Soft Computing (2023) 27:1565–1582
https://doi.org/10.1007/s00500-022-07454-5

APPLICATION OF SOFT COMPUTING

Attribute reduction algorithm of neighborhood rough set based on supervised granulation and its application

Li Zou[1,2,3] · Siyuan Ren[1] · Yibo Sun[2,3] · Xinhua Yang[2,3]

Accepted: 8 August 2022 / Published online: 6 September 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
In neighborhood rough set theory, attribute reduction based on measure of information has important application significance. The influence of different decision classes was not considered for calculation of traditional conditional neighborhood entropy, and the improvement of algorithm based on conditional neighborhood entropy mainly includes of introducing multi granularity and different levels, while the mutual influence between samples with different labels is less considered. To solve this problem, this paper uses the supervised strategy to improve the conditional neighborhood entropy of three-layer granulation. By using two different neighborhood radii to adjust the mutual influence degree of different label samples, and by considering the mutual influence between conditional attributes through the feature complementary relationship, a neighborhood rough set attribute reduction algorithm based on supervised granulation is proposed. Experiment results on UCI data sets show that the proposed algorithm is superior to the traditional conditional neighborhood entropy algorithm in both aspects of reduction rate and reduction accuracy. Finally, the proposed algorithm is applied to the evaluation of fatigue life influencing factors of titanium alloy welded joints. The results of coupling relationship analysis show that the effect of joint type should be most seriously considered in the calculation of stress concentration factor. The results of influencing factors analysis show that the stress range has the highest weight among all the fatigue life influencing factors of titanium alloy welded joint.

Keywords Neighborhood rough set · Attribute reduction · Supervised granulation · Fatigue life · Welded joints

1 Introduction

Attribute reduction (also known as feature reduction) (Zhang and Miao 2014) is a very useful data preprocessing technique, which removes noise, irrelevant or misleading features by mining the importance of feature, and obtains the smallest subset of features from the decision system while maintaining the same classification accuracy. In today's era of data explosion, attribute reduction of data can greatly improve the utilization of data, reduce storage space, save resources, and promote the visualization and understanding of data.

Rough Set Theory (RST) (Pawlak 1982) is an effective feature reduction tool proposed by Professor Zdzisław Pawlak in 1982, and has been widely used in many fields such as fault diagnosis (Xu et al.2020), pattern recognition (Sinha and Namdev 2020), and data mining (Singh and Pamula 2020). RST can obtain the key attributes from the data itself without any prior knowledge, so it is suitable for obtaining the key factors that affect the fatigue life of welded joints from the fatigue test data of welded joints, and can obtain the objective and comprehensive evaluation of the influencing factors of the fatigue life of welded joints. In recent years, RST has been applied to the fatigue life analysis of welded joints (Liu et al.2017; Zou et al. 2019a, b), and the evaluation model and fatigue life

✉ Xinhua Yang
yangxhdl@foxmail.com

[1] Software Technology Institute, Dalian Jiaotong University, Dalian 116028, China

[2] Dalian Key Laboratory of Welded Structures and Its Intelligent Manufacturing Technology (IMT) of Rail Transportation Equipment, Dalian Jiaotong University, Dalian 116028, China

[3] Liaoning Key Laboratory of Welding and Reliability of Rail Transportation Equipment, Dalian Jiaotong University, Dalian 116028, China

prediction model of welded joints are constructed. This paper proposes a neighborhood rough set attribute reduction algorithm based on supervised granulation. The possible influence of different classes in decision attribute from the perspectives of neighborhood partition and calculation of conditional entropy could be considered by the algorithm. And the influence of conditional attributes was also considered by introducing mutual information theory. Based on the proposed algorithm, the evaluation model of influencing factors of titanium alloy welded joint was established. The coupling relationship between influencing factors was discussed. The influencing factors of joint fatigue life were quantitatively evaluated, and the set of key influencing factors of fatigue life was obtained.

2 Related works

Besides RST, information theory is also a method to deal with uncertainty problems. Shannon (1948) first proposed the concept of information entropy in 1948. Information entropy provides the quantitative measurement of information. Miao (1997) merged information theory with rough set theory in 1997, and established the relationship between the roughness of knowledge and information entropy. Using the information quantity of knowledge as the measure of attribute importance, Liang et al. (2001) quantified the relationship between knowledge and information quantity in information system. However, the classical rough set model can only deal with discrete data. For continuous data, it needs to be discretized, which will cause data loss. To solve this problem, Hu et al. (2006, 2008) proposed a neighborhood rough set model based on the definitions of δ neighborhood and neighborhood relations in metric spaces. Then Hu et al. (2009, 2011) generalizes Shannon's information entropy to neighborhood information entropy, and proposes a measure of neighborhood mutual information. According to the measurement attribute method, rough set is mainly divided into algebraic view method (Shen et al.2013) and information view method (Wang and Ou.2008). In which algebra view method calculates the weight value of attribute importance of features by calculating the upper and lower approximation of samples. The information view method is based on the idea of information theory. Through the study of the uncertainty of the universe, it calculates the information entropy or conditional information entropy to get the weight, so as to reduce the attributes of the data.

At present, the main research direction of rough set attribute reduction algorithm is divided into the improvement of different types of data and the improvement of different granulation methods. Due to the complexity and diversity of data, many scholars begin to study different types of data, including dynamic data, incomplete data, mixed data, etc. Then incremental reduction, dynamic reduction, multi decision table reduction and parallel reduction are developed. In the view of algebra, Chu et al. (2020) proposed a three-way clustering algorithm based on neighborhood rough sets for incomplete and attribute-related random large sample data; Deng et al. (2021) proposed F-neighbor rough sets and its reducts for dynamic numerical data, combining the advantages of neighbor rough set and F-rough set; Singh et al. (2020) introduced a novel approach for attribute selection in set-valued information system based on tolerance rough set theory. In terms of information view, Zhao and Qin (2014) proposes an extended rough set model based on neighborhood-tolerance relation for incomplete data mixed by categorical and numerical features, and then proposes conditional entropy of neighborhood tolerance; Sang et al. (2021) proposed incremental feature selection approaches based on a fuzzy dominance neighborhood rough set for dynamic interval-valued ordered data; Wan et al. (2021) proposed a new objective evaluation function of the interactive selection of hybrid features and designed a novel interaction feature selection algorithm based on neighborhood conditional mutual information for hybrid data; Chen et al. (2018a, b, c) proposed a variable precision neighborhood rough set attribute reduction heuristic algorithm based on mutual information entropy for incomplete hybrid decision system; Sun et al. (2020) proposed a novel neighborhood multi-granulation rough sets based attribute reduction method using Lebesgue and entropy measure in incomplete neighborhood decision system; Shu et al. (2020) proposed a neighborhood entropy-based incremental feature selection framework by neighborhood rough set model for dynamic hybrid data with mixed-type features. For multi label data, Qian et al. (2020) integrated label distribution learning into multi label feature selection, and proposed a multi-label feature selection algorithm based on label distribution and feature complementarity.

At present, more and more scholars combine rough set theory with granular computing (Zadeh 1997), and realize the transformation and representation of uncertain knowledge by using different granulation mechanisms. It makes the subsequent calculation start from different levels or granularity, and realizes the characterization of neighborhood information system from multiple perspectives. In the view of algebra, Zhang et al.(2019a, b) developed a novel model called local multi-granulation decision-theoretic rough set in an ordered information system; Zhan and Xu (2018) introduced two types of coverings based (optimistic, pessimistic and variable precision) multi granulation rough fuzzy set models respectively by means of neighborhoods and presented an approach to multiple criteria group decision making problem, and then Zhang et al. (2019a, b)

proposed two types of multi-granulation rough sets model called the optimistic multi-granulation hesitant fuzzy rough sets and pessimistic multi-granulation hesitant fuzzy rough sets; Tsang et al. (2020) investigated the mechanism of multi-level cognitive concept learning method oriented to data sets with fuzziness; Tan et al. (2019) defined several measurements to compare the granularity of neighborhood granulations, using which the granulation selection with multi granulation rough set is characterized; Jiang et al. (2019) proposed a multi-scale based acclerated stragegy for attribute reduction by means of the changing of radius; Chen et al. (2018c) proposed a three-level structure of granules in the neighborhood system: the neighborhood granule, the neighborhood granule swarm and the neighborhood granule library; Chen et al. (2018a) proposed a multi-radius neighborhood rough set weighted feature extraction method for high-resolution remote sensing image classification; Li et al. (2020) proposed a dynamic granularity selection algorithm by introducing local weighted accuracy and local likelihood ratio to compute the weight of granularity. In terms of information view, Zhao et al. (2015) proposed a new complement information entropy model in fuzzy rough set based on arbitrary fuzzy relation, which takes inner-class and outer-class information into consideration; Zhou et al. (2018,2020) applied the idea of three-layer construction to conditional neighborhood entropy; Zhao and Yang (2019) proposed an incremental attribute reduction algorithm for object constantly increasing in numeric information system; Mou et al. (2020) decomposed high classification- based neighborhood approximation condition-entropy and proposed a class-specific attribute reduct based on the new information measure. Mu et al. (2019) establishes double-granule conditional-entropies based on three-level granular structures by improvements of hierarchical granulation.

The parameter of radius in the neighborhood reduction algorithm plays an important role. Different radius results in different reduction result. If the same radius is used, the influence of samples under different labels cannot be fully considered. To solve this problem, Yang et al. (2019) proposed a pseudo-label neighborhood relation. On this basis, Rao et al. (2020) put forward relevant reduction acceleration strategies; Nevertheless, not only is it a time-consuming process for generating pseudo labels of samples, but also the information provided by pseudo labels may be incorrect which will lead to lower quality of neighborhood rough approximations. Jiang et al. (2020) proposed a supervised neighborhood based on the supervised strategy. By using two neighborhood radii, which successfully reduced the interference between samples with different labels.

The traditional rough set reduction algorithm based on information view fails to consider the influence relationship between different decision attributes when calculating conditional neighborhood entropy, and the traditional radius does not take into account the information provided by decision attributes. In this paper, a two-step reduction algorithm is proposed. We combine the supervised strategy (Jiang et al.2020) with the concept of three-level granulation (Zhou et al.2018) in the first step of reduction, and fully consider the influence of decision attributes from the perspective of determination and measurement calculation. In order to further consider the influence of different conditional attributes and eliminate those attributes that are too similar to each other and have little impact on decision attributes, the feature complementary relationship in reference (Qian et al.2020) is introduced in the second step of reduction. The two-step reduction algorithm is called the neighborhood rough set based on supervised granulation (NRSBSG) attribute reduction algorithm. Then, it is applied for the fatigue life influencing factors' analysis of titanium alloy welded joints, coupling relationship between the fatigue life influencing factors of titanium alloy welded joints are studied. The implicit relationship between the influencing factors is researched, and the corresponding intelligent model is constructed. The model is tested by using the fatigue experiment data of titanium alloy welded joints. At last, the analysis system of fatigue life influencing factors of welded joints based on NRSBSG is designed and developed.

The rest of this paper is organized as follows. Section 3, describes the theory of neighborhood rough set reduction by defining some concepts, and the proposed NRSBSG is introduced. Section 4, carries out some experiments on the standard UCI datasets and analyzes the results. Section 5, describes details with the design and implementation of the fatigue life influencing factors analysis system. Section 6, concludes the paper and presents further work in this area.

3 Basic concepts

3.1 Preliminary

Generally, a neighborhood decision system can be denoted as $NDS = (U, C, d, f)$, in which U is the set of nonempty samples, C is the set of conditional attributes and d is the decision attribute. $\forall x \in U$, $d(x)$ indicates the value of x over decision attribute. $IND(d) = \{(x, y) \in U \times U : d(x) = d(y)\}$ indicates the equivalent relation of decision attribute d,$U/IND(d) = \{X_1, X_2, ..., X_q\}$ indicates the sample division of decision attribute d,and X_q is the q decision class with the same label sample. In this Section, some basic

concepts of neighborhood rough set and the proposed NRSBSG are introduced.

3.1.1 Neighborhood relation

Definition 1. (Hu 2006) Distance function.

Given a decision system NDS, $U = \{x_1, x_2, ..., x_i\}$,$\forall B \subseteq C$, the conditional attribute collection $B = \{c_1, c_2, ..., c_n\}$,then the distance function of B is:

$$\Delta_B(x, x_i) = \left(\sum_{k=1}^{n} |f(x, c_k) - f(x_i, c_k)|^p \right)^{\frac{1}{p}} \tag{1}$$

in which $f(x, c_k)$ denotes the value of sample x with respect to conditional attribute c_k. When $p = 1$, it is Manhattan distance, and when $p = 2$, it is Euclidean distance. In this paper, we use Manhattan distance as the distance function.

Definition 2. (Hu 2006) Neighborhood relation.

Given a decision system NDS,$U = \{x_1, x_2, ..., x_i\}$, $x_i \in U$,the neighborhood of x_i can be denoted as:

$$n_B^{\delta}(x_i) = \{x | \Delta_B(x, x_i) < \delta, x \in U\} \tag{2}$$

in which conditional attribute collection $B \subseteq C$,$B = \{c_1, c_2, ..., c_n\}$ indicates the conditional attribute contained in the conditional attribute set B,$\Delta_B(x, x_i)$ indicates the distance between sample x_i and x with respect to conditional attribute set B,δ is the neighborhood radius. The neighborhood relation can be denoted as follows:

$$n_B^{\delta} = \{(x, y) | \Delta_B(x, y) < \delta, \forall x, y \in U\} \tag{3}$$

3.1.2 Neighborhood rough set based on information view

Hu et al. (2009) combined the classical Shannon entropy with neighborhood rough set, studied the correlation measure of neighborhood decision system under the information view, including neighborhood entropy, conditional neighborhood entropy and neighborhood mutual information. It can be directly applied to multi-label data with numerical and discrete characteristics without discretization. The concept is introduced as follows:

Definition 3. (Hu et al. 2009) Conditional neighborhood entropy.

Given a decision system NDS,$\forall A, B \subseteq C$,the conditional neighborhood entropy of conditional attribute set A with respect to conditional attribute set B is defined as:

$$NH_{\delta}(A/B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log \frac{|n_A^{\delta}(x_i) \cap n_B^{\delta}(x_i)|}{|n_B^{\delta}(x_i)|} \tag{4}$$

the conditional neighborhood entropy of decision attribute d with respect to the set of conditional attribute B is defined as:

$$NH_{\delta}(d/B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log \frac{|[x_i]_d \cap n_B^{\delta}(x_i)|}{|n_B^{\delta}(x_i)|} \tag{5}$$

where $[x_i]_d$ is the decision class corresponding to sample x_i.

Definition 4. (Hu et al. 2009) Neighborhood mutual information.

Given a decision system NDS,$\forall A, B \subseteq C$,the neighborhood mutual information of conditional attribute set A with respect to conditional attribute set B is defined as:

$$NH_{\delta}(A; B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log \frac{|n_A^{\delta}(x_i)||n_B^{\delta}(x_i)|}{|U||n_A^{\delta}(x_i) \cap n_B^{\delta}(x_i)|} \tag{6}$$

the neighborhood mutual information of decision attribute d with respect to the set of conditional attribute B is defined as:

$$NH_{\delta}(d; B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log \frac{|[x_i]_d||n_B^{\delta}(x_i)|}{|U||[x_i]_d \cap n_B^{\delta}(x_i)|} \tag{7}$$

If variables B and C are independent of each other, then the value of neighborhood mutual information between B and C is minimum. The value of neighborhood mutual information between B and C is maximum, if B is completely determined by C (Qian et al. 2020).

3.1.3 (Zhou et al. 2018) Conditional neighborhood entropy with granulation monotonicity

Definition 5. Conditional neighborhood entropy with granulation monotonicity.

Given a decision system NDS, $\forall x_i \in U$,$\forall B \subseteq C$ and $U/IND(d) = \{X_1, X_2, ..., X_q\}$, the conditional neighborhood entropy with granulation monotonicity is defined as:

$$\begin{cases} H_{\delta}(X_j/n_B^{\delta}(x_i)) = -\frac{1}{|U|} \log\left(\frac{|n_B^{\delta}(x_i)|^2}{|U|^2} \frac{|X_j \cap n_B^{\delta}(x_i)|}{|n_B^{\delta}(x_i)|}\right), & X_j \cap n_B^{\delta}(x_i) \neq \emptyset \\ H_{\delta}(X_j/n_B^{\delta}(x_i)) = -\frac{1}{|U|} \log\left(\frac{|n_B^{\delta}(x_i)|^2}{|U|^2} \frac{1}{|n_B^{\delta}(x_i)|}\right), & X_j \cap n_B^{\delta}(x_i) = \emptyset \end{cases} \tag{8}$$

3.1.4 Supervision strategy

The traditional neighborhood relation is determined by the distance between two samples and the radius of a single

neighborhood. This method may not be able to express whether the samples with different decision attributes are similar and two samples with different labels will fall into the same neighborhood. In order to solve this problem, the neighborhood relationship based on supervisory decision is proposed in document (Jiang et al. 2020). The neighborhood relationship is introduced as follows:

Definition 6. (Jiang et al. 2020) Supervised neighborhood.

Given a decision system NDS, $\forall x_i \in U, \forall B \subseteq C$,the supervised neighborhood of conditional attribute set B is defined as:

$$n_B^{\delta_I,\delta_O}(x_i) = \{x | x \in U, d(x) = d(x_i) \& \Delta_B(x, x_i) < \delta_I \cup d(x) \neq d(x_i) \& \Delta_B(x, x_i) < \delta_O\} \tag{9}$$

in which the intra class radius δ_I and inter class radius δ_O should satisfy $\delta_O < \delta_I$, which can effectively reduce the impact between different label samples.

3.1.5 Neighborhood rough set based on supervised granulation

Conditional neighborhood entropy with granulation monotonicity takes into account the influence of different decision classes in the calculation of conditional neighborhood entropy. In order to further improve the discriminant performance of neighborhood relations, the supervision strategy is introduced, and the influence relationship between different decision attribute samples is fully considered in the calculation. A neighborhood rough set based on supervised granulation is proposed. The related definitions are introduced as follows:

Definition 7. Conditional neighborhood entropy and neighborhood mutual information based on supervised granulation.

Given a decision system NDS, $\forall x_i \in U, \delta_I, \delta_O \in [0,1], \forall B \subseteq C$,the decision class of the sample is divided into $U/IND(d) = \{X_1, X_2, ..., X_q\}$,the conditional neighborhood entropy of decision attribute d with respect to conditional attribute set B based on supervised granulation is defined as:

$$H_{\delta_I,\delta_O}(d/B) = -\frac{1}{|U|}\sum_{j=1}^{q}\sum_{i=1}^{|U|} H_{\delta_I,\delta_O}(X_j / n_B^{\delta_I,\delta_O}(x_i)) \tag{10}$$

The neighborhood mutual information of decision attribute d with respect to conditional attribute set B based on supervised granulation is defined as:

$$NH_{\delta_I,\delta_O}(d;B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{|[x_i]_d||n_B^{\delta_I,\delta_O}(x_i)|}{|U||[x_i]_d \cap n_B^{\delta_I,\delta_O}(x_i)|} \tag{11}$$

Definition 8. Attribute reduction.

Given a decision system NDS, $\forall x_i \in U, \delta_I, \delta_O \in [0,1], \forall B \subseteq C$,the decision class of the sample is divided into $U/IND(d) = \{X_1, X_2, ..., X_q\}$ when the following two conditions are satisfied, the conditional attribute set B is the attribute reduction of C relative to decision attribute d.

$$\begin{cases} H_{\delta_I,\delta_O}(d/B) \neq H_{\delta_I,\delta_O}(d/B - \{c\}), \ \forall c \in B \\ H_{\delta_I,\delta_O}(d/B) = H_{\delta_I,\delta_O}(d/C) \end{cases} \tag{12}$$

Definition 9. Core.

Given a decision system NDS, $\forall x_i \in U, \delta_I, \delta_O \in [0,1]$ and $\forall c \subseteq C$,the decision class of the sample is divided into $U/IND(d) = \{X_1, X_2, ..., X_q\}$. If $H_{\delta_I,\delta_O}(d/C) \neq H_{\delta_I,\delta_O}(d/C - \{c\})$, then c is the core attribute of the decision system, all core attributes constitute the core of decision system.

Definition 10. Significance.

Given a decision system NDS, supposed $c \in C - B$,then the significance of adding conditional attribute c to conditional attribute set B with respect to decision attribute d is:

$$sig_{out}(c, B, d) = H_{\delta_I,\delta_O}(d/B \cup \{c\}) - H_{\delta_I,\delta_O}(d/B) \tag{13}$$

supposed $c \in B$, then the significance of deleting conditional attribute c to conditional attribute set B with respect to decision attribute d is:

$$sig_{in}(c, B, d) = H_{\delta_I,\delta_O}(d/B) - H_{\delta_I,\delta_O}(d/B - \{c\}) \tag{14}$$

The proof of the monotonicity of conditional neighborhood entropy can be obtained from (Zhou et al. 2018).

3.2 Attribute reduction algorithm of neighborhood rough set based on supervised granulation

In this paper, the algorithm of attribute reduction of neighborhood rough set based on supervised granulation is formed by combining the conditional neighborhood entropy with granulation monotonicity with the supervised strategy, and then considering the interaction between the conditional attributes by introducing the feature complementary relationship. The workflow of the proposed algorithm is shown in Fig. 1.

As is shown in Fig. 1, δ_I, δ_O are intra class radius and inter class radius respectively, calculated by $\delta_I(c_i) = std(c_i)/\lambda, \delta_O(c_i) = a * \delta_I(c_i)$, where $\delta_I(c_i)$ indicates the intra class radius of conditional attributes c_i,$\delta_O(c_i)$ indicates the inter class radius, parameters λ and a control the size of radius. The size of the inter class radius

are determined by the value of a. The value range of a should between [0,1]. The closer it is to 1, the closer the inter class radius is to the intra class radius and the proportion of considering the influence of decision attributes in neighborhood division becomes smaller. No difference in dividing neighborhoods between the five algorithms when the inter class radius is equal to the intra class radius. In order to ensure the intra class radius is greater than the inter class radius. a is taken as 0.5. *sig_ctrl* and *threH* are the significance threshold and complementarity threshold respectively, and the value is a positive number slightly greater than 0. The larger the value of *sig_ctrl*, the less number of the reduction results satisfying the conditions. So, the reduction set will have fewer attribute elements. The larger the value of *threH*, the more similar attributes will be divided into a reduction set. The reduction set will have more attribute elements.

reduction step. In the second step, reduction results are calculated by the mutual information $NH_{\delta_I,\delta_O}(A;B)$ and $NH_{\delta_I,\delta_O}(d;B)$. The attributes which are similar to each other and have little influence on the decision attributes in the reduction set are eliminated. The second step of proposed algorithm corresponds to steps 5,6,7 and 8 in the reduction step. For calculation convenience, the average value R(i) of influence between conditional attribute c_i and other features is calculated. The detailed calculation steps of the algorithm are shown in ALGORITHM.

3.3 Illustrative example

In order to show the calculation process of the proposed algorithm, eight samples are selected from iris dataset from UCI (http://archive.ics.uci.edu/ml/index.php). The specific data are shown in Table 1. Normalization is conducted at

Algorithm: Neighborhood Rough Set Based on Supervised Granulation, NRSBSG

input: Decision system *NDS*, intra class radius δ_I 、inter class radius δ_O 、Mutual information threshold *threH*、significance threshold *sig_ctrl*
output: Reduction set *redSet*

step1 Initialization redSet=$\varnothing$

step2 compute $H_{\delta_I,\delta_O}(d/C)$.

step3 for each $c \in C$

Compute $H_{\delta_I,\delta_O}(d/C-\{c\})$.

Compute $sig(c,C,d)=H_{\delta_I,\delta_O}(d/C)-H_{\delta_I,\delta_O}(d/C-\{c\})$

if $sig(c,C,d)>sig\_ctrl$, $redSet=redSet\cup c$

step4

while $H_{\delta_I,\delta_O}(d/redSet)<H_{\delta_I,\delta_O}(d/C)$

for each $c_i \in C-redSet$ .Compute $sig(c_i,redSet,d)$

Select $c=\arg\max sig(c_i,redSet,d)$, $redSet=redSet\cup c$

step5 for each $c_i,c_j \in redSet$ .Compute $NH_{\delta_I,\delta_O}(c_i;c_j)$

step6 for each $c \in redSet$,Compute $NH_{\delta_I,\delta_O}(c;d)$

step7 for each $j \in redSet$, $j \neq i$,Compute $R(i)=R(i)+NH_{\delta_I,\delta_O}(c_i;c_j)/|redSet|$

step8 for each $c_i \in redSet$, $R(i)-NH_{\delta_I,\delta_O}(c_i;d)>threH$. $redSet=redSet-\{c_i\}$

step9 return redSet

In the first step of proposed algorithm, firstly, the attribute importance is taken as the evaluation criterion to find out the redundant attributes in the whole dataset. In this way, the reduct set can be quickly found in a short time. Then,$H_{\delta_I,\delta_O}(d/redSet)\geq H_{\delta_I,\delta_O}(d/C)$ is taken as the evaluation criterion to ensure the integrity of information. From the point of view of information, the amount of information of the reduced set is not less than that of the original data set by supplementing the reduced set. The first step of proposed algorithm corresponds to steps 1,2,3 and 4 in the

the beginning. The normalization formula is as follows:

$$f(x_i)=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{15}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of samples with the same attribute. The normalized data is shown in Table 2.

Suppose $a=0.5$, $\lambda=2$, neighborhood radius $\delta_I(c_i)$-$=std(c_i)/\lambda$, $\delta_O(c_i)=a*\delta_I(c_i)$ are calculated, $\delta_I(c_1)$-$=0.1930$, $\delta_O(c_1)=0.0965$, $\delta_I(c_2)=0.1485$, $\delta_O(c_2)=0.0743$. According to formula (10), the supervised

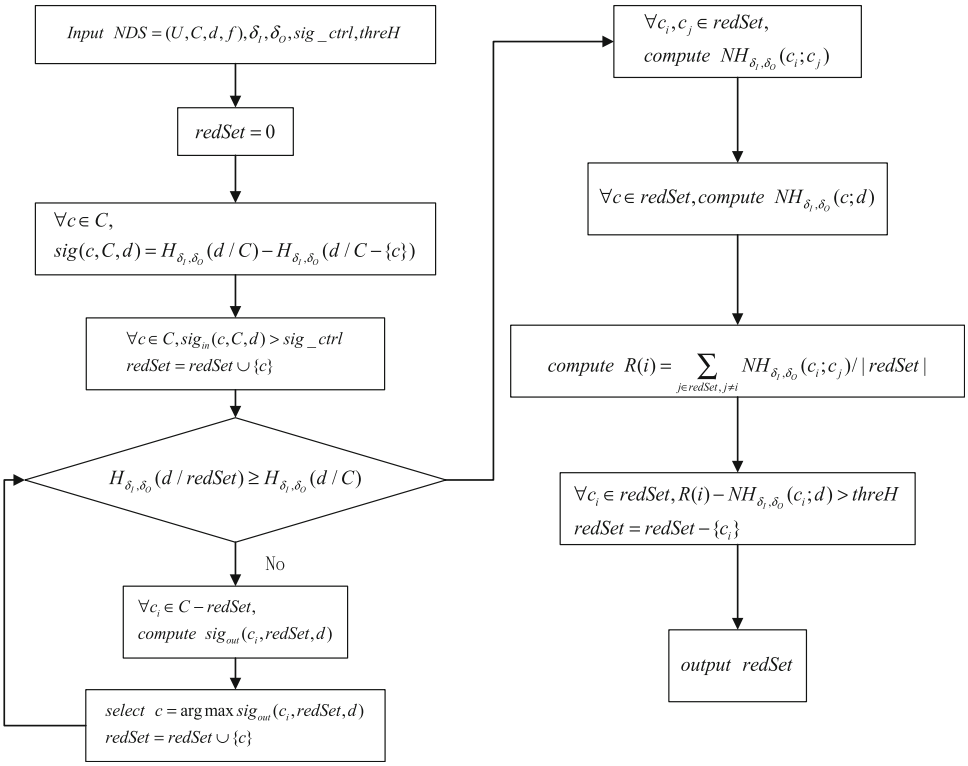


Fig. 1 Attribute reduction process of neighborhood rough set based on supervised

neighborhood of samples in attribute set $B_1 = \{c_1\}$,$B_2 = \{c_2\}$, $B_{12} = \{c_1, c_2\}$ are obtained, as is shown in Table 3.

The decision class of the sample is $U/IND(d) = \{X_1, X_2, X_3\}$, in which $X_1 = \{x_1, x_2\}$, $X_2 = \{x_3, x_4, x_5\}$,$X_3 = \{x_6, x_7, x_8\}$. According to formula (10), the corresponding conditional neighborhood entropy is calculated. Calculating of the neighborhood of samples with respect to attribute set B_1 is detailed as following. The calculation process with respect to B_2 and B_{12} is similar and will not be repeated.

Table 1 Iris data set

Samples	c_1	c_2	c_3	d
x_1	5.1	3.5	1	Setosa
x_2	4.9	3	1	Setosa
x_3	5.8	2.7	1	Virginica
x_4	6.3	3.3	1	Virginica
x_5	7.1	3	2	Virginica
x_6	6.9	3.1	2	Versicolor
x_7	7	3.2	2	Versicolor
x_8	6.4	3.2	1	Versicolor

Table 2 Normalized iris data set

Samples	c_1	c_2	c_3	d
x_1	0.0909	1.0000	1	Setosa
x_2	0	0.3750	1	Setosa
x_3	0.4091	0	1	Virginica
x_4	0.6364	0.7500	1	Virginica
x_5	1.0000	0.3750	2	Virginica
x_6	0.9091	0.5000	2	Versicolor
x_7	0.9545	0.6250	2	Versicolor
x_8	0.6818	0.6250	1	Versicolor

Table 3 Supervised neighborhood

Supervised neighborhood	B_1	B_2	B_{12}
$n_{B_i}^{\delta_l,\delta_o}(x_1)$	$\{x_1, x_2\}$	$\{x_1\}$	$\{x_1\}$
$n_{B_i}^{\delta_l,\delta_o}(x_2)$	$\{x_1, x_2\}$	$\{x_2, x_5\}$	$\{x_2\}$
$n_{B_i}^{\delta_l,\delta_o}(x_3)$	$\{x_3\}$	$\{x_3\}$	$\{x_3\}$
$n_{B_i}^{\delta_l,\delta_o}(x_4)$	$\{x_4, x_8\}$	$\{x_4\}$	$\{x_4\}$
$n_{B_i}^{\delta_l,\delta_o}(x_5)$	$\{x_5, x_6, x_7\}$	$\{x_2, x_5\}$	$\{x_5\}$
$n_{B_i}^{\delta_l,\delta_o}(x_6)$	$\{x_5, x_6, x_7\}$	$\{x_6, x_7, x_8\}$	$\{x_6, x_7\}$
$n_{B_i}^{\delta_l,\delta_o}(x_7)$	$\{x_5, x_6, x_7\}$	$\{x_6, x_7, x_8\}$	$\{x_6, x_7\}$
$n_{B_i}^{\delta_l,\delta_o}(x_8)$	$\{x_4, x_8\}$	$\{x_6, x_7, x_8\}$	$\{x_8\}$

$$X_1 \cap n^{\delta}_{B_1}(x_1) = \{x_1, x_2\},\ X_2 \cap n^{\delta}_{B_1}(x_1) = \{\emptyset\},\ X_3 \cap n^{\delta}_{B_1}(x_1) = \{\emptyset\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_2) = \{x_1, x_2\},\ X_2 \cap n^{\delta}_{B_1}(x_2) = \{\emptyset\},\ X_3 \cap n^{\delta}_{B_1}(x_2) = \{\emptyset\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_3) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_3) = \{x_3\},\ X_3 \cap n^{\delta}_{B_1}(x_3) = \{\emptyset\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_4) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_4) = \{x_4\},\ X_3 \cap n^{\delta}_{B_1}(x_4) = \{x_8\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_5) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_5) = \{x_5\},\ X_3 \cap n^{\delta}_{B_1}(x_5) = \{x_6, x_7\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_6) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_6) = \{x_5\},\ X_3 \cap n^{\delta}_{B_1}(x_6) = \{x_6, x_7\}$$

$$X_1 \cap n^{\delta}_{B_1}(x_7) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_7) = \{x_5\},\ X_3 \cap n^{\delta}_{B_1}(x_7) = \{x_6, x_7\},\ X_1 \cap n^{\delta}_{B_1}(x_8) = \{\emptyset\},\ X_2 \cap n^{\delta}_{B_1}(x_8) = \{x_4\},\ X_3 \cap n^{\delta}_{B_1}(x_8) = \{x_8\}$$

The supervised granulation conditional neighborhood entropy of decision class X_1 with respect to the conditional attribute set B_1 is obtained as follows:

$$H_{\delta_I,\delta_O}(X_1/n^{\delta_I,\delta_O}_{B_1}(x_1)) = -\frac{1}{|U|}\log\left(\frac{\left|n^{\delta}_{B_1}(x_1)\right|^2}{|U|^2}\frac{\left|X_1 \cap n^{\delta}_{B_1}(x_1)\right|}{\left|n^{\delta}_{B_1}(x_1)\right|}\right) = -\frac{1}{8}\log(\frac{2^2}{8^2}\frac{2}{2}) = 0.5$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_2)\right) = -\frac{1}{8}\log\left(\frac{2^2}{8^2}\frac{2}{2}\right) = 0.5$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_3)\right) = -\frac{1}{8}\log\left(\frac{1^2}{8^2}\frac{1}{1}\right) = 0.75$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_4)\right) = -\frac{1}{8}\log\left(\frac{2^2}{8^2}\frac{1}{2}\right) = 0.625$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_5)\right) = -\frac{1}{8}\log\left(\frac{3^2}{8^2}\frac{1}{3}\right) = 0.552$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_6)\right) = -\frac{1}{8}\log\left(\frac{3^2}{8^2}\frac{1}{3}\right) = 0.552$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_7)\right) = -\frac{1}{8}\log\left(\frac{3^2}{8^2}\frac{1}{3}\right) = 0.552$$

$$H_{\delta_I,\delta_O}\left(X_1/n^{\delta_I,\delta_O}_{B_1}(x_8)\right) = -\frac{1}{8}\log\left(\frac{2^2}{8^2}\frac{1}{2}\right) = 0.625$$

The conditional neighborhood entropy of decision class X_1 is:

$$H_{\delta_I,\delta_O}(X_1/B_1) = \sum_{i=1}^{|U|} H_{\delta_I,\delta_O}(X_1/n^{\delta_I,\delta_O}_{B_1}(x_i)) = 4.656$$

The conditional neighborhood entropy of decision class X_2 and X_3 is:

$$H_{\delta_I,\delta_O}(X_2/B_1) = \sum_{i=1}^{|U|} H_{\delta_I,\delta_O}(X_2/n^{\delta_I,\delta_O}_{B_1}(x_i)) = 4.906$$

$$H_{\delta_I,\delta_O}(X_3/B_1) = \sum_{i=1}^{|U|} H_{\delta_I,\delta_O}(X_3/n^{\delta_I,\delta_O}_{B_1}(x_i)) = 4.531$$

Therefore, the conditional neighborhood entropy of the supervised granulation of conditional attribute set B_1 with respect to decision attribute d is:

$$H_{\delta_I,\delta_O}(d/B_1) = \sum_{j=1}^{3} H_{\delta_I,\delta_O}(X_j/B_1) = 14.093$$

Then, the supervised granulation conditional neighborhood entropy of conditional attribute set B_2,B_{12},B_{13},B_{23} and C can be calculated:

$$H_{\delta_I,\delta_O}(d/B_2) = 14.873,\ H_{\delta_I,\delta_O}(d/B_3) = 10.238$$

$$H_{\delta_I,\delta_O}(d/B_{12}) = 17,\ \ H_{\delta_I,\delta_O}(d/B_{13}) = 14.092$$

$$H_{\delta_I,\delta_O}(d/B_{23}) = 17,\ \ H_{\delta_I,\delta_O}(d/C) = 17$$

The significance of attributes c_1,c_2 and c_3 can be calculated by formula (15):

$$\begin{aligned} sig_{in}(c_1, C, d) &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/C - \{c_1\}) \\ &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/B_{23}) = 0 \\ sig_{in}(c_2, C, d) &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/C - \{c_2\}) \\ &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/B_{13}) = 2.908 \\ sig_{in}(c_3, C, d) &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/C - \{c_3\}) \\ &= H_{\delta_I,\delta_O}(d/C) - H_{\delta_I,\delta_O}(d/B_{12}) = 0 \end{aligned}$$

According to the above calculation process, the significance of attribute c_1 and c_3 with respect to d is 0, which is redundant thus could be deleted.

4 Experimental work

Experiments on 10 open UCI data sets are carried out here. The data sets are shown in Table 4.

The effectiveness of the proposed algorithm is verified from two aspects: reduction rate and accuracy of classification. The reduction rate formula is shown in formula (16):

$$Rate = \frac{|C| - |redSet|}{|C|} \tag{16}$$

where $|C|$ indicates the number of original attributes,$|redSet|$ indicates the number of attributes about reduction set.

The accuracy of classification is as following:

Table 4 Dataset used in the experiment

U	Data sets	Samples	Features	Classes
1	Lymphography	114	18	4
2	Wine	178	13	3
3	Ionosphere	351	34	2
4	Breast	277	9	2
5	Zoo	101	16	7
6	Fertility	100	9	2
7	Conn	208	60	2
8	Arrhythmia	452	280	13
9	Parkinson's disease	756	753	2
10	Swarm behavior	2001	2400	2

$$Accuracy = \frac{|U_a|}{|U|} * 100\% \tag{17}$$

where $|U_a|$ indicates the number of samples correctly classified, $|U|$ indicates the total number of samples.

To reduce the adverse effects caused by inconsistent sample data, the preprocessed data are limited to [0,1] by normalization. The normalization formula is shown in (18):

$$f(x_i) = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{18}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of samples under the same attribute respectively.

Comparative experiments are carried out with the same and different neighborhood parameters respectively. The experiment is carried out on an Ali-cloud server computer: using X86 computing architecture, 16vCPU: AMD EPYC™ ROME 7H12, 64 GB running memory. MATLAB R2018b is used as the development tool.

4.1 Reduction results with same parameters

To verify the performance of NRSBSG algorithm, we compared NRSBSG with FARNeMF (Hu 2008), NRSBCE (Zou et al. 2021), IFSANRSR (Zou et al.2019a) and SNBAR (Jiang et al.2020) on the selected 10 data sets. The parameter a in NRSBSG algorithm and SNBAR algorithm is set to be 0.5. And the intra class radius δ_I in NRSBSG is 0.1 which is the same as that in the other four algorithms. In IFSANRSR, the number of iterations is 100, and the number of artificial fish is set to be half of the number of conditional attributes, the maximum field of vision is half of the number of artificial fish, and the maximum movement step is 1 less than the maximum field of view. The reduction results and reduction rate of the five algorithms are shown in Table 5. The accuracy of classification and running time are shown in Table 6.

As could be seen from Table 5, the reduction effect of NRSBSG is similar to the other four algorithms in most data sets when $\delta = 0.1$. In high-dimensional data sets including Arrhythmia, Parkinson's disease and Swarm behavior, the average reduction rate of the proposed algorithm is above 0.95. Compared with NRSBCE and IFSANRSR algorithms, it is slightly better than the two algorithms.

As is shown in Table 6, when $\delta = 0.1$, the accuracy of classification of the proposed NRSBSG algorithm is 0.9504 and 0.9388 on Wine and Ionosphere data sets which is slightly lower than that of NRSBCE algorithm. It has higher accuracy of classification than the other four algorithms on other datasets. The standard deviation of the accuracy of the five algorithms was about 0.04.

As could be seen from Table 6, the FARNeMF algorithm has the shortest running time of the five on the selected ten data sets. The running time of NRSBSG algorithm is shorter than that of IFSANRSR algorithm. The time complexity of NRSBSG algorithm is $O(|U|^2 \cdot |C|^3)$. The time complexity of NRSBCE algorithm is $O(|U|^2 \cdot |C|^2)$(Zou et al. 2021). The time complexity of SNBAR algorithm is $O(|U|^2 \cdot |C|^2)$(Jiang et al.2020). The time complexity of FARNeMF algorithm is $O(|U|log|U| \cdot |C|^2)$(Hu 2008). The time complexity of IFSANRSR algorithm is $O(|Iterations| \cdot |Fish|^2 \cdot |try\_number| \cdot |U|^2 \cdot |C|)$(Zou et al.2019a).

From the above, when $\delta = 0.1$, NRSBSG algorithm has made a progress in the aspect of accuracy of classification at the cost of running time. At the same time, NRSBSG algorithm has also perform well in the reduction rate of high-dimensional data sets.

4.2 Reduction rate evaluation under different parameters

As we know that different value of radius parameter results in different granularity of knowledge. In order to see the change of reduction rate of the four algorithms with different value of radii parameter, 20 radius parameters are selected. The radius range is from 0.05 to 1.0, and the change step is 0.05. The radius parameters δ_I is taken as the horizontal axis, and the value of reduction rate as the vertical axis. Figure 2 shows the comparison result of reduction rate.

It can be seen from Fig. 2, with the increase of radii, the overall trend of reduction rate is gradually decreasing. Compared with the other three algorithms, the maximum reduction rates of NRSBSG algorithm on Breast, Fertility, Ionosphere and Wine data sets are 0.5556, 0.5556, 0.9697 and 0.8462 respectively. While the maximum reduction rates of other three algorithms are 0.5556, 0.5556, 0.8788 and 0.8462. Which are lower or same than that of NRSBSG algorithm. In high-dimensional data sets including

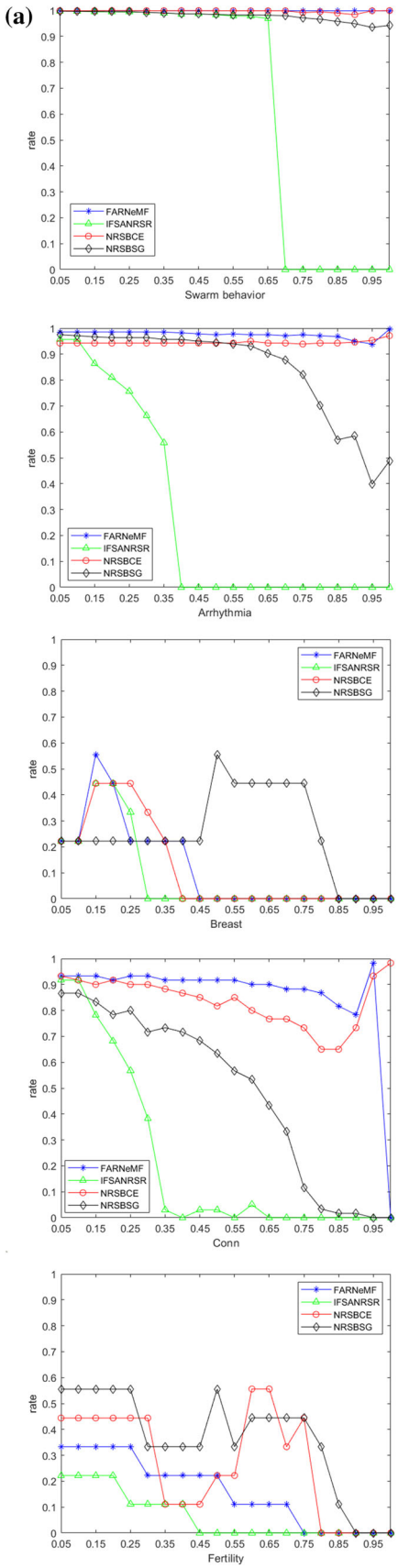
Table 5 Reduction results and reduction rate at δ=0.1

Data sets	NRSBSG		NRSBCE		FARNeMF	
	Reduct	Rate	Reduct	Rate	Reduct	Rate
Lymphography	1,2,7,10,12,13,14,15	0.5556	13	0.9444	2,13,14,15,16,18	0.6667
Wine	1,12,13	0.7692	3,5,6,8,12	0.6154	9,13	0.8462
Ionosphere	2,3,4,6,27,30	0.8182	2,11,17,19	0.8788	2,13,17,23,30	0.8485
Breast	3,4,5,6,7,8,9	0.2222	1,3,4,6,7,8,9	0.2222	1,2,3,4,6,7,8	0.2222
Zoo	1,2,3,8,13	0.6875	6,7	0.875	4,6,8,12,13	0.6875
Fertility	2,3,4,9	0.5556	1,3,4,7,9	0.4444	1,2,3,4,7,9	0.3333
Conn	1,7,16,30,40,43,54,55	0.8667	1,2,4,45,56	0.9167	11,20,35,44	0.9333
Arrhythmia	7,64,78,152,162, 173,250,256	0.9713	23,33,42,43,45,47, 48,57,66, 69,71,76,82,94, 123,143	0.9427	40,64,89,132	0.9857
Parkinson's disease	5,48,63,94,96, 97, 109,170,405	0.988	23,24,36,46,84,347, 348,573, 715,717,722,728, 740,751	0.9814	65,133,373,	0.996
Swarm behavior	214,233,257,507, 1033,1547, 2230,2300,	0.9967	2239,2240,	0.9992	493,901,	0.9992

Data sets	IFSANRSR		SNBAR	
	Reduct	Rate	Reduct	Rate
Lymphography	2,12,13,14,15,16,18	0.6111	2,11,12,13,14,15,18	0.6111
Wine	1,3,9,11,13	0.6154	1,5,7,13	0.692
Ionosphere	4,6,7,13,14,19,23	0.7879	4,5,13,15,23	0.8485
Breast	1,2,3,4,6,7,8	0.2222	1,3,4,5,6,7,8,9	0.1111
Zoo	2,3,4,6,8,13	0.625	4,6,8,12,13	0.6875
Fertility	1,2,3,5,6,7,8	0.2222	1,2,3,5,9	0.4444
Conn	23,28,33,36,45	0.9167	10,14,19,49	0.9333
Arrhythmia	1,7,30,56,67,78,85, 156,167,176,242,251	0.957	3,10,14,109,152,167,250	0.9749
Parkinson's disease	29,41,63,122,124, 229,256, 347,476,578,686,710	0.9841	271,390	0.9973
Swarm behavior	518,724,748,1202,1335,1377, 2137,2150	0.9967	201,933	0.9222

Table 6 Accuracy of classification and running time at δ=0.1

Data sets	NRSBSG		NRSBCE		FARNeMF		IFSANRSR		SNBAR	
	Acc	Time (s)	Acc	Time (s)	Acc	Time (s)	Acc	Time (s)	Acc	Time (s)
Lymphography	0.7838	3.39	0.7761	0.13	0.7704	0.01	0.772	1.3	0.768	0.3
Wine	0.9504	2.05	0.9634	0.11	0.9319	0.03	0.9334	0.9	0.9237	0.4
Ionosphere	0.9388	45.1	0.9563	0.89	0.8983	0.13	0.9003	23.7	0.9173	1.1
Breast	0.737	2.68	0.6863	0.16	0.6898	0.05	0.6793	1.3	0.7032	0.8
Zoo	0.7989	2.18	0.7589	0.02	0.7553	0.01	0.7533	0.7	0.7687	0.1
Fertility	0.8878	1.16	0.8533	0.05	0.8319	0.01	0.8598	0.2	0.8680	0.08
Conn	0.7488	7.2	0.7009	1.2	0.702	0.2	0.6798	8.3	0.6931	2.6
Arrhythmia	0.657	697.8	0.6554	191.5	0.6271	8.4	0.6331	1022.7	0.6277	397.3
Parkinson's disease	0.7659	1233.2	0.681	598.1	0.6773	51.8	0.676	1795.4	0.716	1295.4
Swarm behavior	0.6341	4175.6	0.621	2031.1	0.5802	867.5	0.5637	4938.2	0.5602	4537.2



◀**Fig. 2** Reduction rates of different reduction algorithms in 10 data sets

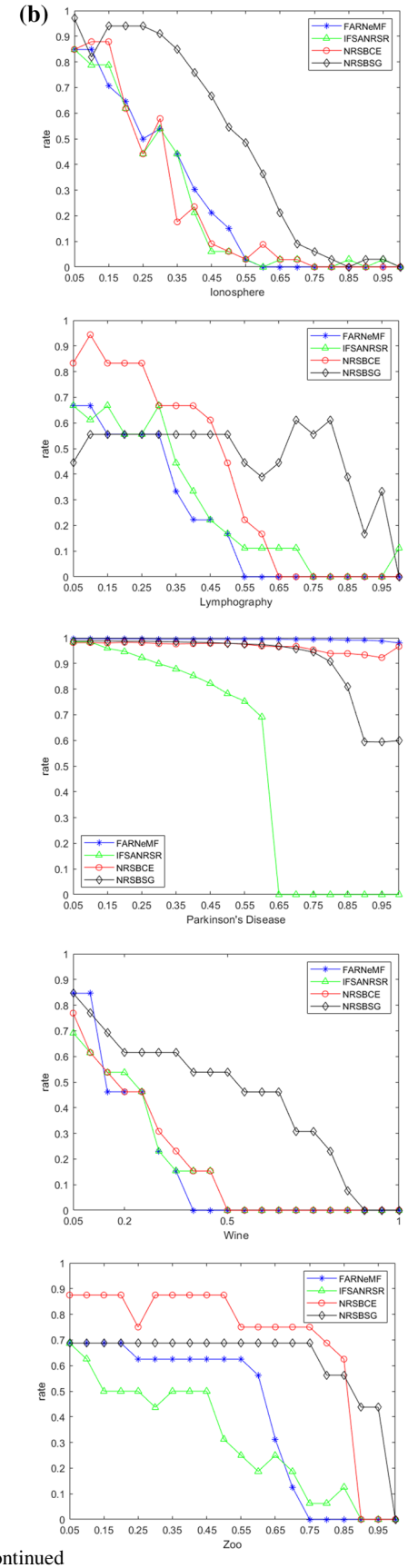
Parkinson's Disease, Arrhythmia and Swarm behavior, the maximum reduction rates of NRSBSG algorithm are close to the other three algorithms. The maximum reduction rate of NRSBSG algorithm on Conn data set is slightly lower than the other three algorithms. Although the maximum reduction rate of NRSBSG algorithm on Zoo data set is lower than NRSBCE algorithm, but it is higher than FARNeMF algorithm and IFSANRSR algorithm. Therefore, in most cases, the NRSBSG algorithm perform better than the other three algorithms in the reduction rate.

4.3 Evaluation of classification accuracy under different parameters

The accuracy of classification is believed more important for algorithm evaluation criteria. In this paper, the support vector machine is used as a classification tool, and the ten-fold cross validation method is used to verify the accuracy of classification of the proposed NRSBSG algorithm. Based on 10 data sets, 20 different neighborhood radius parameters in the reduction experiment are used. The radius values range from 0.05 to 1.0. The radius parameter is taken as the horizontal axis and the value of classification accuracy is taken as the vertical axis. The comparison results are shown in Fig. 3.

According to Fig. 3, the best accuracy of classification can be obtained by using different neighborhood radius parameters in different algorithms. On Breast, Fertility, Ionosphere, Wine, Zoo, Conn and Lymphography data sets, the maximum accuracy of classification of NRSBSG algorithm are 0.7478, 0.8964, 0.9687, 0.9850, 0.8253, 0.7648 and 0.7952 respectively which is higher than that of the other three algorithms. In high-dimensional data sets including Arrhythmia, Parkinson's disease and Swarm behavior, the maximum accuracy of classification of NRSBSG algorithm are 0.6988, 0.8083 and 0.6608 while the maximum accuracy of classification of other three algorithms are 0.6758, 0.7171 and 0.6454. Therefore, the accuracy of classification of NRSBSG algorithm is better than the other three algorithms in most cases.

As could also be seen from Figs. 2 and 3 that the NRSBSG algorithm can obtain higher accuracy of classification and efficiently delete the redundant features from the original data. Compared with the other algorithms, the proposed NRSBSG algorithm mainly carries out attribute reduction from the perspective of information view. Firstly, in the process of dividing the neighborhood, the possible influence among different classes in decision attribute is



◀**Fig. 2** continued

considered by introducing the supervised strategy. Secondly, in the process of entropy calculation, the possible influence among different classes in decision attribute is considered by using conditional neighborhood entropy with granulation monotonicity. Thirdly, the influence of conditional attributes on each other is also considered in attribute reduction through mutual information. Thus, the possible influence of attributes is fully considered from three aspects in the proposed NRSBSG algorithm. So that NRSBSG algorithm is superior to the other algorithms in classification accuracy. At the same time, with the increase of the categories of decision attributes, the influence between different decision attributes will be more and the influence of decision attributes needs to be considered more. NRSBSG algorithm will perform better than other algorithms at this occasion. However, due to the high time complexity, the running time of NRSBSG algorithm will increase. How to decrease the time complexity of the proposed algorithm and how to determine the appropriate neighborhood radius parameters to achieve high accuracy and reduction rate at the same time should be focused in the future.

5 Application

In order to comprehensively analyze the influencing factors of fatigue life of titanium alloy welded joints and the coupling relationship between the influencing factors, an analysis model of fatigue influencing factors of titanium alloy welded joints is established. The NRSBSG algorithm is applied in attribute reduction of the fatigue decision system of titanium alloy welded joints. The key influencing factors of fatigue life of titanium alloy welded joints are determined. At the same time, the coupling relationship between the influencing factors is analyzed based on mutual information theory.

5.1 Fatigue decision system of titanium alloy welded joints

Fatigue test data of titanium alloy welded joints are collected as introduced in reference (Iwata and Matsuoka 2004) and the fatigue database is established as shown in Table 7. There are 43 samples in the database. The database is used as the experimental basis to analyze the influencing factors of fatigue life of titanium welded joints. In total, there are 6 attributes in the established database. Among which, the character of fatigue life is decision attribute, and the other 5 characters including Nominal Stress Range, Thickness, Stress concentration and Equivalent Structural Stress Range are conditional attributes. In the established fatigue database, except the value of Joint

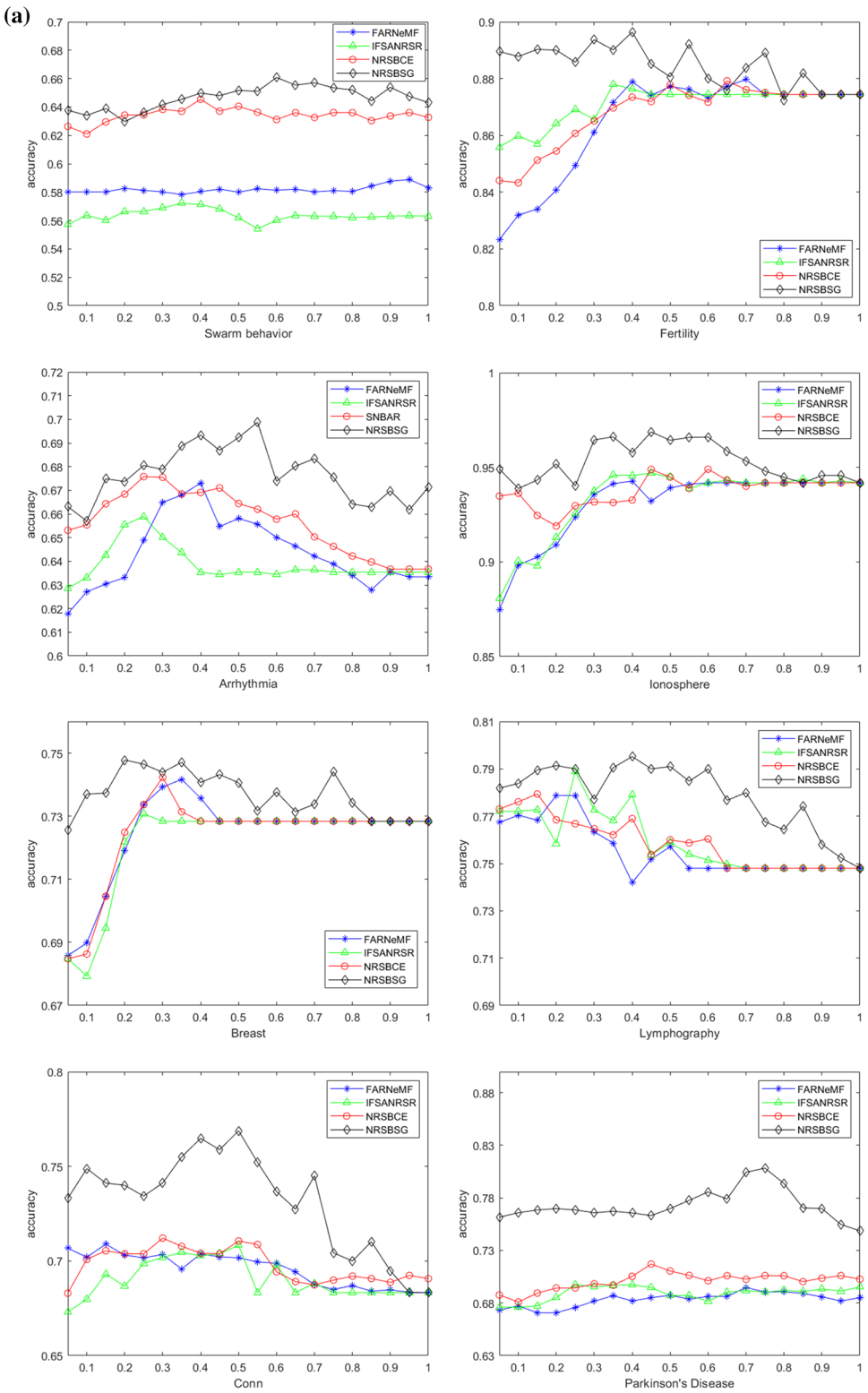


Fig. 3 Accuracy of classification of different reduction algorithms in 10 data sets

Type is discrete, the values of other attributes are all continuous.

Let S_1={Nominal Stress Range (c_1),Thickness (c_2),Stress Concentration Factor (c_3),Joint Type (c_4)} and S_2={Equivalent Structural Stress Range (c_1),Thickness (c_2),Stress Concentration Factor (c_3),Joint Type (c_4)}, by taking the "fatigue life" character as decision attribute, two

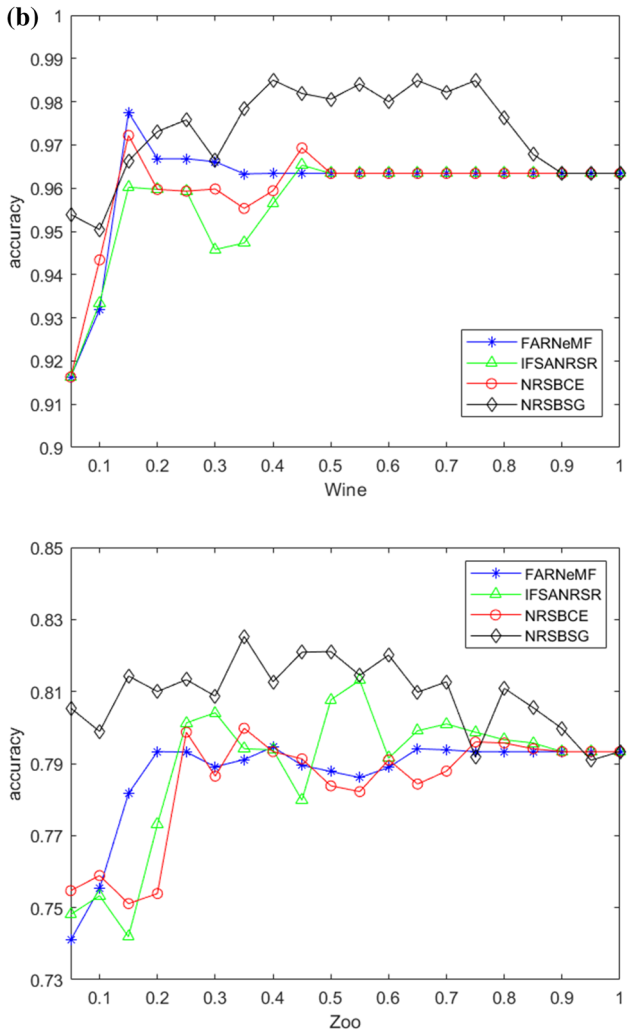


Fig. 3 continued

Table 8 Nominal stress decision system

U	S_1				d
	C_1	C_2	C_3	C_4	
x_1	152	2	1	LB	1,734,430
x_2	166.95	2	1.3789	LT	675,006
x_3	190.8	2	1.3789	LT	568,020
x_4	221.4	2	1	CB	129,860
x_5	230	2	1	CB	105,510
x_6	149.5	2	1.3789	CT	4,213,944
x_7	218.5	2	1.3789	CT	188,779
x_8	92.65	2	2.7535	LL	1,116,913
x_9	102	10	1.37285	LT	2,233,310
x_{10}	204	10	1	CB	7,946,640

fatigue decision systems of titanium alloy welded joints are established. As is shown in Table 8 and Table 9.

The proposed reduction algorithm is applied in the fatigue decision system. Several experiments show that when the radius parameter is 1.2, the accuracy of classification is better. So, the radius parameter is set as 1.2. The other parameters are set as $a = 0.5$, $sig\_ctrl = 0.01$, $threH = 0.3$. The reduction result of the nominal stress decision system is {Nominal Stress Range, Stress Concentration Factor, Joint type}; The reduction result of the equivalent structural stress decision system is {Equivalent structural stress range, Stress concentration factor, Joint type}.

The weight of the conditional attribute is calculated by formula (18) proposed in (Qian et al.2020). Larger value of weight indicates more influence on the fatigue life of the

Table 7 Data set of factors affecting fatigue life of Titanium alloy welded joints

Nominal stress range σ_{NS}/ MPa	Joint type	Thickness t/ mm	Stress concentration factor SCF	Equivalent structural stress range σ_{ES}/ MPa	Fatigue life
152	LB	2	1	170.6142	1,734,430
166.95	LT	2	1.3789	258.3990	675,006
190.8	LT	2	1.3789	295.3132	568,020
221.4	CB	2	1	248.5131	129,860
230	CB	2	1	258.1663	105,510
149.5	CT	2	1.3789	231.3906	4,213,944
218.5	CT	2	1.3789	338.1862	188,779
92.65	LL	2	2.7535	286.3533	1,116,913
102	LT	10	1.37285	205.5370	2,233,310
204	CB	10	1	299.4311	7,946,640

Table 9 Equivalent structural stress decision system

U	S_2				d
	c_1	c_2	c_3	c_4	
x_1	170.6142	2	1	LB	1,734,430
x_2	258.3990	2	1.3789	LT	675,006
x_3	295.3132	2	1.3789	LT	568,020
x_4	248.5131	2	1	CB	129,860
x_5	258.1663	2	1	CB	105,510
x_6	231.3906	2	1.3789	CT	4,213,944
x_7	338.1862	2	1.3789	CT	188,779
x_8	286.3533	2	2.7535	LL	1,116,913
x_9	205.5370	10	1.37285	LT	2,233,310
x_{10}	299.4311	10	1	CB	7,946,640

welded joints. Weight of conditional attributes of the two fatigue decision systems are shown in Table 10 and 11.

$$\omega_j = \frac{NH_{\delta_I,\delta_O}(d;c_j)}{\sum_{i=1}^{|redSet|} NH_{\delta_I,\delta_O}(d;c_i)} \tag{18}$$

Mutual information between different attributes is computed to measure the different influence degree. The smaller of the value of mutual information, the better independent between the two attributes would be. The mutual information between the conditional attributes is computed as shown in Table 12.

According to Table 10, in the nominal stress decision system, the joint type has the largest influence on the fatigue life of titanium alloy welded joint, which is 0.4123. The nominal stress range is 0.3133, and the stress concentration factor is 0.2744. According to Table 11, in the equivalent structural stress decision system, the influence of equivalent structural stress range is the largest, which is 0.4563. The stress concentration factor and joint type are 0.2173 and 0.3264 respectively. Compared with the nominal stress decision system, the influence of stress concentration factor and joint type in the equivalent structure stress decision system are smaller.

According to Table 12, the influence weight of plate thickness on the stress concentration factor is 0.301. The influence weight of joint type on the stress concentration factor is 0.4071. That indicates the stress concentration factor is more easily affected by the joint type than the plate thickness.

5.2 Design and implementation of the influencing factors analysis system

In this work, influencing factors analysis system of welded joint fatigue life is designed and developed by using the proposed NRSBSG algorithm. The development platform of the system is MATLAB 2018b. The attribute reduction algorithms involved in the system are all written in MATLAB. After requirement analysis, the influencing factors analysis system should include 3 modules, including data selection, coupling relationship analysis and quantitative evaluation of influencing factors.

The data selection function enables users to browse and select the fatigue life test data set of welded joints or other data sets for analysis and comparison. The data selection function interface is shown in Fig. 4.

Coupling relationship analysis function enables users to analyze and evaluate the influence degree of different influencing factors in data set, and can see the histogram of

Table 10 The percentage of the influence of each influencing factor on the fatigue life (nominal stress)

	Nominal stress range	Thickness	Stress concentration factor	Joint type
Fatigue life	0.3133	0	0.2744	0.4123

Table 11 The percentage of the influence of each influencing factor on the fatigue life (equivalent structural stress)

	Equivalent structural stress range	Thickness	Stress concentration factor	Joint type
Fatigue life	0.4563	0	0.2173	0.3264

Table 12 The degree of influence between each influencing factor

	Thickness	Stress concentration factor	Joint type
Thickness	0	0.301	0.0014
stress concentration factor	0.301	0	0.4071
Joint type	0.0014	0.4071	0

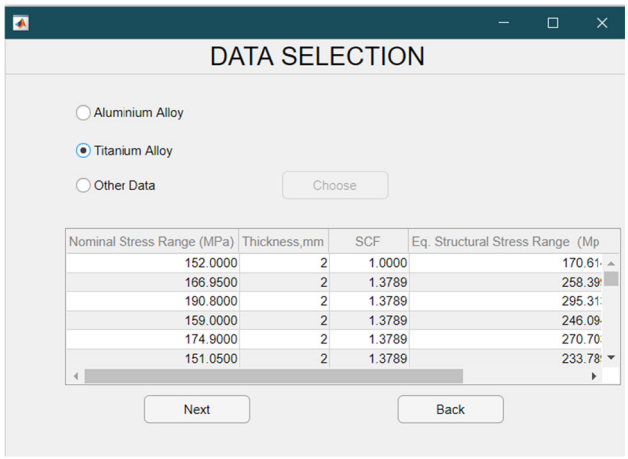


Fig. 4 Data selection interface

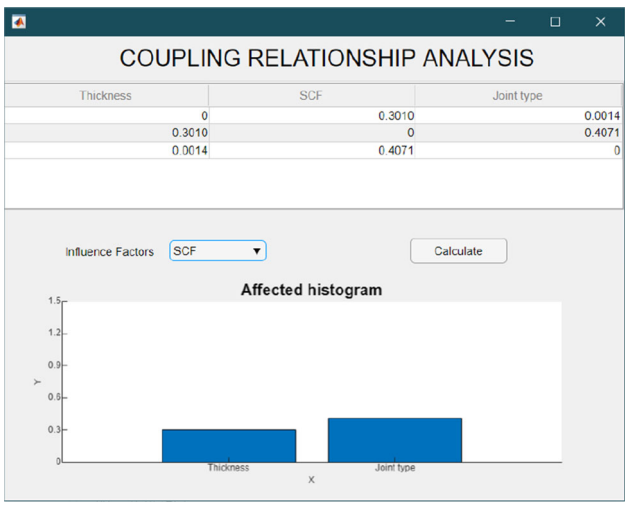


Fig. 5 Coupling relationship analysis interface

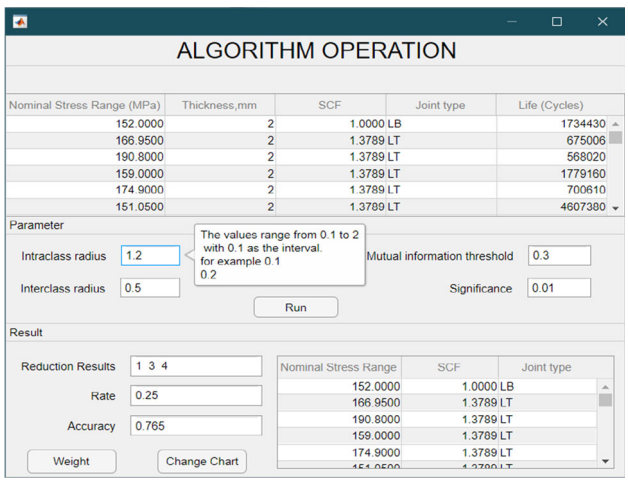


Fig. 6 Algorithm running result interface

influence degree of other factors on the influencing factors by selecting specific influencing factors. As shown in Fig. 5.

In the quantitative evaluation of influencing factors, the reduction result of the data set is calculated and the influence ratio of the reduction result on the fatigue life is obtained. The interface of algorithm operation results is shown in Fig. 6, and the interface of quantitative evaluation of influencing factors is shown in Fig. 7.

6 Conclusion

This paper proposed a neighborhood rough set attribute reduction algorithm NRSBSG, which combines the supervised strategy, three-layer granulation, feature complementary relationship with conditional neighborhood entropy. Experiments on 10 UCI data sets showed that the algorithm improves the reduction rate and accuracy compared with the other three algorithms.

On this basis, a coupling analysis model and quantitative evaluation model of influencing factors of titanium alloy welded joints was proposed. The key factors of titanium alloy welding joints were determined in two stress decision making systems. Through comparative analysis, the effect of equivalent structural stress range on fatigue life of welded joints is greater. It is shown that the equivalent structure stress range can predict the fatigue life of welded joints more accurately. In the coupling relationship of influencing factors, the Stress Concentration Factor is most importantly affected by joint type, which indicates that the influence of joint type should be more considered important when calculating the Stress Concentration Factor.

The design and development of fatigue analysis system in this work provides technical support for fatigue analysis and design of titanium alloy welded joints in industrial

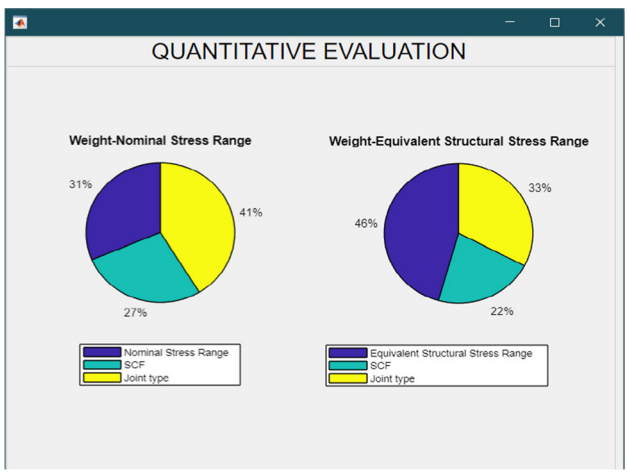


Fig. 7 Quantitative evaluation interface

production. It reduces the labor intensity of designers, improves the working efficiency, and has great practical value and applicability.

The following topics need to be further studied.

1. The inter class radius of supervision strategy is determined by experts. It can be dynamically and automatedly updated by intelligent algorithms in the future.
2. This paper carries on the analysis experiment on static complete data set. In the future, we can analyze on incomplete data set and dynamic data set.
3. In this paper, the coupling relationship analysis and quantitative evaluation of influencing factors are only from aspects of analysis of the fatigue test data. In the future, it can be comprehensively considered and analyzed from the experimental combined with data analyzing aspect.

Funding This work was supported in part by the National Science Foundation of China under Grant 52005071 and 51875072, in part by the Liaoning Provincial Educational Department Project under Grant JDL2020004, and in part by the Liaoning Province "Xingliao Talent Program" project for young top talents under Grant XLYC1807112.

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interests The authors have not disclosed any competing interests.

References

Chen TQ, Liu JH, Zhu F, Wang YH, Liu J, Chen J (2018) A novel multi-radius neighborhood rough set weighted feature extraction method for remote sensing image classification. Geomat Inf Sci Wuhan Univ 43(02):311–317. https://doi.org/10.13203/j.whugis20150290

Chen YC, Li O, Sun Y (2018) Attribute reduction based on clustering discretization and variable precision neighborhood entropy. Control Decis 33(08):1407–1414. https://doi.org/10.13195/j.kzyjc.2017.0512

Chen YM, Qin N, Li W, Xu FF (2018c) Granule structures, distances and measures in neighborhood systems. Knowl-Based Syst 165:268–281. https://doi.org/10.1016/j.knosys.2018.11.032

Chu XL, Sun BZ, Li X, Han KY, Wu JQ, Zhang Y, Huang QC (2020) Neighborhood rough set-based three-way clustering considering attribute correlations: an approach to classification of potential gout groups. Inf Sci 535:28–41. https://doi.org/10.1016/j.ins.2020.05.039

Deng ZX, Zheng ZL, Deng DY (2021) F-neighbor rough sets and its reduction. Acta Automatica Sinica 47(03):695–705. https://doi.org/10.16383/j.aas.c180556

Hu QH, Yu DR, Xie ZX (2006) Neighborhood classifiers. Expert Syst Appl 34(2):866–876. https://doi.org/10.1016/j.eswa.2006.10.043

Hu QH, Yu DR, Xie ZX (2008) Numerical attribute reduction based on neighborhood granulation and rough approximation. J Softw 19(3):640–649

Hu QH, Yu DR (2009) Neighborhood entropy. IEEE Int Conf Mach Learn Cybern. https://doi.org/10.1109/ICMLC.2009.5212245

Hu QH, Zhang L, Zhang D, Pan W, An S, Pedrycz W (2011) Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2011.01.023

Iwata T, Matsuoka K (2004) Fatigue strength of CP grade 2 titanium fillet welded joint for ship structure. Weld World 48(7–8):40–47. https://doi.org/10.1007/BF03266442

Jiang ZH, Wang YB, Xu G, Yang XB, Wang PX (2019) Multi-scale based accelerator for attribute reduction. Comput Sci 46(12):250–256

Jiang ZH, Liu KY, Yang XB, Yu HL, Fujita H, Qian YH (2020) Accelerator for supervised neighborhood based attribute reduction. Int J Approx Reason 119:122–150. https://doi.org/10.1016/j.ijar.2019.12.013

Liu YL, Zou L, Sun YB, Yang XH, Martínez RA (2017) Evaluation model of aluminum alloy welded joint low-cycle fatigue data based on information entropy. Entropy. https://doi.org/10.3390/e19010037

Liang JY, Qu KS, Xu ZB (2001) Reduction of attribute in information systems. Syst Eng-Theory Pract 21(12):76–80. https://doi.org/10.3321/j.issn:1000-6788.2001.12.014

Li L-J, Li M-Z, Mi J-S, Xie B (2020) Dynamic granularity selection based on local weighted accuracy and local likelihood ratio. Applied Soft Computing Journal 89:106087. https://doi.org/10.1016/j.asoc.2020.106087

Miao DQ (1997) Rough set theory and its application in machine learning. Dissertation, Institute of automation, Chinese Academy of Sciences, 1997.

Mou E, Zhang XY, Yao YS, Deng Q (2020) Class-specific attribute reduct and its heuristic algorithm of neighborhood approximation condition-entropy. Comput Eng Appl 56(24):175–180

Mu TP, Zhang XY, Mo ZW (2019) Double-granule conditional-entropies based on three-level granular structures. Entropy 21(7):657. https://doi.org/10.3390/E21070657

Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11(5):341–356. https://doi.org/10.1007/BF01001956

Qian WB, Long XD, Wang YL, Xie YH (2020) Multi-label feature selection based on label distribution and feature complementarity. Appl Soft Comput J 90:106167. https://doi.org/10.1016/j.asoc.2020.106167

Rao XS, Song JJ, Yang XB, Yu HL, Wang PX (2020) Acceleration strategy for attribute reduction based on pseudo-label neighborhood rough set. Comput Eng Des 41(11):3087–93. https://doi.org/10.16208/j.issn1000-7024.2020.11.014

Sang BB, Chen HM, Yang L, Li TR, Xu WH, Luo C (2021) Feature selection for dynamic interval-valued ordered data based on fuzzy dominance neighborhood rough set. Knowl-Based Syst. https://doi.org/10.1016/J.KNOSYS.2021.107223

Shannon CE (1948) A mathematical theory of communication. Wiley, London

Shen XF, Xie J, Liu HF, Xu XY (2013) Improved incremental attribute reduction algorithm based on relative positive region. J Guangxi Normal Univ (Natural Science Edition) 31(03):45–50. https://doi.org/10.16088/j.issn.1001-6600.2013.03.011

Shu WH, Qian WB, Xie YH (2020) Incremental feature selection for dynamic hybrid data using neighborhood rough set. Knowl-Based Syst 194:105516. https://doi.org/10.1016/j.knosys.2020.105516

Sinha AK, Namdev N (2020) Feature selection and pattern recognition for different types of skin disease in human body using the

rough set method. Network Model Anal Health Inf Bioinf 9(1):1–11. https://doi.org/10.1007/s13721-020-00232-z
Singh M, Pamula R (2020) An outlier detection approach in large-scale data stream using rough set. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04421-4
Singh S, Shreevastava S, Som T, Somani G (2020) A fuzzy similarity-based rough set approach for attribute selection in set-valued information systems. Soft Comput 24(6):4675–4691. https://doi.org/10.1007/s00500-019-04228-4
Sun L, Wang LY, Ding WP, Qian YH, Xu JC (2020) Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems. Knowl-Based Syst 192:105373. https://doi.org/10.1016/j.knosys.2019.105373
Tan AH, Wu WZ, Shi SW, Zhao SM (2019) Granulation selection and decision making with multi granulation rough set over two universes. Int J Mach Learn Cybern 10(9):2501–2513. https://doi.org/10.1007/s13042-018-0885-7
Tsang ECC, Fan BJ, Chen DG, Xu WH, Li WT (2020) Multi-level cognitive concept learning method oriented to data sets with fuzziness: a perspective from features. Soft Comput 24(5):3753–3770. https://doi.org/10.1007/s00500-019-04144-7
Wang C, Ou F (2008) An attribute reduction algorithm in rough set theory based on information entropy. In: IEEE international symposium on computational intelligence and design, pp. 3-6
Wan JH, Chen HM, Yuan Z, Li TR, Yang XL, Sang BB (2021) A novel hybrid feature selection method considering feature interaction in neighborhood rough set. Knowl-Based Syst 227(6):107167. https://doi.org/10.1016/J.KNOSYS.2021.107167
Xu YY, Li Y, Wang YJ, Wang C, Zhang GJ (2020) Integrated decision-making method for power transformer fault diagnosis via rough set and DS evidence theories. IET Gener Transm Distrib 14(24):5774–5781. https://doi.org/10.1049/iet-gtd.2020.0552
Yang XB, Liang SC, Yu HL, Gao S, Qian YH (2019) Pseudo-label neighborhood rough set: measures and attribute reductions. Int J Approx Reason 105:112–129. https://doi.org/10.1016/j.ijar.2018.11.010
Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90(2):111–127. https://doi.org/10.1016/S01650114(97)00077-8
Zhan JM, Xu WH (2018) Two types of coverings based multigranulation rough fuzzy sets and applications to decision making. Artif Intell Rev Int Sci Eng J 53(3):167–198. https://doi.org/10.1007/s10462-018-9649-8
Zhang XY, Miao DQ (2014) Reduction target structure-based hierarchical attribute reduction for two-category decision-theoretic rough sets. Inf Sci 277:755–776. https://doi.org/10.1016/j.ins.2014.02.160
Zhang HD, Zhan JM, He YP (2019a) Multi-granulation hesitant fuzzy rough sets and corresponding applications. Soft Comput 23(24):13085–13103. https://doi.org/10.1007/s00500-019-03853-3
Zhang J, Zhang XY, Xu WH, Wu YX (2019b) Local multigranulation decision-theoretic rough set in ordered information systems. Soft Comput 23(24):13247–13261. https://doi.org/10.1007/s00500-019-03868-w
Zhao H, Qin KY (2014) Mixed feature selection in incomplete decision table. Knowledge-Based Systems 57(Feb.):181–190. https://doi.org/10.1016/j.knosys.2013.12.018.
Zhao JY, Zhang ZL, Han CZ, Zhou ZF (2015) Complement information entropy for uncertainty measure in fuzzy rough set and its applications. Soft Comput 19(7):1997–2010. https://doi.org/10.1007/s00500-014-1387-5
Zhao XL, Yang Y (2019) Incremental attribute reduction algorithm based on neighborhood granulation conditional entropy. Control Decis 34(10):2061–72. https://doi.org/10.13195/j.kzyjc.2018.0138
Zhou YH, Zhang XY, Mo ZW (2018) Conditional neighborhood entropy with granulation monotonicity and its relevant attribute reduction. J Comput Res Dev 55(11):2395–2405
Zhou YH, Zhang Q (2020) Three-way neighborhood entropies based on three-layer granular structures. Math Pract Theory 50(14):83–93
Zou L, Li HX, Jiang W, Yang XH (2019a) An improved fish swarm algorithm for neighborhood rough set reduction and its application. IEEE Access 7:90277–90288. https://doi.org/10.1109/ACCESS.2019.2926799
Zou L, Sun YB, Yang XH (2019b) An entropy-based neighborhood rough set and PSO-SVRM model for fatigue life prediction of titanium alloy welded joints. Entropy (Basel, Switzerland) 21(2):117. https://doi.org/10.3390/E21020117
Zou L, Ren S, Li H, Yang XH (2021) An optimization of master S - N curve fitting method based on improved neighborhood rough set. IEEE Access 99:1–1. https://doi.org/10.1109/ACCESS.2021.3049403

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.