

RESEARCH ARTICLE

Attribute Reduction Algorithm for Incomplete Information Systems Based on Intuitive Fuzzy Pairs

WEIHAN LI¹ AND JIANWEI GUO² ¹Haidian District, Beijing 100097, China²School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Jianwei Guo (19114023@bjtu.edu.cn)


ABSTRACT The current attribute reduction algorithms for information systems are difficult to handle imbalanced data with default values. Therefore, to address the shortcomings of traditional attribute reduction algorithms (ARAs) in incomplete information systems, a new algorithm is proposed by introducing intuitive fuzzy pairs (IFP). In addition, a composite minority oversampling technique TampC and Central Limit SMOTE (TampC-CL-SMOTE) is proposed to improve the pre-data sampling method of the algorithm, and its effectiveness is verified by experiments. The experimental results show that the average classification accuracy of the improved attribute reduction algorithm on the naive Bayes classifier is 82.13%, and the average classification accuracy on the support vector machine classifier is 86.48%. In the comparison of operational efficiency, the average running time of the improved attribute reduction algorithm is 5.92 seconds, and the overall consumption of running time is lower than that of the comparison algorithm. Meanwhile, the average accuracy, recall, and F-measure of the algorithm are 76.14%, 78.35%, and 77.19%, respectively. In addition, the G-means of TampC-CL-SMOTE are 2.9% and 5.3% higher than the comparison algorithm, respectively. Overall, the improved attribute reduction algorithm has high efficiency in handling imbalanced data, while the optimization of TampC-CL-SMOTE has effectiveness in practical applications and has advantages in handling high and low imbalanced data in incomplete information environments.

INDEX TERMS Intuitive fuzzy pairs, incomplete information, attribute reduction algorithm, imbalanced data.

I. INTRODUCTION

The development of network technology has led to explosive data growth in various industries. At the same time, the objective impact of noise in real life leads to many incomplete and difficult to determine data in current information systems. To efficiently mine the hidden rules behind data, fuzzy rough set theory is proposed. It is based on the existing knowledge base and uses the form of upper and lower approximation sets to characterize certain uncertain knowledge, ensuring classification performance without requiring additional data [2]. The attribute reduction algorithm means that the classification quality of the related attribute set after attribute reduction

is the same as that of the original attribute set. The goal is to discover some necessary conditional attributes from the set of small item attributes, and based on these conditional attributes, form a classification relative to the decision attribute, which is consistent with the classification relative to the decision attribute formed by all conditional attributes, that is, have the same classification ability as all conditional attributes relative to the decision attribute. Attribute reduction, as an important achievement in fuzzy rough set theory, has gradually been deeply studied in both theoretical and practical development. Attribute reduction refers to the fact that the classification quality of the reduced attribute set is the same as that of the original attribute set. Simply put, attribute reduction is the minimum conditional subset of attributes that does not contain redundant attributes and ensures the

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

correct classification of the information system [3]. In the intelligent processing of decision table information, attribute reduction has gained high practical application value due to this idea. The goal of summarizing attribute reduction in practical applications is to discover some necessary conditional attributes from the set of conditional attributes, and based on these conditional attributes, the classification relative to the decision attribute formed is consistent with the classification relative to the decision attribute formed by all conditional attributes [4], [5]. However, the current ARAs are mainly applied to complete and general data, and there is relatively little research on incomplete or unbalanced data. Meanwhile, the introduction of intuitionistic fuzzy pairs as important attribute values can effectively handle missing and missing values in incomplete intuitionistic fuzzy information systems. At the same time, the current research on attribute reduction algorithms for incomplete information systems has some shortcomings, such as high Time complexity, difficulty in handling unbalanced data with default values, and can only deal with unbalanced data under the condition of complete information. However, in current real life, many cases are under incomplete information conditions, so attribute reduction algorithms for imbalanced data in incomplete information systems urgently need to be improved. Based on this, this paper studies the use of intuitionistic fuzzy pairs to optimize attribute reduction algorithms under incomplete information systems, and proposes a sampling method of the composite minority oversampling technique (TampC and Central Limit SMOTE, TampC-CL-SMOTE) with temporary tags and central limits. Its purpose is to solve the problems of high time complexity and difficulty in handling imbalanced data with default values in current incomplete information system ARAs, and to optimize the shortcomings of the proposed reduction algorithm in handling high imbalanced data.

II. RELATED WORKS

In fuzzy rough set theory, ARAs have always been one of its key research directions, and have high practical application significance in many fields [6]. At the same time, not every attribute is essential in a studied information system, as there are more or less redundant attributes. To obtain non redundant data, attribute reduction is necessary, which is of great significance for the promotion of theory [7]. In addition, some unavoidable errors in daily life can result in incomplete data information, including data loss caused by excessive costs, data loss caused by transmission errors, and data loss caused by inadequate data understanding, resulting in incomplete information systems. It has high practical application significance in many fields. For example, He et al. proposed an attribute reduction algorithm for an incomplete classification decision information system on the basis of fuzzy rough set, aiming at the related problems in attribute selection in rough set theory, thus effectively improving the accuracy of classification data [8]. Song et al. proposed corresponding measurement tools based on attribute similarity to address

the application of uncertainty measurement in attribute reduction. And its effectiveness was verified using the K-means algorithm, providing data support for the promotion and development of set-valued information systems [9]. Bar and Prasad proposed another optimal ARA using the nearest neighbor method to address the complexity of the coarsest granularity index, effectively reducing space utilization and corresponding computational time [10]. Tsai and Hu proposed an ARA for mixed data based on the comparison of different supervised learning technologies, aiming at the related problems when interpolating missing values under the condition of missing attribute values, so as to improve the accuracy of interpolation results and classification accuracy [11]. To improve the quality of teaching in online learning, Han et al. proposed granular adaptive computing techniques based on ARAs, which effectively improved the accuracy of anomaly detection for online learners while strengthening attribute feature selection [12].

In addition, Li et al. proposed an attribute selection method for incomplete interval valued information systems on the basis of incomplete interval data analysis to solve the related problems of attribute selection of information systems in big data processing. Based on this, a corresponding reduction algorithm was proposed, which effectively strengthens the accuracy of attribute selection and improves the efficiency [13]. Karimi and Yahyazade proposed an ARA related to deviation risk based on rough set theory to address the risk issues related to project management, thereby providing assistance in predicting project risk levels [14]. Xu et al. proposed corresponding ARAs for feature selection in fuzzy neighborhood rough set models by utilizing fuzzy neighborhood self information measures, effectively improving the accuracy of classification [15]. Chen et al. proposed a new integrated selector based on attribute selection analysis to solve the problems related to attribute reduction in data dimensionality reduction, thus effectively enhancing the stability in attribute search [16]. Ding et al. proposed a new attribute reduction algorithm based on multi granularity super trust fuzzy rough set to solve the problem of poor knowledge extraction effect of traditional attribute reduction algorithm in incomplete information systems, which effectively solved the problem of Big data analysis and summary data mining [17]. Singh et al. proposed an attribute reduction algorithm for incomplete information systems based on tolerance rough set theory to address the related issues of attribute reduction algorithms in practical applications, effectively improving the prediction accuracy in practical applications [18].

From the research of domestic and foreign scholars, it can be seen that current attribute reduction algorithms have more content in processing complete and general data, while they have less content in processing incomplete and unbalanced data. At the same time, the current research on attribute reduction algorithms for incomplete information systems has some shortcomings, such as high Time complexity, difficulty in handling unbalanced data with default values, and can

only deal with unbalanced data under the condition of complete information. However, in current real life, many cases are under incomplete information conditions, so attribute reduction algorithms for imbalanced data in incomplete information systems urgently need to be improved. Therefore, the research on the use of intuitionistic fuzziness in the proposed attribute reduction algorithm for imbalanced data has certain innovation. It has important theoretical and practical significance for the processing of imbalanced data in incomplete information systems and the expansion of algorithm models. It will lay the foundation for the subsequent application of attribute reduction algorithms in image retrieval and text classification fields, and the introduction of TampC-CL-SMOTE also provides a reference for the processing of high imbalanced data.

III. ANALYSIS OF ARAS FOR INCOMPLETE INFORMATION SYSTEMS UNDER IFP

A. FUZZY ROUGH SET MODEL IN INCOMPLETE INFORMATION SYSTEMS

To solve the problems of high time complexity and difficulty in handling imbalanced data with default values in current incomplete information system ARAs, an improved ARA is proposed based on the research of incomplete information systems with IFP of attributes. The traditional rough set is based on the equivalence relation, so it is not strong in processing noisy data. Many scholars have proposed fuzzy rough set related models and methods for processing imbalanced data to address this issue. However, in practical engineering applications, due to the influence of noise, incomplete information systems are mostly incomplete information systems, and their internal data is in an imbalanced state. Therefore, it is necessary to use relevant methods to process the imbalanced data under the incomplete fuzzy rough set system model [19]. Previous studies have developed a neighborhood fuzzy rough set model, but its applicability is limited to information systems under complete information conditions. Therefore, the research extends the practical process of processing imbalanced data to the fuzzy rough set model applicable to incomplete information systems based on it. Based on this, the study elaborates on the model with two definitions. The first definition is to consider the data attribute set B in the incomplete information system $I = (M, N, \lambda T, E)$, which contains the attribute set λ , while λ is the union of λ^k and λ^b . Among them, B^k contains λ^k , B^b contains λ^b , and any attributes p_i and p_j belong to the set M in incomplete information systems. Therefore, if any attribute belongs to λ^b , the existing equation is shown in equation (1).

$$\begin{aligned} \gamma_{\rho_b}(p_i, p_j) &= \begin{cases} 0, & E_{\rho_b}(p_i) = E_{\rho_b}(p_j) \vee E_{\rho_b}(p_i) = * \vee E_{\rho_b}(p_j) = * \\ 1, & \text{other} \end{cases} \end{aligned} \quad (1)$$

In equation (1), $\gamma_{\rho_b}(p_i, p_j)$ represents the actual distance between incomplete symbolic data objects; E_{ρ_b} represents the

numerical data of the set E . If any attribute belongs to λ^k , the existing equation is shown in equation (2).

$$\begin{aligned} \gamma_{\rho_k}(p_i, p_j) &= \begin{cases} \left| \hat{E}_{\rho_k}(p_i) - \hat{E}_{\rho_k}(p_j) \right|^2, & E_{\rho_k}(p_i) \neq * \wedge E_{\rho_k}(p_j) \neq * \\ 0, & E_{\rho_k}(p_i) \neq * \wedge E_{\rho_k}(p_j) = * \\ 0, & E_{\rho_k}(p_i) = * \wedge E_{\rho_k}(p_j) \neq * \\ 0, & E_{\rho_k}(p_i) = * \wedge E_{\rho_k}(p_j) = * \end{cases} \end{aligned} \quad (2)$$

In equation (2), $\gamma_{\rho_k}(p_i, p_j)$ represents the distance between incomplete numerical data objects. Based on the distance function defined in the first definition, a second definition can be proposed, which is related to the domain tolerance relationship, to provide data support for the subsequent construction of incomplete domain fuzzy rough set models. Based on this, the second definition assumes that λT is the union of the data attribute set B and the attribute set $\{f\}$, and B contains λ , taking into account the incomplete information system $I = (M, N, \lambda T, E)$. In this case, based on the first definition, the definition formula for the domain tolerance correlation between attributes and B in the relevant domain is shown in equation (3).

$$\begin{cases} Q_\lambda^\alpha = \{(p_i, p_j) \in L \mid \Delta_\lambda(p_i, p_j) \leq \alpha\} \\ [p_i]_\lambda^\alpha = \{p_j \mid (p_i, p_j) \in Q_\lambda^\alpha\} \end{cases} \quad (3)$$

Equation (3) defines the formula for domain tolerance correlation in the first row, and the formula for domain tolerance definition in the second row for attribute p_i . Among them, Q_λ^α represents the domain tolerance relationship; L representation theory; α represents the relevant domain radius under the neighborhood tolerance relationship of the incomplete mixed information system, and its value is maintained between [1, 0]. For the target approximation set, its upper and lower approximate expressions in incomplete mixed information systems are shown in equation (4).

$$\begin{cases} \underline{Q}_\lambda^\alpha(Y) = \{p_i \mid [p_i]_\lambda^\alpha \in Y, p_i \in L\} \\ \overline{Q}_\lambda^\alpha(Y) = \{p_i \mid [p_i]_\lambda^\alpha \cap Y \neq \emptyset, p_i \in L\} \end{cases} \quad (4)$$

In equation (4), $\underline{Q}_\lambda^\alpha(Y)$ represents the lower approximation of the target approximation set Y ; $\overline{Q}_\lambda^\alpha(Y)$ represents the upper approximation of the target approximation set Y . Therefore, the formula for defining the dependence of the positive and negative domains of Q_λ^α , boundary domains, and attribute set f on λ in the target approximate solution is shown in equation (5).

$$\gamma_{\rho_k} \begin{cases} S_\lambda^\alpha(Y) = \underline{Q}_\lambda^\alpha(Y) \\ G_\lambda^\alpha(Y) = L - \overline{Q}_\lambda^\alpha(Y) \\ U_\lambda^\alpha(Y) = \overline{Q}_\lambda^\alpha(Y) - \underline{Q}_\lambda^\alpha(Y) \\ \zeta_\lambda = \frac{|n_{Y \in L} S_\lambda^\alpha(Y)|}{|L|} \end{cases} \quad (5)$$

In equation (5), $S_\lambda^\alpha(Y)$ represents the positive domain of Y with respect to Q_λ^α ; $G_\lambda^\alpha(Y)$ represents the negative field of

Y regarding Q_λ^α ; $U_\lambda^\alpha(Y)$ represents the boundary domain of Y regarding Q_λ^α ; ζ_λ represents the dependency of attribute f on λ . Therefore, to reduce the impact of uncertainty in the boundary area, it is particularly important to use incomplete information systems to determine whether the boundary area is majority or minority type by defining the upper and lower boundaries of the boundary area. Based on this theory, the three definitions proposed in the study are shown in Figure 1.

From Figure 1, the first definition first considers a non holonomic information system $I = (M, N, \lambda T = B \cup \{f\}, E)$, and the upper and lower boundary domain definitions in the non holonomic neighborhood fuzzy rough set are expressed as shown in equation (6).

$$\begin{cases} \overline{\Delta}_\lambda^\alpha(Y) = \overline{Q}_\lambda^\alpha(Y) - Y \\ \underline{\Delta}_\lambda^\alpha(Y) = Y - \underline{Q}_\lambda^\alpha(Y) \end{cases} \quad (6)$$

In equation (6), $\overline{\Delta}_\lambda^\alpha(Y)$ represents the upper boundary domain of a non holonomic neighborhood fuzzy rough set; $\underline{\Delta}_\lambda^\alpha(Y)$ represents the lower boundary region of a non holonomic neighborhood fuzzy rough set. Due to space limitations, the study only considers the binary classification problem, that is $L/f = \{Y^+, Y^-\}$. Among them, Y^+ represents positive class (minority class), and Y^- represents negative class (majority class). Therefore, there is a degree of imbalance between positive and negative categories under the subjective question object in a non holonomic mixed information system, as expressed in equation (7).

$$I = |Y^+| / |Y^-| \quad (7)$$

To reduce the impact of boundary region uncertainty, if the boundary region of a certain object is being studied, the upper and lower boundaries of an imperfect neighborhood set can be defined as usual. The second definition considers a non holonomic information system $I = (M, N, \lambda T = B \cup \{f\}, E)$, where the upper and lower boundaries of the neighborhood tolerance class of an object in the attribute set B are expressed as shown in equation (8).

$$\begin{cases} \overline{\Delta}_\lambda^\alpha(Y^+) = \bigcup_{i=1}^{|L|} \overline{\Delta}_\lambda^{\alpha,i}(Y^+) \\ \underline{\Delta}_\lambda^\alpha(Y^+) = \bigcup_{i=1}^{|L|} \underline{\Delta}_\lambda^{\alpha,i}(Y^+) \end{cases} \quad (8)$$

In equation (8), $\overline{\Delta}_\lambda^\alpha(Y^+)$ represents the upper boundary of the neighborhood tolerance class of object p_i under the attribute set B ; $\underline{\Delta}_\lambda^\alpha(Y^+)$ represents the lower boundary of the neighborhood tolerance class of the object p_i in the attribute set B . To reduce the impact of positive categories on classification results under imbalanced data, the upper and lower bounds of the incomplete neighborhood set for each category are used as the estimation basis for classification errors. The third definition considers a non holonomic information system $I = (M, N, \lambda T = B \cup \{f\}, E)$, where the conditional attribute x_j belongs to the attribute set B and Y^+ belongs to the L . At this point, if negative object p_i^- has a high probability of misclassification, the flag value of negative object p_i^-

that may be misclassified under attribute x_j is 1; If the positive object p_i^+ has a high probability of misclassification, then the flag value that the positive object p_i^+ may be misclassified under attribute x_j is 1. Specifically expressed as two, assuming $\chi_i^\alpha = \frac{|\underline{\Delta}_{x_j}^{\alpha,i}(Y^+)|}{(|\underline{\Delta}_{x_j}^{\alpha,i}(Y^+)| + \pi |\overline{\Delta}_{x_j}^{\alpha,i}(Y^+)|)}$, if χ_i^α is greater than 0.5 and $E_f(p_i^-) = E_f(Y^-)$, the negative object p_i^- may be misclassified under attribute x_j with a flag value $\overline{k}_{x_j}^{\alpha,i}$ of 1, otherwise it is 0. Meanwhile, assuming $\varepsilon_i^\alpha = \pi \frac{|\overline{\Delta}_{x_j}^{\alpha,i}(Y^+)|}{(|\underline{\Delta}_{x_j}^{\alpha,i}(Y^+)| + \pi |\overline{\Delta}_{x_j}^{\alpha,i}(Y^+)|)}$, if ε_i^α is greater than 0.5 and $E_f(p_i^+) = E_f(Y^+)$, then the flag value $k_{x_j}^{\alpha,i}$ of the positive object p_i^+ that may be misclassified under attribute x_j is 1, otherwise it is 0. Among them, χ_i^α represents the probability that negative object p_i^- may be misclassified by x_j under attributes; ε_i^α represents the probability that the positive object p_i^+ may be misclassified by x_j under attributes; π represents the normative factor used to balance the uneven distribution between negative and positive classes. Based on this, based on the probability of misclassification, the expression of the number of misclassifications is shown in equation (9).

$$\begin{cases} d_j^- = \sum_{i=1}^{(L)} \overline{k}_{x_j}^{\alpha,i} \\ d_j^+ = \sum_{i=1}^{(L)} k_{x_j}^{\alpha,i} \end{cases} \quad (9)$$

In equation (9), d_j^- represents the actual number of negative object sets misclassified on attribute x_j ; d_j^+ represents the actual number of misclassified positive object sets on attribute x_j .

B. ARA FOR INCOMPLETE INFORMATION SYSTEMS BASED ON IFP

Based on the proposed fuzzy rough set mathematical model in incomplete information systems, an ARA for incomplete information systems is constructed by utilizing IFP. Firstly, the importance of attributes in incomplete intuitive fuzzy information systems is defined based on the conventional methods used in previous literature to handle imbalanced data in complete information systems. The importance of this attribute largely reflects the uncertainty of boundary domains and the imbalance of data [20]. Assuming that the conditional attribute x_j belongs to the attribute set B , the definition and expression of the importance of the attribute are shown in equation (10).

$$\varphi_j = 1 - \frac{\mu \frac{d_j^+}{|Y^+|} + (1 - \mu) \frac{d_j^-}{|Y^-|}}{2} \quad (10)$$

In equation (10), φ_j represents the importance of the conditional attribute x_j ; μ represents the degree of unequal distribution between positive and negative groups. Based on this, the correlation matrix expression of the importance of the attribute set can be $T_\varphi = [\lambda t \ \varphi]$, and since λt is the transpose matrix of the conditional attribute x_j combination,

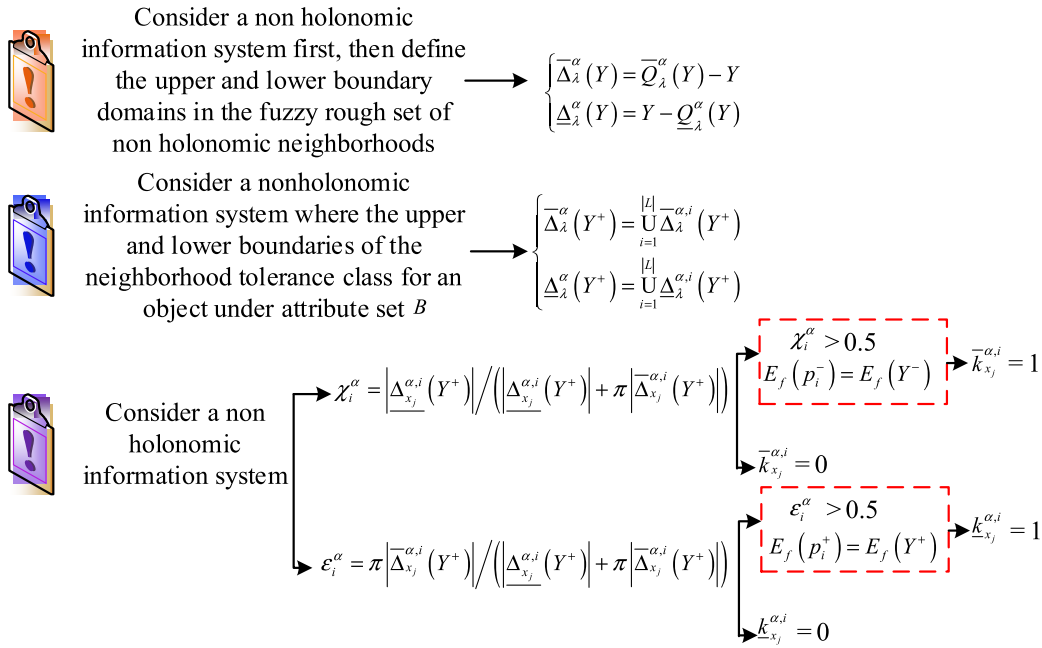


FIGURE 1. Three definitions for solving unbalanced data in incomplete hybrid information systems.

the ordered matrix expression of all attributes can be $\lambda t^\succ = (x_i, x_j, \dots, x_r)^T$, in which case $\varphi_i > \varphi_j$. In the information system of incomplete intuitive fuzzy decision-making, the use of discernibility matrix can be chosen. Assuming that the expression of the incomplete intuitive fuzzy decision system is $I = (M, N, \lambda T = B \cup \{f\}, E)$, then the conventional decision of the incomplete neighborhood class $[p_i]_\lambda^\alpha$ of any object p_i can be expressed as $D_i = \{E_f(p_j) \mid \forall p_j \in [p_i]_\lambda^\alpha\}$.

In addition, in the given incomplete intuitive fuzzy decision information system $I = (M, N, \lambda T = B \cup \{f\}, E)$, if the object p_i does not belong to $[p_i]_\alpha$ and D_i is not equal to D_j , the corresponding equation expression is shown in equation (11).

$$\begin{cases} \text{if } x_k \in B^n, g_{st} = \{x_k \mid |\hat{E}_{x_k}(x_i) - \hat{E}_{x_k}(x_j)| > f\} \\ \text{if } x_k \in B^b, g_{st} = \{x_k \mid \hat{E}_{x_k}(x_i) \neq \hat{E}_{x_k}(x_j)\} \end{cases} \quad (11)$$

In equation (11), B^n and B^b represent subsets of the attribute set B ; g_{st} represents a set of attributes that differ between objects. If it is any other case, g_{st} is an empty set. Based on this, in the given incomplete intuitive fuzzy decision information system $I = (M, N, \lambda T = B \cup \{f\}, E)$ and related discernibility matrix, it is assumed that any g_{st} belongs to the discernibility matrix. If $\exists W \subseteq B$ occurs and the intersection of W and g_{st} is not an empty set, then W is a reduction of the incomplete intuitive fuzzy decision information system. At this point, if g_{st} is a combination set of conditional attribute x_k and is not an empty set, a correlation function can be specified, as expressed in equation (12)

$$\sum g_{st} = x_1 \vee x_2 \vee n \vee x_k \quad (12)$$

In equation (12), $\sum g_{st}$ represents the functional expression of g_{st} . If g_{st} is not an empty set, the value of the function can be given as 1. At this time, the conjunctive normal form expression of the discriminant function is shown in equation (13).

$$DE = \bigwedge_{s=1}^{|\lambda|} \bigwedge_{t=1}^{|\lambda|} \left(\sum g_{st} \right) \quad (13)$$

In equation (13), DE represents the conjunctive normal form of the discriminant function, which can be converted to $DE_{\min} = \bigvee_{r=1}^w W_r$ by the minimum disjunctive normal form transformation. Among them, w represents the total reduction quantity in the incomplete intuitive fuzzy decision information system; W_r represents a reduction among all reductions, so its core attribute set can be represented as $co = \bigcap_{r=1}^w W_r$. Starting from equations (10) to (13), the design of ARAs for imbalanced data in incomplete intuitive fuzzy decision information systems needs to take into account the importance of attribute correlation and the discernibility matrix. The process of this algorithm is shown in Figure 2.

From Figure 2, the algorithm takes parameter specification factors, related domain radii, and the degree of unequal distribution between positive and negative groups in the incomplete intuitionistic fuzzy decision information system $I = (M, N, \lambda T = B \cup \{f\}, E)$ as input values, with the aim of outputting the attribute reduction set R . The process first initializes d_j^- and d_j^+ , and uses $L/\{f\} = \{Y^+, Y^-\}$ to calculate Y^+ on the basis that R is not an empty set. Secondly, for each object p_i , the incomplete neighborhood class, the regularization decision of the incomplete neighborhood class, the upper boundary of the neighborhood tolerance class of the

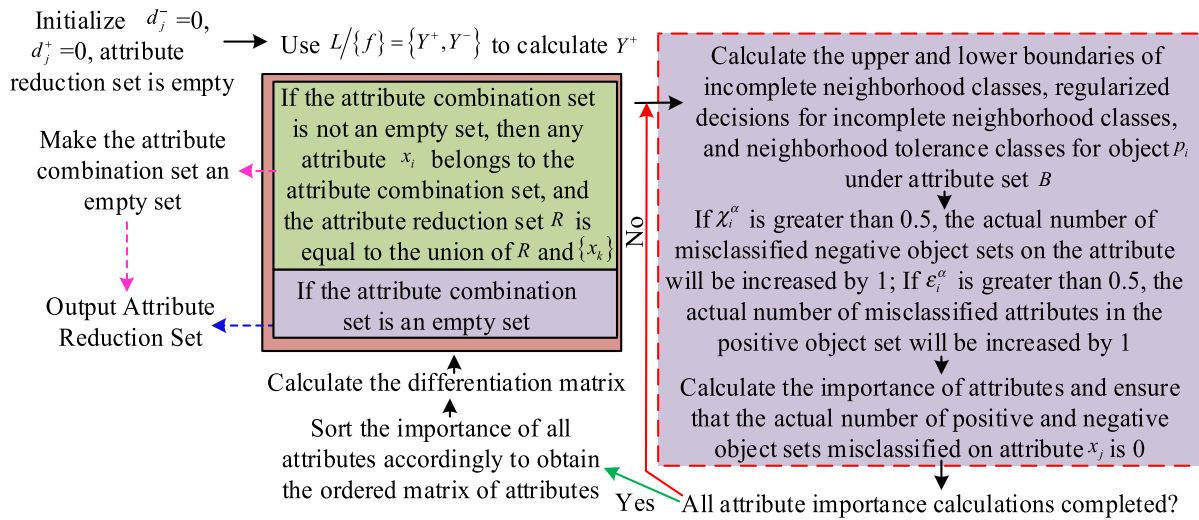


FIGURE 2. Schematic diagram of ARA for incomplete intuitionistic fuzzy information system based on imbalanced data.

object p_i in the attribute set B , and the lower boundary of the neighborhood tolerance class of the object p_i in the attribute set B are calculated. At this point, according to the assumed equation after equation (8), if χ_i^α is greater than 0.5, the actual number of misclassified negative object sets on attribute x_j will be increased by 1; If ϵ_i^α is greater than 0.5, the actual number of misclassified positive object sets on attribute x_j is increased by 1. Calculate the importance of attributes and ensure that the actual number of positive and negative object sets misclassified on attribute x_j is 0. Repeat the calculation until all attribute importance calculations are completed.

Next, rank the importance of all attributes in descending order to obtain the ordered matrix of attributes. Then calculate the discernibility matrix. Finally, for each attribute x_i in the ordered matrix, if the attribute combination set is not an empty set, then any attribute x_i belongs to the attribute combination set, and the attribute reduction set R is equal to the union of R and $\{x_k\}$. At this point, the attribute reduction set is output by making the attribute combination set an empty set; Otherwise, directly end the process and output the attribute reduction set. For incomplete intuitive fuzzy decision information systems, assuming $|L| = n$, $|B| = g$, the relevant time to be calculated through this ARA includes four parts. The first step is to calculate the importance of a neighborhood set and the time spent on each conditional attribute in a scan. Secondly, the time required to calculate the discernibility matrix. Next is the time required to calculate the orderliness matrix. Finally, it is the time it takes to obtain the relevant subset of attributes. Based on this, the time complexity of the four parts is shown in equation (14).

$$\begin{cases} Fi = O(n(n+1)g/2) \\ Se = O(n(n+1)g/2) \\ Th = O(g^2) \\ Fo = O(n^2g) \end{cases} \quad (14)$$

In equation (14), Fi represents the time complexity of the first part; Se represents the time complexity of the second part; Th represents the time complexity of the third part; Fo represents the time complexity of the fourth part. The overall time complexity of the algorithm is expressed in equation (15).

$$Z = O(2n^2g + ng + g^2) \quad (15)$$

In equation (15), Z represents the time complexity of the algorithm; n represents the absolute value of the neighborhood; g represents the absolute value of the B value in the attribute set. Based on this, an example of the algorithm operation is given, and the incomplete intuitive fuzzy decision information system $I = (M, N, \lambda T, E)$ is given. At this point, the neighborhood value and Y^- are assumed to be $\{p_1, p_2, \dots, p_{12}\}$, Y^+ is $\{p_5, p_7, p_{10}\}$, and the attribute set B is $\{p_1, p_2, p_3, p_4, p_5\}$. The value of α is set to 0.2, and the value of π is set to 0.3. After normalizing the relevant attribute values of numerical type, and considering only the attribute p_1 , the neighborhood tolerance class of each object can be calculated according to formula (3). According to equations (5) and (6), the result of misclassification flag values $\underline{k}_{x_j}^{\alpha, i}$ and $\bar{k}_{x_j}^{\alpha, i}$ being 0 can be obtained, which generalizes to all attribute p_k misclassification being 0. Therefore, according to formula (9), the number of negative objects in attribute p_1 that may be misclassified is 3, The number of positive objects in attribute p_1 that may be misclassified is 0. Due to the excessive content involved, the specific process of attribute reduction was simplified into three steps. Firstly, the actual number of misclassified positive and negative object sets on attribute x_j and the importance of the objects are calculated, and an ordered matrix about the attributes is obtained based on the importance of each attribute. The second step is to calculate the discernibility matrix under the incomplete intuitive fuzzy decision information system. Finally, the

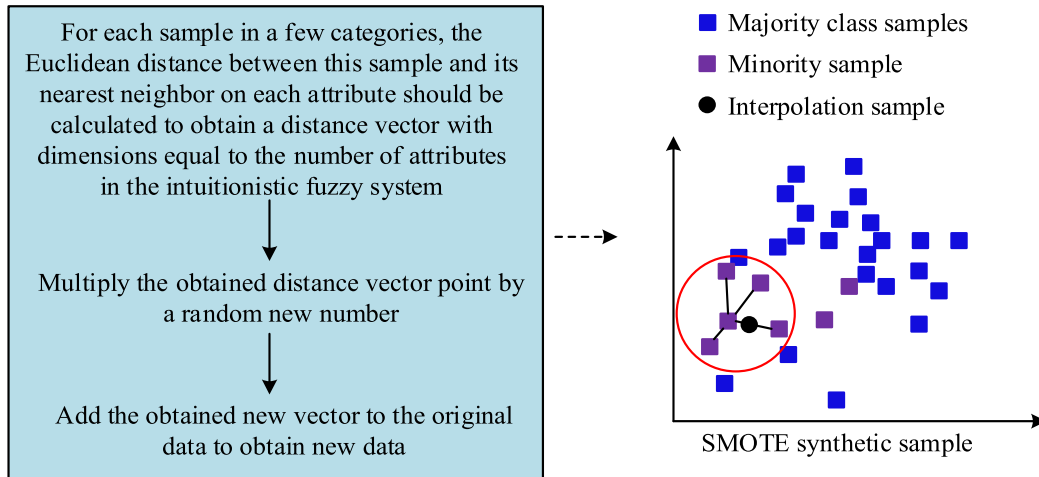


FIGURE 3. Steps for synthesizing new samples using the SMOTE method.

reduction set is calculated by combining the attribute related ordering matrix and atmosphere matrix.

C. OPTIMIZATION ANALYSIS OF ARA BASED ON PRACTICAL LEVEL

In applying the proposed ARA to practical problems, there may be more or less bias towards multi class samples when solving imbalanced data. Therefore, to further enhance the attention of ARAs to minority class samples and reduce decision-making errors of minority class samples, research is conducted before the operation of ARAs. The synthetic minority oversampling technique (SMOTE) is introduced to solve the oversampling problem in data preprocessing. The steps for synthesizing new samples using the SMOTE method are shown in Figure 3.

From Figure 3, for each sample in a few categories, the Euclidean distance between this sample and its nearest neighbor on each attribute needs to be calculated, to obtain a distance vector with dimensions equal to the number of attributes in the intuitionistic fuzzy system. Next, it needs to multiply the obtained distance vector points by a random new number, which maintains a value between 0 and 1. Finally, the new vector obtained is added to the original data to obtain the new data. However, in practical applications, the SMOTE algorithm still has flaws. The SMOTE method uses a method of synthesizing new samples for each minority class sample when synthesizing samples. However, in practical applications, there are often a few minority classes that are outliers. If synthesized around these outliers, it is easy to reduce the effectiveness of the generated new samples. At the same time, the SMOTE method may encounter a boundary between positive and negative classes during the synthesis process. Such boundary samples tend to cause new data to tilt towards the boundary when linear interpolation is performed, and with the increase of synthesis times, the boundary between positive and negative classes tends to blur gradually [21].

Therefore, the study improved the SMOTE algorithm Central Limit SMOTE (CL-SMOTE) by introducing a central limit approach. The sampling rules contained in CL-SMOTE algorithm will consider both down sampling and oversampling, which is a balance between down sampling and oversampling methods. The specific sampling rules are manifested as low imbalanced data when the ratio of majority class to minority class samples is less than or equal to 9. At this time, the ratio of the two is the ratio of the number of samples actually increased by oversampling of minority classes to the number of samples actually decreased by oversampling of majority classes. When the ratio of most samples to minority samples is greater than 9, the data is highly unbalanced. At this time, the ratio of the two is the ratio of the number of samples actually reduced by sampling under most categories to the number of samples actually increased by oversampling of minority categories. It is worth noting that the CL-SMOTE algorithm still has shortcomings in processing high imbalanced data, so the temporary labeling method (TempC) is introduced to further improve it. The TempC method is an imbalanced data processing method based on temporary labeling, which divides the original classification problem into several sub classification problems through two steps, thereby reducing category imbalance in the data. The process of this method is shown in Figure 4.

From Figure 4, the process first involves initializing the temporarily labeled sample set and dividing the original sample set into majority and minority class sample sets. Secondly, multiple nearest neighbor samples of a sample in a minority class are identified in the majority class sample set, and the class of the sample and its multiple nearest neighbor samples are marked as Class C. At the same time, the sample and its multiple nearest neighbor samples are copied into the temporary labeled sample set. Next, a one-step classifier is constructed using the union of the majority class sample set and the minority class sample set, and a two-step classifier is constructed using the temporarily labeled sample set. Finally,

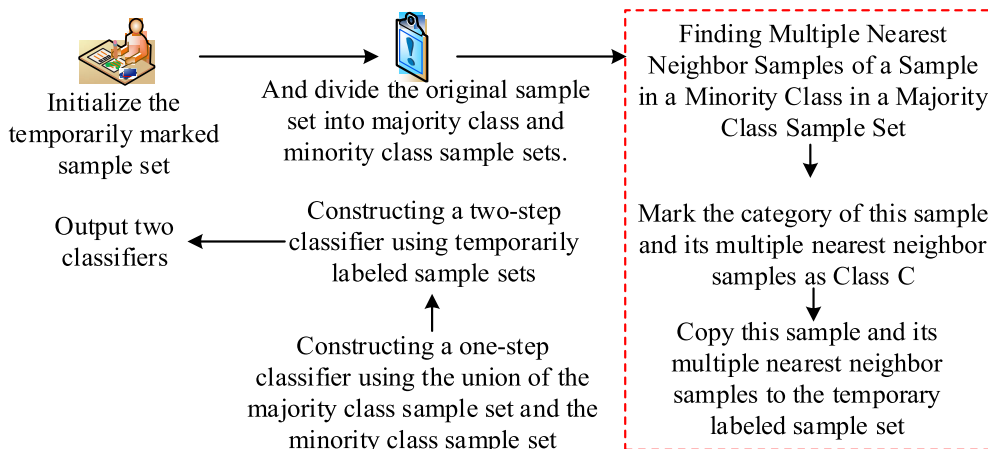


FIGURE 4. Flow diagram of tempC method.

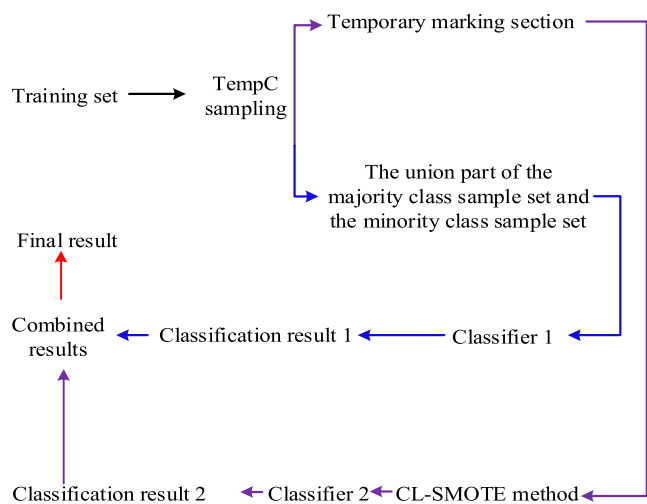


FIGURE 5. The overall process of combining tempC with CL-SMOTE and sampling.

two classifiers are output. From this process, it is evident that the TempC method has significant advantages in handling imbalanced data, effectively reducing the probability of data imbalance and improving the recognition rate of minority class samples. At the same time, it also has high robustness, that is, no matter where the minority class samples are located, it will not affect the actual classification of the two-step classifier. Based on this, the overall process of combining TempC with CL-SMOTE and sampling is shown in Figure 5.

From Figure 5, TempC is first used to process the relevant training samples in the training set into temporary labeled parts and the union part of the majority and the minority class sample set. When conducting specific processing, taking into account the impact of actual imbalanced ratios, when the data is highly imbalanced, selecting too few nearest neighbors will result in difficulties in classifying the union part of the majority and the minority class sample set; When the data is low imbalanced, selecting too many nearest neighbors can

make it difficult to classify the temporary labeled part, and in extreme cases, the union part of the majority and the minority class sample set is completely equivalent to the temporary labeled part. Therefore, the actual number of nearest neighbors should be proportional to the imbalanced ratio of the data. In this case, the grid search method is used to search for the optimal proportion coefficient in two different datasets to determine the closest number of neighbors.

Secondly, it is divided into two steps to train the classifier. The first step is to train the first classifier using the union of the majority and the minority class sample set in the training set, to classify the C-class and majority class samples; The second step is to convert the imbalanced C-class samples into balanced data using the CL-SMOTE method, and then use the processed training set to temporarily mark the part to train a second classifier to classify the majority and minority class samples. Next, it combines the two classifiers, assuming that a sample is classified as Class C by the first classifier and as a minority by the second classifier. The sample will ultimately be classified as a minority class, and in any other case, it will be classified as a majority class. Finally, the actual classification results of the two classifiers are combined to obtain the final result. On the one hand, the TempC-CL-SMOTE joint sampling method has two steps that can effectively handle abnormal samples, thereby transforming the classification problem of abnormal data into an approximate equilibrium classification problem, which can be solved through traditional classification methods. On the other hand, during the sampling, it can effectively reduce the noise caused by oversampling and reduce the information loss caused by undersampling. In summary, the sampling method combining TempC-CL-SMOTE is an effective way to solve the problem of imbalanced data classification.

IV. SIMULATION EXPERIMENT OF ARA

To verify the effectiveness of the proposed ARA, simulation training was conducted using experiments. Eight machine learning related datasets from the University of

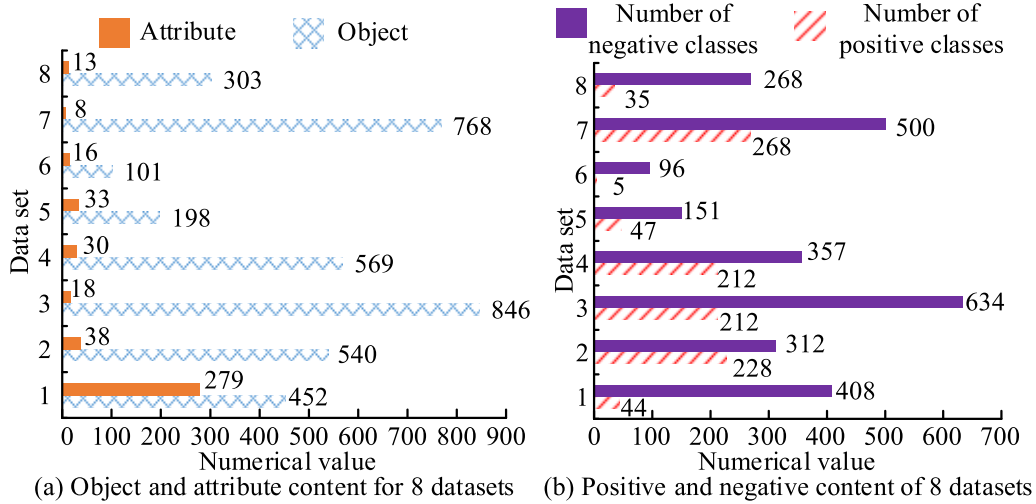


FIGURE 6. Content of experimental related datasets.

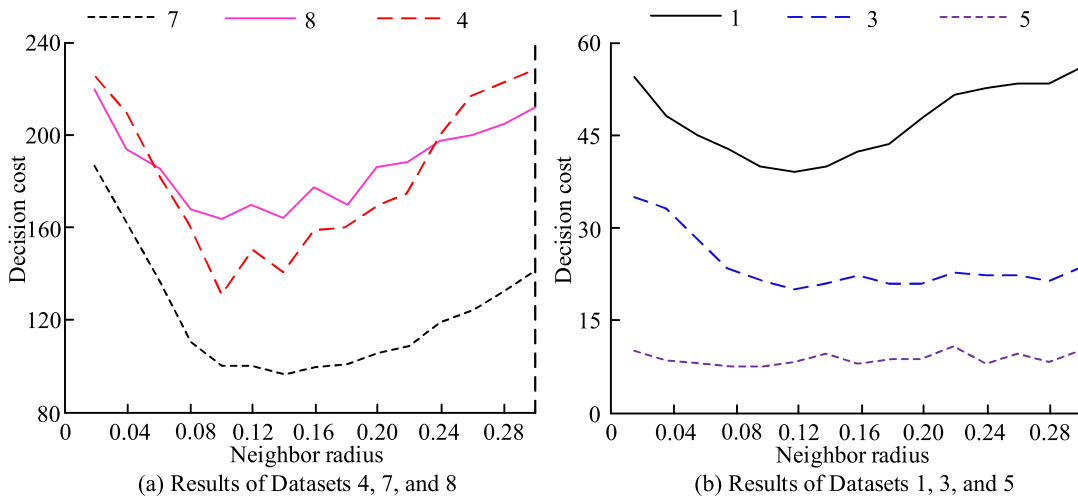


FIGURE 7. Decision costs for different neighborhood radii in local datasets.

California Irvine (UCI) were selected for the experiment, which are arrhythmia, bands, vehicles, breast cancer Wisconsin (WDBC), breast cancer prognosis (WPBC), zoo, Pima, and processed, representing them as 1-8, respectively. Therefore, the content of the dataset is shown in Figure 6.

The data types of the 8 datasets in Figure 6 are roughly divided into mixed, numerical, and symbolic types. Before conducting the experiment, it is necessary to normalize the relevant attribute values with continuity, so that they are between 0 and 1. At the same time, the principle of randomness was used to eliminate 3% of attribute values in the experiment. On this basis, to determine the most suitable neighborhood radius for the research algorithm, the study selected neighborhood radius values between 0.02 and 0.3 at intervals of 0.02 to analyze the decision costs of different neighborhood radii under some datasets (Datasets 2 and 6 were excluded from this experiment due to their overlap with datasets 8 and 1). The results are shown in Figure 7.

From Figure 7, the decision costs of different datasets showed a trend of decreasing first and then increasing. Overall, the optimal experimental effect achieved when the neighborhood radius was 0.12. Therefore, a neighborhood radius of 0.12 was selected for subsequent experiments. On this basis, study the attribute reduction algorithm constructed on the basis of introducing K-nearest neighbor rough set (which integrates K-nearest neighbor and K-nearest neighbor) δ The advantage of nearest neighbors is explained through iterative strategies, which enhances the ability to process heterogeneous data A hybrid data attribute reduction algorithm based on neighborhood rough set combination metrics (this algorithm proposes neighborhood knowledge granularity to evaluate the granulation ability of attributes in hybrid information systems from the perspective of granular computing, and combines neighborhood dependency with neighborhood knowledge granularity to propose neighborhood combination metrics in hybrid information systems, using this metric

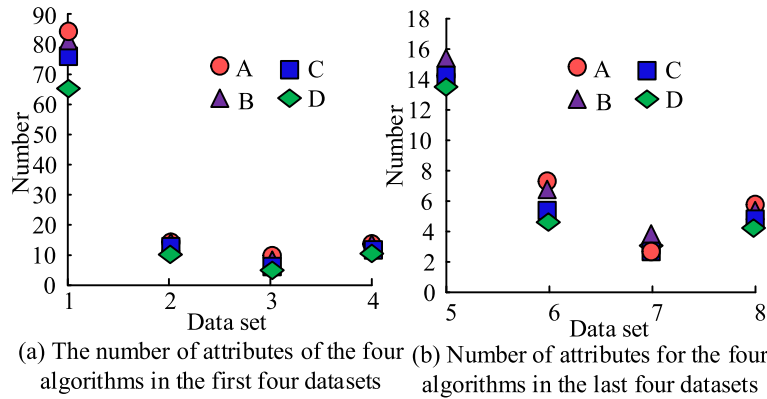


FIGURE 8. Comparison results of attribute quantity of four algorithms in 8 datasets.

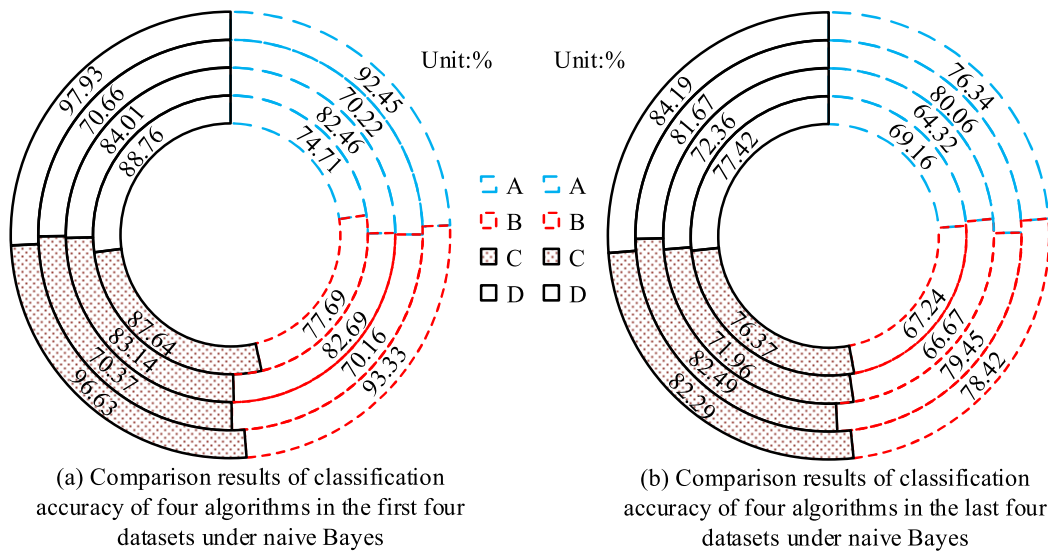


FIGURE 9. Comparison of classification accuracy results of four algorithms in 8 datasets under naive Bayes.

method as a heuristic function) And the unbalanced data attribute reduction algorithm based on neighborhood rough set (this algorithm proposes a Feature selection method using the discernibility matrix by studying the acute offset of the features defined in the upper and lower boundary areas, and thus constructs this algorithm). The three algorithms are represented by A, B, C, and the research algorithm is represented by the letter D. The comparison results of the number of attributes among the four algorithms in 8 datasets are shown in Figure 8.

From Figure 8, the ARA proposed in the study had a smaller number than the comparison algorithm in 7 out of 8 datasets. The number of attributes in datasets numbered 1-8 was 66.3, 10.8, 6.0, 11.7, 13.9, 4.6, 3.2, and 4.4, respectively. Only the dataset with number 7 was higher than the comparison algorithm. Overall, Algorithm A and Algorithm B did not consider the imbalanced state of the dataset. Therefore, the actual attribute selection ability of the algorithm will be reduced to a certain extent when the corresponding attribute

reduction is carried out, which results in a larger actual result of attribute reduction. And algorithm C and the algorithm studied, as they were constructed on imbalanced data, had significantly better results in attribute reduction sets. At the same time, due to the introduction of intuitive fuzziness in the research to construct fuzzy information systems, it had certain advantages in processing missing data, resulting in a reduction performance that is much better than algorithm C. Due to the different processes of attribute evaluation and reduction methods, it is necessary to analyze attribute selection methods using classification methods to verify the superiority of the proposed ARA. Therefore, the study introduces naive Bayes and support vector machines to verify the classification accuracy of the four algorithms. Among them, the classification accuracy comparison results of the four algorithms in 8 datasets under naive Bayes are shown in Figure 9.

In Figure 9 (a), there are datasets numbered 1-4 from the inner ring to the outer ring; In Figure 9 (b), there are datasets numbered 5-8 from the inner ring to the outer ring,

TABLE 1. Comparison of classification accuracy between four algorithms in 8 datasets under support vector machine.

-	A	B	C	D
1	70.30%	71.45%	73.65%	79.71%
2	63.08%	62.50%	79.35%	80.58%
3	76.80%	79.25%	84.74%	85.75%
4	87.75%	86.86%	92.28%	94.40%
5	84.08%	85.67%	87.64%	90.00%
6	77.70%	76.46%	81.69%	82.08%
7	83.28%	86.36%	88.34%	88.05%
8	84.30%	87.37%	90.17%	91.29%

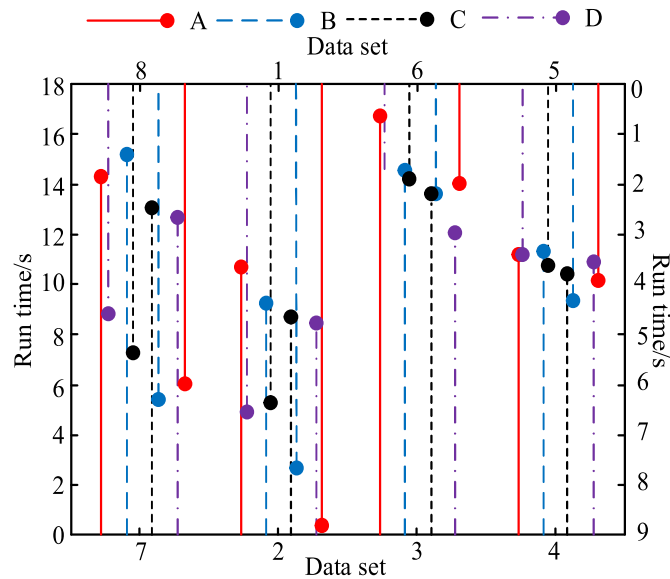


FIGURE 10. Comparison of runtime results of different algorithms on 8 datasets.

respectively. From Figure 9, the classification accuracy of the algorithm studied under Naive Bayes was also higher than that of the comparison algorithm on all datasets except for the dataset numbered 7, with the highest being 97.93% and the lowest being 70.66%. Overall, the algorithm proposed in the study has high effectiveness in improving classification accuracy. To verify this result, the study applied four algorithms to another classifier, and the results are shown in Table 1.

From Table 1, the support vector machine classifier showed the same results as Figure 5, with the highest classification accuracy of 94.40% and the lowest of 80.58%. Overall, compared with other literature, the proposed algorithm could effectively improve the performance of different classifiers on most datasets. Algorithm A and Algorithm B are both ARAs that handle ordinary datasets, but they have limitations in handling imbalanced datasets. The research is based on an incomplete modular intuitive fuzzy information system, and compared to algorithm C, which does not consider default values, it has more advantages. Overall, the algorithm studied is superior to the comparative algorithm. To further verify the superiority of the research algorithm, the study compared the operational efficiency of four algorithms. The results are shown in Figure 10.

From Figure 10, on dataset 2, the algorithm under study ran for 8.02 seconds; The running time on dataset 3 was 12.01 seconds; The running time on dataset 5 was 3.49 seconds; The running time on dataset 6 was 1.72 seconds; The running time on dataset 8 was 4.6 seconds; All were lower than the comparison algorithm C. On dataset 1, the running time was 6.5 seconds; The running time on dataset 4 was 11.01 seconds; Both were lower than Algorithm A and Algorithm B, but slightly higher than Algorithm C. Overall, the average running time of the algorithm proposed in the study was 5.92 seconds, and the overall consumption of running time was lower than that of the comparison algorithm, indicating that it has higher running efficiency. Regarding the relevant uneven data, the study introduced Accuracy, Recall, and F-measure as evaluation indicators. Therefore, the accuracy comparison results of different algorithms on 8 datasets are shown in Figure 11.

From Figure 11, in accuracy comparison, the overall accuracy of the improved attribute reduction algorithm is higher than that of the comparison algorithm. Among them, the accuracy rates of datasets 1-8 were 70.26%, 72.67%, 81.69%, 84.91%, 56.03%, 71.06%, 91.04%, and 81.45%, respectively, with an average accuracy of 76.14%. Overall, the algorithm

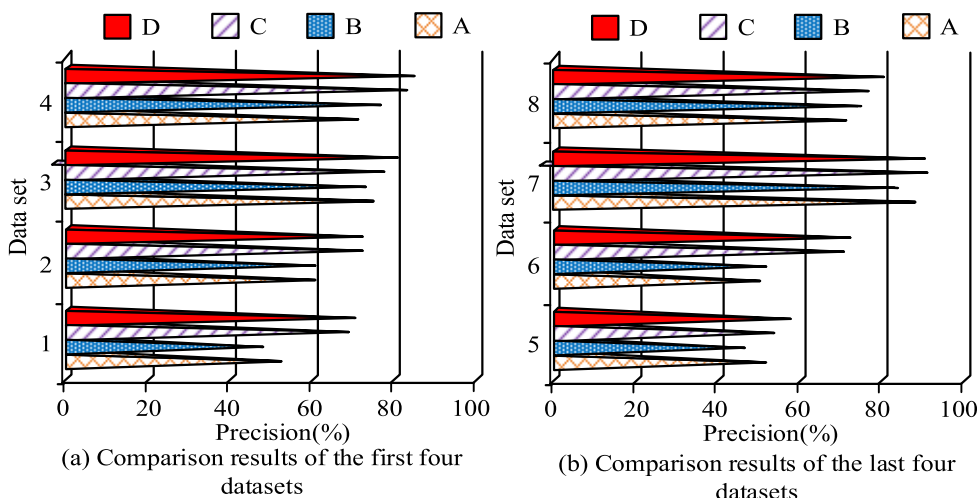


FIGURE 11. Comparison of accuracy results of different algorithms on 8 datasets.

TABLE 2. Comparison of recall rates of different algorithms on 8 datasets.

-	A	B	C	D
1	47.70%	55.95%	75.41%	81.28%
2	64.34%	63.08%	71.06%	72.50%
3	81.97%	80.37%	83.43%	83.84%
4	77.49%	80.48%	83.67%	86.10%
5	48.01%	47.31%	54.74%	58.00%
6	51.24%	49.56%	72.92%	71.10%
7	89.49%	86.71%	90.51%	91.48%
8	71.23%	74.55%	80.90%	82.70%

proposed in the study has a high accuracy, which is significantly superior to Algorithm A and Algorithm B; Except for dataset 7, it outperforms comparison algorithm C and has overall high performance. The comparison results of recall rates between different algorithms on 8 datasets are shown in Table 2.

From Table 2, the recall rate of the algorithms proposed in the construction research of the vast majority of datasets was higher than that of the comparative algorithms. Among them, the highest recall rate was achieved on the dataset numbered 7, at 91.46%; The minimum recall rate was achieved on dataset number 5, which was 57.98. In addition, when only dataset 6 was studied, the recall rate for algorithm C was 71.08%, and algorithm C was 72.90%, the reason is that Algorithm C is also built on imbalanced datasets, and it has certain advantages in processing data such as zoo complete systems. Overall, research on recall rate indicators has high performance. Based on Table 1 and Table 2, it can be seen that the algorithm proposed in the study is generally more suitable for handling imbalanced data in incomplete systems. Finally, the F-measure comparison results of different algorithms on 8 datasets are shown in Figure 12.

From Figure 12, the F-measure value of the improved attribute reduction algorithm is also higher overall than that of the comparison algorithm, only lower than the comparison algorithm C. At this time, the F-measure of the research

algorithm was 71.07%, while the F-measure of algorithm C was 72.17%. In addition, the highest value of the research algorithm appeared on dataset 7, at 91.25%, and the lowest value appeared on dataset 5, at 56.99%. Based on Figures 7, 8, and Table 2, the dataset used in the study contained partially imbalanced data. Therefore, Algorithm A and Algorithm B had limitations in processing data from these datasets, resulting in lower performance than the research algorithm and Algorithm C. In addition, the research algorithm was specifically proposed to solve imbalanced data in incomplete intuitive fuzzy systems. Therefore, compared to Algorithm C, which can only handle imbalanced data in complete systems, the performance of the research algorithm on most datasets was better than Algorithm C. Overall, it is more suitable for handling imbalanced data in incomplete intuitive fuzzy systems. At the same time, the research algorithm has extremely high attribute reduction performance and high effectiveness in the application of incomplete intuitive fuzzy systems.

Finally, in practical applications, the TempC-CL-SMOTE algorithm in the preprocessing was studied and validated. At this time, seven datasets were selected based on the principle of high or low imbalanced data, namely low dimensional and low imbalanced Haberman, Yeast, and Abalone; Low dimensional and high imbalanced Wine; Ionosphere and ZAlizadeh with high dimensionality and low disequilibrium; High dimensional and high imbalanced Data Mining and

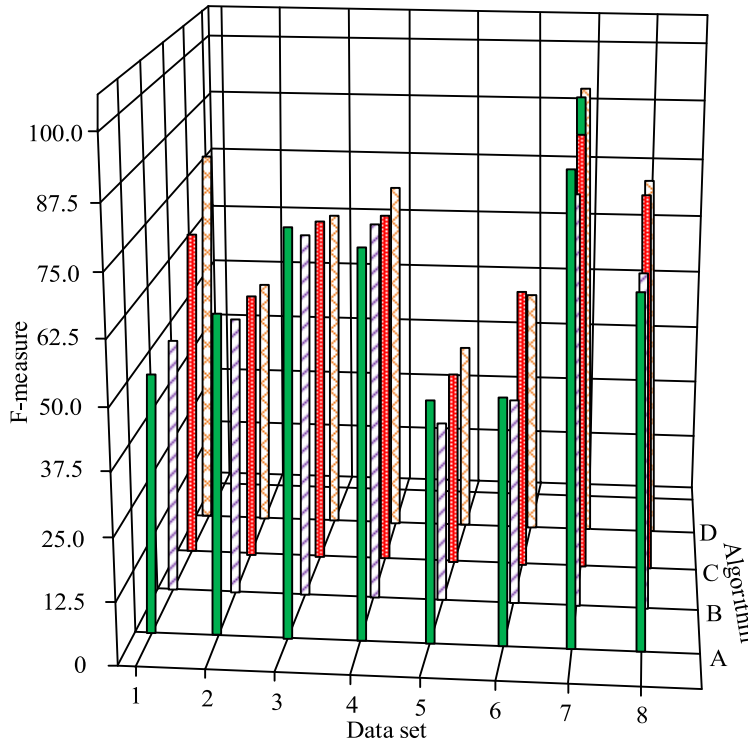


FIGURE 12. Comparison of F-measure results of different algorithms on 8 datasets.

TABLE 3. Comparison results of different sampling methods on 7 datasets.

AUC value comparison results							
-	a	b	c	d	e	f	g
CL-SMOTE	0.62	0.74	0.93	0.74	0.93	0.86	1.00
TempC	0.66	0.79	0.85	0.79	0.91	0.86	0.99
TempC-CL-SMOTE	0.68	0.81	0.95	0.83	0.92	0.89	1.00
F1 value comparison results							
-	a	b	c	d	e	f	g
CL-SMOTE	0.44	0.66	0.18	0.40	0.92	0.82	0.95
TempC	0.47	0.66	0.32	0.48	0.92	0.82	0.97
TempC-CL-SMOTE	0.49	0.68	0.47	0.55	0.92	0.83	0.97
Comparison results of G-means values							
-	a	b	c	d	e	f	g
CL-SMOTE	0.61	0.74	0.93	0.70	0.93	0.86	0.99
TempC	0.66	0.76	0.94	0.80	0.91	0.87	0.99
TempC-CL-SMOTE	0.72	0.81	0.97	0.82	0.92	0.91	0.99

Knowledge Discovery buffer (Kddcupbuffer). It analyzed and compared the area under the curve (AUC), F1 value, and G-mean of CL-SMOTE, TempC, and TempC-CL-SMOTE, represented by a~g, respectively. The results are shown in Table 3.

From Table 3, the TempC-CL-SMOTE method had advantages in three aspects: low dimensional low disequilibrium, low dimensional high disequilibrium, and high dimensional high disequilibrium, compared to the separate use of the CL-SMOTE method and the TempC method. Its performance only on high-dimensional and low imbalanced data was not as good as the CL-SMOTE method. This might be because the TempC method did not play a significant role in processing high-dimensional and low imbalanced

data, but the CL-SMOTE method had good adaptability. Overall, compared to the TempC algorithm, the average AUC value of the TempC-CL-SMOTE algorithm was 3.2% higher; The average F1 value was 4% higher; G-means was 2.9% higher; Compared with CL-SMOTE, AUC and F1 increased by 3.7% and 7.9% respectively, and the G-mean was 5.3% higher. Overall, it has shown better results and provided better support for the implementation of ARAs. To further validate the superiority of the TempC-CL-SMOTE algorithm, the experiment was expanded by applying CL-SMOTE, TempC, and TempC-CL-SMOTE to a new dataset consisting of Page blocks0 and Segment0, respectively. At the same time, an improved version of the previously used dataset was introduced, which compared Pima,

TABLE 4. Comparison results of different algorithms.

F1 value comparison results						
	Page-block s0	Segme nt0	Pi ma	Vehicl e1	Vehicl e0	Yeas t1
-						
CL-SMO TE	0.67	0.97	0.6 5	0.68	0.68	0.66
Temp C	0.69	0.97	0.6 6	0.68	0.69	0.64
Temp C-CL-SMO TE	0.75	0.99	0.6 8	0.70	0.70	0.69
AUC value comparison results						
	Page-block s0	Segme nt0	Pi ma	Vehicl e1	Vehicl e0	Yeas t1
-						
CL-SMO TE	0.78	0.69	0.8 9	0.68	0.66	0.65
Temp C	0.85	0.72	0.9 5	0.71	0.72	0.74
Temp C-CL-SMO TE	0.87	0.75	0.9 9	0.75	0.76	0.78

Vehicle1, Vehicle0, and Yeast1 with F1 and AUC values. The results are shown in Figure 4.

From Table 4, the AUC and F1 values of the TempC-CL-SMOTE algorithm are higher than those of the comparison algorithm, indicating that it has better performance in processing imbalanced datasets. Overall, the improvement direction proposed in the study is correct, and the improved algorithm performance has been significantly improved.

V. CONCLUSION

To solve the problems of high time complexity and difficulty in handling imbalanced data with default values in current incomplete information system ARAs, the introduction of intuitive fuzziness in incomplete information systems was studied, and an ARA was proposed. In response to the shortcomings of this algorithm in handling high imbalanced data, the TempC-CL-SMOTE algorithm was introduced to optimize its preprocessing method for preprocessing data, and its effectiveness was verified through experiments. The experimental results show that in the comparison of attribute quantity, the algorithm proposed in the study has the highest value of 66.3 and the lowest value of 3.2, which is generally lower than the comparison algorithm. In the comparison of classification accuracy, the highest was 97.93% under the naive Bayes classifier. The highest classification accuracy under the support vector machine classifier was 94.40%. In the comparison of operational efficiency, the average running time of the studied algorithm was 5.92 seconds, and

the overall consumption of running time was lower than that of the comparison algorithm. In the comparison of accuracy, the average accuracy of the studied algorithm was 76.14%; In the comparison of recall rates, the average recall rate of the studied algorithm was 78.35%; In the F-measure comparison, the average F-measure value of the studied algorithm was 77.19%. In the validation experiment of the TempC-CL-SMOTE algorithm, its AUC value was 3.2% higher than the TempC algorithm and 3.7% higher than the CL-SMOTE algorithm. Overall, the attribute reduction algorithm proposed in the study outperforms the comparison algorithm in terms of accuracy, recall, and other indicators, and has better performance. The TempC-CL-SMOTE sampling method has high effectiveness in optimizing data sampling, and the combination of the two has practicality in handling imbalanced data. However, the attribute reduction algorithm proposed in the study only involves binary classification, so in subsequent research, it is necessary to study the attribute reduction algorithm of information systems under multi classification. At the same time, the rough set model mainly proposed in the research is mainly used to handle incomplete information. However, in practical life, there is a large amount of mixed symbolic and numerical data, so it is necessary to conduct in-depth research on various types of incomplete rough set models in the future.

REFERENCES

- [1] K. T. Atanassov, "New topological operator over intuitionistic fuzzy sets," *J. Comput. Cognit. Eng.*, pp. 94–102, Apr. 2022.
- [2] M. Tarafdar, X. Page, and M. Marabelli, "Algorithms as co-workers: Human algorithm role interactions in algorithmic work," *Inf. Syst. J.*, vol. 33, no. 2, pp. 232–267, Mar. 2023.
- [3] D. T. Tran and J.-H. Huh, "Building a model to exploit association rules and analyze purchasing behavior based on rough set theory," *J. Supercomput.*, vol. 78, no. 8, pp. 11051–11091, May 2022.
- [4] W. Qian, P. Dong, Y. Wang, S. Dai, and J. Huang, "Local rough set-based feature selection for label distribution learning with incomplete labels," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 8, pp. 2345–2364, Aug. 2022.
- [5] Y. Villuendas-Rey, "Hybrid data selection with preservation rough sets," *Soft Comput.*, vol. 26, no. 21, pp. 11197–11223, Nov. 2022.
- [6] J. Li, Y. Shao, and X. Qi, "On variable-precision-based rough set approach to incomplete interval-valued fuzzy information systems and its applications," *J. Intell. Fuzzy Syst.*, vol. 40, no. 1, pp. 463–475, Jan. 2021.
- [7] J. Yao, Y. Yao, D. Ciucci, and K. Huang, "Granular computing and three-way decisions for cognitive analytics," *Cognit. Comput.*, vol. 14, no. 6, pp. 1801–1804, Nov. 2022.
- [8] J. He, L. Qu, Z. Wang, Y. Chen, D. Luo, and C.-F. Wen, "Attribute reduction in an incomplete categorical decision information system based on fuzzy rough sets," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5313–5348, Oct. 2022.
- [9] Y. Song, D. Luo, N. Xie, and Z. Li, "Uncertainty measurement for incomplete set-valued data with application to attribute reduction," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 10, pp. 3031–3069, Oct. 2022.
- [10] A. Bar and P. S. V. S. S. Prasad, "Approaches for coarsest granularity based near-optimal reduct computation," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 4231–4256, Feb. 2023.
- [11] C.-F. Tsai and Y.-H. Hu, "Empirical comparison of supervised learning techniques for missing value imputation," *Knowl. Inf. Syst.*, vol. 64, no. 4, pp. 1047–1075, Apr. 2022.
- [12] Z. Han, Q. Huang, J. Zhang, C. Huang, H. Wang, and X. Huang, "GA-GWNN: Detecting anomalies of online learners by granular computing and graph wavelet convolutional neural network," *Int. J. Speech Technol.*, vol. 52, no. 11, pp. 13162–13183, Sep. 2022.

- [13] Z. Li, S. Liao, L. Qu, and Y. Song, "Attribute selection approaches for incomplete interval-value data," *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 8775–8792, Apr. 2021.
- [14] T. Karimi and Y. Yahyazade, "Developing a risk assessment model for banking software development projects based on rough-grey set theory," *Grey Sys., Theory Appl.*, vol. 12, no. 3, pp. 574–594, May 2022.
- [15] J. Xu, M. Yuan, and Y. Ma, "Feature selection using self-information and entropy-based uncertainty measure for fuzzy neighborhood rough set," *Complex Intell. Syst.*, vol. 8, no. 1, pp. 287–305, Feb. 2022.
- [16] Y. Chen, P. Wang, X. Yang, and H. Yu, "Bee: Towards a robust attribute reduction," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 12, pp. 3927–3962, Dec. 2022.
- [17] W. Ding, W. Pedrycz, I. Triguero, Z. Cao, and C. T. Lin, "Multigranulation supertrust model for attribute reduction," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 6, pp. 1395–1408, Feb. 2020.
- [18] S. Singh, S. Shreevastava, T. Som, and G. Somani, "A fuzzy similarity-based rough set approach for attribute selection in set-valued information systems," *Soft Comput.*, vol. 24, no. 6, pp. 4675–4691, Mar. 2020.
- [19] P. A. Ejegwa and J. M. Agbetayo, "Similarity-distance decision-making technique and its applications via intuitionistic fuzzy pairs," *J. Comput. Cognit. Eng.*, vol. 2, no. 1, pp. 68–74, Jan. 2022.
- [20] M. Unver, M. Olgun, and E. Ezgi Turkarslan, "Cosine and cotangent similarity measures based on Choquet integral for spherical fuzzy sets and applications to pattern recognition," *J. Comput. Cognit. Eng.*, vol. 1, no. 1, pp. 21–31, Jan. 2022.
- [21] X. Cheng, S. Fu, J. Sun, M. Zuo, and X. Meng, "Trust in online ride-sharing transactions: Impacts of heterogeneous order features," *J. Manage. Inf. Syst.*, vol. 40, no. 1, pp. 183–207, Jan. 2023.



WEIHAN LI was born in December 1978. He received the master's degree. He is currently an Associate Professor. He is engaged in information systems and simulation system research. He has published more than 30 academic papers and more than 20 research projects.



JIANWEI GUO was born in September 1975. He received the master's degree in control science and engineering from the Central University of Finance and Economics. He is currently pursuing the Ph.D. degree with Beijing Jiaotong University. He has 11 academic papers and eight research projects.

...