IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 17, 2020, accepted March 6, 2020, date of publication March 11, 2020, date of current version April 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980162

Research on the Feature Selection Approach Based on IFDS and DPSO With Variable Thresholds in Complex Data Environments

YANAN HU, CHUNSHENG LI, KEJIA ZHANG, MEI WANG, AND YATIAN GAO
School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China

Corresponding author: Chunsheng Li (slee@nepu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61170132, and in part by the Fundamental Research Funds for the Northeast Petroleum University under Grant 2017PYZQL-11 and Grant KYCXTD201903.

ABSTRACT Neighborhood rough model is widely used in feature selection with high dimension, fuzzy, continuous and discrete attributes, incomplete data and so on, and the application of neighborhood rough model depends on neighborhood threshold. In the application of the model, the point-value neighborhood threshold is not adaptive, which leads to low classification accuracy and high time complexity of the algorithm. In order to solve the above problems, a feature selection approach based on IFDS (Incomplete Fuzzy Hybrid Decision System) and DPSO (Discrete Particle Swarm Optimization Algorithm) with variable thresholds is proposed. Firstly, a neighborhood rough model capable of simultaneously processing fuzzy, hybrid and incomplete data was established. The average reachable distance was introduced to construct the attribute neighborhood threshold set and reduce the interference of noise data on classification accuracy. Secondly, we constructed the DPSO particle fitness function using the feature subset length, the significance of the attribute and the negative domain of the neighborhood, and improved the inertia weight computing method, so as to enhance the feature selection speed and the feature subset quality. Finally, the simulation experiment was performed using the real industrial production data. The experiment effect shows that this method has obvious advantages in improving the classification accuracy, optimizing the search speed and the optimal feature subset quality.

INDEX TERMS Feature selection, fuzzy, hybrid and incomplete data, incomplete fuzzy hybrid decision system (IFDS), adaptive neighborhood threshold, discrete particle swarm optimization algorithm (DPSO).

I. INTRODUCTION

"Taking preventive measures" is always better than "fixing the problem later" in industrial production failures. Therefore, how to effectively predict production failures has always been the focus of decision makers. At present, researchers usually predict production failures via mining failure occurrence patterns. This process can be abstractly described as how to select the feature set that can accurately describe the regularity of production failure occurrence and how to determine the feature threshold. The solution to this problem undoubtedly requires the support of a large amount of production monitoring data, while such data has typical complexity, which is specifically characterized by high dimensionality, discrete/continuous data mixing, data missing, etc. As a result, "how to find out the sensitive characteristics of failure occurrence and determine its threshold from complex data" has become one of the challenges in the prediction of production failures, which has always attracted the attention of scientific researchers.

Rough set theory [1] can process fuzzy data with a definite method, and has a good application effect in the fields of pattern recognition, attribute reduction, failure diagnosis, and production abnormality analysis [2]–[8]. However, the classical rough set theory requires to discretize the data in advance when processing continuous data. This process inevitably leads to the loss of information, resulting in the decrease in data identification ability and inaccurate feature selection results. As to this problem, researchers have made a lot of improvements based on the original rough set to improve the accuracy of feature selection.

Hu *et al.* [9] applied the spherical neighborhood theory in the topological space to rough sets and constructed a feature selection algorithm for continuous data, which eliminates

The associate editor coordinating the review of this manuscript and approving it for publication was Rajesh Kumar.

the step of attribute discretization and retains the original meaning of numerical data to the greatest extent. Ma and Li [10] presented a data-driven method for fault detection and diagnosis. Aiming at the problem of low diagnosis accuracy and not easy to obtain incomplete data, the neighborhood rough set and signed directed graph are combined to improve the diagnosis efficiency without prior knowledge. To reduce the impact of noisy data, Suo *et al.* [11] proposed a diagnosis approach based on a variable precision fuzzy neighborhood rough set (VPFNRS) model, which could extract fuzzy rules from hybrid data with noises and make fuzzy diagnosis results based on the extracted fuzzy rule model and the weights of condition attributes. Wang *et al.* [12] designed a variable-precision fuzzy neighborhood rough set model to reduce the possibility of samples being misclassified. At the same time, the dependency between the fuzzy decision and condition attribute was used to evaluate the significance of candidate features, greatly improving the classification performance.

The feature selection of incomplete decision systems is another important application scenario of rough set theory. In such scenario, scientific researchers have proposed some methods to expand and improve rough sets, such as tolerance relation, similarity relation and quantitative tolerance relation [13]–[15], Dai [16] proposed an extended rough set model, i.e., tolerance-fuzzy rough set model to deal with this type of data characterized with numerical attributes and missing values, that is, incomplete numerical data. Zhao and Qin [17] introduced an extended rough set model, which is based on neighborhood-tolerance relations. However, in these feature selection algorithms, the description of the relationship between data is too loose, and there are many misclassification phenomena. Therefore, Zhao *et al.* [18] proposed the neighborhood rough set of IFDS, which realized the processing of lost values and missing values in complex data environments, and achieved good results. The above neighborhood-based rough set research shows that setting appropriate neighborhood thresholds can improve the classification effect, and does not reduce the classification accuracy while selecting fewer features. Therefore, scholars have carried out the research on variable neighborhood thresholds to improve the accuracy of feature selection.

In literature [19], the maximum distance and the minimum distance between the training object and the test object were used to achieve dynamic update of the threshold; In literature [20], the attribute threshold set was used to replace the single attribute threshold, and the corresponding threshold was set for each attribute according to the standard deviation of each attribute. But the influence of abnormal values on the maximum and minimum distances is ignored in the above method. Therefore, how to choose the neighborhood threshold with higher fitness is a key problem to be solved.

Obtaining the optimal feature subset from the high-dimension feature space is an NP-complete problem [21], [22]. Simply using rough set theory for feature selection has such defects as time consuming and that the quality of the optimal solution set cannot be guaranteed. Therefore, scholars at home and abroad have launched research on the feature selection approach combining the rough set theory with heuristic algorithms. Particle Swarm Optimization (PSO) is widely used in the field of feature selection. Compared with the genetic algorithm (GA), PSO does not need to perform complex operations such as crossover and mutation, so it has lower memory occupancy, lower computing cost and faster convergence speed [23]–[25]. Wang *et al.* [26] proposed a feature selection approach based on the rough set and ant colony optimization algorithm, which takes attribute dependency and attribute significance as heuristic factors, the classification quality and feature subset length of rough set as the ant colony update strategy, and employs the data set for test evaluation. The result shows that key features can be obtained from the algorithm, but as the sample size of the data set increases, the efficiency of the algorithm decreases, and it gradually falls into the local optimal state. Chen *et al.* [27] proposed a feature selection algorithm based on neighborhood rough set model and PSO to solve the feature selection of intrusion detection log data, and constructed the fitness function using the positive domain of the attribute subset, the attribute dependency and the number of attributes. However, the algorithm ignores the influence of neighborhood size on feature classification, and the optimal solution quality cannot be guaranteed. Therefore, how to improve the quality of the optimal feature subset is a key problem to be solved.

In view of the above problems, we proposed a hybrid incomplete feature selection approach based on IFDS and DPSO with variable thresholds in this paper. First, the concept of reachable distance was introduced, and the neighborhood threshold was constructed by using the average reachable distance of the attribute, to reduce the influence of abnormal values on the threshold and improve the classification accuracy. Second, the DPSO was used to accelerate the speed of feature selection, and the DPSO inertia weight computing method was improved to enhance the quality of the optimal feature subset. Finally, the proposed algorithm was used to solve the problem of feature selection for oilfield scaling prediction, and to verify the feasibility and effectiveness of the method.

The paper is arranged as follows. Part II is about the basic work, which arranges the application scenarios, describes the basic concepts related to rough sets, and summarizes the key issues to be solved. Part III introduces the fuzzy hybrid incomplete neighborhood rough model, and details the improved method of the variable neighborhood threshold. Part IV describes the hybrid incomplete feature selection approach based on neighborhood reduction and DPSO. Part V analyzes the proposed method experimentally to verify the feasibility of the method. Part VI summarizes the research results.

II. PREPARATION WORK

Before discussion, we gave the definition and basic concepts of the scenario, and described the key problems that need to

TABLE 1. Symbols and interpretations.

Symbols	Interpretations	Symbols	Interpretations
U	universe	$\underline{R}X$	Lower approximations
A	A set of attributes	$\bar{R}X$	upper approximations
KRS	Knowledge representation system	$BA(X)$	The boundary region of X
f	information function of IS	$ND(X)$	the negative domain
V	the attribute value domain of U	$I(A,B)$	A 's inclusion degree in B
IS	Information system	IDS	Incomplete decision system
C	the set of condition attributes	FDS	fuzzy hybrid decision system
D	decision attributes	$IFDS$	Incomplete fuzzy hybrid decision system
B	feature space	$\gamma_B(D)$	The dependency degree of D to B
$\delta_B(x_i)$	the neighborhood information granule	$SIG(a,A,d)$	significance of a in A
δ	neighborhood size	R	A neighborhood relation
Δ_p	Minkowsky distance	NAS	neighborhood approximation space

be solved. For the readers' understanding, the symbols and interpretations in this paper are shown in Table 1.

A. SCENARIOS AND BASIC CONCEPTS

*Definition of Scenario **H**:* It is known that $U = \{x_1, x_2, \cdots, , x_n\}, A = \{a_1, a_2, \ldots, a_m\}$ covers the complete description of any feature x_i. In the knowledge representation system, x_i can be mapped to multiple items of the data entity, and the data reflects the change rule of x_i.

Definition 1: $KRS = (U, A, V, f)$, $V = \{V_{aj}\}$ represents the attribute value domain of U; f represents the mapping of the set of attributes of U to the attribute value, recorded as $f : U \times A \to V$. The data type of V includes discrete type and continuous type.

Information system and decision system are two typical applications of the knowledge representation system. Their difference lies in whether the decision attribute is included. The two systems are abstractly defined as follows respectively.

Definition 2: Information system $IS = (U, C, V, f)$, where $C = \{a\,|a \in C\}$ is the nonempty finite set of the attribute, and is called the condition attribute; V is the attribute value domain of the object, $V_j\,(1 \leq j \leq m)$ stands for the value domain of the attribute a_j; f is the information function of IS, and f_j is the information function of a_j.

Definition 3: Decision system $DS = (U, C \cup D, V, f)$, where C is the condition attribute, $D = \{d\,|d \in D\}$ means the set of decision attributes, and must satisfy $C \cap D = \varnothing, C \neq \varnothing, D \neq \varnothing$; f is the decision function, and can be expressed as $f = \{f_a\,|f_a : U \to V_a, \forall a \in C \cup D\}$, where f_a is the information function of a.

Definition 4: The neighborhood represents the maximum distance between the center of the neighborhood and the boundary. The neighborhood $\delta_B\,(x_i)$ of any x_i in the feature space B can be defined as below:

$$\delta_B\,(x_i) = \left\{x_j \,\middle|\, x_j \in U, \Delta^B\left(x_i, x_j\right) \leq \delta\right\} \tag{1}$$

where, Δ stands for the distance function. For any continuous three features x_{i-1}, x_i, x_{i+1}, Δ meets the following characteristics: (1) nonnegative number, that is, $\Delta\,(x_1, x_2) \geq 0$; (2) commutative law, that is, $\Delta\,(x_1, x_2) = \Delta\,(x_2, x_1)$; (3) $\Delta\,(x_1, x_3) \leq \Delta\,(x_1, x_2) + \Delta\,(x_2, x_3)$. Thus, the distance computing method for N features is shown in the formula (2).

$$\Delta_p\,(x_1, x_2) = \left(\sum_{i=1}^{N} |f\,(x_1, a_i) - f\,(x_2, a_i)|^p\right)^{1/p} \tag{2}$$

where, f (x, a_i) describes the value of the feature x mapped to the attribute a_i. If P = 1, Δ_1 is equivalent to the Manhattan distance; if P = 2, Δ_2 is equivalent to the Euclidean distance; if P = ∞, the function equals to the Chebyshev distance. The distance function is introduced in detail in literature [28].

$\delta_B\,(x_i)$ is the neighborhood information particle centered with sample x_i, and is called the neighborhood particle of x_i; its neighborhood size depends on the threshold δ, the larger the value of δ, the more samples fall into the neighborhood of x_i.

Definition 5: We suppose that $B_1 \subseteq A$ represents the numerical attribute, $B_2 \subseteq A$ the symbolic attribute. The neighborhood particle of the sample x based on the numerical attribute, the symbolic attribute or the hybrid attribute is defined as follows:

$$\delta_{B_1}\,(x) = \left\{x_i \,\middle|\, \Delta_{B_1}\,(x, x_i) \leq \delta, x_i \in U\right\};$$
$$\delta_{B_2}\,(x) = \left\{x_i \,\middle|\, \Delta_{B_2}\,(x, x_i) = 0, x_i \in U\right\};$$
$$\delta_{B_1 \cup B_2}\,(x) = \left\{x_i \,\middle|\, \Delta_{B_1}\,(x, x_i) \leq \delta \wedge \Delta_{B_2}\,(x, x_i) = 0, x_i \in U\right\}.$$

Therefore, we got the following properties:

$$\delta\,(x_i) \neq \varnothing, \quad \text{because } x_i \in \delta\,(x_i);$$
$$x_j \in \delta\,(x_i) \Rightarrow x_i \in \delta\left(x_j\right);$$
$$\bigcup_{i=1}^{n} \delta\,(x_i) = U.$$

So, the neighborhood particle family $\{\delta\,(x_i)\,|i = 1, 2, \ldots, n\}$ constitutes a covering of U. $x_i \neq x_j$ cannot ensure $x_i \notin \delta\left(x_j\right)$, so $\{\delta\,(x_i)\,|i = 1, 2, \ldots, n\}$ generally does not constitute a partition of U.

The neighborhood particle family constructs a neighborhood relation R on the universe of discourse space U, and R can be expressed by the matrix of relation $\mathrm{M}\left(\mathrm{R}\right)=\left(\mathrm{r}_{ij}\right)_{n\times n}$, where:

$$\mathrm{r}_{ij}=\begin{cases}1, & \Delta\left(\mathrm{x}_i,x_j\right)\leq\delta\\ 0, & otherwise\end{cases}\tag{3}$$

The above definition shows that R has reflexivity and symmetry.

The neighborhood of all objects in the universe of discourse forms the granulation of the universe of discourse, $\{\delta\left(x_i\right)|i=1,2,\ldots,n\}$ constitutes the basic concept of the universe of discourse space. Through these concepts, any concept in the space can be approximated.

Definition 6: It is defined that the neighborhood approximation space $NAS=\langle U,R\rangle$, $X\subseteq U$, then the relation between the lower approximation and the upper approximation of X in NAS is defined as below:

$$\begin{aligned}\underline{R}X&=\{x_i\,|\delta(x_i)\subseteq X,\,x_i\in U\}\\ \overline{R}X&=\{x_i\,|\delta(x_i)\cap X\neq\varnothing,\,x_i\in U\}\end{aligned}\tag{4}$$

$\forall X\in U,\underline{R}X\subseteq X\subseteq\overline{R}X$, the approximate boundary of X can be represented as follows:

$$BA(X)=\overline{R}X-\underline{R}X\tag{5}$$

$\underline{R}X$ is the positive domain of X in NAS, X completely contains the maximum union set of the neighborhood information particle of $\underline{R}X$; while $\overline{R}X$ can completely contain the minimum union set of the neighborhood information particle of X; the neighborhood information particle completely unrelated to X is the negative domain of X, which can be shown as below:

$$ND(X)=U-\overline{R}X\tag{6}$$

Attribute Normalization: The data of different attributes in the domain has different dimensions. In order to reduce the influence of different dimensions on classification accuracy, the original data is normalized. The normalization formula is expressed as follows:

$$y=\frac{x-x_{min}}{x_{max}-x_{min}}\tag{7}$$

B. PROBLEM DESCRIPTION

The "optimal" feature selection is to try to satisfy the "fastest search speed requirement" on the premise of giving priority to the "highest accuracy requirement". The "highest accuracy requirement" refers to the highest possible classification accuracy; the "fastest search speed requirement" refers to the fastest possible feature selection speed, that is, the smallest possible time complexity.

The "highest accuracy requirement" is to select the feature subset with the best quality. The conditions for selecting the optimal feature subset are to minimize the number of features obtained by feature selection, maximize the sum of the significance of the obtained feature subsets, and minimize the negative domain of the neighborhood.

"Minimizing the number of features" is expressed as $Arg\,min\left(count\left(F\left(X_i\right)\right)\right)$, where, $F\left(X_i\right)$ is the description of the feature subset of the particle X_i, $count\left(F\left(X_i\right)\right)$ means the number of features selected in the particle X_i.

"Maximizing the significance of the feature" is shown as $Arg\,max\left(\sum_{n=1}^{count(A)}SIG(a_n,A,D)\right)$, where, $SIG\left(a_n,A,D\right)$ is the significance of the attribute a_n, and $\sum_{n=1}^{count(A)}SIG\left(a_n,A,D\right)$ represents the sum of all attribute significance in the set of attributes A.

"Minimizing the negative domain of the neighborhood" is written as $Arg\,min\left(ND(X_i)\right)$.

The neighborhood threshold determines the neighborhood size. However, the degree of neighborhood granulation depends on the neighborhood size, and the upper and lower approximations are generated on the basis of neighborhood granulation. Therefore, the value of the neighborhood threshold δ is very important for the neighborhood system.

Finally, the problem of selecting the optimal feature subset can be described as an optimization problem with $Arg\,min\left(count\left(F\left(X_i\right)\right)\right)$, $Arg\,min\left(ND(X_i)\right)$, $Arg\,max\left(\sum_{n=1}^{count(A)}SIG\left(a_n,A,D\right)\right)$ as the objective function.

Given the above, the key problems to be solved in this paper are summarized as follows:

i. How to select the neighborhood threshold δ for obtaining the highest rough set classification accuracy.

ii. How to quickly obtain the feature subset with the optimal quality from the strategy under multiple optimization objectives and constraints.

III. FUZZY HYBRID INCOMPLETE NEIGHBORHOOD ROUGH MODEL

Scenario H is supposed to be an incomplete decision system, and the attribute types in the scenario include fuzzy and clear, symbolic and numerical, lost and missing types. The incomplete decision system and the model are defined as follows:

A. IFDS

Definition 7: In DS, if $V=Sy\cup Nu$, where, Sy represents the symbolic variable, and Nu the numerical variable, DS is called the hybrid decision system.

If the value domain of any attribute $a\in\left(C\cup D\right)$ in DS is $V_a=\{?,^*\}$, where, "?" stands for the data with lost attribute value, and "*" the data with missing attribute value, that is, it contains the null value, DS is called the incomplete decision system, and expressed as IDS.

Definition 8: If the hybrid decision system $DS=\left(U,C\cup D,V,f,Fd\right)$, where, $Fd=\left\{D_j^-:U\rightarrow[0,1]\,(j\leq r)\right\}$ is the collection of the fuzzy decision set, DS is called the fuzzy objective decision system, and expressed as FDS.

Definition 9: The decision system containing fuzzy and clear data, symbolic and numerical data, lost and missing data is called the incomplete fuzzy hybrid decision system, and expressed as $IFDS$. $IFDS = (U, A, V, f)$, $V = V_A \cup V_D \cup \{?, *\}$.

B. FUZZY HYBRID INCOMPLETE NEIGHBORHOOD ROUGH MODEL

The neighborhood model granulates the universe of discourse by the neighborhood of the object, and takes the neighborhood as the basic information particle. The neighborhood rough model extends the equivalence relation of the traditional rough set, and measures the indiscernible relation with the neighborhood relation, so it can directly process the discrete attribute.

Definition 10: In $IFDS$, we define the numerical attribute as $B_1 \subseteq A$, the symbolic attribute as $B_2 \subseteq A$, the fuzzy attribute as $B_3 \subseteq A$, the lost incomplete attribute as $B_4 \subseteq A$, and the missing incomplete attribute as $B_5 \subseteq A$, then the neighborhood of the sample x is defined as below:

$$\begin{aligned} \delta_{B_1}(x) &= \left\{x_i \middle| \Delta_{B_1}(x, x_i) \le \delta, x_i \in U\right\} \\ \delta_{B_2}(x) &= \left\{x_i \middle| \Delta_{B_2}(x, x_i) = 0, x_i \in U\right\} \\ \delta_{B_3}(x) &= \left\{x_i \middle| \Delta_{B_3}\left[f_l(x), f_l(x_i)\right] = 0, x_i \in U\right\} \\ \delta_{B_4}(x) &= \{x_i | x_i = ?, x_i \in U\} \\ \delta_{B_5}(x) &= \left\{x_i | x_i =^*, x_i \in U\right\} \\ \delta_{B_1 \cup B_2 \cup B_3 \cup B_4 \cup B_5}(x) &= \left\{x_i \middle| \Delta_{B_1}(x, x_i) \le \delta \right. \\ &\quad \wedge \Delta_{B_2}(x, x_i) = 0 \\ &\quad \wedge \Delta_{B_3}[f_l(x), f_l(x_i)] = 0 \\ &\quad \left. \wedge (x_i = ? \parallel x_i =^*), x_i \in U\right\} \end{aligned} \tag{8}$$

Then, we can obtain following properties:

$$\begin{aligned} &\delta(x_i) \ne \varnothing, \quad x_i \in \delta(x_i) \\ &\cup_{i=1}^{n} \delta(x_i) = U \end{aligned}$$

The neighborhood of all objects in the universe of discourse forms the granulation of the universe of discourse and the universe of discourse particle family constitutes the covering of U.

The neighborhood relation R of $IFDS$ is defined as follows:

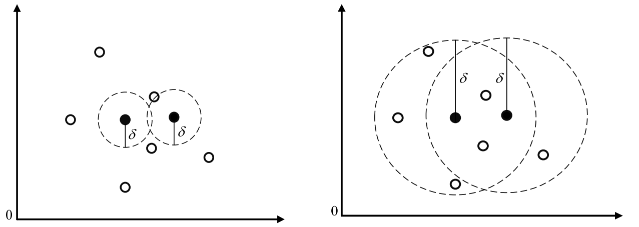
$$\begin{aligned} R(x) = &\left\{(x, y) \in U^2 : \forall a \in X \cap a(x) \ne ? \cap f_i(x) = f_l(y), \right. \\ &\left. a(x) \in \delta(y, a) \cup a(y) \in \delta(x, a) \cup a(x) =^* \cup a(y) =^* \right\} \end{aligned} \tag{9}$$

Due to the reflexivity and symmetry of R, $R(x)$ can be simplified as below:

$$\begin{aligned} R(x) = &\left\{(x, y) \in U^2 : \forall a \in X \cap a(x) \ne ? \cap f_i(x) = \right. \\ &\left. f_l(y), a(x) \in \delta(y, a) \cup a(x) =^* \cup a(y) =^* \right\} \end{aligned} \tag{10}$$

Definition 11: $IFDS = (U, A, D)$, D divides U into P equivalence classes: $X_1, X_2, X_3, \ldots, X_N, B \subseteq A$ generates the neighborhood relation R on U, then the lower approximation and the upper approximation of the neighborhood of the decision D for B can be expressed by the formula (11):

$$\begin{aligned} \underline{R_B}D &= \left\{\underline{R_B}X_1, \underline{R_B}X_2, \ldots, \underline{R_B}X_P\right\} \\ \overline{R_B}D &= \left\{\overline{R_B}X_1, \overline{R_B}X_2, \ldots, \overline{R_B}X_P\right\} \end{aligned} \tag{11}$$



(a) Situation when the value of δ is small

(b) Situation when the value of δ is large

FIGURE 1. Values of the threshold δ.

The decision boundary can be expressed as:

$$BA(D) = \overline{R_B}D - \underline{R_B}D \tag{12}$$

When $BA(D) = \varnothing$, that is, $\overline{R_B}D = \underline{R_B}D$, X is called to be definable on the approximation space $NAS = \langle U, R \rangle$, or else, X is called as the rough set.

The dependency of the decision attribute D on the condition attribute $B \subseteq A$ is expressed as:

$$\gamma_B(D) = {Card(N_B D)}/{Card(U)} \tag{13}$$

When the dependency degree of the attribute is zero, the attribute is redundant. Therefore, the dependency degree reflects the significance of the attribute. The higher the dependency degree of the decision attribute on the condition attribute, the higher the significance of the condition attribute, and vice versa. The significance of the attribute $a \in A$ can be expressed as below:

$$SIG(a, A, D) = \gamma_A(D) - \gamma_{A-a}(D) \tag{14}$$

C. ADAPTIVE NEIGHBORHOOD THRESHOLD

The above research shows that the value of the neighborhood threshold δ affects the classification accuracy of the attribute reduction. Taking the two-dimensional data space in Figure 1 as an example, when the value of the threshold δ is too small, $\delta(x_i)$ only contains the sample x_i, at this time, any two condition attributes in the set of attributes can plan the sample to be tested into the positive domain, which makes the classification accuracy relatively low; when the value of δ is too large, the number of samples included in $\delta(x_i)$ will also increase, and δ may include all samples of the universe of discourse U. Similarly, at this time, any two condition attributes in the set of attributes can include the sample to be tested into the positive domain, leading to the decrease of the dependency degree of attributes. While the data of different attributes has different distribution densities, and setting the same threshold δ will inevitably bring a large error to the attribute reduction. Therefore, we proposed a threshold size setting method based on the density of the attribute, defined a threshold δ suitable for each attribute, and formed a threshold set in this paper.

Definition 12: We suppose that the data set of any attribute $a \in A$ in the universe of discourse U is X, x is a point in X, the parameter ε is the radius, also known as the distance, the parameter $MinPts$ is the minimum number of samples in the neighborhood. The core distance with the minimum

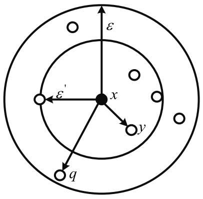


FIGURE 2. Core distance and reachable distance.

neighborhood radius x that makes x the core point is defined by the formula (15):

$$cd\,(x) = \begin{cases} Undefined & |N_\varepsilon\,(x)| < MinPts \\ d\left(x, N_\varepsilon^{MinPts}\,(x)\right) & |N_\varepsilon\,(x)| \geq MinPts \end{cases} \quad (15)$$

where, $N_\varepsilon^j\,(x)$ represents the point neighboring j of the node x in the set $N_\varepsilon\,(x)$; if x is the core point, $cd\,(x) < \varepsilon$. For the point x, its core distance is the minimum ε' that makes x the core point. If x is not the core object, the core distance of x is not defined, as shown in Figure 2.

Definition 13: We set $x, y \in X$, for given parameters ε, $MinPts$, the reachable distance of y relative to x is defined as below:

$$rd\,(y, x) = \begin{cases} Undefined & |N_\varepsilon\,(x)| < MinPts \\ max\,\{cd\,(x)\,, d\,(x, y)\} & |N_\varepsilon\,(x)| \geq MinPts \end{cases} \quad (16)$$

That is, the reachable distance of the point y relative to the point x is the larger Euclidean distance between the core distance of the point x and x, y. if x is not the core object, the reachable distance between x, y is meaningless.

For example, as shown in Figure 3, we set $\varepsilon = 7$, $MinPts = 5$, when x is the core point, the maximum range of x to the nearest five points is $\varepsilon' = 3$, so the core distance of the point x is 3. The reachable distance of the point y relative to the point x is $Max\,\{3, d\,(x, y)\} = 3$, and that of the point q relative to the point x is $Max\,\{3, d\,(x, q)\} = d\,(x, q)$.

According to the above definition, the neighborhood threshold δ of the attribute $a \in A$ is:

$$\delta\,(a) = min_{i=1}^{l}\,(rd\,(x_i))\,, \quad x_i \neq ?,^* \quad (17)$$

where, $l = count\,(a)\,, MinPts = l$, $MinPts$ takes the number of objects that the attribute a contains, and the threshold of the attribute a takes the minimum reachable distance in all objects of the attribute data set.

IV. FEATURE SELECTION OF HYBRID INCOMPLETE DATA BASED ON NEIGHBORHOOD REDUCTION AND DPSO

The key to attribute reduction using DPSO is the construction of the particle encoding mode and the fitness function. The parameter settings that can optimize the convergence speed and improve the quality of the optimal solution are given below, including the particle encoding mode and the fitness function suitable for neighborhood reduction of *IFDS*.

A. PARTICLE ENCODING MODE

The feature selection problem can be described as selecting $M, M \leq count\,(A)$ attributes from the set of attributes A of U to form a subset of attributes that can fully describe the feature of the data approximately. Each attribute has two states, that is, whether it is selected and included into the feature subset. Therefore, each attribute can be defined as a one-dimensional binary variable of the particle, and $count\,(A)$ attributes constitute the $count\,(A)$-dimension discrete binary space of the particle. We define that S indicates the state whether the attributes in the set of attributes A are selected, then the feature subset strategy searched by the particle P can be formally expressed as:

$$P = \left\langle S_1 S_2 S_3 S_4 \cdots S_i \cdots S_{count\,(A)} \right\rangle$$

For example, the expression form of the feature subset strategy containing six attributes is $P = \langle 101011 \rangle$, $S_1, S_3, S_5, S_6 = 1$ denotes that the attribute is selected and included into the feature subset, so the corresponding feature subset is $F\,(P) = \{S_1, S_3, S_5, S_6\}$.

The number of feature subset strategies that may be generated by the above encoding mode is $W = count\,(P) = 2^{count\,(A)}$.

B. FITNESS FUNCTION

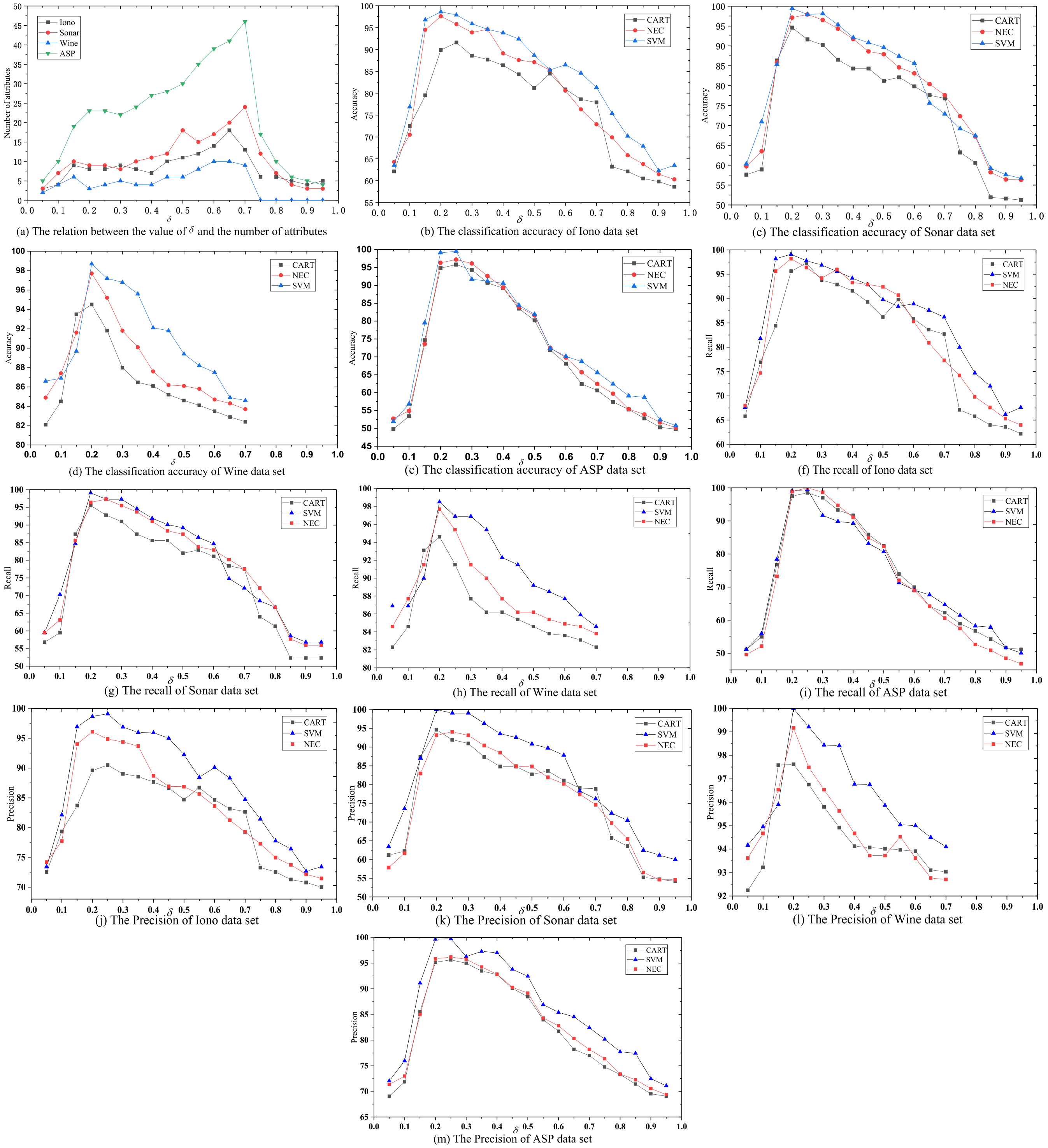
The fitness function is mainly used to evaluate the quality of the particles. The higher the fitness of the particles, the better the quality of the particles. Based on the feature selection problem of *IFDS*, it aims to minimize the number of features obtained by feature selection, and that the obtained feature subset will not reduce the classification accuracy of the sample at the same time, that is, maximizing the sum of the significance of the features in the feature subset, and minimizing the negative domain of the neighborhood. Therefore, the "number of feature subsets", the "significance of feature subset attributes" and "the negative domain of the neighborhood" were selected as objective functions. Solving multi-objective optimization problems usually requires normalizing the objective function. Here, the linear weighting method was used to convert the multi-objective function into the single objective function. We define the weighting factor $w = \{w_c, w_s, w_r\}$ and it satisfies $w_c + w_s + w_r = 1$. Then the fitness function is written as follows:

$$\begin{aligned} P(X_i) &= w_c \times Arg\,min\,(count\,(F(X_i))) + w_s \\ &\quad \times Arg\,max\left(\sum_{n=1}^{count(A)} SIG\,(a_n, A, D)\right) + w_r \\ &\quad \times Arg\,min\,(ND\,(X_i)) \end{aligned} \quad (18)$$

The fitness function can be converted to:

$$\begin{aligned} P\,(X_i) &= w_c \times Arg\,min\,(count\,(F\,(X_i))) + w_s \\ &\quad \times Arg\,min(\frac{1}{\sum_{n=1}^{count(A)} SIG\,(a_n, A, D)}) + w_r \\ &\quad \times Arg\,min\,(ND\,(X_i)) \end{aligned} \quad (19)$$

Therefore, the selection of feature subsets can be described as the problem of obtaining the minimum expectation of the objective function, that is, $Arg\,min\,(P\,(X_i))$.



(a) The relation between the value of δ and the number of attributes

(b) The classification accuracy of Iono data set

(c) The classification accuracy of Sonar data set

(d) The classification accuracy of Wine data set

(e) The classification accuracy of ASP data set

(f) The recall of Iono data set

(g) The recall of Sonar data set

(h) The recall of Wine data set

(i) The recall of ASP data set

(j) The Precision of Iono data set

(k) The Precision of Sonar data set

(l) The Precision of Wine data set

(m) The Precision of ASP data set

FIGURE 3. The relation between the value of δ and the classification accuracy.

C. PARAMETER SETTING

The parameter setting in DPSO is shown below:

We define the dimension as C, the number of particles as K, the number of iterations as I, the flight speed of the particle $u\,(u < K)$ at the $i\,(i < I)$-th time as v_u^i, the position of the particle u at the i-th iteration as $x_{u,i}$, the historically optimal position of the particle u as P_{u-best}, and the historically optimal position of the population as P_{g-best}. Then the speed updating formula and the position updating formula of the particle u are listed as follows:

$$v_u^{i+1} = \omega \times v_u^i + c_1 \times r_1 \left(P_{u-best} - x_{u,i}\right) + c_2 \times r_2 \left(P_{g-best} - x_{u,i}\right) \tag{20}$$

$$x_{u,i+1} = x_{u,i} + v_u^{i+1} \tag{21}$$

c_1, c_2 are called the acceleration constant, r_1, r_2 are the random number, and the value range is [0, 1]. ω is the inertia coefficient used to equalize the local optimum and the global optimum. In order to further improve the convergence speed of the algorithm, a dynamic weight calculation method was proposed, so that the weight decreases linearly with the number of iterations, then the value of ω at the i-th iteration can be expressed as:

$$\omega\left(i\right)=\omega_{max}-\frac{\omega_{max}-\omega_{min}}{D}\times i \tag{22}$$

DPSO will normalize the speed as the basis for updating the particle position. Therefore, the speed is normalized by using the *sigmoid* function, and the processing method is shown as follows:

$$sigmoid\left(v_{u}^{i}\right)=\frac{1}{1+exp\left(-v_{u}^{i}\right)} \tag{23}$$

D. ALGORITHM DESCRIPTION

The algorithm for feature selection using DPSO is described as follows:

Algorithm 1 Hybrid Incomplete Data Feature Selection Algorithm Based on Neighborhood Reduction and DPSO (Called as NRDPSO Algorithm)

Input:

D_Info: It represents the initial feature subset list formed after attribute reduction by neighborhood reduction method of *IFDS*. Each attribute subset in the list is encoded and contains information such as the feature length, the sum of attribute significance, and the negative domain. The list forms a space of discrete points.

P_info: The dynamic information of the particle, including the current speed v, fitness P, negative domain ND, and the historically optimal solution of the particle v_{best}.

Parameter setting: Acceleration constant c_1, c_2; Speed v_{max}, v_{min}; Number of iterations I; Number of particles K; Random number r_1, r_2; Inertia weight $\omega_{max}, \omega_{min}$; Encoding length (number of members) C, Number of test samples SN.

Output:

globe_best: Global optimal solution (optimal feature subset)

```
Begin
for $i = 0$ to $P\_info.length$ do
    set $P\_info[i].v$=Random($v_{min}, v_{max}$); /*Speed of initialized particle */
    set $P\_info[i].x$=Random( D_Info); /*Position of initialized particle*/
    set $P\_info[i].P$=Fitness( $P\_info[i].x$); /*Fitness of initialized particle*/
    set $P\_info[i].v\_best$=$P\_info[i].P$; /*Historically optimal solution of initialized particle*/
end
Update( globe_best, P_info); /*Update the global optimal solution by P_info*/
for $i = 1$ to $I$ do
    set $\omega = \omega_{max} - (\omega_{max} - \omega_{min}) \times i/I$; /*Update inertia weight*/
    for $j = 1$ to $K$ do
        set
            $P\_info[j].v = \omega \times P\_info[j].v + c_1 \times r_1 (P\_info[j].v_{best} - P\_info[i].P) + c_2 \times r_2 (globe\_best - P\_info[i].P)$; /*Update speed*/
        set $sigmoid(P\_info[j].v) = \frac{1}{1+exp(-P\_info[j].v)}$;/*Update speed*/
    set $P\_info[i].x = P\_info[i].x + P\_info[i].v$;/*Update position*/
    set $P\_info[i].P = Fitness(P\_info[i].x)$; /*According to the current fitness of the particle*/
    Update($P\_info[i].v_{best}$, $P\_info[i].x$) /*Update the historically optimal solution of initialized particle by $P\_info[i].x$*/
    Update( globe_best, P_info); /*Update the global optimal solution by P_info*/
```

E. COMPUTATION COMPLEXITY ANALYSIS OF THE ALGORITHM

In the NRDPSO algorithm, the number of particles and the number of iterations affect the search range and coverage of the algorithm respectively. Increasing the number of iterations properly can improve the quality of the optimal feature subset to some extent; increasing the number of particles can enlarge the search scope and reduce the possibility of falling into local optimum.

During the algorithm execution, the time complexity of the initialization population is $O\left(C\times K\right)$; When calculating the fitness value of each particle, it is necessary to construct an appropriate neighborhood threshold for the attributes in each particle, and recalculate the neighborhood of the sample. The time complexity at this stage is $O\left(C\times SN^{2}\times K\right)$; the time complexity of updating particle speed and position is $O\left(C\times K\right)$.

To sum up, the time complexity of the algorithm NRDPSO is $O\left(C\times K+I\times\left(C\times SN^{2}\times K+C\times K\right)\right)$.

V. EXPERIMENTS AND ANALYSIS

The comparative experiment method was used to verify the advantages of the proposed method in improving the classification accuracy, optimizing the search speed and the quality of the optimal feature subsets. The experimental design is briefly described as follows:

i. Clarify the experimental preparation, prepare the experimental data, set the environmental parameters, and describe the comparative experimental methods.

ii. About the experimental effect analysis, analyze and compare the number of feature selection and classification accuracy of NRDPSO algorithm based on different neighborhood threshold values; compare the convergence speed of the feature selection approach by the IFDS with variable thresholds based on different heuristic algorithms; compare the influence of different weight coefficients in heuristic algorithms on the optimal feature subset.

Algorithm 1 *(Continued.)* Hybrid Incomplete Data Feature Selection Algorithm Based on Neighborhood Reduction and DPSO (Called as NRDPSO Algorithm)

```
    end for
  end for
  return globe_best
End
```

TABLE 2. Description of UCI data set.

Data Set	Abbreviation	Number of Instances	Numerical Features	Classes
Ionosphere	Iono	351	34	2
Sonar,Mines vs.Rocks	Sonar	208	60	2
Wine recognition	Wine	178	13	3

In this paper, Accuracy, Recall, and Precision were used as evaluation indexes for the merits of the algorithm.

$$\text{Accuracy}: A = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Recall}: R = \frac{TP}{TP + FN}$$

$$\text{Precision}: P = \frac{TP}{TP + FP}$$

where *TP*(True Positive) means that the real category is Positive, and the prediction category is Positive. *FP*(False Positive) indicates that the real category is negative, and the prediction category is Positive. *TN* (True Negative) means that the True category is Negative, and the prediction category is Negative. *FN* (False Negative) example, the real category is positive, and the prediction category is negative.

A. EXPERIMENTAL PREPARATION

1) DATA PREPARATION

Three kinds of data were selected from the UCI data set to verify the effectiveness of the NRDPSO algorithm proposed in this paper. The data description is shown in Table 2.

At the same time, in order to verify the real application effect of NRDPSO, the proposed method was applied to the production field of the petroleum industry to solve the problem of scaling prediction of the strong alkali ASP flooding production well.

The scaling related data of the strong alkali ASP flooding production well has the characteristics of large volume, high dimension, mixed type, dynamic nature and multiple data types. At the same time, in the process of water ion testing of the oilfield, water sampling is usually performed manually, and water quality data is obtained by manual and instrument collaboration. This process is often accompanied by problems such as unapproved data, uncalibrated test equipment, and test results subject to subjective human factors, which lead to incomplete ion test data. Ion test data contained fuzzy and clear data, symbolic and numerical data, lost and missing data. This scenario satisfies the research scenario of the paper.

A total of 3,337 scaling times from 2014 to 2018 were selected as the data set (ASP for short). The detailed description is shown in Table 3.

The experimental data contains 81 attributes, including 9 symbolic attributes and 69 numerical attributes. The experimental data consists of both complete and incomplete data; 70% of the experimental data were used as training samples and 30% as test samples. The experimental data attributes were numbered in the order of description and formed a set of attributes, then we got the set of condition attributes $C = \{a_1, a_2, a_3, a_4, \ldots, a_{81}\}$ and the set of decision attributes $D = \{d_1, d_2\}$. The hybrid incomplete decision table of the scaling prediction data is shown in Table 4.

2) NORMALIZATION PROCESSING

The scaling related data has mixed and incomplete characteristics and the data of different attributes have greatly different dimensions. Therefore, the experimental data were normalized by using the formula (7), and the attribute values were normalized to [0, 1], so as to reduce the effect of attribute dimension on classification accuracy. The decision table of the experimental data after normalization is shown in Table 5.

B. ENVIRONMENT AND PARAMETER SETTING

1) EXPERIMENTAL ENVIRONMENT

We adopted Windows Server 2008R2, CPU2.4GHz, memory 8GB, 64-bit operating system for the experiment, and Matlab2017b as the experimental platform.

2) PARAMETER SETTING

The parameter setting of NRDPSO algorithm is shown in Table 6.

We set the value range of the number of iterations as [100,1000], and the value step size as 50; the value range of the number of particles K is set as [5,50], and the growth step size as 5; the maximum and minimum flight speeds of particles are set as $v_{max} = 4, v_{min} = -4$; the acceleration constant is set according to the common approach as $c_1 = c_2 = 2.0$; the inertia coefficient is set as $\omega_{max} = 2$, $\omega_{min} = 0.5$.

TABLE 3. Description of the data set.

Data Name	Number of Attributes	Number of Symbolic Attributes	Number of Numerical Attributes	Complete or Not	Training Samples	Test Samples
Basic Data	20	4	13	YES	2337	1000
Geological Data	13	2	11	YES		
Well History Data	35	1	34	NO		
Ion Test Data	13	2	11	NO		

TABLE 4. Experimental data hybrid incomplete decision table.

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7		a_9	a_{10}	$\cdots$	a_{78}	a_{79}	a_{80}	a_{81}	d_1	d_2
1	—	641.55	1956.73	852.89	12.39	31.02	15.68	1723.25	5233.52	0.00	$\cdots$	94.92	1.6	8.64	11.85	Y	1
2	—	2214.74	570.41	922.52	11.82	49.30	11.96	2437.38	6218.12	28.83	$\cdots$	314.42	6.8	10.05	30.59	N	
3	421.00	4243.41	0.00	1081.23	74.45	0.00	0.00	4558.60	10378.69	31.38	$\cdots$	642.90	—	11.21	891.44	N	
4	149.35	4715.77	0.00	779.90	23.53	0.00	0.00	4333.43	10001.98	43.25	$\cdots$	619.60	—	10.76	768.57	N	
5	0.00	631.41	2781.96	984.80	24.02	30.06	12.16	2125.43	6589.84	10.18	$\cdots$	629.27	—	8.58	32.78	Y	3
6	790.28	4135.68	0.00	911.07	49.47	0.00	0.00	4853.00	10739.50	110.24	$\cdots$	662.00	—	11.46	1513.26	N	
7	430.86	5182.13	0.00	1036.20	94.62	0.00	0.00	5271.83	12015.64	100.06	$\cdots$	783.29	—	11.52	1027.11	N	
8	0.00	1689.26	2008.90	1017.77	69.64	43.69	20.54	2656.50	7506.30	82.26	$\cdots$	509.70	—	13.16	600.30	Y	2
9	—	647.02	1435.19	885.90	108.07	15.03	6.08	1634.61	4731.89	—	$\cdots$	29.05	1.0	8.55	—	N	
10	—	251.60	2462.03	765.86	35.73	44.73	6.03	1593.44	5159.43	1.70	$\cdots$	274.36	—	9.11	—	Y	3
11	—	660.46	2462.03	696.24	23.82	64.61	6.03	1829.83	5743.03	—		327.92	—	8.5	—	N	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

TABLE 5. Experimental data normalization decision table.

U	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	$\cdots$	a_{78}	a_{79}	a_{80}	a_{81}	d_1	d_2
1	—	0.0497	0.1951	0.372	0.0029	0.3	0.3846	0.0486	0.0577	0.00	$\cdots$	0.0629	0.048	0.1357	0.00	Y	1
2	—	0.1842	0.0569	0.4225	0.001	0.4767	0.2934	0.1106	0.0876	0.0815	$\cdots$	0.2368	0.464	0.3809	0.0032	N	
3	0.1657	0.3575	0.00	0.5375	0.211	0.00	0.00	0.2947	0.2139	0.0887	$\cdots$	0.4971	—	0.5826	0.1494	N	
4	0.0588	0.3979	0.00	0.3191	0.0402	0.00	0.00	0.2751	0.2025	0.1223	$\cdots$	0.4786	—	0.5043	0.1285	N	
5	0.00	0.0488	0.2774	0.4676	0.0419	0.2907	0.2983	0.0835	0.0988	0.0288	$\cdots$	0.4863	—	0.1252	0.0036	Y	3
6	0.3111	0.3483	0.00	0.4142	0.1272	0.00	0.00	0.3202	0.2249	0.3117	$\cdots$	0.5122	—	0.6261	0.255	N	
7	0.1696	0.4378	0.00	0.5049	0.2786	0.00	0.00	0.3565	0.2636	0.283	$\cdots$	0.6083	—	0.6365	0.1724	N	
8	0.00	0.1392	0.2003	0.4915	0.1948	0.4225	0.5038	0.1296	0.1267	0.2326	$\cdots$	0.3915	—	0.9217	0.1	Y	2
9	—	0.0502	0.1431	0.3959	0.3237	0.1453	0.1491	0.0409	0.0424	—	$\cdots$	0.0107	0.00	0.12	—	N	
10	—	0.0164	0.2455	0.3089	0.0811	0.4326	0.1479	0.0374	0.0554	0.0048	$\cdots$	0.2051	—	0.2174	—	Y	3
11	—	0.0513	0.2455	0.2585	0.0412	0.6248	0.1479	0.0579	0.0731	—		0.2475	—	0.1113	—	N	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

TABLE 6. Parameter setting of NRDPSO algorithm.

Parameter	I	K	v_{max}	v_{min}	c_1	c_2	ω_{max}	ω_{min}	w_c	w_s	w_r
Parameter value	[0,100]	[5,50]	4	-4	2.0	2.0	2	0.5	0.6	0.3	0.1

The weight coefficient of the fitness function is set as $w_r = 0.1$, $w_c = 0.6$, $w_s = 0.3$.

C. DESCRIPTION OF THE COMPARATIVE EXPERIMENT

DPSO and PSO were selected as the comparative algorithm, and the classification ability of the algorithm was verified using CART, NEC and SVM classifiers. The classification accuracy with the 10-fold cross method was used to evaluate the quality of feature selection. We define that $N1$ represents the number of original features; $N2$ the number of selected features; $Accuracy1$ the classification accuracy of the original features; $Accuracy2$ the classification accuracy of the selected feature subset.

1) CONTRAST EXPERIMENT OF NRDPSO BASED ON DIFFERENT NEIGHBORHOOD THRESHOLD VALUES

First, the neighborhood threshold δ was defined as a fixed value, which was applied to NRDPSO. The value of δ changed from 0 to 1 in steps of 0.05, and the number of feature selections and the corresponding classification accuracy δ were compared with the changes in the neighborhood threshold. Figure 3 (a) identifies the relation between the value of δ and the number of attributes. Figures 3(b), 4(c), 4(d), and 4(e) indicate the corresponding classification accuracy changes of the Iono, Sonar, Wine, and ASP data sets under the CART, NEC, and SVM classifiers, with the changes of δ.

Figure 3 shows that when $\delta < 0.1$ or $\delta > 0.75$, a small number of features is obtained by the NRDPSO reduction algorithm, and the corresponding classification accuracy is relatively low; when the value of δ is between [0.15, 0.4], the number of selected features is relatively optimal, and the classification accuracy, recall and precision are relatively high. This proves that the value of the neighborhood threshold δ determines the number of feature selection and the classification accuracy.

Secondly, the dynamic threshold value method proposed in literature [19] and [20] and the variable threshold definition method proposed in this paper were selected and applied to NRDPSO to perform the attribute reduction and compare with the classification accuracy of the original data.

TABLE 7. The number of features and classification accuracy based on the neighborhood threshold value method in literature [19].

Data	Feature		CART		NEC		SVM	
	N1	N2	Accuracy1	Accuracy2	Accuracy1	Accuracy2	Accuracy1	Accuracy2
Iono	34	10	0.8755±0.0693	0.9026±0.0127	0.8921±0.0182	0.8975±0.0050	0.9379±0.0508	0.8803±0.0078
Sonar	60	11	0.7207±0.1394	0.7337±0.0340	0.7968±0.0839	0.8226±0.0092	0.8510±0.0949	0.7721± .0223
Wine	13	3	0.8986±0.0635	0.9371±0.0102	0.9453±0.0621	0.9586±0.0121	0.9889±0.0234	0.9899±0.0024
ASP	81	28	0.7309±0.2342	0.7834±0.0987	0.8134±0.0367	0.8624±0.0387	0.8627±0.0874	0.8821±0.0765

TABLE 8. The number of features and classification accuracy based on the neighborhood threshold value method in literature [20].

Data	Feature		CART		NEC		SVM	
	N1	N2	Accuracy1	Accuracy2	Accuracy1	Accuracy2	Accuracy1	Accuracy2
Iono	34	9	0.8755±0.0693	0.9074±0.0396	0.8921±0.0182	0.8995±0.0610	0.9379±0.0508	0.9176±0.0483
Sonar	60	9	0.7207±0.1394	0.7520±0.0673	0.7968±0.0839	0.8220±0.0135	0.8510±0.0949	0.8374±0.0825
Wine	13	4	0.8986±0.0635	0.9065±0.0395	0.9453±0.0621	0.9521±0.0112	0.9889±0.0234	0.9860±0.0481
ASP	81	25	0.7309±0.2342	0.7583±0.0421	0.8134±0.0367	0.8261±0.0196	0.8627±0.0874	0.8719±0.0643

TABLE 9. The number of features and classification accuracy based on variable thresholds.

Data	Feature		CART		NEC		SVM	
	N1	N2	Accuracy1	Accuracy2	Accuracy1	Accuracy2	Accuracy1	Accuracy2
Iono	34	8	0.8755±0.0693	0.9113±0.0146	0.8921±0.0182	0.9198±0.0365	0.9379±0.0508	0.9435±0.0152
Sonar	60	10	0.7207±0.1394	0.7597±0.0583	0.7968±0.0839	0.8492±0.0197	0.8510±0.0949	0.8906±0.0998
Wine	13	3	0.8986±0.0635	0.9412±0.0138	0.9453±0.0621	0.9514±0.0218	0.9889±0.0234	0.9898±0.0531
ASP	81	23	0.7309±0.2342	0.7981±0.0687	0.8134±0.0367	0.8321±0.0119	0.8627±0.0874	0.9019±0.0762

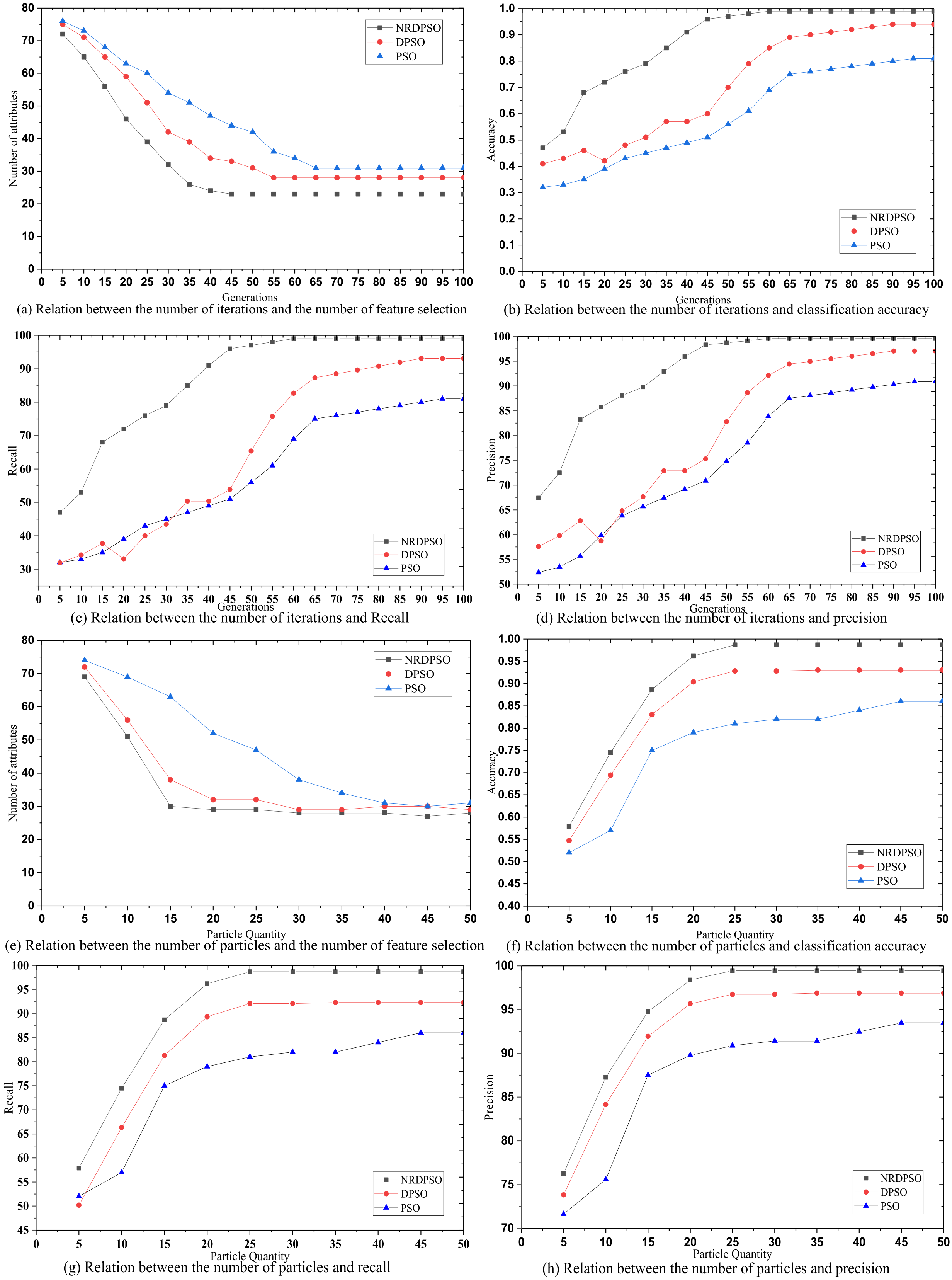
Table 7 exhibits the comparison of the number of features and the classification accuracy using the threshold value method proposed in the literature [19] with the original data. Table 8 shows the number of features and the classification accuracy obtained after attribute reduction by the dynamic threshold value method proposed in the literature [20]. Table 9 provides the number of features and the classification accuracy after attribute reduction by the variable threshold feature selection approach proposed in this paper.

According to Tables 7-9, the three algorithms have effectively reduced the number of features. By contrast, the number of features obtained by the method in literature [20] is small and the classification accuracy is relatively high. The classification accuracy obtained by the method in literature [19] is also relatively high, but the number of features after reduction is relatively large. The variable-precision neighborhood threshold value method proposed in this paper not only obtains the highest classification accuracy, but also has the smallest number of features after reduction.

2) CONTRAST EXPERIMENT OF FEATURE SELECTION BY IFDS WITH VARIABLE THRESHOLDS BASED ON DIFFERENT HEURISTIC ALGORITHMS

Aiming at the proposed IFDS method with variable thresholds, the DPSO algorithm with improved weight coefficients and the traditional DPSO and PSO heuristic algorithms were used for attribute reduction of ASP data to verify the search performance of the algorithm used in this paper. Figure 4 shows the changes in the number of feature selection and classification accuracy with the changes in the number of iterations and the number of particles, respectively.

According to the experimental results of the number of iterations, the following conclusions are obtained: As shown in Figure 4(a), the three algorithms of NRDPSO, DPSO and PSO converge and obtain the optimal solution after the 45th, 55th and 65th iterations, respectively. While the optimal solution quality and the convergence speed of NRDPSO are better than that of DPSO and PSO; as the number of iterations increases, the classification accuracy obtained by the NRDPSO algorithm is 98.9%, the recall is 99%, and the precision is 99.57%, which is significantly better than that obtained by the other two, as shown in Figure 4(b), 4(c), 4(d). Meanwhile, the recall of the NRDPSO algorithm is 98.7%, as shown in Figure 4(g). The precision is 99.45%, as shown in Figure 4(h). The experimental results of analyzing the number of different particles are summarized as follows: The two algorithms will quickly obtain the optimal solution when the number of particles is large enough, as shown in Figure 4(e).



(a) Relation between the number of iterations and the number of feature selection

(b) Relation between the number of iterations and classification accuracy

(c) Relation between the number of iterations and Recall

(d) Relation between the number of iterations and precision

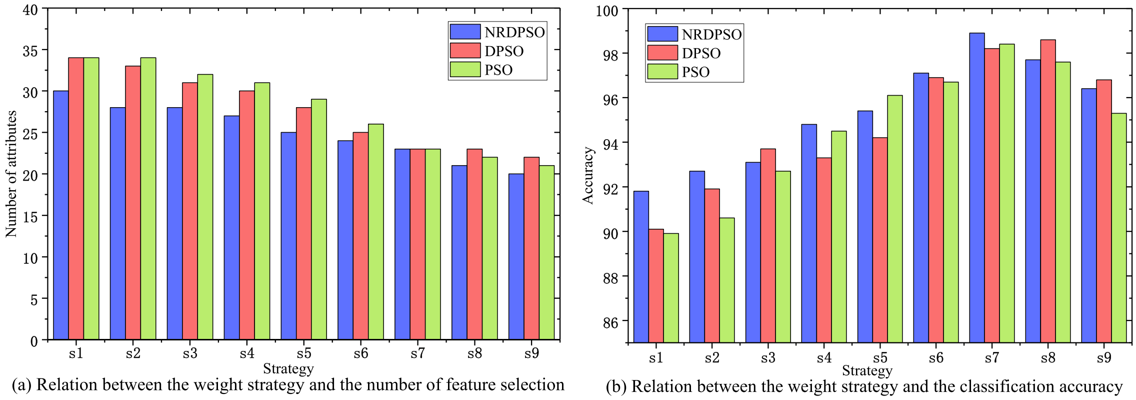
(e) Relation between the number of particles and the number of feature selection

(f) Relation between the number of particles and classification accuracy

(g) Relation between the number of particles and recall

(h) Relation between the number of particles and precision

FIGURE 4. Performance comparison of three algorithms under different number of iterations and number of particles.



(a) Relation between the weight strategy and the number of feature selection

(b) Relation between the weight strategy and the classification accuracy

FIGURE 5. Performance comparison of three algorithms under different weight strategies.

which proves that the more the number of particles, the faster the optimal solution is updated. At the same time, it is further verified that the optimal solution quality of NRDPSO is always slightly superior to that of DPSO and PSO, as shown in Figures 4(c) and 4(d).

3) EXPERIMENT ON THE INFLUENCE OF DIFFERENT WEIGHT COEFFICIENTS ON THE OPTIMAL FEATURE SUBSET

With ASP as the experimental data, the number of features and the classification accuracy of features selected by NRDPSO algorithm were compared and analyzed under different weight coefficient values, to analyze the influence of different weight coefficients on the quality of feature subsets. During the search of feature subsets, the most important objective is to minimize the number of features obtained in the optimal feature subset and obtain the highest feature quality. Therefore, in the experiment, we set $w_r = 0.1$, the weight coefficient value range of w_c, w_s as [0.1,0.8], and $w_c + w_s + w_r = 1$, the values of w_c, w_s given are shown in Table 10. At the same time, the iteration coefficient is set as $I = 100$, $K = 25$.

Figure 5 shows the influence of the weight strategy of fitness function on the number of feature selection and the classification accuracy of feature selection.

Figure 5(a) shows that when the strategy is S7, the number of feature selection obtained reaches the optimal number of 23, and when the strategy is larger than S7, the number of feature selected by the objective gradually decreases; when the strategy is less than S7 and larger than S1, the number of features selected by the objective gradually decreases and approaches to the optimal number of features. Figure 5(b) indicates that when the strategy is S7, the feature subset obtained by the objective has the optimal classification accuracy, the highest of the three algorithms is 98.9%; as the strategy changes, the classification accuracy of the feature subset selected by the objective is inversely proportional to the number of feature subsets.

4) COMPARATIVE EXPERIMENTS ON EVALUATION INDEXES OF DIFFERENT FEATURE SELECTION ALGORITHMS

In the experiment, the algorithm proposed in this paper (IDFS-VTRS), the Dynamic variable precision rough set model of mixed information system (MIS-DPRS) proposed in the literature [29], the feature selection method adopted the binary particle swarm optimization based on the mutation operator [30], best first forward search (BFFS) and best first backward search (BFBS) algorithm provided by Weka were used to select the characteristics of the test data. The classification and evaluation indexes of each algorithm are shown in table 11.

TABLE 10. Values of w_c, w_s.

Strategy	w_c	w_s	Strategy	w_c	w_s
S1	0.1	0.8	S6	0.5	0.4
S2	0.2	0.7	S7	0.6	0.3
S3	0.3	0.6	S8	0.7	0.2
S4	0.4	0.5	S9	0.8	0.1
S5	0.45	0.45			

Table 11 shows that the evaluation indexes of Accuracy, Recall and Precision of the IFDS-VTRS algorithm are the best among the four-test data. The optimal value of Accuracy is 98.7%, and the optimal amount of Precision is 99.5%, which proves the effectiveness of the proposed algorithm. The Precision of the MIS-DPRS algorithm is lower than MBPSO-FS in the Wine test data but higher than MBPSO-FS in other test data and Evaluation indexes, which proves that MIS-DPRS algorithm is better than MBPSO-FS. The evaluation indexes of BFFS and BFBS are lower than the other three algorithms.

Through theoretical description and experimental verification, the following conclusions are drawn:

Conclusion 1: The value of the neighborhood threshold δ determines the number and classification accuracy of feature selection.

Conclusion 2: The NRDPSO algorithm based on variable thresholds has a higher classification accuracy and a relatively small number of features when solving the problem of feature selection in complex environments.

Conclusion 3: As the weight strategy changes, the classification accuracy of the feature subset selected by the objective is inversely proportional to the number of feature subsets.

TABLE 11. Evaluation index of feature selection algorithm.

Data	Algorithm	Accuracy	Recall	Precision
Iono	IFDS-VTRS	98.6	98.2	99.5
	MIS-DPRS	96.5	96.9	97.8
	MBPSO-FS	95.2	96.4	97.5
	BFFS	91.7	92.9	94.1
	BFBS	90.8	92.4	93.3
Sonar	IFDS-VTRS	98.5	98.2	99.1
	MIS-DPRS	95.6	95.5	96.4
	MBPSO-FS	94.7	93.9	95.6
	BFFS	91.3	91.9	91.9
	BFBS	90.3	89.2	92.5
Wine	IFDS-VTRS	98.7	98.5	99.2
	MIS-DPRS	97.1	97.7	98.4
	MBPSO-FS	96.3	96.7	98.7
	BFFS	91.5	94.6	93.9
	BFBS	89.8	91.5	94.4
ASP	IFDS-VTRS	97.3	97.3	98.8
	MIS-DPRS	96.8	96.8	98.6
	MBPSO-FS	97.1	97.4	97.9
	BFFS	94.6	94.9	97.3
	BFBS	94.5	94.7	97.4

VI. CONCLUSION

A feature selection approach based on IFDS and DPSO with variable thresholds was proposed to solve the problem of feature selection in complex environments in this paper.

The average reachable distance was introduced to construct the attribute neighborhood threshold set, and constructed the DPSO particle fitness function using the feature subset length, the significance of the attribute and the negative domain of the neighborhood, and improved the inertia weight computing method, so as to enhanced the feature selection speed and the feature subset quality. Furthermore, we demonstrate the application of this method to feature selection of real industrial production data. Experimental analysis shows that the feature selection method based on variable threshold IFDS and DPSO can screen out the optimal feature subset, maintain or even significantly improve the classification accuracy of reduced data, and verify the effectiveness of the method.

REFERENCES

[1] Z. Pawlak, “Rough sets and intelligent data analysis,” *Inf. Sci.*, vol. 147, nos. 1–4, pp. 1–12, Nov. 2002.

[2] J. Zhang, J.-S. Wong, Y. Pan, and T. Li, “A parallel matrix-based method for computing approximations in incomplete information systems,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 326–339, Feb. 2015.

[3] W.-Z. Wu, Y. Qian, T.-J. Li, and S.-M. Gu, “On rule acquisition in incomplete multi-scale decision tables,” *Inf. Sci.*, vol. 378, pp. 282–302, Feb. 2017.

[4] T. Jing, F. Long, and X. Shi, “Research of Aircraft Integrated Drive Generator fault diagnostic decision based on attribute reduction in rough sets,” in *Proc. IEEE Int. Conf. Aircr. Utility Syst. (AUS)*, Oct. 2016, pp. 393–397.

[5] J. Hongqiang, L. Jia, W. Ding, and N. Li, “Subjective logic application research—Rough set belief rule model and trust evaluation model of the intrusion detection system,” *Optik*, vol. 126, no. 24, pp. 5593–5599, 2015.

[6] A. Banerjee and P. Maji, “Rough sets and stomped normal distribution for simultaneous segmentation and bias field correction in brain MR images,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5764–5776, Dec. 2015.

[7] Y. Yang, D. Chen, H. Wang, and X. Wang, “Incremental perspective for feature selection based on fuzzy rough sets,” *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1257–1273, Jun. 2018.

[8] C. Wang, Y. Qi, M. Shao, Q. Hu, D. Chen, Y. Qian, and Y. Lin, “A fitting model for feature selection with fuzzy rough sets,” *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 741–753, Aug. 2017.

[9] Q. Hu, D. Yu, J. Liu, and C. Wu, “Neighborhood rough set based heterogeneous feature subset selection,” *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, 2008.

[10] X. Ma and D. Li, “A hybrid fault diagnosis method based on fuzzy signed directed graph and neighborhood rough set,” in *Proc. 6th Data Driven Control Learn. Syst. (DDCLS)*, May 2017, pp. 253–258.

[11] M. Suo, M. Zhang, D. Zhou, B. Zhu, and S. Li, “Fault diagnosis of satellite power system using variable precision fuzzy neighborhood rough set,” in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 7301–7306.

[12] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, “Feature subset selection based on fuzzy neighborhood rough sets,” *Knowl.-Based Syst.*, vol. 111, pp. 173–179, Nov. 2016.

[13] M. Kryszkiewicz, “Rough set approach to incomplete information systems,” *Inf. Sci.*, vol. 112, nos. 1–4, pp. 39–49, Dec. 1998.

[14] J. Stefanowski and A. Tsoukiàs, “On the extension of rough sets under incomplete information,” in *Proc. Int. Workshop Rough Sets, Fuzzy Sets, Data Mining, Granular-Soft Comput.*, 1999, pp. 73–81.

[15] G. Wang, “Extension of rough set under incomplete information systems,” in *Proc. IEEE World Congr. Comput. Intell. IEEE Int. Conf. Fuzzy Syst.*, May 2020, pp. 1098–1103.

[16] J. Dai, “Rough set approach to incomplete numerical data,” *Inf. Sci.*, vol. 241, pp. 43–57, Aug. 2013.

[17] H. Zhao and K. Qin, “Mixed feature selection in incomplete decision table,” *Knowl.-Based Syst.*, vol. 57, pp. 181–190, Feb. 2014.

[18] B. T. Zhao, X. J. Chen, and Q. S. Zeng, “Approach for incomplete fuzzy hybrid decision system on neighborhood rough set,” *J. Jilin Univ. (Eng. Technol. Ed.)*, vol. 41, no. 3, pp. 721–727, 2011.

[19] W. Li, Z. Huang, X. Jia, and X. Cai, “Neighborhood based decision-theoretic rough set models,” *Int. J. Approx. Reasoning*, vol. 69, pp. 1–17, Feb. 2016.

[20] D.-W. Zhang, P. Wang, J. Qiu, and Y. Jiang, “An improved approach to feature selection,” in *Proc. Int. Conf. Mach. Learn. Cybern.*, Jul. 2010, vol. 5, no. 32, pp. 488–493.

[21] I. A. Gheyas and L. S. Smith, “Feature subset selection in large dimensionality domains,” *Pattern Recognit.*, vol. 43, no. 1, pp. 5–13, 2010.

[22] H. H. Inbarani, A. T. Azar, and G. Jothi, “Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis,” *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 175–185, Jan. 2014.

[23] L. Cervante, B. Xue, and L. Shang, “A multi-objective feature selection approach based on binary pso and rough set theory,” in *Proc. Eur. Conf. Evol. Comput. Combinat. Optim.* Berlin, Germany: Springer, 2013, pp. 25–36.

[24] A. Stevanovic, B. Xue, and M. Zhang, “Feature selection based on PSO and decision-theoretic rough set model,” in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2013, pp. 2840–2847.

[25] M. Adamczyk, “Parallel feature selection algorithm based on rough sets and particle swarm optimization,” in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2014, pp. 43–50.

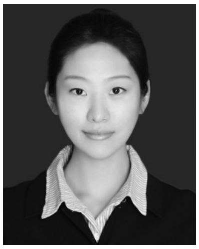
[26] W. Lu, Q. Tao-Rong, H. Niu, and L. Ping, “A method for feature selection based on rough sets and ant colonyoptimization algorithm,” *J. Nanjing Univ. (Natural Sci.)*, vol. 46, no. 5, pp. 487–493, 2010.

[27] S. T. Chen, G. L. Chen, W. Z. Guo, and Y. H. Liu, “Feature selection of the intrusion detection data based on particle swarm optimization and neighborhood reduction,” *J. Comput. Res. Develop.*, vol. 47, no. 7, pp. 1261–1267, 2010.

[28] D. R. Wilson and T. R. Martinez, “Improved heterogeneous distance functions,” *J. Artif. Intell. Res.*, vol. 6, pp. 1–34, Jan. 1997.

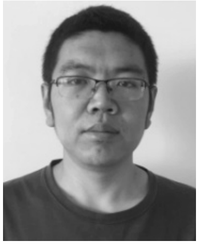
[29] Z. Yang and B. Z. Qiu, "Dynamic variable precision rough set model of mixed information system," *Control Decis.*, vol. 35, no. 2, pp. 297–308, 2020.

[30] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowl.-Based Syst.*, vol. 64, pp. 22–31, Jul. 2014.



YANAN HU received the B.S. degree in educational technology from Jiamusi University, China, in 2014, and the M.S. degree in educational technology from Northeast Petroleum University, China, in 2017, where she is currently pursuing the Ph.D. degree in geological resources and geological engineering. Her research interests include multiagent systems, swarm intelligence, and feature selection.

CHUNSHENG LI received the Ph.D. degree in computer application from the University of Technology Sydney, Sydney, in 2005. He is currently working as a Professor with the School of Computer and Information Technology, Northeast Petroleum University, China, where he is also the Dean of the School of Computer and Information Technology. His research interests include multiagent system methodology and data mining.



KEJIA ZHANG received the Ph.D. degree in petroleum engineering computing technology from Northeast Petroleum University, China, in 2016. He is currently working as a Teacher with the School of Computer and Information Technology, Northeast Petroleum University. His research interests include multiagent systems, mechanism design, and cooperative game. He is a member of CCF.

MEI WANG was born in 1976. She is currently a Professor. Her main research interests include machine learning, model selection, and kernel methods. She has received the funding from the National Natural Science Foundation of China (51774090), in 2017.

YATIAN GAO was born in 1979. She is currently an Associate Professor with the School of Computer and Information Technology, Northeast Petroleum University, China. Her research interests include big data and data mining.