A Novel Unsupervised Approach to Heterogeneous Feature Selection Based on Fuzzy Mutual Information

Zhong Yuan, Hongmei Chen, *Member, IEEE*, Pengfei Zhang, Jihong Wan and Tianrui Li, *Senior Member, IEEE*

***Abstract*—Aiming at the problem of effectively selecting relevant features from heterogeneous data without decision, a novel feature selection approach is studied based on fuzzy mutual information in fuzzy rough set theory. First, the fuzzy relevance of each feature is defined by using fuzzy mutual information, and then the fuzzy conditional relevance is further given. Next, the fuzzy redundancy is defined by using the difference between the fuzzy relevance and the fuzzy conditional relevance. Thereby the evaluation index of the feature importance is obtained by using the idea of unsupervised minimum redundancy and maximum relevance. Finally, a fuzzy mutual information-based unsupervised feature selection (FMIUFS) algorithm is designed to select feature sequences. Extensive experiments are conducted on public datasets and six unsupervised feature selection algorithms are compared. The selected features are evaluated by classification, clustering, and outlier detection methods. Experimental results show that the proposed algorithm can select fewer heterogeneous features to maintain or improve the performance of learning algorithms.**

***Index Terms*—Fuzzy rough set theory, unsupervised feature selection, fuzzy mutual information, minimum redundancy, maximum relevance, heterogeneous data**

I. Introduction

THE goal of unsupervised feature selection (also called attribute reduction) is to find a minimum feature subset in the unlabeled high-dimensional feature space, so that the probability distribution of the data is as close as possible to the original distribution obtained by using all the features [1]–[3]. To the best of our knowledge, most of the existing unsupervised feature selection is only applicable to a single feature type (numerical or nominal). However, in practical applications, data usually exists in the form of heterogeneous (mixed or hybrid) attributes. For example, in a hospital dataset, nominal (categorical) (such as gender, color, and ethnicity) and numerical (weight, height, and temperature) features usually exist simultaneously. Generally, there are two types of methods to deal with heterogeneous data [4], [5]. One is to convert heterogeneous data into homogeneous data. It mainly includes two conversion methods: 1) Using discretization algorithm to convert corresponding numerical features into nominal features. But discretization may lead to the loss of the neighborhood information and ordered structure of real-valued features in real space. 2) Replacing the nominal features with different integers. However, it may be meaningless to use substitute value to measure nominal features. Another type of methods uses different standards to process different types of features, and then integrates different processing results [5].

As a mathematical framework that can handle uncertainty data, the rough set model has been successfully applied in many fields [6]–[10]. However, this model builds a learning model based on equivalence relations and is only suitable for nominal attributes. In response to the above-mentioned problems, Dubois and Prade proposed a Fuzzy Rough Set (FRS) model [11], [12]. It provides an effective tool that can overcome the discretization problem, and can be directly applied to feature selection of numerical or mixed attributes. In FRS model, a fuzzy similarity relation is defined to measure the similarity between data objects. Numerical feature values no longer need to be discretized. Currently, the research on FRS theory mainly involves the model extension of FRS and its application. On the one hand, inspired by the original FRS model, a series of FRS extensions were proposed [13]–[15]. On the other hand, fuzzy rough sets have been successfully applied in many fields, such as feature selection [4], [16]–[18], rule extraction [7], [19], classification tree induction [20], and medicine analysis [21].

Recently, the feature selection methods based on FRSs have received widespread attention. Generally, the feature selection methods using fuzzy rough sets can be divided into fuzzy dependency-based [22], fuzzy discernibility matrix-based [23], and fuzzy uncertainty measure-based [24] methods. For example, Jensen and Shen first extended the dependency function in classical rough sets to fuzzy rough sets, and proposed a related feature selection algorithm [22]. In [25]–[27], the selection algorithms based on FRS are further studied, and some heuristic algorithms are developed to select the optimal feature. Further, Hu et al. used the dependency function to design a hybrid attribute reduction algorithm [28]. Aiming at the shortcoming that the classical FRS model can only maintain the maximum dependency, Wang et al. introduced a new FRS model for feature selection [29]. Furthermore, Wang et al. again introduced distance metric to fuzzy rough sets, and then proposed a new method of attribute reduction

This work was supported by the National Natural Science Foundation of China (61976182, 62076171, 61773324, and 61602327), the Key Techniques of Integrated Operation and Maintenance for Urban Rail Train Dispatching Control System based on Artificial Intelligence (2019YFH0097), the Applied Basic Research Programs of Science and Technology Department of Sichuan Province (2019YJ0084), and Sichuan Key R&D project (2020YFG0035). (Corresponding author: Hongmei Chen)

Z. Yuan, H. M. Chen, P. F. Zhang, J. H. Wan, and T. R. Li, are with the School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China (e-mail: yuanzhong2799@foxmail.com; hmchen@swjtu.edu.cn; feifeihappy55@163.com; jhwan@my.swjtu.edu.cn; trli@swjtu.edu.cn).

based on dependency [30]. In [31], a new type of fuzzy similarity relation is used to control the similarity between objects, and then a novel dependency-based attribute reduction is proposed for dealing with categorical data. In addition, the fuzzy discernibility matrix-based methods have also been studied. Chen et al. introduced a method based on classical discernibility matrix to fuzzy rough sets, and proposed the corresponding attribute reduction algorithm [23], [32], [33]. Dai et al. proposed the concept of reducing maximum discernibility pairs in FRS, and then proposed two attribute selection algorithms based on fuzzy discernibility matrix [34].

Information entropy originated from information theory and can establish an effective competition mechanism to measure uncertainty. Information entropy is introduced into FRS for uncertainty measurement and representation, resulting in different forms [24], [28], [35]–[37]. For example, regarding the importance of fuzzy relations, Yager first introduced the concept of information entropy into fuzzy similarity relations, and then discussed the uncertainty measure of fuzzy information [35]. Mi et al. gave a new fuzzy entropy and applied it to the FRS based on partition for the first time [36]. What's more, Hu et al. studied fuzzy information entropy to represent the uncertainty in the fuzzy approximation space and applied it to supervised feature selection [24], [28]. An et al. combined fuzzy mutual information with the idea of minimum Redundancy and Maximum Relevance (mRMR) and proposed heterogeneous feature selection based on fuzzy mutual information [38], [39]. Based on the concept of information gain ratio in decision tree theory, Dai et al. proposed an attribute selection method based on fuzzy gain ratio under the framework of FRS theory by using fuzzy mutual information [40]. Taking into account the deficiencies of [24], Zhang et al. introduced a new fuzzy conditional entropy algorithm for feature selection in FRSs [5]. Lin et al. presented fuzzy mutual information in multi-label learning to evaluate the quality of features [41]. Recently, Dai et al. proposed a feature selection strategy based on fuzzy conditional mutual information via normative fuzzy information weight [42].

However, it is worth noting that the above feature selection methods are all supervised. For the supervised method, the decision information of an object must be known in advance. However, there are many data without decision information in life. In fact, obtaining decision information is sometimes very expensive. As far as we know, there is only a small amount of work using FRS theory for unsupervised feature selection [24], [43], [44]. For example, Hu et al. first used fuzzy information entropy to define the attribute significance for unsupervised feature selection [24]. Ganivada et al. studied an unsupervised feature selection based on granular neural network using a FRS model [43]. In [44], Mac Parthaláin et al. introduced some unsupervised feature selection methods based on FRSs.

Based on the above analyses, this paper proposes an unsupervised heterogeneous feature selection method based on fuzzy mutual information. First, the fuzzy mutual information is employed to define the fuzzy relevance of each feature, and thus the feature with the largest fuzzy relevance is selected. Then, the fuzzy conditional relevance is defined to characterize the relevance of a feature when a certain feature is known, and thus fuzzy redundancy is defined to indicate the redundancy of a candidate feature. Furthermore, the evaluation index of feature importance is constructed for the selection of subsequent feature by using the idea of unsupervised minimum redundancy and maximum relevance. Finally, a Fuzzy Mutual Information-based Unsupervised Feature Selection (FMIUFS) algorithm is designed to obtain an ordered feature sequence. twenty-four real datasets are used to compare and analyze algorithm FMIUFS with some existing algorithms. Experimental results show that the algorithm can select fewer features to maintain or improve performances of classification, clustering, and outlier detection for numerical, categorical, and heterogeneous data.

The rest of this paper is organized as follows. In the second section, we introduce the preliminary knowledge on FRS theory. In the third section, we propose unsupervised hybrid feature selection based on fuzzy mutual information and design the corresponding algorithm. Experimental results are given in the fourth section. Finally, the fifth section summarizes the full text.

II. Preliminaries

This section introduces some related knowledge about fuzzy rough set theory through [38], [45].

A. Fuzzy relation

Let $FIS = \langle U, C\rangle$ be a Fuzzy Information System (FIS), where U is a set of non-empty finite objects; C is a set of non-empty finite conditional features.

A fuzzy relation $\mathcal{R}$ on $U \times U$ is defined as $\mathcal{R} : U \times U \to [0, 1]$. For $\forall(x, y) \in U \times U$, the membership degree $\mathcal{R}(x, y)$ indicates the degree to which x and y have a relationship $\mathcal{R}$. The set of all fuzzy relations on $U \times U$ is denoted as $\mathcal{F}(U \times U)$. Obviously, the fuzzy relation is a special kind of fuzzy sets.

Suppose $\mathcal{R} \in \mathcal{F}(U \times U)$, for $\forall x, y \in U$, if it meets the following conditions. 1) Reflexivity: $\mathcal{R}(x, x) = 1$; 2) Symmetry: $\mathcal{R}(x, y) = \mathcal{R}(y, x)$; 3) Transitivity: $\mathcal{R}(x, z) \geq \sup \min_{y\in U}\{\mathcal{R}(x, y), \mathcal{R}(y, z)\}$, then $\mathcal{R}$ is called a fuzzy equivalence relation on U. Besides, if $\mathcal{R}$ only satisfies 1) and 2), then $\mathcal{R}$ is called a fuzzy similarity relation on U. For $\mathcal{R}_1, \mathcal{R}_2 \in \mathcal{F}(U)$, we have 1) $\mathcal{R}_1(x, y) \leq \mathcal{R}_2(x, y) \Rightarrow \mathcal{R}_1 \subseteq \mathcal{R}_2$; 2) $(\mathcal{R}_1 \cap \mathcal{R}_2)(x, y) = \min\{\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)\}$; 3) $(\mathcal{R}_1 \cup \mathcal{R}_2)(x, y) = \max\{\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)\}$.

B. Fuzzy rough set

FRS model was first proposed by Dubois and Prade [11], [12], which is defined as follows.

Definition 1: Let $\mathcal{R}$ be a fuzzy equivalence relation on U. For $\forall\mathcal{X} \in \mathcal{F}(U)$, the lower approximation $\underline{\mathcal{R}}\mathcal{X}$ and upper approximation $\overline{\mathcal{R}}\mathcal{X}$ of $\mathcal{X}$ are a pair of fuzzy sets on U whose membership functions respectively are

$$\underline{\mathcal{R}}\mathcal{X}(x) = \inf_{y\in U} \max\{1 - \mathcal{R}(x, y), \mathcal{X}(y)\}, \quad (1)$$

$$\overline{\mathcal{R}}\mathcal{X}(x) = \sup_{y\in U} \min\{\mathcal{R}(x, y), \mathcal{X}(y)\}. \quad (2)$$

C. Fuzzy information measures

Let $U = \{x_1, x_2, \ldots, x_n\}$. For $\forall B \subseteq C = \{c_1, c_2, \ldots, c_m\}$, B can induce a fuzzy similarity relation $\mathcal{R}_B$ on U. It can be represented by fuzzy relation matrix $M(\mathcal{R}_B) = (r_{ij}^B)_{n\times n}$, where $r_{ij}^B = \mathcal{R}_B(x_i, x_j)$, each row $(r_{i1}^B, r_{i2}^B, \ldots, r_{in}^B)$ represents a fuzzy set.

The fuzzy set induced by $\mathcal{R}_B$ is defined as

$$[x_i]_{\mathcal{R}_B} = \frac{r_{i1}^B}{x_1} + \frac{r_{i2}^B}{x_2} + \cdots + \frac{r_{in}^B}{x_n} = (r_{i1}^B, r_{i2}^B, \ldots, r_{in}^B). \quad (3)$$

Without causing confusion, $\mathcal{R}_B$ can be replaced with B.

Let $B = \{c_{k_1}, c_{k_2}, \cdots, c_{k_h}\}(1 \leq h \leq m)$. Obviously, $[x_i]_B$ is a fuzzy set on B. We have $[x_i]_B(x_j) = \mathcal{R}_B(x_i, x_j) = r_{ij}^B$. If $\mathcal{R}_B(x_i, x_j) = 1$, it means that x_j must belong to $[x_i]_B$; if $\mathcal{R}_B(x_i, x_j) = 0$, it means that x_j definitely does not belong to $[x_i]_B$. For the determination of $[x_i]_B$, there are several commonly used methods: 1) Intersection method [24], 2) Distance method [30], and 3) Correlation coefficient method [46]. In this paper, the intersection method is used, which is calculated as $[x_i]_B = \bigcap_{l=1}^{h} [x_i]_{c_{k_l}}$. The cardinality of $[x_i]_B$ is defined as $|[x_i]_B| = \sum_{j=1}^{n} r_{ij}^B = \sum_{j=1}^{n} \mathcal{R}_B(x_i, x_j)$. Obviously, $1 \leq |[x_i]_B| \leq n$.

Hu et al. studied fuzzy information entropy in FRS theory and applied it to feature selection [39], [45]. Let $B, E \subseteq C$, some fuzzy information measures are defined as follows.

Definition 2: [45] The fuzzy information entropy of B is defined as

$$FE(B) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{|[x_i]_B|}{|U|}. \quad (4)$$

If $\mathcal{R}_B$ degenerates into a classical equivalence relation, then the above definition is the same as the classical information entropy.

Property 1: [46] if $B_1 \subseteq B_2 \subseteq C$, then $FE(B_1) \leq FE(B_2)$.

The Property 1 reflects that the fuzzy information entropy changes monotonously with features. It can be seen that the increase or decrease of features in FIS makes the fuzzy information entropy and uncertainty change.

Definition 3: [45] The fuzzy joint entropy of B and E is defined as

$$FE(B, E) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{|[x_i]_B \cap [x_i]_E|}{|U|}. \quad (5)$$

After a feature subset B is known, the uncertainty of the feature subset E can be measured by fuzzy conditional entropy.

Definition 4: [45] The fuzzy conditional entropy of B on E is defined as

$$FE(B|E) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{|[x_i]_B \cap [x_i]_E|}{|[x_i]_E|}. \quad (6)$$

Property 2: [45] $FE(B|E) = FE(B, E) - FE(E)$.

Definition 5: [39] The fuzzy mutual information of B and E is defined as

$$FMI(B; E) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{|[x_i]_B| \times |[x_i]_E|}{|U| \times |[x_i]_B \cap [x_i]_E|}. \quad (7)$$

Property 3: [39] Some corresponding properties are as follows.

1) $FMI(B; E) = FMI(E; B)$;
2) $FMI(B; E) = FE(B) - FE(B|E)$;
3) $FMI(B; E) = FE(B) + FE(E) - FE(B, E)$.

Fuzzy mutual information can be regarded as the variation of the uncertainty of the feature subset B after the feature subset E is known, that is, the amount of information contained in both.

Through the above Definitions 2-5, we find that fuzzy mutual information overcomes the limitations of classical mutual information when processing heterogeneous data. Therefore, the fuzzy information entropy can be used to construct heterogeneous feature selection method.

D. The mRMR method based on fuzzy mutual information

The mRMR method was first proposed by Peng et al. [47], which is considered to be an effective feature selection method. An et al. combined fuzzy mutual information with the idea of mRMR method and proposed heterogeneous feature selection based on fuzzy mutual information [38], [39]. In their method, given the set S_{k-1} with $k-1$ features selected, the kth feature can be determined by

$$\max_{c \in C - S_{k-1}} \{FMI(c; d) - \frac{1}{k-1}\sum_{c' \in S_{k-1}} FMI(c; c')\}. \quad (8)$$

In the above formula, the first term represents the relevance between the candidate feature c and the decision feature d, which ensures the feature that has the greatest relevance with the decision feature d to be chosen. Since there may be redundancy when only the maximum relevance condition is used, the second term further considers the redundancy between the candidate feature c and the selected feature subset S_{k-1}. Because when two features are highly correlated, removing one of them may have little effect on the distinguishing ability of a FIS.

However, the decision information is unknown in unsupervised feature selection. For this reason, Xu et al. proposed an unsupervised feature selection method based on mutual information [48]. The proposed algorithm can handle both non-numerical and numerical data. For non-numerical data, the mutual information can be calculated directly, but for numerical data, the mutual information is difficult to be calculated because it is difficult to estimate the probability distribution density of a numerical variable from a limited number of objects [39], [48]. In this case, the density estimation method can be used to approximate the mutual information. But if we want to get an accurate estimate, a lot of data is required, which results in a lot of time and storage consumption. In addition, data discretization can also be used as a preprocessing step. But discretization may cause the loss of the neighborhood information and ordered structure of real-valued features in

real space. In response to the above problems, this paper proposes an unsupervised feature selection based on fuzzy mutual information for heterogeneous data in FRS theory.

III. Unsupervised heterogeneous feature selection based on fuzzy mutual information

In this section, we first construct a FIS, then propose an unsupervised heterogeneous feature selection strategy based on fuzzy mutual information, and finally design the corresponding algorithm.

A. Fuzzy information system

In order to obtain accurate data processing results, the min-max normalization is employed to normalize the original numerical attribute value in this paper. After normalization, the range of these attribute values is [0,1].

Given a $FIS = \langle U, C\rangle$, and $U = \{x_1, x_2, \ldots, x_n\}$, $C = \{c_1, c_2, \ldots, c_m\}$. In order to effectively process categorical, numerical, and mixed data, the fuzzy similarity degree $r_{ij}^{c_k}$ between x_i and x_j on c_k is calculated as [49], [50]

$$r_{ij}^{c_k} = \begin{cases} 1, c_k(x_i) = c_k(x_j) \text{ and } c_k \text{ is categorical;} \\ 0, c_k(x_i) \neq c_k(x_j) \text{ and } c_k \text{ is categorical;} \\ 1 - |c_k(x_i) - c_k(x_j)|, |c_k(x_i) - c_k(x_j)| \leq \varepsilon_{c_k} \\ \text{ and } c_k \text{ is numerical;} \\ 0, |c_k(x_i) - c_k(x_j)| > \varepsilon_{c_k} \text{ and } c_k \text{ is numerical,} \end{cases} \tag{9}$$

where $c_k(x)$ denotes the value of x under c_k.

It can be seen from Formula 9 that when the absolute value of the difference between $c_k(x_i)$ and $c_k(x_j)$ is greater than a certain threshold, we consider the similarity between x_i and x_j to be 0. This processing method reflects people's tolerance to data noise in numerical features. Among them, ε_{c_k} is the adaptive fuzzy radius, which is calculated by $\varepsilon_{c_k} = \frac{std(c_k)}{\lambda}$, where $std(c_k)$ is the standard deviation of the attribute value of c_k. The preset parameter λ is used to adjust the fuzzy radius.

Through the above normalization, fuzzy similarity degree and adaptive fuzzy radius, a specific FIS can be obtained. It lays the foundation for the establishment of heterogenous feature selection model.

B. Feature selection strategy

From the perspective of information theory, the goal of feature selection is to find a feature subset that contains most or all of the information in the original feature set. The selected feature subset can minimize the uncertainty of other unselected features. Therefore, in unsupervised learning, the feature selection algorithm should select those features that have the greatest mutual information with other features. In order to characterize the relevance between features, based on fuzzy mutual information, we first define the fuzzy relevance of features.

Definition 6: (**Fuzzy relevance**) The fuzzy relevance of a feature c_k is defined as

$$FRel(c_k) = \frac{1}{m}\sum_{s=1}^{m} FMI(c_k; c_s). \tag{10}$$

Property 4: Obviously, the following equation holds.

$$FRel(c_k) = \frac{1}{m}(FE(c_k) + \sum_{s=1, s\neq k}^{m} FMI(c_k; c_s)). \tag{11}$$

In Property 4, $FE(c_k)$ represents the fuzzy information contained in the feature c_k. The larger the value of $FE(c_k)$, the more information the feature c_k can provide to a learning algorithm. $\sum_{s=1, s\neq k}^{m} FMI(c_k; c_s)$ represents the reduction in the information contained in other features after the information of the known feature c_k, that is, the amount of information provided by the feature c_k and other features. The larger the value of $\sum_{s=1, s\neq k}^{m} FMI(c_k; c_s)$, the less information other features can provide to the leaning algorithm. If the feature with the largest fuzzy relevance is selected, the lose in information is minimized. Therefore, the feature selection strategy in this paper starts with an empty set S, then selects one feature in each iteration. In Step 1, the feature c_{ℓ_1} with largest fuzzy relevance is selected firstly, which satisfies the following conditions:

$$c_{\ell_1} = arg\max_{c\in C}\{FRel(c)\}. \tag{12}$$

Since the feature c_{ℓ_1} has the largest fuzzy relevance, it can minimize the uncertainty of other features in the feature set. In other words, when only one feature is selected, c_{ℓ_1} provides the most information for the FIS. At this time, let $S_1 = \langle c_{\ell_1}\rangle$.

Example 1: A fuzzy information system $FIS = \langle U, C\rangle$ is on the left of Table I, which mainly involves heterogenous feature data. Herein, $U = \{x_1, x_2, \cdots, x_6\}$, $C = \{c_1, c_2, c_3, c_4\}$. Among them, c_1, c_2 are nominal features, and c_3, c_4 are numerical features.

TABLE I: Initial and normalization IS

U	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
x_1	b	D	8	0.6	b	D	0.8750	0.6250
x_2	c	A	4	0.6	c	A	0.3750	0.6250
x_3	c	B	6	0.5	c	B	0.6250	0.5000
x_4	a	B	3	0.3	a	B	0.2500	0.2500
x_5	a	C	1	0.1	a	C	0	0
x_6	c	A	9	0.9	c	A	1	1

First, the original numerical data is normalized by the min-max normalization, and the normalized result is shown on the right side of Table I. The standard deviations of numeric features c_3 and c_4 are $std(c_3) \approx 0.3492, std(c_4) \approx 0.3146$. Let $\lambda = 1$, the fuzzy radii are calculated as $\varepsilon_{c_3} \approx 0.3492, \varepsilon_{c_4} \approx 3146$.

By Definition 6, the fuzzy relevance of feature c_k is calculated as follows.
$FRel(c_1) \approx 1.0617$; $FRel(c_2) \approx 1.1959$; $FRel(c_3) \approx 0.9679$; $FRel(c_4) \approx 1.0721$.

Next, the feature with the largest fuzzy relevance is selected and added to S, that is, the feature c_2 is added to S. Then, we have $S_1 = \langle c_2\rangle$.

Suppose S_u denotes the currently unselected features, $S_{r-1} = \langle c_{\ell_1}, c_{\ell_2}, \ldots, c_{\ell_{r-1}}\rangle$ denotes the selected features. The selection of the rth feature adopts the following strategy:

Compared with other features in S_u, c_{ℓ_r} should be as close as possible to the entire feature set. It is relevant, and it should be minimally redundant with the selected features in S_{r-1}. To this end, we further define the fuzzy redundancy of features.

Given $c \in S_u$, for $\forall c_{\ell_s} \in S_{r-1}$, we first assume that the fuzzy relevance of feature c_{ℓ_s} is proportional to the value of its fuzzy entropy $FE(c_{\ell_s})$. In addition, the conditional fuzzy entropy of c_{ℓ_s} relative to c is $FE(c_{\ell_s}|c)$. Obviously, if c is selected to join S_{r-1}, the amount of information provided by the feature c_{ℓ_s} will change due to the addition of c, so its fuzzy relevance should also change. Based on the above analysis, the concept of fuzzy conditional relevance is given below.

Definition 7: (**Fuzzy conditional relevance**) The fuzzy conditional relevance of feature c relative to c_{ℓ_s} is defined as

$$FRel(c_{\ell_s}|c) = \frac{FE(c_{\ell_s}|c)}{FE(c_{\ell_s})} FRel(c_{\ell_s}). \tag{13}$$

Further, the difference between the two can be defined as fuzzy redundancy.

Definition 8: (**Fuzzy redundancy**) The fuzzy redundancy between feature c and c_{ℓ_s} is defined as

$$FRed(c, c_{\ell_s}) = FRel(c_{\ell_s}) - FRel(c_{\ell_s}|c). \tag{14}$$

Therefore, when selecting the rth feature, this paper comprehensively considers the fuzzy relevance of candidate features and its fuzzy redundancy, and obtains the unsupervised fuzzy minimum redundancy and maximum redundancy (UFmRMR) evaluation index of feature importance.

$$c_{\ell_r} = arg \max_{c \in S_u} \{FRel(c) - \frac{1}{|S_{r-1}|} \sum_{s=1}^{|S_{r-1}|} FRed(c, c_{\ell_s})\}. \tag{15}$$

Then the rth feature can be selected as c_{ℓ_r}, because it can reduce the uncertainty of other features to the greatest extent and contains only very little fuzzy redundant information. When selecting subsequent features, we use similar strategies to select them one by one. Finally, an ordered feature sequence is obtained.

Example 2: The continued of Example 1. Given $c \in S_u = \{c_1, c_3, c_4\}$, according to the Definition 7, the fuzzy conditional relevance of c relative to c_2 is calculated as follows.
$FRel(c_2|c_1) \approx 0.4941, FRel(c_2|c_3) \approx 0.7495, FE(c_2|c_4) \approx 0.5579$.

According to Definition 8, the fuzzy redundancy between c and c_2 is calculated as follows.
$FRed(c_1, c_2) = 0.7019, FRed(c_3, c_2) = 0.4464,$
$FRed(c_4, c_2) = 0.6380$.

From this, we have
$FRel(c_1) - FRed(c_1, c_2) = 0.3598;$
$FRel(c_3) - FRed(c_3, c_2) = 0.5214;$
$FRel(c_4) - FRed(c_4, c_2) = 0.4341.$

Next, the feature c_3 is selected and added to S_1, that is, $S_2 = \langle c_2, c_3 \rangle, S_u = \{c_1, c_4\}$.

We use a similar strategy to select features one by one. Finally, an ordered feature sequence $S = \langle c_2, c_3, c_1, c_4 \rangle$ is obtained.

Algorithm 1: FMIUFS algorithm

Input: $IS = \langle U, C \rangle$, threshold value λ, $|C| = m$
Output: An ordered feature sequence S

$S \leftarrow \emptyset, S_u \leftarrow C$;
for $k \leftarrow 1$ *to* m **do**
  Calculate the fuzzy relation matrix $M_{\mathcal{R}_{c_k}}$;
  Calculate the fuzzy entropy $FE(c_k)$;
end
for $k \leftarrow 1$ *to* m **do**
  for $s \leftarrow 1$ *to* m **do**
    Calculate the fuzzy joint entropy $FE(c_k, c_s)$;
    Calculate the fuzzy mutual information $FMI(c_k; c_s)$;
  end
end
for $k \leftarrow 1$ *to* m **do**
  Calculate the fuzzy relevance $FRel(c_k)$;
end
Select feature c_{ℓ_1} so that $FRel(c_{\ell_1})$ has the maximum value;
$S \leftarrow S \cup \{c_{\ell_1}\}, S_u \leftarrow S_u - \{c_{\ell_1}\}$;
while $|S_u| \neq 0$ **do**
  for $l \leftarrow 1$ *to* $|S_u|$ **do**
    for $s \leftarrow 1$ *to* $|S|$ **do**
      Calculate the fuzzy redundancy $FRed(c_l, c_{\ell_s})$;
    end
  end
  Select feature c_{ℓ_r} so that $FRel(c_{\ell_r}) - \frac{1}{|S|} \sum_{s=1}^{|S|} FRed(c_{\ell_r}, c_{\ell_s})$ has the maximum value;
  $S \leftarrow S \cup \{c_{\ell_r}\}, S_u \leftarrow S_u - \{c_{\ell_r}\}$;
end
return S.

C. Feature selection algorithm

In this selection, we design a FMIUFS algorithm and analyze the time complexity.

Algorithm 1 starts from an empty set. First, the fuzzy relevance of each feature is calculated, and then the feature with the largest fuzzy relevance is selected and added to the feature subset S. Then, the UFmRMR evaluation index is used to select subsequent features until an ordered feature sequence is obtained.

In Algorithm 1, the number of loop for Steps 2-5 is m, the number of loop for Step 3 is $n \times n$, the number of loop for Step 4 is n, the number of the loop for Steps 6-11 is $m \times m$, the number of the loop for Step 8 is n, the number of the loop for Steps 12-14 is m, and the number of loop in Steps 18-22 is $|S_u| \times |S|$. Therefore, the total number of loop of Algorithm 1 is $m \times (n \times n + n) + m \times m \times n + m \times m + |S_u| \times |S|$. Therefore, in the worst case, the time complexity of Algorithm 1 is $O(mn(m + n))$.

IV. Experiments

This section evaluates algorithm FMIUFS through the performance of classification, clustering, and outlier detection. To this end, twenty-four datasets (including numerical, categorical, and mixed attributes) are selected from the UCI database for experiments [51]. The maximum probability value method is adopted to fill in the missing values. The description of eighteen datasets used for classification and clustering is shown in Table II. Besides, the selected datasets are often used to evaluate classification and clustering algorithms. Therefore, the downsampling method proposed in [52] is adopted to

TABLE II: The description of datasets for classification and clustering

No	Datasets	Number of objects	Number of conditional features		Decision classes
			Numerical	Nominal	
1	Ecoli	336	7	0	8
2	Glass	214	10	0	6
3	Iris	150	4	0	3
4	Wbc	699	9	0	2
5	Wdbc	569	31	0	2
6	Abalone	4177	7	1	28
7	Autos	205	15	10	6
8	Credit	690	6	9	2
9	Heart	270	6	7	2
10	Horse	368	7	20	2
11	Labor	57	8	8	2
12	Sick	3772	6	23	2
13	Chess	3196	0	36	2
14	Lymphography	148	0	18	4
15	Mushroom	8124	0	22	2
16	Soybean-large	683	0	35	19
17	Vote	435	0	16	2
18	Zoo	101	0	16	7

TABLE III: The description of datasets for outlier detection

No	Datasets	Abbreviation	Number of conditional features		Number of objects	Number of outliers
			Numerical	Nominal		
1	Credit	Cre	6	9	425	42
2	German	Ger	7	13	714	14
3	Heart	Hea	6	7	166	16
4	Hepatitis	Hep	6	13	94	9
5	Diabetes	Dia	8	0	526	26
6	Ionosphere	Ion	34	0	249	24
7	Iris	Iri	4	0	111	11
8	Pima	Pim	9	0	555	55
9	Sonar	Son	60	0	107	10
10	Wbc	Wb	9	0	483	39
11	Wdbc	Wdb	31	0	396	39
12	Yeast	Yea	8	0	1141	5
13	Lymphography	Lym	0	8	148	6
14	Mushroom	Mus	0	22	4429	221

obtain some effective datasets for detecting outliers. Following the preprocessing method in [50], the description of the final fourteen datasets for outlier detection[1] is shown in Table III. Among them, the decision classes of all datasets are removed during feature selection.

On these datasets listed in Tables II and III, we compared algorithm FMIUFS with performances of Feature Similarity-based Feature Selection (FSFS) [53], SPECtral analysis-based (SPEC) [54], Mutual Information-based (MI) [48], Unsupervised Spectral Feature Selection Method (USFSM) [55], Feature Ranking (FR) [56], UnSupervised Quick Reduct (USQR) [57], Unsupervised Fuzzy Rough-based Feature Selection (UFRFS) [44], and Fuzzy Entropy-based Unsupervised Attribute Recuction (FEUAR) [24] algorithms. Among them, algorithms FSFS and SPEC are usually applied to numerical data. Algorithm MI is difficult to process numerical data. Algorithms UFRFS and FEUAR are suitable for heterogeneous data. Algorithm USQR is an attribute reduction method based on classic rough sets, which only apply to nominal data. However, numerical data requires to be discretized. Algorithms UFRFS and FEUAR are methods based on FRS theory, where algorithm UFRFS only considers numerical data, and algorithm FEUAR is suitable for mixed data. Algorithms USFSM and FR are also suitable for mixed attribute data. Algorithm USFSM is inspired by the spectral feature selection, and the kernel function and a new spectral-based feature evaluation measure are used to quantify the correlation of features. Algorithm FR sorts the features by normalizing mutual information.

[1]https://github.com/BElloney/Outlier-detection

A. Experiment preparation

The classification performance of the above comparison algorithms is evaluated by calling classification trees (CT) and naive bayes (NB) algorithms provided by MATLAB R2015b. The 10-fold cross-validation is adopted to implement experiments. The original dataset is randomly divided into 10 subsets, of which 9 subsets are used as training data, and

the remaining a subset is used as test data. We repeat the experiment 10 times, and the average and standard deviation of the classification accuracy are calculated as the final result.

For clustering experiments, the clustering effects of the above comparison algorithms is also evaluated by calling algorithm k-Means in MATLAB R2015b. The parameters of all clustering experiments are kept as initial values. The number of clusters is set to the number of real decision classes. The clustering accuracy (ACC) [58] and adjusted rand index (ARI) [59], [60] are used to evaluate the performance of clustering. The larger the ACC or ARI of a clustering algorithm, the better its performance. Since randomness exits in the results of clustering algorithm, we repeat the experiment 10 times, and the average and standard deviation of the clustering accuracy are recorded as the final result.

In addition, for outlier detection experiments, k-nearest neighbor outlier (kNNO) [61] algorithm is used to test the outlier detection effect of the above comparison algorithms. Follow the strategy in [50], Precision (P) is used to evaluate the effectiveness of outlier detection. The larger the value of $P(t)$, the better the result of outlier detection. Specifically, the sequence number t is set as the number of outliers. We repeat algorithm kNNO for different feature subsets, and the optimal final detection results of k are calculated in the range of $[1, n/4]$ with step size 1.

In the experiments, for algorithms FSFS and SPEC, all different nominal attribute values are replaced with different integer values, and all attribute values are normalized to the interval $[0, 1]$ using min-max normalization. Following the preferred parameter settings in [53], the feature similarity is calculated by the "Maximal Information Compression Index" in algorithm FSFS. In addition, for algorithm SPEC, its style is set as -1. Generally, algorithms FSFS, SPEC, MI, USFSM, FR, and the proposed algorithm FMIUFS output a sequence with m features in descending order of feature importance for all datasets. In order to compare these algorithms more reasonably, the first $k(k = 1, 2, ..., m-1)$ features are selected iteratively to evaluate the performance. Finally, $(m-1)$ results can be obtained. The best result among the $(m-1)$ results is selected to compare these feature selection algorithms. Since algorithms MI and USQR only applicable to categorical attributes, discretization technology needs to be used to preprocess the numerical attributes. Generally speaking, different discretization methods may result in the selection of different feature subsets. The experimental results in [24] show that the FCM discretization method provides a better performance than equal width and equal frequency. Therefore, the FCM discretization technology is employed to discretize numerical data [62]. The numerical attributes are discretized into four intervals [29]. For algorithm FMIUFS, we calculate the optimal feature subset of λ in the range of [0.1,2] with step size 0.1.

B. Classification results

The numbers of optimal feature subsets based on different classification algorithms are shown in Table IV. The underline "–" indicates that the corresponding feature selection algorithm does not remove any features. Through Tables IV, we have the following findings.

1) Most feature selection algorithms can remove some irrelevant or redundant features in most cases. However, algorithms USQR, UFRFS, and FEUAR obtain the entire conditional feature on some datasets such as Iris, Wbc, and Abalone. It shows that these algorithms cannot effectively remove redundant features on some datasets.
2) In most cases, for the same feature selection algorithm, different classification algorithms may require different feature subsets to obtain optimal performance. For example, for algorithms CT and NB, the numbers of optimal feature subsets of algorithm FMIUFS are different on 17 datasets.
3) For some datasets, the numbers of optimal feature subsets of the same algorithm may be the same under different classification algorithms. For example, for dataset Ecoli, the numbers of optimal feature subsets of algorithm FSFS are 6 under the two classification algorithms.
4) From the average number of selected features, algorithm FMIUFS achieves a smaller value on the two classification algorithms. For example, for algorithm CT, the average feature number of algorithm FMIUFS is 11.9, while those of algorithms FSFS, SPEC, MI, USFSM, and FR are 14.7, 15.4, 14.7, 11.9, and 14.2, respectively.

Tables V and VI give comparisons of the classification accuracies of feature subsets based on algorithms CT and NB, respectively. The optimal classification accuracies of different selection algorithms are in boldface. From Tables V and VI, it can be seen that algorithm FMIUFS achieves better classification accuracy in most cases. The experimental results are discussed as follows.

1) Table V shows that algorithm FMIUFS achieves the best classification accuracy on 11 datasets. However, algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR achieve optimal classification accuracy on only 1, 2, 1, 3, 4, 1, 1, and 1 datasets, respectively.
2) From the results in Table VI, it can be seen that algorithm FMIUFS achieves the best result on 14 datasets. However, algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR achieve the best result on 1, 1, 1, 1, 2, 0, 0, and 0 datasets, respectively.
3) From the average classification accuracy, algorithm FMIUFS achieves the maximum value on two classification algorithms. In addition, from the perspective of standard deviation, the standard deviation of the proposed algorithm FMIUFS is relatively small. It indicates that the results of algorithm FMIUFS are relatively stable.

Both the replacement and discretization of data values may lead to changes in the data structure, which in turn leads to the loss of information. On datasets with numerical attributes, algorithms MI and USQR show relatively poor classification accuracy. This is because the numerical features are discretized before experiments. The classification performance of algorithms FSFS, SPEC, and UFRFS is affected by replacing the nominal attributes with integer values. However, for algorithm

TABLE IV: Number of optimal feature subsets based on different classification algorithms

Dataset	Original features	FSFS		SPEC		MI		USFSM		FR		USQR	UFRFS	FEUAR	FMIUFS	
		CT	NB	CT	NB	CT	NB	CT	NB	CT	NB	All	All	All	CT	NB
Ecoli	7	6	6	6	6	6	6	5	5	6	6	5	7	7	6	5
Glass	10	5	1	5	8	9	1	9	9	9	6	9	8	10	3	1
Iris	4	3	3	1	1	2	2	3	3	1	1	4	4	4	3	1
Wbc	9	5	8	7	8	5	8	3	6	6	7	9	9	9	2	8
Wdbc	31	28	30	27	28	27	12	18	30	28	29	13	9	31	7	30
Abalone	8	2	3	7	7	5	2	1	2	1	2	8	8	8	7	2
Autos	25	23	14	22	21	22	24	14	23	24	23	11	16	22	20	23
Credit	15	10	10	13	14	8	11	12	14	9	14	13	14	14	5	10
Heart	13	11	11	6	12	9	10	12	12	4	4	10	13	13	5	5
Horse	27	24	21	22	9	19	4	2	1	21	10	3	21	6	25	18
Labor	16	9	14	10	15	7	11	15	13	6	8	9	14	14	4	7
Sick	29	28	1	26	1	28	1	25	5	28	1	20	27	27	24	17
Chess	36	35	34	35	35	35	35	33	16	35	35	26	32	33	35	33
Lymphography	18	9	10	12	17	15	16	4	2	12	13	10	16	10	15	14
Mushroom	22	17	12	18	21	20	20	11	3	13	4	14	18	15	6	3
Soybean-large	35	27	26	34	34	29	34	20	25	31	34	12	23	18	27	25
Vote	16	15	9	14	14	3	3	15	12	14	3	15	16	16	15	3
Zoo	16	8	13	13	14	15	11	12	15	7	8	11	13	11	6	10
Average	18.7	14.7	12.6	15.4	14.7	14.7	11.7	11.9	10.9	14.2	11.6	11.2	14.9	14.9	11.9	11.9

TABLE V: The comparison of classification accuracy of reduced data based on CT algorithm(%)

Datasets	Original data	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Ecoli	80.71±0.93	80.06±1.36	77.26±0.73	81.76±1.19	81.79±1.25	81.37±0.89	80.98±1.02	–	–	**81.90±1.09**
Glass	98.79±0.33	98.69±0.48	98.41±0.39	98.79±0.24	97.66±0.58	97.85±0.24	98.60±0.00	98.55±0.27	–	**98.83±0.40**
Iris	94.93±0.78	92.73±1.39	95.60±0.34	95.27±1.11	93.27±1.35	95.40±0.21	–	–	–	**95.73±0.47**
Wbc	93.33±0.69	94.68±0.63	94.28±0.43	94.13±0.42	94.64±0.62	94.08±0.61	–	–	–	**95.58±0.55**
Wdbc	92.20±0.67	92.99±0.85	93.20±0.67	92.99±0.94	92.85±0.87	93.92±0.84	92.67±0.69	93.69±0.65	–	**93.99±0.75**
Abalone	21.08±0.59	23.23±0.37	21.41±0.19	21.18±0.68	**24.84±0.36**	24.09±0.18	–	–	–	21.34±0.45
Autos	72.93±1.56	74.63±2.22	74.78±2.58	76.20±1.73	75.56±2.08	75.66±2.67	69.02±1.51	66.68±0.98	73.56±2.05	**78.78±1.72**
Credit	81.41±0.86	82.00±0.89	82.16±0.93	84.19±0.67	81.23±0.98	83.13±0.95	81.94±0.83	80.54±0.96	81.12±1.07	**85.93±0.61**
Heart	76.41±0.99	76.33±1.42	78.15±1.40	78.59±1.04	76.33±2.26	**84.59±0.47**	74.30±2.67	–	–	83.19±1.09
Horse	77.96±1.51	78.94±1.77	78.61±1.03	80.84±1.90	**81.39±0.75**	79.95±1.15	69.08±1.81	78.10±1.61	78.45±1.56	79.51±1.18
Labor	84.21±2.86	83.68±2.75	**91.05±1.54**	89.12±2.72	79.82±2.77	89.30±1.29	77.37±3.35	81.58±3.33	83.51±1.89	87.89±1.54
Sick	98.85±0.13	98.85±0.15	98.92±0.07	98.48±0.08	98.92±0.13	98.48±0.12	98.84±0.14	98.82±0.09	98.85±0.06	**98.93±0.10**
Chess	99.39±0.07	99.09±0.12	99.35±0.08	99.43±0.07	99.33±0.09	99.33±0.08	98.26±0.10	99.10±0.10	99.11±0.07	**99.45±0.06**
Lymphography	75.27±2.78	79.93±2.07	79.86±1.96	76.76±1.36	78.78±2.49	78.18±2.09	76.49±1.68	77.70±1.74	77.50±3.20	76.49±1.71
Mushroom	100.00±0.00	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**
Soybean-large	90.82±1.04	90.76±0.37	90.73±0.60	90.07±0.51	91.65±0.60	**92.01±0.71**	83.66±1.02	89.63±0.63	85.04±0.90	91.68±0.40
Vote	95.08±0.53	89.40±0.63	89.22±0.80	95.63±0.00	95.59±0.47	**95.70±0.41**	89.43±0.58	–	–	95.43±0.37
Zoo	89.80±0.67	93.17±0.31	90.79±0.67	91.58±0.84	93.66±0.83	92.08±0.00	86.93±1.38	87.33±1.38	87.13±0.93	**96.04±0.00**
Average	84.62±0.94	84.95±0.99	85.21±0.80	85.83±0.86	85.41±1.03	86.39±0.72	–	–	–	**86.71±0.69**

TABLE VI: The comparison of classification accuracy of reduced data based on NB algorithm(%)

Datasets	Original data	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Ecoli	84.49±0.82	83.45±0.61	79.20±0.53	84.85±0.45	84.67±0.77	82.98±0.78	84.67±0.70	–	–	**84.88±0.48**
Glass	90.42±0.97	98.27±0.62	90.93±0.97	97.85±0.80	91.73±0.73	93.69±0.40	92.48±0.84	88.60±0.63	–	**98.41±0.39**
Iris	95.87±0.53	89.27±0.80	**96.00±0.00**	**96.00±0.00**	89.27±1.11	**96.00±0.00**	–	–	–	**96.00±0.00**
Wbc	96.48±0.17	96.54±0.06	95.67±0.17	95.51±0.20	96.29±0.11	95.59±0.11	–	–	–	**96.58±0.05**
Wdbc	93.67±0.32	93.43±0.35	93.84±0.39	94.24±0.14	94.43±0.17	94.02±0.41	93.20±0.25	93.33±0.26	–	**94.53±0.19**
Abalone	25.00±0.15	26.35±0.19	25.03±0.13	26.72±0.12	26.61±0.19	26.49±0.18	–	–	–	**26.73±0.09**
Autos	59.46±1.43	**63.51±1.26**	62.05±1.17	58.10±1.67	57.95±1.49	58.34±1.22	61.51±0.90	57.85±1.13	59.32±1.28	61.76±1.81
Credit	67.08±0.49	81.29±0.36	66.10±0.92	67.07±1.17	66.49±0.64	66.71±0.32	82.57±0.47	67.74±0.61	67.63±0.56	**85.14±0.53**
Heart	79.81±0.53	80.11±0.72	80.22±0.53	81.00±0.63	79.93±0.65	83.00±0.56	76.85±0.73	–	–	**83.26±0.62**
Horse	65.46±1.39	69.29±0.88	67.58±0.54	66.22±0.13	66.03±0.48	66.60±0.43	64.29±0.94	60.65±1.23	60.00±1.09	**70.52±1.13**
Labor	85.44±2.62	86.14±2.80	84.21±3.60	84.91±2.06	85.44±3.51	85.44±2.49	87.72±1.85	88.42±1.23	87.72±2.48	**92.11±1.49**
Sick	93.73±0.02	93.88±0.00	93.88±0.00	93.88±0.00	**96.75±0.03**	93.88±0.00	93.74±0.02	93.73±0.03	93.71±0.05	93.89±0.01
Chess	57.37±0.16	72.01±0.20	57.46±0.22	57.21±0.34	74.50±0.18	72.11±0.57	66.90±0.15	76.19±0.31	74.79±0.41	**76.29±0.15**
Lymphography	80.34±0.67	80.88±1.86	82.09±1.66	81.69±1.61	74.93±2.15	82.50±1.51	79.86±1.77	80.20±2.16	79.26±1.83	80.95±1.62
Mushroom	93.58±0.03	97.65±0.02	93.60±0.03	94.61±0.01	98.52±0.00	98.52±0.00	95.32±0.02	94.19±0.03	95.32±0.02	**98.82±0.00**
Soybean-large	42.61±0.61	46.65±0.25	41.63±0.32	37.19±0.93	46.50±0.47	42.49±0.64	43.66±0.43	46.73±0.32	44.76±0.51	**46.91±0.38**
Vote	89.70±0.15	88.64±0.22	87.95±0.16	94.71±0.00	92.34±0.24	94.71±0.00	87.10±0.17	–	–	**94.74±0.22**
Zoo	83.07±0.56	82.77±1.16	79.31±1.27	85.84±0.48	81.09±1.27	**86.14±0.00**	71.09±1.02	70.89±0.51	71.19±0.73	85.85±0.67
Average	76.87±0.65	79.45±0.69	76.49±0.70	77.64±0.60	77.97±0.79	78.85±0.53	–	–	–	**81.52±0.55**

FMIUFS, neither replacement nor discretization needs to be done, so that more real information about the data can be retained. Therefore, algorithm FMIUFS has relatively better classification results. In addition, although algorithms USFSM, FR, and FEUAR do not require corresponding preprocessing, it does not obtain a better classification result. This may be because it only considers the relevance of features and does not consider the redundancy of features.

To sum up, algorithm FMIUFS can obtain a relatively small feature subset and improve or maintain the classification

TABLE VII: Number of optimal feature subsets based on algorithm k-Means

Dataset	Original features	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Ecoli	7	5	6	6	6	6	5	7	7	6
Glass	10	4	7	5	9	7	9	8	10	7
Iris	4	2	1	2	3	1	4	4	4	1
Wbc	9	8	6	8	5	7	9	9	9	6
Wdbc	31	30	29	28	27	29	13	9	31	4
Abalone	8	1	1	7	4	7	8	8	8	1
Autos	25	18	13	23	16	16	11	16	22	20
Credit	15	12	12	8	12	7	13	14	14	4
Heart	13	11	12	10	8	5	10	13	13	11
Horse	27	3	4	4	2	26	3	21	6	1
Labor	16	13	12	8	5	9	9	14	14	12
Sick	29	23	1	20	1	1	20	27	27	2
Chess	36	14	25	35	4	15	26	32	33	30
Lymphography	18	1	1	3	1	1	10	16	10	9
Mushroom	22	9	19	17	11	17	14	18	15	8
Soybean-large	35	33	34	13	33	24	12	23	18	9
Vote	16	15	14	5	12	5	15	16	16	3
Zoo	16	13	15	8	2	9	11	13	11	10
Average	18.7	11.9	11.8	11.7	8.9	10.7	11.2	14.9	14.9	8.0

TABLE VIII: The comparison of clustering accuracy of reduced data based on k-Means algorithm(%)

Datasets	Original data	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Ecoli	59.23±6.83	58.81±5.48	55.71±5.25	59.46±11.20	58.48±6.01	60.42±6.14	50.86±4.72	–	–	**61.43±3.79**
Glass	60.33±9.18	71.12±4.30	61.54±5.45	70.28±4.06	62.90±10.17	67.76±9.84	58.93±3.57	57.66±6.78	–	**71.92±7.61**
Iris	85.53±9.91	80.00±0.00	**96.00±0.00**	**96.00±0.00**	83.33±0.00	**96.00±0.00**	–	–	–	**96.00±0.00**
Wbc	95.85±0.00	95.85±0.00	95.28±0.00	95.31±0.09	95.61±0.07	95.57±0.00	–	–	–	**96.28±0.00**
Wdbc	92.79±0.00	92.27±0.00	92.62±0.00	92.79±0.00	93.23±0.09	92.79±0.00	92.62±0.00	88.51±9.11	–	**94.02±0.00**
Abalone	15.00±0.79	19.37±0.05	19.37±0.04	14.59±0.63	15.90±0.56	15.08±1.01	–	–	–	**19.37±0.04**
Autos	40.29±2.82	43.41±3.09	41.41±2.21	42.10±2.33	**44.68±2.56**	42.88±3.26	41.85±3.68	40.88±3.40	39.85±3.67	44.54±1.98
Credit	65.96±11.56	80.03±10.16	68.91±17.50	80.33±6.10	**85.51±0.00**	79.36±12.95	78.23±5.60	78.75±5.98	68.01±14.17	**85.51±0.00**
Heart	75.56±8.59	76.22±6.40	75.33±8.47	76.78±6.42	77.96±2.25	78.30±3.04	63.19±7.49	–	–	**80.81±1.80**
Horse	53.53±0.00	60.76±6.45	66.33±0.09	65.14±1.46	**74.10±0.13**	56.14±5.29	65.95±1.98	56.47±5.04	62.04±3.17	66.58±0.00
Labor	84.04±8.49	81.75±11.73	79.65±6.57	83.16±10.57	74.91±5.43	78.60±6.18	70.35±7.37	78.07±11.67	79.30±15.42	**87.02±4.15**
Sick	75.44±10.21	83.48±8.91	**93.90±0.00**	84.53±2.58	93.85±0.00	93.85±0.00	76.25±6.44	73.83±7.12	77.59±9.13	84.07±0.08
Chess	52.57±3.05	**59.96±0.37**	58.63±8.44	54.38±4.24	59.82±7.58	56.16±4.26	52.25±2.43	54.52±4.55	52.25±2.44	56.14±4.39
Lymphography	46.82±3.98	**70.27±0.00**	56.76±0.00	64.59±7.13	61.49±0.00	54.73±0.00	49.39±2.28	49.59±2.74	48.58±2.98	53.58±2.21
Mushroom	72.22±12.18	74.40±0.00	75.13±12.81	76.00±9.29	83.90±0.00	83.52±3.86	71.15±8.99	74.42±14.80	75.61±7.49	**83.90±0.00**
Soybean-large	57.10±3.22	58.20±3.07	58.13±3.49	60.34±3.53	57.55±2.68	60.69±2.25	43.69±2.55	53.72±2.60	46.44±3.96	**62.68±2.68**
Vote	88.05±0.00	85.06±0.00	85.75±0.00	88.51±0.00	87.70±0.20	88.51±0.00	85.06±0.00	–	–	**88.87±10.47**
Zoo	75.84±7.54	74.65±8.47	78.42±9.10	**89.70±3.00**	76.53±2.80	89.21±3.00	61.49±10.83	73.47±6.97	63.37±9.19	81.88±4.45
Average	66.45±5.46	70.31±3.80	69.94±4.41	71.89±4.03	71.53±2.25	71.64±3.39	–	–	–	**73.03±2.43**

TABLE IX: The comparison of ARI of reduced data based on k-Means algorithm(%)

Datasets	Original data	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Ecoli	0.4395±0.0478	0.4797±0.1044	0.4158±0.0740	0.4448±0.0673	0.4365±0.0770	0.4098±0.0284	0.4195±0.0628	–	–	**0.4870±0.0493**
Glass	0.4015±0.1110	0.5327±0.1016	0.4833±0.1325	0.5095±0.1139	0.4182±0.1242	0.4799±0.1839	0.4492±0.1114	0.4318±0.1414	–	**0.5328±0.1677**
Iris	0.6830±0.0895	0.5537±0.0149	0.8857±0.0001	0.8857±0.0000	0.6014±0.0606	**0.8858±0.0000**	–	–	–	**0.8858±0.0000**
Wbc	0.8391±0.0000	0.8389±0.0000	0.8179±0.0000	0.8189±0.0033	0.8301±0.0026	0.8284±0.0000	–	–	–	**0.8551±0.0000**
Wdbc	0.7302±0.0000	0.7122±0.0000	0.7279±0.0031	0.7302±0.0000	0.7460±0.0032	0.7298±0.0000	0.7244±0.0000	0.6855±0.0031	–	**0.7734±0.0000**
Abalone	0.0510±0.0031	0.0386±0.0003	0.0386±0.0003	0.0479±0.0020	**0.0521±0.0025**	0.0461±0.0044	–	–	–	0.0385±0.0003
Autos	0.1531±0.0150	0.1510±0.0128	0.1504±0.0148	0.1531±0.0187	**0.1746±0.0257**	0.1646±0.0175	0.1532±0.0227	0.1582±0.0118	0.1666±0.0136	0.1554±0.0126
Credit	0.2873±0.1857	0.3317±0.2089	0.0742±0.1586	0.3276±0.1268	**0.5036±0.0000**	0.4038±0.2103	0.1788±0.2110	0.2117±0.2027	0.2469±0.1945	**0.5036±0.0000**
Heart	0.2740±0.1342	0.2609±0.1350	0.3329±0.0319	0.2970±0.1139	0.3207±0.0307	0.2830±0.0686	0.1355±0.0729	–	–	**0.3593±0.0518**
Horse	-0.0080±0.0006	0.0020±0.0104	0.0489±0.0182	0.0351±0.0109	**0.1227±0.0839**	-0.0080±0.0006	0.0161±0.0207	-0.0059±0.0005	0.0153±0.0227	0.0591±0.0014
Labor	0.1858±0.1305	0.3049±0.2528	0.2974±0.2639	0.3819±0.1785	0.2920±0.1286	0.2621±0.1544	0.0356±0.0901	0.3026±0.2742	0.3407±0.2568	**0.4412±0.2733**
Sick	-0.0217±0.0332	-0.0253±0.0310	**0.0076±0.0000**	-0.0274±0.0272	-0.0005±0.0000	-0.0005±0.0000	-0.0095±0.0378	-0.0283±0.0389	-0.0064±0.0317	0.0014±0.0726
Chess	0.0178±0.0178	0.0326±0.0116	0.0429±0.0580	0.0083±0.0170	**0.0549±0.0573**	0.0082±0.0123	0.0002±0.0000	0.0064±0.0146	0.0060±0.0133	0.0082±0.0150
Lymphography	0.1523±0.0697	0.2106±0.0023	0.0927±0.0000	0.2073±0.0287	**0.3052±0.0000**	0.0424±0.0015	0.1566±0.0414	0.1213±0.0449	0.1515±0.0360	0.1239±0.0490
Mushroom	0.4146±0.1981	0.2152±0.0719	0.2164±0.2127	0.2247±0.2008	0.4596±0.0000	0.2888±0.2148	0.2822±0.1114	0.3045±0.2232	0.2825±0.1254	**0.4596±0.0000**
Soybean-large	0.4077±0.0386	0.4076±0.0259	0.3942±0.0484	0.3959±0.0150	0.4030±0.0297	0.4175±0.0287	0.2855±0.0282	0.3824±0.0310	0.3115±0.0180	**0.4573±0.0634**
Vote	0.5758±0.0034	0.4904±0.0000	0.5140±0.0034	0.5919±0.0000	0.5676±0.0067	0.5919±0.0000	0.4904±0.0000	–	–	**0.7109±0.0000**
Zoo	0.7407±0.1321	0.6359±0.1246	0.6037±0.1424	**0.8632±0.0815**	0.7285±0.0775	0.8617±0.0639	0.5382±0.0979	0.6264±0.1024	0.5431±0.1245	0.8085±0.0532
Average	0.3513±0.0672	0.3430±0.0616	0.3414±0.0646	0.3831±0.0559	0.3898±0.0395	0.3720±0.0550	–	–	–	**0.4256±0.0450**

accuracy. Therefore, it is suitable for feature selection of classification on mixed attribute datasets.

C. Clustering results

The number of optimal feature subsets for algorithm k-Means is presented in Table VII. The underline "–" indicates that the corresponding feature selection algorithm does not re-

TABLE X: Number of optimal feature subsets based on algorithm kNNO

Dataset	Original features	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Cre	15	13	4	14	4	3	12	12	13	7
Ger	20	18	18	4	7	9	12	19	18	12
Hea	13	11	9	8	11	12	9	13	13	10
Hep	19	18	6	2	10	1	9	15	19	9
Dia	8	4	5	3	6	5	8	8	8	7
Ion	34	9	22	6	10	7	16	8	33	12
Iri	4	3	2	2	3	2	4	4	4	2
Pim	9	7	8	6	8	6	8	8	8	6
Son	60	3	6	36	58	50	7	7	60	13
Wb	9	5	6	8	6	7	9	9	9	5
Wdb	31	30	20	9	27	5	12	9	31	2
Yea	8	5	1	5	1	5	5	8	6	7
Lym	18	17	1	2	10	6	10	16	10	17
Mus	22	17	9	17	7	8	13	18	14	10
Average	19.3	11.4	8.4	8.7	12.0	9.0	9.6	11.0	17.6	8.5

move any features. Through Table VII, we have the following conclusions.

1) Most feature feature algorithms can effectively remove candidate features. Algorithms USQR, UFRFS, and FEUAR do not remove candidate features on some data sets, such as Iris, Wbc, and Abalone.
2) The proposed algorithm FMIUFS obtains the smallest number of selected features on 8 data sets. However, algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR obtains the smallest number of selected features on only 4, 4, 0, 6, 4, 2, 0, and 0 datasets, respectively. It shows that algorithm FMIUFS can obtain better clustering results with a smaller number of selected features, such as datasets Iris, wdbc and Abalone.
3) From the average number of selected features, the average number of selected features of algorithm FMIUFS is 8.0, which is smaller than those of other algorithms.

Tables VIII and IX respectively give comparisons of the optimal clustering ACC and ARI based on algorithm k-Means, and the best clustering ACCs and ARIs of different feature selection algorithms are in boldface. Through Tables VIII and IX, the corresponding analyses are as follows.

1) From Tables VIII and IX, it can be seen that algorithm FMIUFS can improve or maintain the clustering ACC and ARI of the original data on all datasets.
2) Table VIII shows that algorithm FMIUFS achieves the best clustering ACC on 12 datasets. However, for algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR, only 2, 2, 2, 3, 2, 0, 0, and 0 datasets achieve the best ACC.
3) Through Table IX, we can see that algorithm FMIUFS achieves the best clustering ARI on 11 datasets. However, algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR achieve the best clustering ARI on 0, 1, 1, 6, 1, 0, 0, and 0 datasets, respectively.
4) From Tables VIII and IX, it can be seen that algorithm FMIUFS achieves the maximum average values for clustering ACC and ARI on algorithm k-Means. In addition, the relatively small standard deviation also indicates the stability of the results of the proposed algorithm FMIUFS.

In summary, algorithm FMIUFS achieves better clustering result. It can obtain a relatively small subset of attributes and improve or maintain the clustering result. Therefore, it is suitable for mixed feature selection in case of clustering.

D. Outlier detection results

The number of optimal feature subset on algorithm kNNO is presented in Table X. The underline "–" indicates that the corresponding feature selection algorithm does not remove any features. Through Table X, we can see that most feature selection algorithms can remove some candidate features for all datasets. Among them, the proposed algorithm FMIUFS obtains a smaller number of preferred features in most cases. In addition, algorithm FMIUFS also achieves a relatively small average number of selected features

Table XI gives a comparison of outlier detection precision based on algorithm kNNO. On 14 datasets of outlier detection, it can be seen from Table XI that most comparison algorithms can maintain or improve the original outlier detection precision. Among them, algorithm FMIUFS achieves the best precision on 10 datasets. However, algorithms FSFS, SPEC, MI, USFSM, FR, USQR, UFRFS, and FEUAR only achieve the best clustering ARI on 5, 5, 5, 5, 8, 2, 2, and 2 datasets, respectively.

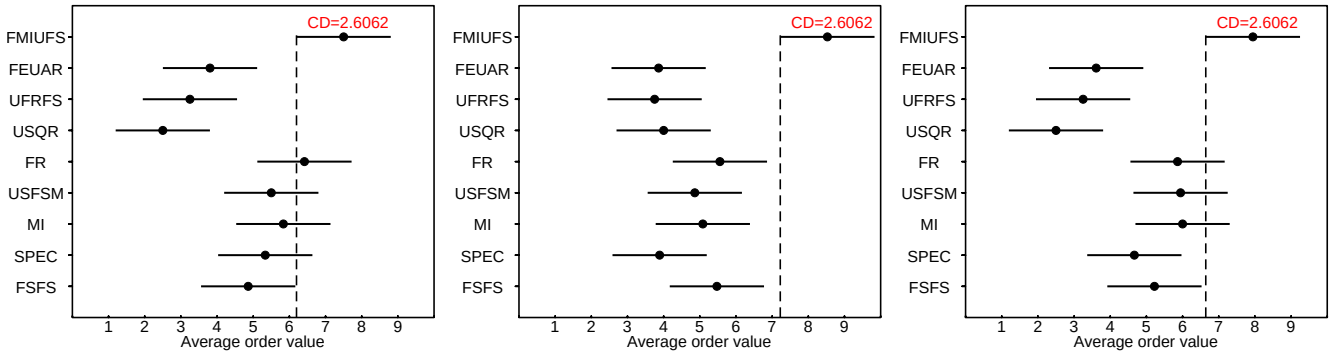
In summary, algorithm FMIUFS can remove irrelevant or redundant features and improve or maintain the precision of outlier detection algorithms.

E. Statistical test

What's more, Friedman's test [63] and Nemenyi's post-hoc test [64] are applied to evaluate the statistical significance of the results. Before using Friedman's test, the accuracy of each

TABLE XI: The comparison of detection precision of reduced data based on algorithm kNNO (%)

Datasets	Original data	FSFS	SPEC	MI	USFSM	FR	USQR	UFRFS	FEUAR	FMIUFS
Cre	64.29	64.29	**80.95**	66.67	76.19	**80.95**	66.67	66.67	71.43	71.43
Ger	42.86	42.86	42.86	35.71	**50.00**	**50.00**	28.57	42.86	35.71	**50.00**
Hea	81.25	75.00	**81.25**	75.00	**81.25**	**81.25**	50.00	–	–	**81.25**
Hep	88.89	55.56	55.56	**88.89**	**88.89**	**88.89**	33.33	44.44	–	**88.89**
Dia	46.15	**53.85**	50.00	46.15	50.00	42.31	–	–	–	46.15
Ion	100.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Iri	72.73	72.73	**100.00**	**100.00**	72.73	**100.00**	–	–	–	**100.00**
Pim	58.18	**61.82**	60.00	50.91	60.00	60.00	58.18	58.18	58.18	58.18
Son	90.00	80.00	**90.00**	**90.00**	**90.00**	80.00	80.00	60.00	–	**90.00**
Wb	87.18	87.18	87.18	87.18	84.62	89.74	–	–	–	**92.31**
Wdb	87.18	87.18	89.74	92.31	92.31	94.87	89.74	71.79	–	**97.44**
Yea	20.00	**60.00**	0.00	**60.00**	40.00	**60.00**	0.00	–	0.00	**60.00**
Lym	66.67	**100.00**	66.67	66.67	83.33	66.67	**100.00**	**100.00**	**100.00**	**100.00**
Mus	34.39	49.32	71.04	49.32	59.73	**92.76**	35.29	34.39	36.20	59.28
Average	67.13	70.70	69.66	72.06	73.50	77.67	–	–	–	**78.21**



(a) CT (b) NB (c) k-Means

Fig. 1: Nemenyi's test figures on algorithms CT, NB, and k-Means

algorithm on all datasets is sorted from low to high, and the sequence number is assigned $(1, 2, \cdots)$. Among them, if the accuracy of the two algorithms is the same, the ordinal values are equally divided. Then, Friedman's test is used to determine whether these algorithms have the same performance. Suppose we compare M algorithms on N datasets, and let r_i represent the average ordinal value of the ith algorithm, then Friedman's test is calculated as follows.

$$\begin{aligned} \tau_F &= \frac{(N-1)\tau_{\chi^2}}{N(M-1)-\tau_{\chi^2}} \text{ and} \\ \tau_{\chi^2} &= \frac{12N}{M(M+1)}\left(\sum_{i=1}^{M} r_i^2 - \frac{M(M+1)^2}{4}\right). \end{aligned} \quad (16)$$

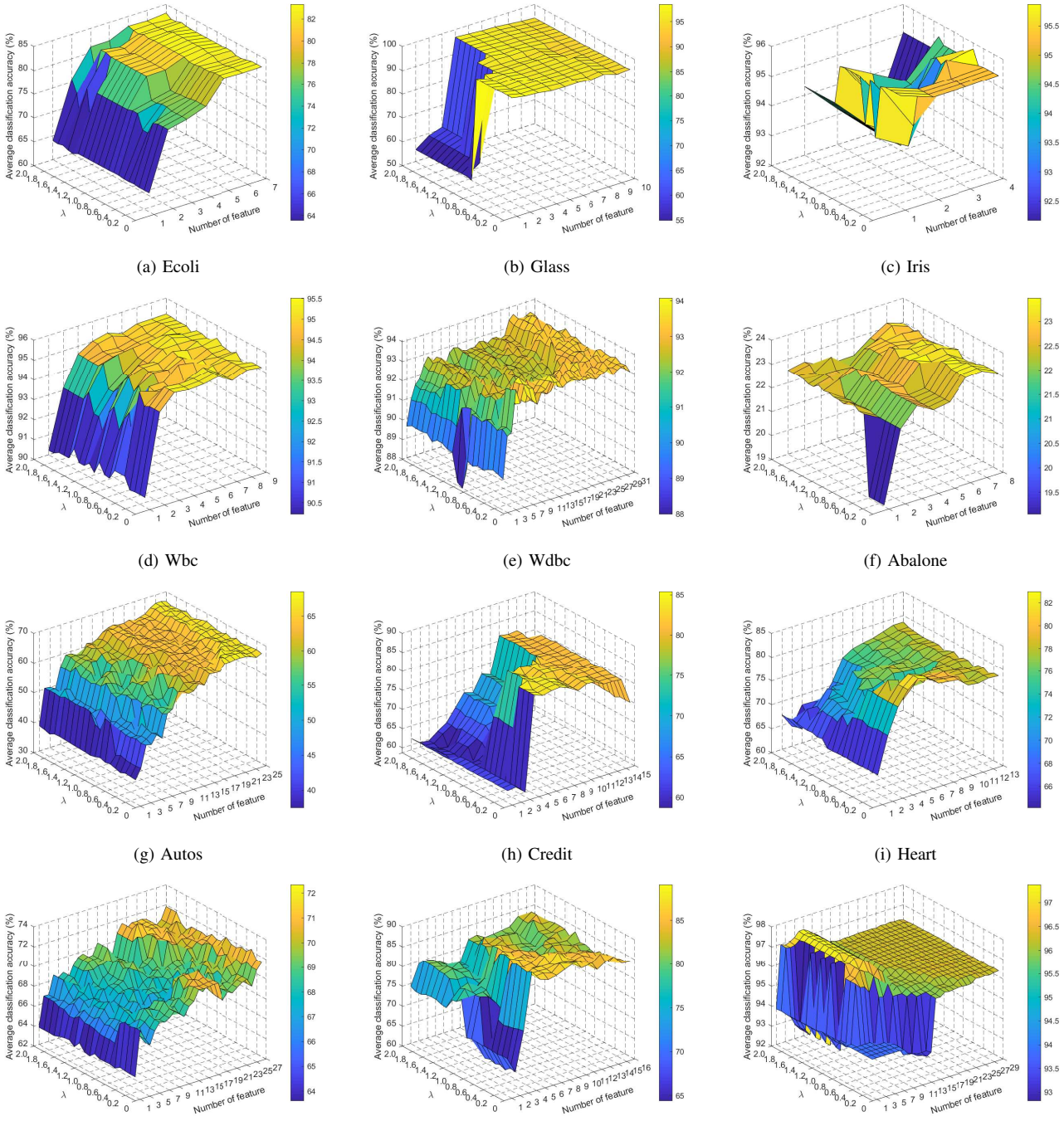
τ_F obeys the F distribution with $(M-1)$ and $(M-1)(N-1)$ degrees of freedom. If the null hypothesis of "all algorithms have the same performance" is rejected, it means that the performance of the algorithms is significantly different. At this time, a post-hoc test needs to be used to further distinguish these feature selection algorithms. Nemenyi's post-hoc test is commonly used. In Nemenyi's test, the critical difference (CD) of the average ordinal value is calculated by the following formula.

$$CD_\alpha = q_\alpha \sqrt{\frac{M(M+1)}{6N}}, \quad (17)$$

where q_α is the critical value of Tukey's distribution, which can be found in [64].

Further, Nemenyi's test figure is used to more intuitively represent the significant differences between the two algorithms [65]. In Nemenyi's test figure, for each algorithm, a dot is used to show its average ordinal value, and a horizontal line segment with the dot as the center is used to indicate the size of CD. If the horizontal line segments of the two algorithms do not overlap, it means that there is a significant difference between the two algorithms, otherwise it means that there is no significant difference.

Specifically, in order to facilitate comparison, original accuracy is used to fill all entries that have no value in Tables V, VI, and VIII. Finally, we can get $M = 9$ and $N = 18$, the τ_F distribution has 8 and 136 degrees of freedom. According to Friedman's test, τ_F of different learning algorithms and critical value (significance level α is 0.1) are shown in Table XII. According to Table XII, when $\alpha = 0.1$, each value of τ_F on algorithms CT, NB, and k-Means is greater than the critical value 1.7158. Therefore, the null hypothesis that "all



(a) Ecoli (b) Glass (c) Iris

(d) Wbc (e) Wdbc (f) Abalone

(g) Autos (h) Credit (i) Heart

(j) Horse (k) Labor (l) Sick

Fig. 2: The average classification accuracy varies with the parameter λ and the number of features

algorithms have the same performance" is rejected algorithms CT, NB, and k-Means. It shows that the performance of all feature selection algorithms is significantly different on algorithms CT, NB, and k-Means. At this time, a post-hoc test needs to be used to further distinguish these feature selection algorithms.

For significance level $\alpha = 0.1$, the corresponding critical distance $CD_{0.1} = 2.6062$ can be obtained. Finally, Nemenyi's test figures on three learning algorithms are shown in Fig. 1. From Fig. 1, we can see that algorithm FMIUFS is statistically significantly different from most other algorithms. For example, it can be seen from Fig. 1(b) that the horizontal line segments of algorithm FMIUFS and other algorithms have

TABLE XII: τ_F on algorithms CT, NB, and k-Means

Algorithms	τ_F	Critical value ($\alpha = 0.1$)
CT	8.5453	
NB	7.3421	1.7158
k-Means	10.3497	

no overlapping area, which shows that algorithm FMIUFS and other algorithms are statistically significantly different on algorithm NB. Fig. 1(c) demonstrates that algorithm FMIUFS is statistically better than algorithms FSFS, SPEC, USQR, UFRFS, and FEUAR on algorithm k-Means, respectively. However, there is no consistent evidence to indicate the statistical differences from algorithms MI, USFSM, and FR on algorithm k-Means. Besides, the average order value of three learning algorithms in Fig. 1 is also relatively large, which also shows the effectiveness of algorithm FMIUFS.

F. Parametric sensitivity analyses

The threshold λ plays an important role in algorithm FMIUFS. It can be used as a parameter to control the fuzzy granularity of data analysis. We can obtain different feature sequences at each granularity level. Different feature subsets may be obtained by feature selection algorithms. This section takes the classification results as an example to analyze the sensitivity of parameters λ and the number of selected features.

The relations among parameters λ , the number of selected features, and the classification accuracy are depicted in Fig. 2. Obviously, it can be seen that the classification performance is different on different datasets. Through Fig. 2, the specific analyses are as follows.

1) In most datasets, as the number of features increases, the classification accuracy increases rapidly at first, and then remains unchanged or even decreases, such as datasets Ecoli, Wbc, and Auto. This is because selected redundant or irrelevant features cannot provide new information to a learning algorithm and may even mislead the learning algorithm.
2) For datasets Glass, Iris and Credit, the classification algorithm can obtain the same classification accuracy on the reduced data with a small number of features as on the original data.
3) On some datasets, when algorithm FMIUFS is used to obtain the first few features in the feature sequence, the classification accuracy is higher than that on the original data, such as Heart, Labor, and Sick datasets. These results show that algorithm FMIUFS really puts important features that are representative and rich in information in the head of the feature sequence.
4) We can see that with the increase of λ on some datasets, such as datasets Abalone, Glass, and Sick, the average classification accuracy does not change much. However, for Autos, Horse and Labor datasets, as λ increases, the average classification accuracy fluctuates greatly. It shows that λ is too small or too large to make the algorithm optimal.
5) From Fig. 2, it can be seen that for most datasets, the optimal value can be obtained under multiple parameter λ values. For each dataset, we can choose the appropriate value of λ to achieve better performance according to Fig. 2.

Through the above analyses, we can see that the experimental performance is sensitive to the parameter λ and the number of selected features, which is still need to be further studied. However, under the appropriate λ value and selected feature data conditions, algorithm FMIUFS can obtain better results in most cases. In summary, algorithm FMIUFS is feasible for feature selection of three learning algorithms.

V. Conclusion

This paper proposes an unsupervised method for heterogeneous feature selection based on fuzzy mutual information. This method not only utilizes the ability of fuzzy information theory to effectively process uncertainty data, but also comprehensively considers the relevance and redundancy of features. Further, the hybrid fuzzy similarity measure is used to calculate the fuzzy similarity relation, which makes the proposed method suitable for nominal, numerical and heterogeneous attribute data. In addition, raw data does not need to be replaced and discretized, which can reduce data processing time and information loss. For this method, the corresponding algorithm FMIUFS is designed. Algorithm FMIUFS and the existing algorithms are compared and analyzed on UCI datasets. The results show that the proposed algorithm can select fewer features to maintain or improve the performance of learning algorithms, and it is suitable for mixed attribute data.

In future work, some other granular computing models (such as neighborhood rough set models) can also be employed for unsupervised feature selection.

References

[1] P. F. Zhu, W. M. Zuo, L. Zhang, Q. H. Hu, and S. C. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.
[2] P. F. Zhu, Q. Xu, Q. H. Hu, and C. Q. Zhang, "Co-regularized unsupervised feature selection," *Neurocomputing*, vol. 275, pp. 2855–2863, 2018.
[3] Y. Zhang, Q. Wang, D. W. Gong, and X. F. Song, "Nonnegative laplacian embedding guided subspace learning for unsupervised feature selection," *Pattern Recognition*, vol. 93, pp. 337–352, 2019.
[4] D. G. Chen and Y. Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 5, pp. 1325–1334, 2014.
[5] X. Zhang, C. L. Mei, D. G. Chen, and J. H. Li, "Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy," *Pattern Recognition*, vol. 56, pp. 1–15, 2016.
[6] H. M. Chen, T. R. Li, R. Da, J. H. Lin, and C. X. Hu, "A rough-set based incremental approach for updating approximations under dynamic maintenance environments," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 2, pp. 274–284, 2013.
[7] T. P. Hong, T. T. Wang, S. L. Wang, and B. C. Chien, "Learning a coverage set of maximally general fuzzy rules by rough sets," *Expert Systems with Applications*, vol. 19, no. 2, pp. 97–103, 2000.
[8] H. M. Chen, T. R. Li, C. Luo, S. J. Horng, and G. Y. Wang, "A decision-theoretic rough set approach for dynamic data mining," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 1958–1970, 2015.
[9] X. Y. Zhang, H. Yao, Z. Y. Lv, and D. Q. Miao, "Class-specific information measures and attribute reducts for hierarchy and systematicness," *Information Sciences*, vol. 563, pp. 196–225, 2021.
[10] X. Y. Zhang, H. Y. Gou, Z. Y. Lv, and D. Q. Miao, "Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system," *Knowledge-Based Systems*, vol. 217, p. 106799, 2021.
[11] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General System*, vol. 17, no. 2-3, pp. 191–209, 1990.
[12] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," in *Intelligent Decision Support*, pp. 203–232, Springer, 1992.
[13] W. Z. Wu, J. S. Mi, and W. X. Zhang, "Generalized fuzzy rough sets," *Information sciences*, vol. 151, pp. 263–282, 2003.

[14] D. S. Yeung, D. G. Chen, E. C. C. Tsang, J. W. T. Lee, and X. Z. Wang, "On the generalization of fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 3, pp. 343–361, 2005.

[15] J. S. Mi, Y. Leung, H. Y. Zhao, and T. Feng, "Generalized fuzzy rough sets determined by a triangular norm," *Information Sciences*, vol. 178, no. 16, pp. 3203–3213, 2008.

[16] Q. H. Hu, S. An, X. Yu, and D. R. Yu, "Robust fuzzy rough classifiers," *Fuzzy Sets and Systems*, vol. 183, no. 1, pp. 26–43, 2011.

[17] S. Zhao, H. Chen, C. Li, X. Du, and H. Sun, "A novel approach to building a robust fuzzy rough classifier," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 4, pp. 769–786, 2014.

[18] P. Maji, "Fuzzy–rough supervised attribute clustering algorithm and classification of microarray data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 222–233, 2010.

[19] X. Z. Wang, E. C. C. Tsang, S. Y. Zhao, D. G. Chen, and D. S. Yeung, "Learning fuzzy rules from fuzzy samples based on rough set technique," *Information Sciences*, vol. 177, no. 20, pp. 4493–4514, 2007.

[20] R. B. Bhatt and M. Gopal, "Frct: fuzzy-rough classification trees," *Pattern Analysis and Applications*, vol. 11, no. 1, pp. 73–88, 2008.

[21] A. Hassanien, "Fuzzy rough sets hybrid scheme for breast cancer detection," *Image and Vision Computing*, vol. 25, no. 2, pp. 172–183, 2007.

[22] R. Jensen and Q. Shen, "Fuzzy–rough attribute reduction with application to web categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469–485, 2004.

[23] E. C. Tsang, D. Chen, D. S. Yeung, X.-Z. Wang, and J. W. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Transactions on Fuzzy systems*, vol. 16, no. 5, pp. 1130–1141, 2008.

[24] Q. H. Hu, D. R. Yu, and Z. X. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern Recognition Letters*, vol. 27, no. 5, pp. 414–423, 2006.

[25] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches," *IEEE Transactions on Knowledge & Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[26] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 73–89, 2007.

[27] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2008.

[28] Q. H. Hu, Z. X. Xie, and D. R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognition*, vol. 40, no. 12, pp. 3509–3521, 2007.

[29] C. Z. Wang, Y. L. Qi, M. W. Shao, Q. H. Hu, D. G. Chen, Y. H. Qian, and Y. J. Lin, "A fitting model for feature selection with fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 741–753, 2017.

[30] C. Z. Wang, Y. Huang, M. W. Shao, and X. D. Fan, "Fuzzy rough set-based attribute reduction using distance measures," *Knowledge-Based Systems*, vol. 164, pp. 205–212, 2019.

[31] C. Z. Wang, Y. Wang, M. W. Shao, Y. H. Qian, and D. G. Chen, "Fuzzy rough attribute reduction for categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 5, pp. 818–830, 2020.

[32] D. G. Chen, Q. H. Hu, and Y. P. Yang, "Parameterized attribute reduction with gaussian kernel based fuzzy rough sets," *Information Sciences*, vol. 181, no. 23, pp. 5169–5179, 2011.

[33] D. G. Chen, L. Zhang, S. Y. Zhao, Q. H. Hu, and P. F. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 2, pp. 385–389, 2012.

[34] J. H. Dai, H. Hu, W. Z. Wu, Y. H. Qian, and D. Huang, "Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 2174–2187, 2018.

[35] R. R. Yager, "Entropy measures under similarity relations," *International Journal Of General System*, vol. 20, no. 4, pp. 341–358, 1992.

[36] J. S. Mi, Y. Leung, and W. Z. Wu, "An uncertainty measure in partition-based fuzzy rough sets," *International Journal of General Systems*, vol. 34, no. 1, pp. 77–90, 2005.

[37] C. Z. Wang, Y. Huang, M. W. Shao, and D. G. Chen, "Uncertainty measures for general fuzzy relations," *Fuzzy Sets and Systems*, vol. 360, pp. 82–96, 2019.

[38] S. An, Q. H. Hu, and D. R. Yu, "Fuzzy entropy based max-relevancy and min-redundancy feature selection," in *2008 IEEE International Conference on Granular Computing*, pp. 101–106, IEEE, 2008.

[39] D. R. Yu, S. An, and Q. H. Hu, "Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection," *International Journal of Computational Intelligence Systems*, vol. 4, no. 4, pp. 619–633, 2011.

[40] J. H. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, no. 1, pp. 211–221, 2013.

[41] 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.

[42] J. H. Dai and J. L. Chen, "Feature selection via normative fuzzy information weight with application in biological data classification," *Applied Soft Computing*, pp. 106–299, 2020.

[43] A. Ganivada, S. S. Ray, and S. K. Pal, "Fuzzy rough sets, and a granular neural network for unsupervised feature selection," *Neural Networks*, vol. 48, pp. 91–108, 2013.

[44] N. Mac Parthaláin and R. Jensen, "Unsupervised fuzzy-rough set-based dimensionality reduction," *Information Sciences*, vol. 229, pp. 106–121, 2013.

[45] Q. H. Hu, D. R. Yu, Z. X. Xie, and J. F. Liu, "Fuzzy probabilistic approximation spaces and their information measures," *IEEE transactions on fuzzy systems*, vol. 14, no. 2, pp. 191–201, 2006.

[46] D. R. Yu, Q. H. Hu, and C. X. Wu, "Uncertainty measures for fuzzy relations and their applications," *Applied soft computing*, vol. 7, no. 3, pp. 1135–1143, 2007.

[47] 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, 2005.

[48] J. L. Xu, Y. M. Zhou, L. Chen, and B. W. Xu, "An unsupervised feature selection approach based on mutual information," *Journal of Computer Research and Development*, vol. 49, no. 2, p. 372, 2012.

[49] Z. Yuan, X. Y. Zhang, and S. Feng, "Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures," *Expert Systems with Applications*, vol. 112, pp. 243–257, 2018.

[50] Z. Yuan, H. M. Chen, T. R. Li, J. Liu, and S. Wang, "Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection," *Fuzzy Sets and Systems*, vol. 421, pp. 1–28, 2021.

[51] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017.

[52] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[53] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[54] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, 2007.

[55] S. Solorio-Fernández, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A new unsupervised spectral feature selection method for mixed data: a filter approach," *Pattern Recognition*, vol. 72, pp. 314–326, 2017.

[56] A. Chaudhuri, D. Samanta, and M. Sarma, "Two-stage approach to feature set optimization for unsupervised dataset with heterogeneous attributes," *Expert Systems with Applications*, vol. 172, p. 114563, 2021.

[57] C. Velayutham and K. Thangavel, "Unsupervised quick reduct algorithm using rough set theory," *Journal of Electronic Science and Technology*, vol. 9, no. 3, pp. 193–201, 2011.

[58] P. F. Zhu, W. C. Zhu, Q. H. Hu, C. Q. Zhang, and W. M. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognition*, vol. 66, pp. 364–374, 2017.

[59] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[60] S. Zhang, H.-S. Wong, and Y. Shen, "Generalized adjusted rand indices for cluster ensembles," *Pattern Recognition*, vol. 45, no. 6, pp. 2214–2226, 2012.

[61] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 427–438, 2000.

[62] D. R. Yu, Q. H. Hu, and W. Bao, "Combining rough set methodology and fuzzy clustering for knowledge discovery from quantitative data," *Proceedings of the CSEE*, vol. 24, no. 6, pp. 205–210, 2004.

[63] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[64] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[65] Z. Yuan, H. M. Chen, P. Xie, P. F. Zhang, J. Liu, and T. R. Li, "Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions," *Applied Soft Computing*, vol. 107, p. 107353, 2021.

Zhong Yuan received the B.S. degree in mathematics from Sichuan Minzu College, Kangding, China, in 2015. He received the M.S. degree in mathematics from Sichuan Normal University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with Southwest Jiaotong University, Chengdu, China. His research interests include rough sets, granular computing, and data mining.

Hongmei Chen (Member, IEEE) received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2000. She received the Ph.D. degree from the Southwest Jiaotong University, Chengdu, China, in 2013. She is currently a professor at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. Her research interests include the areas of data mining, pattern recognition, fuzzy sets, and rough sets. Her research interests include data mining, pattern recognition, fuzzy sets, and rough sets.

Pengfei Zhang received the M.Sc. degree in mathematics from the Guangxi University for Nationalities, Nanning, China, in 2019. He is currently pursuing the Ph.D. degree with Southwest Jiaotong University, Chengdu, China. His research interests include granular computing, rough set theory, data mining, and information fusion.

Jihong Wan received the B.S. degree from the Zhengzhou Normal University, Zhengzhou, China, in 2016, and the M.S. degree from the Xihua University, Chengdu, China, in 2019. She is currently pursuing the Ph. D. degree with the Southwest Jiaotong University, Chengdu, China. Her research interests include data mining, granular computing, rough sets, and social networks, etc. She has published several papers in journals such as IEEE Transactions on Cybernetics, Knowledge-Based Systems, Neurocomputing, etc.

Tianrui Li (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Southwest Jiaotong University, Chengdu, China, in 1992, 1995, and 2002, respectively. He was a postdoctoral researcher at the Belgian Nuclear Research Centre (SCK.CEN), Belgium from 2005-2006, a visiting professor at Hasselt University, Belgium, in 2008 and the University of Technology, Sydney, Australia in 2009. Currently, he is a professor and the director of the Key Lab of Cloud Computing and Intelligent Technology, Southwest Jiaotong University, China. He has authored or coauthored more than 300 research papers in refereed journals and conferences. He is a senior member of the IEEE. His research interests include big data, cloud computing, data mining, granular computing, and rough sets.