Proceedings of the 4th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/67/2024.18711

A Classification Model for Rheumatoid Arthritis Patients Based on Rough Sets and Three-Way Multi-Attribute Decision Making

Weiting Wang[1,2], Yan Zhang[1,3,*]

[1]Guangdong University of Finance and Economics, Guangzhou, China

[2]1246360264@qq.com
[3]zhangyan@gdufe.edu.cn
*corresponding author

Abstract. In clinical diagnosis, mainstream machine learning models often face common issues of incomplete data and multi-source heterogeneity when classifying patients. To address this, the paper defines distance measures for different attribute types in incomplete heterogeneous information systems, enabling effective handling of such information. First, based on the mixed distance between two objects, a binary relation in incomplete heterogeneous information systems was derived, and a multi-granularity rough set model for incomplete heterogeneous information was constructed. Second, on the basis of the multi-granularity rough set model, three-way multi-attribute decision-making was introduced, building a three-way multi-attribute decision-making classification model that considers risk costs and learning costs in the medical classification process. Finally, the proposed model was validated using real clinical data in the experimental section, demonstrating its effectiveness in classifying rheumatoid arthritis patients and improving classification accuracy.

Keywords: multi-granularity rough set, three-way multi-attribute decision-making, rheumatoid arthritis, classification.

1. Introduction

Rheumatoid arthritis (RA), a chronic inflammatory joint disease that severely impacts the quality of life of patients globally, has become a significant public health issue due to its high prevalence and extensive societal implications. RA is a chronic, progressive, systemic autoimmune disease that primarily affects joints, causing pain, swelling, stiffness, and functional loss. Beyond its detrimental effect on patients' quality of life, RA imposes a substantial economic burden on society and families. According to the World Health Organization (WHO), approximately 1% of the global population suffers from RA. In recent years, with the rapid aging of the population, the number of RA cases has been increasing annually. For instance, data from China's National Health Commission indicate that there are already millions of RA patients in China, and this number continues to rise each year. The severity of RA varies among patients, ranging from mild discomfort to severe joint damage, significantly affecting their daily lives and ability to work. Therefore, effectively classifying and managing RA patients, and developing personalized treatment plans, have become key challenges and focal points in current medical research.

DOI: 10.54254/2753-8818/67/2024.18711

However, the clinical diagnosis of RA is highly complex. Influencing factors typically include the patient's overall symptoms, tongue characteristics, pulse patterns, and various medical instrument data. These data, due to their diverse sources and structures, present the characteristics of multi-source heterogeneity, complicating the precise classification of RA. Moreover, due to the difficulty in standardizing and generalizing clinical experience among physicians, diagnostic discrepancies for RA may arise between different doctors and medical institutions. To address these challenges, classification algorithms and data mining theories can now be utilized to identify similarities between patients and cases in existing medical databases, enabling more accurate classification of RA severity. This not only reduces the risk and time costs associated with clinical diagnosis but also helps avoid the time and financial losses caused by "over-diagnosis."

We collaborated with a hospital in Guangdong Province to collect real electronic medical records from over 1,000 patients, which we have named the Rheumatoid Arthritis Patient Dataset. Based on this dataset, we employed a research method combining multi-granularity rough sets and three-way multi-attribute decision-making to classify RA patients. The RA patient dataset contains attributes such as the degree of joint pain or swelling, tongue coating characteristics, pulse patterns, erythrocyte sedimentation rate (ESR), and C-reactive protein (CRP) levels. In traditional Chinese medicine, RA is diagnosed by evaluating the joints' condition in conjunction with overall symptoms, tongue appearance, and pulse readings to determine factors like cold-heat, deficiency-excess, yin-yang, and interior-exterior syndromes. In contrast, Western medicine relies on key indicators from medical instruments and clinical tests, such as ESR and CRP, to confirm RA. These attributes are critical to understanding RA and involve various complex binary relationships, such as equivalence relations, neighborhood relations, and dominance relations. The traditional rough set model, however, cannot adequately handle the multi-source heterogeneous data needed for RA patient classification. Therefore, effectively processing diverse and complex medical data is the primary problem this paper aims to solve.

To address this issue, Zhang Meng et al. [1] extended the traditional rough set by defining distance measures between two objects based on different data type attribute sets within the information system to handle complex medical data. Additionally, they derived a binary relation on the set of objects from the mixed distance between two objects in a heterogeneous information system, constructing a rough set model based on heterogeneous information. Building upon this, the present paper further investigates situations involving dominance relation attributes, defining Hamming distance to measure the distance between two objects in the dominance relation attribute set, thus constructing a rough set model for incomplete heterogeneous information. Moreover, as the classification of RA patients is a typical multi-granularity problem, this paper further develops a multi-granularity rough set model for multi-source heterogeneous information systems, which includes both optimistic and pessimistic rough set models, laying the theoretical foundation for the subsequent three-way multi-attribute decision classification.

After solving the issue of incomplete and multi-source heterogeneous data in medical classification, the next step is to address the classification of rheumatoid arthritis patients. Disease classification involves grading and categorizing patients based on the severity and urgency of their condition, which is crucial for the rational allocation of medical resources, improving diagnostic and treatment efficiency, and developing personalized treatment plans. Both domestically and internationally, research on disease classification has been a significant topic in the medical field, and considerable achievements have been made in patient classification studies. Researchers have effectively enhanced the accuracy and efficiency of patient classification by employing various machine learning algorithms, such as decision trees [2], support vector machines [3], and random forests [4]. Additionally, methods based on deep learning [5] have gradually become a research focus, significantly improving the predictive power of classification models. However, research on incomplete heterogeneous datasets and cost-sensitive medical data remains insufficient. Based on the rough set model for multi-source heterogeneous information systems, this paper introduces a three-way multi-attribute decision-making method to construct a three-way multi-attribute classification model. This model considers the risk costs and learning costs in disease prediction, significantly improving the accuracy of RA classification.

DOI: 10.54254/2753-8818/67/2024.18711

Effective patient classification and clinical diagnosis can help medical institutions enhance the quality of patient care, optimize resource allocation, and provide better medical services. This study aims to develop a novel RA patient classification model that integrates multi-granularity rough set theory with a three-way multi-attribute decision-making mechanism, with the goal of overcoming the limitations of single or traditional classification methods in complex clinical settings, and improving the scientific validity and accuracy of RA patient classification. Through in-depth exploration and comprehensive evaluation of various clinical indicators, this model is expected to assist physicians in identifying different subtypes of RA earlier and formulating highly targeted and effective individualized treatment plans based on the characteristics of the disease. Ultimately, this will contribute to improving the overall diagnosis and treatment of rheumatoid arthritis and enhancing patient health outcomes.

2. Rough Set Model Based on Incomplete and Heterogeneous Information

Given a decision information table $S =< U, A = C \cup D, V, f >$, where U is a non-empty finite set of objects, referred to as the universe; A is a non-empty finite set of attributes, with C being the set of condition attributes and D the set of decision attributes; for each attribute $a \in A$, there exists a corresponding information function $f: U \times A \to V$, where $V_a = \{a(x) | x \in U\}$ represents the value range of attribute a.

Definition 1: Let (U, A) be a complete heterogeneous information system. For $P \subseteq A$ and $x_i, x_j \in U$, the binary equivalence relation of attribute set P, denoted as $IND(P)$, is defined as:

$$IND(P) = \{(x_i, x_j) \in U \times U \mid a(x_i) = a(x_j), \forall a \in P\}$$

Definition 2: Let (U, A) be a complete heterogeneous information system. For $P \subseteq A$ and $x_i, x_j \in U$, the equivalence class of x_i with respect to attribute set P, denoted as $[x_i]_p$, is defined as:

$$[x_i]_p = \{x_j \in U \mid (x_i, x_j) \in IND(P)\}$$

Due to the incompleteness and complexity of medical data, it is often difficult to satisfy the condition $a(x_i) = a(x_j)$, making it impossible to meet the binary equivalence relation $IND(P)$ in the information system. As pointed out in [1], the relationship between $a(x_i)$ and $a(x_j)$ can generally be considered from two aspects: similarity, which the aforementioned binary relation reflects; and dissimilarity, which can be measured using a distance function. When $a(x_i) = a(x_j)$, the distance between the two is regarded as 0.

Given the incomplete and heterogeneous nature of medical data, we now propose a distance function between two objects in the attribute set of an incomplete heterogeneous information system. Based on this, we derive the mixed distance between two objects in an incomplete heterogeneous information system. Furthermore, we construct a rough set model based on this system and define the lower and upper approximations of the model.

Definition 3: For $\forall x_i, x_j \in U$, let $a(x_i)$ and $a(x_j)$ represent the values of samples x_i and x_j on attribute a, respectively. The distance between samples x_i and x_j is defined as follows:

$$d(a(x_i),a(x_j)) = \begin{cases} 1, & x_i \text{ or } x_j \text{ has missing values on attribute a} \\ db(a(x_i), a(x_j)), & \text{an equivalence relation attribute} \\ dr(a(x_i), a(x_j)), & \text{a neighborhood relation attribute} \\ ds(a(x_i), a(x_j)), & \text{a dominance relation attribute} \end{cases}$$

Definition 4: Equivalence Relation Attribute Distance db

Let (U, A) be an incomplete heterogeneous information system. For $x_i, x_j \in U$, the equivalence relation attribute distance db is defined as:

$$db = \begin{cases} 0, a(x_i) = a(x_j) \\ 1, a(x_i) \neq a(x_j) \end{cases}$$

Definition 5: Neighborhood Relation Attribute Distance dr

Let (U, A) be an incomplete heterogeneous information system. For non-missing values $x_i, x_j \in U$, the neighborhood relation attribute distance dr is defined as:

DOI: 10.54254/2753-8818/67/2024.18711

$$dr = \frac{\sqrt{2}}{2}\sqrt{(a(x_i)^l - a(x_j)^l)^2 + (a(x_i)^r - a(x_j)^r)^2}, x_i, x_j \in U$$

where $a(x_i)^l$ and $a(x_i)^r$ are the lower and upper bounds of the neighborhood relation attribute for $a(x_i)$, and $a(x_j)^l$ and $a(x_j)^r$ are the lower and upper bounds for $a(x_j)$.

Definition 6: Dominance Relation Attribute Distance ds

Let (U, A) be an incomplete heterogeneous information system. For non-missing values $x_i, x_j \in U$, the dominance relation attribute distance ds is defined as the Hamming distance. When the one-hot encoding of $a(x_i)$ and $a(x_j)$ differ at the same position, the distance increases by 1.

Definition 7: Mixed Distance dm

Let (U, A) be an incomplete heterogeneous information system. For $a \in A$ and $x_i, x_j \in U$, the mixed distance dm between x_i and x_j in the entire system is defined as:

$$dm = \sqrt{\frac{1}{|A|}\sum_{a \in A} d^2(a(x_i), a(x_j))}$$

Definition 8

Let (U, A) be an incomplete heterogeneous information system. For $\theta \in [0,1]$, the binary relation on the universe U is denoted as T_A^θ, defined as:

$$T_A^\theta = \{(x_i, x_j) \in U \times U | d_A(x_i, x_j) \leq \theta\}$$

For any $x_i (x_i \in A)$, its equivalence class is defined as:

$$[x_i]_A^\theta = \{x_j | (x_i, x_j) \in T_A^\theta\}$$

Definition 9

Let (U, A) be an incomplete heterogeneous information system. For $X \subseteq U$, the upper and lower approximations of the rough set model based on the incomplete heterogeneous information system are defined as:

$$\overline{T_A^\theta} X = \{x \in U | T_A^\theta \cap X = \varnothing\},$$
$$\underline{T_A^\theta} X = \{x \in U | T_A^\theta \subseteq X\}.$$

For any $X \subseteq U$, the corresponding upper and lower approximations divide the universe U into three mutually exclusive regions: the positive region POS(X), the boundary region BND(X), and the negative region NEG(X), which are defined as:

$$POS(X) = \underline{T_A^\theta} X,$$
$$NEG(X) = U - \overline{T_A^\theta} X$$
$$BND(X) = \overline{T_A^\theta} X - \underline{T_A^\theta} X$$

In the probabilistic rough set model based on incomplete heterogeneous information, the conditional probability of an object x belonging to X given that it belongs to the equivalence class $[x]_A^\theta$ is:

$$P(X | [x]_A^\theta) = \frac{|X \cap [x]_A^\theta|}{|[x]_A^\theta|}$$

The following equivalent conditions can be derived:

$$P(X|[x]_A^\theta) = 1 \Leftrightarrow [x]_A^\theta \subseteq X,$$
$$P(X|[x]_A^\theta) = 0 \Leftrightarrow [x]_A^\theta \cap X = \varnothing,$$
$$0 < P(X|[x]_A^\theta) < 1 \Leftrightarrow [x]_A^\theta \cap X \neq \varnothing \wedge \neg([x] \subseteq X)$$

Based on the definitions of positive, boundary, and negative regions, the equivalent expressions are:

$$POS(X) = \{x \in U | P(X|[x]_A^\theta) = 1\},$$
$$BND(X) = \{x \in U | 0 < P(X|[x]_A^\theta) < 1\},$$
$$NEG(X) = \{x \in U | P(X|[x]_A^\theta) = 0\}.$$

DOI: 10.54254/2753-8818/67/2024.18711

In the decision rough set model based on incomplete heterogeneous information, threshold pairs (α, β) are introduced to partition the universe U, where $0 \leq \beta < \alpha \leq 1$. The calculation and interpretation of thresholds are given through Bayesian decision theory. Therefore, the definitions of the positive, negative, and boundary regions can be expressed as:

$$\mathrm{POS}(X) = \left\{x \in U \middle| P(X|[x]_A^{\theta}) \geq \alpha\right\}$$
$$\mathrm{NEG}(X) = \left\{x \in U \middle| P(X|[x]_A^{\theta}) \leq \beta\right\}$$
$$\mathrm{BND}(X) = \left\{x \in U \middle| \beta < P(X|[x]_A^{\theta}) < \alpha\right\}$$

3. Multi-attribute Three-way Classification Model for Rheumatoid Arthritis Patients

In the clinical diagnosis of rheumatoid arthritis (RA), traditional two-way decision models have limitations and can no longer meet the complex demands of modern medical decision-making. To improve diagnostic accuracy and reduce the risk of misdiagnosis, this paper proposes a novel decision framework based on a multi-attribute three-way decision model. This framework leverages minimum-risk Bayesian decision-making and rough set theory, classifying patients into three categories—severe, mild, and delayed diagnosis—by defining positive, negative, and boundary regions, thus accounting for costs and risks in the decision process. To determine the optimal decision threshold, a genetic algorithm is employed to iteratively optimize the objective function value. Ultimately, the optimal threshold pair is combined with a multi-granulation rough set model to construct a complete and efficient decision framework, enabling precise classification of RA patients' disease severity. This effectively reduces the high-risk costs of misclassification or missed diagnoses and enhances the practicality and stability of the classification system.

3.1. Multi-attribute Three-way Decision Model

Traditional two-way decisions are based on simple "yes" or "no" outcomes. Applied to the classification of RA patients, this approach divides patients into those requiring hospitalization and those who do not—corresponding to severe and mild patients, respectively. However, in real medical decision-making, it is not always possible to make a binary decision. In cost-sensitive clinical decisions, diagnosing a mild patient as severe incurs additional time and financial costs for the patient, while diagnosing a severe patient as mild can result in unpredictable harm to the patient's health. Therefore, binary decisions, if misjudged, carry a high-risk cost. In contrast, three-way decisions account for decision-making costs and introduce delayed decisions, reducing the risks associated with binary decision-making. Based on a multi-granulation rough set model with incomplete heterogeneous information, this paper constructs a multi-attribute three-way decision model, which classifies patients into severe, mild, and delayed diagnosis categories according to risk costs in the clinical decision-making process.

Definition 10: Let $U = \{x_1, x_2, \ldots, x_n\}$ be a finite non-empty set, and A a finite condition set. For the condition set A, three-way decision-making maps the set U into three mutually exclusive regions—POS (positive region), NEG (negative region), and BND (boundary region)—using a mapping function f, expressed as $U \xrightarrow{f} \{\mathrm{POS}, \mathrm{NEG}, \mathrm{BND}\}$, where POS, NEG, and BND are subsets of U and satisfy the conditions: $\mathrm{POS} \cap \mathrm{NEG} = \emptyset, \mathrm{POS} \cap \mathrm{BND} = \emptyset, \mathrm{NEG} \cap \mathrm{BND} = \emptyset$.

Definition 11: Let $S = (U, A)$ be an information system, where $U = \{x_1, x_2, \ldots, x_{|U|}\}$ is a finite non-empty set of objects, referred to as the universe or object space, and $A = \{a_1, a_2, \ldots, a_{|A|}\}$ is a finite non-empty set of attributes, where elements in A are called attributes. For each attribute $a \in A$, there is a mapping a: $a: U \to a(U)$, and $a(U) = \{a(u) | u \in U\}$ is referred to as the value domain of attribute a. For $P \subseteq A$, and $x_i, x_j \in U$, any subset R satisfying $\emptyset \neq R \subseteq A$, the equivalence relation of attributes P on U, denoted $\mathrm{IND}(P)$, is defined as $\mathrm{IND}(P) = \left\{(x_i, x_j) \in U \times U \middle| a(x_i) = a(x_j), \forall a \in P\right\}$.

Definition 12

Let the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ be a finite set of m states, and $A = \{a_1, a_2, \ldots, a_n\}$ be a finite set of n possible actions. $(\omega_i|x)$ is the conditional probability of x under the given state ω_i, and $\lambda(a_j|\omega_i)$

DOI: 10.54254/2753-8818/67/2024.18711

represents the loss or cost of taking action a_j under state ω_i. For an object x, if action a_j is chosen, the expected loss is:

$$R(a_j|x) = \sum_{i=1}^{m} \lambda(a_j|\omega_j) \cdot P(\omega_i$$

That is, for each object x, the loss function for each action can be calculated, and the expected loss for taking different actions under different states can be determined, allowing the selection of the action with the minimum conditional risk.

The rough set decision-making process uses two state sets and three action sets to describe the decision process. The state set $\Omega = \{X, \neg X\}$ represents whether an event belongs to X or not. The action set $A = \{a_P, a_B, a_N\}$ corresponds to three actions: accept the event, delay the decision, or reject the event. Considering that different actions may incur different losses, let λ_{PP}, λ_{BP}, λ_{NP} represent the losses when x belongs to X, and actions $a_P, a_B,$ and a_N are taken, respectively. Similarly, λ_{PN}, λ_{BN}, λ_{NN} represent the losses when x does not belong to X, and actions $a_P, a_B,$ and a_N are taken, respectively. Thus, given the equivalence class description $[x]_R$ of object x, the expected loss for taking actions $a_P, a_B,$ and a_N can be expressed as:

$$R(a_P|[x]R) = \lambda_{PP}P(X|[x]_R) + \lambda_{PN}P(\neg X|[x]_R)$$
$$R(a_B|[x]R) = \lambda_{BP}P(X|[x]_R) + \lambda_{BN}P(\neg X|[x]_R)$$
$$R(a_N|[x]R) = \lambda_{NP}P(X|[x]_R) + \lambda_{NN}P(\neg X|[x]_R)$$

where the conditional probability $P(X|Y) = \frac{|X \cap Y|}{|Y|}$.

According to the Bayesian minimum-cost decision principle, the following rules can be derived:

(P):If $R(a_P|[x]_R) < R(a_B|[x]_R), R(a_P|[x]R) < R(a_N|[x]_R)$,thus, $x \in \mathrm{POS}(X)$.

(B):If $R(a_B|[x]_R) < R(a_P|[x]_R), R(a_B|[x]R) < R(a_N|[x]_R)$,thus, $x \in \mathrm{BND}(X)$.

(N):If $R(a_N|[x]_R) < R(a_P|[x]_R), R(a_N|[x]R) < R(a_B|[x]_R)$,thus, $x \in \mathrm{NEG}(X)$.

Through a series of derivations, the threshold formulas for the cost function can be obtained as:

$$\alpha = (1 + \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{PN} - \lambda_{BN}})^{-1} = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}$$
$$\beta = (1 + \frac{\lambda_{NP} - \lambda_{PP}}{\lambda_{BN} - \lambda_{NN}})^{-1} = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}$$

The threshold pair (α, β) satisfies $0 \leq \beta < \alpha \leq 1$, and the final decision rules are:

$$\mathrm{POS}(X) = \{x \in U | P(X|[x]_R) \geq \alpha\}$$
$$\mathrm{BND}(X) = \{x \in U | \beta < P(X|[x]_R) < \alpha\}$$
$$\mathrm{NEG}(X) = \{x \in U | P(X|[x]_R) \leq \beta\}$$

These three parts correspond to the three different classifications of the three-way decision model. For the RA diagnosis of a given patient, the doctor's attitude can be either "agree" or "disagree," denoted as X, ¬X. Thus, when using the three-way decision-making method for diagnosis, there are three possible outcomes: mild, delayed diagnosis, and severe, represented as a_P, a_B, a_N, respectively, as shown in Table 1.

Table 1. Loss Values Corresponding to Each Classification Decision

Decision Action	R	$\bar{R}$
Mild P	λ_{PP}	λ_{PN}
Moderate (Delayed Diagnosis) B	λ_{BP}	λ_{BN}
Severe N	λ_{NP}	λ_{NN}

Note: The loss value refers to the risk associated with the patient's recovery as assessed by the doctor.

Based on the loss values corresponding to clinical diagnostic decisions and the threshold calculation formula, two thresholds, α and β, can be calculated. The diagnostic choice for each patient is then determined according to the decision rules. When P≥α, the diagnosis is classified as mild; when P≤ β,

DOI: 10.54254/2753-8818/67/2024.18711

the diagnosis is classified as severe; when P lies between α and β, the diagnosis is delayed. This method effectively reduces decision costs and enhances the specificity of clinical treatments.

3.2. Genetic Algorithm for Determining the Optimal Threshold Pair in Three-Way Decision-Making

When classifying using the three-way multi-attribute decision model, each decision action carries a corresponding risk loss. Minimizing the total risk loss is one of the current challenges to be addressed [7-8]. How to minimize the decision risk loss in three-way multi-attribute decision-making can be solved by setting appropriate thresholds. However, these thresholds usually need to be set by experts with domain-specific knowledge, which hinders the application of this model in clinical decision-making in medicine. In response, some scholars have proposed algorithms to automatically generate the optimal threshold pairs for three-way decision-making, such as adaptive algorithms, grid search, simulated annealing, and artificial fish swarm algorithms. Based on these algorithms, this paper attempts to apply a genetic algorithm to automatically generate the optimal threshold pairs for three-way decision-making.

The genetic algorithm is a heuristic optimization algorithm that solves problems by simulating biological evolution. Through simulating operations such as genetic inheritance, crossover, and mutation, genetic algorithms produce and improve a population of candidate solutions, thereby gradually optimizing the objective function. The pseudocode of the algorithm is described as follows:

```
Input: a set of probabilities.
Output: a threshold pair.
BEGIN
population = initialize_population()
iteration = 0
while iteration < max_iterations and not is_converged(population):
        fitness_values = calculate_fitness(population)
    selected_population = selection(population, fitness_values)
    offspring = crossover(selected_population)
    mutated_offspring = mutation(offspring)
    elite_individuals = select_elite(population, fitness_values)
    population = update_population(mutated_offspring, elite_individuals)
    iteration += 1
    optimal_threshold_pair = population[fitness_values.index(max(fitness_values))]
END BEGIN
```

By combining the optimal threshold pair with the multi-granularity rough set model based on incomplete heterogeneous information, a complete and efficient decision framework is constructed. In this framework, heterogeneous medical data are first effectively processed and refined through the multi-granularity rough set model. Then, the three-way multi-attribute decision model is used to make the final classification decision based on the pre-calculated optimal threshold pair. This enables the precise categorization of the severity of rheumatoid arthritis in patients, significantly reducing the high-risk costs associated with misdiagnosis and missed diagnosis while improving the practicality and stability of the classification system.

4. Integrating Multi-Granularity Rough Set with Three-Way Multi-Attribute Decision-Making for Incomplete Heterogeneous Information-Based Arthritis Classification

4.1. Problem Description

In the medical classification context proposed in this paper, real clinical medical data are used to validate the classification method. The data used in this paper include the diagnostic data of 1,015 patients with rheumatoid arthritis. This dataset not only covers the clinical signs of the patients, such as the degree of joint swelling and pain, but also includes blood test indicators like erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP), as well as traditional Chinese medicine-specific diagnoses, such as tongue

DOI: 10.54254/2753-8818/67/2024.18711

diagnosis and pulse patterns. These data originate from various sources and consist of diverse attribute types, including continuous quantitative indicators and discrete qualitative descriptions, forming a clear heterogeneous information structure.

To address this situation, the proposed method first establishes a multi-granularity rough set model based on incomplete heterogeneous information. By defining distance measures between different data types and attributes, the method effectively handles heterogeneous data from different dimensions. Based on this, the constructed multi-granularity rough set model can depict and analyze the condition of rheumatoid arthritis patients at multiple levels based on data granularity. Subsequently, three-way multi-attribute decision-making theory is introduced to address the cost-sensitive nature of risk in medical decision-making. Within this framework, the three-way multi-attribute decision classification model is constructed to comprehensively consider the impact of various indicators on the severity of the disease, as well as the potential medical costs associated with misdiagnosis and missed diagnosis, thus achieving a more precise and rational classification of rheumatoid arthritis patients.

4.2. Result Analysis

This section provides an in-depth analysis of the application of the integrated multi-granularity rough set and three-way multi-attribute decision model in the classification of rheumatoid arthritis. The experiment was validated using a real clinical dataset of rheumatoid arthritis patients, exploring the model's performance advantages in handling incomplete heterogeneous information and the improvement in classification accuracy when considering both risk cost and learning cost.

In the experimental phase, to optimize the accuracy and risk sensitivity of rheumatoid arthritis classification, we utilized the powerful optimization tool, the genetic algorithm, to determine the optimal threshold pair for three-way decision-making. After a series of iterations and selection processes, the genetic algorithm successfully identified the ideal threshold pair for the rheumatoid arthritis classification task as (0.70, 0.69). These two values represent the critical points in the three-way decision framework for distinguishing between mild, delayed diagnosis, and severe cases. By applying this optimal threshold pair to the three-way multi-attribute decision classifier, we conducted a detailed evaluation and validation of the classifier's performance. The specific results are as follows:

Table 2. Comparison of Classification Performance of Different Classifiers

Evaluation Metrics	Three-Way Multi-Attribute Decision	SVM	Random Forest	KNN	Neural Network
Accuracy	0.936	0.876	0.880	0.861	0.792
Precision	0.952	0.893	0.811	0.882	0.854
Recall	0.899	0.792	0.794	0.890	0.795
F1 Score	0.947	0.833	0.890	0.912	0.880

A rigorous comparative experiment was conducted between the three-way multi-attribute decision classifier based on genetic algorithms and traditional machine learning classifiers (such as SVM, Random Forest, KNN, and Neural Networks). The experimental results indicate that the three-way multi-attribute decision classifier demonstrated significant advantages across various evaluation metrics, including accuracy, precision, recall, and F1 score, achieving an accuracy of 93.6%, which far exceeds the performance of other machine learning classifiers.

This exceptional performance can be attributed to several factors: First, the three-way decision model considers the risk costs and learning costs involved in the medical decision-making process, allowing for the adoption of a delayed diagnosis strategy when making precise decisions is difficult, effectively avoiding the potential for overdiagnosis or treatment delays. Second, the integration of multi-granularity rough set theory enables flexible responses to heterogeneous data, allowing for reasonable integration of information at different granularity levels, thereby enhancing the model's generalization capability and stability when handling complex clinical data.

DOI: 10.54254/2753-8818/67/2024.18711

Moreover, it is worth noting that while the three-way decision classifier performs excellently overall, there may be some uncertainty in the classification results for patient samples that fall within the boundary region. Therefore, future research will consider introducing advanced strategies such as sequential three-way decision-making to further refine the classification of samples in the boundary area, aiming to more accurately distinguish the severity of patients' conditions while maintaining classification accuracy, thus providing strong support for clinicians in formulating more targeted treatment plans.

5. Conclusion and Outlook

This study fully integrates methods such as multi-granularity rough sets, three-way decisions, and multi-attribute decisions, proposing a multi-attribute classification method based on multi-granularity rough sets, which improves the performance of this classification method in handling incomplete heterogeneous information. It addresses the cost-sensitive issues in the medical classification process, enhances classification accuracy, and reduces decision costs. Ultimately, this method is applied to clinical practical issues, improving triage efficiency and promoting the dissemination of medical experience from a data mining perspective.

Generally, after classifying medical datasets using the three-way decision method, further division is still required for those patient samples categorized as belonging to the boundary region, typically through traditional experience, expert discussions, or sequential three-way decision methods. In future research, the consideration will be to apply the sequential three-way decision method to further refine the classification of boundary region samples until all samples can be distinctly categorized into definite severe and definite mild categories according to disease severity.

References

[1] Zhang, M., Sun, B., Wang, T., Chu, X., & Tong, S. (2022). Integration of rough sets and GRA for heterogeneous multi-criteria three-way recommendations and their application in healthcare recommendations. Control and Decision, 37(07), 1883-1893. https://doi.org/10.13195/j.kzyjc.2020.1631

[2] Kiran, S., Reddy, G. R., Girija, S. P., et al. (2023). A gradient boosted decision tree with binary spotted hyena optimizer for cardiovascular disease detection and classification. Healthcare Analytics, 3, 100173.

[3] Wu, G., Li, C., Yin, L., et al. (2023). Comparison between support vector machine (SVM) and deep belief network (DBN) for multi-classification of Raman spectroscopy for cervical diseases. Photodiagnosis and Photodynamic Therapy, 42, 103340.

[4] Parameswari, A., Kumar, K. V., & Gopinath, S. (2022). Thermal analysis of Alzheimer's disease prediction using random forest classification model. Materials Today: Proceedings, 66, 815-821.

[5] Hatami, M., Yaghmaee, F., & Ebrahimpour, R. (2024). Investigating the potential of reinforcement learning and deep learning in improving Alzheimer's disease classification. Neurocomputing, 597, 128119.

[6] Yao, Y. (2009). Three-way decisions with probabilistic rough sets. Information Sciences, 180(3).

[7] Liu, D., Yao, Y., & Li, T. (2011). Three-way investment decisions with decision-theoretic rough sets. International Journal of Computational Intelligence Systems, 4(1), 66-74.

[8] Zhang, Y., Zou, H., & Zhao, S. (2015). Cost-sensitive three-way decision model based on CCA. Journal of Nanjing University (Natural Sciences), 51(02), 447-452.