Knowledge-Based Systems 231 (2021) 107398

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Fuzzy complementary entropy using hybrid-kernel function and its unsupervised attribute reduction

Zhong Yuan, Hongmei Chen *, Xiaoling Yang, Tianrui Li, Keyu Liu

School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China

ARTICLE INFO

Article history:
Received 19 April 2021
Received in revised form 8 July 2021
Accepted 12 August 2021
Available online 23 August 2021

Keywords:
Fuzzy rough set theory
Hybrid-kernel function
Complementary entropy
Uncertainty measure
Unsupervised attribute reduction

ABSTRACT

Fuzzy rough set theory has been proved to be an effective tool to deal with uncertainty data. Some different forms of fuzzy uncertainty measures have been introduced in fuzzy rough set theory, such as fuzzy information entropy, fuzzy conditional entropy, and fuzzy mutual information. However, as far as we know, most of the above fuzzy conditional entropy and fuzzy mutual information are non-monotonic, which may lead to a non-convergent learning algorithm. For this reason, this paper proposes a novel fuzzy complementary entropy based on the hybrid-kernel function. Then, based on the proposed fuzzy complementary entropy, some corresponding uncertainty measures are also proposed. Furthermore, fuzzy complementary conditional entropy and fuzzy complementary mutual information are proved to change monotonously with attributes. Finally, based on the proposed uncertainty measures, three kinds of evaluation criteria for unsupervised hybrid attribute reduction are defined and a generalized attribute reduction algorithm is designed. The experimental results show that the proposed method is an effective scheme for reducing hybrid attributes.

1. Introduction

Fuzzy rough set (FRS) theory was first proposed by Dubois and Prade [1,2], which is an effective mathematical model. It makes up for the shortcomings of the discretization method, so it can directly process numeric or hybrid attribute data. The researches on FRS theory mainly involve the expansion and application of models [3]. Inspired by the original FRS model, a series of FRS extensions [4–6] were proposed. For example, Mi and Zhang proposed a new definition of FRS based on residual implication and its duality [4]. Yeung et al. proposed two FRS methods based on arbitrary fuzzy relations [6]. In recent years, FRS theory has been successfully applied in the fields of outlier detection [7,8], attribute reduction (or feature selection) [9–13], group decision making [14,15], and rule extraction [16,17].

Most FRS models are constructed in the fuzzy granular space generated by fuzzy $T-$ equivalence relations, but how to effectively generate fuzzy relations from data has not been systematically discussed. The method of generating fuzzy relations from the data directly affects the performance of the model [18,19]. Therefore, it is very important to develop a systematic and effective method to determine fuzzy relations. The kernel-based method transforms the nonlinear learning problem in low-dimensional space into a linear learning problem in high-dimensional space [18]. In this way, many linear learning algorithms can be employed to handle non-linear tasks, such as kernel perceptron [20], kernel discrimination analysis [21], and kernel matching pursuit [22], etc.

Moser proved that the kernel that satisfies reflexivity and symmetry is at least $T_{cos}-$transitive [5,23]. The fuzzy relation calculated by this kind of kernel function is the fuzzy $T_{cos}-$equivalence relation. For this reason, Hu et al. used the fuzzy relation generated by the kernel function to granulate the universe and then constructed different kernel-based FRSs [18,19]. These models build a bridge between the kernel method and the rough set theory. Furthermore, attribute reductions based on kernelized FRS were also studied [18,19,24–27]. For example, Chen et al. studied the attribute reduction of FRSs based on Gaussian kernel and then designed an attribute reduction algorithm using fuzzy discernibility matrix [24]. Li et al. used the kernel function to process the feature space and then proposed a feature selection to multi-label learning based on the kernelized FRS [25]. Hu et al. further constructed a multi-kernel FRS model and designed an attribute reduction algorithm for large-scale multi-modal data [26]. Recently, Rao et al. put forward an accelerated method of attribute reduction based on Gaussian kernel FRS [27].

* Corresponding author.
E-mail addresses: yuanzhong2799@foxmail.com (Z. Yuan), hmchen@swjtu.edu.cn (H. Chen), yangxl@my.swjtu.edu.cn (X. Yang), trli@swjtu.edu.cn (T. Li), keyu_liu@my.swjtu.edu.cn (K. Liu).

https://doi.org/10.1016/j.knosys.2021.107398
0950-7051/

Information entropy comes from information theory, which can effectively measure the uncertainty of a system. [28]. Considering 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 [29,30]. Mi et al. gave a new fuzzy entropy and applied it to the partition-based FRS for the first time [31]. Since then, different forms of fuzzy uncertainty measures have been proposed in FRSs [32–34], such as fuzzy information entropy [32], fuzzy conditional entropy [35] and fuzzy mutual information [34]. Among them, only some fuzzy uncertainty measures are applied to attribute reduction. For example, Hu et al. used the information entropy given in [29,30] to redefine joint entropy and conditional entropy and applied them to hybrid feature selection [9,32]. Yu et al. defined a generalized fuzzy entropy based on fuzzy relations and used it to construct the evaluation criteria of the significance of attributes [36]. Taking into account the lack of fuzzy entropy defined in [9], Zhang et al. proposed a monotonic fuzzy conditional entropy for hybrid feature selection [35]. Based on the monotonic fuzzy conditional entropy proposed above, Zhang et al. further studied active incremental feature selection [37]. Wang et al. proposed fuzzy entropy and some related metrics under generalized fuzzy relations and designed corresponding algorithms for feature selection [38]. Under the framework of FRS theory, Dai et al. proposed an attribute selection method based on fuzzy information gain ratio by using the concept of gain ratio in decision tree theory [39]. Zhao et al. proposed a complementary information entropy model based on arbitrary fuzzy relations in FRSs and applied the information entropy to feature selection [40]. Lin et al. introduced fuzzy mutual information in multi-label learning to evaluate the quality of features [41]. Recently, based on regulating fuzzy information weights, Dai et al. designed a feature selection strategy based on fuzzy conditional mutual information [34]. Wang et al. combined fuzzy rough approximation with the concept of self-information, and constructed four monotonic fuzzy uncertainty measures for feature selection [42].

However, as far as we know, most of the fuzzy conditional entropy and mutual information defined by the above fuzzy uncertainty measures are non-monotonic, which may lead to a non-convergent learning algorithm. Besides, the definition of the fuzzy uncertainty measures mentioned above cannot reflect the uncertainty of the system well. The reason is that when calculating the fuzzy entropy, the intersection operation is used to calculate the fuzzy similarity relation. Such a strategy may reduce the ability to distinguish objects in high-dimensional sample spaces [43]. Therefore, the fuzzy similarity relations may not truly reflect the relationship between two objects. What is more, the above attribute reduction methods are all supervised, and the label information of the data object must be known in advance. Therefore, they cannot be applied to attribute reduction without labels (or decisions). As far as we know, there is only a small amount of work using FRS theory for unsupervised feature selection [9,44,45]. For example, Hu et al. first used fuzzy information entropy to define the attribute significance for unsupervised feature selection [9]. Ganivada et al. studied an unsupervised attribute reduction based on the granular neural network using FRS model [44]. In [45], Mac Parthaláin et al. introduced an unsupervised feature selection method based on FRSs.

A novel fuzzy complementary entropy is proposed by using a hybrid-kernel method in this paper, which considers the complementarity of fuzzy information. Then, based on the proposed fuzzy complementary entropy, fuzzy complementary joint entropy, fuzzy complementary conditional entropy, and fuzzy complementary mutual information are also proposed. Based on the proposed uncertainty measures, three unsupervised hybrid attribute reduction methods are proposed. Finally, a generalized unsupervised heuristic reduction algorithm is designed. Specifically, our main contributions in this paper are as follows.

(1) The hybrid-kernel function is defined to calculate the degree of fuzzy similarity between two objects, which makes the proposed fuzzy uncertainty measure suitable for hybrid data;
(2) It is proved that fuzzy complementary conditional entropy and fuzzy complementary mutual information change monotonously with attributes;
(3) The proposed uncertainty measures are applied to unsupervised hybrid attribute reduction;
(4) Twenty-three actual data sets are used to compare and analyze the proposed attribute reduction method and existing algorithms;
(5) The experimental results show that the proposed method is an effective scheme for reducing mixed attributes.

The rest of this paper is organized as follows. In the next section, the preliminary knowledge about kernelized FRSs and fuzzy uncertainty measures are introduced. In the third section, a hybrid-kernel based fuzzy complementary entropy is defined, and some related properties are discussed. Furthermore, based on the proposed uncertainty measures, three unsupervised hybrid attribute reduction methods are proposed and a generalized unsupervised reduction algorithm is designed. The experimental results are given in the fourth section. The fifth section summarizes the paper.

2. Preliminaries

In fuzzy information theory, data can be imported into an information system $IS = \langle U, A\rangle$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a non-empty finite set of objects, A is a non-empty finite set of attributes. When $A = C \cup D$ and $C \cap D = \emptyset$, the information system is called a decision system (DS), where C represents conditional attributes, and D represents decision attributes. If $\mathcal{X}$ is a mapping from U to $[0, 1]$, then $\mathcal{X}$ is called a fuzzy set on U. $\forall x \in U$, $\mathcal{X}(x)$ is called the membership degree of x to $\mathcal{X}$. The set of all fuzzy sets on U is denoted as $\mathcal{F}(U)$. Let $\mathcal{X}, \mathcal{Y} \in \mathcal{F}(U)$, some operations are defined as follows.

(1) $\mathcal{X} = \mathcal{Y} \Leftrightarrow \mathcal{X}(x) = \mathcal{Y}(x), \forall x \in U$;
(2) $\mathcal{X} \subseteq \mathcal{Y} \Leftrightarrow \mathcal{X}(x) \leq \mathcal{Y}(x), \forall x \in U$;
(3) $(\mathcal{X} \cup \mathcal{Y})(x) = \max\{\mathcal{X}(x), \mathcal{Y}(x)\} = \mathcal{X}(x) \vee \mathcal{Y}(x), \forall x \in U$;
(4) $(\mathcal{X} \cap \mathcal{Y})(x) = \min\{\mathcal{X}(x), \mathcal{Y}(x)\} = \mathcal{X}(x) \wedge \mathcal{Y}(x), \forall x \in U$;
(5) $\mathcal{X}^c(x) = 1 - \mathcal{X}(x), \forall x \in U$.

Given $a \in [0, 1]$, a binary operator $T : [0, 1]^2 \to [0, 1]$ is called the triangular norm ($T-$ norm) on [0,1], if it satisfies commutativity, associativity, monotonicity, and $T(a, 1) = a$. If a binary operator $S : [0, 1]^2 \to [0, 1]$ satisfies the first three conditions above and $S(a, 0) = a$, then the binary operator S is called the triangle conorm on [0,1]. Negator operator N is a descending map on $[0, 1] \to [0, 1]$ that satisfies the boundary conditions $N(0) = 1$ and $\mathcal{N}(1) = 0$.

A fuzzy relation $\mathcal{R}$ on $U \times U$ is defined as $\mathcal{R} : U \times U \to [0, 1]$. $\forall x, y, z \in U$, if it meets the following conditions:

(1) Reflexivity: $\mathcal{R}(x, x) = 1, \forall x \in U$;
(2) Symmetry: $\mathcal{R}(x, y) = \mathcal{R}(y, x), \forall x, y \in U$;
(3) $T-$transitivity: $\mathcal{R}(x, z) \geq T_{y\in U}(\mathcal{R}(x, y), \mathcal{R}(y, z)), \forall x, y, z \in U$,

then $\mathcal{R}$ is called the fuzzy $T-$ equivalence relation on U; if $T =$ min, then $\mathcal{R}$ is called the fuzzy equivalence relation on U.

Moser et al. established a closed relationship between the fuzzy similarity relation and the kernel function [5,23]. The kernel function is defined as follows.

Definition 1 ([5]). Given a non-empty finite set of objects U, if a real-valued function $k : U \times U \to R$ satisfies positive-semidefinite and symmetry, then k is called a kernel function.

Theorem 1 ([5]). *Any kernel function k whose value range is in the unit interval and satisfies $k(x,x) = 1$, then k is at least T_{cos}-transitive, where*

$$T_{cos}(x,y) = \max\{xy - \sqrt{1-x^2}\sqrt{1-y^2}, 0\}. \tag{1}$$

According to Definition 1 and Theorem 1, it can be known that some kernel functions satisfy reflexivity, symmetry, and $T_{cos}-$transitivity. The relationship between objects is a fuzzy $T_{cos}-$equivalence relation when they are calculated by such kernel function. Therefore, the kernel function can be used for relation extraction in FRS theory. Hu et al. replaced the fuzzy relations in FRSs with kernel functions and obtained the following definition of kernelized FRSs [19].

Definition 2 ([19]). Given a non-empty finite set of objects U and a kernel function k that satisfies reflexivity, symmetry, and T_{cos}-transitivity. For $\mathcal{X} \in \mathcal{F}(U)$, the membership degrees of fuzzy lower approximation and upper approximation of $\mathcal{X}$ are defined as

$$\underline{k_S}\mathcal{X}(x) = \inf_{y\in U} S(N(k(x,y)), \mathcal{X}(y)); \tag{2}$$

$$\overline{k_T}\mathcal{X}(x) = \sup_{y\in U} T(k(x,y), \mathcal{X}(y)). \tag{3}$$

Information entropy can be used to represent the amount of information in an information system. Information entropy is introduced into FRS theory for related uncertainty measurement and expression, resulting in different forms [32,33,46]. Hu et al. proposed a fuzzy information entropy and used it to characterize the uncertainty of FRSs [32]. Let $G(\mathcal{R}) = \{[x_1]_{\mathcal{R}}, [x_2]_{\mathcal{R}}, \ldots, [x_n]_{\mathcal{R}}\}$ denote the fuzzy information granular generated by the fuzzy relation $\mathcal{R}$.

Definition 3 ([32]). The fuzzy information entropy on $\mathcal{R}$ is defined as

$$E(\mathcal{R}) = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|[x_i]_{\mathcal{R}}|}{|U|}. \tag{4}$$

Qian et al. defined another form of fuzzy complementary information entropy [33], which is defined as follows.

Definition 4 ([33]). The fuzzy complementary entropy on $\mathcal{R}$ is defined as

$$CE(\mathcal{R}) = \frac{1}{|U|}\sum_{i=1}^{n}\left(1 - \frac{|[x_i]_{\mathcal{R}}|}{|U|}\right). \tag{5}$$

However, the intersection operation is used to calculate fuzzy similarity relations in most fuzzy information entropies. Such a calculation strategy may reduce the ability to distinguish objects in high-dimensional spaces [43]. Therefore, the fuzzy similarity relation may not truly reflect the relationship between two objects.

Example 1. An information system $IS = \langle U, C\rangle$ is on the left of Table 1, which involves mixed attribute data. Herein, $U = \{x_1, x_2, \ldots, x_6\}$, $C = \{c_1, c_2, c_3, c_4\}$. Among them, c_1, c_2 are numeric attributes, c_3, c_4 are nominal attributes.

Table 1
Initial and normalized IS.

$U\backslash A$	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
x_1	3	0.6	b	D	0.2857	0.7143	b	D
x_2	4	0.8	c	A	0.4286	1.0000	c	A
x_3	1	0.5	c	B	0.0000	0.5714	c	B
x_4	8	0.3	b	B	1.0000	0.2857	b	B
x_5	6	0.1	b	C	0.7143	0.0000	b	C
x_6	7	0.4	c	A	0.8571	0.4286	c	A

The min–max normalization is employed to normalize the raw numerical data, and the normalization results are shown on the right side of Table 1. Let $\mathcal{R}_B$ denote the fuzzy similarity relation induced by $B \subseteq C$. We take the objects x_2 and x_6 as examples to illustrate the problems of the intersection operation. First, the fuzzy similarity degrees of x_2 and x_6 with respect to all single attribute are calculated as $\mathcal{R}_{c_1}(x_2,x_6) = 1 - |0.4286 - 0.8571| = 0.5715$, $\mathcal{R}_{c_2}(x_2,x_6) = 0.4286$, $\mathcal{R}_{c_3}(x_2,x_6) = 1$, $\mathcal{R}_{c_4}(x_2,x_6) = 1$. Then, the intersection operation is adopted to compute $\mathcal{R}_C$, i.e., $\mathcal{R}_C(x_2,x_6) = \min_{k=1}^{4}\mathcal{R}_{c_k}(x_2,x_6) = 0.4286 = \mathcal{R}_{c_2}(x_2,x_6)$.

Through the above calculation, it can be found that the fuzzy similarity degree between x_2 and x_6 on C is only determined by the minimum value $\mathcal{R}_{c_2}(x_2,x_6)$. Obviously, it is not reasonable. Intuitively, the entire attribute set C is easier to distinguish between the objects x_2 and x_6 than the single attribute c_2, that is, their fuzzy similarity degree should become smaller.

In addition, the developed fuzzy complementary conditional entropy and complementary mutual information defined above are also non-monotonic. In order to solve the above problems, we propose a fuzzy complementary entropy based on hybrid-kernel function and apply it to unsupervised attribute reduction.

3. Fuzzy complementary entropy using hybrid-kernel function and its unsupervised attribute reduction

In this section, we fuse the idea of a hybrid-kernel function with fuzzy complementary entropy and propose fuzzy complementary entropy based on the hybrid-kernel function. First, the hybrid-kernel function is introduced to calculate the fuzzy similarity relation under hybrid attribute subsets. Then, the fuzzy complementary entropy based on the hybrid-kernel function and its development measures are defined. Finally, the proposed uncertainty measures are applied to unsupervised attribute reduction.

3.1. Hybrid-kernel function

Suppose $U = \{x_1, x_2, \ldots, x_n\}$, $C = \{c_1, c_2, \ldots, c_m\}$, for $B = \{k_1, k_2, \ldots, k_l\}(1 \le l \le m) \subseteq C$, the description information of x_i on B is denoted through a vector $\langle x_{ik_1}, x_{ik_2}, \ldots, x_{ik_l}\rangle$.

There are a lot of nominal, numerical, and mixed attribute data in real life. For a categorical attribute subset $B^c \subseteq C$, the matching kernel [22,26] is employed to calculate the fuzzy similarity relation between x_i and x_j on B^c, which is defined as follows.

$$k_{B^c}(x_i,x_j) = \begin{cases} 0, \textit{if } x_i \neq x_j; \\ 1, \textit{if } x_i = x_j. \end{cases} \tag{6}$$

For a numerical attribute subset $B^n \subseteq C$, the Gaussian kernel is used to extract the fuzzy similarity relation between x_i and x_j with respect to B^n, which is calculated as follows.

$$k_{B^n}(x_i,x_j) = \exp\left(-\frac{\|x_i - x_j\|_{B^n}^2}{2\sigma^2}\right), \tag{7}$$

where $\|x_i - x_j\|_{B^n}$ is Euclidean distance between x_i and x_j on B^n. σ is used to adjust the granularity of the fuzzy approximation space based on the hybrid-kernel function, which reflects people's tolerance to data noise in numerical attributes. Obviously, the greater the value of σ, the greater the $k_{B^n}(x_i, x_j)$.

Definition 5. For a hybrid attribute subset $B = B^c \cup B^n \subseteq C$, the hybrid-kernel function $k_{B^c\cup B^n}(x_i, x_j)$ is defined by the following algebraic product.

$$k_B(x_i, x_j) = k_{B^c\cup B^n}(x_i, x_j) = k_{B^c}(x_i, x_j) \cdot k_{B^n}(x_i, x_j). \tag{8}$$

Property 1. *$\forall B \subseteq C$, we have*

$$k_B(x_i, x_j) = \prod_{b\in B} k_b(x_i, x_j). \tag{9}$$

Proof. Since $k_{B^c}(x_i, x_j) = \prod_{b_1\in B^c} k_{b_1}(x_i, x_j)$ and $k_{B^n}(x_i, x_j) = \prod_{b_2\in B^n} k_{b_2}(x_i, x_j)$, it is easy to prove that $\forall B \subseteq C$, there is $k_B(x_i, x_j) = \prod_{b\in B} k_b(x_i, x_j)$.

Property 2. *If $B \subseteq E \subseteq C$, then $k_B \supseteq k_E$.*

Proof. It is easy to prove that it is true.

It is easy to see that the above hybrid-kernel function satisfies the properties of reflexivity, symmetry, and $T_{cos}-$ transitivity. Therefore, the fuzzy relation calculated by the above hybrid-kernel functions is the fuzzy $T_{cos}-$ equivalence relation.

3.2. Fuzzy complementary entropy using hybrid-kernel function

Suppose $U = \{x_1, x_2, \dots, x_n\}$, $\forall B \subseteq C$, B can induce a hybrid-kernel T_{cos}-equivalence relation k_B on U. It can be expressed by fuzzy relation matrix $M_{k_B} = (r_{ij}^{k_B})_{n\times n}$, where $r_{ij}^{k_B} = k_B(x_i, x_j)$, each row $(r_{i1}^{k_B}, r_{i2}^{k_B}, \dots, r_{in}^{k_B})$ represents a fuzzy set. Obviously, $\forall B, E \subseteq C$, there is $M_{k_{B\cup E}} = M_{k_B} \otimes M_{k_E}$, where $\otimes$ represents the algebraic product, namely $(r_{ij}^{k_{B\cup E}})_{n\times n} = (r_{ij}^{k_B} \cdot r_{ij}^{k_E})_{n\times n}$. Without causing confusion, we also use B instead of k_B.

The fuzzy information granular containing x_i induced by B is defined as $[x_i]_B = \frac{r_{i1}^B}{x_1} + \frac{r_{i2}^B}{x_2} + \cdots + \frac{r_{in}^B}{x_n} = (r_{i1}^B, r_{i2}^B, \dots, r_{in}^B)$. Obviously, $[x_i]_B$ is a fuzzy set about B. We have $[x_i]_B(x_j) = k_B(x_i, x_j) = r_{ij}^B$. The cardinality of the fuzzy set $[x_i]_{k_B}$ is defined as $|[x_i]_B| = \sum_{j=1}^n r_{ij}^B = \sum_{j=1}^n k_B(x_i, x_j)$. Obviously, we have $1 \le |[x_i]_B| \le n$.

Definition 6. The fuzzy complementary entropy of B is defined as

$$CH(B) = CH(k_B) = \frac{1}{|U|}\sum_{i=1}^n \frac{|[x_i]_B^c|}{|U|}, \tag{10}$$

where $[x_i]_B^c$ denotes the complementary set of $[x_i]_B$, i.e., for $x \in U$, we have $[x_i]_B^c(x) = 1 - [x_i]_B(x)$.

It is easy to get $0 \le CH(B) \le 1 - \frac{1}{|U|}$, if and only if $\forall x, y \in U$, $k_B(x, y) = 1$, i.e., $|[x_i]_B^c| = |U^c| = 0$, so $CH(B) = 0$. In this case all object pairs are indistinguishable. Therefore, the granulation space is the coarsest at this time. On the contrary, $\forall x \ne y$, $k_B(x, y) = 0$, that is, $|[x_i]_B^c| = |\{x_i\}^c| = n - 1$, so $CH(B) = 1 - \frac{1}{|U|}$. At this time, the granulation space is the smallest.

The above fuzzy complementary entropy considers the complementarity of an information system. If k_B degenerates into a crisp equivalence relation, the above-mentioned fuzzy complementary entropy is the same as the classical complementary entropy mentioned in [46].

Property 3. *If $k_{c_s} \subseteq k_{c_t}$, then $CH(k_{c_s}) \ge CH(k_{c_t})$.*

Proof. Given $k_{c_s} \subseteq k_{c_t}$, so $\forall r_{ij}^{c_s}, r_{ij}^{c_t}$, there is $\forall r_{ij}^{c_s} \le r_{ij}^{c_t}$. Then, there is $|[x_i]_{c_s}| = \sum_{j=1}^n r_{ij}^{c_s} \le \sum_{j=1}^n r_{ij}^{c_t} = |[x_i]_{c_t}|$, i.e., $\frac{|[x_i]_{c_s}^c|}{|U|} \ge \frac{|[x_i]_{c_t}^c|}{|U|}$. Therefore, $CH(k_{c_s}) \ge CH(k_{c_t})$.

Property 4. *Let $B \subseteq E \subseteq C$, we have $CH(B) \le CH(E)$.*

Proof. Given $B \subseteq E \subseteq C$, by Property 2, there is $k_B \supseteq k_E$. By Property 3, there is $CH(B) \le CH(E)$.

Property 4 reflects that the fuzzy complementary entropy changes monotonously with attribute. As the attributes increase, the fuzzy complementary entropy increases, and the uncertainty increases. Conversely, the fuzzy complementary entropy becomes smaller and the uncertainty decreases. It can be seen that the increase or decrease of attributes in the fuzzy information system makes the fuzzy complementary entropy and uncertainty change. σ is a parameter used in Eq. (7). $CH(B)$ denotes the fuzzy complementary entropy when σ is used. The property of the fuzzy complementary entropy with different σ is given below.

Property 5. *Let $B \subseteq C$ be a subset of attributes with numeric attributes. If $\sigma_1 \le \sigma_2$, then $CH(B^1) \ge CH(B^2)$. Where B^1 and B^2 respectively represent the fuzzy relations on B calculated through σ_1 and σ_2.*

Proof. Given that $B \subseteq C$ is a subset of attributes with numerical attributes, if $\sigma_1 \le \sigma_2$, then $\exp\left(-\frac{\|x_i - x_j\|_B^2}{2\sigma_1^2}\right) \le \exp\left(-\frac{\|x_i - x_j\|_B^2}{2\sigma_2^2}\right)$, so $k_{B^1} \subseteq k_{B^2}$. Therefore, $CH(B^1) \ge CH(B^2)$.

From Property 5 that under the same attribute, a larger kernel parameter σ will lead to a coarser granularity. The greater the value of σ, the greater the similarity between objects. In this case, it is difficult to distinguish arbitrary objects from other objects. As a result, the fuzzy complementary entropy is lower.

Based on the above definition of fuzzy complementary entropy, fuzzy complementary joint entropy, conditional entropy, and mutual information are further defined.

Definition 7. The fuzzy complementary joint entropy of B and E is defined as

$$CH(E, B) = CH(k_{E\cup B}) = \frac{1}{|U|}\sum_{i=1}^n \frac{|([x_i]_E \otimes [x_i]_B)^c|}{|U|}. \tag{11}$$

Definition 8. The fuzzy complementary conditional entropy of E on B is defined as

$$CH(E|B) = \frac{1}{|U|}\sum_{i=1}^n \frac{|[x_i]_B \otimes [x_i]_E^c|}{|U|}. \tag{12}$$

$CH(E|B)$ describes the uncertainty of E when B is known.

Property 6. *$CH(E|B) = CH(E, B) - CH(B)$.*

Proof. According to the above definition, we have

$$CH(B, E) - CH(B) = \frac{1}{|U|}\sum_{i=1}^n \left(\frac{|([x_i]_B \otimes [x_i]_E)^c| - |[x_i]_B^c|}{|U|}\right)$$
$$= \frac{1}{|U|}\sum_{i=1}^n \left(\frac{\sum_{j=1}^n (1 - r_{ij}^B \cdot r_{ij}^E) - \sum_{j=1}^n (1 - r_{ij}^B)}{|U|}\right)$$
$$= \frac{1}{|U|}\sum_{i=1}^n \left(\frac{\sum_{j=1}^n (r_{ij}^B - r_{ij}^B \cdot r_{ij}^E)}{|U|}\right)$$
$$= \frac{1}{|U|}\sum_{i=1}^n \left(\frac{\sum_{j=1}^n [r_{ij}^B \cdot (1 - r_{ij}^E)]}{|U|}\right) = CH(E|B).$$

Property 7. *If $B_1 \subseteq B_2 \subseteq C$, then $CH(E|B_1) \geq CH(E|B_2)$.*

Proof. Let $B_1 \subseteq B_2 \subseteq C$, so there is $k_{B_1} \supseteq k_{B_2}$. So there is $r_{ij}^{B_1} - r_{ij}^{B_2} \geq 0$.

$$CH(E|B_1) - CH(E|B_2) = \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{|[x_i]_{B_1} \otimes [x_i]_E^c| - |[x_i]_{B_2} \otimes [x_i]_E^c|}{|U|}\right)$$
$$= \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}[r_{ij}^{B_1} \cdot (1 - r_{ij}^E)] - \sum_{j=1}^{n}[r_{ij}^{B_2} \cdot (1 - r_{ij}^E)]}{|U|}\right)$$
$$= \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}[(r_{ij}^{B_1} - r_{ij}^{B_2}) \cdot (1 - r_{ij}^E)]}{|U|}\right) \geq 0.$$

Therefore, $CH(E|B_1) \geq CH(E|B_2)$.

Definition 9. The fuzzy complementary mutual information of B and E is defined as

$$CMI(E; B) = \frac{1}{|U|}\sum_{i=1}^{n}\frac{|[x_i]_B^c \otimes [x_i]_E^c|}{|U|}. \quad (13)$$

$CMI(E; B)$ measures the amount of information that B and E have in common, and it can reflect the correlation between these two attribute subsets.

Property 8. *$\forall B, E \subseteq C$, we have*

(1) $CMI(E; B) = CMI(B; E)$;
(2) $CMI(E; B) = CH(E) - CH(E|B) = CH(B) - CH(B|E)$;
(3) $CMI(E; B) = CH(E) + CH(B) - CH(E, B)$.

Proof. From Definition 9, it is obvious that $CMI(B; E) = CMI(E; B)$, so Eq. (1) holds.

Since $CMI(E; B) = \frac{1}{|U|}\sum_{i=1}^{n}\frac{|[x_i]_B^c \otimes [x_i]_E^c|}{|U|} = \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}[(1-r_{ij}^B)\cdot(1-r_{ij}^E)]}{|U|}\right)$
$= \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}(1-r_{ij}^E) - \sum_{j=1}^{n} r_{ij}^B \cdot (1-r_{ij}^E)}{|U|}\right) = \frac{1}{|U|}\sum_{i=1}^{n}\left(\frac{|[x_i]_E^c| - |[x_i]_B \otimes [x_i]_E^c|}{|U|}\right)$
$= \frac{1}{|U|}\sum_{i=1}^{n}\frac{|[x_i]_E^c|}{|U|} - \frac{1}{|U|}\sum_{i=1}^{n}\frac{|[x_i]_B \otimes [x_i]_E^c|}{|U|} = CH(E) - CH(E|B)$.

Similarly, $CMI(E; B) = CH(B) - CH(B|E)$, so Eq. (2) holds.

From Property 6, there is $CMI(E; B) = CH(E) - CH(E|B) = CH(E) + CH(B) - CH(E, B)$. Therefore, Eq. (3) holds.

In Eq. (2) of Property 8, the complementary mutual information entropy of E and B is the complementary entropy of E minus the complementary conditional entropy of E about B. The complementary mutual information entropy of E and B can be regarded as the fuzzy complementary information contained in E and B. Therefore, it can reflect the degree of correlation between the two attribute subsets E and B, which is completely consistent with the conclusion of other forms of mutual information entropy.

Property 9. *If $B_1 \subseteq B_2 \subseteq C$, then $CMI(E; B_1) \leq CMI(E; B_2)$.*

Proof. Let $B_1 \subseteq B_2 \subseteq C$, so there is $k_{B_1} \supseteq k_{B_2}$. By Properties 6 and 7, there are $CMI(E; B_1) - CMI(E; B_2) = CH(E) - CH(E|B_1) - [CH(E) - CH(E|B_2)] = CH(E|B_2) - CH(E|B_1) \leq 0$.

Therefore, $CMI(E; B_1) \leq CMI(E; B_2)$.

Properties 4, 7, and 9 indicate the monotonicity of fuzzy complementary entropy, complementary conditional entropy, and complementary mutual information and the size of attribute subsets. These properties are essential to design a forward search algorithm. Because it ensures that candidate attribute is added to the selected attribute subset will not reduce (or increase) the information of the new attribute subset. Therefore, the uncertainty measure defined above is used as an evaluation criterion for unsupervised attribute reduction, and its search stop criterion is easy to implement.

3.3. Unsupervised attribute reduction method

The key problem of attribute reduction is to establish the evaluation function of attribute quality. In this section, the significance evaluation functions of attributes based on fuzzy complementary entropy, complementary conditional entropy, and complementary mutual information are constructed respectively.

As pointed out in Section 3.2, the proposed fuzzy complementary entropy can be used to measure the discernibility of relations or attributes. However, it only reflects the importance of attribute subset B. Obviously, this is not sufficient. The fuzzy complementary conditional entropy reflects the uncertainty change of the feature subset E after the fuzzy information entropy of a feature subset B is known. Therefore, it can be used to express the correlation between the conditional attribute subsets B and E.

In order to calculate the fuzzy complementary conditional entropy $CH(E|B)$, we should not try to check all the subsets E of C, because there are $2^{|C|}$ the attribute subsets of C, so we will get $2^{|C|}$ fuzzy complementary conditional entropy. It is impractical to calculate the fuzzy complementary conditional entropy of all subsets, because the time complexity will be exponential with respect to $|C|$. Therefore, in the following we only consider the fuzzy complementary conditional entropy of B relative to some attribute subsets, that is, the fuzzy complementary conditional entropy of all single attribute subsets $\{c_k\}$. For this reason, the average fuzzy complementary conditional entropy is defined as follows.

Definition 10. The average fuzzy complementary conditional entropy of B on all single attribute subsets is defined as

$$ACH(B) = \frac{1}{|C|}\sum_{k=1}^{m} CH(\{c_k\}|B). \quad (14)$$

$ACH(B)$ represents the average value of the conditional entropy of the attribute subset B and all single attributes, which can characterize the correlation between B and all single attributes. The greater the value of $ACH(B)$, the greater the correlation between B and all single attributes.

Property 10. *If $B_1 \subseteq B_2 \subseteq C$, then $ACH(B_1) \geq ACH(B_2)$.*

Proof. It is easy to get from Property 7.

The complementary mutual information entropy of E and B reflects the amount of fuzzy complementary information jointly contained in E and B. Therefore, it reflects the degree of correlation between the two attribute sets E and B. In the same way, the average fuzzy complementary mutual information is defined as follows.

Definition 11. The average fuzzy complementary mutual information of B on all single attribute subsets is defined as

$$ACMI(B) = \frac{1}{|C|}\sum_{k=1}^{m} CMI(\{c_k\}; B). \quad (15)$$

$ACMI(B)$ represents the average value of the mutual information of the attribute subset B and all single attribute subsets, which can characterize the correlation between B and all single attributes. The greater the value of $ACMI(B)$, the greater the correlation between B and all single attributes.

Property 11. *If $B_1 \subseteq B_2$, then $ACMI(B_1) \leq ACMI(B_2)$.*

Proof. It is easy to get from Property 9.

When the data set contains numerical attributes, it can be seen from Eqs. (7) and (8) that when $k_{B^c}(x_i, x_j) = 1$ and $k_{B^n}(x_i, x_j) \to 0$, there are $k_{B^c \cup B^n}(x_i, x_j) \to 0$ and $k_{B^n \cup B^c}(x_i, x_j) \neq 0$. Therefore, it is not possible to directly define attribute reduction by keeping the uncertainty measure unchanged. Instead, the attribute reduction can be defined by the idea of allowing the uncertainty measure to change within a small disturbance, that is, a parameter $\varepsilon > 0$ can be introduced to define the following $\varepsilon-$ attribute reduction.

Definition 12. Given $b \in B$, if $|UM(B) - UM(B - b)| > \varepsilon$, it is said that b is $\varepsilon-$indispensable in B, Otherwise $|UM(B) - UM(B - b)| \leq \varepsilon$, it is said that b is $\varepsilon-$redundant in B, where $UM = CH, ACH$, or $ACMI$. If $\forall b \in B$ is $\varepsilon-$indispensable, it is said that B is $\varepsilon-$independent.

Definition 13. Let $B \subseteq C$, we say that B is a $\varepsilon-$reduction of C, if B satisfies

(1) $|UM(C) - UM(B)| \leq \varepsilon$;
(2) $\forall b \in B, |UM(B) - UM(B - \{b\})| > \varepsilon$ $(r = 1, 2, 3)$,

The above-defined condition (1) requires that the reduction cannot reduce the system's distinguishing ability within $\varepsilon-$tolerance, and the reduction should have the distinguishing ability within $\varepsilon-$tolerance with all the conditional attributes in the system. Condition (2) requires that there is no $\varepsilon-$redundant attribute in a reduction, and all attributes should be $\varepsilon-$indispensable.

According to Properties 4, 10, and 11, a new conditional attribute is added to the information system, the entropy value will increase (or decrease) monotonically, which reflects that the increase (or decrease) of information will lead to the change of the distinguishing ability. Therefore, the three kinds of significance of attributes can be defined as follows.

Definition 14. $\forall c \in C - B$, based on fuzzy complementary entropy, fuzzy complementary conditional entropy, and fuzzy complementary mutual information, the three kinds of significance of c on attribute subset B are defined as, respectively.

$$sig_1(c, B) = CH(B \cup \{c\}) - CH(B); \tag{16}$$
$$sig_2(c, B) = ACH(B) - ACH(B \cup \{c\}); \tag{17}$$
$$sig_3(c, B) = ACMI(B \cup \{c\}) - ACMI(B). \tag{18}$$

The above definition is applicable to unsupervised attribute reduction. $sig_r(c, B)$ $(r = 1, 2, 3)$ is used to measure the change in discernibility introduced by attribute c.

Property 12. *It is easy to prove that the definitions of $sig_2(c, B)$ and $sig_3(c, B)$ are equivalent.*

3.4. Unsupervised attribute reduction algorithm

Based on the above definition of significance, we design a generalized unsupervised heuristic algorithm and analyze its complexity.

Algorithm 1 takes an empty set as the starting point. First, the fuzzy relation matrix for each attribute is calculated. Then, the attribute significances of all remaining attributes are calculated in each iteration. Finally, the attribute with the greatest significance is selected to be added to the reduction set until the significances of all remaining attributes are less than or equal to a given threshold ε.

Algorithm 1: A generalized heuristic unsupervised attribute reduction algorithm

Input: $IS = \langle U, C\rangle$, r, threshold σ, ε, $|C| = m$
Output: A Reduct (Red)
1 $Red \leftarrow \emptyset$, Label$\leftarrow 1$, $B \leftarrow C - Red$;
2 **for** $k \leftarrow 1$ *to* m **do**
3   Compute fuzzy relation matrix $M_{k_{c_k}}$;
4 **end**
5 **while** Label **do**
6   //Let $B = \{c_{k_1}, c_{k_2}, \cdots, c_{k_h}\}$
7   **for** $l \leftarrow 1$ *to* h **do**
8   Compute significance $sig_r(k_l, B)$;
9   **end**
10   Select attribute $c_{k_{l'}}$, s.t. $sig_r(k_{l'}, B)$ with maximum value;
11   **if** $sig_r(k_{l'}, B) > \varepsilon$ **then**
12         $Red \leftarrow Red \cup \{c_{k_{l'}}\}$, $B \leftarrow B - \{c_{k_{l'}}\}$;
13   **else**
14   Label$\leftarrow 0$;
15   **end**
16   **end**
17 **end**
18 **if** $|Red| == m$ **then**
19   **return** $Red \leftarrow Red(1 : m - 1)$;
20   **else**
21   **return** Red;
22   **end**
23 **end**

In Algorithm 1, the number of cycles in Steps 2–4 is m, the number of cycles in Step 3 is $n \times n$, the number of cycles in Steps 5–17 is h, and the number of cycles in Step 8 is m. Thus, the total number of cycles of the algorithm 1 is $m \times n \times n + h \times m$. Therefore, in the worst case, the time complexity of the algorithm 1 is $O(mn^2)$.

3.5. An illustrative example

This section takes the fuzzy complementary mutual information entropy as an example to illustrate the proposed method.

Example 2. The continuation of Example 1. Let $Red = \emptyset$, $B = C$, and $\sigma = 0.1, \varepsilon = 0.001$. For each $c_j \in C$, the fuzzy relation matrices are as follows.

$$M_{k_{c_1}} = \begin{bmatrix} 1 & 0.3604 & 0.0169 & 0 & 0.0001 & 0 \\ 0.3604 & 1 & 0.0001 & 0 & 0.0169 & 0.0001 \\ 0.0169 & 0.0001 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.0169 & 0.3604 \\ 0.0001 & 0.0169 & 0 & 0.0169 & 1 & 0.3604 \\ 0 & 0.0001 & 0 & 0.3604 & 0.3604 & 1 \end{bmatrix};$$

$$M_{k_{c_2}} = \begin{bmatrix} 1 & 0.0169 & 0.3604 & 0.0001 & 0 & 0.0169 \\ 0.0169 & 1 & 0.0001 & 0 & 0 & 0 \\ 0.3604 & 0.0001 & 1 & 0.0169 & 0 & 0.3604 \\ 0.0001 & 0 & 0.0169 & 1 & 0.0169 & 0.3604 \\ 0 & 0 & 0 & 0.0169 & 1 & 0.0001 \\ 0.0169 & 0 & 0.3604 & 0.3604 & 0.0001 & 1 \end{bmatrix};$$

$$M_{k_{c_3}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix};$$

$$M_{k_{c_4}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

According to Definition 6, the fuzzy complementary entropy of each single attribute is calculated as follows.

$CH(c_1) = \frac{1}{|U|}\sum_{i=1}^{n}\frac{|[x_i]^c_{c_1}|}{|U|} = \frac{1}{6}(1-\frac{1.3774}{6})+\frac{1}{6}(1-\frac{1.3775}{6})+\frac{1}{6}(1-\frac{1.0170}{6})+\frac{1}{6}(1-\frac{1.3773}{6})+\frac{1}{6}(1-\frac{1.3943}{6})+\frac{1}{6}(1-\frac{1.7210}{6}) \approx 0.7704$,

Similarly, there are $CH(c_2) \approx 0.7695$, $CH(c_3) \approx 0.5000$ and $CH(c_4) \approx 0.7222$.

Through Eq. (11), the fuzzy complementary joint entropy of each single attribute is calculated as follows.

$CH(c_1,c_1) \approx 0.7704$, $CH(c_1,c_2) \approx 0.8254$, $CH(c_1,c_3) \approx 0.8324$, $CH(c_1,c_4) \approx 0.8333$;
$CH(c_2,c_1) \approx 0.8254$, $CH(c_2,c_2) \approx 0.7695$, $CH(c_2,c_3) \approx 0.8124$, $CH(c_2,c_4) \approx 0.8324$;
$CH(c_3,c_1) \approx 0.8324$, $CH(c_3,c_2) \approx 0.8124$, $CH(c_3,c_3) \approx 0.5000$, $CH(c_3,c_4) \approx 0.7778$;
$CH(c_4,c_1) \approx 0.8333$, $CH(c_4,c_2) \approx 0.8324$, $CH(c_4,c_3) \approx 0.7778$, $CH(c_4,c_4) \approx 0.7222$.

According to Property 8, the fuzzy complementary mutual information of each single attribute is calculated as follows.

$CMI(c_1;c_1) \approx 0.7704$, $CMI(c_1;c_2) \approx 0.7145$, $CMI(c_1;c_3) \approx 0.4380$, $CMI(c_1;c_4) \approx 0.6593$;
$CMI(c_2;c_1) \approx 0.7145$, $CMI(c_2;c_2) \approx 0.7695$, $CMI(c_2;c_3) \approx 0.4571$, $CMI(c_2;c_4) \approx 0.6593$;
$CMI(c_3;c_1) \approx 0.4380$, $CMI(c_3;c_2) \approx 0.4571$, $CMI(c_3;c_3) \approx 0.5000$, $CMI(c_3;c_4) \approx 0.4444$;
$CMI(c_4;c_1) \approx 0.6593$, $CMI(c_4;c_2) \approx 0.6593$, $CMI(c_4;c_3) \approx 0.4444$, $CMI(c_4;c_4) \approx 0.7222$.

By Eq. (15), the average fuzzy complementary mutual information of feature c_k is calculated as follows.

$ACMI(c_1) = \frac{1}{4}\sum_{k=1}^{4} CMI(c_k;c_1) = \frac{1}{4}(0.7292+0.7145+0.4380+0.6593) \approx 0.6456$; $ACMI(c_2) \approx 0.6501$;
$ACMI(c_3) \approx 0.4599$; $ACMI(c_4) \approx 0.6213$.

Therefore, the significance of each attribute is calculated as follows (At this time $ACMI(Red) = 0$).

$sig_3(c_1,Red) = ACMI(Red \cup c_1) - ACMI(Red) = 0.6456$;
$sig_3(c_2,Red) = ACMI(Red \cup c_2) - ACMI(Red) = 0.6501$;
$sig_3(c_3,Red) = ACMI(Red \cup c_3) - ACMI(Red) = 0.4599$;
$sig_3(c_4,Red) = ACMI(Red \cup c_4) - ACMI(Red) = 0.6213$.

Next, the attribute c_2 with the greatest significance is selected to add to Red, and we get $Red = \{c_2\}$, $B = C - \{c_2\} = \{c_1, c_3, c_4\}$.

In the same way, we can calculate

$sig_3(c_1,Red) = ACMI(Red \cup c_1) - ACMI(Red) = 0.6853 - 0.6501 = 0.0352$;
$sig_3(c_3,Red) = ACMI(Red \cup c_3) - ACMI(Red) = 0.6766 - 0.6501 = 0.0265$;
$sig_3(c_4,Red) = ACMI(Red \cup c_4) - ACMI(Red) = 0.6898 - 0.6501 = 0.0397$.

At this point, we can get $Red = \{c_2, c_4\}$, $B = \{c_1, c_3\}$.

In the same way, we can calculate

$sig_3(c_1,Red) = ACMI(Red \cup c_1) - ACMI(Red) = 0.6905 - 0.6898 = 0.0007 < 0.001$;
$sig_3(c_3,Red) = ACMI(Red \cup c_3) - ACMI(Red) = 0.6905 - 0.6898 = 0.0007 < 0.001$.

Therefore, $Red = \{c_2, c_4\}$.

4. Experiments

In this section, in order to verify the feasibility and effectiveness of the proposed method, we conducted a comparative analysis of clustering experiments with some existing algorithms. These comparative algorithms are presented as follows.

(1) Original attributes: The clustering results on original data set with all attribute are used.
(2) Variance-based (Var) method [47]: The attributes are arranged in descending order according to the variance of each attribute, thereby outputting an attribute sequence.
(3) Correlation-based Unsupervised Feature Selection (CUFS) [48]: This algorithm sorts features according to pairwise correlations, and then can get a feature sequence.
(4) Laplacian Score (LS)-based method [49]: algorithm LS calculates the feature significance by the consistency between the feature and the Laplacian matrix, and then a feature sequence can be obtained.
(5) Feature Similarity-based Feature Selection (FSFS) [50]: algorithm FSFS constructs evaluation indicators for features

Table 2
Basic information for data sets.

No.	Data sets	Abbr.	$\|U\|$	$\|C\|$	Classes	Data type
1	Abalone	Abal	4177	8	28	Hybird
2	Bands	Band	531	39	2	Hybird
3	Chess	Chess	3196	36	2	Nominal
4	Contraceptive Method Choice	CMC	1473	9	3	Hybird
5	Credit approval	Cred	690	15	2	Hybird
6	German	Germ	1000	20	2	Hybird
7	Hepatitis	Hepa	155	19	2	Hybird
8	Horse	Horse	368	27	2	Hybird
9	Ionosphere	Iono	351	33	2	Numeric
10	Labor	Labor	57	16	2	Hybird
11	Lymphography	Lymp	148	18	4	Nominal
12	Monks	Monk	432	6	2	Nominal
13	Movement_libras	Move	360	90	15	Numeric
14	Parkinsons	Park	195	22	2	Numeric
15	Primary_tumor	Prim	339	17	21	Nominal
16	Sick	Sick	3772	29	2	Hybird
17	Sonar	Sonar	208	60	2	Numeric
18	Spect	Spect	267	22	2	Nominal
19	Waveform	Wave	5000	21	3	Numeric
20	Wisconsin Breast Cancer	WBC	699	9	2	Numeric
21	Wisconsin Diagnostic Breast Cancer	WDBC	569	31	2	Numeric
22	Winequality_white	Wine	4898	11	7	Numeric
23	Wisconsin Prognostic Breast Cancer	WPBC	198	34	2	Numeric

Table 3
The average number of selected attributes.

Data set	Original attribute	Others	UFRFS	FEUAR	HKCMI
Abal	8	3.0	8.0	8.0	3.0
Band	39	3.7	16.0	19.0	3.7
Chess	36	7.3	32.0	33.0	7.3
CMC	9	3.0	9.0	9.0	3.0
Cred	15	6.7	14.0	14.0	6.7
Germ	20	2.0	20.0	19.0	2.0
Hepa	19	9.0	18.0	19.0	9.0
Horse	27	1.0	21.0	6.0	1.0
Iono	33	25.3	12.0	33.0	25.3
Labor	16	12.7	14.0	14.0	12.7
Lymp	18	8.0	16.0	10.0	8.0
Monk	6	3.0	6.0	6.0	3.0
Move	90	39.7	5.0	90.0	39.7
Park	22	7.0	8.0	22.0	7.0
Prim	17	11.3	17.0	16.0	11.3
Sick	29	15.0	27.0	27.0	15.0
Sonar	60	6.0	8.0	60.0	6.0
Spect	22	8.0	20.0	20.0	8.0
Wave	21	7.0	12.0	21.0	7.0
WBC	9	4.3	9.0	9.0	4.3
WDBC	31	4.0	9.0	31.0	4.0
Wine	11	6.0	10.0	11.0	6.0
WPBC	34	13.3	8.0	34.0	13.3
Average	25.7	9.0	13.9	23.1	9.0

Table 4
The clustering accuracy of reduced data based on k-Means (%).

Data sets	Original data	Var	CUFS	LS	FSFS	USFSM	FR	UFRFS	FEUAR	HKCMI	(ε, σ)
Abal	15.14 ± 0.74	19.36 ± 0.04	19.34 ± 0.04	19.35 ± 0.04	**19.37 ± 0.05**	14.13 ± 0.47	12.60 ± 0.35	15.57 ± 0.29	15.46 ± 0.42	19.34 ± 0.03	$(10^{-2}, 0.4)$
Band	57.08 ± 2.58	55.29 ± 2.03	54.88 ± 5.06	58.10 ± 4.46	63.99 ± 0.36	55.63 ± 2.10	**64.22 ± 0.00**	54.24 ± 1.19	57.18 ± 4.19	62.86 ± 0.19	$(10^{-5}, 0.1)$
Chess	53.24 ± 3.64	56.99 ± 3.95	**57.39 ± 6.14**	54.01 ± 0.88	52.82 ± 1.29	55.25 ± 7.14	52.50 ± 2.82	54.52 ± 4.55	52.25 ± 2.44	56.57 ± 4.57	$(10^{-3}, –)$
CMC	38.87 ± 1.17	39.54 ± 0.57	37.01 ± 0.73	38.91 ± 0.47	44.15 ± 1.42	40.26 ± 1.08	39.04 ± 1.07	40.51 ± 1.92	40.06 ± 1.74	**45.89 ± 0.00**	$(10^{-2}, 0.1)$
Cred	75.43 ± 8.29	71.35 ± 12.65	67.49 ± 8.24	76.97 ± 11.48	67.38 ± 9.59	59.35 ± 2.70	78.65 ± 4.93	**78.75 ± 5.98**	68.01 ± 14.17	78.20 ± 4.80	$(10^{-4}, 0.1)$
Germ	53.30 ± 3.66	52.93 ± 1.31	**67.58 ± 1.40**	54.75 ± 5.41	63.78 ± 4.60	67.15 ± 0.05	64.11 ± 4.61	55.89 ± 5.94	53.43 ± 3.89	61.67 ± 4.64	$(10^{-2}, 0.7)$
Hepa	61.29 ± 0.00	61.29 ± 0.00	58.58 ± 0.27	59.87 ± 2.08	58.45 ± 5.43	**64.58 ± 4.81**	63.68 ± 2.60	60.65 ± 0.00	60.65 ± 2.04	61.29 ± 0.00	$(10^{-4}, 1.0)$
Horse	53.53 ± 0.00	53.26 ± 0.00	55.60 ± 0.14	53.26 ± 0.00	53.26 ± 0.00	60.00 ± 0.64	51.63 ± 0.00	56.47 ± 5.04	62.04 ± 3.17	**66.58 ± 0.00**	$(10^{-2}, 0.1)$
Iono	71.23 ± 0.00	70.37 ± 0.00	70.85 ± 3.61	70.09 ± 0.00	65.73 ± 4.86	58.09 ± 6.21	69.52 ± 0.00	**72.11 ± 3.65**	71.23 ± 0.00	71.62 ± 0.65	$(10^{-4}, 0.1)$
Labor	75.44 ± 7.58	78.60 ± 5.78	81.93 ± 5.43	78.60 ± 6.66	81.75 ± 11.73	73.86 ± 12.74	70.18 ± 11.00	78.07 ± 11.67	79.30 ± 15.42	**82.11 ± 12.56**	$(10^{-5}, 0.3)$
Lymp	46.28 ± 6.62	45.00 ± 6.72	50.27 ± 5.62	44.93 ± 4.77	42.43 ± 5.23	45.41 ± 5.08	45.61 ± 2.09	49.59 ± 2.74	48.58 ± 2.98	**53.78 ± 5.02**	$(10^{-4}, –)$
Monk	57.87 ± 11.86	50.37 ± 0.57	54.44 ± 10.16	56.25 ± 6.91	54.79 ± 5.18	62.22 ± 9.17	51.02 ± 1.92	55.56 ± 6.35	52.31 ± 3.14	**64.72 ± 8.72**	$(10^{-2}, –)$
Move	44.33 ± 2.89	44.36 ± 1.32	44.42 ± 1.91	43.56 ± 1.61	44.36 ± 1.34	43.86 ± 3.05	43.61 ± 2.24	30.44 ± 1.30	44.25 ± 1.72	**45.44 ± 1.32**	$(10^{-5}, 0.9)$
Park	63.08 ± 0.00	73.74 ± 1.49	50.77 ± 0.00	56.41 ± 6.65	**76.46 ± 0.16**	68.62 ± 0.22	55.38 ± 0.00	76.36 ± 0.85	63.08 ± 0.00	76.31 ± 1.05	$(10^{-3}, 0.2)$
Prim	28.82 ± 1.44	27.99 ± 1.70	26.25 ± 1.47	29.32 ± 1.52	**31.09 ± 1.73**	29.12 ± 1.74	28.41 ± 1.53	29.17 ± 1.07	29.73 ± 1.83	29.44 ± 2.26	$(10^{-5}, –)$
Sick	80.38 ± 8.69	74.51 ± 12.86	72.05 ± 2.57	76.35 ± 12.19	78.43 ± 10.92	75.32 ± 8.24	75.09 ± 11.50	73.83 ± 7.12	77.59 ± 9.13	**80.38 ± 9.81**	$(10^{-5}, 0.2)$
Sonar	53.80 ± 1.05	63.46 ± 0.00	53.85 ± 0.00	52.40 ± 0.00	55.77 ± 0.00	54.18 ± 1.24	56.63 ± 1.58	56.25 ± 0.00	54.66 ± 1.20	**64.71 ± 0.46**	$(10^{-2}, 0.1)$
Spect	56.33 ± 0.32	59.48 ± 4.75	62.02 ± 3.87	55.96 ± 2.68	55.99 ± 3.22	61.42 ± 0.00	57.98 ± 4.59	56.67 ± 0.95	57.12 ± 1.09	**62.10 ± 2.61**	$(10^{-3}, –)$
Wave	50.12 ± 0.00	51.67 ± 0.04	39.21 ± 0.12	51.81 ± 0.17	34.59 ± 0.04	34.36 ± 0.14	**52.18 ± 0.17**	50.58 ± 0.00	50.12 ± 0.00	51.88 ± 0.19	$(10^{-2}, 0.5)$
WBC	95.85 ± 0.00	95.14 ± 0.00	91.27 ± 0.00	95.14 ± 0.00	94.28 ± 0.00	94.23 ± 0.07	92.13 ± 0.00	95.85 ± 0.00	95.85 ± 0.00	**95.99 ± 0.00**	$(10^{-2}, 0.4)$
WDBC	92.79 ± 0.00	93.15 ± 0.00	73.01 ± 3.67	89.81 ± 0.00	72.30 ± 0.15	72.93 ± 5.19	86.29 ± 0.00	88.51 ± 9.11	92.79 ± 0.00	**95.08 ± 0.00**	$(10^{-2}, 0.2)$
Wine	25.86 ± 1.36	27.75 ± 2.29	24.32 ± 0.95	28.04 ± 1.18	24.77 ± 1.06	26.43 ± 1.39	21.60 ± 1.26	**28.71 ± 2.39**	28.12 ± 1.80	28.43 ± 2.69	$(10^{-4}, 0.6)$
WPBC	60.10 ± 0.00	60.10 ± 0.00	60.45 ± 6.81	59.60 ± 0.00	58.43 ± 4.65	55.51 ± 2.37	60.10 ± 0.00	61.26 ± 9.20	60.10 ± 0.00	**65.20 ± 2.36**	$(10^{-3}, 0.3)$
Average	56.96 ± 2.69	57.64 ± 2.52	55.35 ± 2.97	56.67 ± 3.01	56.28 ± 3.18	55.30 ± 3.30	56.18 ± 2.36	57.37 ± 3.54	57.13 ± 3.06	**61.72 ± 2.78**	–

Table 5
The clustering accuracy of reduced data based on k-Medoids (%).

Data sets	Original data	Var	CUFS	LS	FSFS	USFSM	FR	UFRFS	FEUAR	HKCMI	(ε, σ)
Abal	15.78 ± 0.75	**19.30 ± 0.00**	**19.30 ± 0.00**	**19.30 ± 0.00**	**19.30 ± 0.00**	14.39 ± 0.58	12.24 ± 0.53	15.25 ± 0.64	15.69 ± 0.69	**19.30 ± 0.00**	$(10^{-2}, 0.4)$
Band	56.50 ± 0.00	56.50 ± 0.00	53.39 ± 0.89	56.50 ± 0.00	60.92 ± 1.89	55.20 ± 2.09	**64.22 ± 0.00**	54.80 ± 0.00	56.31 ± 0.00	**64.11 ± 0.10**	$(10^{-5}, 0.3)$
Chess	56.75 ± 3.54	56.23 ± 2.98	54.99 ± 4.71	56.22 ± 2.98	**57.12 ± 3.63**	55.76 ± 4.44	55.83 ± 4.17	51.87 ± 0.92	54.27 ± 3.26	55.36 ± 3.30	$(10^{-4}, –)$
CMC	39.29 ± 0.28	39.51 ± 0.00	37.20 ± 0.00	39.16 ± 0.07	41.47 ± 2.44	40.02 ± 0.85	40.21 ± 0.98	39.24 ± 0.29	39.35 ± 0.26	**45.33 ± 0.71**	$(10^{-2}, 0.1)$
Cred	74.75 ± 9.88	76.88 ± 8.02	73.33 ± 0.00	74.75 ± 9.88	53.48 ± 0.00	59.78 ± 3.52	**85.51 ± 0.00**	74.75 ± 9.88	77.29 ± 6.74	**79.42 ± 0.00**	$(10^{-4}, 0.3)$
Germ	54.00 ± 0.00	52.20 ± 0.00	66.60 ± 0.00	52.20 ± 0.00	**66.93 ± 0.22**	66.50 ± 0.00	64.80 ± 0.00	54.00 ± 0.00	53.70 ± 0.00	65.20 ± 0.00	$(10^{-2}, 0.7)$
Hepa	61.94 ± 0.00	62.90 ± 1.02	58.06 ± 0.00	58.71 ± 0.00	54.84 ± 0.00	55.48 ± 0.00	60.77 ± 1.00	60.52 ± 2.99	61.94 ± 0.00	**64.52 ± 3.33**	$(10^{-3}, 0.5)$
Horse	53.80 ± 0.00	53.26 ± 0.00	55.60 ± 0.14	53.26 ± 0.00	53.26 ± 0.00	59.40 ± 0.34	51.63 ± 0.00	53.80 ± 0.00	65.33 ± 2.64	**66.58 ± 0.00**	$(10^{-2}, 0.1)$
Iono	70.94 ± 0.00	**71.23 ± 0.00**	**71.23 ± 0.00**	**71.23 ± 0.00**	68.66 ± 0.00	70.94 ± 0.00	70.94 ± 0.00	70.83 ± 3.53	70.94 ± 0.00	**71.23 ± 0.00**	$(10^{-4}, 0.7)$
Labor	73.68 ± 3.70	77.19 ± 0.00	77.19 ± 0.00	77.19 ± 0.00	**87.72 ± 0.00**	67.02 ± 2.72	**77.19 ± 0.00**	75.44 ± 0.00	74.56 ± 2.77	77.19 ± 0.00	$(10^{-5}, 0.5)$
Lymp	45.95 ± 0.00	44.86 ± 3.65	52.70 ± 0.00	42.70 ± 3.75	45.41 ± 0.70	36.55 ± 1.08	45.14 ± 0.43	46.22 ± 1.69	50.00 ± 0.00	**59.46 ± 0.00**	$(10^{-4}, –)$
Monk	56.94 ± 0.00	50.28 ± 0.59	50.00 ± 0.00	56.94 ± 0.00	56.94 ± 0.00	**67.78 ± 9.18**	50.00 ± 0.00	56.94 ± 0.00	56.94 ± 0.00	**63.98 ± 10.77**	$(10^{-2}, –)$
Move	46.36 ± 1.05	35.83 ± 1.84	39.44 ± 1.58	35.94 ± 1.21	39.64 ± 1.06	35.78 ± 2.32	30.78 ± 1.39	31.14 ± 0.72	45.69 ± 1.17	**47.11 ± 2.02**	$(10^{-3}, 0.4)$
Park	67.69 ± 0.00	71.28 ± 0.00	55.38 ± 0.00	72.31 ± 0.00	69.54 ± 1.73	67.69 ± 0.00	56.92 ± 0.00	68.21 ± 0.00	67.69 ± 0.00	**78.56 ± 4.11**	$(10^{-3}, 0.1)$
Prim	30.06 ± 1.11	29.65 ± 0.54	26.58 ± 0.80	29.06 ± 1.01	**30.97 ± 1.03**	28.47 ± 0.74	28.38 ± 1.01	29.97 ± 1.26	30.21 ± 0.99	29.68 ± 0.80	$(10^{-5}, –)$
Sick	79.90 ± 9.97	79.90 ± 9.97	77.57 ± 6.47	**85.69 ± 6.10**	73.98 ± 9.34	73.80 ± 7.74	79.03 ± 9.26	82.90 ± 7.74	84.29 ± 7.13	80.38 ± 7.28	$(10^{-5}, 0.4)$
Sonar	54.81 ± 0.00	58.17 ± 0.00	50.96 ± 0.00	60.10 ± 0.00	54.04 ± 1.86	58.08 ± 3.85	55.67 ± 0.30	53.37 ± 0.00	54.81 ± 0.00	**64.90 ± 0.00**	$(10^{-5}, 0.1)$
Spect	52.06 ± 0.00	56.18 ± 0.00	53.93 ± 0.00	56.18 ± 0.00	**61.57 ± 5.61**	60.30 ± 0.00	54.31 ± 0.00	50.56 ± 0.00	50.56 ± 0.00	60.67 ± 1.93	$(10^{-3}, –)$
Wave	52.89 ± 0.30	55.16 ± 6.32	57.59 ± 7.84	53.06 ± 0.69	54.55 ± 0.74	**60.75 ± 5.58**	54.89 ± 2.74	52.15 ± 2.42	53.02 ± 0.51	**58.34 ± 0.00**	$(10^{-4}, 0.2)$
WBC	95.57 ± 0.00	95.85 ± 0.00	95.42 ± 0.00	95.85 ± 0.00	93.28 ± 0.00	94.56 ± 0.00	92.06 ± 0.08	95.57 ± 0.00	95.57 ± 0.00	**96.14 ± 0.00**	$(10^{-3}, 0.3)$
WDBC	92.27 ± 0.00	92.97 ± 0.00	72.76 ± 0.00	89.81 ± 0.00	71.25 ± 0.22	75.04 ± 0.00	86.29 ± 0.00	90.51 ± 0.00	92.27 ± 0.00	**94.90 ± 0.00**	$(10^{-2}, 0.2)$
Wine	28.49 ± 2.68	27.90 ± 1.75	25.23 ± 1.11	27.53 ± 0.61	24.59 ± 1.13	26.72 ± 2.40	22.06 ± 0.88	27.68 ± 2.00	26.86 ± 1.93	**29.23 ± 2.45**	$(10^{-4}, 0.6)$
WPBC	67.93 ± 5.06	69.60 ± 4.69	65.66 ± 1.06	68.28 ± 8.30	70.20 ± 0.00	60.00 ± 3.77	67.07 ± 10.73	70.71 ± 0.00	69.85 ± 4.64	**72.22 ± 0.00**	$(10^{-5}, 0.2)$
Average	57.75 ± 1.67	57.95 ± 1.80	56.09 ± 1.07	57.91 ± 1.50	56.94 ± 1.37	56.09 ± 2.23	56.78 ± 1.46	56.80 ± 1.48	58.57 ± 1.42	**62.77 ± 1.60**	–

through feature similarity to perform feature selection. It outputs a feature sequence sorted according to feature scores.

(6) Unsupervised Spectral Feature Selection Method (USFSM) [51]: The algorithm quantifies feature correlation through the combined use of kernels and new Spectral-based feature evaluation measures. It outputs a feature ranking.
(7) Feature Ranking (FR) [52]: All features are arranged in descending order of their average shared information.
(8) Unsupervised Fuzzy Rough-based Feature Selection (UFRFS) [45]: This algorithm uses fuzzy positive region to define the significance of attributes, and then outputs a feature subset.
(9) Fuzzy Entropy-based Unsupervised Attribute Reduction (FEUAR) [9]: algorithm FEUAR uses fuzzy entropy to construct attribute significance to perform unsupervised attribute reduction. Finally, a reduced subset is also obtained.
(10) The Hybrid-Kernel based fuzzy Complementary Mutual Information (HKCMI) method is proposed in this paper.

We first compare the number of reduction sets. Then, the accuracy of clustering based on these comparative algorithms is given. Finally, the influence of two parameters on the proposed algorithm is discussed.

4.1. Experimental description

Three clustering algorithms are used to evaluate the clustering effects of these comparison algorithms, including k-Means, k-Medoids, and Fuzzy C-Means (FCM) clustering. The number of clusters is set to the true number of decision classes in each data set. The clustering algorithm returns the prediction labels of the data set. The clustering accuracy is used to evaluate the performance of clustering [53]. Because these clustering algorithms use random parameters, we repeat the experiment 10 times. Finally, the average and standard deviation of the clustering accuracy are calculated as the final results.

Twenty-three data sets selected from UCI machine learning library[1] are used for experiments. The basic information of these data sets is shown in Table 2. For the missing values in some

[1] http://archive.ics.uci.edu/ml.

Table 6
The clustering accuracy of reduced data based on FCM (%).

Data sets	Original data	Var	CUFS	LS	FSFS	USFSM	FR	UFRFS	FEUAR	HKCMI	(ε, σ)
Abal	15.30 ± 0.94	14.68 ± 0.68	15.02 ± 0.53	14.87 ± 0.76	15.30 ± 0.69	12.09 ± 0.41	14.41 ± 0.54	14.54 ± 1.13	14.49 ± 0.93	**15.50 ± 0.91**	$(10^{-3}, 0.2)$
Band	55.40 ± 0.08	56.50 ± 0.00	52.54 ± 0.00	56.31 ± 0.60	60.45 ± 0.00	54.24 ± 0.00	64.22 ± 0.00	54.61 ± 0.00	55.97 ± 1.67	**62.71 ± 0.00**	$(10^{-5}, 0.1)$
Chess	52.36 ± 2.72	52.79 ± 2.03	53.07 ± 0.00	54.66 ± 0.00	50.22 ± 0.00	51.60 ± 0.00	51.16 ± 0.00	51.66 ± 2.42	52.93 ± 3.34	**55.10 ± 0.00**	$(10^{-2}, –)$
CMC	39.73 ± 0.70	39.02 ± 0.42	36.86 ± 0.00	39.24 ± 0.00	42.23 ± 0.00	40.04 ± 0.07	40.19 ± 0.00	39.83 ± 0.91	39.95 ± 0.94	**45.82 ± 0.00**	$(10^{-2}, 0.1)$
Cred	80.43 ± 0.00	80.14 ± 0.00	53.48 ± 0.00	79.42 ± 0.00	53.48 ± 0.00	57.54 ± 0.00	54.78 ± 0.00	80.58 ± 0.00	80.58 ± 0.00	**83.65 ± 0.06**	$(10^{-3}, 0.1)$
Germ	52.73 ± 0.76	52.20 ± 0.00	**66.60 ± 0.00**	52.20 ± 0.00	63.70 ± 0.00	**67.05 ± 0.05**	64.80 ± 0.00	52.52 ± 0.84	52.60 ± 0.35	61.70 ± 0.00	$(10^{-2}, 0.7)$
Hepa	61.94 ± 0.00	60.65 ± 0.00	**64.52 ± 0.00**	60.65 ± 0.00	**64.52 ± 0.00**	**66.45 ± 0.00**	64.52 ± 0.00	61.94 ± 0.00	61.94 ± 0.00	62.58 ± 0.00	$(10^{-4}, 0.6)$
Horse	51.88 ± 0.09	26.68 ± 34.10	55.71 ± 0.00	33.29 ± 34.80	46.49 ± 31.90	59.24 ± 0.00	64.84 ± 4.64	52.99 ± 0.00	60.33 ± 0.00	**66.58 ± 0.00**	$(10^{-2}, 0.1)$
Iono	70.94 ± 0.00	70.94 ± 0.00	70.94 ± 0.00	70.94 ± 0.00	68.38 ± 0.00	**71.23 ± 0.00**	70.94 ± 0.00	67.52 ± 0.00	70.94 ± 0.00	**71.23 ± 0.00**	$(10^{-5}, 0.2)$
Labor	81.75 ± 7.81	82.81 ± 9.71	82.63 ± 5.39	77.54 ± 11.66	84.56 ± 5.08	78.25 ± 9.90	77.89 ± 6.63	82.81 ± 5.35	83.68 ± 4.97	**86.84 ± 6.93**	$(10^{-5}, 0.4)$
Lymp	48.24 ± 3.19	51.82 ± 0.46	45.27 ± 2.85	50.88 ± 0.33	37.43 ± 0.35	52.70 ± 0.00	34.46 ± 0.00	49.26 ± 0.50	59.05 ± 4.39	**67.43 ± 6.37**	$(10^{-5}, –)$
Monk	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	62.78 ± 8.61	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	**66.67 ± 8.65**	$(10^{-2}, –)$
Move	15.89 ± 1.07	32.50 ± 0.91	25.42 ± 1.11	33.61 ± 1.26	**41.42 ± 1.25**	33.33 ± 1.68	33.33 ± 2.23	31.58 ± 1.08	16.61 ± 1.54	38.61 ± 0.87	$(10^{-3}, 0.9)$
Park	67.18 ± 0.00	72.31 ± 0.00	56.92 ± 0.00	55.38 ± 0.00	67.18 ± 0.00	67.69 ± 0.00	51.28 ± 0.00	**75.38 ± 0.00**	67.18 ± 0.00	74.87 ± 0.00	$(10^{-3}, 0.2)$
Prim	24.34 ± 1.89	23.24 ± 8.08	19.23 ± 10.06	**29.68 ± 1.56**	28.38 ± 0.30	26.55 ± 0.24	25.25 ± 1.63	24.93 ± 1.74	24.31 ± 1.31	27.43 ± 0.00	$(10^{-2}, –)$
Sick	63.88 ± 0.30	68.32 ± 0.00	73.65 ± 0.00	**84.04 ± 0.00**	68.32 ± 0.00	**93.88 ± 0.00**	93.88 ± 0.00	63.97 ± 0.00	63.97 ± 0.00	73.65 ± 0.00	$(10^{-2}, 0.3)$
Sonar	55.63 ± 0.82	63.46 ± 0.00	54.33 ± 0.00	50.48 ± 0.00	55.77 ± 0.00	54.81 ± 0.00	59.13 ± 0.00	56.73 ± 0.00	55.34 ± 0.15	**65.87 ± 0.00**	$(10^{-2}, 0.1)$
Spect	60.49 ± 0.59	61.80 ± 0.00	**65.17 ± 0.00**	**65.17 ± 0.00**	56.55 ± 0.00	61.42 ± 0.00	56.55 ± 0.00	60.22 ± 0.95	59.93 ± 0.00	64.79 ± 0.00	$(10^{-3}, –)$
Wave	49.30 ± 0.00	**51.44 ± 0.00**	39.36 ± 0.00	**51.44 ± 0.00**	34.51 ± 0.02	34.28 ± 0.05	**52.38 ± 0.05**	50.08 ± 0.01	49.30 ± 0.00	**51.44 ± 0.00**	$(10^{-2}, 0.5)$
WBC	95.28 ± 0.00	94.99 ± 0.00	94.13 ± 0.00	94.99 ± 0.00	94.28 ± 0.00	93.85 ± 0.00	91.70 ± 0.00	95.28 ± 0.00	95.28 ± 0.00	**95.85 ± 0.00**	$(10^{-2}, 0.6)$
WDBC	92.79 ± 0.00	93.32 ± 0.00	72.93 ± 0.00	90.33 ± 0.00	71.53 ± 0.00	76.27 ± 0.00	86.99 ± 0.00	91.74 ± 0.00	92.79 ± 0.00	**94.90 ± 0.00**	$(10^{-2}, 0.2)$
Wine	30.73 ± 0.65	30.70 ± 1.42	28.62 ± 0.60	26.32 ± 0.78	**34.54 ± 2.25**	31.34 ± 1.63	31.27 ± 0.75	30.83 ± 1.07	29.65 ± 0.86	31.09 ± 0.55	$(10^{-3}, 0.2)$
WPBC	58.08 ± 0.00	62.12 ± 0.00	60.66 ± 0.60	**67.68 ± 0.00**	62.68 ± 1.76	60.10 ± 0.00	61.62 ± 0.00	59.14 ± 3.59	58.18 ± 0.32	63.23 ± 2.34	$(10^{-3}, 1.0)$
Average	55.40 ± 0.94	56.19 ± 2.51	53.78 ± 0.92	56.05 ± 2.25	54.43 ± 1.90	56.81 ± 0.98	56.33 ± 0.72	56.44 ± 0.85	56.30 ± 0.90	**62.07 ± 1.16**	–

data sets, this paper uses the maximum probability value method to fill in them. In addition, all numerical attribute values are normalized to the interval [0,1] through min–max normalization.

In the experiment, there are two parameters σ and ε in algorithm HKCMI. The parameter σ is introduced to control the fuzzy similarity between objects, which has a great impact on the performance of the algorithm. The parameter ε is used as the stopping condition of the algorithm. For a given data set, if the value of the parameter ε decreases, the number of selected attributes will not decrease. Generally speaking, different σ and ε will lead to different clustering accuracy. Therefore, we adjust the parameter value to make σ change from 0 to 1 with step size 0.1 and make $\varepsilon \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Finally, an optimal attribute subset is selected for each data set. For algorithms Var, CUFS, LS, and FSFS, 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 experimental settings in [53], when algorithm LS calculates the neighbor graph matrix, the neighborhood size is fixed at 5. For algorithm FSFS, the feature similarity is calculated through the "maximum information compression index". Generally, algorithms Var, CUFS, LS, FSFS, USFSM, and FR output a sequence of attributes. To facilitate the comparison with algorithm HKCMI, we select the attribute subsequence equal to the number of attributes selected by algorithm HKCMI.

4.2. Experimental result

Table 3 gives a comparison of the average size of selected attributes under different algorithms. Through Table 3, we can see that these reduction methods can effectively remove attributes. In most cases, the number of attributes selected by algorithm HKCMI is less than or equal to the other six algorithms. For example, for the data set Germ, the number of attributes selected by algorithm HKCMI is 2, which is significantly smaller than the number of attributes selected by algorithms UFRFS and FEUAR. In addition, for the average value, algorithm HKCMI is also significantly smaller than or equal to other algorithms. This shows that the proposed algorithm can remove redundant attributes more effectively.

Tables 4–6 respectively show the clustering accuracy of the original data and the reduced data set based on these seven algorithms. The number in bold highlights the highest cluster accuracy in the reduced data set. From Tables 4–6, it can be seen that algorithm HKCMI can improve or maintain the clustering accuracy of the original data on all data sets. Among 69 records in Tables 4–6, algorithm HKCMI has 45 records to achieve the best clustering accuracy. However, for algorithms Var, CUFS, LS, FSFS, UFRFS, and FEUAR, only 4, 7, 8, 13, 3, and 0 records achieve the best accuracy. What is more, the average clustering accuracy of algorithm HKCMI is better than all other attribute reduction algorithms in the three clustering algorithms.

In summary, algorithm HKCMI achieves better clustering performance. It can obtain a relatively small subset of attributes and improve or maintain the accuracy of clustering. These 23 data sets include nominal, numerical, and mixed attribute data sets. Therefore, algorithm HKCMI is suitable for attribute reduction of multiple attribute types on clustering tasks.

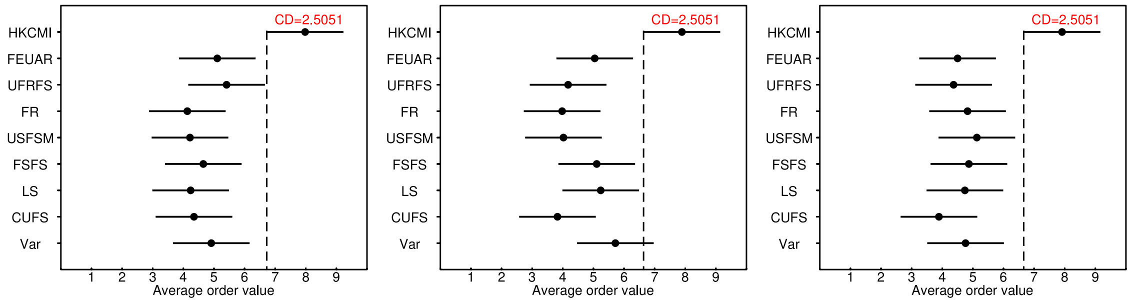
What is more, Friedman's test [54] and Nemenyi's post-hoc test [55] are applied to evaluate the statistical significance of the results. Before using Friedman's test, the accuracy of each algorithm on all data sets is sorted from low to high, and the sequence number is assigned $(1, 2, \ldots)$. Among 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 data sets, and let r_i represent the average ordinal value of the ith algorithm, then Friedman's test is calculated as follows.

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(M-1)-\tau_{\chi^2}} \text{ and}$$
$$\tau_{\chi^2} = \frac{12N}{M(M+1)}\left(\sum_{i=1}^{M} r_i^2 - \frac{M(M+1)^2}{4}\right). \tag{19}$$

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.

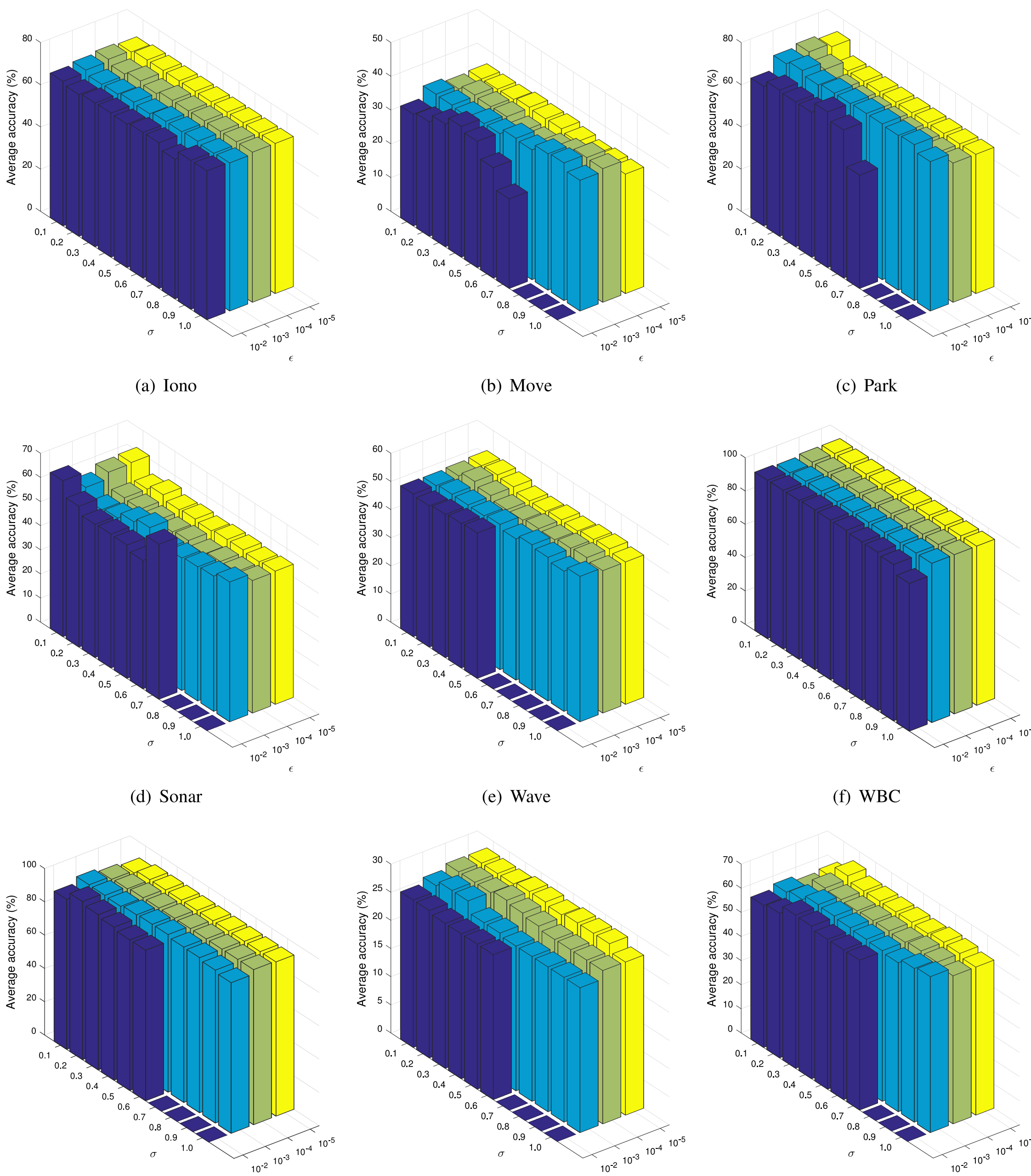
$$CD_\alpha = q_\alpha \sqrt{\frac{M(M+1)}{6N}}, \tag{20}$$

where q_α is the critical value of Tukey's distribution, which can be found in [55].



(a) k-Means (b) k-Medoids (c) FCM

Fig. 1. Nemenyi's test figures on algorithms k-Means, k-Medoids, and FCM.



(a) Iono (b) Move (c) Park

(d) Sonar (e) Wave (f) WBC

(g) WDBC (h) Wine (i) WPBC

Fig. 2. The average accuracy varies with the parameters σ and ε on numerical attribute data sets.

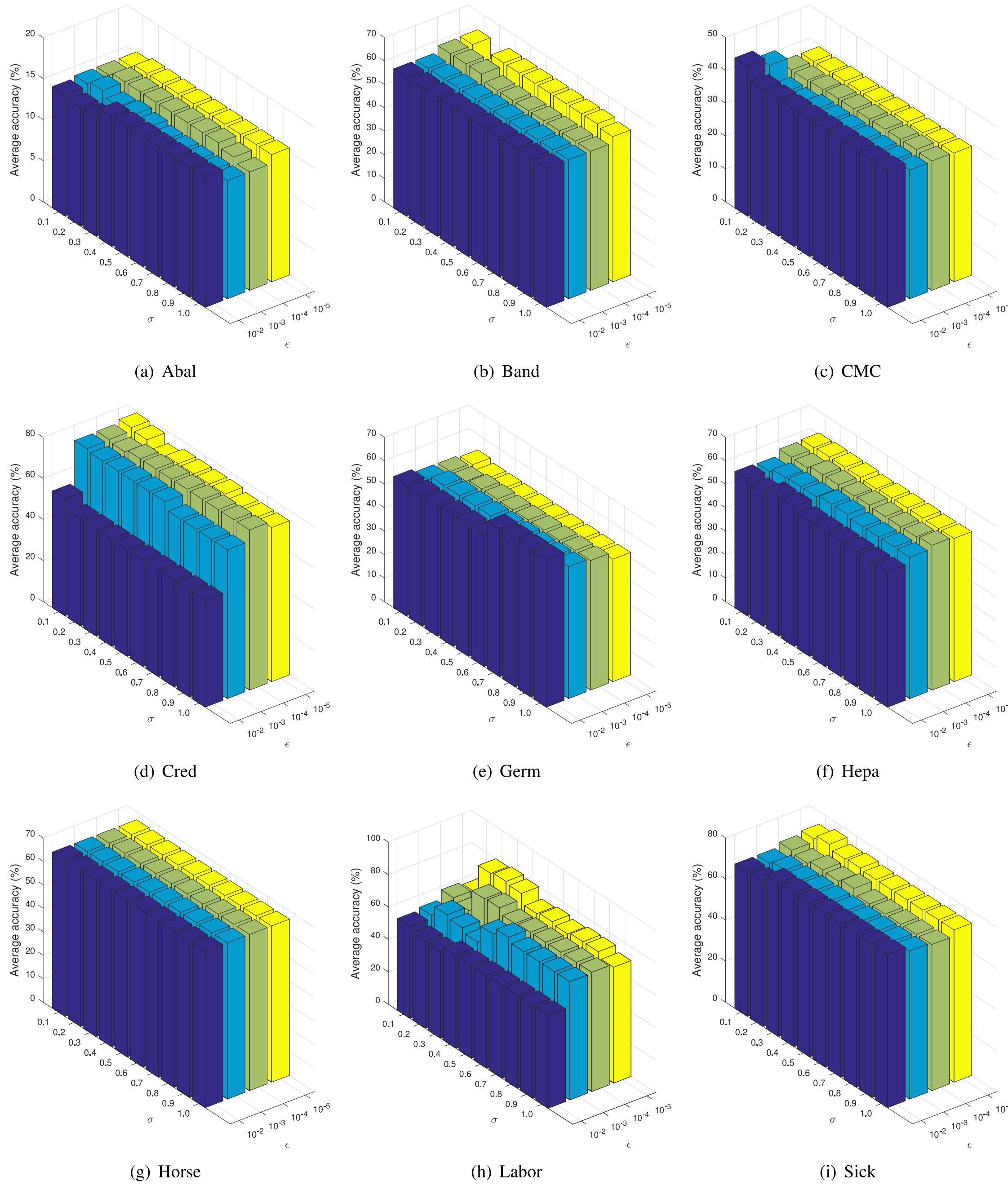


Fig. 3. The average accuracy varies with the parameters σ and ε on hybrid attribute data sets.

Further, Nemenyi's test figure is used to more intuitively represent the significant differences between the two algorithms [3]. 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.

Form Tables 4–6, we can get $M = 9$ and $N = 23$, the τ_F distribution has 8 and 176 degrees of freedom. According to Friedman's test, τ_F of different learning algorithms and critical value (significance level α is 0.05) are shown in Table 7. According to Table 7, when $\alpha = 0.05$, each value of τ_F on algorithms k-Means, k-Mediods, and FCM is greater than the critical value 1.9913. Therefore, the null hypothesis that "all algorithms have the same performance" is rejected on algorithms k-Means, k-Mediods, and FCM. It shows that the performance of all feature selection algorithms is significantly different on algorithms k-Means, k-Mediods, and FCM. At this time, a post-hoc test needs to be used to further distinguish these feature selection algorithms.

Table 7
τ_F on algorithms k-Means, k-Mediods, and FCM.

Algorithms	τ_F	Critical value ($\alpha = 0.05$)
k-Means	5.2468	
k-Mediods	6.0891	1.9913
FCM	4.6947	

For significance level $\alpha = 0.05$, the corresponding critical distance $CD_{0.05} = 2.5051$ 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 HKCMI is statistically significantly different from most other algorithms. The specific analyses are as follows.

For example, it can be seen from Figs. 1(a) and 1(c) that the horizontal line segments of algorithm HKCMI and other algorithms have no overlapping area on algorithms k-Means and FCM, which shows that algorithm HKCMI and other algorithms are statistically significantly different on algorithms k-Means and FCM. Fig. 1(b) demonstrates that algorithm HKCMI is statistically better than algorithms CUFS, LS, FSFS, USFSM, FR, UFRFS, and FEUAR on algorithm k-Medoids, respectively. However, there is no consistent evidence to indicate the statistical differences from algorithm Var on algorithm k-Medoids. Besides, the average order value of three learning algorithms in Fig. 1 is also relatively large, which also shows the effectiveness of algorithm HKCMI.

4.3. Experimental parameter

In order to analyze the sensitivity of the performance of the proposed algorithm HKCMI to the parameters σ and ε, the average clustering accuracy of the three clustering algorithms varies with the parameters σ and ε, as shown in Figs. 2–3. Obviously, it can be seen that different parameter combinations may lead to different clustering accuracy. Through Figs. 2–3, we can see that when $\varepsilon = 10^{-2}$, with the increase of σ, the final average accuracy becomes 0 on data sets Move, Park, Sonar, Wave, WDBC, Wine, and WPBC. The proposed algorithm HKCMI may unable to make effective attribute selection when the parameter σ is too small. On data sets Iono, WBC, Band, and Horse, the clustering average accuracy of the proposed algorithm HKCMI is relatively stable. For most data sets, the optimal value can be obtained under a variety of parameter combinations. For each data set, we can choose the appropriate combination of parameters σ and ε to achieve better performance. For example, on the data set CMC, when $\sigma = 0.1$ and $\varepsilon = 10^{-2}$, the proposed algorithm HKCMI achieves the best performance.

Through the above analysis, it can be seen that the experimental performance has a certain sensitivity to the parameters σ and ε. Therefore, it is necessary to adjust the parameters in the proposed method. However, under the condition of the appropriate parameter combination value, algorithm HKCMI can obtain better results in most cases. In summary, algorithm HKCMI is feasible and effective for the attribute selection of numerical and mixed attribute clustering algorithms.

5. Conclusion

This paper proposes a new fuzzy complementary entropy metric by using the hybrid-kernel function. Corresponding fuzzy complementary conditional entropy and fuzzy complementary mutual information are proved to be monotonic about attributes. Furthermore, three evaluation criteria for unsupervised hybrid attribute reduction based on fuzzy complementary entropy, complementary conditional entropy, and fuzzy complementary mutual information are constructed respectively. Among them, the one based on complementary conditional entropy and fuzzy complementary mutual information is proved to be equivalent. The corresponding generalized unsupervised heuristic algorithm is designed. Finally, the method based on fuzzy complementary mutual information and the existing methods are compared and analyzed on 23 public data sets. The experimental results show that the proposed method is an effective scheme for reducing hybrid attributes.

In the algorithm proposed in this paper, the remaining attributes are directly removed. In fact, if the complementarity of these remaining attributes is further considered in the attribute reduction process, it will theoretically make the reduction performance better. Therefore, in future work, the complementarity of attributes can be further considered. In addition, the idea of multi-granularity attribute reduction [56] can be further introduced into the unsupervised attribute reduction proposed in this paper.

CRediT authorship contribution statement

Zhong Yuan: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Hongmei Chen:** Supervision, Resources, Project administration, Funding acquisition, Writing – review & editing. **Xiaoling Yang:** Structuralization, Validation, Writing – review & editing. **Tianrui Li:** Supervision, Resources, Funding acquisition. **Keyu Liu:** Structuralization, Review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61976182, 62076171, 61976245, and 61876157), the Key program for International S&T Cooperation of Sichuan Province, China (2019YFH0097), and Sichuan Key R&D project, China (2020YFG0035).

References

[1] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (2–3) (1990) 191–209.
[2] D. Dubois, H. Prade, Putting rough sets and fuzzy sets together, in: Intelligent Decision Support, Springer, 1992, pp. 203–232.
[3] Z. Yuan, H.M. Chen, P. Xie, P.F. Zhang, J. Liu, T.R. Li, Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, Appl. Soft Comput. 107 (2021) 107353.
[4] J.S. Mi, W.X. Zhang, An axiomatic characterization of a fuzzy generalization of rough sets, Inform. Sci. 160 (1–4) (2004) 235–249.
[5] B. Moser, On the T-transitivity of kernels, Fuzzy Sets and Systems 157 (13) (2006) 1787–1796.
[6] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee, X.z. Wang, On the generalization of fuzzy rough sets, IEEE Trans. Fuzzy Syst. 13 (3) (2005) 343–361.
[7] Z. Yuan, H.M. Chen, T.R. Li, J. Liu, S. Wang, Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection, Fuzzy Sets and Systems 421 (2021) 1–28.
[8] Z. Yuan, H.M. Chen, T.R. Li, B.B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Trans. Cybern. (2021) http://dx.doi.org/10.1109/TCYB.2021.3058780.
[9] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (5) (2006) 414–423.
[10] C.Z. Wang, M.W. Shao, Q. He, Y.H. Qian, Y.L. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (2016) 173–179.
[11] J.J. Song, E.C. Tsang, D.G. Chen, X.B. Yang, Minimal decision cost reduct in fuzzy decision-theoretic rough set model, Knowl.-Based Syst. 126 (2017) 104–112.
[12] X. Zhang, C.L. Mei, D.G. Chen, Y.Y. Yang, A fuzzy rough set-based feature selection method using representative instances, Knowl.-Based Syst. 151 (2018) 216–229.
[13] Y.J. Lin, Y.W. Li, C.X. Wang, J.K. Chen, Attribute reduction for multi-label learning with fuzzy rough set, Knowl.-Based Syst. 152 (2018) 51–61.
[14] B.Z. Sun, W.M. Ma, Y.H. Qian, Multigranulation fuzzy rough set over two universes and its application to decision making, Knowl.-Based Syst. 123 (2017) 61–74.
[15] J.M. Zhan, B.Z. Sun, J.C.R. Alcantud, Covering based multigranulation (i, t)-fuzzy rough set models and applications in multi-attribute group decision-making, Inform. Sci. 476 (2019) 290–318.

[16] H.X. Bai, Y. Ge, J.F. Wang, D.Y. Li, Y.L. Liao, X.Y. Zheng, A method for extracting rules from spatial data based on rough fuzzy sets, Knowl.-Based Syst. 57 (2014) 28–40.
[17] X.Z. Wang, E.C.C. Tsang, S.Y. Zhao, D.G. Chen, D.S. Yeung, Learning fuzzy rules from fuzzy samples based on rough set technique, Inform. Sci. 177 (20) (2007) 4493–4514.
[18] Q.H. Hu, L. Zhang, D.G. Chen, W. Pedrycz, D.R. Yu, Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications, Internat. J. Approx. Reason. 51 (4) (2010) 453–471.
[19] Q.H. Hu, D.R. Yu, W. Pedrycz, D.G. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (11) (2010) 1649–1667.
[20] J.H. Chen, C.S. Chen, Fuzzy kernel perceptron, IEEE Trans. Neural Netw. 13 (6) (2002) 1364–1373.
[21] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[22] V. Popovici, S. Bengio, J.-P. Thiran, Kernel matching pursuit for large datasets, Pattern Recognit. 38 (12) (2005) 2385–2390.
[23] B. Moser, On representing and generating kernels by fuzzy equivalence relations, J. Mach. Learn. Res. 7 (Dec) (2006) 2603–2620.
[24] D.G. Chen, Q.H. Hu, Y.P. Yang, Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets, Inform. Sci. 181 (23) (2011) 5169–5179.
[25] Y.W. Li, Y.J. Lin, J.H. Liu, W. Weng, Z.K. Shi, S.X. Wu, Feature selection for multi-label learning based on kernelized fuzzy rough sets, Neurocomputing 318 (2018) 271–286.
[26] Q.H. Hu, L.J. Zhang, Y.C. Zhou, W. Pedrycz, Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (1) (2017) 226–238.
[27] X.S. Rao, K.Y. Liu, J.J. Song, X.B. Yang, Y.H. Qian, Gaussian kernel fuzzy rough based attribute reduction: An acceleration approach, J. Intell. Fuzzy Systems 39 (1) (2020) 679–695.
[28] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[29] R.R. Yager, Entropy measures under similarity relations, Int. J. Gen. Syst. 20 (4) (1992) 341–358.
[30] R.R. Yager, Uncertainty representation using fuzzy measures, IEEE Trans. Syst. Man Cybern. B 32 (1) (2002) 13–20.
[31] J.S. Mi, Y. Leung, W.Z. Wu, An uncertainty measure in partition-based fuzzy rough sets, Int. J. Gen. Syst. 34 (1) (2005) 77–90.
[32] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.
[33] Y.H. Qian, J.Y. Liang, W. Wei Zhi, C.Y. Dang, Information granularity in fuzzy binary GrC model, IEEE Trans. Fuzzy Syst. 19 (2) (2010) 253–264.
[34] J.H. Dai, J.L. Chen, Feature selection via normative fuzzy information weight with application in biological data classification, Appl. Soft Comput. (2020) 106–299.
[35] X. Zhang, C. Mei, D. Chen, J. Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[36] D.R. Yu, Q.H. Hu, C.X. Wu, Uncertainty measures for fuzzy relations and their applications, Appl. Soft Comput. 7 (3) (2007) 1135–1143.
[37] X. Zhang, C.L. Mei, D.G. Chen, Y.Y. Yang, J.H. Li, Active incremental feature selection using a fuzzy-rough-set-based information entropy, IEEE Trans. Fuzzy Syst. 28 (5) (2019) 901–915.
[38] C.Z. Wang, Y. Huang, M.W. Shao, D.G. Chen, Uncertainty measures for general fuzzy relations, Fuzzy Sets and Systems 360 (2019) 82–96.
[39] J.H. Dai, Q. Xu, Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, Appl. Soft Comput. 13 (1) (2013) 211–221.
[40] J.Y. Zhao, Z.L. Zhang, C.Z. Han, Z.F. Zhou, Complement information entropy for uncertainty measure in fuzzy rough set and its applications, Soft Comput. 19 (7) (2015) 1997–2010.
[41] Y.J. Lin, Q.H. Hu, J.H. Liu, J.J. Li, X.D. Wu, Streaming feature selection for multilabel learning based on fuzzy mutual information, IEEE Trans. Fuzzy Syst. 25 (6) (2017) 1491–1507.
[42] C.Z. Wang, Y. Huang, W.P. Ding, Z.H. Cao, Attribute reduction with fuzzy rough self-information measures, Inform. Sci. 49 (5) (2021) 68–86.
[43] C.Z. Wang, Y. Huang, M.W. Shao, X.D. Fan, Fuzzy rough set-based attribute reduction using distance measures, Knowl.-Based Syst. 164 (2019) 205–212.
[44] A. Ganivada, S.S. Ray, S.K. Pal, Fuzzy rough sets, and a granular neural network for unsupervised feature selection, Neural Netw. 48 (2013) 91–108.
[45] N. Mac ParthaláIn, R. Jensen, Unsupervised fuzzy-rough set-based dimensionality reduction, Inform. Sci. 229 (2013) 106–121.
[46] J.Y. Liang, K.-S. Chin, C.Y. Dang, R.C. Yam, A new method for measuring uncertainty and fuzziness in rough set theory, Int. J. Gen. Syst. 31 (4) (2002) 331–342.
[47] W.J. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components, J. R. Stat. Soc. Ser. C. Appl. Stat. 36 (1) (1987) 22–33.
[48] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[49] X.F. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Advances in Neural Information Processing Systems, 2006, pp. 507–514.
[50] P. Mitra, C. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 301–312.
[51] S. Solorio-Fernández, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, A new unsupervised spectral feature selection method for mixed data: A filter approach, Pattern Recognit. 72 (2017) 314–326.
[52] A. Chaudhuri, D. Samanta, M. Sarma, Two-stage approach to feature set optimization for unsupervised dataset with heterogeneous attributes, Expert Syst. Appl. 172 (2021) 114563.
[53] P.F. Zhu, W.C. Zhu, Q.H. Hu, C.Q. Zhang, W.M. Zuo, Subspace clustering guided unsupervised feature selection, Pattern Recognit. 66 (2017) 364–374.
[54] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[55] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[56] K.Y. Liu, X.B. Yang, H. Fujita, D. Liu, X. Yang, Y.H. Qian, An efficient selector for multi-granularity attribute reduction, Inform. Sci. 505 (2019) 457–472.