Fuzzy Sets and Systems 526 (2026) 109665

Contents lists available at ScienceDirect

Fuzzy Sets and Systems

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

Feature subset selection using fuzzy scale entropy-Based uncertainty measures for multi-scale fuzzy relation decision systems

Jiaying Wang [a], Zhehuang Huang [a,*], Zhifeng Weng [a,b], Jinjin Li [c]

[a] School of Mathematics Sciences, Huaqiao University, Quanzhou, Fujian, 362021, China
[b] Fujian Province University Key Laboratory of Computation Science, School of Mathematical Sciences, Huaqiao University, Quanzhou, 362021, China
[c] School of Mathematics Sciences and Statistics, Minnan Normal University, Zhangzhou, Fujian, 363000, China

ARTICLE INFO

Keywords:
Multi-scale decision systems
Fuzzy entropy
Uncertainty measure
Feature selection

ABSTRACT

As a typical multi-granularity data analysis model, multi-scale decision systems have received widespread attention from researchers in recent years. However, most multi-scale models struggle to handle continuous data and fail to accurately characterize the differences between samples in complex scenes. Moreover, there is a lack of investigation on fuzzy multi-scale uncertainty measures, as well as their application in dimension reduction. Motivated by these issues, we put forth a new multi-scale fuzzy relation decision system and investigate the uncertainty measures for fuzzy relation families at different scales. To this end, δ-fuzzy similarity relationship is presented to characterize the correlation of target objects. Fuzzy scale entropy is then proposed to reflect the distinguishing ability of fuzzy relation families with different scales. Some variants of the uncertainty measure, such as joint fuzzy scale entropy, conditional fuzzy scale entropy, and mutual fuzzy scale entropy, are then presented to reveal the relationship between the distinguishing ability of feature subsets. Finally, a knowledge reduction algorithm for multi-scale fuzzy relation decision systems is developed from the perspective of maintaining the distinguishing ability. Extensive experiments on 16 public datasets exhibit that our model can effectively reduce redundant features from different scales, and demonstrates competitive classification performance compared with four state-of-the-art dimension reduction algorithms.

1. Introduction

Feature subset selection is a fundamental variable selection process in artificial intelligence systems. It aims to reduce the computational complexity of data processing and enhance the generalization capability of classification models by eliminating non-essential and redundant features [1,2]. Feature evaluation is a crucial step in feature subset selection, which can optimize the identification of important features. So far, a variety of feature evaluation methods have been put forward, including consistency-based measures [3], classifier error rate [4], fuzzy relevance and redundancy [5], information-gain [6,7], and so on. All these evaluation methods can effectively assess the importance of features, thereby obtaining the optimal subset with strong distinguishing ability and low redundancy.

* Corresponding author at: School of Mathematics Sciences, Huaqiao University, Quanzhou, 362021, Fujian, China.
E-mail address: startstart1@163.com (Z. Huang).

https://doi.org/10.1016/j.fss.2025.109665
Received 19 July 2025; Received in revised form 30 September 2025; Accepted 29 October 2025
Available online 4 November 2025

Granular computing [8] is a powerful paradigm for knowledge representation and data analysis, enabling the decomposition of complex problems into manageable "granules". It provides a framework for multi-perspective data modeling, integrating methodologies such as rough sets, fuzzy sets, and quotient space theory. Typical granular computing models have been developed to leverage this paradigm, such as multi-granularity rough set [9,10], multi-scale rough sets [11,12], and multi-neighborhood rough sets [13]. These methods facilitate hierarchical data analysis by integrating information across different granular levels, enhancing the ability to handle complex datasets and extract meaningful insights.

Multi-scale decision systems formalized by Wu and Leung [14] in 2011 is an innovative granular computing paradigm. Its core idea is decomposing attributes into hierarchical scales, enabling flexible data representation from fine to coarse granularity. In recent years, multi-scale decision systems have attracted considerable attention and gradually become a research hotspot. Li and Hu [15, 16] proposed a stepwise optimal scale selection algorithm based on attribute significance, iteratively evaluating scale combinations to maximize classification efficiency. Huang and Li [17] developed covering multi-scale systems that utilize covering relations to construct granular spaces for incomplete data. Xie et al. [18] introduced set-valued multi-scale systems where set-valued attribute representations were applied to address incomplete information. Key research directions in multi-scale systems focus on optimal scale selection and decision rule acquisition. She et al. [19] delved into the selection of optimal cuts in complete multi-scale decision tables, contributing to the optimization of hierarchical scale structures. Zhang et al. [20] integrated three-way decision theory with Hasse diagrams to visualize and optimize scale combinations in complex systems. For decision rule extraction, Wu et al. [21] developed an edge discovery algorithm for incomplete multi-scale decision tables. Zhan et al. [22] established a complex proportional assessment strategy that has been used to link multi-scale systems with expert decision-making frameworks. Deng et al. [23] designed a three-way decision methodology with multi-scale decision information systems. She et al. [24] explored generalization reducts and optimal cuts, demonstrating the model's adaptability. Xia et al. [25] introduced three-way approximations fusion with granular-ball computing, providing a novel framework for uncertainty processing in multi-scale systems.

In recent years, an endless stream of multi-scale models and multi-scale data analysis methods have emerged. Yang et al. [26] solved classification learning tasks at different scales. Deng et al. [27] studied granular ball-based feature subset selection for incomplete generalized double multi-scale decision tables. Zhang et al. [28] put forward a multi-scale information fusion-based multiple correlations for unsupervised attribute selection, which is also an important exploration in the field of multi-scale decision systems. Wang et al. [29] further investigated incomplete generalized multi-scale ordered information systems, integrating optimal scale combination selection with practical classification tasks. Xiao et al. [30] introduced a sequential three-way decision framework for group consensus under interval multi-scale systems. In addition, researchers have applied multi-scale thinking to different fields and achieved many outstanding results [31–36].

As an uncertainty measure, information entropy serves in granular computing to characterize knowledge uncertainty, evaluate feature significance, and analyze multi-granular data. Miao and Wang [37] first introduced Shannon's entropy into rough sets to establish an information-entropy-based uncertainty measure framework; Qian and Liang [38] proposed combinatorial granular entropy and ordered granular entropy for discrete variables analysis; Dai et al. [39] developed a conditional entropy model for incomplete systems. In fuzzy set theory, the applications of information entropy focus on similarity measure and fuzzy partitioning. Yager [40] extended entropy to fuzzy similarity relations; Bertoluzza [41] studied uncertainty measures for fuzzy partitions; Zhu et al. [42] constructed a cross-scale granular entropy model. In [43], fuzzy relative entropy is presented to evaluate the importance of features in outlier detection. Xu et al. [44] discussed a multi-criteria decision-making model by means of intuitive fuzzy ternary ranking and intuitive fuzzy entropy. To evaluate the uncertainty of multi-label data, Hamidzadeh et al. [45] proposed an entropy-based optimization function via integrating fuzzy rough sets, kernel fuzzy rough sets, and genetic algorithm. Yang et al. [46] investigated a multi-neighborhood entropy, and applied it to unsupervised outlier detection. In [47], an incremental feature selection algorithm is developed by integrating local neighborhood rough sets and composite entropy. Zhang and Zhao [48] explored a fuzzy neighborhood joint entropy for fuzzy decision systems. In [49], a multi-perspective dynamic neighborhood entropy is presented to evaluate data uncertainty. Xie et al. [50] designed an optimal scale combination selection strategy by using combinatorial entropy.

The motivations of this study are elaborated from the following three dimensions.

(1) Most multi-scale decision systems struggle to handle continuous data, failing to effectively characterize the differences between samples in complex scenarios. Although a small amount of research has involved fuzzy multi-scale scenes [12]. But they lack a method for constructing multi-scale information with clear semantics.

(2) Most of existing uncertainty measures are with single-scale [46–49]. They are difficult to effectively evaluate the uncertainty of feature subsets at different scales. Recently, Xie et al. [50] proposed a multi-scale entropy for optimal scale selection. However, this method cannot handle fuzzy data and is difficult to accurately characterize target objects.

(3) There exists a notable lack of exploration into fuzzy multi-scale uncertainty measures, as well as their application in dimensionality reduction. In practical applications, it is difficult to capture the dynamic discriminative ability of feature subsets across scales.

These gaps motivate us to explore multi-scale fuzzy relation decision systems and study uncertainty measures with different scales. The main contributions of this research are reflected in three aspects.

(1) A δ-fuzzy relation is introduced to construct a novel multi-scale fuzzy decision system, addressing the limitations of traditional models in handling continuous data and maintaining dynamic balance of multi-scale information. This system enables precise characterization of continuous features and fuzzy dependencies.

(2) Fuzzy scale entropy and some of its variants, such as joint fuzzy scale entropy, conditional fuzzy scale entropy, and mutual fuzzy scale entropy, are then presented reflect the distinguishing ability of feature subsets at different scales. Compared to single-scale

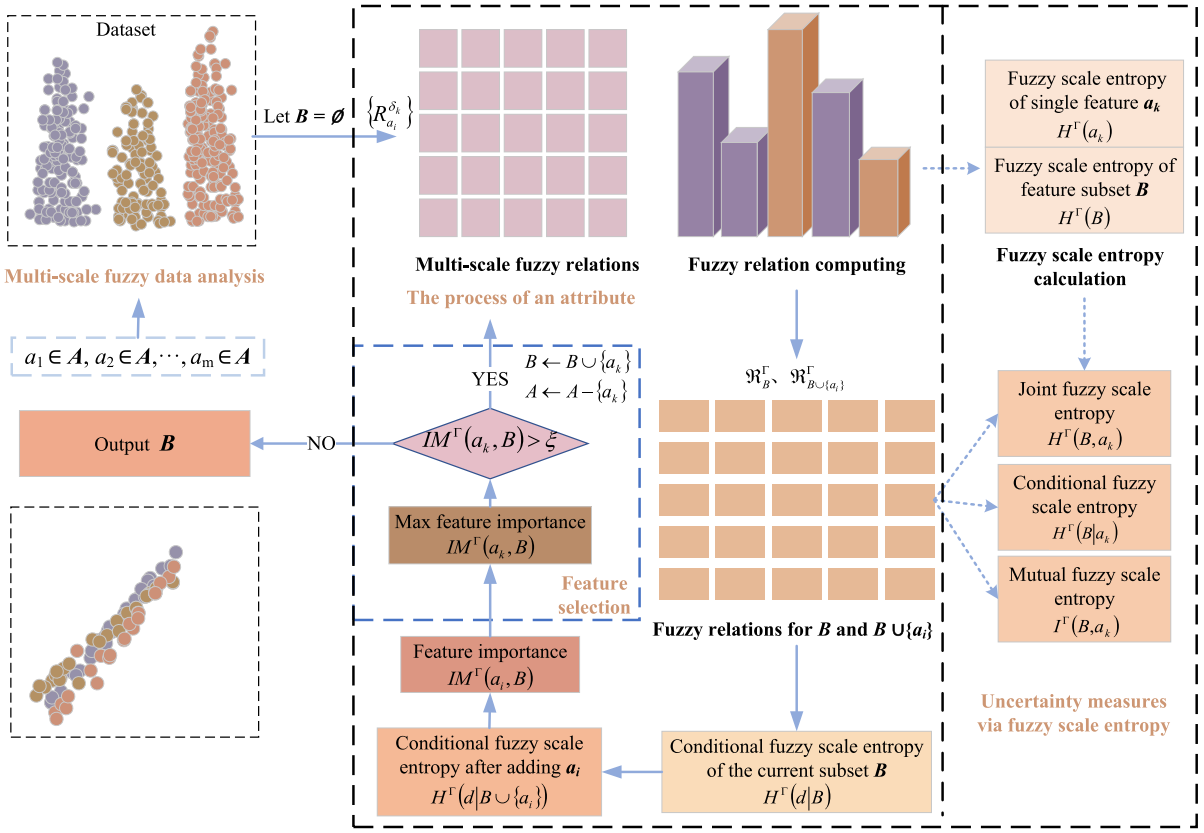


Fig. 1. The framework diagram of MSFRE.

uncertainty measures [46]-[49], the fuzzy scale entropy can characterize the differences between samples at multiple scales, thus fitting the data more accurately.

(3) A knowledge reduction algorithm for multi-scale fuzzy relation decision systems (MSFRE) is developed from the perspective of maintaining the distinguishing ability.

The remaining part of this paper is structured as follows: Section 2 reviews the relevant research on multi-scale rough sets and information entropy; Section 3 presents the multi-scale fuzzy relation decision system and its uncertainty measures; Section 4 elaborates in detail the MSFRE algorithm; Section 5 validates the effectiveness of the algorithm through experiments; finally, the contributions of the whole paper are summarized.

As shown in Fig. 1, the overall process of MSFRE consists of four steps: (1) Multi-scale data analysis; (2) Single feature processing; (3) Uncertainty measures via fuzzy scale entropy; (4) Feature subset selection.

2. Basic knowledge

In the section, we introduce some basic notions about multi-scale information system and Shannon entropy.

2.1. Multi-Scale information systems

Multi-scale information systems, originally introduced by Wu and Leung [11,14], represent a specific form of multigranulation model that has increasingly become the focus of extensive academic exploration.

Let $U = \{x_1, x_2, \ldots, x_n\}$ be a universe of discourse, and $A = \{a_1, a_2, \ldots, a_m\}$ be the set of attributes, then (U, A) is called an information system. Additionally, (U, A, d) is referred to as a decision system, where d is the decision attribute.

For $a \in A$, R_a is defined as a binary relation on $U \times U$, that is $R_a : U \times U \to \{0, 1\}$, and is specifically defined as follows.

$$R_a(x_i, x_j) = \begin{cases} 1, & a(x_i) = a(x_j) \\ 0, & \text{else} \end{cases}, \tag{1}$$

where $x_i, x_j \in U$, R_a is called the indiscernibility relation induced by the attribute a. Especially, R_d is the indiscernibility relation induced by the decision attribute d,

$$R_d(x_i, x_j) = \begin{cases} 1, & d(x_i) = d(x_j) \\ 0, & \text{else} \end{cases}. \tag{2}$$

Definition 1. [14] Let (U, A) be an information system, and each attribute $a_i \in A$ haves I scales. Then, a multi-scale information system can be expressed as $(U, \{a_i^j \mid j = 1, 2, \dots, I; i = 1, 2, \dots, m\})$. For $i = 1, 2, \dots, m$ and $1 \le j \le I - 1$, there exists a surjective mapping $h_i^{j,j+1} : W_i^j \to W_i^{j+1}$, such that $a_i^{j+1} = h_i^{j,j+1} \circ a_i^j$, namely $a_i^{j+1}(x) = h_i^{j,j+1}(a_i^j(x))$ for $x \in U$, where W_i^j denotes the domain of a_i^j, and $h_i^{j,j+1}$ is called a granularity transformation mapping.

Scale variations in multi-scale information systems are represented by these mappings.

Definition 2. [33] Let $(U, \{a_i^j \mid j = 1, 2, \dots, I; i = 1, 2, \dots, m\})$ be a multi-scale information system. $D = \{d\}$ is a non-empty finite set containing a single decision attribute d with n distinct scales $\{d^t \mid t = 1, 2, \dots, n\}$. Then,

$$S = \left(U, \left\{ a_i^j \mid j = 1, 2, \dots, I; i = 1, 2, \dots, m \right\} \cup \left\{ d^t \mid t = 1, 2, \dots, n \right\}\right)$$

is called a generalized multi-scale decision system.

By extending partitions to coverings, Huang and Li [17] formalized multi-scale covering decision systems.

Definition 3. [17] A multi-scale covering information system is a pair $S = (U, \mathcal{F}) = \left(U, \left\{ F_j^k \mid k = 1, 2, \dots, l_j; j = 1, 2, \dots, m \right\}\right)$, where $U = \{x_1, x_2, \dots, x_n\}$ is a universe, $\mathcal{F}$ is a family of coverings of U, each $\mathcal{F}_j = \left\{ F_j^1, F_j^2, \dots, F_j^{l_j} \right\} \subseteq \mathcal{F}$ has l_j levels of scales. If $i \le h$, then F_j^i is finer than F_j^h.

2.2. Shannon entropy

Given an approximation space (U, Π), where the partition Π consists of several blocks B_i, $1 \le i \le k$. If Π is regarded as a random variable with possible values $\{B_1, B_2, \dots, B_k\}$, then the probability of Π taking the value B_i is denoted as

$$P(B_i) = \frac{|B_i|}{|U|}, \quad i = 1, 2, \dots, k, \tag{3}$$

where $|B_i|$ is the cardinality of B_i.

The Shannon information entropy of Π is defined as

$$H(\Pi) = \sum_{i=1}^{k} -P(B_i) \log P(B_i), \tag{4}$$

Information entropy reflects the uncertainty characteristics of the approximate space. With the help of information entropy, a comparative analysis of the two partitioning situations can be achieved.

For the partitions $\Pi_1 = \{B_1, B_2, \dots, B_k\}$ and $\Pi_2 = \{C_1, C_2, \dots, C_l\}$, the conditional entropy of Π_1 relative to Π_2 can be expressed as

$$H(\Pi_1|\Pi_2) = -\sum_{i=1}^{k} \sum_{j=1}^{l} P(B_i \cap C_j) \log P(B_i|C_j). \tag{5}$$

The conditional entropy $H(\Pi_1|\Pi_2)$ represents the remaining uncertainty of partition Π_1 when partition Π_2 is known.

The mutual information between Π_1 and Π_2 can be calculated as follows

$$M(\Pi_1, \Pi_2) = \sum_{i=1}^{k} \sum_{j=1}^{l} P(B_i \cap C_j) \log \frac{P(B_i \cap C_j)}{P(B_i)P(C_j)}. \tag{6}$$

Here, the mutual information $M(\Pi_1, \Pi_2)$ quantifies the level of association between the two partitions.

3. Multi-scale fuzzy relation decision system

In this Section, we first introduced δ-fuzzy similarity relation to characterize the correlation between samples. Based on this, a novel multi-scale decision system, i.e., multi-scale fuzzy relation decision system is formulated. Fuzzy scale entropy and some of its variants are then presented characterize the differences in classification ability of feature subsets at different scales.

Definition 4. Let (U, A) be an information system. For any $x, y \in U$, $a \in A$, and $\delta \in [0, 1]$, the δ-fuzzy similarity relation induced by a is defined as

$$R_a^\delta(x, y) = \begin{cases} 1 - |a(x) - a(y)|, & |a(x) - a(y)| \le \delta \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

where $a(x)$, $a(y)$ are the normalized attribute values, δ is the scale parameter.

Table 1
Original Decision System.

U/A	a_1	a_2	a_3	d
x_1	0.4	0.6	0.2	1
x_2	0.5	0.3	0.3	1
x_3	0.6	0.5	0.5	2

The smaller δ is, the stricter the similarity judgment becomes. Therefore, by adjusting δ, we can obtain a series of fuzzy similarity relationships with different scales.

Definition 5. Let $B \subseteq A$, $R_B^\delta = \bigcap_{a\in B} R_a^\delta$ is called the δ-fuzzy similarity relation induced by B.

We can easily obtain the following properties.

Proposition 1. *If $B_1 \subseteq B_2$, then $R_{B_2}^\delta \subseteq R_{B_1}^\delta$.*

Proposition 2. *If $\delta_1 \leq \delta_2$, then $R_B^{\delta_1} \subseteq R_B^{\delta_2}$.*

Proposition 1 indicates that the more features there are, the more detailed the fuzzy similarity relationship is, and the stronger the discrimination ability of the feature subset is.

Proposition 2 indicates that the smaller δ is, the more refined the characterization of the relationship between samples becomes. In this way, by taking different values of δ, it is easy to obtain a series of fuzzy similarity relations at different scales.

Definition 6. Let (U, A) be an information system. For each attribute $a_i \in A$, we formulate a family of fuzzy similarity relations $\mathbb{R}_{a_i} = \{R_{a_i}^{\delta_1}, R_{a_i}^{\delta_2}, \dots, R_{a_i}^{\delta_I}\}$, where δ_k $(k = 1, 2, \dots, I)$ represents different scales, and $\delta_1 < \delta_2 < \cdots < \delta_I$. The tuple $(U, \{R_{a_i}^{\delta_k} \mid k = 1, 2, \dots, I, i = 1, 2, \dots, m\})$ is called a multi-scale fuzzy relation information system. Furthermore, $(U, \{R_{a_i}^{\delta_k} \mid k = 1, 2, \dots, I, i = 1, 2, \dots, m\}, R_d)$ is refered to a multi-scale fuzzy relation decision system.

For any $a_i \in A$, we have $R_{a_i}^{\delta_1} \subseteq R_{a_i}^{\delta_2} \subseteq \cdots \subseteq R_{a_i}^{\delta_I}$. This exhibits a hierarchical structure from finer to coarser of fuzzy relations.

Example 1. A decision system is presented in Table 1. Let $\delta_1 = 0.1$, $\delta_2 = 0.2$, $\delta_3 = 0.3$, and $I = 3$.

By formula (7), it can be obtained that

$$R_{a_1}^{\delta_1} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0.9 \\ 0 & 0.9 & 1 \end{pmatrix}, \quad R_{a_1}^{\delta_2} = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{pmatrix}, \quad R_{a_1}^{\delta_3} = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{pmatrix},$$

$$R_{a_2}^{\delta_1} = \begin{pmatrix} 1 & 0 & 0.9 \\ 0 & 1 & 0 \\ 0.9 & 0 & 1 \end{pmatrix}, \quad R_{a_2}^{\delta_2} = \begin{pmatrix} 1 & 0 & 0.9 \\ 0 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{pmatrix}, \quad R_{a_2}^{\delta_3} = \begin{pmatrix} 1 & 0.7 & 0.9 \\ 0.7 & 1 & 0.8 \\ 0.9 & 0.8 & 1 \end{pmatrix},$$

$$R_{a_3}^{\delta_1} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_{a_3}^{\delta_2} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}, \quad R_{a_3}^{\delta_3} = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.8 \\ 0.7 & 0.8 & 1 \end{pmatrix},$$

$$R_d = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We can easily see that $R_{a_i}^{\delta_1} \subseteq R_{a_i}^{\delta_2} \subseteq R_{a_i}^{\delta_3}$, $i = 1, 2, 3$.

Thus, $(U, \{R_{a_1}^{\delta_1}, R_{a_1}^{\delta_2}, R_{a_1}^{\delta_3}, R_{a_2}^{\delta_1}, R_{a_2}^{\delta_2}, R_{a_2}^{\delta_3}, R_{a_3}^{\delta_1}, R_{a_3}^{\delta_2}, R_{a_3}^{\delta_3}\}, R_d)$ is multi-scale fuzzy relation decision system.

Definition 7. Let $(U, \{R_{a_i}^{\delta_k} | k = 1, 2, \cdots, I, i = 1, 2, \cdots, m\}, R_d)$ be a multi-scale fuzzy relation decision system. Let $l_i \in \{1, 2, \cdots, I\}$ denote the selected scale of a_i. We call the scale index set $\Gamma = (l_1, l_2, \cdots, l_m)$ a scale combination of A, The collection of all scale combinations is denoted by $\mathcal{T} = \{(l_1, l_2, \cdots, l_m) \mid l_i \in \{1, 2, \cdots, I\}, i = 1, 2, \cdots, m\}$.

For a given scale combination $(l_1, l_2, \cdots, l_m)$, we obtain a single-scale fuzzy relation decision system $(U, \{R_{a_1}^{\delta_{l_1}}, R_{a_2}^{\delta_{l_2}} \dots, R_{a_m}^{\delta_{l_m}}\}, R_d)$.

Let $\Gamma_1 = (l_1^1, l_2^1, \cdots, l_m^1)$ and $\Gamma_2 = (l_1^2, l_2^2, \cdots, l_m^2) \in \mathcal{T}$, a partial order relation is defined as follows

(1) Γ_1 is finer than or equal to Γ_2 (denoted as $\Gamma_1 \preceq \Gamma_2$) if and only if $l_j^1 \leq l_j^2$ for all $j \in \{1, 2, \dots, m\}$.

(2) Γ_1 is strictly finer than Γ_2 (denoted as $\Gamma_1 \prec \Gamma_2$) if $\Gamma_1 \preceq \Gamma_2$ and there exists at least one index $t \in \{1, 2, \dots, m\}$ such that $l_t^1 < l_t^2$.

Definition 8. Let $B \subseteq A$, $\Gamma = (l_1, l_2, \cdots, l_m) \in \mathcal{T}$. The fuzzy similarity relation induced by B is denoted as $\mathfrak{R}_B^\Gamma = \bigcap_{a_k\in B} R_{a_k}^{\delta_{l_k}}$, where l_k is the scale of a_k under Γ.

Proposition 3. *If $\Gamma_1 \preceq \Gamma_2$, then $\mathfrak{R}_B^{\Gamma_1} \subseteq \mathfrak{R}_B^{\Gamma_2}$.*

Proof. Given $\Gamma_1 = (l_1^1, l_2^1, \dots, l_m^1)$ and $\Gamma_2 = (l_1^2, l_2^2, \dots, l_m^2)$ with $\Gamma_1 \preceq \Gamma_2$. We have $\delta_{l_k^1} \leq \delta_{l_k^2}$. Thus, $R_{a_k}^{\delta_{l_k^1}} \subseteq R_{a_k}^{\delta_{l_k^2}}$ for all $a_k \in B$. By Definition 8, we obtain that $\mathfrak{R}_B^{\Gamma_1} \subseteq \mathfrak{R}_B^{\Gamma_2}$. □

From Proposition 3, we can see that the hierarchical characteristic of the multi-scale fuzzy relation: a fine-grained scale not only inherits the original fuzzy partition but also can further expand the new discrimination boundaries through refinement.

Proposition 4. *If $B_1 \subseteq B_2$, then $\mathfrak{R}_{B_2}^{\Gamma} \subseteq \mathfrak{R}_{B_1}^{\Gamma}$.*

Proof. By Definition 8, we have $\mathfrak{R}_{B_1}^{\Gamma} = \bigcap_{a_k \in B_1} R_{a_k}^{\delta_{l_k}}$, $\mathfrak{R}_{B_2}^{\Gamma} = \bigcap_{a_k \in B_2} R_{a_k}^{\delta_{l_k}}$. Since $B_1 \subseteq B_2$, for any $a \in B_1$, we obtain that $a \in B_2$. It follows that $\bigcap_{a_k \in B_2} R_{a_k}^{\delta_{l_k}} \subseteq \bigcap_{a_k \in B_1} R_{a_k}^{\delta_{l_k}}$. Hence $\mathfrak{R}_{B_2}^{\Gamma} \subseteq \mathfrak{R}_{B_1}^{\Gamma}$. □

Proposition 4 depicts the association between the inclusion relation of attribute subsets and the inclusion relation of corresponding multi-scale fuzzy relations.

Example 2. Continued from Example 1, we have obtained a multi-scale fuzzy relation decision system $(U, \{R_{a_1}^{\delta_1}, R_{a_1}^{\delta_2}, R_{a_1}^{\delta_3}, R_{a_2}^{\delta_1}, R_{a_2}^{\delta_2}, R_{a_2}^{\delta_3}, R_{a_3}^{\delta_1}, R_{a_3}^{\delta_2}, R_{a_3}^{\delta_3}\}, R_d)$. Let $B = \{a_1, a_2\}$, $C = \{a_2, a_3\}$, $\Gamma_1 = (1, 1, 2)$, $\Gamma_2 = (2, 1, 3)$.

It is calculated by Definition 8 that

$$\mathfrak{R}_B^{\Gamma_1} = R_{a_1}^{\delta_1} \cap R_{a_2}^{\delta_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathfrak{R}_B^{\Gamma_2} = R_{a_1}^{\delta_2} \cap R_{a_2}^{\delta_1} = \begin{pmatrix} 1 & 0 & 0.8 \\ 0 & 1 & 0 \\ 0.8 & 0 & 1 \end{pmatrix},$$

$$\mathfrak{R}_C^{\Gamma_1} = R_{a_2}^{\delta_1} \cap R_{a_3}^{\delta_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathfrak{R}_C^{\Gamma_2} = R_{a_2}^{\delta_1} \cap R_{a_3}^{\delta_3} = \begin{pmatrix} 1 & 0 & 0.7 \\ 0 & 1 & 0 \\ 0.7 & 0 & 1 \end{pmatrix}.$$

Currently, most uncertainty measures are based on distinguishing information at a single scale. However, different scale combinations contain different sample discrimination information. In fact, the granularities of sample description vary among different scales, and a single scale is difficult to fully reflect the characteristics of the data.

Definition 9. Let $\Gamma = (l_1, l_2, \cdots, l_m) \in \mathcal{T}$ and $B \subseteq A$. The fuzzy scale entropy of B is defined as

$$H^{\Gamma}(B) = -\log \frac{|\mathfrak{R}_B^{\Gamma}|}{|U|^2} \tag{8}$$

where $|\mathfrak{R}_B^{\Gamma}|$ represents the cardinality of $\mathfrak{R}_B^{\Gamma}$.

The fuzzy scale entropy of B assesses the discrimination ability of the fuzzy similarity relation family under a specific scale combination. It represents a mapping from the feature space into the real space: $H : (B, \Gamma) \to \mathbb{R}^+$, in which $\mathbb{R}^+$ is the domain of non-negative real numbers. Using this mapping, the larger its value is, the stronger the discrimination ability of the fuzzy similarity relation family is, and the smaller the uncertainty is.

Based on the inequality $|\mathfrak{R}_B^{\Gamma}| \leq n$, the non-negativity $H^{\Gamma}(B) \geq 0$ can be derived. Correspondingly, when $|\mathfrak{R}_B^{\Gamma}| = n$, the system reaches the maximum uncertainty $H^{\Gamma}(B) = \log n$; when $|\mathfrak{R}_B^{\Gamma}| = n^2$, the system has a deterministic state $H^{\Gamma}(B) = 0$.

By using fuzzy scale entropy, one can analyze the discrimination ability of fuzzy similarity relations for samples under different scale combinations, and then screening out the optimal scale combination is of great significance for accurately depicting samples. Next, we will delve into analyze the specific impacts of different scales on the discriminative information they contain.

Proposition 5. *Let $\Gamma_1, \Gamma_2 \in \mathcal{T}$. If $\Gamma_1 \preceq \Gamma_2$, then $H^{\Gamma_1}(B) \geq H^{\Gamma_2}(B)$.*

Proof. Since $\Gamma_1 \preceq \Gamma_2$, by Proposition 3, we have $\mathfrak{R}_B^{\Gamma_1} \subseteq \mathfrak{R}_B^{\Gamma_2}$. Hence, $|\mathfrak{R}_B^{\Gamma_1}| \leq |\mathfrak{R}_B^{\Gamma_2}|$. Then, by Definition 9, we obtain $H^{\Gamma_1}(B) \geq H^{\Gamma_2}(B)$. □

This proposition indicates that the finer the scale, the larger the fuzzy scale entropy, and the stronger the ability of the fuzzy relation family to distinguish samples.

Proposition 6. *If $B_1 \subseteq B_2$, then $H^{\Gamma}(B_1) \leq H^{\Gamma}(B_2)$.*

Proof. Since $B_1 \subseteq B_2$, we have $\mathfrak{R}_{B_2}^{\Gamma} \subseteq \mathfrak{R}_{B_1}^{\Gamma}$. So, $|\mathfrak{R}_{B_2}^{\Gamma}| \leq |\mathfrak{R}_{B_1}^{\Gamma}|$. By Definition 9, we obtain $H^{\Gamma}(B_1) \leq H^{\Gamma}(B_2)$. □

This proposition shows that under the same scale combination, the larger the size of the feature subset, the higher its corresponding fuzzy scale entropy and the stronger its discrimination ability.

Definition 10. Let $B_1, B_2 \subseteq A$ and $\Gamma \in \mathcal{T}$, the joint fuzzy scale entropy of B_1 and B_2 is defined as

$$H^{\Gamma}(B_1, B_2) = -\log \frac{|\mathfrak{R}_{B_1}^{\Gamma} \cap \mathfrak{R}_{B_2}^{\Gamma}|}{|U|^2}. \tag{9}$$

The joint fuzzy scale entropy $H^{\Gamma}(B_1, B_2)$ represents the ability of the fuzzy similarity relations to jointly distinguish samples under B_1 and B_2.

Proposition 7. *If $B_1, B_2 \subseteq A$, and $\Gamma \in \mathcal{T}$. Then $H^{\Gamma}(B_1, B_2) \geq \max(H^{\Gamma}(B_1), H^{\Gamma}(B_2))$.*

Proof. From Definitions 9 and 10, we have $H^{\Gamma}(B_1) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1}|}{|U|^2}$, $H^{\Gamma}(B_1, B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|U|^2}$. Since $\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2} \subseteq \mathfrak{R}^{\Gamma}_{B_1}$, we get $|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}| \leq |\mathfrak{R}^{\Gamma}_{B_1}|$, thus $H^{\Gamma}(B_1, B_2) \geq H^{\Gamma}(B_1)$. Similarly, we can obtain $H^{\Gamma}(B_1, B_2) \geq H^{\Gamma}(B_2)$. So $H^{\Gamma}(B_1, B_2) \geq \max(H^{\Gamma}(B_1), H^{\Gamma}(B_2))$. □

Obviously, the joint fuzzy scale entropy of B_1 and B_2 exceeds the fuzzy scale entropy of any single one. This indicates that the discriminative ability of the joint features increases with the addition of new features. This is because by introducing new features, we can obtain a more refined fuzzy similarity relation.

Proposition 8. *Let $\Gamma_1, \Gamma_2 \in \mathcal{T}$, $\Gamma_1 \preceq \Gamma_2$, and $B_1, B_2 \subseteq A$. Then $H^{\Gamma_1}(B_1, B_2) \geq H^{\Gamma_2}(B_1, B_2)$.*

Proof. From Definitions 10, we have $H^{\Gamma_1}(B_1, B_2) = -\log \frac{|\mathfrak{R}^{\Gamma_1}_{B_1} \cap \mathfrak{R}^{\Gamma_1}_{B_2}|}{|U|^2}$, $H^{\Gamma_2}(B_1, B_2) = -\log \frac{|\mathfrak{R}^{\Gamma_2}_{B_1} \cap \mathfrak{R}^{\Gamma_2}_{B_2}|}{|U|^2}$. Since $\Gamma_1 \preceq \Gamma_2$, by Proposition 3, we have $\mathfrak{R}^{\Gamma_1}_{B_1} \cap \mathfrak{R}^{\Gamma_1}_{B_2} \subseteq \mathfrak{R}^{\Gamma_2}_{B_1} \cap \mathfrak{R}^{\Gamma_2}_{B_2}$. So $H^{\Gamma_1}(B_1, B_2) \geq H^{\Gamma_2}(B_1, B_2)$. □

This proposition reveals a key-scale-fineness-driven distinction mechanism in joint fuzzy scale entropy. When Γ_1 (finer scale) and Γ_2 (coarser scale), act on the same feature subsets B_1, B_2, the finer Γ_1 induces a smaller indistinguishable relation.

Proposition 9. *If $B_1 \subseteq B_2$, then $H^{\Gamma}(B_1, B_2) = H^{\Gamma}(B_2)$.*

This proposition indicates that when newly added features are contained within other existing features, the addition of these features will not enhance the discriminative ability. In this case, the original feature subset already implicitly contains the uncertainty information.

Definition 11. Let $B_1, B_2 \subseteq A$ and $\Gamma \in \mathcal{T}$, the conditional fuzzy scale entropy of B_1 with respect to B_2 is

$$H^{\Gamma}(B_1|B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|\mathfrak{R}^{\Gamma}_{B_2}|} \tag{10}$$

The conditional fuzzy scale entropy reflects the increment of the ability of feature subset B_1 to distinguish samples under the condition of feature subset B_2. When $B_1 \subseteq B_2$, $H^{\Gamma}(B_1|B_2) = 0$.

Proposition 10. *Let $B_1, B_2 \subseteq A$, then $H^{\Gamma}(B_1|B_2) = H^{\Gamma}(B_1, B_2) - H^{\Gamma}(B_2)$.*

Proof. From Definitions 9 and 10, we have $H^{\Gamma}(B_1, B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|U|^2}$, $H^{\Gamma}(B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_2}|}{|U|^2}$. Since $H^{\Gamma}(B_1, B_2) - H^{\Gamma}(B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|U|^2} + \log \frac{|\mathfrak{R}^{\Gamma}_{B_2}|}{|U|^2} = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|\mathfrak{R}^{\Gamma}_{B_2}|}$, and $H^{\Gamma}(B_1|B_2) = -\log \frac{|\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|\mathfrak{R}^{\Gamma}_{B_2}|}$. Thus, $H^{\Gamma}(B_1|B_2) = H^{\Gamma}(B_1, B_2) - H^{\Gamma}(B_2)$. □

The conditional fuzzy scale entropy $H^{\Gamma}(B_1, B_2)$ represents the additional discriminative information for samples when B_1 is the known feature subset B_2. This reflects the improvement in the ability to distinguish samples with the addition of a new feature subset, based on the information of the existing feature subset.

Proposition 11. *Let $\Gamma \in \mathcal{T}$, $B_1 \subseteq B_2$, then $H^{\Gamma}(B_1|B_2) = 0$.*

This proposition point implies that when B_2 contains B_1, B_1 does not add additional discriminative information for samples.

Definition 12. Let $B_1, B_2 \subseteq A$ and $\Gamma \in \mathcal{T}$, the mutual fuzzy scale entropy of B_1 and B_2 is defined as

$$I^{\Gamma}(B_1, B_2) = \log \frac{|U|^2 |\mathfrak{R}^{\Gamma}_{B_1} \cap \mathfrak{R}^{\Gamma}_{B_2}|}{|\mathfrak{R}^{\Gamma}_{B_1}| \cdot |\mathfrak{R}^{\Gamma}_{B_2}|} \tag{11}$$

The mutual fuzzy scale entropy measures the common discriminative information between the two feature subsets B_1 and B_2.

Proposition 12. *Let $B_1, B_2 \subseteq A$ and $\Gamma \in \mathcal{T}$. Then the following properties hold.*
(1) $I^{\Gamma}(B_1, B_2) = H^{\Gamma}(B_1) + H^{\Gamma}(B_2) - H^{\Gamma}(B_1, B_2)$;
(2) $I^{\Gamma}(B_1, B_2) = H^{\Gamma}(B_1) - H^{\Gamma}(B_1|B_2) = H^{\Gamma}(B_2) - H^{\Gamma}(B_2|B_1)$;

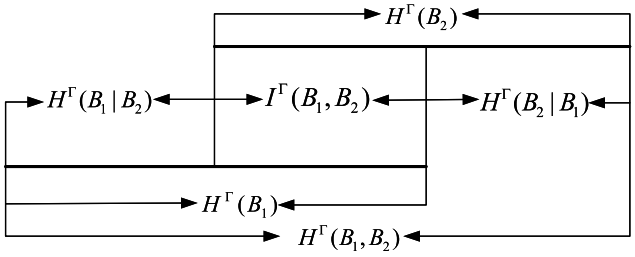


Fig. 2. Relationship diagram of fuzzy scale entropy and its variants.

Proof. (1) $H^\Gamma(B_1) + H^\Gamma(B_2) - H^\Gamma(B_1, B_2) = -\log \frac{|\mathfrak{R}^\Gamma_{B_1}|}{|U|^2} - \log \frac{|\mathfrak{R}^\Gamma_{B_2}|}{|U|^2} + \log \frac{|\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|U|^2} = \log \left(\frac{|U|^2}{|\mathfrak{R}^\Gamma_{B_1}|} \cdot \frac{|U|^2}{|\mathfrak{R}^\Gamma_{B_2}|} \cdot \frac{|\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|U|^2} \right) = \log \frac{|U|^2 |\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|\mathfrak{R}^\Gamma_{B_1}| \cdot |\mathfrak{R}^\Gamma_{B_2}|} = I^\Gamma(B_1, B_2)$

(2) $H^\Gamma(B_1) - H^\Gamma(B_1|B_2) = -\log \frac{|\mathfrak{R}^\Gamma_{B_1}|}{|U|^2} + \log \frac{|\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|\mathfrak{R}^\Gamma_{B_2}|} = \log \frac{|U|^2 |\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|\mathfrak{R}^\Gamma_{B_1}| \cdot |\mathfrak{R}^\Gamma_{B_2}|} = I^\Gamma(B_1, B_2)$

Similarly, it can be proved that $I^\Gamma(B_1, B_2) = H^\Gamma(B_2) - I^\Gamma(B_2|B_1)$. □

The first item indicates that the mutual fuzzy scale entropy is the difference between the sum of the scale entropies and the joint fuzzy scale entropy. The second item shows that the mutual fuzzy scale entropy is the difference between the fuzzy scale entropy of one of the two feature subsets and their conditional fuzzy scale entropy. This reflects that the mutual fuzzy scale entropy is the common part of the discriminative information of the two feature subsets. The relationships among fuzzy scale entropy, conditional fuzzy scale entropy, and mutual fuzzy scale entropy can be illustrated by Fig. 2.

Example 3. Continuing from Example 1 and Example 2. Let $\Gamma = (1, 2, 1)$, $B_1 = \{a_1, a_2\}$ and $B_2 = \{a_1, a_3\}$. we have

$$\mathfrak{R}^\Gamma_{B_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}, \quad \mathfrak{R}^\Gamma_{B_2} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

By Definition 9, we have

$$H^\Gamma(B_1) = -\log \frac{|\mathfrak{R}^\Gamma_{B_1}|}{|U|^2} = -\log \frac{4.6}{3^2} \approx 0.290,$$

$$H^\Gamma(B_2) = -\log \frac{|\mathfrak{R}^\Gamma_{B_2}|}{|U|^2} \approx 0.273.$$

We can obtain by Definition 10 that

$$H^\Gamma(B_1, B_2) = -\log \frac{|\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|U|^2} = -\log \frac{3}{3^2} = \log 3 \approx 0.477.$$

By Definition 11, we have

$$H^\Gamma(B_1|B_2) = -\log \frac{|\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|\mathfrak{R}^\Gamma_{B_2}|} = -\log \frac{3}{4.8} \approx 0.204.$$

We can obtain by Definition 12 that

$$I^\Gamma(B_1, B_2) = \log \frac{|U|^2 |\mathfrak{R}^\Gamma_{B_1} \cap \mathfrak{R}^\Gamma_{B_2}|}{|\mathfrak{R}^\Gamma_{B_1}||\mathfrak{R}^\Gamma_{B_2}|} = \log \frac{3^2 \times 3}{4.6 \times 4.8} \approx 0.087.$$

These uncertainty measures provide a theoretical basis for feature selection. By measuring the discrimination ability of samples under different scale combinations, it helps to excavate the potential information of feature subsets.

4. Feature selection algorithm based on multi-scale fuzzy relation entropy

In this section, we utilize multi-scale fuzzy relation entropy to develop a novel feature selection algorithm.

Theorem 1. *Let $S = (U, \{R_{a_i}^{\delta_k} \mid k = 1, 2, \dots, I, i = 1, 2, \dots, m\}, R_d)$ be a multi-scale fuzzy relation decision system with $\mathfrak{R}_B^\Gamma \subseteq R_d$, then $H^\Gamma(d|B) = 0$.*

Proof. Since $\mathfrak{R}_B^\Gamma \subseteq R_d$. We have $H^\Gamma(d|B) = -\log \dfrac{|R_B^\Gamma \cap R_d|}{|R_B^\Gamma|} = -\log \dfrac{|R_B^\Gamma|}{|R_B^\Gamma|} = 0.$ □

In the multi-scale fuzzy decision system, conditional fuzzy scale entropy can be used to measure the distinguishing ability of feature subsets. The smaller the conditional fuzzy scale entropy of a feature subset is, the stronger its distinguishing ability for samples is, and the higher the importance of this feature subset is. The decrease in conditional fuzzy scale entropy reflects the improvement of the distinguishing ability of the new feature subset.

In classification learning, it is necessary to seek a reduct of the multi-scale fuzzy relation family. The sufficiency requires that the selected multi-scale fuzzy relation family maintains the maximum distinguishing ability under different decisions, and the necessity requires that there are no redundant fuzzy relations in the selected relation family. Based on this, the definition of the reduct of the multi-scale fuzzy relation family is given as follows.

Definition 13. Let $S = (U, \{R_{a_i}^{\delta_k} | k = 1, 2, \cdots, I, i = 1, 2, \cdots, m\}, R_d)$ be a multi-scale fuzzy relation decision system, $B \subseteq A$, and $\Gamma \in \mathcal{T}$. We say B is a reduct of A relative to d if B satisfies

(1)$H^\Gamma(d|B) \le H^\Gamma(d|A)$;

(2)$H^\Gamma(d|B - \{a_i\}) > H^\Gamma(d|B),\ \forall a_i \in B$.

Obviously, a reduct of B relative to d is the minimal feature subsets, which can keep or decrease the conditional discernibility measure.

As mentioned above, the conditional fuzzy scale entropy can quantify the distinguishing ability of feature subsets. The smaller this value, the stronger the distinguishing ability. The improvement of the overall distinguishing ability by the newly added fuzzy relations can be characterized by the decrease in the conditional fuzzy scale entropy. Therefore, the importance of a feature can be defined as follows.

Definition 14. Let $B \subseteq A$, $a_i \in A - B$ and $\Gamma \in \mathcal{T}$. the importance of a_i related to B is defined as

$$IM^\Gamma(a_i, B) = H^\Gamma(d|B) - H^\Gamma(d|B \cup \{a_i\}). \tag{12}$$

When $B = \emptyset$, we define $H^\Gamma(d|B) = H^\Gamma(d)$. If $IM^\Gamma(a_i, B) > \xi$ (threshold ξ), it indicates that after adding the feature a_i, the discrimination ability of the feature subset is improved, and the larger the value is, the more important the feature a_i is. This index provides a theoretical basis for feature selection by quantifying the gain of conditional fuzzy scale entropy.

Next, we design a feature selection algorithm, which performs feature subset selection and screens out optimal features by conditional fuzzy scale entropy.

Algorithm 1 Feature Selection with Multi-scale Fuzzy Relation Entropy (MSFRE).

Input: A multi-scale fuzzy relation decision system
$S = (U, \{R_{a_i}^{\delta_k} \mid k = 1, 2, \dots, I, i = 1, 2, \dots, m\}, R_d)$, and $\Gamma \in \mathcal{T}$.
Output: Optimal feature subset B.
1: Let $B = \emptyset$;
2: Compute the conditional fuzzy scale entropy $H^\Gamma(d|B)$;
3: **For each** feature $a_i \in A - B$
4: Compute $\mathfrak{R}_B^\Gamma$;
5: Compute $\mathfrak{R}_{B\cup\{a_i\}}^\Gamma$;
6: Compute $H^\Gamma(d|B \cup \{a_i\})$ according to Definition 11;
7: Compute the importance $IM^\Gamma(a_i, B)$ according to Definition 14;
8: **End For**
9: Find a_k with the maximum value $IM^\Gamma(a_k, B)$;
10: **If** $IM^\Gamma(a_k, B) > \xi$
11: Let $B \leftarrow B \cup \{a_k\}$ and $A \leftarrow A - \{a_k\}$;
12: Goto Step 2;
13: **Else**
14: Output B and terminate the algorithm;
15: **End If**

Next, we discuss the time complexity of the new algorithm.

In step 5, the fuzzy similarity relation for a_i is calculated with the time complexity $O(n^2)$. In step 6, the computation of the conditional fuzzy scale entropy can be obtained in $O(n^2)$. In steps 10–12, each a_i in A must be evaluated in the worst case within $O(m^2n^2)$. Thus, the overall time complexity of MSFRE is $O(m^2n^2)$.

Next, we use an example to illustrate the idea of this algorithm.

Example 4. Continuing from Example 1, let $\Gamma = (3, 1, 2)$, and $\xi = 0.05$.

Initialize $B = \varnothing$. By Definition 9, we have

$$H^{\Gamma}(d) = -\log \frac{|R_d|}{|U|^2} = -\log \frac{5}{9} \approx 0.255.$$

According to Algorithm 1, we compute the fuzzy similar relation for each attribute.

$$R^{\Gamma}_{a_1} = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{pmatrix}, \quad R^{\Gamma}_{a_2} = \begin{pmatrix} 1 & 0 & 0.9 \\ 0 & 1 & 0 \\ 0.9 & 0 & 1 \end{pmatrix},$$
$$R^{\Gamma}_{a_3} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}, \quad R^{\Gamma}_{a_1} \cap R_d = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$R^{\Gamma}_{a_2} \cap R_d = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R^{\Gamma}_{a_3} \cap R_d = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is calculated that

$$H^{\Gamma}(d|\{a_1\}) = -log \frac{|R^{\Gamma}_{a_1} \cap R_d|}{|R^{\Gamma}_{a_1}|} = -\log \frac{4.8}{8.2} \approx 0.233,$$
$$H^{\Gamma}(d|\{a_2\}) \approx 0.204, H^{\Gamma}(d|\{a_3\}) \approx 0.125.$$

We can obtain by Definition 14 that

$$IM^{\Gamma}(a_1, B) = 0.255 - 0.233 = 0.022,$$
$$IM^{\Gamma}(a_2, B) = 0.051, IM^{\Gamma}(a_3, B) = 0.130.$$

From steps 10–12, we should set $B = \{a_3\}$. Since $IM^{\Gamma}(a_3, B) > \xi$, it must go to the next cycle. We can easily obtain that $H^{\Gamma}(d|\{a_3\}) = 0.125$. Consequently,

$$R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_3} = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{pmatrix}, \quad R^{\Gamma}_{a_2} \cap R^{\Gamma}_{a_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_3} \cap R_d = \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R^{\Gamma}_{a_2} \cap R^{\Gamma}_{a_3} \cap R_d = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It follows that

$$H^{\Gamma}(d|\{a_1, a_3\}) = -\log \frac{|R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_3} \cap R_d|}{|R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_3}|} \approx 0.125, \ H^{\Gamma}(d|\{a_2, a_3\}) = 0,$$
$$IM^{\Gamma}(a_1, \{a_3\}) = 0.125 - 0.125 = 0, \ IM^{\Gamma}(a_2, \{a_3\}) = 0.125.$$

We then verify that a_2 is with the maximum of significance. Thus, we update $B = \{a_2, a_3\}$. Since $IM^{\Gamma}(a_2, a_3) > \xi$, we turn to the next cycle. It is calculated that $H^{\Gamma}(d|\{a_2, a_3\}) = 0$.

Subsequently,

$$R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_2} \cap R^{\Gamma}_{a_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ R^{\Gamma}_{a_1} \cap R^{\Gamma}_{a_2} \cap R^{\Gamma}_{a_3} \cap R_d = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Successively, we compute that

$H^{\Gamma}(d|\{a_1, a_2, a_3\}) = 0$, $IM^{\Gamma}(a_1, \{a_2, a_3\}) = 0$.

Based on the fact that $IM^{\Gamma}(a_1, \{a_2, a_3\}) < \xi$, we determine that $B = \{a_2, a_3\}$ is the optimal feature subset.

The MSFRE algorithm realizes the efficient selection of feature subsets under multiple scales by quantifying the changes in conditional scale entropy. In Example 4, the algorithm screens out the smallest and most effective feature subset by comparing the importance of different features, embodying the core idea of optimizing feature combinations based on uncertainty measures.

5. Experimental results and analysis

In this section, to verify the performance of the proposed algorithm, we compare it with four state-of-the-art feature selection models: Neighborhood Rough Set (NRS)[51], Correlation-based Feature Selection (CFS) [52], Fuzzy Neighborhood Rough Set (FNRS)[7], and Directed Fuzzy Rough Set (DRFFS)[34]. First, we compare the classification accuracies of different datasets. Then, the sizes of the attribute subsets selected by these data reduction algorithms are shown. Finally, we further evaluate the effectiveness of the proposed model through significance tests.

Table 2
Dataset Description.

NO.	Dataset	Instances	Attributes	classes
1	Iris	150	4	3
2	Newthyroid	215	5	3
3	Bupa	345	6	2
4	Car	1728	6	4
5	Glass	214	9	7
6	Wine	178	13	3
7	Australian	690	14	2
8	Wdbc	569	30	2
9	Ionosphere	351	34	2
10	Dermatology	358	34	6
11	Spectfheart	267	44	2
12	Sonar	208	60	2
13	Toxicity	171	1203	2
14	Lung	203	3312	5
15	Arcene	200	10,000	2
16	SMK_CAN_187	187	19,993	2

Ten-fold cross-validation is adopted to evaluate these algorithms. First, we use two classic classifiers, namely the K-Nearest Neighbors (KNN) ($K = 5$) and Classification and Regression Tree (CART), to assess the classification accuracy on the dimensionality-reduced data. In MSFRE algorithm, for each attribute, we adopt 10 scales ($\delta_i = i/10$, where $i \in \{1, 2, \dots, 10\}$). To simplify the experimental setup, all attributes are constrained to adopt the same scale parameter, ensuring synchronous adjustment of their scales across the entire dataset. The threshold ξ was set as 0.01 for high-dimensional data and 0.001 for low-dimensional data. Subsequently, we calculate all classification results of different classifiers through ten-fold cross-validation. The classification performances of different feature selection algorithms are compared by the average accuracy (mean ± standard deviation) and the size of the selected attribute subsets.

Sixteen datasets are used in the experimental analysis, including UCI Machine Learning Repository[1] (Iris, Car, Glass, Wine, Australian, Ionosphere, Spectfheart, Sonar, Toxicity), Scikit-Feature Datasets[2] (Arcene, SMK_CAN_187), ELVIRA Biomedical Data Set Repository[3] (Lung) and KEEL Datesets Repository[4] (Newthyroid, Bupa, Dermatology, Wdbc). An overview of these datasets is presented in Table 2. All data are normalized using the min-max normalization method, with the formula as follows

$$a'(x) = \frac{a(x) - \min a}{\max a - \min a}, \tag{13}$$

where $a(x)$ is the raw value of object x on feature a, and $\min a$ and $\max a$ denote the minimum and maximum values of feature a across all objects in the dataset, respectively. Through this method, the data are normalized to the interval $[0, 1]$.

(1) Comparison of Classification Accuracy

Tables 3 and 4 list the means and standard deviations of classification accuracies under the 5NN and CART classifiers, respectively. The values that are bold and underlined represent the highest accuracy on the dimensionality-reduced data.

- The proposed MSFRE outperforms other algorithms on most datasets. For the 5NN classifier, MSFRE achieves the highest accuracy 8 times; for the CART classifier, this algorithm achieves the highest accuracy 9 times.
- Although MSFRE performs poorly on some datasets (such as Car), compared with the original data, the average classification accuracy of this algorithm has improved: the 5NN classifier has improved by 4.58 %, and the CART classifier has improved by 5.66 %.
- In terms of the standard deviation of classification accuracy, MSFRE ranks second only to DRFFS under the 5NN classifier. This indicates that MSFRE has relatively high stability in the classification of dimensionality-reduced data.
- Compared with other algorithms, MSFRE performs excellently in extracting important features. The features selected at the optimal scale combination have stronger generalization ability in classification learning. For the CART classifier, compared to DRFFS, the classification accuracy of MSFRE has increased by 4.20 %; compared to NRS, it has increased by 3.99 %; compared to CFS, it has increased by 3.65 %; and compared to FNRS, it has increased by 3.64 %.

(2) Size of the Selected Feature Subset

From the perspective of knowledge reduction, these five algorithms can efficiently select key attributes and significantly reduce the number of attributes. Table 5 shows the average size of the selected feature subsets when the average accuracy reaches its peak under two classifiers through ten-fold cross-validation for different algorithms. Compared with the original data, the number of attributes selected by each algorithm has been greatly reduced.

[1] [Online]. Available: http://archive.ics.uci.edu/ml/index.php.
[2] [Online]. Available: https://jundongl.github.io/scikit-feature/datasets.HTML.
[3] [Online]. Available: http://leo.ugr.es/elvira/DBCRepository.
[4] [Online]. Available: https://sci2s.ugr.es/keel/index.php.

Table 3
Comparison of Classification Accuracy of Dimensionality-Reduced Data Using 5NN(%).

Dataset	Raw data	DRFFS	NRS	CFS	FNRS	MSFRE
Iris	94.33 ± 3.87	94.56 ± 3.78	**96.21 ± 4.66**	95.33 ± 4.50	96.00 ± 5.62	94.67±5.26
Newthyroid	90.93 ± 4.02	93.02 ± 2.45	93.01 ± 4.52	93.05 ± 5.83	93.51 ± 6.24	**94.91 ± 2.57**
Bupa	58.99 ± 3.21	**63.19 ± 4.44**	53.97 ± 12.26	55.61 ± 8.44	53.01 ± 12.09	57.94±8.87
Car	68.83 ± 1.25	68.93 ± 1.93	68.92 ± 4.33	**70.02 ± 0.17**	70.01 ± 1.17	69.61±3.03
Glass	65.24 ± 8.64	64.29 ± 6.04	63.66 ± 11.88	**71.54 ± 7.75**	66.86 ± 5.75	66.39±11.34
Wine	94.86 ± 2.95	94.86 ± 2.63	**96.67 ± 4.68**	93.24 ± 6.36	90.88 ± 10.93	96.50±7.43
Australian	75.29 ± 2.69	**86.01 ± 2.86**	71.16 ± 13.12	84.49 ± 0.70	84.35 ± 0.75	83.33±2.29
Wdbc	90.37 ± 0.88	95.04 ± 2.01	**95.38 ± 2.51**	93.85 ± 3.81	90.50 ± 2.25	95.25±3.10
Ionosphere	82.73 ± 3.63	84.57 ± 3.97	88.17 ± 4.21	88.75 ± 4.47	83.48 ± 3.01	**89.17 ± 4.93**
Dermatology	95.07 ± 2.59	93.38 ± 12.90	91.33 ± 3.86	96.31 ± 5.29	95.76 ± 4.21	**96.93 ± 3.32**
Spectfheart	74.09 ± 4.52	76.09 ± 4.94	75.28 ± 7.30	76.48 ± 7.43	**78.66 ± 2.40**	76.89±7.21
Sonar	81.46 ± 2.62	76.59 ± 6.32	81.38 ± 12.14	78.79 ± 10.06	74.17 ± 11.99	**86.55 ± 5.39**
Toxicity	59.71 ± 8.09	57.35 ± 5.92	62.20 ± 12.82	62.03 ± 8.17	56.18 ± 11.64	**62.61 ± 7.11**
Lung	84.50 ± 3.50	85.15 ± 7.17	85.24 ± 9.27	86.76 ± 6.99	87.74 ± 5.19	**93.62 ± 4.04**
Arcene	62.50 ± 5.77	63.50 ± 9.27	68.00 ± 8.23	68.54 ± 9.07	68.50 ± 8.58	**73.00 ± 11.11**
SMK_CAN_187	61.62 ± 5.67	63.51±4.81	64.71 ± 11.18	65.71 ± 8.87	65.79 ± 10.73	**76.37 ± 10.77**
Average	77.53±3.99	78.75±5.09	78.46±7.94	80.03±6.12	78.46±6.41	**82.11 ± 6.11**

Table 4
Comparison of Classification Accuracy of Dimensionality-Reduced Data Using CART(%).

Dataset	Raw data	DRFFS	NRS	CFS	FNRS	MSFRE
Iris	93.00 ± 1.76	94.33 ± 3.53	93.33 ± 7.70	93.83 ± 7.03	93.23 ± 7.03	**96.00 ± 4.66**
Newthyroid	91.16 ± 3.92	92.56 ± 4.63	93.51 ± 4.38	**95.35 ± 4.93**	94.89 ± 7.43	93.92±3.85
Bupa	62.80 ± 4.94	63.48 ± 4.67	58.59 ± 11.04	57.09 ± 6.83	55.95 ± 7.60	**63.50 ± 6.36**
Car	80.93 ± 1.31	81.22 ± 1.52	**87.78 ± 4.26**	84.02 ± 0.17	86.52 ± 2.39	82.87±2.22
Glass	63.57 ± 11.28	65.00 ± 6.55	60.82 ± 9.06	69.61 ± 6.78	64.89 ± 7.23	**70.63 ± 9.60**
Wine	88.86 ± 6.09	**92.86 ± 2.95**	92.61±6.65	92.12 ± 5.40	92.09 ± 6.13	90.52 ± 5.24
Australian	81.96 ± 2.50	83.33 ± 3.58	69.96 ± 9.14	83.51 ± 3.55	72.61 ± 3.71	**83.91 ± 4.55**
Wdbc	92.83 ± 3.30	93.01 ± 2.65	92.09 ± 2.65	90.86 ± 3.69	**97.69 ± 2.52**	95.78±2.77
Ionosphere	85.98 ± 4.70	86.57 ± 4.72	86.62 ± 5.36	92.60 ± 5.07	**93.13 ± 3.50**	92.58±3.37
Dermatology	70.23 ± 3.28	70.85 ± 13.88	94.13 ± 3.82	78.48 ± 4.39	86.58 ± 5.72	**95.27 ± 3.22**
Spectfheart	72.64 ± 6.43	73.21 ± 3.75	72.26 ± 5.15	72.68 ± 7.34	76.07 ± 4.14	**80.51 ± 7.86**
Sonar	71.71 ± 5.66	74.39 ± 6.00	73.50 ± 8.58	74.98 ± 11.51	74.40 ± 11.59	**78.38 ± 5.58**
Toxicity	52.65 ± 9.35	60.00 ± 8.20	**64.35 ± 7.43**	53.73 ± 13.05	52.71 ± 16.13	60.13±10.90
Lung	89.00 ± 3.37	84.75 ± 8.62	84.81 ± 6.95	82.76 ± 7.88	87.26 ± 7.62	**93.17 ± 6.95**
Arcene	69.75 ± 4.78	67.50 ± 11.68	66.50 ± 11.32	71.00 ± 10.22	68.50 ± 11.07	**76.50 ± 11.07**
SMK_CAN_187	60.81 ± 9.56	**68.08 ± 4.80**	63.64 ± 10.72	67.31 ± 10.93	63.63 ± 12.37	64.74±8.66
Average	76.74±5.14	78.20±5.73	78.41±7.14	78.75±6.80	78.76±7.26	**82.40 ± 6.05**

Table 5
Average Size of Feature Subsets Selected by Different Algorithms.

Dataset	Raw data	DRFFS	NRS	CFS	FNRS	MSFRE
Iris	4	4	4	4	2	2
Newthyroid	5	5	5	4	5	4
Bupa	6	5	3	4	3	3
Car	6	6	2	5	1	4
Glass	9	9	8	6	8	6
Wine	13	6	9	5	5	8
Australian	14	12	3	1	1	4
Wdbc	30	10	13	3	1	4
Ionosphere	34	15	12	3	2	3
Dermatology	34	7	10	6	13	7
Spectfheart	44	5	24	5	1	1
Sonar	60	8	12	7	19	1
Toxicity	1203	2	12	8	3	15
Lung	3312	3	7	11	6	10
Arcene	10000	13	10	4	7	8
SMK_CAN_187	19993	2	9	9	9	24
Average	2172.9	7.0	8.8	5.2	5.3	6.5

Table 6
Optimal Attribute Subsets of DRFFS and MSFRE Algorithms.

Dataset	DRFFS	MSFRE
Iris	3, 4, 1, 2	4, 3
Newthyroid	4, 5, 1, 2, 3	2, 5, 1, 4
Bupa	3, 6, 1, 2, 4	2, 5, 1
Car	1, 2, 3, 4, 5, 6	6, 5, 4, 3
Glass	3, 4, 5, 1, 2, 6, 7, 8, 9	3, 4, 1, 9, 5, 2
Wine	1, 5, 7, 10, 12, 13	12, 13, 1, 2, 5, 11, 10, 7
Australian	5, 6, 1, 2, 4, 13, 14, 8, 9, 10, 11, 12	8, 9, 12, 4
Wdbc	2, 9, 12, 19, 22, 23, 25, 28, 29, 30	28, 21, 8, 22
Ionosphere	1, 3, 6, 7, 9, 14, 16, 17, 20, 25, 26, 28, 30, 33 ,34	28, 16, 4
Dermatology	8, 19, 5, 6, 28, 29, 34	16, 28, 15, 29, 5, 31, 3
Spectfheart	6, 26, 32, 40, 41	11
Sonar	1, 27, 30, 11, 36, 45, 12, 53	10
Toxicity	39, 204	585, 561, 701, 359, 889, 26, 227, 1145, 100, 236, 626, 840, 870, 627, 312
Lung	491, 2153, 2841	1464, 3178, 614, 2579, 1792, 2750, 1111, 580, 2702, 1765
Arcene	1601, 1780, 5580, 6657, 2496, 2533, 3008, 3358, 4645, 5333, 6929, 8186, 7813	5818, 766, 9986, 9699, 8769, 7205, 8146, 2101
SMK_CAN_187	5702, 6109	6676, 17896, 10156, 12503, 13576, 1887, 11471, 11111, 11394, 2131, 18779, 4161, 12075, 1850, 7557, 19672, 16262, 7229, 8531, 18940, 16072, 5702, 17546, 16583

It can be seen from Table 5 that these reduction methods can effectively reduce the number of attributes. In most cases, MSFRE selects fewer features than the other four algorithms. For the Car dataset, MSFRE selects fewer than DRFFS and CFS. For the Glass dataset, MSFRE selects 6 features, same as CFS. It selects fewer than DRFFS, NRS, and FNRS. This shows that the proposed algorithm is more effective in reducing redundant attributes.

For low-to-medium dimensional data (e.g., Spectfheart with 44 raw features or Sonar with 60 raw features), critical discriminative information resides in a small number of high-quality features. MSFRE's mechanism identifies these information-dense features, enabling it to maintain peak accuracy with just one feature.

For ultra-high-dimensional data (e.g., SMK_CAN_187 with 19,993 raw features), discriminative information is dispersed across weakly correlated subsets. Selecting 2 features with DRFFS or 9 features with NRS leads to information loss due to excessive dimensionality reduction. In contrast, MSFRE selects 24 features, covering all necessary discriminative subsets while balancing dimensionality reduction and performance.

In summary, MSFRE does not merely pursue minimal dimensionality. It adapts to the data's information distribution and task requirements, aligning with the goal of knowledge reduction. This highlights the importance of adaptive feature selection, especially when balancing reduction and information integrity in high-dimensional data.

The optimal feature subsets of DRFFS and MSFRE are shown in Table 6. Table 6 lists the optimal attribute subsets when DRFFS and MSFRE achieve the highest accuracy. It can be found that on most datasets, such as Iris, Newthyroid, Puba, etc., the optimal features selected by MSFRE are mostly subsets of the optimal features selected by DRFFS. Take the Iris dataset as an example. DRFFS selects features (3, 4, 1, 2), while MSFRE selects features (4, 3). The feature set of MSFRE is contained within the feature set of DRFFS. For other datasets, such as Car and Glass, although the optimal features selected by the two algorithms differ, there are common features in the feature subsets. This indicates that for a given classification task, there are multiple feature subsets with acceptable classification capabilities.

Overall, this shows that MSFRE tends to select more concise subsets in feature selection, possibly having certain advantages in removing redundant features. However, the situations vary across different datasets, which also reflects the diversity in feature selection strategies and effects among different algorithms.

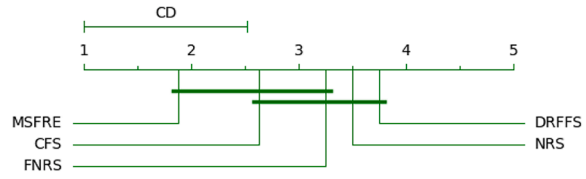
(3) Statistical tests

Additionally, the Friedman statistical test [53], the Nemenyi statistical test [54] and the Bonferroni-Dunn statistical test [54] are used to assess the statistical differences between MSFRE and other algorithms. These two significance test methods are widely used because they can utilize all the information in the related samples.

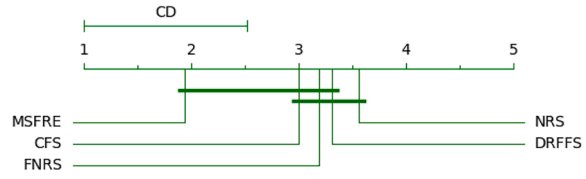
Let t be the number of algorithms for comparison, n be the number of datasets, and A_r be the mean rank of algorithm r. By Friedman test, when the original hypothesis holds, F_F follows a Fisher distribution

$$F_F = \frac{(n-1)\chi_F^2}{n(t-1)-\chi_F^2} \sim F(t-1,(t-1)(n-1)) \tag{14}$$

where $\chi_F^2 = \frac{12n}{t(t+1)}(\sum_{r=1}^{t} A_r^2 - \frac{t(t+1)^2}{4})$.

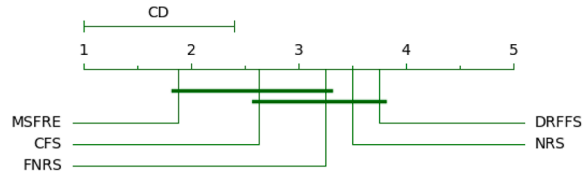


(a) 5NN classifier

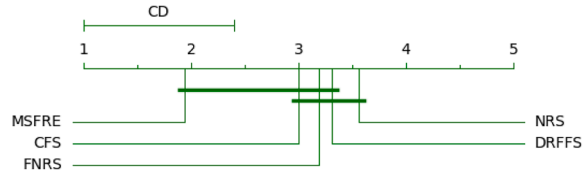


(b) CART classifier

Fig. 3. CD diagram of Nemenyi test.



(a) 5NN classifier



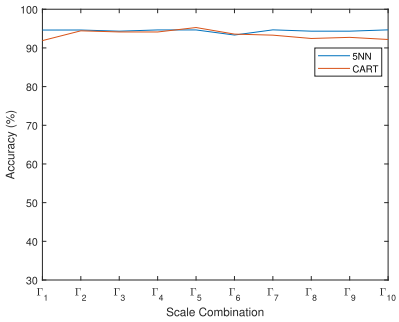
(b) CART classifier

Fig. 4. CD diagram of Bonferroni-Dunn test.

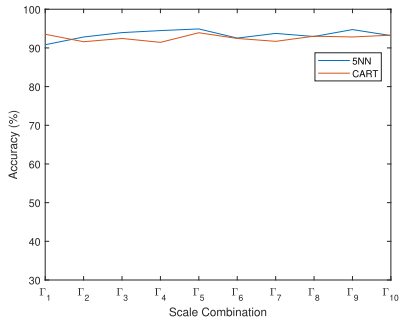
Table 7
Statistical Test of Five Models with Two Classifiers.

Classifiers	Average Ranking					χ_F^2	F_F
	DRFFS	NRS	CFS	FNRS	MSFRE		
5NN	3.75	3.50	2.63	3.25	1.88	14.60	4.43
CART	3.31	3.56	3.00	3.19	1.94	10.10	2.81

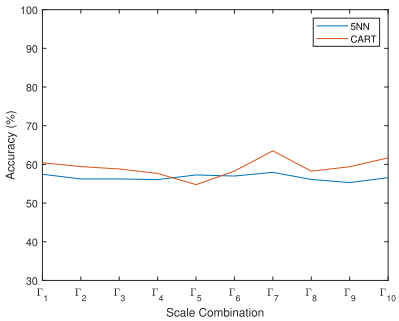
Table 7 lists the statistical test of five models with two classifiers. At level $\alpha = 0.05$, the critical value of $F(4.60)$ is 2.5252. The F_F values under 5NN and CART classifiers are 4.43 and 2.81, respectively, which are both greater than the critical value. Form Table 7, we reject the original hypothesis at $\alpha = 0.05$. In this case, the five algorithms are considered to be markedly different in terms of classification performance. Therefore, the Nemenyi test and the Bonferroni–Dunn test are commonly used to further explore the



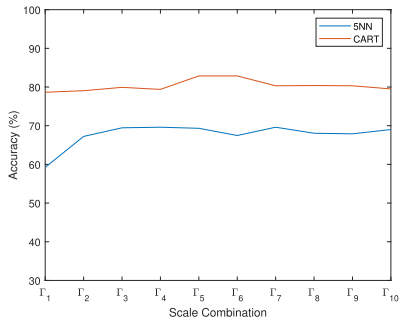
(a). Iris dataset



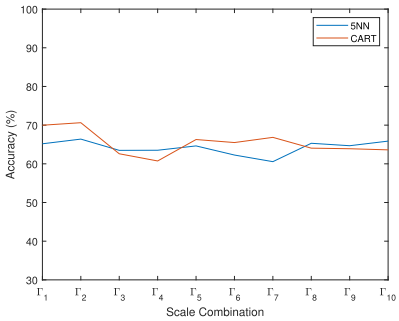
(b). Newthyroid dataset



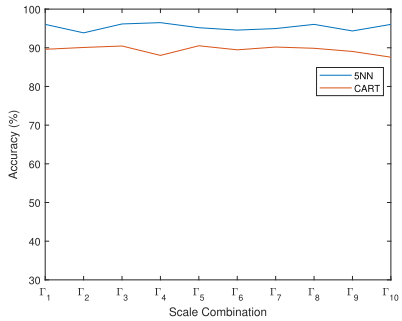
(c). Bupa dataset



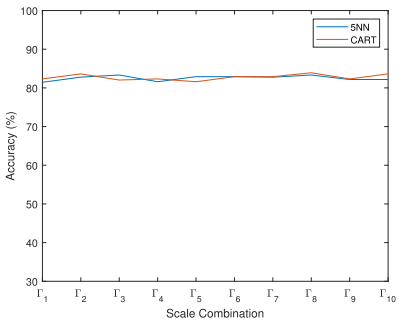
(d). Car dataset



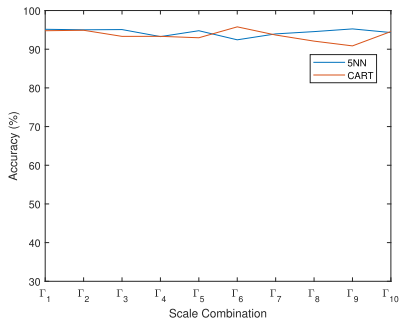
(e). Glass dataset



(f). Wine dataset

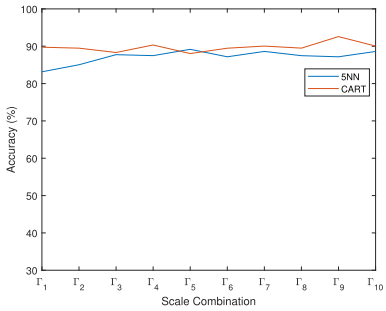


(g). Australian dataset

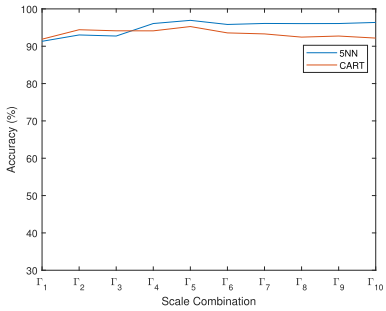


(h). Wdbc dataset

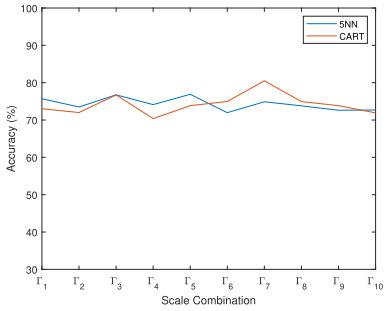
Fig. 5. The classification accuracy of different scale combinations.



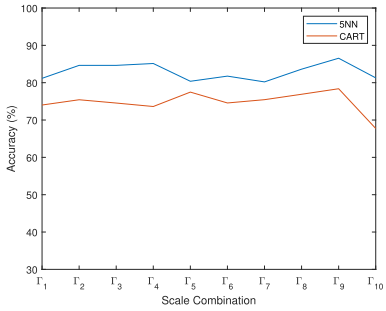
(i). Ionosphere dataset



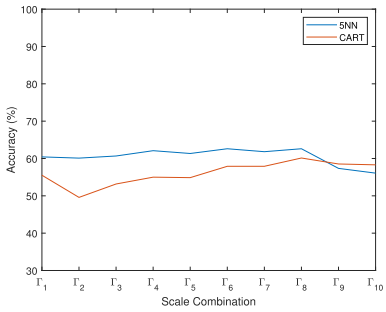
(j). Dermatology dataset



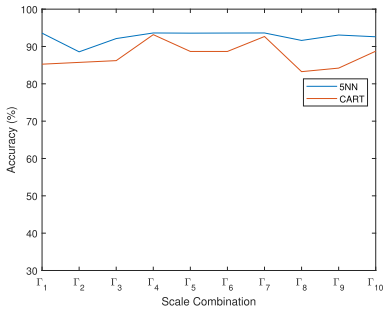
(k). Spectfheart dataset



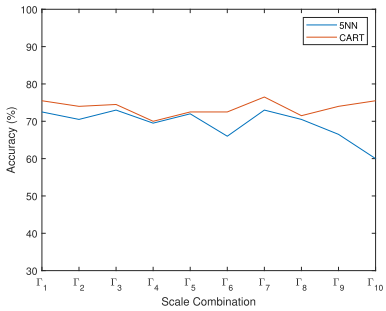
(l). Sonar dataset



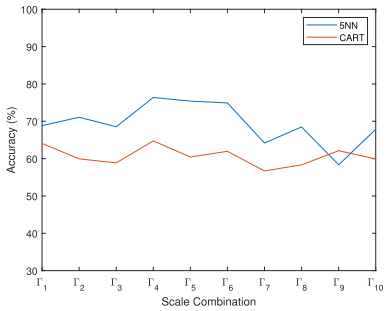
(m). Toxicity dataset



(n). Lung dataset



(o). Arcene dataset



(p). SMK_CAN_187 dataset

Fig. 5. Continued.

differences in performance of the five models. The critical distance of the test is

$$\mathrm{CD}_\alpha = c_\alpha \sqrt{\frac{t(t+1)}{6n}} \tag{15}$$

where c_α is a critical value.

In reference [54], the CD diagram is used to intuitively demonstrate the differences. If there is not a line connecting two algorithms, they are significantly different. The CD diagram of the Nemenyi test and the Bonferroni-Dunn test is shown in Fig. 3 and Fig. 4, the Nemenyi test result is $c_{0.05} = 2.728$, and $\mathrm{CD}_{0.05} = 1.52$ ($t = 5$, $n = 16$). For the Bonferroni-Dunn test, we can calculated $c_{0.05} = 2.343$, and $\mathrm{CD}_{0.05} = 1.31$, ($t = 5$, $n = 16$). From Fig. 3 and Fig. 4, the following conclusions can be drawn.

- For the 5NN classifier: Both the Nemenyi test and Bonferroni-Dunn test yield consistent results: at the significance level $\alpha = 0.05$, MSFRE significantly outperforms NRS and DRFFS, and exhibits competitive levels of performance compared to CFS and FNRS.
- For the CART classifier: under the Nemenyi test and Bonferroni-Dunn test, at a significance level of $\alpha = 0.05$, MSFRE significantly outperforms NRS.

Overall, MSFRE provides competitive performance compared with the other four algorithms. Finally, Fig. 5 shows the classification accuracy curves of the 5NN classifier and the CART classifier when the scale combination changes. For simplicity, Γ_1 denotes that the scale of each attribute is set to the first one simultaneously, that is, $\delta_1 = 0.1$. Γ_2 denotes that the scale of each attribute is set to the second one simultaneously, that is, $\delta_2 = 0.2$. Similarly, other scale combinations are defined in the same way, where Γ_1 is the finest scale, Γ_{10} is the coarsest scale, and $\Gamma_2 - \Gamma_9$ are intermediate scales. From these figures, the following observations can be drawn.

• Most datasets achieve peak accuracy and stability at intermediate scales ($\Gamma_3 - \Gamma_7$). For instance, the CART classifier on the Iris dataset reaches 96.00 % accuracy at Γ_4, with accuracy declining when deviating from this scale. This demonstrates that MSFRE effectively balances details and noise, ensuring the algorithm's stability.

• High-dimensional datasets are sensitive to fine scales ($\Gamma_1 - \Gamma_4$). For example, the 5NN classifier on the Arcene dataset achieves 73.00 % accuracy at Γ_3, while accuracy plummets to 60.00 % at the coarse scale (Γ_{10}). Fine scales preserve weak discriminative signals in high-dimensional data, whereas coarse scales discard critical information.

• Discrete datasets exhibit sensitivity to coarse scales ($\Gamma_7 - \Gamma_{10}$). Their performance improves only at coarse scales: the 5NN classifier on the Car dataset reaches 69.61 % accuracy at Γ_7, while at the fine scale (Γ_1), it overfits to 59.22 % due to noise. Coarse scales filter interference and focus on core discriminative features.

Consequently, intermediate scales are most appropriate for continuous data, fine scales for high-dimensional data, and coarse scales for discrete data. In practical applications, scales can be matched to dataset characteristics accordingly, which aligns with human multi-granularity thinking and enhances classification efficiency.

6. Conclusion and future work

In this article, we have introduced the idea of δ-fuzzy relation into multi-scale models and formalized a new decision system model, i.e., multi-scale fuzzy relation decision system. Fuzzy scale entropy and some of its variants are then presented to characterize the differences in classification ability of feature subsets at different scales. Moreover, feature selection is addressed from the view of maintaining the classification ability of fuzzy relations family, and MSFRE is designed to reduce redundant features by quantifying conditional fuzzy scale entropy. Compared to single-scale fuzzy rough methods, fuzzy scale entropy effectively captures differences in classification capabilities across multiple scales. Extensive numerical experiments showed that the presented MSFRE can effectively reduce redundant attributes, and obtain better classification accuracy. Future work may include.

(1) At present, most multi-scale models are designed for static data, while real data is often dynamically changing. Therefore, considering dynamic update algorithms for optimal scales and exploring dynamic acquisition of decision rules would be a meaningful task.

(2) An adaptive scale selection mechanism will be developed to automatically optimize the scale parameter δ_k based on data characteristics, reducing manual tuning dependency and improving model scalability for high dimensional datasets.

CRediT authorship contribution statement

Jiaying Wang: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation; **Zhehuang Huang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization; **Zhifeng Weng:** Writing – review & editing, Validation, Supervision, Funding acquisition; **Jinjin Li:** Writing – review & editing, Supervision, Project administration, Methodology.

Data availability

Data will be made available on request.

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 (No. 12271191), and the University Industry-University-Research Joint Innovation Project Plan in Fujian Province (2024H6027).

References

[1] X. L. Yang, H. M. Chen, T. R. Li, S. Feng, J. H. Wan, Y. Y. Yao, Adaptive feature selection with weighted fuzzy rough sets for noisy data, Fuzzy Sets Syst. 517 (2025) 109499.
[2] Z. H. Huang, J. J. Li, Covering based multi-granulation rough fuzzy sets with applications to feature selection, Expert Syst. Appl. 238 (2024) 121908.
[3] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.
[4] C. Z. Wang, Y. H. Qian, W. P. Ding, X. D. Fan, Feature selection with fuzzy-rough minimum classification error criterion, IEEE Trans. Fuzzy Syst. 30 (2022) 2930–2942.
[5] K. Y. Liu, T. R. Li, X. B. Yang, H. M. Chen, J. Wang, Z. X. Deng, Semifree: semisupervised feature selection with fuzzy relevance and redundancy, IEEE Trans. Fuzzy Syst. 31 (2023) 3384–3396.
[6] Z. H. Huang, J. J. Li, C. Z. Wang, Robust feature selection using multigranulation variable-precision distinguishing indicators for fuzzy covering decision systems, IEEE Trans. Syst. Man, Cybern. Syst. 54 (2024) 903–914.
[7] 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.
[8] T. Y. Lin, Granular computing on binary relations II: rough set representations and belief functions, rough sets knowl, 1998, pp. 121–140.
[9] Y. X. Chen, Z. H. Huang, J. J. Li, Fuzzy neighborhood based variable-precision granular-ball rough sets with applications to feature selection, Fuzzy Sets Syst. 512 (2025) 109382.
[10] C. Gao, X. F. Tan, J. Zhou, W. P. Ding, W. Pedrycz, Fuzzy granule density-based outlier detection with multi-scale granular balls, IEEE Trans. Knowl. Data Eng. 37 (2025) 1182–1197.
[11] W. Z. Wu, Y. Leung, Optimal scale selection for multi-scale decision tables, Int. J. Approx. Reason. 54 (2013) 1107–1129.
[12] Z. H. Huang, J. J. Li, Feature subset selection with multi-scale fuzzy granulation, IEEE Trans. Artif. Intell. 4 (2023) 121–134.
[13] L. Sun, S. S. Si, W. P. Ding, X. Y. Wang, J. C. Xu, TFSFB: two-stage feature selection via fusing fuzzy multi-neighborhood rough set with binary whale optimization for imbalanced data, Inf. Fusion 95 (2023) 91–108.
[14] W. Z. Wu, Y. Leung, Theory and applications of granular labeled partitions in multi-scale decision tables, Inf. Sci. 181 (2011) 3878–3897.
[15] F. Li, B. Q. Hu, J. Wang, Stepwise optimal scale selection for multi-scale decision tables via attribute significance, Knowl. Based Syst. 129 (2017) 4–16.
[16] F. Li, B. Q. Hu, A new approach of optimal scale selection to multi-scale decision tables, Inf. Sci. 381 (2017) 193–208.
[17] Z. H. Huang, J. J. Li, Multi-scale covering rough sets with applications to data classification, Appl. Soft Comput. 110 (2021) 107736.
[18] Z. H. Xie, W. Z. Wu, L. X. Wang, Optimal scale combination selection for incomplete generalized multi-scale ordered information systems, IEEE Trans. Artif. Intell. 4 (2023) 432–446.
[19] Y. H. She, Z. J. Zhao, M. T. Hu, W. L. Zheng, X. L. He, On selection of optimal cuts in complete multi-scale decision tables, Artif. Intell. Rev. 54 (2021) 6125–6148.
[20] Q. H. Zhang, Y. L. Cheng, F. Zhao, G. Y. Wang, S. Y. Xia, Optimal scale combination selection integrating three-way decision with hasse diagram, IEEE Trans. Neural Netw. Learn. Syst. 33 (2022) 2437–2449.
[21] W. Z. Wu, Y. H. Qian, T. J. Li, S. M. Gu, On rule acquisition in incomplete multi-scale decision tables, Inf. Sci. 378 (2017) 282–302.
[22] J. M. Zhan, K. Zhang, W. Z. Wu, An investigation on wu-leung multi-scale information systems and multi-expert group decision-making, Expert Syst. Appl. 170 (2021) 114542.
[23] J. Deng, J. Zhan, W. Z. Wu, A three-way decision methodology to multi-attribute decision-making in multi-scale decision information systems, Inf. Sci. 568 (2021) 175–198.
[24] Y. H. She, Z. H. Qian, X. L. He, J. T. Wang, T. Qian, W. L. Zheng, On generalization reducts in multi-scale decision tables, Inf. Sci. 555 (2021) 104–124.
[25] D. Xia, G. Wang, Q. Zhang, J. Yang, S. Xia, Three-way approximations fusion with granular-ball computing, IEEE Trans. Fuzzy Syst. 32 (2024) 5963–5977.
[26] Y. Yang, Q. H. Zhang, F. Zhao, Y. L. Cheng, Q. Xie, G. Y. Wang, Optimal scale combination selection based on genetic algorithm in generalized multi-scale decision systems for classification, Inf. Sci. 693 (2025) 121685.
[27] J. Deng, L. Wei, C. J. Qiu, L. J. Zhang, Granular ball-based feature subset selection for incomplete generalized double multi-scale decision tables, Int. J. Mach. Learn. 2025, p. 101007.
[28] P. F. Zhang, D. X. Wang, Z. Yu, Y. J. Zhang, T. Jiang, T. R. Li, A multi-scale information fusion-based multiple correlations for unsupervised attribute selection, Inf. Fusion, 106 (2024) 102276.
[29] X. Wang, F. Yin, M. Yin, Y. Yang, Incomplete generalized multi-scale ordered information systems and optimal scale combination selection, in: Rd International Symposium on Computer Technology and Information Science, 2023, pp. 796–802.
[30] Y. B. Xiao, J. M. Zhan, C. Zhang, P. D. Liu, A sequential three-way decision-based group consensus method with regret theory under interval multi-scale decision information systems, IEEE Trans. Emerg. Top. Comput. Intell. 8 (2024) 1670–1686.
[31] R. Li, C. Zhang, D. Y. Li, W. T. Li, J. M. Zhan, Improved evidential three-way decisions in incomplete multi-scale information systems, Int. J. Approx. Reasoning. 2025, p. 109417.
[32] T. Y. Wang, B. Yang, Optimal scale selection of dynamic incomplete generalized multi-scale fuzzy ordered decision systems based on rough fuzzy sets, Fuzzy Sets Syst. 515 (2025) 109420.
[33] Z. H. Huang, J. J. Li, W. Z. Dai, R. D. Lin, Generalized multi-scale decision tables with multi-scale decision attributes, Int. J. Approx. Reasoning. 2019, pp. 194–208.
[34] C. Wang, S. An, W. Ding, Y. Qian, Feature selection and classification based on directed fuzzy rough sets, IEEE Trans. Syst. Man, Cybern. Syst. 55 (2025) 699–711.
[35] Y. B. Xiao, X. L. Ma, J. M. Zhan, Group decision-making in heterogeneous multi-scale information fusion: integrating overconfident and non-cooperative behaviors, Inf. Fusion 125 (2026) 103401.
[36] L. X. Wang, W. Z. Wu, Z. H. Xie, A. H. Tan, Optimal scale combination selection in generalized multi-scale hybrid decision systems, Inf. Sci. 689 (2025) 121429.
[37] D. Miao, G. Wang, Information-theoretic characterization of knowledge in rough set theory, Int. J. Gen. Syst. 27 (1998) 479–488.
[38] Y. H. Qian, J. Y. Liang, An entropy-based knowledge granulation for evaluating the uncertainty of discrete variables in knowledge-distributing ability, Int. J. Gen. Syst. 34 (2005) 449–465.
[39] J. S. Dai, Y. H. Chen, D. Y. Li, A new-type of conditional entropy for measuring the importance of attributes in incomplete decision systems, Knowl. Based Syst. 24 (2011) 1303–1310.
[40] R. R. Yager, On the entropy of fuzzy sets, Inf. Sci. 18 (1979) 1–15.
[41] C. Bertoluzza, On the measure of the uncertainty of fuzzy partitions, Fuzzy Sets Syst. 97 (1998) 287–292.
[42] Z. Zhu, C. Zhang, J. Dai, Fuzzy information quantity measurement and feature selection by macrogranular entropy, IEEE. Trans. Artif. Intell. 2025, p. 101109.
[43] B. Y. Chen, Z. Yuan, Z. Liu, D. Z. Peng, Y. X. Li, C. Liu, G. D. Duan, Outlier detection in mixed-attribute data: a semi-supervised approach with fuzzy approximations and relative entropy, Int. J. Approx. Reason. 179 (2025) 109373.
[44] C. L. Xu, Y. S. Lan, A multi-criteria decision-making method based on intuitionistic fuzzy entropy and three-way ranking topsis with applications, Int. J. Fuzzy Syst. 24 (2025) 1954.
[45] J. Hamidzadeh, Z. Mehravaran, A. Harati, Feature selection by utilizing kernel-based fuzzy rough set and entropy-based non-dominated sorting genetic algorithm in multi-label data, Knowl. Inf. Syst. 67 (2025) 3789–3819.

[46] S. Y. Yang, Z. Yuan, C. Luo, H. M. Chen, D. Z. Peng, Fuzzy multi-neighborhood entropy-based interactive feature selection for unsupervised outlier detection, Appl. Soft Comput. 169 (2025) 112572.
[47] W. H. Xu, W. R. Ye, Incremental feature selection: parallel approach with local neighborhood rough sets and composite entropy, Pattern Recognit. 159 (2025) 111141.
[48] X. Y. Zhang, W. C. Zhao, Uncertainty measures and feature selection based on composite entropy for generalized multigranulation fuzzy neighborhood rough set, Fuzzy Sets Syst. 486 (2024) 108971.
[49] J. C. Xu, M. X. Ma, S. Zhang, W. L. Niu, Feature selection based on multi-perspective dynamic neighbourhood entropy measures in a dynamic neighbourhood rough set, Appl. Intell. 25 (2025) 55–441.
[50] Z. H. Xie, W. Z. Wu, L. X. Wang, A. H. Tan, Entropy based optimal scale selection and attribute reduction in multi-scale interval-set decision tables, Int. J. Mach. Learn. Cybern. 15 (2024) 3005–3026.
[51] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (2008) 3577–3594.
[52] C. Siregar, B. C. Octariadi, Feature selection for sambas traditional fabric 'kain Lunggi' using correlation-based featured selection, in: International Conference on Data and Software Engineering (ICoDSE), Pontianak, Indonesia, 2019, pp. 1–5.
[53] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1940) 86–92.
[54] O. J. Dunn, Multiple comparisons among means, J. Amer. Stat. Assoc. 56 (1961) 52–64.