional entropy are combined to form a mixed measure, and the relevant proof process is given.

Multi-label fuzzy neighborhood approximation accuracy and fuzzy decision

Definition 1 A multi-label fuzzy neighborhood decision system can be denoted as MFNDS $=< U, A \bigcup D, T, \delta >$. $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of samples; $A = \{a_1, a_2, \ldots a_m\}$ indicates a set of features; $D = \{d_1, d_2, \ldots, d_t\}$ represents a set of labels; $T = \{(x_i, A(x_i), D(x_i))|x_i \in U\}$, $\forall x_i \in U$, it allows $A(x_i) = (a_1(x_i), a_2(x_i), \ldots, a_m(x_i))$, where $a_m(x_i)$ is the value of the sample x_i in the feature a_m, $D(x_i) = (d_1(x_i), d_2(x_i), \ldots, d_t(x_i))$, where $d_j(x_i) = \{0, 1\}$, $d_j(x_i)$ indicates whether the sample x_i contains label d_j, if x_i contains label d_j, then $d_j(x_i) = 1$; otherwise, $d_j(x_i) = 0$.

Definition 2 Given MFNDS $=< U, A \bigcup D, T, \delta >$, let $\{D_0^1, D_1^1, D_0^2, D_1^2, \ldots, D_1^t\}$ denote a label determined coverage of U, then the parameterized fuzzy decision is constructed as follows:

$$\tilde{D}_p^{\mathrm{j}}(x) = \frac{\left|[x]_A(y) \bigcap D_p^j\right|}{|[x]_A(y)|}, \tag{8}$$

where D_p^j represents a sample set which is p in the column of the label d_j, $j = 1, 2, \ldots, t$, $p = 0, 1$.

$$\tilde{D}_p^j = \{\tilde{D}_p^j(x_1), \tilde{D}_p^j(x_2), \ldots, \tilde{D}_p^j(x_n)\}, \tag{9}$$

where $\tilde{D}_{\mathrm{p}}^j(x_i)$ is the fuzzy membership degree of x_i with respect to D_{p}^j; $\tilde{D}_{\mathrm{p}}^j$ is the fuzzy set of the equivalence decision class of the samples.

$$\tilde{D}_p(x_i) = \frac{1}{t}\sum_{j=1}^{t} \tilde{D}_p^j(x_i), \tag{10}$$

$$\tilde{D}_p = \{\tilde{D}_p(x_1), \tilde{D}_p(x_2), \cdots, \tilde{D}_p(x_n)\}, \tag{11}$$

where $\tilde{D}_{\mathrm{p}}(x_i)$ is the fuzzy set of the sample x_i which belongs label p.

$$\tilde{D} = \{\tilde{D}_0^T, \tilde{D}_1^T\}, \tag{12}$$

where $\{\tilde{D}_0, \tilde{D}_1\}$ is the fuzzy decision of the samples induced by D.

Definition 3 [49] Let F' and R' are the two fuzzy sets, the inclusion degree between F' and R' can be defined as

$$P(F', R') = \frac{\left|F' \bigcap R'\right|}{|U|}, \tag{13}$$

where $P(F', R')$ represents the inclusion degree of fuzzy set F' in fuzzy set R', $\left|F' \bigcap R'\right|$ represents the number of samples whose membership degree of fuzzy set F' is not greater than that of fuzzy set R'.

Example 1 Given a set $U = \{x_1, x_2, \ldots, x_6\}$, F' and R' are two fuzzy sets defined on U, which represent the membership degree of samples separately, as follows:

$$F' = \left\{\frac{0.7}{x_1}, \frac{0.9}{x_2}, \frac{0.4}{x_3}, \frac{0.3}{x_4}, \frac{0.6}{x_5}, \frac{0.5}{x_6}\right\},$$
$$R' = \left\{\frac{0.5}{x_1}, \frac{0.9}{x_2}, \frac{0.7}{x_3}, \frac{0.6}{x_4}, \frac{0.3}{x_5}, \frac{0.4}{x_6}\right\}.$$

So, we can get

$$\left|F' \bigcap R'\right| = |x_2, x_3, x_4| = 3,$$
$$\left|R' \bigcap F'\right| = |x_1, x_2, x_5, x_6| = 4.$$

Definition 4 Given MFNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, $D = \{d_1, d_2, \ldots, d_t\}$ represents a set of labels; δ is called the fuzzy neighborhood radius and satisfies $0 \leq \delta \leq 1$. For $\forall x, y \in U$, the parameterized fuzzy neighborhood information granule is constructed as follows:

$$\delta_B(x) = [x]_B^{\delta}(y) = \begin{cases} R_B(x, y), R_B(x, y) \geq 1 - \delta \\ 0, \quad R_B(x, y) < 1 - \delta, \end{cases}, \tag{14}$$

where R_B is the fuzzy similarity relation induced by B on U, when $B_1 \subseteq B_2$, $R_{B_2} \subseteq R_{B_1}$; when $\delta_1 \leq \delta_2$, for $\forall x \in U$, $[x]_B^{\delta_1} \subseteq [x]_B^{\delta_2}$.

Definition 5 Given $MFNDS = < U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, δ is called the fuzzy neighborhood radius; $\left\{\tilde{D}_0, \tilde{D}_1\right\}$ is the fuzzy decision of samples induced by D. The upper and lower approximations of the fuzzy neighborhood of D is relative to B are defined, separately, as

$$\overline{R_B^\delta}(D) = \left\{\overline{R_B^\delta}(\tilde{D}_1), \overline{R_B^\delta}(\tilde{D}_2), \ldots \overline{R_B^\delta}(\tilde{D}_p)\right\}, \tag{15}$$

$$\underline{R_B^\delta}(D) = \left\{\underline{R_B^\delta}(\tilde{D}_1), \underline{R_B^\delta}(\tilde{D}_2), \ldots \underline{R_B^\delta}(\tilde{D}_p)\right\}, \tag{16}$$

where

$$\overline{R_B^\delta}(\tilde{D}_p) = \left\{x \in U \middle| P(\delta_B(x), \tilde{D}_p) > \beta\right\}, 0 \leq \beta < 0.5, \tag{17}$$

$$\underline{R_B^\delta}(\tilde{D}_p) = \left\{x \in U \middle| P(\delta_B(x), \tilde{D}_p) \geq \alpha\right\}, 0.5 \leq \alpha \leq 1. \tag{18}$$

Definition 6 Given MFNDS $= < U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, δ is called the fuzzy neighborhood radius; $\left\{\tilde{D}_0, \tilde{D}_1\right\}$ is the fuzzy decision of samples induced by D; R_B is the fuzzy similarity relation induced by B on U. The fuzzy neighborhood approximation accuracy is defined as

$$\alpha_B^\delta(D) = \frac{\sum_{p=1}^r \left|\underline{R_B^\delta}(\tilde{D}_p)\right|}{\sum_{p=1}^r \left|\overline{R_B^\delta}(\tilde{D}_p)\right|}, \tag{19}$$

where $|.|$ represents the cardinality of the set. $\left|\underline{R_B^\delta}(\tilde{D}_p)\right| \leq \left|\overline{R_B^\delta}(\tilde{D}_p)\right|$, so $0 \leq \alpha_B^\delta(D) \leq 1$.

Property 1 Given MFNDS $= < U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, δ_1 and δ_2 are two fuzzy neighborhood radii, if $\delta_1 \leq \delta_2$, then $\alpha_B^{\delta_2}(D) \leq \alpha_B^{\delta_1}(D)$.

Proof For $\forall x \in U$, according to Definition 4, the fuzzy neighborhood information granule satisfies the relation is obtained $[x]_B^{\delta_1} \subseteq [x]_B^{\delta_2}$, then $\underline{R_B^{\delta_2}}(\tilde{D}_p) \subseteq \underline{R_B^{\delta_1}}(\tilde{D}_p)$, $\overline{R_B^{\delta_1}}(\tilde{D}_p) \subseteq \overline{R_B^{\delta_2}}(\tilde{D}_p)$, so there is $\alpha_B^{\delta_2}(D) \leq \alpha_B^{\delta_1}(D)$. □

Property 2 Given MFNDS $= < U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, δ is a fuzzy neighborhood radius, if $B_1 \subseteq B_2$, we can get the property: $\alpha_{B_1}^\delta(D) \leq \alpha_{B_2}^\delta(D)$.

Proof Since $B_1 \subseteq B_2$, according to the fuzzy neighborhood granule satisfies the relation is obtained $[x]_{B_2}^\delta \subseteq [x]_{B_1}^\delta$, then according to Definitions 5 and 6, we have $\underline{R_{B_1}^\delta}(\tilde{D}_p) \subseteq \underline{R_{B_2}^\delta}(\tilde{D}_p)$, $\overline{R_{B_2}^\delta}(\tilde{D}_p) \subseteq \overline{R_{B_1}^\delta}(\tilde{D}_p)$. Then, $\alpha_{B_1}^\delta(D) \leq \alpha_{B_2}^\delta(D)$ holds. □

Example 2 Given a multi-label decision table MDT=$< U, A \bigcup D >$ to display in Table 1, $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ represents a sample set, $A = \{a_1, a_2, a_3\}$ means a feature set, $D = \{d_1, d_2, d_3\}$ indicates a label set, R_A is based on the fuzzy similarity relation induced by A, let the value of fuzzy neighborhood radius be 0.

Table 1 A multi-label decision table

U	a_1	a_2	a_3	d_1	d_2	d_3
x_1	6.2	0.6	2	0	1	0
x_2	11.2	1.3	1	1	0	1
x_3	4.3	1.3	0.6	0	0	1
x_4	5.8	0.3	2	1	1	0
x_5	0.7	1.3	1	1	1	0
x_6	2.9	0.4	1	0	1	1

The data in Table 1 were normalized according to literature [24], so that the numerical value was within the range of [0,1]. The fuzzy similarity relationship R_{a_k} between the samples x_i and x_j relative to the attribute a_k is calculated by

$$R_{a_k} = 1 - \left|x_{ik} - x_{jk}\right|, \tag{20}$$

where $a_k \in A$, $k = 1, 2, 3$, $x_i, x_j \in U$, $i = 1, 2, 3, 4, 5, 6$, $j = 1, 2, 3, 4, 5, 6$. So, we can obtain the fuzzy similarity matrix $[x]_{a_k}(y)$ about the attribute a_k, and $[x]_A(y) = \min_{a_k \in A}\left([x]_{a_k}(y)\right)$. Because the fuzzy similarity relation R_{a_k} satisfies the reflexivity, $R_{a_k} = 1$ when $i = j$, then we can get

$$[x]_A(y) = \begin{bmatrix} 1 & 0.2857 & 0 & 0.7 & 0.2857 & 0.2857 \\ 0.2857 & 1 & 0.3429 & 0 & 0 & 0.1 \\ 0 & 0.3429 & 1 & 0 & 0.6571 & 0.1 \\ 0.7 & 0 & 0 & 1 & 0 & 0.2857 \\ 0.2857 & 0 & 0.6571 & 0 & 1 & 0.1 \\ 0.2857 & 0.1 & 0.1 & 0.2857 & 0.1 & 1 \end{bmatrix}.$$

The fuzzy decision under the labels d_1, d_2, d_3 are calculated as follows:

$$D_1^1 = \{x_2, x_4, x_5\}, \qquad D_0^1 = \{x_1, x_3, x_6\};$$
$$D_1^2 = \{x_1, x_4, x_5, x_6\}, \qquad D_0^2 = \{x_2, x_3\};$$
$$D_1^3 = \{x_2, x_3, x_6\}, \qquad D_0^3 = \{x_1, x_4, x_5\};$$

where D_r^j represents the sample set of the value is p under the label d_j, where $j = 1, 2, 3$, $p = 0, 1$. According to Definition 2, we can obtain

$$\begin{aligned}
\tilde{D}_0^1 &= \{\tilde{D}_0^1(x_1), \tilde{D}_0^1(x_2), \tilde{D}_0^1(x_3), \tilde{D}_0^1(x_4), \tilde{D}_0^1(x_5), \tilde{D}_0^1(x_6)\} \\
&= \{0.5028, 0.4215, 0.5238, 0.4964, 0.5105, 0.7405\}, \\
\tilde{D}_1^1 &= \{\tilde{D}_1^1(x_1), \tilde{D}_1^1(x_2), \tilde{D}_1^1(x_3), \tilde{D}_1^1(x_4), \tilde{D}_1^1(x_5), \tilde{D}_1^1(x_6)\} \\
&= \{0.4972, 0.5785, 0.4762, 0.5036, 0.4895, 0.2595\}; \\
\tilde{D}_0^2 &= \{\tilde{D}_0^2(x_1), \tilde{D}_0^2(x_2), \tilde{D}_0^2(x_3), \tilde{D}_0^2(x_4), \tilde{D}_0^2(x_5), \tilde{D}_0^2(x_6)\} \\
&= \{0.1117, 0.7769, 0.6395, 0, 0.3217, 0.1069\}, \\
\tilde{D}_1^2 &= \{\tilde{D}_1^2(x_1), \tilde{D}_1^2(x_2), \tilde{D}_1^2(x_3), \tilde{D}_1^2(x_4), \tilde{D}_1^2(x_5), \tilde{D}_1^2(x_6)\} \\
&= \{0.8883, 0.2231, 0.3605, 1, 0.6783, 0.8931\}; \\
\tilde{D}_0^3 &= \{\tilde{D}_0^3(x_1), \tilde{D}_0^3(x_2), \tilde{D}_0^3(x_3), \tilde{D}_0^3(x_4), \tilde{D}_0^3(x_5), \tilde{D}_0^3(x_6)\} \\
&= \{0.7765, 0.1653, 0.3129, 0.8561, 0.6294, 0.3588\}, \\
\tilde{D}_1^3 &= \{\tilde{D}_1^3(x_1), \tilde{D}_1^3(x_2), \tilde{D}_1^3(x_3), \tilde{D}_1^3(x_4), \tilde{D}_1^3(x_5), \tilde{D}_1^3(x_6)\} \\
&= \{0.2235, 0.8347, 0.6871, 0.1439, 0.3706, 0.6412\}.
\end{aligned}$$

Then we can get

$$\begin{aligned}
\tilde{D}_0 &= \frac{1}{3}\sum_{j=1}^{3} \tilde{D}_0^j \\
&= \{0.4637, 0.4546, 0.4921, 0.4508, 0.4872, 0.4021\}; \\
\tilde{D}_1 &= \frac{1}{3}\sum_{j=1}^{3} \tilde{D}_1^j \\
&= \{0.5363, 0.5454, 0.5079, 0.5492, 0.5128, 0.5979\}.
\end{aligned}$$

From the above, we can derive that $\tilde{D}_0(x) + \tilde{D}_1(x) = 1$, so the eventual fuzzy decision of entire label space is

$$\tilde{D} = \{\tilde{D}_0^T, \tilde{D}_1^T\} = \begin{bmatrix} 0.4637 & 0.5363 \\ 0.4546 & 0.5454 \\ 0.4921 & 0.5079 \\ 0.4508 & 0.5492 \\ 0.4872 & 0.5128 \\ 0.4021 & 0.5979 \end{bmatrix}.$$

Multi-label fuzzy neighborhood conditional entropy

Definition 7 Suppose there exists $\mathrm{MFNDS} =< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, δ is the neighborhood radius, then fuzzy neighborhood entropy of B is defined as

$$E_{fn}(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{|\delta_B(x_i)|}{|U|}, \tag{21}$$

where $|\delta_B(x_i)|$ represents the number of nonzero values in the fuzzy neighborhood particle of an object x_i, then $\frac{|\delta_B(x_i)|}{|U|}$ represents the probability of the number of nonzero values in fuzzy neighborhood granule $|\delta_B(x_i)|$ in U.

Definition 8 Suppose there exists $\mathrm{MFNDS} =< U, A \bigcup D, \delta >$ with $\forall B_1, B_2 \subseteq A$, $\delta_{B_1}(x)$ and $\delta_{B_2}(x)$ are fuzzy neighborhood granules, then the fuzzy neighborhood joint entropy of B_1 and B_2 is defined as

$$E_{fn}(B_1, B_2) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{B_1 \bigcup B_2}(x_i)\right|}{|U|}. \tag{22}$$

Definition 9 Suppose there exists $\mathrm{MFNDS} =< U, A \bigcup D, \delta >$ with $\forall B_1, B_2 \subseteq A$, $\delta_{B_1}(x)$ and $\delta_{B_2}(x)$ are fuzzy neighborhood granules, then the fuzzy neighborhood conditional entropy of B_1 and B_2 is defined as

$$E_{fn}(B_1 | B_2) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{B_1 \bigcup B_2}(x_i)\right|}{\left|\delta_{B_2}(x_i)\right|}. \tag{23}$$

Property 3 Suppose there exists $\mathrm{MFNDS} =< U, A \bigcup D, \delta >$ with $\forall B_1, B_2 \subseteq A$, $\delta_{B_1}(x)$ and $\delta_{B_2}(x)$ are fuzzy neighborhood granules, for $\forall B_1, B_2 \subseteq A$. The property is as follows:

$$E_{fn}(B_1 | B_2) = E_{fn}(B_1, B_2) - E_{fn}(B_2).$$

Proof According to Definitions 6 and 7, it can be proved

$$\begin{aligned}
&E_{fn}(B_1, B_2) - E_{fn}(B_2) \\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{B_1 \bigcup B_2}(x_i)\right|}{|U|} + \frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{B_2}(x_i)\right|}{|U|} \\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \left(\frac{\left|\delta_{B_1 \bigcup B_2}(x_i)\right|}{|U|} \cdot \frac{|U|}{\left|\delta_{B_2}(x_i)\right|} \right) \\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{B_1 \bigcup B_2}(x_i)\right|}{\left|\delta_{B_2}(x_i)\right|}.
\end{aligned}$$

Then, from Definition 8, it follows that $E_{fn}(B_1 | B_2) = E_{fn}(B_1, B_2) - E_{fn}(B_2)$. □

Definition 10 Suppose there exists $\mathrm{MFNDS} =< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, $\delta_B(x)$ is the fuzzy neighborhood granule, $\tilde{D} = \{\tilde{D}_0, \tilde{D}_1\}$ is a fuzzy decision, then the conditional entropy of decision attribute set D on feature subset B is defined as

$$\begin{aligned}
&E_{fn}(D | B) \\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|}\sum_{p=0}^{1} \frac{\left|\delta_{B \bigcup \tilde{D}_p}(x_i)\right|}{|\delta_B(x_i)|} \\
&\quad = -\frac{1}{|U|}\sum_{i=1}^{|U|}\sum_{p=0}^{1} \log_2 \frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|},
\end{aligned} \tag{24}$$

where $|\delta_B(x_i)|$ represents the number of nonzero values in the fuzzy neighborhood particle of an object x_i, then $\left|\delta_B(x_i) \bigcap \tilde{D}_p\right|$ represents the number of nonzero values of samples whose membership degree of $\delta_B(x_i)$ is not greater than $\tilde{D}_p$.

The feature selection only from the algebraic or information viewpoint is limited. For the algebraic viewpoint, feature selection under the definition of information theory may also exist redundancy features, for information theory, feature selection under definitions of algebraic viewpoint, conditional entropy may have changed. So we combine the approximate precision from the algebraic viewpoint with the measurement method of conditional entropy from information theory to calculate the importance degree of candidate features.

Definition 11 Given MFNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, $\delta_B(x)$ is the fuzzy neighborhood granule, $\tilde{D} = \{\tilde{D}_0, \tilde{D}_1\}$ is a fuzzy decision, then the mixed measure based on the approximate accuracy of the fuzzy neighborhood and the conditional entropy of the fuzzy neighborhood is defined as

$$\begin{aligned} & EM_{fn}(D|B) \\ & = \alpha_B^\delta(D) \cdot E_{fn}(D|B) \\ & = -\frac{\alpha_B^\delta(D)}{|U|} \sum_{\mathrm{i}=1}^{|U|} \sum_{p=0}^{1} \log_2 \frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|}. \end{aligned} \tag{25}$$

Property 4 Given MFNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, $\delta_B(x)$ is the fuzzy neighborhood granule, $\tilde{D} = \{\tilde{D}_0, \tilde{D}_1\}$ is a fuzzy decision, then $EM_{fn}(D|B) \geq 0$.

Proof Assume $EM_{fn}(D|B) < 0$, then $\log_2 \frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|} > 0$, then $\frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|} > 1$; therefore, $|\delta_B(x_i)| < \left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|$, but this is obviously not established. So, $\frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|} \leq 1$, that is, $\log_2 \frac{\left|\delta_B(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_B(x_i)|} \leq 0$. Therefore, $EM_{fn}(D|\ B) \geq 0$. □

Property 5 Let MFNDS $=< U, A \bigcup D, \delta >$ be a multi-label fuzzy neighborhood decision system, for $\forall B_1, B_2 \subseteq A$, if $B_1 \subseteq B_2$, according to Property 2, $\delta_{B_2}(x) \subseteq \delta_{B_1}(x)$, then $EM_{fn}(D|B_1) \geq EM_{fn}(D|B_2)$, if and only if $\delta_{B_1}(x) = \delta_{B_2}(x)$, then the equal sign is true.

Proof According to Eq. (26), $EM_{fn}(D|B_1) \geq EM_{fn}(D|B_2)$ holds.

$$\begin{aligned} & EM_{fn}(D|B_1) - EM_{fn}(D|B_2) \\ & = \left(-\frac{\alpha_{B_1}^\delta(D)}{|U|} \sum_{\mathrm{i}=1}^{|U|} \sum_{p=0}^{1} \log_2 \frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_{B_1}(x_i)|}\right) \\ & \quad - \left(-\frac{\alpha_{B_2}^\delta(D)}{|U|} \sum_{\mathrm{i}=1}^{|U|} \sum_{p=0}^{1} \log_2 \frac{\left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_{B_2}(x_i)|}\right) \\ & = -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{p=0}^{1} \left(\alpha_{B_1}^\delta(D) \log_2 \left(\frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_{B_1}(x_i)|}\right)\right. \\ & \quad \left. -\alpha_{B_2}^\delta(D) \log_2 \left(\frac{\left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_{B_2}(x_i)|}\right)\right) \\ & = -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{p=0}^{1} \left(\log_2 \left(\frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|}{|\delta_{B_1}(x_i)|}\right)^{\alpha_{B_1}^\delta(D)}\right. \\ & \quad \left. +\log_2 \left(\frac{|\delta_{B_2}(x_i)|}{\left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|}\right)^{\alpha_{B_2}^\delta(D)}\right) \\ & = -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{p=0}^{1} \log_2 \\ & \quad \times \left(\frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_1}^\delta(D)}}{|\delta_{B_1}(x_i)|^{\alpha_{B_1}^\delta(D)}} \cdot \frac{|\delta_{B_2}(x_i)|^{\alpha_{B_2}^\delta(D)}}{\left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_2}^\delta(D)}}\right) \\ & = -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{p=0}^{1} \log_2 \frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_1}^\delta(D)} \cdot |\delta_{B_2}(x_i)|^{\alpha_{B_2}^\delta(D)}}{|\delta_{B_1}(x_i)|^{\alpha_{B_1}^\delta(D)} \cdot \left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_2}^\delta(D)}} \\ & \geq -\frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{p=0}^{1} \log_2 \frac{\left|\delta_{B_1}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_1}^\delta(D)} \cdot |\delta_{B_2}(x_i)|^{\alpha_{B_2}^\delta(D)}}{|\delta_{B_2}(x_i)|^{\alpha_{B_2}^\delta(D)} \cdot \left|\delta_{B_2}(x_i) \bigcap \tilde{D}_p(x_i)\right|^{\alpha_{B_2}^\delta(D)}} \\ & \geq 0. \end{aligned} \tag{26}$$

□

Property 6 Given MFNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A, \forall a \in B$, if $EM_{fn}(D|B - \{a\}) = EM_{fn}(D|B)$, then the feature a is unnecessary.

Proof Assume there exists $a \in B$ satisfies $EM_{fn}(D|B - \{a\}) = EM_{fn}(D|B)$ and the feature a is necessary. We can see from previous knowledge that $\delta_{B-\{a\}}(x) \neq \delta_B(x)$, and $B - \{a\} \subseteq B$, according to Property 5, we know that $EM_{fn}(D|B - \{a\}) > EM_{fn}(D|B)$, this contradicts the hypothesis. So, for $\forall a \in B$, if $EM_{fn}(D|\ B - \{a\}) = EM_{fn}(D|B)$, then the feature a is unnecessary. □

Definition 12 Given MFNDS $=< U, A \bigcup D, \delta >$with $\forall B \subseteq A$,we call B a reduction of A in the fuzzy neighborhood decision information system, relative to decision class D when it satisfies that

(1) $EM_{fn}(D|B) = EM_{fn}(D|A)$;
(2) $\forall a \in B, EM_{fn}(D|B - \{a\}) > EM_{fn}(D|B)$.

Definition 13 Given MFNDS $=< U, A \bigcup D, \delta >$ with $\forall B \subseteq A$, the importance of feature for $a \in B$ relative to D is expressed as

$$SIG(a, B, D) = EM_{fn}(D\,|B - \{a\}) - EM_{fn}(D\,|B\,). \tag{27}$$

To get a reduced subset, two preconditions from the Definition 12 must be met. However, there are many redundant and unrelated features in the multi-label datasets, and searching for the minimum reduced subset is an NP-complete problem. Therefore, we set a threshold value λ to control subset selection before selecting the final feature subset. If the difference of the mixing measure between the current feature subset and the original feature subset is less than λ, then a relatively approximate reduced subset *Red* is selected, which shall meet the following requirements:

$$EM_{fn}(D\,|Red\,) - EM_{fn}(D\,|A\,) \le \lambda. \tag{28}$$

Then the importance of feature for $R \in A - Red$ relative to D is expressed as

$$\begin{aligned} &SIG\,(R, Red, D) \\ &\quad = EM_{fn}(D\,|Red\,) - EM_{fn}(D\,|Red \bigcup \{R\}). \end{aligned} \tag{29}$$

Remark 1 Sun et al. [51] considered that the upper and lower approximations of rough set belong to the viewpoint of algebraic theory, and information entropy and its extension belong to the viewpoint of information theory. Then Definition 6 shows the fuzzy neighborhood approximate accuracy $\alpha_B^\delta(D)$ from the algebraic point of view, and Definition 10 shows the conditional entropy $E_{fn}(D\,|B\,)$ of the feature subset B of the fuzzy information decision set $\tilde{D}$ from the information theory. Therefore, Definition 11 measures the uncertainty of the multi-label fuzzy neighborhood decision systems from both the algebraic view and information view.

Multi-label feature selection algorithm based on fuzzy neighborhood rough sets

According to the relevant definitions in the third section, this paper constructs a multi-label feature selection algorithm based on fuzzy neighborhood rough sets. To clearly understand our proposed algorithm, the process of feature selection for multi-label classification is described by the framework is shown in Fig. 1.

In Algorithm 1, a multi-label feature selection algorithm (MFSFN) is proposed based on fuzzy neighborhood rough sets, assume that the multi-label fuzzy neighborhood decision system contains n is the size of samples, m is the number of features and t is the number of labels which have $|D|$ decision classes. The time complexity on the calculation of the fuzzy similarity relation is $\mathrm{O}(\frac{1}{2}n^2m)$, which is the basis for the calculation of the fuzzy decision is $O(tn\,|D|)$ in Steps 4–6, the time complexity on calculation of the approximation accuracy is $\mathrm{O}(nm\,|D|)$ in Step 7 and the time complexity on calculation of the fuzzy neighborhood conditional entropy is $\mathrm{O}(nm\,|D|)$ in Steps 8–9. In Steps 11–24, assume the size of selected subset is r, its time complexity is $\mathrm{O}(mr\,|D|)$. Therefore, the worst time complexity of MFSFN is approximately $\mathrm{O}(\frac{1}{2}n^2m{+}tn\,|D| + nm\,|D| + mr\,|D|)$. Since the decision classes in the proposed algorithm is constant and that is $|D| = 2$, the total computational time complexity of Algorithm 1 is $\mathrm{O}(\frac{1}{2}n^2m)$.

Algorithm 1 MFSFN

```
Require: $MFNDS < U, A \bigcup D >$:a multi-label fuzzy neighborhood decision system; $\delta$: a fuzzy neighborhood radius; $\lambda$: a parameter to control the selection of feature subset.
Ensure: $Red$: the reduced subset.
1: Initialize $red = \emptyset$, $B = A - red$
2: for $\forall a \in B$ do
3:   Compute the fuzzy similarity relation $R_a$
4:   for $\forall d_j \in D$ do
5:     Compute the fuzzy decision $\tilde{D}$
6:   end for
7:   Compute the accuracy of fuzzy neighborhood approximation $\alpha_B^\delta(D)$
8:   Compute the fuzzy neighborhood conditional entropy $E_{fn}(D\,|B\,)$
9:   Compute the mixed metric$EM_{fn}(D\,|B\,)$
10: end for
11: while $EM_{fn}(D\,|Red\,) \neq EM_{fn}(D\,|A\,)$ do
12:   for $\forall a \in B$ do
13:     Compute $SIG(a, B, D)$
14:     Find $SIG(a, B, D)$=max$\{a\,|SIG(a, B, D)\}$
15:     if $SIG(a, B, D) > 0$ then
16:       Let $Red = Red \bigcup \{a\}$, and compute $EM_{fn}(D\,|Red\,)$
17:       if $EM_{fn}(D\,|Red\,) - EM_{fn}(D\,|A\,) \le \lambda$ then
18:         Return $Red$
19:       end if
20:     end if
21:     $B = B - Red$
22:   end for
23: end while
24: return $Red$
```

Experimental results and analysis

Experimental preparation

The main goal of feature selection is to select fewer feature subsets and achieve higher classification performance. To prove the validity and classification performance of our method, we select ten multi-label datasets of four different fields from http://mulan.sourceforge.net/datasets.html and http://www.uco.es/kdis/mllresources/. The Flags dataset contains details of some countries and their flags; Cal500 is a music dataset, composed of 502 songs; Emotions is about the music fragments that can cause emotions; Scene stores

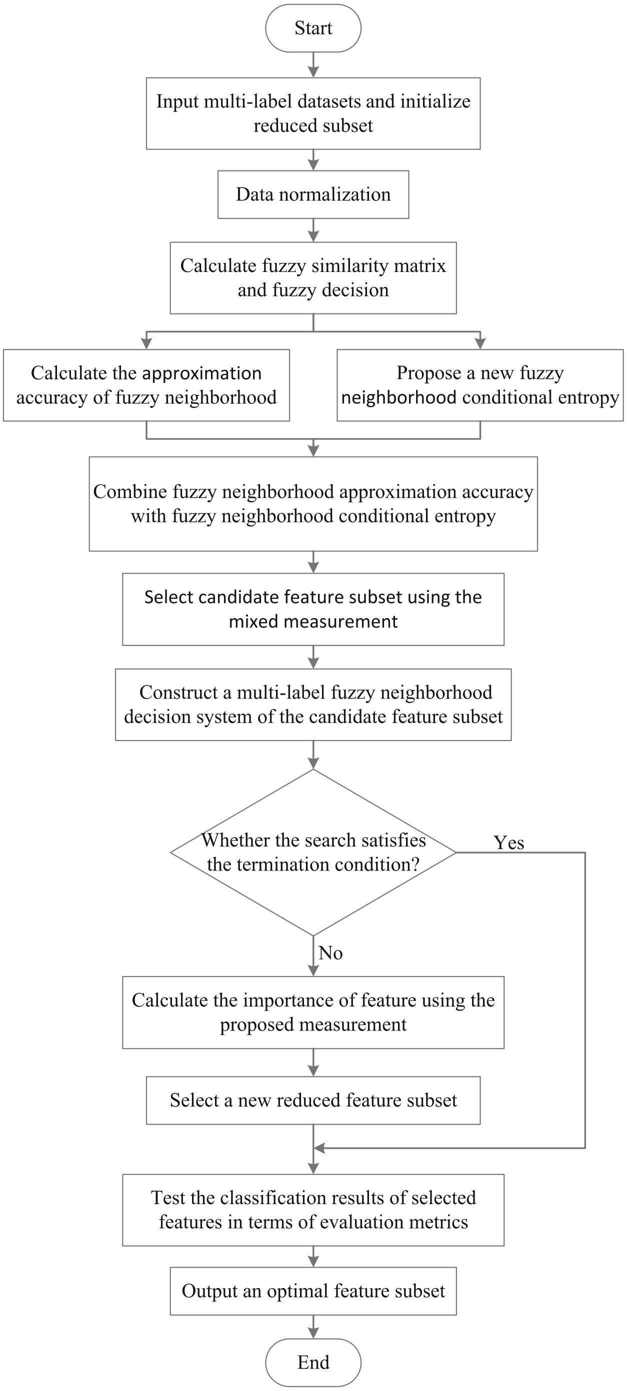

Start
Input multi-label datasets and initialize reduced subset
Data normalization
Calculate fuzzy similarity matrix and fuzzy decision
Calculate the approximation accuracy of fuzzy neighborhood
Propose a new fuzzy neighborhood conditional entropy
Combine fuzzy neighborhood approximation accuracy with fuzzy neighborhood conditional entropy
Select candidate feature subset using the mixed measurement
Construct a multi-label fuzzy neighborhood decision system of the candidate feature subset
Whether the search satisfies the termination condition?
Yes
No
Calculate the importance of feature using the proposed measurement
Select a new reduced feature subset
Test the classification results of selected features in terms of evaluation metrics
Output an optimal feature subset
End

Fig. 1 The framework of proposed multi-label feature selection algorithm

pattern information for a series of scenes; Yeast contains the biological information about gene microarray data and phylogenetic spectrum; the BBC and Guardian datasets include 654 news articles covering 416 distinct news stories; the Gnegative, Plant and Virus datasets are used to predict the subcellular locations of proteins according to their sequences, where Gnegative stores 1392 sequences for Gram-negative bacterial species, Plant contains 978 sequences for plant species, Virus contains 207 sequences for virus species. The basic information description of these datasets, including the size of samples set, the dimensionality of attributes set, the cardinality of the labels set, the domains of ten multi-label datasets, which are demonstrated in Table 2, where LC $(D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{t}\left[d_j(x_i) = +1\right]$ is cardinality of the labels; LD$(D) = \frac{1}{nt}\sum_{i=1}^{n}\sum_{j=1}^{t}\left[d_i(x_i) = +1\right]$ is density of the labels; $[d_j(x_i) = +1]$ denotes that the sample x_i is associated with the label d_j. When $\left[d_j(x_i) = +1\right]$ holds, $[\cdot]$ equal to 1, otherwise it is 0 [52].

The following all experiments were performed using MATLAB R2016b on Windows10 with the experimental platform of Inter(R) Core(TM) i5-8500 CPU at 3.00 GHz, memory 16.00 GB. Two classifiers MLKNN [52] and MLFE [53], which are used to prove the classification performance of MFSFN. The smoothing factor is equal to 1, and the size of nearest neighbor K is equal to 10 in MLKNN and MLFE [54]. We select several the common evaluation indexes of multi-label classification to evaluate the classification performance based on our proposed method in multi-label learning, including the number of selected features (N), average precision (AP), coverage (CV), Hamming loss (HL), one error (OE), ranking loss (RL), macro-averaging F1 (MacF1) and micro-averaging F1 (MicF1) [25,36,40,54], each of these indexes measures different aspects of the classification performance. The higher the value of AP, CV, MacF1 and MicF1 are, the better the classification performance is, and the lower the CV, OE, RL and HL are, the better the classification performance is. In the following experimental results, "↑" represents "the larger the better", and "↓" represents "the smaller the better". The number in bold indicates that this algorithm is better than other algorithms in the corresponding index.

Parameter discussion

Since parameters δ and λ will impact the classification performance of the MFSFN, to obtain the best classification results, in this subsection we will demonstrate the influence of parameters on the feature selection results. The parameter δ represents the fuzzy neighborhood radius, and the parameter λ is threshold to control the selection of feature subset. In this paper, we set the variation range of δ be [0,0.5] with step size of 0.05, and the variation range of λ is [0,1] with the step size of 0.05. As shown in Figs. 2 and 3, where the X-axis refers to the neighborhood radius δ, the Y-axis refers to λ that controls the selection of feature subset. We select the Scene dataset by our proposed algorithm MFSFN to demonstrate the training process that is the selection of parameters δ and λ under two classifiers MLKNN and MLFE. Finally, we select

Table 2 Description of the ten multi-label datasets

No.	Datasets	Samples	Features	Lables	LC	LD	Domain
1	Flags	194	19	7	3.392	0.485	Images
2	Cal500	502	68	174	26.044	0.150	Music
3	Emotions	593	72	6	1.868	0.311	Music
4	Scene	2407	294	6	1.074	0.179	Images
5	Yeast	2417	103	14	4.237	0.303	Biology
6	BBC	352	1000	6	1.125	0.188	Text
7	Gnegative	1392	440	8	1.046	0.131	Biology
8	Virus	207	440	6	1.217	0.203	Biology
9	Plant	978	440	12	1.069	0.089	Biology
10	Guardian	302	1000	6	1.126	0.188	Text

the most appropriate parameters for each multi-label dataset are shown in Tables 3 and 4.

The purpose of first portion is to analysis the change of evaluation indexes with parameters under classifier MLKNN. Figure 2 illustrates the change of each evaluation index with the parameters on the Scene dataset. For the Scene dataset, when $\delta = 0.15$, $\lambda = 0.65$, the five evaluation indexes AP, CV, RL, OE and N are the most appropriate. Therefore, the following will take $\delta = 0.15$, $\lambda = 0.65$ as the best parameter on the Scene dataset. Using the same process to obtain the best parameters of the other nine datasets from Table 2. The parameter values and evaluation index values are displayed in Table 3.

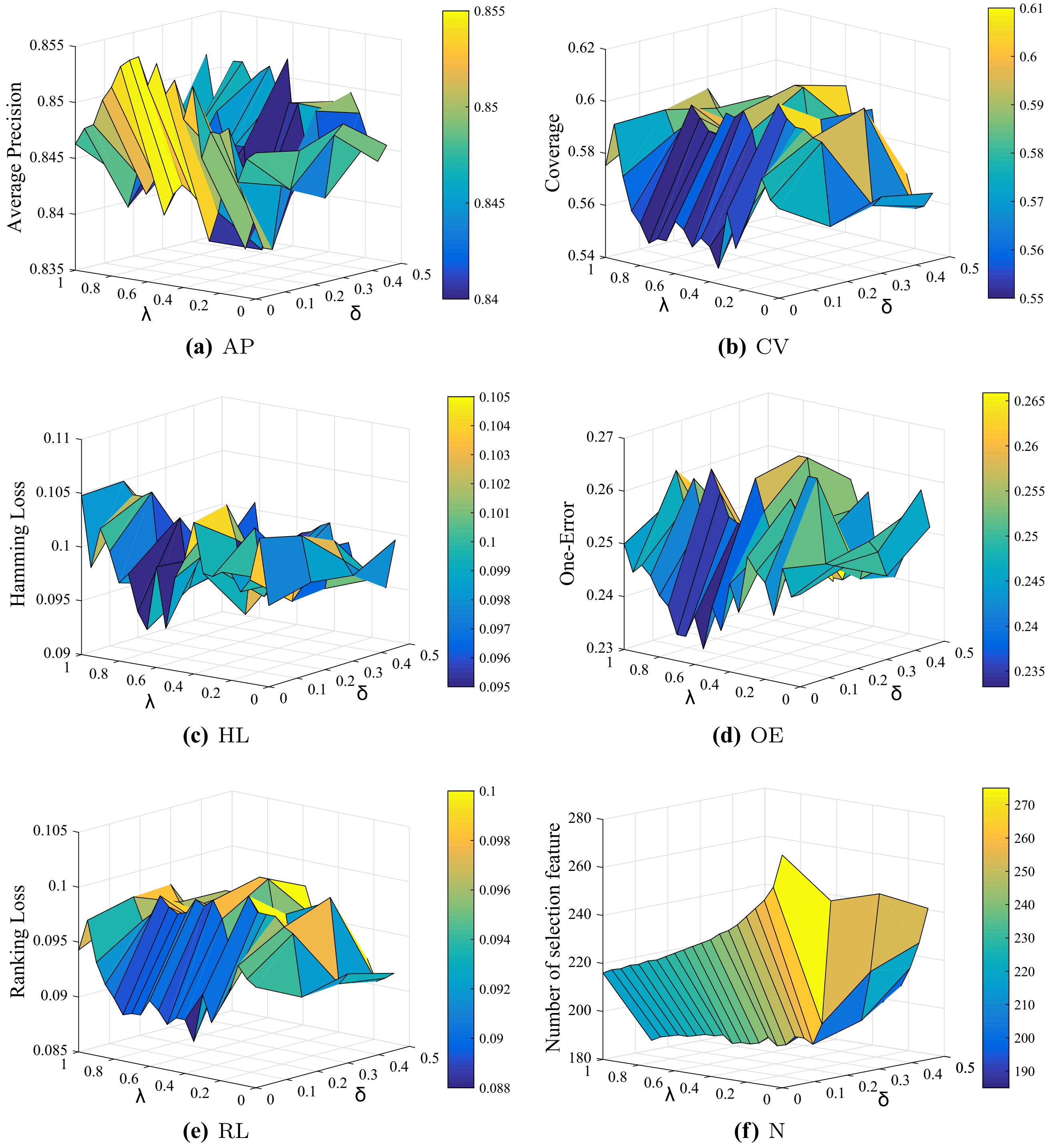
The second portion of this subsection is to analysis change of evaluation indexes with parameters under classifier MLFE. Figure 3 demonstrates the change of each evaluation index with parameters on the Scene dataset. For the Scene dataset, when $\delta = 0.05$, $\lambda = 1$, the eight evaluation indexes N, AP, HL, CV, OE, RL, MacF1 and MicF1 are the optimal value. Therefore, the following will take $\delta = 0.05$, $\lambda = 1$ as the best parameters on the Scene dataset, use the same procedure to get the best parameters of the other nine datasets from Table 2. The parameter values and each evaluation index value are shown in Table 4.

Comparison results of methods under MLKNN

This subsection exhibitions the comparison results of our proposed method with other related algorithms under MLKNN. First, our improved algorithm is compared with eight most advanced multi-label feature selection algorithms on the Scene dataset, including MLNB [55], MDDMspc [56], MDDMproj [56], PMU [57], RF-ML [58], MFNMIopt [25], MFNMIneu [25], MFNMIpes [25] were tested in aspects of AP, CV, HL and RL. Using the experimental techniques and results provided in [25], where μ is set as 0.5 in MDDMspc. The parameters δ and λ of MFSFN in the experiment select the optimal parameter values in Table 3. As shown in Table 5, it is the experimental result of comparing MFSFN on the Scene dataset with the other eight algorithms. The AP value of MFSFN is optimal, which is 0.0117 higher than MFNMIopt. On the CV index, MFSFN achieves lowest on the eight algorithms, which is 0.0292 lower than MDDMspc. The RL value of MFSFN is lower than other seven algorithms, where MFSFN is 0.0043 lower than MLNB. In terms of HL, MFSFN is compared with the other eight algorithms on the Scene dataset is ranked 2nd, and MFSFN is only 0.0002 higher than MFNMIopt, but MFSFN has obvious advantages over MFNMIopt for indexes AP, CV and RL. Obviously, for the Scene dataset, MFSFN achieves better results in each evaluation indications compared with other eight algorithms, and the validity of the selected parameters δ and λ is proved.

This part of the subsection adopts the classifier MLKNN and proves the validity of MFSFN in the aspects of N, AP,OE, CV and HL. Our method is compared with ParetoFS [59], ELA-CHI [60], PPT-CHI [61], and MUCO [62] on the Scene and Yeast datasets, the experimental techniques and results in reference [59] are used, as shown in Tables 6 and 7.

According to the experimental results in Table 6, the AP index of the proposed algorithm yields the most competitive performance on five algorithms. On the CV index, MFSFN has obvious advantages over other algorithms, MFSFN is 0.6526 lower than ELA-CHI. On the OE index, the proposed method achieves higher performance than other algorithms, which is 0.2649 lower than the algorithm ELA-CHI. On the HL index, the proposed algorithm obtains better results than other algorithms, and MFSFN is 0.0934 lower than MUCO. The number of selected features obtains fewest by algorithm ParetoFS, but our proposed method performs fairly better than ParetoFS in the aspects of AP, CV, HL and OE. In Table 7, we can observe that AP of the proposed algorithm has obvious advantages over other algorithms on the Yeast dataset, MFSFN is at least 0.0023 and at most 0.0248 larger than other algorithms. For CV, MFSFN achieves superior performance than other algorithms except for algorithm



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

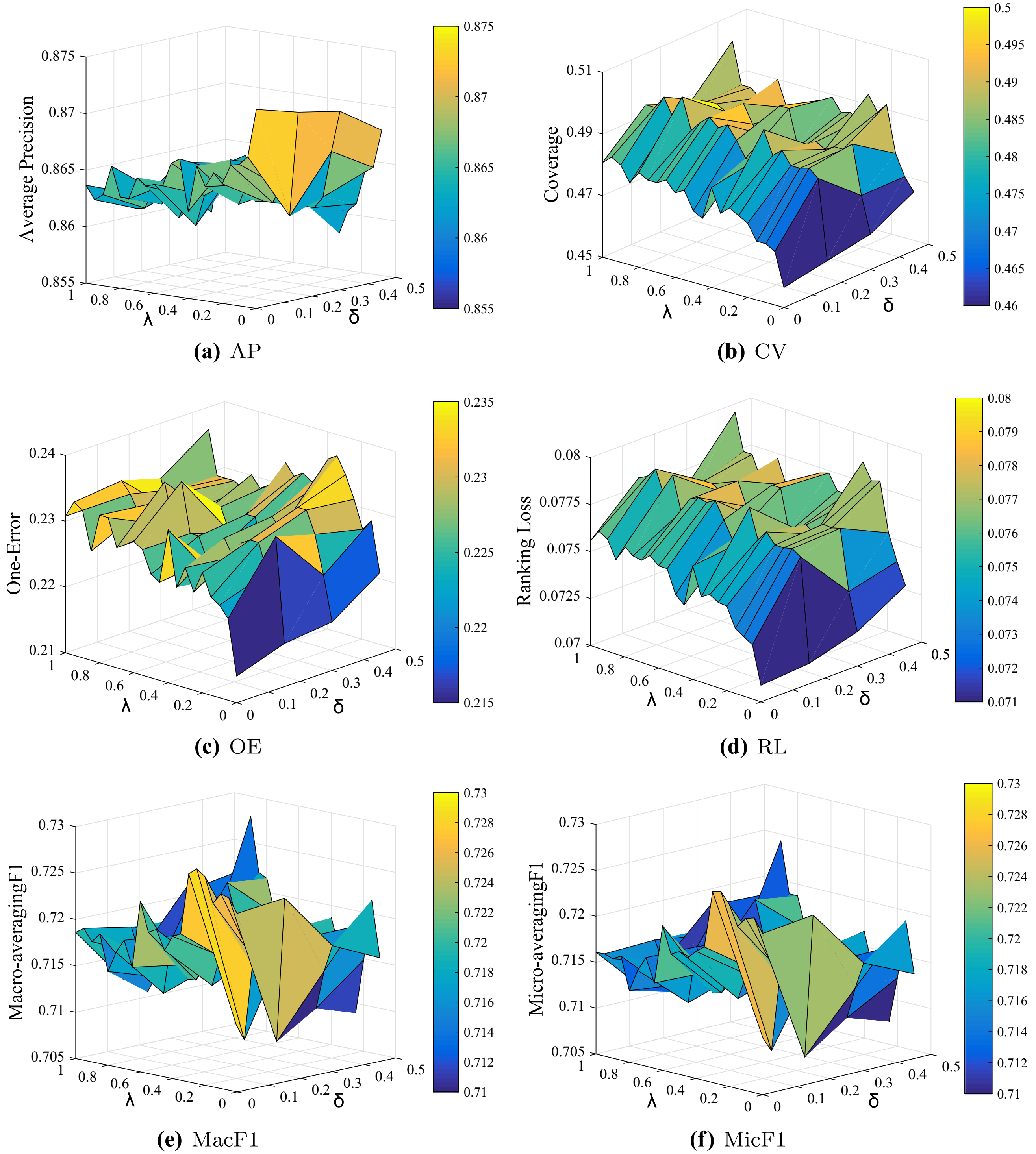
(c) HL **(d)** OE

(e) RL **(f)** N

Fig. 2 Variation of each evaluation index with parameters δ and λ on Scene dataset

ParetoFS, and ranks 2nd, but MFSFN performs better than ParetoFS in aspects of AP, OE and HL. As a whole, our proposed algorithm MFSFN has better classification performance than other algorithms on the Scene and Yeast datasets, the validity of the selected parameters δ and λ is proved.

Then seven multi-label datasets: Flags, Yeast, Plant, Gnegative, Virus, BBC and Guardian are selected from Table 2, carry out a series of experiments which compare the proposed algorithm MFSFN with the six advanced related algorithms, including RF-ML, PMU, MDDMproj, MDDMspc, FSRD [63] and MFSMR [20], the experimental techniques and



(a) AP

(b) CV

(c) OE

(d) RL

(e) MacF1

(f) MicF1

Fig. 3 Variation of each evaluation index with parameters δ and λ on Scene dataset

Table 3 The evaluation results of the ten datasets under classifier MLKNN

Datasets	(δ, λ)	N	AP (↑)	CV (↓)	RL (↓)	OE (↓)
Scene	(0.15,0.65)	197	0.8513	0.5711	0.0933	0.2383
Emations	(0.3,0.2)	43	0.7266	2.1461	0.2436	0.3539
Yeast	(0.3,0.9)	95	0.7607	6.3599	0.1706	0.2312
Flags	(0.3,0.15)	11	0.8357	3.8	0.2069	0.0923
Cal500	(0.05,0.3)	31	0.4841	132.07	0.1908	0.1111
Plant	(0,0)	158	0.5441	0.3538	0.2063	0.6462
Gnegative	(0,0)	216	0.7875	0.8399	0.1062	0.3237
Virus	(0.05,0)	62	0.6905	1.2530	0.2093	0.4819
BBC	(0,0.05)	48	0.5035	2.2768	0.4183	0.6964
Guardian	(0,0.05)	39	0.5012	2.1633	0.3959	0.7143

Table 4 The evaluation results of the ten datasets under classifier MLFE

Dataset	(δ, λ)	N	AP (↑)	CV (↓)	HL (↓)	RL (↓)	OE (↓)	MacF1 (↑)	MicF1 (↑)
Scene	(0.05,1)	184	0.8622	0.4933	0.0925	0.0780	0.2308	0.7124	0.7105
Emations	(0.05,0.15)	43	0.7751	1.7303	0.2228	0.1729	0.3202	0.5972	0.6124
Yeast	(0.3,0.85)	95	0.7560	6.4962	0.2056	0.1766	0.2388	0.4090	0.6228
Flags	(0.35,0.55)	12	0.8268	3.7385	0.2484	0.2162	0.1538	0.6368	0.7341
Cal500	(0.1,1)	21	0.4646	135.81	0.1535	0.2069	0.1786	0.1182	0.3702
Plant	(0,0)	158	0.5868	1.9	0.1644	0.0917	0.6103	0.1589	0.2909
Gnegative	(0.05,0)	198	0.8237	0.7374	0.0920	0.0764	0.2698	0.3745	0.6480
Virus	(0.05,0)	62	0.6538	1.5181	0.2653	0.2169	0.5181	0.3624	0.4194
BBC	(0.1,0.05)	64	0.5339	2.1429	0.3844	0.2426	0.6518	0.2553	0.2819
Guardian	(0.05,0.05)	46	0.5322	2.0816	0.3842	0.2364	0.6733	0.2215	0.2486

Table 5 The comparative of evaluation results among nine methods on the Scene dataset

Methods	AP (↑)	CV (↓)	RL (↓)	HL (↓)
MLNB	0.8351	0.5936	0.0976	0.0984
MDDMspc	0.8313	0.6212	0.1036	0.1028
MDDMproj	0.8383	0.6003	0.0990	0.1040
PMU	0.8277	0.6355	0.1066	0.1052
RF-ML	0.7933	0.7575	0.1307	0.1200
MFNMIopt	0.8396	0.6087	0.1015	**0.0964**
MFNMIneu	0.8302	0.6388	0.1074	0.1019
MFNMIopt	0.8169	0.6873	0.1711	0.1088
MFSFN	**0.8513**	**0.5711**	**0.0933**	0.0966

Table 6 The comparative of evaluation results among five methods on the Scene dataset

Methods	N	AP (↑)	CV (↓)	OE (↓)	HL (↓)
ParetoFS	**59**	0.7942	0.6947	0.3451	0.1225
ELA-CHI	100	0.6765	1.2237	0.5032	0.1605
PPT-CH	100	0.6972	1.1507	0.4720	0.1552
MUCO	87	0.7838	0.7617	0.3587	0.1900
MFSFN	197	**0.8513**	**0.5711**	**0.2383**	**0.0966**
Mean	109	0.7606	0.8804	0.3835	0.1450

Table 7 The comparative of evaluation results among five methods on the Yeast dataset

Methods	N	AP (↑)	CV (↓)	OE (↓)	HL (↓)
ParetoFS	56	0.7584	**6.3313**	0.2366	0.1976
ELA-CHI	100	0.7544	6.3906	0.2395	0.1984
PPT-CH	100	0.7529	6.4201	0.2415	0.1996
MUCO	**30**	0.7359	6.6838	0.2563	0.2052
MFSFN	95	**0.7607**	6.3599	**0.2312**	**0.1969**
Mean	76	0.7525	6.4371	0.2410	0.1995

results in reference [63] are used, and in reference [20], the number of missing labels is set up to 0. In the aspects of AP, CV, OE and RL, the results of this classification are demonstrated in Tables 8, 9, 10 and 11.

In Table 8, the OE index of MFSFN performs obvious advantages compared with other algorithms on most of the datasets, which exhibits superior performance against other algorithms on four datasets: Flags, Yeast, Plant and Virus, the highest value on Guardian dataset is achieved by FSRD and the highest value on Gnegative and BBC is achieved by MFSMR, the other four algorithms do not achieve optimal performance on all datasets. As an example, with the respect to the Flags dataset, the AP value of MFSFN is 0.8357, which compares better than other five algorithms with 0.8093 for MDDMproj, 0.8226 for MDDMspc, 0.7970 for PUM, 0.8148 for RF-ML,0.8288 for FSRD and 0.8182 for MFSMR. It is evident that the proposed method has the advantage over other methods.

From Table 9, in the CV index, MFSFN has obvious advantages compared with other five algorithms on the Yeast and Plant datasets. On the Gnegative dataset, MFSFN is inferior to MFSMR and MDDMproj, but it has obvious advantages over the other four algorithms. On the Virus dataset, the CV of MFSFN is 1.2530, is in close proximity to the lowest CV value of FSRD, 1.2417, which represents that our method has certain competitiveness with other methods. Additionally, MDDMproj, PMU and RF-ML do not outperform the other algorithms on any dataset. The CV value of MFSFN on the Guardian dataset is slightly lower than the algorithm FSRD, and ranks 2nd. In short, our proposed method is superior to other algorithms in most cases.

As seen from Table 10, on the datasets Yeast and Plant, the RL of the proposed method is obviously better than other algorithms. On the Virus dataset, MFSFN was 0.0186 lower than MDDMSPC and 0.0031 lower than the algorithm RF-ML. On the datasets Gnegative and Guardian, the RL value of MFSFN ranks 2nd. It is clear that the 2nd best performance for MFSFN is slightly inferior to FSRD or MFSMR, but better than other five algorithms.

As shown in Table 11, the OE index of MFSFN performs exhibits superior performance against other algorithms on three datasets Flags, Plant,and Virus. On the Yeast dataset, the best performance of OE is FSRD, our method is only 0.0072 larger than FSRD. On the dataset BBC, MFSFN is larger 0.268 of the lowest value which is achieved by the algorithm MFSMR and ranks 2nd. On the dataset Guardian, the proposed algorithm is slightly inferior to MDDMspc and RF-ML, but MFSFN is about 0.037 lower than PMU. On the whole, our proposed method is fairly well to other algorithms. From Tables 8, 9, 10 and 11, comprehensive analysis shows that our algorithm has higher classification performance than other algorithms in AP, CV, RL and OE.

To verify the validity and stability of proposed algorithm MFSFN, the experimental comparisons for multi-label classification on the selected features are carried out by fivefold cross-validation. We select four multi-label datasets of different fields from Table 2, including Yeast, Emotions, Scene and Cal500 datasets. Combine the proposed algorithm MFSFN with MUCO, MDDMproj, MDDMspc, PMU, MFS-KA [64] and RFNMIFS [39] in four multi-label datasets. The six comparison algorithms verify the validity of our proposed algorithm using the classification in AP, CV, OE, RL and HL measures, and using the experimental techniques and results in the literature [39], The results of classification are demonstrated in Tables 12, 13, 14, 15 and 16. From Table 12, the index AP of MFSFN apparently outperforms other algorithms on the four datasets Yeast, Emotions, Cal500 and Scene; as an example, with respect to the Scene dataset, the maximum value of MFSFN is 0.0099 lower than MDDMspc, and the minimum value of MFSFN is 0.0941 higher than MDDMspc. Thus, MFSFN obtained better classification performance than other algorithms on AP. As can be seen from Table 13 that the CV value of MFSFN has a significant advantage over other algorithms on the three datasets: Yeast, Emotions and Scene. On the Cal500 dataset, the proposed algorithm MFSFN is 0.0917 higher than the minimum value of RFNMIFS, but the maximum value of MFSFN is 0.6717 lower than RFNMIFS, so MFSFN is more stable than RFNMIFS. As shown in Table 14, for the Yeast, Emotions and Scene datasets in metrics of OE, MFSFN achieves the lowest mean values. On the Cal500 dataset, the lowest value of RFNMIFS is 0.0143 lower than that of the MFSFN algorithm, but the highest value of RFNMIFS is 0.0047 higher than that of MFSFN, which proves that the stability of MFSFN is stronger than other algorithms. It can be seen from Table 15 that the RL of MFSFN is significantly better than other the six algorithms and obtains satisfactory results on the four datasets. From Table 16, the HL of MFSFN is better than other algorithms on the Yeast, Scene, Emotions and Cal500 four datasets. The results show that our algorithm can not only eliminate the redundant features on the four datasets, but also achieve better performance than other six algorithms in terms of AP, CV, OE, RL and HL.

Comparison results of methods under MLFE

This subsection illustrates the performance of the proposed method by comparing with other methods under classifier MLFE. We select three datasets from Table 2, including Flags, Yeast and Scene. MFSFN is compared with six most advanced multi-label feature selection methods, including PCT-CHI2 [19], CSFS [65], SFUS [66], Avg.CHI [67], MCLS [54], and RFNMIFS, on three multi-label datasets. The algorithm MFSFN is tested on the aspects of AP, CV, OE, RL, MacF1, and MicF1, the experimental techniques and results in reference [39] are used, as shown in Tables 17, 18 and 19. MFSFN prevails over other algorithms for the optimal mean values in the each evaluation index. It can be seen from Table 17 that the six metrics of MFSFN are better than other algorithms in the Flags dataset. The CV, MacF1 and MicF1 value of MFSFN have obvious advantages against the other six algorithms. On the RL index, MFSFN is 0.0172 higher than the lowest value of RFNMIFS and 0.0112 lower than its highest value. On the whole, MFSFN has better classifica-

Table 8 AP (↑) index of the seven methods on the seven datasets

Datasets	MDDMproj	MDDMspc	PMU	RF-ML	FSRD	MFSMR	MFSFN
Flags	0.8093	0.8226	0.7970	0.8148	0.8288	0.8182	**0.8357**
Yeast	0.7447	0.7469	0.7507	0.7571	0.7584	**0.7581**	**0.7607**
Plant	0.5182	0.5299	0.5276	0.5136	0.5322	0.5307	**0.5441**
Gnegative	0.7770	0.7670	0.7642	0.7753	0.7785	**0.8074**	0.7875
Virus	0.6853	0.6420	0.6534	0.6845	0.6886	0.6829	**0.6905**
BBC	0.4692	0.4867	0.4850	0.4995	0.5034	**0.5299**	0.5035
Guardian	0.4910	0.5036	0.4713	0.5037	**0.5057**	0.4928	0.5012

Table 9 CV (↓) index of the seven methods on the seven datasets

Datasets	MDDMproj	MDDMspc	PMU	RF-ML	FSRD	MFSMR	MFSFN
Flags	3.7329	3.6779	3.8782	3.7787	**3.6671**	3.7231	3.8
Yeast	6.4753	6.5020	6.4108	6.3888	6.4053	6.3795	**6.3599**
Plant	2.5865	2.5918	2.5929	2.6530	2.5558	2.4590	**2.3538**
Gnegative	0.8270	0.8904	0.9110	0.8457	0.8414	**0.7230**	0.8399
Virus	1.2471	1.3536	1.3000	1.3090	**1.2417**	1.2771	1.2530
BBC	2.3210	2.3138	2.3252	2.2537	2.2400	**2.0714**	2.2768
Guardian	2.2298	2.2402	2.3622	2.2473	**2.1510**	2.3656	2.1633

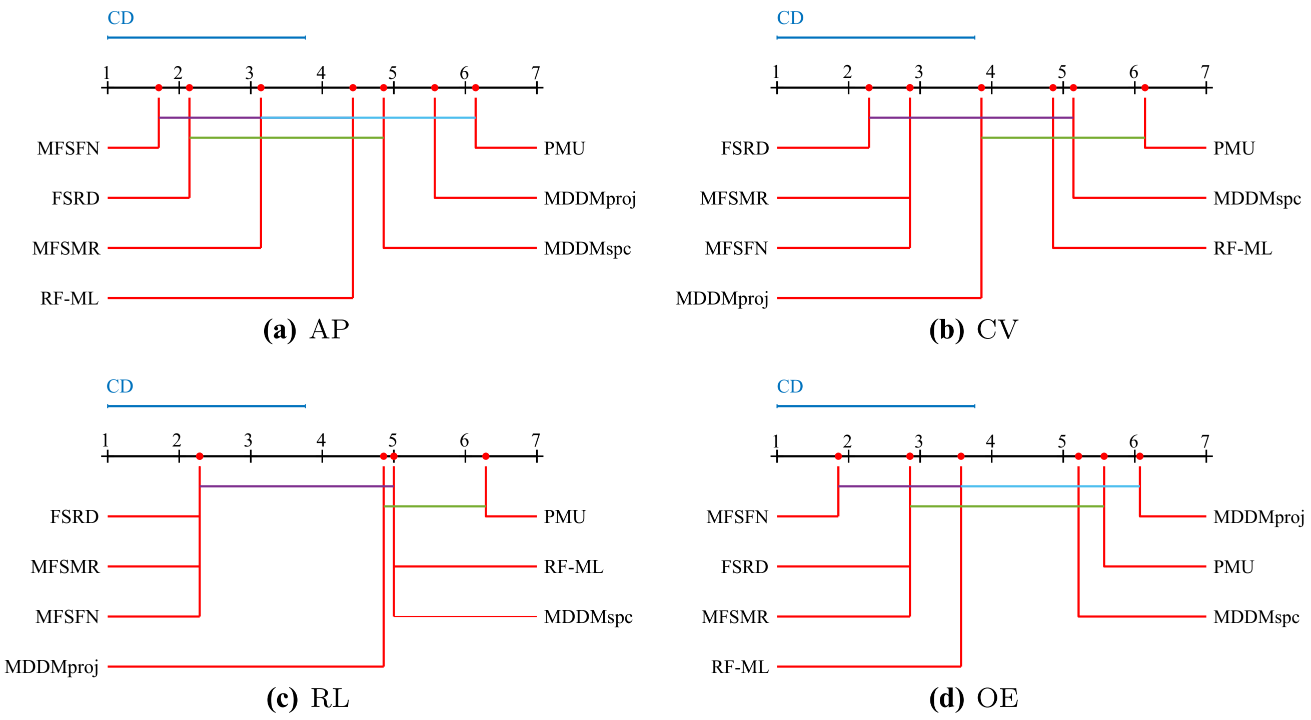
Table 10 RL (↓) index of the seven methods on the seven datasets

Datasets	MDDMproj	MDDMspc	PMU	RF-ML	FSRD	MFSMR	MFSFN
Flags	0.2121	0.2061	0.2359	0.2135	**0.1990**	0.2056	0.2069
Yeast	0.1806	0.1796	0.1755	0.1731	0.1752	0.1716	**0.1706**
Plant	0.2232	0.2226	0.2233	0.2299	0.2198	0.2158	**0.2063**
Gnegative	0.1078	0.1170	0.1201	0.1112	0.1097	**0.0904**	0.1062
Virus	0.2045	0.2279	0.2183	0.2124	**0.2022**	0.2126	0.2093
BBC	0.4343	0.4291	0.4311	0.4193	0.4129	**0.3725**	0.4183
Guardian	0.4146	0.4127	0.4379	0.4165	**0.3931**	0.4023	0.3959

tion performance on the Flags dataset. From Table 18, the AP, MacF1 and MicF1 indexes of MFSFN are better than other algorithms on the Yeast dataset; on the CV, MFSFN is 0.0176 larger than the lowest value of RFNMIFS and 0.2832 lower than the highest value of RFNMIFS, so MFSFN has better performance for CV. For the OE index, MFSFN is 0.0056 higher than the lowest value of RFNMIFS, but 0.0164 lower than the highest value, in short, MFSFN still has advantages in OE measurement. In the terms of RL, MFSFN is 0.0010 higher than the lowest value of RFNMIFS, but 0.0216 lower than its highest value. MFSFN is more stable than other algorithms. It can be seen from Table 19 that the six indicators of MFSFN are significantly better than other algorithms on the Scene dataset. Based on the above analysis of the classification results of the three datasets on MLFE classifier, MFSFN algorithm can not only effectively eliminate the redundant features of the three datasets, but also has higher classification performance than other algorithms.

Table 11 OE (↓) index of the seven methods on the seven datasets

Datasets	MDDMproj	MDDMspc	PMU	RF-ML	FSRD	MFSMR	MFSFN
Flags	0.2450	0.2274	0.2368	0.2213	0.2058	0.2154	**0.0923**
Yeast	0.2491	0.2491	0.2420	0.2246	**0.2230**	0.2334	0.2312
Plant	0.6871	0.6708	0.6626	0.6892	0.6677	0.6693	**0.6462**
Gnegative	0.3506	0.3621	0.3656	0.3492	0.3456	**0.3040**	0.3237
Virus	0.5069	0.5888	0.5745	0.4967	0.5067	0.4940	**0.4819**
BBC	0.7640	0.7246	0.7303	0.7106	0.7102	**0.6696**	0.6964
Guardian	0.7381	0.7051	0.7513	**0.7048**	0.7182	0.7347	0.7143



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

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

Fig. 4 Comparison of the seven methods with Bonferroni–Dunn test under MLKNN

Statistical analysis

To systematically analyze the classification performance of MFSFN and intuitively display the statistical performance of each evaluation index under various comparison algorithms, Friedman statistical test [68] and Bonferroni–Dunn test [69] are used in this section. Friedman test is demonstrated as follows

$$\chi_F^2 = \frac{12T}{M(M+1)}\left(\sum_{i=1}^{M} R_i^2 - \frac{M(M+1)^2}{4}\right), \tag{30}$$

$$F_F = \frac{(T-1)\chi_F^2}{T(M-1)-\chi_F^2}, \tag{31}$$

where M and T are the numbers of methods and datasets, respectively, and R_i is the average order value of the $i-th$ method in all datasets. In the Bonferroni–Dunn test, the average rank difference between methods is calculated to evaluate whether there are significant differences between methods. The critical difference is expressed as follows

$$(CD)_\alpha = q_\alpha\sqrt{\frac{M(M+1)}{6T}}, \tag{32}$$

where q_α indicates the critical tabulated value of the test, and α represents the significance level. According to the statistical tests in references [36,70], the mean order of all datasets is obtained by averaging all levels on each metric. The optimal value under each index is set to the rank of 1, the second is set to the rank of 2, and so on. With CD value chart is used to visually display MFSFN correlations with other algorithms, each of these algorithms, the average ranking of each method is drew along the axis, in which the rank value on the axis increases from left to right. The MFSFN and compared algorithms are linked together with a thick line if the mean rank difference between these algorithms is within a criticality difference, indicating there is no significant difference between algorithms; otherwise, any algorithm that is not connected together will be considered markedly different from the other algorithms.

From the classification results in Tables 8, 9, 10 and 11, we can get the average ranking of MFSFN and six comparison algorithms on the four aspects of AP, CV, RL and OE under the MLKNN classifier, and the corresponding F_F values are demonstrated in Table 20. When the significance level $\alpha = 0.1$, each indicator rejects the zero hypotheses that seven algorithms have the same performance under the Friedman test. At that time, $q_\alpha = 2.394$, then CD = 2.7644 ($M = 7, T = 7$). The accuracy comparison of seven algorithms by Bonferroni–Dunn test is demonstrated in Fig. 4. It can be obtained from Fig. 4 that MFSFN is significantly better than other algorithms in AP and OE evaluation indicators. From Fig. 4a, for the metric AP, the proposed algorithm has obvious advantages compared with PMU, MDDMspc, MDDMproj, RF-ML, and MFSFN is no significant difference with algorithms FSRD and MFSMR. It can be seen

Table 12 AP (↑) index of the seven methods on the four datasets

Datasets	MUCO	MDDMproj	MDDMspc	PMU	MFS-KA	RFNMIFS	MFSFN
Yeast	0.7390 ± 0.0103	0.7394 ± 0.0015	0.7432 ± 0.0157	0.7315 ± 0.0105	0.7410 ± 0.0102	0.7500 ± 0.0111	**0.7664 ± 0.0037**
Emotions	0.6967 ± 0.0073	0.7141 ± 0.0197	0.7134 ± 0.0270	0.7474 ± 0.0106	0.7510 ± 0.0060	0.7512 ± 0.0054	**0.7987 ± 0.0050**
Scene	0.8167 ± 0.0203	0.8138 ± 0.0412	0.8113 ± 0.0574	0.7521 ± 0.0692	0.8336 ± 0.0095	0.8343 ± 0.0164	**0.8534 ± 0.0054**
Cal500	0.4762 ± 0.0005	0.4762 ± 0.0014	0.4758 ± 0.0012	0.4762 ± 0.0013	0.4795 ± 0.0014	0.4819 ± 0.0011	**0.4921 ± 0.0022**

Table 13 CV (↓) index of the seven methods on the four datasets

Datasets	MUCO	MDDMproj	MDDMspc	PMU	MFS-KA	RFNMIFS	MFSFN
Yeast	6.6124 ± 0.1233	6.6266 ± 0.1340	6.5936 ± 0.1545	6.6759 ± 0.1100	6.5462 ± 0.0928	6.5032 ± 0.1069	**6.2531 ± 0.0244**
Emotions	2.4629 ± 0.0800	2.3441 ± 0.1464	2.3581 ± 0.1811	2.1155 ± 0.0393	1.9801 ± 0.0577	1.9512 ± 0.0613	**1.7989 ± 0.0400**
Scene	0.6610 ± 0.0617	0.6992 ± 0.0617	0.6962 ± 0.2250	0.9391 ± 0.2789	0.5840 ± 0.0257	0.5796 ± 0.0386	**0.5087 ± 0.0155**
Cal500	131.90 ± 0.0734	131.85 ± 0.3252	131.81 ± 0.3104	131.85 ± 0.3263	131.45 ± 0.4842	130.61 ± 0.7721	**130.32 ± 0.3904**

Table 14 OE (↓) index of the seven methods on the four datasets

Datasets	MUCO	MDDMproj	MDDMspc	PMU	MFS-KA	RFNMIFS	MFSFN
Yeast	0.2526 ± 0.0078	0.2502 ± 0.0100	0.2457 ± 0.0113	0.2547 ± 0.0066	0.2437 ± 0.0065	0.2383 ± 0.0078	**0.2280 ± 0.0080**
Emotions	0.4059 ± 0.0050	0.3977 ± 0.0223	0.3911 ± 0.0206	0.3672 ± 0.0175	0.3593 ± 0.0193	0.3505 ± 0.0191	**0.2771 ± 0.0113**
Scene	0.3015 ± 0.0348	0.2988 ± 0.0601	0.3067 ± 0.0845	0.3928 ± 0.1005	0.2736 ± 0.0092	0.2692 ± 0.0302	**0.2466 ± 0.0092**
Cal500	0.1217 ± 0.0044	0.1243 ± 0.0055	0.1237 ± 0.0070	0.1243 ± 0.0055	0.1245 ± 0.0071	0.1172 ± 0.0073	**0.1175 ± 0.0023**

Table 15 RL (↓) index of the seven methods on the four datasets

Datasets	MUCO	MDDMproj	MDDMspc	PMU	MFS-KA	RFNMIFS	MFSFN
Yeast	0.1869 ± 0.0086	0.1875 ± 0.0110	0.1839 ± 0.0113	0.1919 ± 0.0082	0.1819 ± 0.0070	0.1794 ± 0.0083	**0.1655 ± 0.0019**
Emotions	0.2778 ± 0.0127	0.2541 ± 0.0028	0.2573 ± 0.0351	0.2076 ± 0.0089	0.2068 ± 0.0103	0.2033 ± 0.0089	**0.1636 ± 0.0064**
Scene	0.1114 ± 0.0127	0.1190 ± 0.0326	0.1183 ± 0.0448	0.1668 ± 0.0551	0.1000 ± 0.0056	0.0990 ± 0.0014	**0.0841 ± 0.0032**
Cal500	0.1929 ± 0.0003	0.1928 ± 0.0002	0.1927 ± 0.0002	0.1928 ± 0.0002	0.1902 ± 0.0010	0.1894 ± 0.0011	**0.1833 ± 0.0009**

Table 16 HL (↓) index of the seven methods on the four datasets

Datasets	MUCO	MDDMproj	MDDMspc	PMU	MFS-KA	RFNMIFS	MFSFN
Yeast	0.2082 ± 0.0068	0.2097 ± 0.0087	0.2077 ± 0.0101	0.2149 ± 0.0069	0.2053 ± 0.0044	0.2052 ± 0.0066	**0.1927 ± 0.0018**
Emotions	0.2723 ± 0.0117	0.2592 ± 0.0151	0.2663 ± 0.0279	0.2496 ± 0.0064	0.2402 ± 0.0066	**0.2324 ± 0.0085**	0.3108 ± 0.0058
Scene	0.1026 ± 0.0267	0.0983 ± 0.0197	0.0998 ± 0.0258	0.1331 ± 0.0267	0.1004 ± 0.0051	0.0978 ± 0.0111	**0.0931 ± 0.0005**
Cal500	0.1439 ± 0.0007	0.1456 ± 0.0015	0.1456 ± 0.0015	0.1334 ± 0.0007	0.1245 ± 0.0071	**0.1172 ± 0.0073**	0.1674 ± 0.0004

Table 17 Classification results of the seven methods on the Flags dataset

Index	PCT-CHI2	MCLS	CSFS	SFUS	Avg.CHI	RFNMIFS	MFSFN
AP (↑)	0.7691 ± 0.0041	0.7955 ± 0.0159	0.7675 ± 0.0083	0.7747 ± 0.0037	0.7663 ± 0.0036	0.8121 ± 0.0058	**0.8152 ± 0.0030**
CV (↓)	4.0379 ± 0.0507	3.8392 ± 0.1339	4.0309 ± 0.0475	4.0243 ± 0.0167	4.0364 ± 0.0266	3.7885 ± 0.1037	**3.7480 ± 0.0236**
OE (↓)	0.3036 ± 0.0177	0.2500 ± 0.0435	0.3056 ± 0.0265	0.2958 ± 0.0231	0.3062 ± 0.0165	0.2231 ± 0.0189	**0.2134 ± 0.0102**
RL (↓)	0.2619 ± 0.0049	0.2279 ± 0.0200	0.2640 ± 0.0089	0.2595 ± 0.0030	0.2641 ± 0.0042	0.2191 ± 0.0094	**0.2126 ± 0.0047**
MacF1 (↑)	0.3847 ± 0.0238	0.5349 ± 0.1088	0.4343 ± 0.0118	0.4227 ± 0.0119	0.4361 ± 0.0108	0.4458 ± 0.0214	**0.6281 ± 0.0183**
MicF1 (↑)	0.6327 ± 0.0093	0.6787 ± 0.0438	0.5990 ± 0.0329	0.6305 ± 0.0111	0.6325 ± 0.0075	0.6668 ± 0.0168	**0.7256 ± 0.0087**

Table 18 Classification results of the seven methods on the Yeast dataset

Index	PCT-CHI2	MCLS	CSFS	SFUS	Avg.CHI	RFNMIFS	MFSFN
AP (↑)	0.7249 ± 0.0174	0.7533 ± 0.0266	0.7058 ± 0.0191	0.7278 ± 0.0193	0.7296 ± 0.0214	0.7533 ± 0.0147	**0.7659 ± 0.0032**
CV (↓)	6.6500 ± 0.2305	6.4776 ± 0.4069	6.7498 ± 0.2083	6.7011 ± 0.1849	6.6707 ± 0.2605	6.4228 ± 0.1638	**6.2900 ± 0.0134**
OE (↓)	0.3068 ± 0.0183	0.2593 ± 0.0372	0.2807 ± 0.0167	0.2768 ± 0.0193	0.2793 ± 0.0155	0.2449 ± 0.0178	**0.2395 ± 0.0068**
RL (↓)	0.1936 ± 0.0125	0.1800 ± 0.0231	0.2051 ± 0.0153	0.1946 ± 0.0153	0.1952 ± 0.0124	0.1758 ± 0.0131	**0.1655 ± 0.0018**
MacF1 (↑)	0.3424 ± 0.0606	0.3849 ± 0.0654	0.3546 ± 0.0776	0.3567 ± 0.0594	0.3425 ± 0.0639	0.3883 ± 0.0375	**0.4242 ± 0.0043**
MicF1 (↑)	0.5933 ± 0.0378	0.6157 ± 0.0404	0.5766 ± 0.0401	0.5938 ± 0.0336	0.5955 ± 0.0371	0.6186 ± 0.0185	**0.6437 ± 0.0042**

Table 19 Classification results of the seven methods on the Scene dataset

Index	PCT-CHI2	MCLS	CSFS	SFUS	Avg.CHI	RFNMIFS	MFSFN
AP (↑)	0.7166 ± 0.1050	0.7742 ± 0.1255	0.7393 ± 0.1263	0.7830 ± 0.1250	0.7191 ± 0.1064	0.8067 ± 0.0708	**0.8857 ± 0.0006**
CV (↓)	1.0208 ± 0.4689	0.7750 ± 0.5889	0.8338 ± 0.6157	0.7292 ± 0.5652	0.9913 ± 0.4808	0.7227 ± 0.2539	**0.3940 ± 0.0008**
OE (↓)	0.4372 ± 0.1498	0.3369 ± 0.1881	0.3992 ± 0.1848	0.3206 ± 0.1859	0.4258 ± 0.1558	0.3147 ± 0.1111	**0.1965 ± 0.0012**
RL (↓)	0.1912 ± 0.0938	0.1396 ± 0.1144	0.2824 ± 0.1742	0.1279 ± 0.1165	0.1792 ± 0.0989	0.1207 ± 0.0518	**0.0619 ± 0.0003**
MacF1 (↑)	0.4658 ± 0.1999	0.5717 ± 0.2463	0.4455 ± 0.3019	0.6013 ± 0.2458	0.4625 ± 0.1942	0.5857 ± 0.1247	**0.7552 ± 0.0080**
MicF1 (↑)	0.4642 ± 0.1986	0.5842 ± 0.2474	0.4998 ± 0.2400	0.6018 ± 0.2426	0.4698 ± 0.1982	0.5845 ± 0.1230	**0.7500 ± 0.0062**

Table 20 Values of the four evaluation indexes with F_F under MLKNN

	AP	CV	OE	RL
χ_F^2	18.3552	18.3625	28.3385	25.3740
F_F	4.6577	4.6610	12.4459	9.1570

Table 21 Values of the five evaluation indexes with F_F under MLKNN

	AP	CV	OE	RL	HL
χ_F^2	17.3571	20.4643	16.2589	20.3839	9.0804
F_F	7.8387	17.3636	6.3010	16.91111	1.8259

from Fig. 4b, in the aspect of CV index, there is no significant difference with algorithm MFSMR, and it has obvious advantages over algorithms MDDMproj, PMU, RF-ML and MDDMspc, and there is no definite evidence to prove that MDDMproj, RF-ML, MDDMspc and PMU have prominent differences. As shown in Fig. 4c, in metric of RL index, the algorithms MFSFN, FSRD and MFSMR is significantly better than MDDMproj, RF-ML, MDDMspc and PMU, and there is no consistent evidence to prove a statistical equivalence between MDDMproj, PMU, RF-ML and MDDMspc. From Fig. 4d, the OE index of the algorithm MFSFN is distinctly better than other algorithms, and and the distinction among the performance of FSRD, MFSMR, RF-ML, PMU and MDDMspc is insignificant. To sum up, the proposed algorithm has more excellent classification performance than other algorithms.

From the classification results which are illustrated in Tables 12, 13, 14, 15 and 16, we can get the average ranking of the proposed method and six comparison algorithms on the five aspects of AP, CV, HL, OE and RL under the MLKNN classifier, and the corresponding F_F values are displayed in Table 21. When the significance level $\alpha = 0.1$, each indicator rejects the zero hypotheses that seven algorithms have the same performance under the Friedman test. At that time, q_α= 2.394, then CD = 3.6569 ($M = 7, T = 4$). The accuracy comparison of seven algorithms by the Bonferroni–Dunn test is demonstrated in Fig. 5. It can be seen from Fig. 5 that MFSFN is significantly better than other algorithms in each index. Fig. 5a illustrates that in terms of AP, MFSFN achieves significantly better than four algorithms PMU, MDDMspc, MDDMproj and MUCO and obtains comparable results against MFS-KA and RFNMIFS. As can be seen from Fig. 5b, d, the CV and RL of algorithm MFSFN outperforms PMU, MUCO and MDDMproj and comparable to MDDMspc, MFS-KA and RFNMIFS, and there is no full evidence to demonstrate a statistical equivalence with RFNMIFS, MFS-KA, MDDMspc, MDDMproj, MUCO and PMU. As can be obtained from Fig. 5c, for the index OE, MFSFN is significantly better than other algorithms and comparable to RFNMIFS, MFS-KA and MDDMspc, and there is no consistent evidence to indicate a statistical equivalence with RFNMIFS, MFS-KA, MDDMspc, PMU and MDDMproj, and there is no concrete evidence to determine the significant difference among MFS-KA, MDDMspc, PMU, MDDMproj and MUCO. It can be obtained from Fig. 5e that HL index of MFSFN is more excellent than MDDMspc, MDDMproj, MUCO and PMU. In general, MFSFN has strong classification performance compared with other algorithms under classifier MLKNN.

The classification results in Tables 17, 18 and 19 were statistically tested under the classifier MLFE. The F_F values of the six metrics are listed in Table 22. When $\alpha = 0.1$, $q_\alpha = 2.394$, then CD = 4.2226 ($M = 7, T = 3$). The test results are demonstrated in Fig. 6. As can be obtained from Fig. 6a, for the AP index, MFSFN performs better than PCT-CHI2 and CSFS and is comparable to RFNMIFS, MCLS, SFUS and Avg.CHI. As can be seen from Fig. 6b, e, there is not enough evidence to suggest a statistical equivalence among MFSFN and RFNMIFS, MCLS, SFUS, CSFS and Avg.CHI in the aspects of CV and MacF1, and it is significantly superior to PCT-CHI2. As can be seen from Fig. 6c, there is no obvious difference between algorithm MFSFN and RFNMIFS, MCLS, SFUS and CSFS in the OE index, and it is superior to algorithms Avg.CHI and PCT-CHI2. As can be seen from Fig. 6d, for RL index, MFSFN is comparable to RFNMIFS, MCLS, SFUS and PCT-CHI2, and performs better than Avg.CHI and CSFS. As can be seen from Fig. 6f, in metric of MicF1, algorithm MFSFN is comparable to RFNMIFS, MCLS, SFUS, PCT-CHI2 and Avg.CHI, and is significantly superior to CSFS. Therefore, under the classifier MLFE, the algorithm MFSFN has more excellent performance compared with the other algorithms in general.

Conclusion

In this article, a multi-label feature selection method based on fuzzy neighborhood rough sets was improved by combining information view with the algebraic view, which achieved highly classification performance in the multi-label fuzzy neighborhood decision system. First, a new multi-label fuzzy neighborhood rough set model was proposed by combining NRS with FRS. Second, the fuzzy similarity matrix was obtained by computing the similarity between samples under different condition attributes, and a new multi-label fuzzy decision was proposed and the fuzzy neighborhood approximation accuracy was defined. Then, the fuzzy neighborhood conditional entropy was introduced, according to the concept of information entropy in information theory, and a hybrid metric was designed by combining the fuzzy neighborhood approximate accuracy with the fuzzy neighborhood condi-

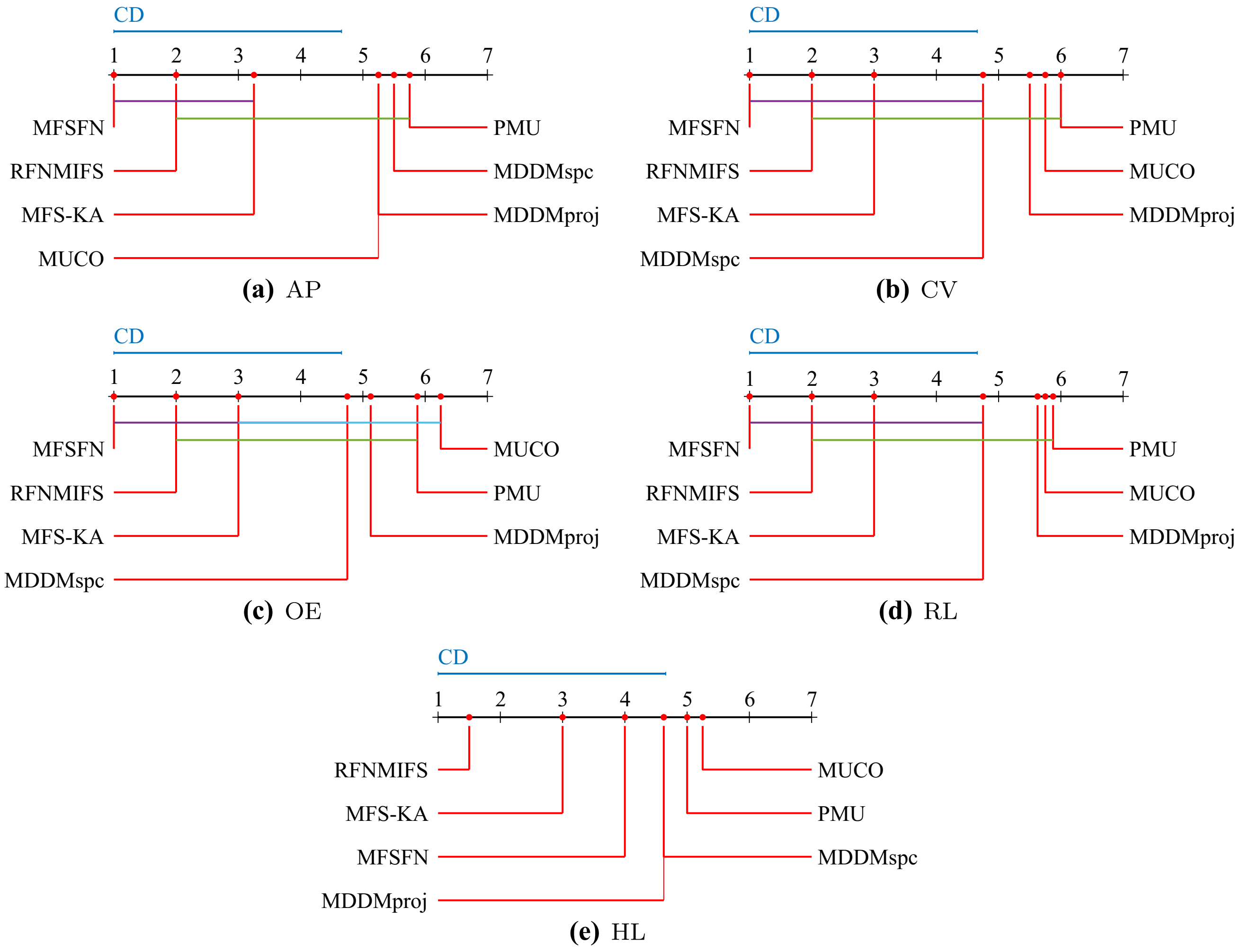


Fig. 5 Comparison of the seven methods with Bonferroni–Dunn test under MLKNN

Table 22 Values of the six evaluation indexes with F_F under the MLFE

	AP	CV	OE	RL	MacF1	MicF1
χ_F^2	13.1057	14.8729	14.0943	16.4429	13.9857	13.5615
F_F	5.3555	9.5122	7.2173	21.1195	6.9680	6.1108

tional entropy, to measure the importance of each attribute. Finally, a multi-label feature selection method based on fuzzy neighborhood rough sets was developed, a novel forward search algorithm for multi-label feature selection is provided. A series of experiments on ten multi-label datasets verify the effectiveness of the proposed algorithm in multi-label classification. In our future work, we will seek multi-label feature selection method of higher classification performance, and more efficient search strategies.

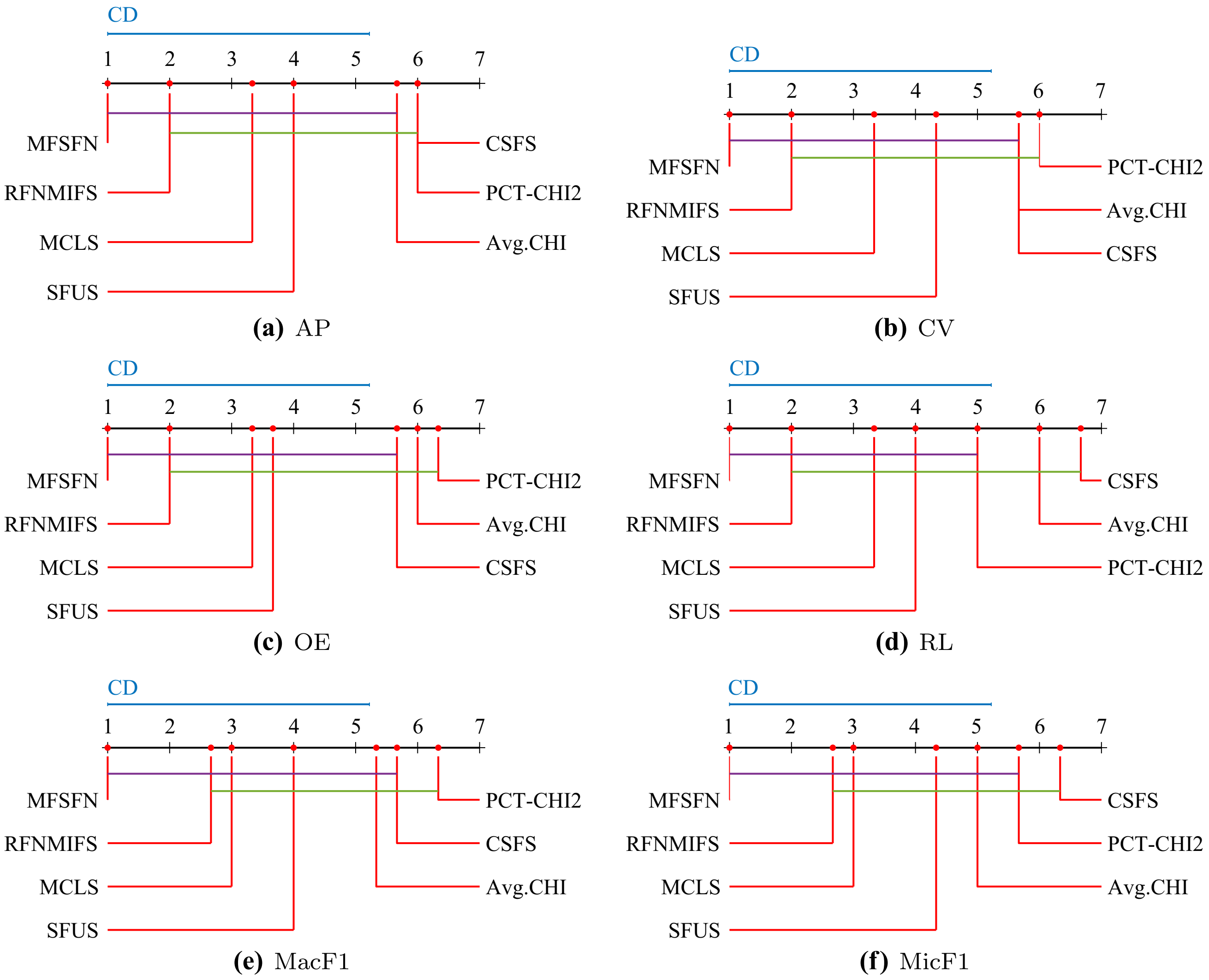


Fig. 6 Comparison of the seven methods with Bonferroni–Dunn test under MLFE

Acknowledgements This research was funded by the National Natural Science Foundation of China under Grants 61976082, 62076089 and 62002103; and the Key Scientific and Technological Project of Henan Province under Grant 212102210136.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Che X-Y, Chen D-G, Mi J-S (2019) A novel approach for learning label correlation with application to feature selection of multi-label data. Inf Sci 512:795–812
2. Huang M-M, Sun L, Xu J-C, Zhang S-G (2020) Multilabel feature selection using Relief and minimum redundancy maximum relevance based on neighborhood rough sets. IEEE Access 8(99):62011–62031
3. Che X-Y, Chen D-G, Mi J-S (2021) Feature distribution-based label correlation in multi-label classification. Int J Mach Learn Cybern 12(6):1705–1719
4. Qian W-B, Huang J-T, Wang Y-L, Xie Y-H (2021) Label distribution feature selection for multi-label classification with rough set. Int J Approx Reason 128:32–55

5. Zhang P, Gao W-F (2021) Feature relevance term variation for multi-label feature selection. Appl Intell. https://doi.org/10.1007/s10489-020-02129-w
6. Fujita H, Gaeta A, Loia V, Orciuoli F (2019) Hypotheses analysis and assessment in counterterrorism activities: A method based on OWA and fuzzy probabilistic rough sets. IEEE Trans Fuzzy Syst 28(5):831–845
7. Yue X-D, Chen Y-F, Miao D-Q, Fujita H (2020) Fuzzy neighborhood covering for three-way classification. Inf Sci 507:795–808
8. Zhang M-L, Zhou Z-H (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
9. Zhang M-L, Zhou Z-H (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
10. Boutell M-R, Luo J, Shen X, Brown C-M (2004) Learning multi-label scene classification. Pattern Recogn 37(9):1757–1771
11. Arslan S, Ozturk C (2019) Multi hive artificial bee colony programming for high dimensional symbolic regression with feature selection. Appl Soft Comput 78:515–527
12. Chen S-B, Zhang Y-M, Ding H-Q, Zhang J, Luo B (2019) Extended adaptive Lasso for multi-class and multi-label feature selection. Knowl Based Syst 173:28–36
13. Jiang Z-H, Liu K-Y, Yang X-B, Yu H-L, Fujita H, Qian Y-H (2020) Accelerator for supervised neighborhood based attribute reduction. Int J Approx Reason 119:122–150
14. Xu T-T, Zhao L (2020) A structure-induced framework for multi-label feature selection with highly incomplete labels. IEEE Access 8:71229–71230
15. Zhang P, Gao W-F, Hu J-C, Li Y-H (2021) Multi-label feature selection based on the division of label topics. Inf Sci 553:129–153
16. Fan Y-L, Liu J-H, Weng W, Chen B-H, Chen Y-N, Wu S-X (2021) Multi-label feature selection with local discriminant model and label correlations. Neurocomputing 442:98–115
17. Liang M-S, Mi J-S, Feng T (2019) Optimal granulation selection for multi-label data based on multi-granulation rough sets. Granul Comput 4(3):323–335
18. Dong H-B, Sun J, Li T, Ding R, Sun X-H (2020) A multi-objective algorithm for multi-label filter feature selection problem. Appl Intell. https://doi.org/10.1007/s10489-020-01785-2
19. Sun L, Wang L-Y, Ding W-P, Qian Y-H, Xu J-C (2020) Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. IEEE Trans Fuzzy Syst 29(1):19–33
20. Sun L, Yin T-Y, Ding W-P, Qian Y-H, Xu J-C (2021) Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ.2021.3053844
21. Ding W-P, Lin C-T, Cao Z-H (2018) Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping PSO with nearest-neighbor memeplexes. IEEE Trans Cybern 49(7):2744–2757
22. Li A-D, Xue B, Zhang M-G (2021) Improved binary particle swarm optimization for feature selection with new initialization and search space reduction strategies. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2021.107302
23. Zhang J, Luo Z-M, Li C-D, Zhou C-G, Li S-Z (2019) Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn 95:136–150
24. Xu J-C, Wang Y, Mu H-Y, Huang F-Z (2018) Feature genes selection based on fuzzy neighborhood conditional entropy. J Intell Fuzzy Syst 36(1):117–126
25. Lin Y-J, Hu Q-H, Liu J-H, Chen J-K, Duan J (2016) Multi-label feature selection based on neighborhood mutual information. Appl Soft Comput 38:244–256
26. Wang C-Z, Huang Y, Shao M-W, Hu Q-H, Chen D-G (2019) Feature selection based on neighborhood self-information. IEEE Trans Cybern 50(9):4031–4042
27. Sha Z-C, Liu Z-M, Ma C, Chen J (2021) Feature selection for multi-label classification by maximizing full-dimensional conditional mutual information. Appl Intell 51(1):326–340
28. Qian W-B, Huang J-T, Wang Y-L, Shu W-H (2020) Mutual information-based label distribution feature selection for multi-label learning. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2020.105684
29. Li L, Liu H-W, Ma Z-J, Mo Y-C, Duan Z-J, Zhou J-Q, Zhao J-M (2014) Multi-label feature selection via information gain, vol 8933. Springer International Publishing, pp 345–355
30. Gao W-F, Hu J-C, Li Y-H, Zhang P (2020) Feature redundancy based on interaction information for multi-label feature selection. IEEE Access 8:146050–146064
31. Xu J-C, Yuan M, Ma Y-Y (2021) Feature selection using self-information and entropy-based uncertainty measure for fuzzy neighborhood rough set. Complex Intell Syst. https://doi.org/10.1007/s40747-021-00356-3
32. Qian W-B, Yu S-D, Yang J, Wang Y-L, Zhang J-H (2020) Multi-label feature selection based on information entropy fusion in multi-source decision system. Evol Intell 13(2):255–268
33. Chen H-M, Li T-R, Cai Y, Luo C, Fujita H (2016) Parallel attribute reduction in dominance-based neighborhood rough set. Inf Sci 373:351–368
34. Lin Y-J, Li Y-W, Wang C-X, Chen J-K (2018) Attribute reduction for multi-label learning with fuzzy rough set. Knowl Based Syst 152:51–61
35. Li Y-W, Lin Y-J, Liu J-H, Weng W, Shi Z-K, Wu S-X (2018) Feature selection for multi-label learning based on kernelized fuzzy rough sets. Neurocomputing 318:271–286
36. Sun L, Yin T-Y, Ding W-P, Xu J-C (2019) Hybrid multilabel feature selection using BPSO and neighborhood rough sets for multilabel neighborhood decision systems. IEEE Access 7:175793–175815
37. Wang C-Z, Qi Y-L, Shao M-W, Hu Q-H, Chen D-G, Qian Y-H, Lin Y-J (2017) A fitting model for feature selection with fuzzy rough sets. IEEE Trans Fuzzy Syst 25(4):741–753
38. Che X-Y, Chen D-G, Mi J-S (2021) Label correlation in multi-label classification using local attribute reductions with fuzzy rough sets. Fuzzy Sets Syst. https://doi.org/10.1016/j.fss.2021.03.016
39. Sun L, Yin T-Y, Ding W-P, Qian Y-H, Xu J-C (2020) Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems. Inf Sci 537:401–424
40. Huang M-M, Sun L, Xu J-C, Zhang S-G (2020) Multilabel feature selection using Relief and minimum redundancy maximum relevance based on neighborhood rough sets. IEEE Access 8(99):62011–62031
41. Xie Y-H, Li D-L, Zhang D-Z, Shuang H (2018) An improved multi-label relief feature selection algorithm for unbalanced datasets. Adv Intell Syst Comput 686:141–151
42. Cai Y-P, Yang M, Gao Y, Yin H-J (2015) ReliefF-based multi-label feature selection. Int J Database Theory Appl 8(4):307–318
43. Gao W, Zhou Z-H (2013) On the consistency of multi-label learning. Artif Intell 199:22–44
44. Zhao D-W, Gao Q-W, Lu Y-X, Sun D, Cheng Y-S (2021) Consistency and diversity neural network multi-view multi-label learning. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2021.106841
45. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356
46. Duan J, Hu Q-H, Zhang L-J, Qian Y-H, Li D-Y (2015) Feature selection for multi-label classification based on neighborhood rough sets. Comput Res Dev 52(1):56–65
47. Lin Y-J, Li Y-W, Wang C-X, Chen J-K (2018) Attribute reduction for multi-label learning with fuzzy rough set. Knowl Based Syst 152:51–61

48. Chen P-P, Lin M-L, Liu J-H (2020) Multi-label attribute reduction based on variable precision fuzzy neighborhood rough set. IEEE Access 8:133565–133576
49. Wang C-Z, Shao M-W, He Q, Qian Y-H, Qi Y-L (2016) Feature subset selection based on fuzzy neighborhood rough sets. Knowl Based Syst 111:173–179
50. Shannon C-E (2001) A mathematical theory of communication. ACM Sigmobile Mob Comput Commun Rev 5(1):3–55
51. Sun L, Zhang X-Y, Qian Y-H, Xu J-C, Zhang S-G (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Inf Sci 502:18–41
52. He Z-F, Yang M, Liu H-D (2014) Joint learning of multi-label classification and label correlations. J Soft 25(9):1967–1981
53. Zhang Q-W, Zhang Y, Zhang M-L (2018) Feature-induced labeling information enrichment for multi-label learning. In: Thirty-second AAAI conference on artificial intelligence, Hilton. AAAI 2018, February 2-7, pp 4446–4453 (2018)
54. Huang R, Jiang W-D, Sun G-L (2018) Manifold-based constraint Laplacian Score for multi-label feature selection. Pattern Recogn Lett 112:346–352
55. Zhang M-L, Pena J-M, Robles V (2009) Feature selection for multi-label naive Bayes classification. Inf Sci 179(19):3218–3229
56. Zhang Y, Zhou Z-H (2008) Multi-Label dimensionality reduction via dependence maximization. In: Proceedings of the twenty-third AAAI conference on artificial intelligence, July 13–17, 2008. Chicago, vol 3, pp 1503–1505
57. Lee J, Kim D-W (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recogn Lett 34(3):349–357
58. Doquire G, Verleysen M (2011) Feature selection for multi-label classification problems. In: Eleventh international work-conference on artificial neural networks, June 8–10, vol 6691, no 1, pp 9–16
59. Shima K, Hossein N-P (2019) A label-specific multi-label feature selection algorithm based on Pareto dominance concept. Pattern Recogn 88:654–667
60. Chen W-Z, Yan J, Zhang B-Y, Chen Z (2007) Yang Q (2007) Document transformation for multi-label feature selection in text categorization. In: 7th IEEE international conference on data mining, October 28–31, pp 451–456
61. Read J (2008) A pruned problem transformation method for multi-label classification. In: 6th New Zealand computer science research student conference, April 14–18, 2008, pp 143–150
62. Lin Y-J, Hu Q-H, Liu J-H, Li J-J, Wu X-D (2017) Streaming feature selection for multi-label learning based on fuzzy mutual information. IEEE Trans Fuzzy Syst 25(6):1491–1507
63. Reyes O, Morell C, Ventura S (2015) Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. Neurocomputing 161:168–182
64. Chen L-L, Chen D-G (2019) Alignment based feature selection for multi-label learning. Neural Process Lett 50(3):2323–2344
65. Chang X-J, Nie F-P, Yang Y, Huang H (2014) A convex formation for semi-supervised multi-label feature selection. In: Twenty-eight AAAI conference on artifical intelligence, July 27–31, 2014, Québec City
66. Ma Z-G, Nie F-P, Yang Y, Uijlings J, Sebe N (2012) Web image annotation via subspace-sparsity collaborated feature selection. IEEE Trans Multimed 14(4):1021–1030
67. Lim H, Lee J, Kim D-W (2017) Optimization approach for feature selection in multi-label classification. Pattern Recogn Lett 89:25–30
68. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat 11(1):86–92
69. Dunn O-J (1961) Multiple comparisons among means. Publ Am Stat Assoc 56(293):52–64
70. Sun L, Wang L-Y, Qian Y-H, Xu J-C, Zhang S-G (2019) Feature selection using Lebesgue and entropy measures for incomplete neighborhood decision systems. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.104942

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