Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060833

ARTICLE

A Method for Fast Feature Selection Utilizing Cross-Similarity within the Context of Fuzzy Relations

Wenchang Yu[1], Xiaoqin Ma[1,2], Zheqing Zhang[1] and Qinli Zhang[1,2,*]

[1]School of Big Data and Artificial Intelligence, Chizhou University, Chizhou, 247000, China

[2]Anhui Education Big Data Intelligent Perception and Application Engineering Research Center, Anhui Provincial Joint Construction Key Laboratory of Intelligent Education Equipment and Technology, Chizhou, 247000, China

*Corresponding Author: Qinli Zhang. Email: zqlynu@163.com

Received: 11 November 2024; Accepted: 28 January 2025; Published: 26 March 2025

ABSTRACT: Feature selection methods rooted in rough sets confront two notable limitations: their high computational complexity and sensitivity to noise, rendering them impractical for managing large-scale and noisy datasets. The primary issue stems from these methods' undue reliance on all samples. To overcome these challenges, we introduce the concept of cross-similarity grounded in a robust fuzzy relation and design a rapid and robust feature selection algorithm. Firstly, we construct a robust fuzzy relation by introducing a truncation parameter. Then, based on this fuzzy relation, we propose the concept of cross-similarity, which emphasizes the sample-to-sample similarity relations that uniquely determine feature importance, rather than considering all such relations equally. After studying the manifestations and properties of cross-similarity across different fuzzy granularities, we propose a forward greedy feature selection algorithm that leverages cross-similarity as the foundation for information measurement. This algorithm significantly reduces the time complexity from $O(m^2n^2)$ to $O(mn^2)$. Experimental findings reveal that the average runtime of five state-of-the-art comparison algorithms is roughly 3.7 times longer than our algorithm, while our algorithm achieves an average accuracy that surpasses those of the five comparison algorithms by approximately 3.52%. This underscores the effectiveness of our approach. This paper paves the way for applying feature selection algorithms grounded in fuzzy rough sets to large-scale gene datasets.

KEYWORDS: Fuzzy rough sets; feature selection; cross-similarity; fuzzy relations

1 Introduction

With the evolution of information technology, a substantial amount of information is captured in the form of data. These datasets contain a substantial quantity of redundant information. In data mining and model training, these redundant features can increase computational complexity and even result in model distortion, ultimately affecting the model's predictive capability. Feature selection (or attribute selection, attribute reduction) is a crucial technique to minimize redundant features, aiming to identify the smallest subset of features that retain the distinguishability of the data.

Rough set theory [1] enjoys widespread application across diverse fields [2–4], with feature selection emerging as a particularly prominent research area within its scope. However, a limitation of rough set theory is its inability to handle continuous data directly. When dealing with features with continuous values, discretization is necessary, which can lead to the loss of valuable data information. To address these limitations of traditional rough set theory, Dubois et al. [5] introduced a novel concept by integrating fuzzy

sets with rough sets, resulting in the emergence of a fuzzy rough set (*FRS*). This extension of Pawlak's rough set model to the fuzzy domain represents a significant advancement in the field. Since then, researchers have conducted thorough investigations into fuzzy rough set theory, further refining its theoretical framework, alleviating constraints in practical applications, mitigating the impact of noisy data, and enhancing its overall robustness. Some researchers [6,7] have endeavored to generalize the fuzzy relationships among samples in fuzzy rough sets, broadening their scope and enhancing the model's adaptability. Additionally, they have gone beyond the traditional definitions of the upper and lower approximation operators, giving the model a more comprehensive and flexible framework. These innovative extensions not only increase the model's flexibility but also enable it to handle a wider range of practical applications with greater precision and adaptability. Despite the efforts of some researchers [8–11] to design more reasonable relationships and upper-lower approximation operations in order to mitigate the impact of noisy data and enhance the robustness of these fuzzy rough set models, they are often accompanied by increased computational complexity. Some researchers [12–16] have broadened the concept of fuzzy partition to fuzzy covering, thereby easing the constraints on the model's application and enhancing its adaptability to diverse scenarios. This generalization not only facilitates a more flexible application of the model but also empowers it to tackle a broader spectrum of practical problems with enhanced precision and accuracy.

With the evolution of fuzzy rough set theory, numerous feature selection algorithms have emerged, built upon this theory. These algorithms aim to identify and extract salient features from datasets, aiding in their effective representation and interpretation. Jensen et al. [17] introduced the concepts of the fuzzy positive region and the fuzzy-rough dependency function, subsequently developing a feature selection model based on this dependency. Yang et al. [18] proposed a granular matrix and a granular reduction model, informed by fuzzy β-covering theory. Huang et al. [19,20] introduced a novel discernibility measure and a novel fitting model for attribute reduction, utilizing the fuzzy β-neighborhood to capture the distinguishing power of a fuzzy covering family. Huang et al. [21] devised a robust rough set model that integrates fuzzy rough sets, covering-based rough sets, and multi-granulation rough sets. Furthermore, he successfully employed this model in the realm of feature selection. Wan et al. [22] proposed a novel interactive and complementary feature selection approach, which is based on the fuzzy multi-neighborhood rough set model, specifically designed for use with fuzzy and uncertain data. Hu et al. [23] developed a fuzzy multi-objective feature selection method, incorporating a fuzzy dominance relationship and particle swarm optimization. Deng et al. [24] formulated the dual similarity by leveraging the concept of neighborhood fuzzy entropy, and subsequently utilized it for feature selection in label distribution learning. Dai et al. [25] introduced the concept of multi-fuzzy β-covering approximation spaces and subsequently proposed a forward greedy fuzzy β-covering reduction algorithm, which is based on the monotone conditional entropy in a multi-fuzzy β-covering approximation space.

Current feature selection algorithms predominantly employ forward greedy feature selection, that follows a sequential process, consisting of several distinct steps:

1) Determine an evaluation metric.
2) Model initialization: commences with an empty set as the selected feature subset and a candidate feature subset consisting of all features.
3) Feature selection: in each iteration, incorporate one feature from the candidate feature subset to the already selected feature subset according to the evaluation metric. This process continues until the stopping criteria are satisfied.

When dealing with a dataset containing n features and m samples, the forward greedy feature selection algorithm, in the worst-case scenario, necessitates the calculation of the evaluation metric $n(n-1)/2$ times. Moreover, for each evaluation of the evaluation metric, the calculation of relationships between every pair

of samples in the dataset is essential, which necessitates calculating m^2 relationships between samples, such as distances or similarities. Consequently, the time complexity of these algorithms typically lies in the order of $O(m^2n^2)$ or higher. A significant challenge faced by these feature selection algorithms is their high computational complexity, which makes it challenging for them to effectively handle large sample sizes or high-dimensional data in practical settings.

Through rigorous research and meticulous analysis, we have identified the following issues with existing uncertainty measures and their corresponding feature selection algorithms:

1) Existing feature selection algorithms must consider the relationships between all samples when calculating uncertainty measures (based on dependency or conditional information entropy).
2) Existing sample relationships may exhibit weak associations due to noisy data, which can result in the selection of redundant features.
3) Existing feature selection algorithms struggle to process large sample sizes or high-dimensional datasets effectively due to their high computational complexity.

Drawing upon the insights obtained, we introduce a novel and efficient feature selection algorithm that is grounded in cross-similarity. This algorithm represents a significant advancement in addressing the above challenges associated with feature selection. The primary contributions of this paper are outlined as follows:

1) We establish a fuzzy binary relation known as the "cross-similarity relation", which is highly sensitive to samples and relationships that impact the uncertainty of the decision tables while ignoring samples that are already consistent in the decision table.
2) Using the cross-similarity relation as the uncertainty measure, we propose a new feature selection algorithm. Compared to existing feature selection algorithms, our algorithm reduces the time complexity from $O(m^2n^2)$ to $O(mn^2)$.
3) To mitigate the impact of noise-induced weak relationships, we introduce a truncation parameter that effectively reduces the influence of noisy data on the feature selection algorithm.

The paper is organized as follows: Section 2 covers the basics of similar relations and fuzzy rough sets. Section 3 introduces cross-similarity relations and their significance, along with their key properties. Section 4 details our new forward greedy feature selection model using fuzzy cross-similarity for data analysis. Section 5 presents experiments proving our method's effectiveness, stability, and efficiency. Finally, Section 6 summarizes our findings, limitations of the algorithm, and further research directions.

2 Preliminaries

In this section, we will delve into the fundamental concepts of similar relations and fuzzy rough sets.

2.1 Similar Relations

In an information table, a binary relation among samples can be constructed based on the feature set. This binary relation is generally either a similarity relation [5,20,25] or a distance relation [7,26].

Definition 1. *Let $U = \{x_1, x_2, \ldots, x_m\}$ be a set of samples, $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be a set of conditional features, and $R = \{r_1, r_2, \ldots, r_n\}$ be a set of similar relations between samples, where $r_i \in R(i = 1, 2, \cdots, n)$ is a fuzzy compatible relation derived from the corresponding feature $a_i \in A$, signifying that $r_i(x_j, x_k) \in [0,1](x_j, x_k \in U)$ and r_i fulfills the properties of reflexivity and symmetry. Then, the pair (U, R) is referred to as a fuzzy compatible relation space (FCRS).*

Definition 2. *Let $(U; R)$ be an FCRS, $x, y \in U, \mathcal{G} \subseteq R$, and $Sim_{\mathcal{G}}(x, y) = \bigwedge_{r_i \in \mathcal{G}} r_i(x, y)$. Then $Sim_{\mathcal{G}}$ is called a multi-attribute fuzzy compatible relation concerning $\mathcal{G}$.*

Definition 3. *A parameterized multi-attribute fuzzy compatible relation is defined as follows:*

$$Sim_{\mathcal{G}}^{\lambda}(x,y) = \begin{cases} 0, Sim_{\mathcal{G}}^{\lambda}(x,y) < \lambda \\ Sim_{\mathcal{G}}^{\lambda}(x,y), Sim_{\mathcal{G}}^{\lambda}(x,y) \geq \lambda \end{cases} \tag{1}$$

where $0 \leq \lambda < 1$.

The truncation parameter λ can be used to eliminate weak similarity relationships introduced by noisy data in the dataset. $\mathrm{Sim}_{\mathcal{G}}^{\lambda}$ is capable of filtering out weak similarity relationships that exert influence on data uncertainty.

Proposition 4. *Let* (U,R) *be an FCRS. It holds that, for any* $x, y \in U$,

1) *if* $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq R$, *then* $Sim_{\mathcal{G}_1}^{\lambda}(x,y) \geq Sim_{\mathcal{G}_2}^{\lambda}(x,y)$;
2) *if* $\lambda_1 \leq \lambda_2$, *then* $Sim_{\mathcal{G}}^{\lambda_1}(x,y) \geq Sim_{\mathcal{G}}^{\lambda_2}(x,y)$.

Proof. 1) For any $\mathrm{r_i} \in \mathcal{G}_1$, $\mathcal{G}_1 \subseteq \mathcal{G}_2 \Longrightarrow \mathrm{r_i} \in \mathcal{G}_2$, $\mathrm{Sim}_{\mathcal{G}_1}^{\lambda}(\mathrm{x},\mathrm{y}) = \bigwedge_{\mathrm{r_i} \in \mathcal{G}_1} \mathrm{r_i}(\mathrm{x},\mathrm{y}) \geq \bigwedge_{\mathrm{r_i} \in \mathcal{G}_2} \mathrm{r_i}(\mathrm{x},\mathrm{y}) = \mathrm{Sim}_{\mathcal{G}_2}^{\lambda}(\mathrm{x},\mathrm{y})$.

2) If $\mathrm{Sim}_{\mathcal{G}}(\mathrm{x},\mathrm{y}) < \lambda_1 \leq \lambda_2$, then $\mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(\mathrm{x},\mathrm{y}) = \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(\mathrm{x},\mathrm{y}) = 0$; if $\lambda_1 \leq \mathrm{Sim}_{\mathcal{G}}(\mathrm{x},\mathrm{y}) < \lambda_2$, $\mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(\mathrm{x},\mathrm{y}) = \mathrm{Sim}_{\mathcal{G}}(\mathrm{x},\mathrm{y}) > \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(\mathrm{x},\mathrm{y}) = 0$; if $\lambda_1 \leq \lambda_2 \leq \mathrm{Sim}_{\mathcal{G}}(\mathrm{x},\mathrm{y})$, then $\mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(\mathrm{x},\mathrm{y}) = \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(\mathrm{x},\mathrm{y}) = \mathrm{Sim}_{\mathcal{G}}(\mathrm{x},\mathrm{y})$. In summary, $\mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(\mathrm{x},\mathrm{y}) \geq \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(\mathrm{x},\mathrm{y})$. □

2.2 Fuzzy Rough Sets

Definition 5. *Let* (U,R) *be an FCRS, with D serving as the decision feature. Then the triple* (U,R,D) *is known as a fuzzy decision table (FDT).*

Let $\mathrm{U/D} = \{\mathrm{D_1}, \mathrm{D_2}, \ldots, \mathrm{D_p}\}$ denote the ensemble of equivalent classes that are categorized by D. $\mathrm{D_i} \in \mathrm{U/D}(\mathrm{i} = 1, 2, \ldots, \mathrm{p})$ can also be conceptualized as a fuzzy set, characterized by clear boundaries.

$$D_i(x) = \begin{cases} 0, x \notin D_i \\ 1, x \in D_i \end{cases} \tag{2}$$

Definition 6. *Let* (U,R) *be an FCRS,* $\forall A \in \mathcal{F}(U)$, $\forall \mathcal{G} \subseteq R$, *and* $\forall \lambda \in [0,1]$. *Then the lower and upper approximations of A concerning* $\mathcal{G}$ *and* λ *are respectively as follows:*

$$\underline{Apr}_{\mathcal{G}}^{\lambda}(A)(x) = \bigwedge_{y \in U} \left\{ A(y) \bigvee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y)) \right\} \tag{3}$$

$$\overline{Apr}_{\mathcal{G}}^{\lambda}(A)(x) = \bigvee_{y \in \mathrm{U}} \left\{ A(y) \bigwedge \mathrm{Sim}_{\mathcal{G}}^{\lambda}(x,y) \right\} \tag{4}$$

Then the quadruple $(\mathrm{U}, \mathcal{G}, \mathrm{A}, \lambda)$ is called a fuzzy rough set model (FRSM) concerning $\mathcal{G}$ and λ.

Proposition 7. *Let* $(U, \mathcal{G}, A, \lambda)$ *be an FRSM. Then the following conclusions hold:*

1) *if* $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{G}$, *then* $\underline{Apr}_{\mathcal{G}_1}^{\lambda}(A) \subseteq \underline{Apr}_{\mathcal{G}_2}^{\lambda}(A)$;
2) *if* $\lambda_1 \leq \lambda_2$, *then* $\underline{Apr}_{\mathcal{G}}^{\lambda_1}(A) \subseteq \underline{Apr}_{\mathcal{G}}^{\lambda_2}(A)$.

Proof. 1) Base on Proposition 4, for any $\mathrm{x},\mathrm{y} \in \mathrm{U}$, if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{G}$, then $\mathrm{Sim}_{\mathcal{G}_1}^{\lambda}(\mathrm{x},\mathrm{y}) \geq \mathrm{Sim}_{\mathcal{G}_2}^{\lambda}(\mathrm{x},\mathrm{y}) \Rightarrow \underline{\mathrm{Apr}}_{\mathcal{G}_1}^{\lambda}(\mathrm{A})(\mathrm{x}) = \bigwedge_{\mathrm{y} \in \mathrm{U}} \{\mathrm{A}(\mathrm{y}) \bigvee (1 - \mathrm{Sim}_{\mathcal{G}_1}^{\lambda}(\mathrm{x},\mathrm{y}))\} \leq \bigwedge_{\mathrm{y} \in \mathrm{U}} \{\mathrm{A}(\mathrm{y}) \vee (1 - \mathrm{Sim}_{\mathcal{G}_2}^{\lambda}(\mathrm{x},\mathrm{y}))\} = \underline{\mathrm{Apr}}_{\mathcal{G}_2}^{\lambda}(\mathrm{A})(\mathrm{x}) \Rightarrow \underline{\mathrm{Apr}}_{\mathcal{G}_1}^{\beta,\lambda}(\mathrm{A}) \subseteq \underline{\mathrm{Apr}}_{\mathcal{G}_2}^{\beta,\lambda}(\mathrm{A})$.

2) Base on Proposition 4, for any $x, y \in U$, if $\lambda_1 \le \lambda_2$, then $\mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(x,y) \ge \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(x,y) \Rightarrow \underline{\mathrm{Apr}}_{\mathcal{G}}^{\lambda_1}(A)(x) = \bigwedge_{y\in U}\{A(y) \vee (1 - \mathrm{Sim}_{\mathcal{G}}^{\lambda_1}(x,y))\} \le \bigwedge_{y\in U}\{A(y) \vee (1 - \mathrm{Sim}_{\mathcal{G}}^{\lambda_2}(x,y))\} = \underline{\mathrm{Apr}}_{\mathcal{G}}^{\lambda_2}(A)(x) \Rightarrow \underline{\mathrm{Apr}}_{\mathcal{G}}^{\lambda_1}(A) \subseteq \underline{\mathrm{Apr}}_{\mathcal{G}}^{\lambda_2}(A)$. □

Proposition 8. *Let $(U, \mathcal{G}, A, \lambda)$ be an FRSM. Then the following conclusions hold:*

1) *if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{G}$, then $\overline{Apr}_{\mathcal{G}_1}^{\lambda}(A) \supseteq \overline{Apr}_{\mathcal{G}_2}^{\lambda}(A)$;*
2) *if $\lambda_1 \le \lambda_2$, then $\overline{Apr}_{\mathcal{G}}^{\lambda_1}(A) \supseteq \overline{Apr}_{\mathcal{G}}^{\lambda_2}(A)$.*

Proof of Proposition 8 follows a similar approach to Proposition 7.

Proposition 9. *Let $(U, \mathcal{G}, A, \lambda)$ be an FRSM. Then $\underline{Apr}_{\mathcal{G}}^{\lambda}(A) \subseteq A \subseteq \overline{Apr}_{\mathcal{G}}^{\lambda}(A)$.*

Proof. Based on Definitions 2.2 and 2.7, for any $x \in \mathbf{U}$, $Sim_{\mathcal{G}}^{\lambda}(x,x) = 1$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(A)(x) = \bigwedge_{y\in U}\{A(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} \le A(x) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,x)) = A(x)$ and $\overline{Apr}_{\mathcal{G}}^{\lambda}(A)(x) = \bigvee_{y\in U}\{A(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} \ge A(x) \wedge Sim_{\mathcal{G}}^{\lambda}(x,x) = A(x)$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(A)(x) \le A(x) \le \overline{Apr}_{\mathcal{G}}^{\lambda}(A)(x)$, so we have $\underline{Apr}_{\mathcal{G}}^{\lambda}(A) \subseteq A \subseteq \overline{Apr}_{\mathcal{G}}^{\lambda}(A)$. □

Definition 10. *Let (U, R, D) be an FDT, $\forall \mathcal{G} \subseteq R$ and $\forall \lambda \in [0,1)$. If $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x)$ $(\forall D_i \in U/D)$, x is referred to as a consistent sample regarding $\mathcal{G}$ and λ. If all samples of U demonstrate consistency, the FDT is considered a consistent one.*

Proposition 11. *Let (U, R, D) be an FDT, $\forall A \in \mathcal{F}(U)$, $\forall \mathcal{G} \subseteq R$, and $\forall \lambda \in [0,1]$. Then the following conclusions are true:*

1) *if $Sim_{\mathcal{G}}^{\lambda}(x,y) > 0 (\forall y \in U) \Rightarrow D_i(x) = D_i(y)\ (i = 1, 2, \ldots, p)$, then the sample x is a consistent sample;*
2) *if $Sim_{\mathcal{G}}^{\lambda}(x,y) > 0 (\forall y \in U) \Rightarrow \exists D_i \in U/D$ such that $D_i(x) \ne D_i(y)$, then the sample x is an inconsistent sample.*

Proof. 1) For $D_i(x) = 1$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigwedge_{y\in U}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} = \{\bigwedge_{y\in D_i}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\}\} \wedge \{\bigwedge_{y\in U-D_i}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\}\}$. For $\in D_i \Rightarrow D_i(y) = 1 \Rightarrow \bigwedge_{y\in D_i}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} = 1$; for $y \notin D_i \Rightarrow Sim_{\mathcal{G}}^{\lambda}(x,y) = 0 \Rightarrow \bigwedge_{y\in U-D_i}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} = 1$. $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 1$ holds. $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigvee_{y\in U}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} \ge D_i(x) \wedge Sim_{\mathcal{G}}^{\lambda}(x,x) = 1$, and $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) \in [0,1]$, so $\forall y \in U, \overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 1 = \underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x)$.

For $D_i(x) = 0$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigwedge_{y\in U}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} \le D_i(x) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,x)) = 0$, and $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) \ge 0$, so $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 0$; $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigvee_{y\in U}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} = \{\bigvee_{y\in D_i}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\}\} \vee \{\bigvee_{y\in U-D_i}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\}\}$. For $y \in D_i \Rightarrow Sim_{\mathcal{G}}^{\lambda}(x,y) = 0 \Rightarrow \bigvee_{y\in D_i}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} = 0$; and for $y \notin D_i \Rightarrow D_i(y) = 0 \Rightarrow \bigvee_{y\in U-D_i}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} = 0$. $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 0$ holds. $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 0 = \overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x)$.

Hence, the sample x is a consistent sample.

2) For $D_i(x) = 1$, $D_i(y) = 0$ and $Sim_{\mathcal{G}}^{\lambda}(x,y) > 0$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigwedge_{y\in U}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} \le D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y)) < 1$; $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigvee_{y\in U}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} \ge D_i(x) \wedge Sim_{\mathcal{G}}^{\lambda}(x,x) = 1$ and $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) \in [0,1]$, $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = 1 > \underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x)$.

For $D_i(x) = 0$, $D_i(y) = 1$ and $Sim_{\mathcal{G}}^{\lambda}(x,y) > 0$, $\underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigwedge_{y\in U}\{D_i(y) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,y))\} \le D_i(x) \vee (1 - Sim_{\mathcal{G}}^{\lambda}(x,x)) = 0$; $\overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) = \bigvee_{y\in U}\{D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y)\} \ge D_i(y) \wedge Sim_{\mathcal{G}}^{\lambda}(x,y) = Sim_{\mathcal{G}}^{\lambda}(x,y) > 0, so \overline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x) > \underline{Apr}_{\mathcal{G}}^{\lambda}(D_i)(x)$.

Hence, the sample x is an inconsistent sample. □

3 Cross-Similarity

According to Proposition 11, the non-zero similarity of samples that are inconsistent originates from distinct classes. Hence, the focus should be exclusively on the similarity relationships that exist between samples belonging to different classes. Now, it's time to delve into the intriguing concept of cross-similarity.

Definition 12. *Let* (U, R, D) *be an FDT,* $\mathcal{G} \subseteq R$, $x \in U$, $y \in U$, *and* $0 \leq \lambda \leq 1$. *The cross-similarity between* x *and* y *is defined as follows:*

$$CS^{\lambda}_{\mathcal{G},D}(x,y) = \begin{cases} 0, D(x) = D(y) \\ Sim^{\lambda}_{\mathcal{G}}(x,y), D(x) \neq D(y). \end{cases} \tag{5}$$

One can consider the FDT as a graph.

- The samples in U are visualized as nodes in the graph.
- The parameterized multi-attribute similarity relations, designated as $\mathrm{Sim}^{\lambda}_{\mathcal{G}}$, are conceptually understood as undirected weighted edges connecting nodes within the graph. If $\mathrm{Sim}^{\lambda}_{\mathcal{G}} = 0$, then there is no edge connecting node x and node y in the graph.
- The equivalence relations derived from D can be visualized as distinct enclosed areas within the graph.

By Definition 12, we narrow our focus to the similarity relations among samples belonging to distinct target partitions. The term "cross" denotes the intersection of the edge connecting samples of different target partitions and the boundary of enclosed areas, from a graphical perspective.

Definition 13. *Let* (U, R, D) *be an FDT,* $\mathcal{G} \subseteq R$, $x \in U$, *and* $0 \leq \lambda \leq 1$. *The fuzzy set with respect to* x *is defined as follows:*

$$CS^{\lambda}_{\mathcal{G},D}(x) = \frac{CS^{\lambda}_{\mathcal{G},D}(x)(x_1)}{x_1} + \frac{CS^{\lambda}_{\mathcal{G},D}(x)(x_2)}{x_2} + \cdots + \frac{CS^{\lambda}_{\mathcal{G},D}(x)(x_m)}{x_m} \tag{6}$$

The fuzzy set can be conceptualized as the fuzzy information granule of x, with its cardinality represented as follows:

$$\left|CS^{\lambda}_{\mathcal{G},D}(x)\right| = \sum_{i\in[1,m]} CS^{\lambda}_{\mathcal{G},D}(x)(x_i) \tag{7}$$

Let

$$CS^{\lambda}_{G,D}(U) = \begin{bmatrix} CS^{\lambda}_{G,D}(x_1,x_1) & CS^{\lambda}_{G,D}(x_1,x_2) & \cdots & CS^{\lambda}_{G,D}(x_1,x_m) \\ 0 & CS^{\lambda}_{G,D}(x_1,x_2) & \cdots & CS^{\lambda}_{G,D}(x_1,x_m) \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & CS^{\lambda}_{G,D}(x_m,x_m) \end{bmatrix}$$

and

$$\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| = \sum_{i\in[1,m],j\in[i,m]} CS^{\lambda}_{\mathcal{G},D}(x_i)(x_j) \tag{8}$$

$\left|CS^{\lambda}_{\mathcal{G},D}(U)\right|$ represents the cross-similarity level of U.

Proposition 14. *Let* (U, R, D) *be an FDT,* $x \in U$, $y \in U$, $0 \leq \lambda \leq 1$ *and* $\mathcal{G} \subseteq R$. *The following conclusions are true:*

1) $CS^{\lambda}_{\mathcal{G},D}(x,y) = 0$;
2) $CS^{\lambda}_{\mathcal{G},D}(x,y) = CS^{\lambda}_{\mathcal{G},D}(y,x)$;
3) $\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| = \frac{1}{2}\sum_{i\in[1,m]}\left|CS^{\lambda}_{\mathcal{G},D}(x_i)\right|$;
4) $0 \le \left|CS^{\lambda}_{\mathcal{G},D}(x)\right| \le m-1$;
5) $0 \le \left|CS^{\lambda}_{\mathcal{G},D}(U)\right| \le \frac{m(m-1)}{2}$.

Proof. 1) By Definition 12, it is evident that $\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x},\mathrm{y}) = 0$;

2) Under Definitions 2.2–2.4 and Definition 12, it is evident that this conclusion holds true;

3) According to 1), 2), and Eq. (8), it is evident that this conclusion holds true;

4) Since $\mathrm{Sim}^{\lambda}_{\mathcal{G}}(\mathrm{x},\mathrm{y}) \in [0,1]$ and $\mathrm{Sim}^{\lambda}_{\mathcal{G}}(\mathrm{x},\mathrm{x}) = 0$, this conclusion is true;

5) In accordance with 3) and 4), $\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right| = \frac{1}{2}\sum_{\mathrm{i}\in[1,\mathrm{m}]}\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x_i})\right| \ge 1/2 \times \mathrm{m} \times 0 = 0$ and $\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right| = \frac{1}{2}\sum_{\mathrm{i}\in[1,\mathrm{m}]}\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x_i})\right| \le \frac{1}{2} \times \mathrm{m} \times (\mathrm{m}-1) \le \frac{\mathrm{m}(\mathrm{m}-1)}{2}$. □

Theorem 15. *Let (U,R,D) be an FDT, $x \in U$, $y \in U$, $\lambda \in [0,1]$ and $\mathcal{G} \subseteq R$. The following conclusions hold true:*

1) *if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq R$, then $CS^{\lambda}_{\mathcal{G}_1,D}(x,y) \ge CS^{\lambda}_{\mathcal{G}_2,D}(x,y)$;*
2) *if $0 \le \lambda_1 \le \lambda_2 \le 1$, then $CS^{\lambda_1}_{\mathcal{G},D}(x)(y) \ge CS^{\lambda_2}_{\mathcal{G},D}(x)(y)$.*

Proof. 1) In accordance with Definition 12 and Proposition 4, it is evident that this conclusion holds true.

2) According to Definition 12 and Proposition 4, it is straightforward. □

Theorem 16. *Let (U,R,D) be an FDT, $x \in U$, $\lambda \in [0,1]$ and $\mathcal{G} \subseteq R$. The following conclusions hold true:*

1) *if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq R$, then $|CS^{\lambda}_{\mathcal{G}_1,D}(x)| \ge |CS^{\lambda}_{\mathcal{G}_2,D}(x)|$;*
2) *if $0 \le \lambda_1 \le \lambda_2 \le 1$, then $|CS^{\lambda_1}_{\mathcal{G},D}(x)| \ge |CS^{\lambda_2}_{\mathcal{G},D}(x)|$.*

Proof. According to Theorem 15 and Eq. (7), it is straightforward to demonstrate the validity of these conclusions. □

Theorem 17. *Let (U,R,D) be an FDT, $x \in U$, $\lambda \in [0,1]$ and $\mathcal{G} \subseteq R$. The following conclusions are true:*

1) *if $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq R$, then $|CS^{\lambda}_{\mathcal{G}_1,D}(U)| \ge |CS^{\lambda}_{\mathcal{G}_2,D}(U)|$;*
2) *if $0 \le \lambda_1 \le \lambda_2 \le 1$, then $|CS^{\lambda_1}_{\mathcal{G},D}(U)| \ge |CS^{\lambda_2}_{\mathcal{G},D}(U)|$.*

Proof. According to Theorem 16 Proposition 14, and Eq. (8), it is straightforward to demonstrate the validity of these conclusions. □

Example 1. *Let (U,R,D) be an FDT, where $U = \{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$ and $U/D = \{D_1 = \{x_2,x_3\}, D_2 = \{x_1,x_7\}, D_3 = \{x_4,x_5,x_6\}\}$. $Sim^{\lambda}_{\mathcal{G}}$ is shown in Table 1, where $\mathcal{G} \subseteq R$ and $0 \le \lambda < 1$.*

Target partition D_1 can be represented as a fuzzy set $D_1 = \frac{0}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0}{x_7}$. The lower and upper approximations of $\mathrm{D_1}$ under FCRS $(\mathrm{U},\mathcal{G})$ and parament λ are:

$$\underline{Apr}^{\lambda}_{\mathcal{G}}(D_1) = \frac{0}{x_1} + \frac{0.5}{x_2} + \frac{0.3}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0}{x_7}, \overline{Apr}^{\lambda}_{\mathcal{G}}(D_1) = \frac{0.5}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \frac{0.6}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0.7}{x_7}.$$

Obviously, $\underline{Apr}^{\lambda}_{\mathcal{G}}(D_1) \subset \overline{Apr}^{\lambda}_{\mathcal{G}}(D_1)$, there is uncertainty in the FCRS $(U,\mathcal{G})$ for Target partition D_1. For a decision table, if there is an approximate inequality between the upper and lower parts of a decision partition, it is claimed that the decision table has uncertainty.

By Definition 13, the fuzzy set associated with each sample point is computed as follows:

$$\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x_1}) = \frac{0}{\mathrm{x_1}} + \frac{0.5}{\mathrm{x_2}} + \frac{0}{\mathrm{x_3}} + \frac{0}{\mathrm{x_4}} + \frac{0}{\mathrm{x_5}} + \frac{0}{\mathrm{x_6}} + \frac{0}{\mathrm{x_7}}, \left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x_1})\right| = 0.5;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_2) = \frac{0.5}{x_1} + \frac{0}{x_2} + \frac{0}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_2)\right| = 0.5;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_3) = \frac{0}{x_1} + \frac{0}{x_2} + \frac{0}{x_3} + \frac{0.6}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0.7}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_3)\right| = 1.3;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_4) = \frac{0}{x_1} + \frac{0}{x_2} + \frac{0.6}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_4)\right| = 0.6;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_5) = \frac{0}{x_1} + \frac{0}{x_2} + \frac{0}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_5)\right| = 0;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_6) = \frac{0}{x_1} + \frac{0}{x_2} + \frac{0}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0}{x_6} + \frac{0.4}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_6)\right| = 0.4;$$

$$CS^{\lambda}_{\mathcal{G},D}(x_7) = \frac{0}{x_1} + \frac{0}{x_2} + \frac{0.7}{x_3} + \frac{0}{x_4} + \frac{0}{x_5} + \frac{0.4}{x_6} + \frac{0}{x_7}, \left|CS^{\lambda}_{\mathcal{G},D}(x_7)\right| = 1.1.$$

Table 1: $Sim^{\lambda}_{\mathcal{G}}$

$Sim^{\lambda}_{\mathcal{G}}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1	0.5	0	0	0	0	0.8
x_2	0.5	1	0.6	0	0	0	0
x_3	0	0.6	1	0.6	0	0	0.7
x_4	0	0	0.6	1	0.8	0	0
x_5	0	0	0	0.8	1	0.5	0
x_6	0	0	0	0	0.5	1	0.4
x_7	0.8	0	0.7	0	0	0.4	1

Therefore, based on Definition 12, the FDT (U, R, D) under $\mathcal{G}$ and λ can be represented in Fig. 1. Each sample is represented as a node, a similarity relationship greater than 0 corresponds to an edge, and a target partition is depicted as an area enclosed by a dashed line in the graph. Cross-similarity relationships within similar relationships are highlighted in bold red.

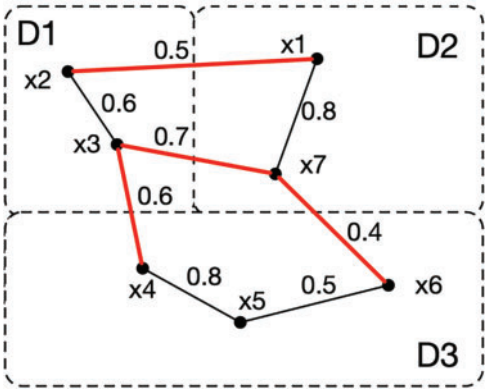

Figure 1: Graphical representation of the FDT

The cross-similarity matrix $CS^{\lambda}_{\mathcal{G},D}(U)$ of the FDT is shown as follows:

$$\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U}) = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore, the cross-similarity level of U is:

$$\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right| = 1/2 \sum_{\mathrm{x}\in\mathrm{U}} \left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{x})\right| = 2.2$$

Theorem 18. *Let* (U, R, D) *be an FDT,* $\lambda \in [0,1]$ *and* $\mathcal{G} \subseteq R$. *If* $\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| = 0$, *the FDT is consistent:*

Proof. $\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| = 0 \Longleftrightarrow \forall x, y \in U, \quad CS^{\lambda}_{\mathcal{G},D}(x,y) = 0 \Longleftrightarrow if\, Sim^{\lambda}_{\mathcal{G}}(x,y) > 0 \Rightarrow D(x) = D(y)$. Therefore, the FDT is consistent according to Proposition 11 1). □

Theorem 19. *Let* (U, R, D) *be an FDT,* $\lambda \in [0,1]$ *and* $\mathcal{G} \subseteq R$. *If* $\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| > 0$, *the FDT is inconsistent.*

Proof. $\left|CS^{\lambda}_{\mathcal{G},D}(U)\right| > 0 \Rightarrow \exists x, y \in U, \quad CS^{\lambda}_{\mathcal{G},D}(x,y) > 0 \Rightarrow \exists D_l \in U/D, \quad D_l(x) \neq D_l(y)$ and $Sim^{\lambda}_{\mathcal{G}}(x,y) > 0$. Therefore, the FDT is inconsistent according to Proposition 11 2). □

4 Feature Selection Based on Cross-Similarity

As $\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right|$ decreases, the similarity between samples belonging to different classes diminishes. In other words, a low $\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right|$ suggests that $\mathcal{G}$ exhibits strong discriminatory capabilities. Consequently, $\left|\mathrm{CS}^{\lambda}_{\mathcal{G},\mathrm{D}}(\mathrm{U})\right|$ serves as an effective tool for measuring the uncertainty associated with the FDT. The framework diagram of feature selection based on cross-similarity is shown in Fig. 2.

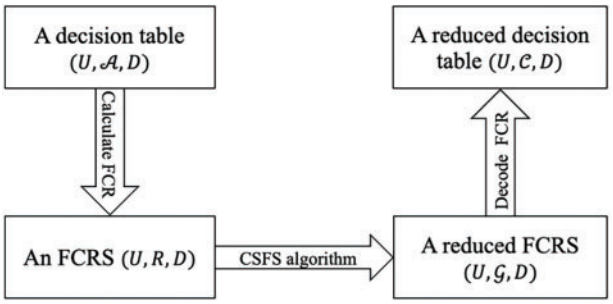

Figure 2: A framework diagram of feature selection

Definition 20. *Let* (U, R, D) *be an FDT,* $\lambda \in [0,1]$ *and* $red \subseteq R$. *The* red *is referred to as a reduction of* R, *if*

1) $\left|CS^{\lambda}_{red,D}(U)\right| = \left|CS^{\lambda}_{R,D}(U)\right|$;
2) $\left|CS^{\lambda}_{red-\{r\},D}(U)\right| > \left|CS^{\lambda}_{red,D}(U)\right|, \forall r \in red.$

Definition 21. *Let* (U, R, D) *be an FDT,* $\lambda \in [0,1]$, $red \subseteq R$ *and* $r \in R - red$. *The significance of* r *with respect to* red *is defined as follows:*

$$DC_\lambda(r, red, D) = \left|CS^\lambda_{red,D}(U)\right| - \left|CS^\lambda_{red\cup\{r\},D}(U)\right| \tag{9}$$

The larger $DC_\lambda(r, red, D)$ is, the smaller $\left|CS^\lambda_{red\cup\{r\},D}(U)\right|$ becomes, making the r more signicant.

Algorithm 1, a cross-similarity based feature selection (CSFS) algorithm, is a forward greedy feature selection algorithm.

Algorithm 1: A forward greedy feature selection based on cross-similarity (CSFS)

Input: An FDT (U, R, D) derived from $\mathcal{A}$ and λ.
Output: A reduct red.

Initialization: $red = \varnothing$, $U^* = U$, $R^* = R$
for each $a \in \mathcal{A}$ **do**
  Calculate r_a based on a;
  Calculate the parameterized relation $\mathrm{Sim}^\lambda_{\{r\}}(x, y)$;
  Calculate the corresponding cross-similarity matrix $CS^\lambda_{\{r\},D}(U^*)$;
end
n matrices $CS^\lambda_{S_{\{r\}},D}(U^*)$ based on $\mathcal{A}$ are generated.
while $R^* \neq \varnothing$ **and** $U^* \neq \varnothing$ **do**
  for each r $\in R^*$ **do**
    Calculate $|CS^\lambda_{red\cup\{r\},D}(U)| = \left|CS^\lambda_{red,D}(U) \wedge CS^\lambda_{\{r\},D}(U)\right|$
  end
  Find $r' \in R^*$ such that the minimum value is $|CS^\lambda_{red\cup\{r'\},D}(U^*)|$;
  Calculate $DC_\lambda(r', red, D) = |CS^\lambda_{red,D}(U^*)| - |CS^\lambda_{red\cup\{r'\},D}(U^*)|$
  if $DC_\lambda(r', red, D) > 0$ **then**
    $red \leftarrow red \cup \{r'\}$;
    $R^* \leftarrow R^* - \{r'\}$;
    $U^* \leftarrow U^* - \left\{x \middle| \left|CS^\lambda_{red\cup\{r'\},D}(x)\right| = 0\right\}$
  else
    break;
  end
end
return red

Considering that as more features are chosen, the number of relationships to be considered when calculating cross-similarity progressively diminishes, and only the relationships between samples located in different categories are considered, the number of samples gradually decreases. It can be assumed that the number of samples decreases in geometric progression with a common ratio of $k(0 < k < 1)$. Then $m + km + k^2m + \cdots + k^tm = \left[(1 - k^{t+1})/(1 - k)\right]m = cm$. Therefore, the time complexity of Algorithm 1 is improved to $O(mn^2)$. In contrast, the time complexity of the feature selection algorithm based on positive domain, dependency, and conditional entropy is $O(m^2n^2)$.

Example 2. *A decision information table is shown in Table 2, where* $U = \{x_1, x_2, x_3, x_4, x_5, x_5, x_7\}$, $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$ *and* $D = \{p, n\}$. *Let* $U^* = U$ *and* $R^* = R$.

Table 2: A decision information table

U	x_1	x_2	x_3	x_4	x_5	x_6	x_7
a_1	0.3	0.4	0	1	0.4	0.5	0.75
a_2	1	0.5	0	0.5	0.2	0.8	0.3
a_3	0.7	0.3	1	0	0.3	0.8	0.25
a_4	0.8	1	0	0.2	0.7	0.65	0.6
D	p	p	p	n	n	n	n

In this paper, the value of r_i is determined through the utilization of the following equation:

$$r_{a_i}(x,y) = max\left\{min\left\{\frac{a_i(x)-a_i(y)+\sigma_{a_i}}{\sigma_{a_i}}, \frac{a_i(y)-a_i(x)+\sigma_{a_i}}{\sigma_{a_i}}\right\}, 0\right\} \tag{10}$$

where σ_{a_i} is the standard deviation of the attribute a_i. By calculations, it is determined that $\sigma_{a_1} = 0.32$, Let $\lambda = 0.4$. Then r_{a_1}, $Sim^{\lambda}_{\{r_{a_1}\}}$ and $CS^{\lambda}_{\{r_{a_1}\}}$ are shown in Table 3.

Table 3: r_{a_1}, $Sim^{\lambda}_{\{r_{a_1}\}}$ and $CS^{\lambda}_{\{r_{a_1}\}}$

	r_{a_1}							$Sim^{\lambda}_{\{r_{a_1}\}}$							$CS^{\lambda}_{\{r_{a_1}\}}(U)$							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1	0.69	0.06	0	0.69	0.38	0	1	0.69	0	0	0.69	0	0	0	0	0	0	0.69	0	0
x_2	0.69	1	0	0	1	0.69	0	0.69	1	0	0	1	0.69	0	0	0	0	0	1	0.69	0
x_3	0.06	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
x_4	0	0	0	1	0	0	0.22	0	0	0	1	0	0	0	0	0	0	0	0	0	0
x_5	0.69	1	0	0	1	0.69	0	0.69	1	0	0	1	0.69	0	0	0	0	0	0	0	0
x_6	0.38	0.69	0	0	0.69	1	0.22	0	0.69	0	0	0.69	1	0	0	0	0	0	0	0	0
x_7	0	0	0	0.22	0	0.22	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0

The samples correspond to the nodes, similarity relations correspond to edges, bold red edges highlight the cross-similarity relations, and an area enclosed by a dashed line represents the target partition. The cross-similarity relation derived from a_1 is shown in Fig. 3.

Similarly, $CS^{\lambda}_{\{r_{a_2}\}}$, $CS^{\lambda}_{\{r_{a_3}\}}$ and $CS^{\lambda}_{\{r_{a_4}\}}$ can be determined, as shown in Table 4.

If $U^* = \varnothing$, let $Sim^{\lambda}_{red}(x,y) = 0$. Therefore, $\left|CS^{\lambda}_{red,D}(x)\right| = 0$. If $red = \varnothing$, let $Sim^{\lambda}_{red}(x,y) = 1$. $\left|CS^{\lambda}_{red,D}(x)\right| = 12$.

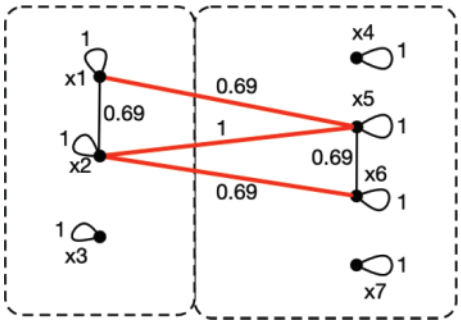

Figure 3: Cross-similarity graph representation

Table 4: Cross-similarity relations $CS^{\lambda}_{\{r_{a_2}\}}$, $CS^{\lambda}_{\{r_{a_3}\}}$ and $CS^{\lambda}_{\{r_{a_4}\}}$

	$CS^{\lambda}_{\{r_{a_2}\}}(U)$							$CS^{\lambda}_{\{r_{a_3}\}}(U)$							$CS^{\lambda}_{\{r_{a_4}\}}(U)$							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0	0	0	0	0	0.43	0	0	0	0	0	0.72	0	0	0	0	0	0	0.71	0.57	0.43
x_2	0	0	0	1	0	0	0.43	0	0	0	0	1	0	0.86	0	0	0	0	0	0	0
x_3	0	0	0	0	0.43	0	0	0	0	0	0	0	0.44	0	0	0	0	0.43	0	0	0
x_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

According to Tables 3 and 4, we obtain $|CS^{\lambda}_{red\cup\{r_{a_1}\}}(U)| = 2.38$, $|CS^{\lambda}_{red\cup\{r_{a_2}\}}(U)| = 2.29$, $|CS^{\lambda}_{red\cup\{r_{a_3}\}}(U)| = 3.02$ and $|CS^{\lambda}_{red\cup\{r_{a_4}\}}(U)| = 2.14$. Therefore, the minimum value is $|CS^{\lambda}_{red\cup\{r_{a_4}\}}(U)|$ and $DC_{\lambda}(r_{a_4}, red, D) = 12 - 2.14 = 9.86 > 0$. Thus, $red = \{r_{a_4}\}$, $R^* = \{r_{a_1}, r_{a_2}, r_{a_3}\}$ and $U^* = \{x_1, x_3, x_4, x_5, x_6, x_7\}$. Subsequently, we can obtain the following values by calculation. $\left|CS^{\lambda}_{red\cup\{r_{a_1}\}}(U)\right| = 0.69$, $\left|CS^{\lambda}_{red\cup\{r_{a_2}\}}(U)\right| = 0.43$ and $\left|CS^{\lambda}_{red\cup\{r_{a_3}\}}(U)\right| = 0.57$. Therefore, the minimum value is $\left|CS^{\lambda}_{red\cup\{r_{a_2}\}}(U)\right|$ and $DC_{\lambda}(r_{a_2}, red, D) = 2.14 - 0.43 = 1.71 > 0$. Thus, $red = \{r_{a_2}, r_{a_4}\}$, $R^* = \{r_{a_1}, r_{a_3}\}$ and $U^* = \{x_1, x_6\}$. Proceeding further, we have $\left|CS^{\lambda}_{red\cup\{r_{a_1}\}}(U)\right| = 0$ and $\left|CS^{\lambda}_{red\cup\{r_{a_3}\}}(U)\right| = 0.43$. Therefore, the minimum value is $\left|CS^{\lambda}_{red\cup\{r_{a_1}\}}(U)\right|$ and $DC_{\lambda}(r_{a_2}, red, D) = 0.43 - 0 = 0.43 > 0$. Thus, $red = \{r_{a_1}, r_{a_2}, r_{a_4}\}$, $R^* = \{r_{a_3}\}$ and $U^* = \varnothing$. As $DC_{\lambda}(r_{a_3}, red, D) = 0 - 0 = 0$, the feature selection process is terminated. As a result, the final feature selection outcome is $red = \{r_{a_1}, r_{a_2}, r_{a_4}\}$. Given $m = 7$, $k = 6/7$ and $c = 3$. Then, the descending sequence of sample sizes fulfills the condition: $7 + 6 + 2 + 0 \leq m + km + k^2m + k^3m \leq cm << m^2$.

5 Experimental Analysis

To establish the validity of CSFS, we conduct rigorous experimental comparisons between CSFS and five other state-of-the-art feature selection algorithms.

The following provides brief introductions to the five comparison algorithms, highlighting their key features and characteristics:

- FRFS [15] (2024): It is a classic feature selection algorithm based on the dependency between the decision attribute and the conditional attributes subset.
- FBC [18] (2020): It proposes a forward attribute selection algorithm based on conditional discernibility measure which can reflect the change of distinguishing ability caused by different attribute subsets.
- NEFE [22] (2023): It provides a feature selection algorithm for label distribution using dual-similarity-based neighborhood fuzzy entropy.
- MFBC [23] (2021): It is a feature selection algorithm based on the conditional information entropy under multi-fuzzy β-covering approximation space.
- HANDI [24] (2022): It proposes a feature selection algorithm based on a neighborhood discrimination index, which can characterize the distinguishing information of a neighborhood relation.

5.1 Experimental Design

5.1.1 Evaluation Criteria

In the following three dimensions, we comprehensively evaluate the six algorithms, emphasizing their individual strengths and limitations.

- The size of the selected feature subset.
- The time cost of the feature selection process.
- The performance of the selected feature subset in classification.

5.1.2 Classifiers

To conduct a rigorous and comprehensive evaluation of the efficacy of these six algorithms, we have chosen KNN (k = 5), SVM, and DT (CART) as the primary classification algorithms for assessing their classification accuracy. All other parameters of the three classifiers are set to default values. All testing procedures are conducted using ten-fold cross-validation for rigorous evaluation.

5.1.3 Experimental Environment

All experiments are implemented by Python 3.8.16 and scikit-learn 1.3.2, and run on a computer with an Intel(R) Xeon(R) Gold 5318Y CPU at 2.1 GHz, 128 GB RAM, and the operation system is Ubuntu18.04.1 LTS.

5.1.4 Datasets

In our experiments, we utilized 10 publicly available datasets as benchmarks to assess the performance of various algorithms. The first six datasets, being of low dimension, were procured from the UCI Machine Learning Repository [27], whereas the remaining four high-dimensional datasets were retrieved from the KRBD Repository [28]. For a comprehensive overview of these datasets, please refer to Table 5, which provides detailed descriptions of each dataset's characteristics. If there are missing values in the dataset, we will fill them with the average value of samples that belong to the same class and have the same feature as the missing value.

Table 5: Description of datasets

No.	Dataset	Samples	Attributes	Classes	Brief description
1	Wine	178	13	3	Using chemical analysis to determine the origin of wines.
2	Glass	214	9	6	Determine 6 categories of glass based on oxide content.
3	Ionosphere	315	34	2	Classification of radar returns from the ionosphere.
4	MuskV1	476	166	2	Predict whether new molecules will be in the musk or non-musk category.
5	German	1000	20	2	Evaluate good or bad credit risks based on a set of attribute descriptions.
6	WDBC	569	30	2	Diagnostic Wisconsin Breast Cancer Database.
7	DLBCL Stanford	47	4026	2	High-dimensional biomedical data sets taken from the Kent Ridge Biomedical Data Set Repository, including gene expression data, protein profiling data, and genomic sequence data that are related to classification tasks.
8	DLBCL Tumor	77	7129	2	
9	DLBCL NIH	240	7399	2	
10	AMLALL	72	7129	2	

5.1.5 Parameter Settings

Among the six algorithms, FRFS stands alone as the sole algorithm that does not involve any parameters. To ensure the fairness and integrity of comparative experiments, we employ a meticulous grid search approach to identify the optimal parameter values for all algorithms encompassing parameters. The step size of the CSFS algorithm has been set to 0.01, while the step sizes of other algorithms have been configured to 0.1. As per the specifications provided in the respective references for each algorithm, the search scope of the parameters is summarized in Table 6. The reason why the other algorithms set the step size to 0.1 instead of 0.01 is that if they also set the step size to 0.01, the experiments would not be able to complete. Initially, we outline the time taken by each algorithm in their search for optimal parameters. Subsequently, we present the optimal parameter values that were chosen based on their classification accuracy.

Table 6: Parameter setting of experiments algorithm

No.	Algorithm	Parameter
1	CSFS	λ from 0.01 to 0.99 with a step of 0.01
2	MFBC	β from 0.5 to 0.9 with a step of 0.1 λ from 0.1 to 0.5 with a step of 0.1
3	NFEM	delta from 0.1 to 0.9 with a step of 0.1
4	FBC	β from 0.5 to 0.9 with a step of 0.1 λ from 0.1 to 0.5 with a step of 0.1

(Continued)

Table 6 (continued)

No.	Algorithm	Parameter
5	HANDI	epsilon from 0.1 to 0.9 with a step of 0.1
6	FRFS	–

Note: Symbol '–' indicates that the algorithm has no parameter.

Table 7 provides an enumeration of the total time consumed by each algorithm on each dataset when running all possible parameters under the parameter configuration outlined in Table 7. Notably, CSFS stands out in terms of efficiency, consistently achieving the lowest average time for parameter search across all datasets.

Table 7: Time consumption for iterating through all parameters (second)

Dataset	*FRFS*	*HANDI*	*FBC*	*NFEM*	*MFBC*	*CSFS*
Wine	186	14	20	30	132	54
Glass	369	18	21	18	305	41
Ionosphere	4965	431	707	641	3464	548
MuskV1	226,604	5154	7742	83,534	37,082	4705
German	50,361	1332	2464	2088	63,577	1919
WDBC	20,172	787	1371	1277	8957	1254
DLBCL Stanford	151,897	2132	306	147,142	1439	670
DLBCL Tumor	1,831,943	8447	1761	1,497,466	8753	2754
DLBCL NIH	–	148,991	34,715	–	190,671	35,208
AMLALL	1,508,175	5414	1322	1,287,427	7071	2465
Average	421,630.2	17,272	5042.9	335,513.7	32,145.1	4961.8

Note: Symbol '–' indicates that the corresponding algorithm failed to complete execution on the dataset within 30 days. The runtime of FRFS is defined as the time required for it to complete a single feature selection.

5.2 Result Analysis

5.2.1 The Size of the Selected Feature Subset

Table 8 presents the count of features chosen by six feature selection algorithms across ten datasets. FRFS and NFEM were terminated early due to excessive runtime, hence their outcomes are undisclosed.

Table 8 indicates the following findings:

- In comparison to the original feature count in Table 5, it is noteworthy that these feature selection algorithms effectively select salient features and reduce data redundancy.
- The optimal feature subsets for different classifiers can vary.
- CSFS selects fewer features than FRFS, HANDI, NFEM, and MFBC, yet slightly more than FBC.

Table 8: Selected feature size for optimal classification performance

Dataset	KNN						SVM						DT					
	FRFS	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**	**FRFS**	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**	**FRFS**	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**
Wine	12	8	8	13	8	8	12	11	6	13	13	9	12	5	6	10	6	4
Glass	8	5	7	8	9	5	8	7	6	7	9	6	8	7	7	6	8	5
Ionosphere	16	12	5	31	22	4	16	9	14	31	22	24	16	12	10	31	31	16
MuskV1	86	7	14	48	15	20	86	30	14	85	15	17	86	30	12	26	13	7
German	3	18	13	15	20	13	3	17	15	13	16	9	3	17	13	19	16	16
WDBC	27	8	20	25	26	20	27	8	19	25	11	14	27	9	6	26	14	9
DLBCL	14	11	3	38	4	4	14	11	3	38	4	3	14	75	4	38	3	2
Stanford DLBCL Tumor	20	4	4	88	4	5	20	73	4	16	3	4	20	4	4	11	4	3
DLBCL NIH	–	12	8	–	7	8	–	41	8	–	7	8	–	12	7	–	6	9
AMLALL	12	4	3	129	3	3	12	11	3	129	3	3	12	4	3	72	3	3
Average	22	8.9	8.5	43.89	11.8	9	22	21.8	9.2	39.67	10.3	9.7	22	17.5	7.2	26.56	10.4	7.4

Note: Symbol '–' indicates that the corresponding algorithms failed to complete execution on the dataset within 30 days.

Table 9: Comparison of feature selection time under optimal classification parameters (second)

Dataset	KNN						SVM						DT					
	FRFS	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**	**FRFS**	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**	**FRFS**	**HANDI**	**FBC**	**NFEM**	**MFBC**	**CSFS**
Wine	186	1.76	1.11	2.79	3.18	0.462	186	2.13	0.887	2.79	3.54	0.506	186	1.31	0.889	7.36	2.71	0.491
Glass	369	1.99	0.792	2.06	8.06	0.563	369	1.93	0.796	2.01	8.06	0.407	369	1.93	0.792	1.99	7.82	0.392
Ionosphere	4965	31.7	9.67	74.5	70.9	2.74	4965	20.5	35.9	74.5	70.9	7.92	4965	31.7	22.1	74.5	75.9	6.42
MuskV1	226,604	433	391	14,215	555	81.3	226,604	464	383	9296	555	67.6	226,604	464	326	8093	500	35.8
German	50,361	142	112	282	2444	22.5	50,361	162	119	401	2435	18.9	50,361	162	112	214	2435	23.8
WDBC	20,172	109	119	138	114	22.1	20,172	109	109	138	66	14.3	20,172	109	31.7	142	79	11.3
DLBCL Stanford	151,897	76.3	13.5	19462	50.6	7.73	151,897	76.3	13.5	19,462	50.6	7.4	151,897	464	17.9	19,462	40.8	6.21
DLBCL Tumor	1,831,943	146	82.4	207,232	209	35.3	1,831,943	2661	82.4	138,007	171	31	1,831,943	146	82.4	138,555	209	29
DLBCL NIH	–	2774	2000	–	3814	457	–	8714	2000	–	3814	453	–	2774	1709	–	3391	502
AMLALL	1,508,175	138	54.5	258,374	159	25.5	1,508,175	357	53.6	258,374	159	25.6	1,508,175	138	53.4	163,424	159	25.5
Average	421,630	385	278	55,531	743	65.6	421,630	1257	280	47,306	733	62.7	421,630	429	236	36,664	690	64.1

Note: Symbol '–' indicates that the corresponding algorithms failed to complete execution on the dataset within 30 days.

5.2.2 The Runtime of the Algorithms

Table 9 and Fig. 4 present a comparison of the time taken by the six feature selection algorithms to execute a feature selection using parameters that achieved the highest classification accuracy for each classifier.

Table 9 and Fig. 4a–c indicate the following findings:

- When it comes to time efficiency, CSFS outperforms all other algorithms, exhibiting a clear advantage. The scale of the vertical axis representing time consumption in the graph is exponentially increasing.
- As the sample size increases, the advantage of CSFS becomes increasingly apparent. This observation is consistent with the theoretical time complexity of CSFS, which is $O(mn^2)$.

It is worth noting that for the DLBCL NIH dataset with 240 samples and 7399 conditional attributes, as the dimensionality increases, the time required for feature selection by the comparison algorithms is significantly higher than that of the CSFS algorithm. The NFEM algorithm is even unable to produce results within an acceptable time frame.

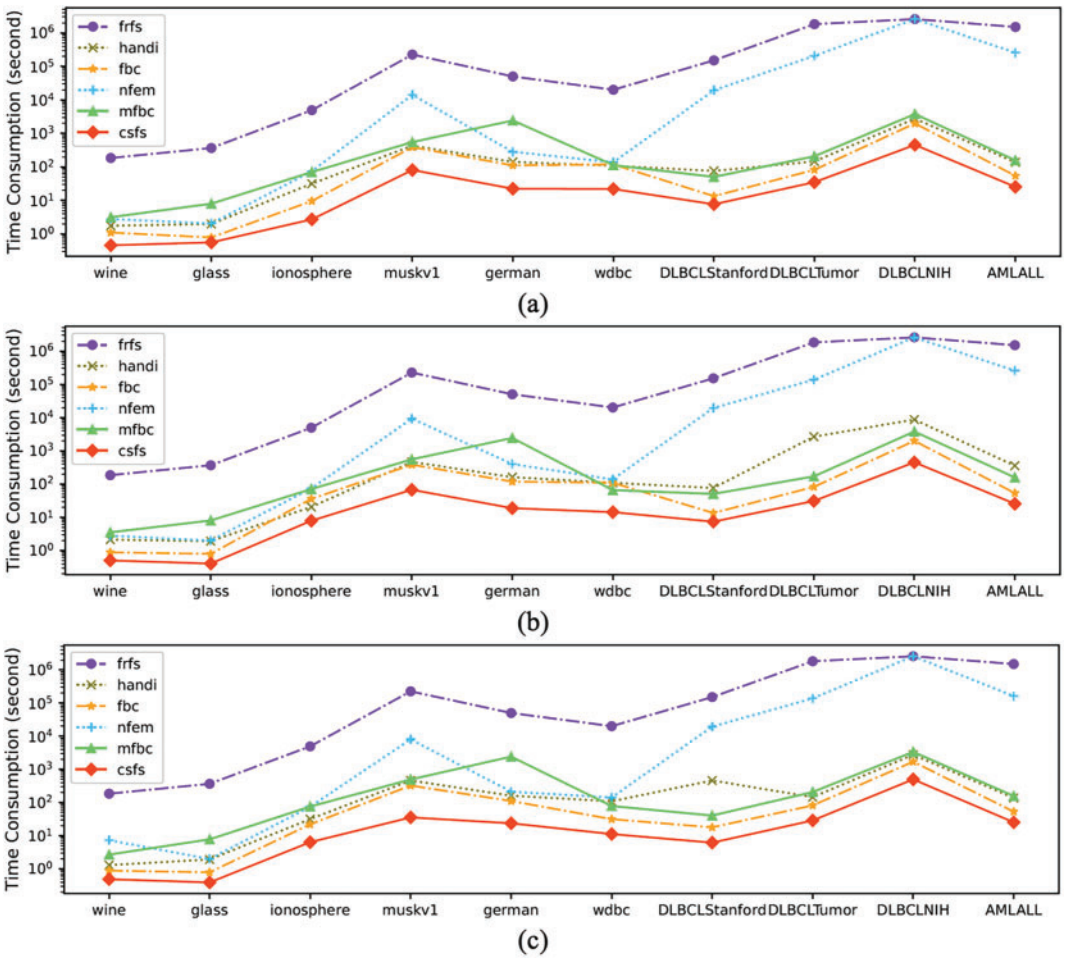

Figure 4: The time consumption of the feature selection process across various datasets. (a) KNN classifier with the optimal parameter. (b) SVM classifier with the optimal parameter. (c) DT classifier with the optimal parameter. FRFS and NFEM failed to complete execution on the DLBCLNIH dataset within 30 days and the runtime is set to 2,592,000 s (30 days)

5.2.3 Classification Accuracy

Tables 10–12 present the classification accuracy of three distinct classifiers across the selected feature subsets of the six algorithms. The highest classification accuracy is boldly displayed.

Tables 10–12 clearly indicate the following findings:

- Through 30 rigorous evaluations, algorithm CSFS has emerged as the clear leader, achieving the highest classification accuracy on 19 occasions. In comparison, algorithm FBC excelled on 11 occasions, while algorithms HANDI and MFBC delivered the best results on 8 and 3 occasions, respectively. Algorithms FRFS and NFEM lagged behind, achieving the highest accuracy on just 2 occasions each.
- The standard deviation values clearly indicate that the features selected by algorithm CSFS exhibit excellent representativeness, providing a stable and accurate representation of the data.

Table 10: Comparison of the classification accuracy achieved by KNN on 10 datasets, utilizing six distinct feature selection algorithms (%)

Dataset	FRFS	HANDI	FBC	NFEM	MFBC	CSFS
Wine	96.6 ± 2.78	**98.33 ± 2.55**	**98.33 ± 2.55**	96.05 ± 3.59	**98.33 ± 2.55**	**98.33 ± 2.55**
Glass	66.86 ± 7.79	69.59 ± 9.28	68.25 ± 8.6	69.68 ± 9.2	69.65 ± 8.5	**71.97 ± 9.5**
Ionosphere	84.9 ± 6.14	88.33 ± 5.16	89.74 ± 7.03	85.19 ± 6.96	83.78 ± 6.33	**90.89 ± 5.22**
MuskV1	76.97 ± 12.14	75.25 ± 7.37	**80.26 ± 5.65**	73.41 ± 12.11	76.28 ± 6.5	79.44 ± 5.87
German	65.8 ± 3.92	**72.8 ± 6.24**	72.0 ± 6.86	72.4 ± 7.58	71.9 ± 5.97	72.3 ± 3.77
WDBC	96.83 ± 2.2	96.3 ± 2.01	96.83 ± 2.07	**96.84 ± 1.89**	95.96 ± 1.37	95.96 ± 2.08
DLBCL Stanford	**98.0 ± 6.0**	**98.0 ± 6.0**	97.5 ± 7.5	61.0 ± 13.75	95.5 ± 9.07	**98.0 ± 6.0**
DLBCL Tumor	92.32 ± 8.42	96.07 ± 6.02	96.07 ± 6.02	79.64 ± 11.24	96.07 ± 6.02	**97.32 ± 5.37**
DLBCL NIH	–	68.75 ± 12.11	67.5 ± 13.15	–	63.33 ± 11.76	**67.92 ± 7.69**
AMLALL	98.75 ± 3.75	98.57 ± 4.29	100.0 ± 0.0	87.32 ± 7.68	98.57 ± 4.29	**100.0 ± 0.0**

Note: Symbol '–' indicates that the corresponding algorithms failed to complete execution on the dataset within 30 days. The bold font indicates the highest accuracy achieved on this dataset.

Table 11: Comparison of the classification accuracy achieved by SVM on 10 datasets, utilizing six distinct feature selection algorithms (%)

Dataset	FRFS	HANDI	FBC	NFEM	MFBC	CSFS
Wine	98.33 ± 2.55	98.33 ± 2.55	98.33 ± 2.55	98.33 ± 2.55	98.33 ± 2.55	**99.44 ± 1.67**
Glass	64.05 ± 9.46	65.89 ± 8.89	63.1 ± 10.86	65.89 ± 8.89	64.98 ± 9.73	**66.41 ± 8.76**
Ionosphere	92.87 ± 6.3	94.88 ± 3.55	**95.44 ± 3.43**	93.73 ± 4.75	95.15 ± 4.25	94.59 ± 4.32
MuskV1	78.43 ± 9.9	79.47 ± 8.49	**80.28 ± 7.11**	79.06 ± 9.54	77.94 ± 5.54	79.85 ± 6.29
German	72.0 ± 1.9	74.2 ± 6.11	74.4 ± 6.12	74.3 ± 8.52	73.6 ± 7.98	**74.8 ± 6.46**
WDBC	**97.89 ± 2.33**	97.37 ± 1.41	**97.89 ± 2.05**	**97.89 ± 2.05**	97.01 ± 2.25	97.36 ± 1.97
DLBCL Stanford	97.5 ± 7.5	**100.0 ± 0.0**	98.0 ± 6.0	69.5 ± 15.24	96.0 ± 8.0	98.0 ± 6.0
DLBCL Tumor	92.5 ± 8.29	97.5 ± 5.0	94.64 ± 6.59	87.32 ± 11.19	94.82 ± 6.36	**97.5 ± 5.0**
DLBCL NIH	–	68.75 ± 12.25	65.83 ± 14.04	–	63.33 ± 10.34	**69.17 ± 10.74**
AMLALL	95.89 ± 6.29	**100.0 ± 0.0**	**100.0 ± 0.0**	87.68 ± 9.46	98.57 ± 4.29	**100.0 ± 0.0**

Note: Symbol '–' indicates that the corresponding algorithms failed to complete execution on the dataset within 30 days. The bold font indicates the highest accuracy achieved on this dataset.

Table 12: Comparison of the classification accuracy achieved by DT on 10 datasets, utilizing six distinct feature selection algorithms (%)

Dataset	FRFS	HANDI	FBC	NFEM	MFBC	CSFS
Wine	85.3 ± 69.57	90.52 ± 9.3	93.89 ± 6.78	88.27 ± 11.2	92.19 ± 7.53	**95.0 ± 5.24**
Glass	59.74 ± 14.31	64.39 ± 18.17	63.94 ± 13.6	64.44 ± 10.81	62.14 ± 15.21	**65.97 ± 12.4**
Ionosphere	92.31 ± 3.62	93.17 ± 2.9	**93.44 ± 3.85**	90.87 ± 5.98	90.03 ± 6.29	93.17 ± 4.07
MuskV1	65.69 ± 14.84	68.7 ± 18.02	**75.05 ± 8.88**	73.32 ± 7.8	72.97 ± 10.42	73.54 ± 4.58
German	63.1 ± 5.07	68.7 ± 5.57	68.8 ± 3.71	68.2 ± 6.71	**69.1 ± 1.7**	**69.1 ± 1.7**
WDBC	91.93 ± 3.43	93.68 ± 2.5	**95.08 ± 2.8**	93.15 ± 2.76	94.19 ± 3.07	94.73 ± 3.59
DLBCL Stanford	94.0 ± 9.17	**96.0 ± 8.0**	**96.0 ± 8.0**	77.01 ± 6.61	93.0 ± 10.77	**96.0 ± 8.0**
DLBCL Tumor	89.82 ± 9.42	**94.64 ± 6.59**	91.96 ± 6.6	81.61 ± 13.17	91.96 ± 6.6	92.5 ± 11.46
DLBCLNIH	–	62.92 ± 10.94	**63.75 ± 6.47**	–	58.33 ± 7.68	**63.75 ± 6.19**
AMLALL	94.46 ± 6.8	**97.14 ± 5.71**	95.71 ± 9.15	87.5 ± 7.66	**97.14 ± 5.71**	**97.14 ± 5.71**

Note: Symbol '–' indicates that the corresponding algorithms failed to complete execution on the dataset within 30 days. The bold font indicates the highest accuracy achieved on this dataset.

Compared with other comparison methods, the CSFS algorithm focuses more attention on samples that have not yet been perfectly partitioned under current conditions when measuring the uncertainty of target attributes in feature subsets and is more sensitive to improving the uncertainty of such samples when selecting subsequent features.

5.3 Parameter Analysis

The value of parameter 'λ' in the CSFS algorithm impacts the outcome of feature selection. In Fig. 5a–d, we can observe the correlation between the number of selected features and the value of parameter 'λ' in the CSFS algorithm. In Fig. 6a–d, we can observe the correlation between the classification accuracy of selected features and the value of parameter 'λ' in the CSFS algorithm.

Figs. 5 and 6 clearly indicate the following findings:

- As the value of parameter 'λ' increases, the overall tendency is for the number of selected features to decrease. This is because with an increase in the value of 'λ', the number of cross-similar relationships being removed also rises.
- When the value of parameter 'λ' surpasses a certain threshold, the classification accuracy undergoes a drastic drop. This is because at high values of 'λ', crucial information is also being eliminated, leading to a significant decline in performance.
- The parameter 'λ' serves to eliminate noisy and redundant relationships, hence the value of 'λ' should generally not be set too high.

Let's illustrate the impact of the truncation parameter λ on noise reduction in feature selection through a specific case. For the gene dataset DLBCL NIH, which contains noisy data, by observing Fig. 6c, we find that as λ increases from a small value to a large one, the accuracy of the selected attribute subset in the classifier first increases and then decreases. This indicates that within a certain range, the parameter λ can effectively eliminate the impact of noise and identify a truly meaningful subset of attributes for target classification.

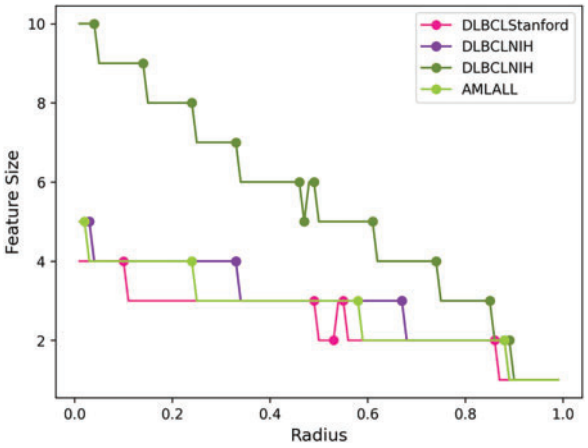

Figure 5: The impact of parameter 'λ' on the number of selected features

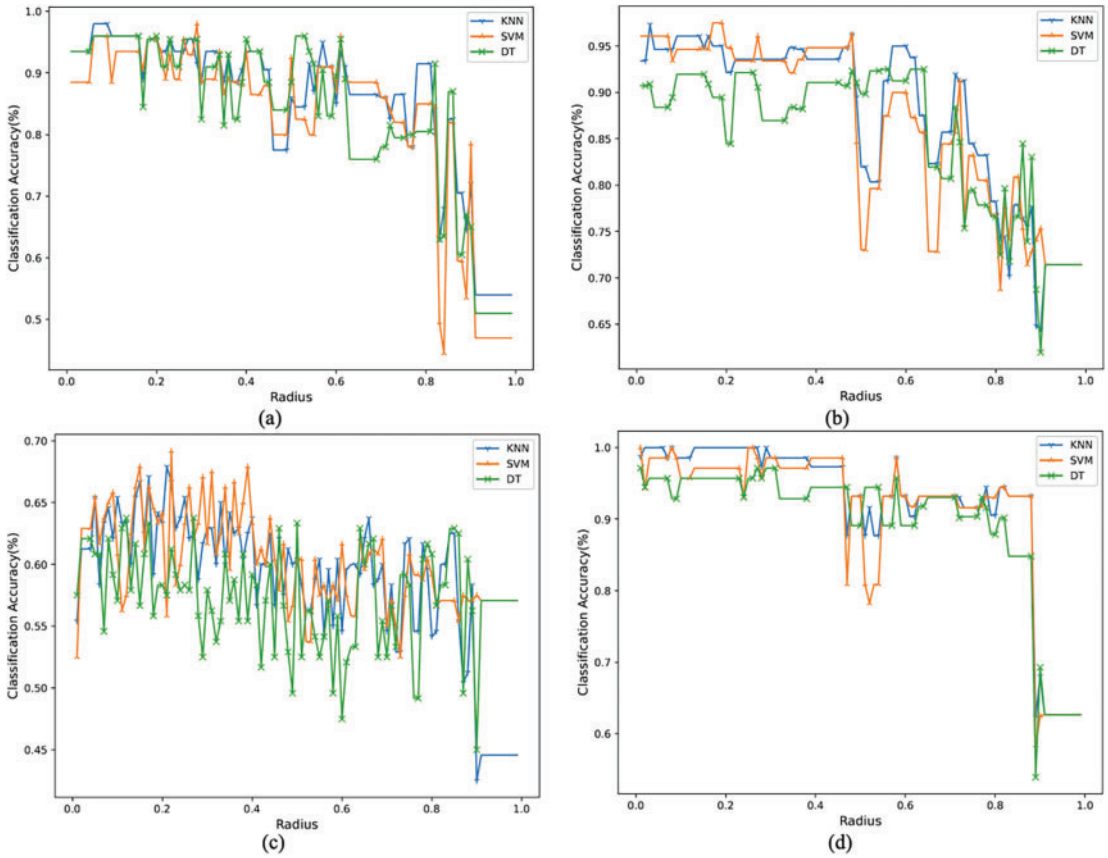

Figure 6: The impact of parameter 'λ' on the classification accuracy of selected features. (a) DLBCL Stanford (b) DLBCL Tumor (c) DLBCL NIH (d) AMLALL

6 Conclusions and Discussions

In this article, we introduce a novel and unique metric for quantifying data uncertainty, termed cross-similarity. This metric iteratively emphasizes similarity relations that substantially influence the uncertainty of fuzzy decision tables. Leveraging this metric, we have devised a forward greedy feature selection algorithm, known as CSFS. The CSFS algorithm significantly reduces the time complexity of feature selection algorithms based on fuzzy rough sets, transforming it from O (m^2n^2) to O (mn^2), thereby greatly enhancing computational efficiency. Experimental results validate that the CSFS algorithm not only accelerates the feature selection process but also accurately identifies the most representative features. Consequently, our algorithm effectively tackles the challenges posed by the high computational complexity of existing fuzzy rough set-based feature selection algorithms.

However, the CSFS algorithm also has some limitations. Firstly, when integrating multi-attribute similarity relationships, it relies on the minimum similarity relationship, making the algorithm sensitive to outlier features. Secondly, the uncertainty measure is based on cross-similarity within the decision table, thus its application is limited to comparing the uncertainty of different attribute subsets within the same decision table, rather than across different decision tables.

In response to the above limitations, our future research endeavors will encompass the following: Firstly, we aim to delve into multi-attribute similarity metrics that consider sample distribution or incorporate linear reconstruction of samples prior to similarity assessment. Both approaches hold promise in reducing the impact of outlier features, ultimately bolstering the algorithm's robustness. Secondly, to navigate the trade-off between computational complexity and robustness, we will explore the potential of parallel computing. Lastly, we will undertake further research into normalized uncertainty measures grounded in cross-similarity, with the objective of expanding their applicability across diverse scenarios.

Acknowledgement: The authors would like to thank the anonymous reviewers for their valuable comments.

Funding Statement: This work was supported by the Anhui Provincial Department of Education University Research Project (2024AH051375), Research Project of Chizhou University (CZ2022ZRZ06), Anhui Province Natural Science Research Project of Colleges and Universities (2024AH051368) and Excellent Scientific Research and Innovation Team of Anhui Colleges (2022AH010098).

Author Contributions: The authors confirm their contribution to the paper as follows: study conception and design: Wenchang Yu; analysis and interpretation of results: Xiaoqin Ma, Zheqing Zhang; draft manuscript preparation: Qinli Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

FRS	Fuzzy rough set
FCRS	Fuzzy compatible relation space
FDT	Fuzzy decision table
FRSM	Fuzzy rough set model
CSFS	Cross-similarity based feature selection

References

1. Pawlak Z. Rough sets. Int J Comput Inf Sci. 1982;11(4):341–56. doi:10.1007/BF01001956.
2. Dinçer H, Yüksel S, An J, Mikhaylov A. Quantum and AI-based uncertainties for impact-relation map of multidimensional NFT investment decisions. Finance Res Lett. 2024;66(3):105723. doi:10.1016/j.frl.2024.105723.
3. Kamran M, Salamat N, Jana C, Xin Q. Decision-making technique with neutrosophic Z-rough set approach for sustainable industry evaluation using sine trigonometric operators. Appl Soft Comput. 2025;169(2):112539. doi:10.1016/j.asoc.2024.112539.
4. Banerjee M, Mitra S, Banka H. Evolutionary rough feature selection in gene expression data. IEEE Trans Syst Man Cybern C Appl Rev. 2007;37(4):622–32. doi:10.1109/TSMCC.2007.897498.
5. Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst. 1990;17(3):191–209. doi:10.1080/03081079008935107.
6. Mi J, Leung Y, Zhao H, Feng T. Generalized fuzzy rough sets determined by a triangular norm. Inf Sci. 2008;178(12):3203–13. doi:10.1016/j.ins.2008.03.013.
7. Yao W, Zhang G, Zhou C. Real-valued hemimetric-based fuzzy rough sets and an application to contour extraction of digital surfaces. Fuzzy Sets Syst. 2023;459(1–3):201–19. doi:10.1016/j.fss.2022.07.010.
8. Zhang QL, Chen YY, Zhang GQ, Li ZW, Chen LJ, Wen CF. New uncertainty measurement for categorical data based on fuzzy information structures: an application in attribute reduction. Inf Sci. 2021;580(5):541–77. doi:10.1016/j.ins.2021.08.089.
9. Yao Y, Zhao Y. Attribute reduction in decision-theoretic rough set models. Inf Sci. 2008;178(17):3356–73. doi:10.1016/j.ins.2008.05.010.
10. Zhao S, Tsang E, Chen D, Wang X. Building a rule-based classifier—a fuzzy-rough set approach. IEEE Trans Knowl Data Eng. 2009;22(4):624–38. doi:10.1109/TKDE.2009.118.
11. Hu Q, Zhang L, An S, Zhang D, Yu D. On robust fuzzy rough set models. IEEE Trans Fuzzy Syst. 2011;20(4):636–51. doi:10.1109/TFUZZ.2011.2181180.
12. Ma L. Two fuzzy covering rough set models and their generalizations over fuzzy lattices. Fuzzy Sets Syst. 2016;294(11):1–17. doi:10.1016/j.fss.2015.05.002.
13. Deer L, Cornelis C. A comprehensive study of fuzzy covering-based rough set models: definitions, properties and interrelationships. Fuzzy Sets Syst. 2018;336(16):1–26. doi:10.1016/j.fss.2017.06.010.
14. Yang B, Hu B. On some types of fuzzy covering-based rough sets. Fuzzy Sets Syst. 2017;312:36–65. doi:10.1016/j.fss.2016.10.009.
15. Zhang Q, Yan S, Peng Y, Li Z. Attribute reduction algorithms with an anti-noise mechanism for hybrid data based on fuzzy evidence theory. Eng Appl Artif Intell. 2024;129:107659. doi:10.1016/j.engappai.2023.107659.
16. Zhang L, Zhan J. Fuzzy soft β-covering based fuzzy rough sets and corresponding decision-making applications. Int J Mach Learn Cybern. 2019;10(6):1487–502. doi:10.1007/s13042-018-0828-3.
17. Jensen R, Shen Q. Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst. 2007;15(1):73–89. doi:10.1109/TFUZZ.2006.889761.
18. Yang T, Zhong X, Lang G, Qian Y, Dai J. Granular matrix: a new approach for granular structure reduction and redundancy evaluation. IEEE Trans Fuzzy Syst. 2020;28(12):3133–44. doi:10.1109/TFUZZ.2020.2984198.
19. Huang Z, Li J. A fitting model for attribute reduction with fuzzy XX-covering. Fuzzy Sets Syst. 2021;413:114–37. doi:10.1016/j.fss.2020.07.010.
20. Huang Z, Li J. Discernibility measures for fuzzy β covering and their application. IEEE Trans Cybern. 2021;52(4):9722–35. doi:10.1109/TCYB.2021.3054742.
21. Huang Z, Li J, Qian Y. Noise-tolerant fuzzy-β-covering-based multigranulation rough sets and feature subset selection. IEEE Trans Fuzzy Syst. 2021;30(9):2721–35. doi:10.1109/TFUZZ.2021.3093202.
22. Wan J, Chen H, Li T, Yuan Z, Liu J, Huang W, et al. Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures. IEEE Trans Cybern. 2023;53(2):1208–21. doi:10.1109/TCYB.2021.3112203.
23. Hu Y, Zhang Y, Gong D. Multiobjective particle swarm optimization for feature selection with fuzzy cost. IEEE Trans Cybern. 2021;51(3):874–88. doi:10.1109/TCYB.2020.3015756.

24. Deng Z, Li T, Deng D, Liu K, Zhang P, Zhang S, et al. Feature selection for label distribution learning using dual-similarity based neighborhood fuzzy entropy. Inf Sci. 2022;615(10):385–404. doi:10.1016/j.ins.2022.10.054.
25. Dai J, Zou X, Qian Y, Wang X. Multifuzzy β-covering approximation spaces and their information measures. IEEE Trans Fuzzy Syst. 2022;31(5):955–69. doi:10.1109/TFUZZ.2022.3193448.
26. Wang C, Hu Q, Wang X, Chen D, Qian Y, Dong Z. Feature selection based on neighborhood discrimination index. IEEE Trans Neural Netw Learn Syst. 2017;29(8):2986–99. doi:10.1109/TNNLS.2017.2710422.
27. UCI machine learning repository [Internet]. Irvine, CA, USA: University of California; [cited 2024 Oct 1]. Available from: http://archive.ics.uci.edu.
28. Kent Ridge bio-medical data set repository [Internet]. Granada, Spain: University of Granada; [cited 2024 Oct 1]. Available from: https://leo.ugr.es/elvira/DBCRepository.