Graph-Based Unsupervised Feature Selection for Interval-Valued Information System

Weihua Xu, Man Huang, Zongying Jiang, and Yuhua Qian, *Member, IEEE*

***Abstract*— Feature selection has become one of the hot research topics in the era of big data. At the same time, as an extension of single-valued data, interval-valued data with its inherent uncertainty tend to be more applicable than single-valued data in some fields for characterizing inaccurate and ambiguous information, such as medical test results and qualified product indicators. However, there are relatively few studies on unsupervised attribute reduction for interval-valued information systems (IVISs), and it remains to be studied how to effectively control the dramatic increase of time cost in feature selection of large sample datasets. For these reasons, we propose a feature selection method for IVISs based on graph theory. Then, the model complexity could be greatly reduced after we utilize the properties of the matrix power series to optimize the calculation of the original model. Our approach can be divided into two steps. The first is feature ranking with the principles of relevance and nonredundancy, and the second is selecting top-ranked attributes when the number of features to keep is fixed as a priori. In this article, experiments are performed on 14 public datasets and the corresponding seven comparative algorithms. The results of the experiments verify that our algorithm is effective and efficient for feature selection in IVISs.**

***Index Terms*— Feature ranking, graph theory, interval-valued information system (IVIS), matrix power series (MPS).**

I. Introduction

In THE era of big data, data mining is usually used to extract useful information from various complex data, among which interval-valued data are widely used in different fields, such as medical test results [1] and range of temperature [2]. Interval-valued data can express fuzzy information or changeable information because of its own characteristics of uncertainty. As a common method of dimension reduction in data mining, feature selection [3], by eliminating irrelevant or redundant [4] features, is to find the optimal feature subset for improving the accuracy of the model classification and reducing the running time of classification. It is widely known that feature selection can be divided into three types [5] according to its form.

Manuscript received 23 August 2022; revised 11 January 2023; accepted 25 March 2023. This work was supported by the National Natural Science Foundation of China under Grant 61976245. *(Corresponding author: Weihua Xu.)*

Weihua Xu, Man Huang, and Zongying Jiang are with the College of Artificial Intelligence, Southwest University, Chongqing 400715, China (e-mail: chxuwh@gmail.com; huangman1027@163.com; jzyswu@163.com).

Yuhua Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan, Shanxi 030006, China (e-mail: jinchengqyh@126.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3263684.

Digital Object Identifier 10.1109/TNNLS.2023.3263684

The first type is filter methods [6] that score each feature according to the principle of divergence or correlation and set a threshold or the number of features to select a feature subset. They can be divided into univariate filter methods and multivariate filter methods [7]. The univariate filter methods do not need to consider the relationship between the features. They sort the features according to the correlation or mutual information between the feature and the label for decision and then filter out the least relevant features. Multivariate filter methods need to consider the relationship between features. For unsupervised feature selection, filter methods are mainly based on the principle of local preservation found through clustering. Here are several significant research results in this regard. The maximum dual interaction and maximum feature relevance [8] propose the forward filter feature selection method with mutual information, which considers relevance, redundancy, and feature interaction simultaneously. In addition, in the ensemble feature selection framework, Wang et al. [9] take fast correlation based on the filter to capture the nonlinear relationships among features and introduce five aggregators to combine multiple feature subsets into a final set and return it. The multiobjective mutual information (MOMI) method [10] is a multiobjective feature selection algorithm based on mutual information with the principles of minimal redundancy and maximum relevancy with the target class. Compared with the other two types, filter methods are classifier independent and have better computational efficiency. Our approach as a filter method gets fast since it uses intrinsic properties of the data instead of relying on clustering, and the model complexity could be greatly reduced after we utilize the properties of the matrix power series (MPS) to optimize the calculation of the original model.

The second type is the wrapper approach [11], which uses machine learning algorithms to evaluate the effect of feature subsets. They can detect the interaction between two or more features and select the feature subset to optimize the effect of the model. This is a combination of feature subset searching and evaluation metrics. The former provides candidate subsets of new features, and the latter trains a model based on the new subsets of features and evaluates them on the validation set. The entire process tends to be computationally expensive, and it is easy to overfit when the samples are not sufficient. Manikandan et al. [12] incorporate a wrapper-based subset selection technique for selecting a subset from the high dimensional datasets, but the feature subsets are given to the classifier iteratively until the maximum accuracy is

obtained for finding the optimal threshold value. Therefore, there is no doubt that the high computational complexity becomes its main drawback. The feature selection method based on a multilayer perception neural network [13] can select essential features and discard derogatory and indifferent features. In addition, this method is extended by equipping it with a mechanism to deal with redundancy. However, it cannot control the level of redundancy maintaining the desired level of performance of the classifier.

The third type is the embedded methods [14], which embed feature selection into the model construction process. The feature selection with constrained $\ell_{2,0}$-norm and optimized graph [15] unifies feature selection and similarity matrix construction into a general framework instead of independently performing the two-stage process. This approach is showed that the result tends to be well when the feature number is large. In addition, the unsupervised feature selection [16] operates on a group of features initially and then updates the selection when a better group appears, which exploits the combination effect of the features by $\ell_{2,0}$ norm. However, it seems a trouble of solving such optimization problems with the reasons of the nonsmooth objective function and nonconvex constraints.

The rough set theory (RST) proposed by Pawlak [17] belongs to a kind of granular computing theory, which can be used to deal with the uncertainty and inconsistency of information. It is worth mentioning that RST has become one of the research focuses in the field of artificial intelligence not only in theoretical research but also in its application in the world nowadays. The interval-valued dataset is called the interval-valued information system (IVIS) [18], and feature selection is also called attribute reduction in RST. RST-based attribute reduction methods are directly applicable to interval-valued datasets [19], [20], [21], [22], [23], [24] and single-point datasets [25], [26], [27]. However, these methods usually encounter the dilemma of high computational complexity.

Using graphs to represent the relationships between data is a common and highly efficient method [28], [29], [30], [31]. Therefore, graph-based feature selection methods become more and more widely applicable because of their good performance. The graph-based feature selection algorithm mainly includes two stages: graph construction and feature evaluation. The structure graph is constructed to provide as close as possible to the actual sample distribution for the feature subset evaluation stage so that, in the subsequent feature selection, as many features containing important information as possible and redundant features are removed.

Our proposed method belongs to a kind of filter method for IVISs. In the process of attribute reduction, this article starts with the importance of the attribute to the entire information system, which is mainly affected by two factors. One is the range of variation of the attribute values in IVISs, and the other is the correlation between attributes. For the first factor, if the value of an attribute fluctuates greatly in the information system, we believe that the attribute gets significant for this IVIS. On the contrary, if the fluctuation is small, the attribute is considered nonsignificant. Therefore, this factor can be regarded as the correlation between a certain attribute and the entire information system. For the second factor, if an attribute is not correlated with other attributes for the same IVIS, the attribute becomes crucial because it reflects its uniqueness under this IVIS. On the contrary, if an attribute is strongly related to the other attribute for the same IVIS, the function of this attribute could be replaced by other attributes, so this attribute seems unconsidered. Therefore, the attribute subset selected by combining these two factors tends to be relevant to the entire information system but not redundant. Then, a weighted undirected fully connected graph is constructed based on the graph theory to transform the information in IVIS into an adjacency matrix of the graph starting from the above two factors. In the process of calculating the importance of each feature by using this adjacency matrix, the calculation complexity is greatly reduced by using the convergence properties of MPS.

The main contributions of this study are given as follows.

1) This article proposes a new feature selection method based on graph theory for IVISs with the principles of relevance and nonredundancy. In addition, we utilize the properties of the MPS to reduce the model complexity greatly, so our model still has less reduction time in the case of large samples.
2) Comparative experiments are carried out on 14 public datasets, and the results show the effectiveness and efficiency of the proposed algorithm against the corresponding seven comparative algorithms.

This article is organized by the following five sections. In Section II, we mention some relevant preliminaries on graph theory for IVISs and MPS. In Section III, we construct the graph for IVISs and design the unsupervised attribute reduction method, as well as the corresponding algorithm. The experiments to verify the efficiency and practicability of this new way are presented in Section IV. Section V is the conclusions and future work of this article.

II. Preliminaries

In this section, we introduce IVISs and then review some definitions of the graph theory, MPS, and the Spearman rank correlation coefficient (SRCC).

A. Interval-Valued Information System

Let $U = \{x_1, x_2, \ldots, x_n\}$ be an object set and $A = \{a_1, a_2, \ldots, a_n\}$ be an attribute set. IVIS $= \langle U, A, F\rangle$ is called an IVIS, and for any $x_i \in U, a_k \in A$, $f(x_i, a_k)$ is an interval-valued number, i.e., $f(x_i, a_k) = (f^l_{a_k}(x_i), f^r_{a_k}(x_i))$. $f^l_{a_k}(x_i)$ and $f^r_{a_k}(x_i)$ are called the left and right boundaries of $f(x_i, a_k)$, respectively, and they can also be denoted by f^l_{ik} *and* f^r_{ik}. In an IVIS, for any $x_i \in U, a_k \in A$, $f(x_i, a_k)$ degenerates to a single value when $f^l_{a_k}(x_i) = f^r_{a_k}(x_i)$. Therefore, a single-valued decision system is a special form of the IVIS.

According to the literature [32], given an interval-valued ordered information system (IVOIS) $\mathrm{IS}^{\preccurlyeq} = \langle U, A, F\rangle, \forall B \subseteq A$, the dominance relation $D_B^{\preccurlyeq}$ is defined as

$$D_B^{\preccurlyeq} = \{(x_i, x_j) \in U \times U \mid f^l_{a_k}(x_i) \le f^l_{a_k}(x_j), f^r_{a_k}(x_i) \le f^r_{a_k}(x_j), \forall a_k \in B\}. \quad (1)$$

From (1), we easily find that the dominance relation $D_B^{\preccurlyeq}$ is reflexive, asymmetric, and transitive. Then, the dominating and dominated sets of $x_i \in U$ in terms of B are defined as

$$D_B^+(x_i) = \left\{x_j \in U | x_i D_B^{\preccurlyeq} x_j\right\}$$
$$D_B^-(x_i) = \left\{x_j \in U | x_j D_B^{\preccurlyeq} x_i\right\}$$

which are called knowledge granules induced by $D_B^{\preccurlyeq}$. Therefore, we can know the ordinal position of an object under an attribute by calculating the cardinality of the corresponding dominating and dominated sets.

B. Graph Theory

In the light of literature [33], the set of features $H = \{h_1, \ldots, h_n\}$ can be represented by a node set with the same amount from a weighted undirected fully connected graph $G = (V, E)$. What is more, the set of edges E formed by connecting the points in the graph indicates the relationships between features. The adjacency matrix B of the graph G is used to describe the information contained by relationships between nodes, where the elements $B(i, j)$, $1 \leqslant i, j \leqslant n$, demonstrate the confidence that features h_i and h_j (thenodes$\vec{v}_i$and$\vec{v}_j$) are both good candidates to be selected. The element $B(i, j)$ can be expressed by an associated weight function $\omega(\cdot, \cdot)$

$$B(i, j) = \omega(i, j).$$

For the unsupervised feature selection, $\omega(\cdot, \cdot)$ is a weighted linear combination of the two statistics of the features h_i and h_j, defined as

$$\omega_U\left(\vec{v}_i, \vec{v}_j\right) = \alpha L_{ij} + (1 - \alpha) D_{ij}.$$

L_{ij} indicates the relevance of feature h_i to the information system, while D_{ij} represents the irrelevance between feature h_i and other features. Therefore, $\omega_U(\vec{v}_i, \vec{v}_j)$ can be used to measure the importance of the feature in terms of the principles of relevance and nonredundancy.

C. Matrix Power Series

Given a sequence of matrices $Q_0, Q_1, Q_2, \ldots, Q_k, \ldots$, where $Q_k = (q_{ij}^{(k)}) \in C^{m \times n}$, then the sum formula $Q_0 + Q_1 + Q_2 + \cdots + Q_k + \cdots$ is called an MPS [34], which can be written as $\sum_{k=0}^{\infty} Q_k$.

For any positive integer N, let $S^{(N)} = \sum_{k=0}^{N} Q_k$ be the partial sum of the matrix series. If the matrix sequence $\{S^{(N)}\}$, which is made of $S^{(N)} = \sum_{k=0}^{N} Q_k$, converges [35] and has a limit, that is, $\lim_{N \to \infty} S^{(N)} = S$, then the matrix series $\sum_{k=0}^{\infty} Q_k$ converges, and let call S the sum of this matrix series, that is, $S = \sum_{k=0}^{\infty} Q_k$. It is worth noting that, if the matrix series $\sum_{k=0}^{\infty} Q_k$ converges to S, we will get $m \times n$ convergent numerical series, that is,

$$\sum_{k=0}^{\infty} Q_k = S \Leftrightarrow \lim_{N \to \infty} \sum_{k=0}^{N} Q_k = S$$
$$\Leftrightarrow \lim_{N \to \infty} \sum_{k=0}^{N} a_{ij}^{(k)} = s_{ij}$$
$$\Leftrightarrow \sum_{k=0}^{\infty} a_{ij}^{(k)} = s_{ij}.$$

Using the convergence properties of MPS can greatly simplify the calculation when summing some specific matrix series.

D. Spearman Rank Correlation Coefficient

Let ρ be an SRCC [36] of two vectors a and b, where

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}. \tag{2}$$

n is the sample size, and d represents a sort difference sequence of a and b. From (2), we can conclude that $\rho \in [-1, 1]$, and the closer the absolute value of ρ is to 1, the stronger the correlation between a and b.

Example 1: Here are two stocks a and b in different sectors, given the monthly returns of these two stocks over the past six months

$$a = (1.23\%, 2.58\%, -0.03\%, 0.97\%, -1.45\%, 1.63\%)$$
$$b = (0.32\%, 0.56\%, 0.13\%, 3.28\%, 2.99\%, 4.01\%).$$

Then, the sort difference sequence of a and b is

$$d = (4, 6, 2, 3, 1, 5) - (2, 3, 1, 5, 4, 6) = (2, 3, 1, -2, -3, -1)$$

and the SRCC is computed as follows:

$$\rho = 1 - \frac{6 \times \left((2)^2 + (3)^2 + 1^2 + (-2)^2 + (-3)^2 + (-1)^2\right)}{6 \times \left(6^2 - 1\right)}$$
$$= 0.2.$$

It shows that the monthly returns of these two stocks have a slight trend of change in the same direction in the past six months.

III. Our Approach

The proposed feature selection method can be constructed in three steps, as shown in Fig. 1.

A. Digital Feature of Interval-Valued Random Variable

In order to facilitate the selection and definition of correlation indicators and redundancy indicators for IVIS in subsequent research, the following numerical characteristic definitions of interval-valued random variables are given first.

Definition 1: Let $g = (g^-, g^+)$ and $h = (h^-, h^+)$ be two interval values. The distance [23] between two interval values g and h is redefined as follows:

$$d(g, h) = \frac{1}{\sqrt{3}}\left(|g^- - h^-|^2 + |\frac{g^- + g^+}{2} - \frac{h^- + h^+}{2}|^2 + |g^+ - h^+|^2\right)^{\frac{1}{2}}.$$

Definition 2: In the probability distribution, let Y be an interval-valued random variable. The population mean $E(Y)$,

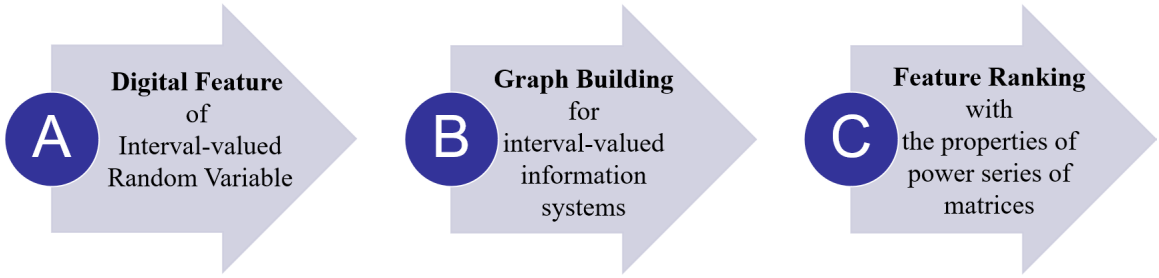

Fig. 1. Construction process of our approach.

the population variance $D(Y)$, and the standard deviation of Y are denoted as follows:

$$E(Y) = E\left(\left(Y^l, Y^r\right)\right) = \left(E\left(Y^l\right), E\left(Y^R\right)\right) = \left(\frac{1}{N}\sum_{i=1}^{N} y_i^l, \frac{1}{N}\sum_{i=1}^{N} y_i^r\right)$$

$$D(Y) = \frac{1}{N}\sum_{i=1}^{N} (d(y_i, E(Y)))^2$$

$$\sigma(Y) = \sqrt{D(Y)} = \left(\frac{1}{N}\sum_{i=1}^{N} (d(y_i, E(Y)))^2\right)^{1/2}$$

where N indicates the size of the population. Similarly, if n is equal to the sample size, then sample mean $\overline{Y}$ and sample variance S^2 could be defined as

$$\overline{Y} = \left(\overline{Y^l}, \overline{Y^r}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} {y_i}^l, \frac{1}{n}\sum_{i=1}^{n} {y_i}^r\right)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(d\left(y_i, \overline{Y}\right)\right)^2$$

$$S(Y) = \sqrt{S^2(Y)} = \left(\frac{1}{n-1}\sum_{i=1}^{n} \left(d\left(y_i, \overline{Y}\right)\right)^2\right)^{1/2}.$$

In practice, the values of sample statistics are often used to estimate population parameters. Based on the previous definition [32], given an IVOIS $\mathrm{IS}^{\preccurlyeq} = \langle U, A, F\rangle, \forall B \subseteq A$, the dominance relation $D_B^{\preccurlyeq}$ is defined as

$$D_B^{\preccurlyeq} = \left\{\left(x_i, x_j\right) \in U \times U \middle| v_{a_k}^l(x_i) \le v_{a_k}^l\left(x_j\right), v_{a_k}^r(x_i) \le v_{a_k}^r\left(x_j\right), \forall a_k \in B\right\}.$$

However, this definition cannot sort two nested interval values. In addition, a certain degree of deviation should be allowed when ordering sample values. Therefore, we introduce a new method of ranking interval values.

Definition 3: Let $g = (g^-, g^+)$ and $h = (h^-, h^+)$ be two interval values. $g \preccurlyeq f$ if and only if

$$\left(\frac{h^+ + h^-}{2} - \frac{g^+ + g^-}{2}\right) > \beta$$

where β is a nonnegative number, which represents the tolerance of deviation.

Definition 4: Given an IVOIS $\mathrm{IS}^{\preccurlyeq} = \langle U, A, F\rangle, \forall B \subseteq A$, the dominance relation $D_B^{\preccurlyeq}$ is defined as

$$D_B^{\preccurlyeq} = \left\{\left(x_i, x_j\right) \in U \times U \middle| \frac{f_{a_k}^l\left(x_j\right) + f_{a_k}^r\left(x_j\right)}{2} - \frac{f_{a_k}^l(x_i) + f_{a_k}^r(x_i)}{2} \ge \beta, \beta \ge 0\right\}.$$

The dominating and dominated sets of $x_i \in U$ in terms of B are defined as

$$D_B^+(x_i) = \left\{x_j \in U | x_i D_B^{\preccurlyeq} x_j\right\}$$

$$D_B^-(x_i) = \left\{x_j \in U | x_j D_B^{\preccurlyeq} x_i\right\}$$

which are called knowledge granules induced by $D_B^{\preccurlyeq}$. We can use the cardinality of the dominating or dominated sets of $x_i \in U$ in terms of one feature h_j to know the sort position of the object x_i in the attribute h_j. Therefore, for an IVOIS $\mathrm{IS}^{\preccurlyeq} = \langle U, A, F\rangle, \forall h_i, h_j \in A$, SRCC of h_i and h_j is computed as

$$\rho_{i,j} = 1 - \frac{6\sum_{k=1}^{n}\left(\left|D_{\{h_i\}}^-(x_k)\right| - \left|D_{\{h_j\}}^-(x_k)\right|\right)^2}{n\left(n^2-1\right)}.$$

B. Graph Building for IVIS

In the unsupervised feature selection for IVIS, we build upon a weighted undirected fully connected graph $G = (V, E)$. Using node set $V = \{\vec{v}_1, \ldots, \vec{v}_n\}$ represents the set of features $A = \{h_1, \ldots, h_n\}$ from the IVIS. In addition, using the set of edges E formed by connecting the points in the graph indicates the relationship between features. Then, the adjacency matrix B of the graph G is used to describe the information contained by relationships between features, where the element $B(i, j)$, $1 \leqslant i, j \leqslant n$, demonstrates the confidence that features h_i and h_j (the nodes$\vec{v}_i$ and $\vec{v}_j$) are both good candidates to be selected. The element $B(i, j)$ can be expressed by an associated weight function $\omega(\cdot, \cdot)$, defined as

$$\omega_U\left(\vec{v}_i, \vec{v}_j\right) = \alpha L_{ij} + (1-\alpha) D_{ij}. \tag{3}$$

L_{ij} shows the maximum normalized standard deviation of the two features distributions on IVIS, i.e., $L_{ij} = \max(S_i, S_j)$, where S_i is the standard deviation over the samples $\{h_i\}$, which is normalized to the range $[0, 1]$ by the maximum standard deviation over the attribute set F. L_{ij} reflects the

principle of relevance, which means the bigger, the better. As for the principle of nonredundancy, there is no doubt that we expect as small as possible. Therefore, in order to keep the monotony of the criteria $\omega(i, j)$, the second term is the opposite of the correlation $D_{ij} = 1 - |\rho(h_i, h_j)|$. α is a parameter $\in [0, 1]$, with its value being estimated during the experiments by fivefold cross-validating on the training set for the classification tasks.

C. Feature Ranking Procedure

The feature ranking procedure for an IVIS can be explained with the properties of the power series of matrices. Let $\eta = \{\vec{v}_0 = i, \vec{v}_1, \ldots, \vec{v}_{l-1}, \vec{v}_l = j\}$ represent a path of length l between nodes $\vec{v}_0$ and $\vec{v}_l$, where $\vec{v}_0$ and $\vec{v}_l$ denote features h_i and h_j on a weighted undirected fully connected graph $G = (V, E)$. Provided that the number of the nodes n in the graph G is more than the length l of the path η, in this case, the path can simply be deemed as a subset of the feature set A.

Then, the overall weight matched with η is defined as

$$\vartheta_\eta = \prod_{k=0}^{l-1} B\left(\vec{v}_k, \vec{v}_{k+1}\right)$$

with ϑ_η interpreting all the feature pairs that belong to it. On account of that, there can be more than one path of length l connecting nodes $\vec{v}_0$ and $\vec{v}_l$, let us define the set $\Phi^l_{i,j}$ consisting of all the paths of length l between two nodes $\vec{v}_0$ and $\vec{v}_l$. Then, we use the following sum:

$$\Gamma_l(i, j) = \sum_{\eta \in \Phi^l_{i,j}} \vartheta_\eta$$

to express the contribution of all these paths. Following the notion of matrix algebra, we can conclude that

$$\Gamma_l = B^l.$$

From the perspective of feature selection, feature h_i and feature h_j can be related directly or through other features. When holding the number of mediations connecting h_i and h_j to be a constant $l-1$, $\Phi^l_{i,j}$ denotes all these possible situations. Therefore, we can evaluate the single feature score for the feature h_i at a given path length l as

$$\varrho_l(i) = \sum_{j \in V} \Gamma_l(i, j) = \sum_{j \in V} B^l(i, j).$$

In addition, combining all the possible path lengths means that the set of paths can be seen as all the subsets of feature set A, which can be calculated as

$$\varrho(i) = = \sum_{l=1}^{\infty} \varrho_l(i) = \sum_{l=1}^{\infty} \left(\sum_{j \in V} \Gamma_l(i, j) \right) = \sum_{l=1}^{\infty} \left(\sum_{j \in V} B^l(i, j) \right).$$

Equation (4) evaluates the value of feature h_i when considered in every subset of features. It is obvious that the higher $\varrho(i)$, the better. However, directly computing ϱ in terms of (4) is impractical because the computation tends to be infinite. As a result, it is apparent to explore whether we could simplify the computation by applying the property of power series convergence in algebra. Let Θ be the power series of adjacency matrix B: $\Theta = \sum_{l=1}^{\infty} B^l$. Then, $\varrho(i)$ can be shown as

$$\varrho(i) = \sum_{l=1}^{\infty} \varrho_l(i) = \left[\left(\sum_{l=1}^{\infty} B^l \right) \boldsymbol{I} \right]_i = [\Theta \boldsymbol{I}]_i$$

where $\boldsymbol{I}$ is a column vector whose each component equals 1, and the square bracket denotes the ith component of this vector.

The problem is that the power series $\sum_{l=1}^{\infty} B^l$ may not be convergent. Therefore, we should use the method of regularization in the form of generating functions by assigning a consistent value for the sum of possible divergent power series. There exist different forms of generating functions [37], [38] and the generating function that we define for the $l-$path as

$$\check{\varrho}(i) = \sum_{l=1}^{\infty} \varsigma^l \varrho_l(i) = \sum_{l=1}^{\infty} \sum_{j \in V} \varsigma^l \Gamma_l(i, j)$$

where ς is a real-valued regularization factor and ς^l can be considered as the weight for paths of length l. Let $\delta(B)$ be the spectral radius of B, so parameter ς was defined as $\varsigma = 0.9/\delta(B)$ in order to ensure that the infinite sum $\sum_{l=1}^{\infty} B^l$ would be convergent. The effectiveness of this method is proven as follows.

Proof [39]: Given the unit matrix $\boldsymbol{E}$ and the eigenvalues $\{\epsilon_0, \ldots, \epsilon_{n-1}\}$ of matrix B, obtaining from linear algebra, we can define the spectral radius $\delta(B) = \max_{\epsilon_i \in \{\epsilon_0, \ldots, \epsilon_{n-1}\}} (|\epsilon_i|)$. Then, $\lim_{l \to \infty} B^l = 0 \Leftrightarrow \delta(B) < 1 \Leftrightarrow \sum_{l=1}^{\infty} B^l = (\boldsymbol{E} - B)^{-1} - \boldsymbol{E}$, and Gelfand's formula shows that, for matrix B, we have $\delta(B) = \lim_{l \to \infty} \|B^l\|^{1/l}$. Following this, for matrix B and M, we have $\delta(BM) \leq \delta(B)\delta(M)$. When $M = \varsigma \boldsymbol{E}$, we obtain $\delta(\varsigma B) = \delta((\varsigma \boldsymbol{E})B) \leq \delta(\varsigma \boldsymbol{E})\delta(B) = \varsigma \delta(B)$. Thus, if we choose ς, which could satisfy $0 < \varsigma < (1/\delta(B))$, the power series of matrix in the definition of $\check{\varrho}(i)$ would be convergent because $0 < \delta(\varsigma B) = \varsigma \delta(B) < (1/\delta(B))\delta(B) = 1$. Then, we can get $\check{\Theta} = \sum_{l=1}^{\infty} (\varsigma B)^l = (\boldsymbol{E} - \varsigma B)^{-1} - \boldsymbol{E}$. Definitely, we use $\varsigma = (0.9/\delta(B)), \varsigma \in (0, (1/\delta(B)))$ for all the experiments in this article.

As a result, $\check{\varrho}(i)$ can be concisely computed with the convergence property of the power series of a matrix, and we can acquire the single score for each feature by marginalizing this quantity

$$\check{\varrho}(i) = \left[\check{\Theta} \boldsymbol{I}\right]_i.$$

According to the definition of single scores of features for IVISs in this article, the highest value means that the corresponding feature tends to be most relevant but least redundant compared with others. Therefore, the top-ranked features should be retained as the reduction set for IVISs, which can be obtained from the decreasing order of $\check{\varrho}$ vector. Algorithm 1 completes feature ranking by following this logic. The computational complexity of Algorithm 1 is $\mathcal{O}(mn^3)$, where m is the number of samples and n is the number of

initial features. From this complexity, we can see that our algorithm has a great advantage in reducing time under large sample data.

Algorithm 1 Unsupervised Feature Ranking for IVIS

Input : $IVIS = \langle U, A, F\rangle, \alpha, \beta$; where $U = \{x_1, x_2, \ldots, x_m\}$, $A = \{h_1, h_2, \ldots, h_n\}$, $\alpha \in [0, 1]$, $\beta \geq 0$;
Output : $\breve{\varrho}$, the final scores for each feature and select the features with top scores

```
for i ← 1 to n do
    for k ← 1 to m do
        calculate d(h_i(x_k), h_i^-);
    end
    compute S(h_i);
end
for i ← 1 to n do
    for j ← 1 to n do
        for k ← 1 to m do
            calculate D^-_{h_i}(x_k) and D^-_{h_j}(x_k);
        end
        compute ρ(h_i, h_j);
        D_ij ← 1 − |ρ(h_i, h_j)|;
        L_ij ← max(S(h_i), S(h_j));
        B(i, j) ← αL_ij + (1 − α)D_ij;
    end
end
compute adjacency matrix B;
ς = 0.9/δ(B);
Θ̆ = (E − ςB)^{-1} − E;
ϱ̆(i) = [Θ̆I]_i.
return ϱ̆
```

Next, a small example is given to explain the algorithm presented above.

Example 2: Table I shows an IVIS $= (U, A, F)$. Then, we have

$$
\begin{aligned}
&S(h_1) = 0.1099, \quad S(h_2) = 0.1034, \quad S(h_3) = 0.2265\\
&S(h_4) = 0.1713, \quad S(h_5) = 0.2692, \quad S(h_6) = 0.2214\\
&L_{12} = L_{21} = 0.0390, \quad L_{13} = L_{31} = 0.7426\\
&L_{14} = L_{41} = 0.4096, \quad L_{15} = L_{51} = 1.0000\\
&L_{16} = L_{61} = 0.7119, \quad L_{23} = L_{32} = 0.7426\\
&L_{24} = L_{42} = 0.4096, \quad L_{25} = L_{52} = 1.0000\\
&L_{26} = L_{62} = 0.7119, \quad L_{34} = L_{43} = 0.7426\\
&L_{35} = L_{53} = 1.0000, \quad L_{36} = L_{63} = 0.7426\\
&L_{45} = L_{54} = 1.0000, \quad L_{46} = L_{64} = 0.7119\\
&L_{56} = L_{65} = 1.0000.
\end{aligned}
$$

When $\beta = 0.01$ and $\alpha = 0.7$, following Definition 4, we could get

$$
\begin{aligned}
&\rho_{1,1} = \rho_{2,2} = \rho_{3,3} = \rho_{4,4} = \rho_{5,5} = \rho_{6,6} = 1.0000\\
&\rho_{1,2} = \rho_{2,1} = 0.9940, \quad \rho_{1,3} = \rho_{3,1} = 0.6429\\
&\rho_{1,4} = \rho_{4,1} = -0.2679, \quad \rho_{1,5} = \rho_{5,1} = 0.1905\\
&\rho_{1,6} = \rho_{6,1} = 0.6310, \quad \rho_{2,3} = \rho_{3,2} = 0.6667\\
&\rho_{2,4} = \rho_{4,2} = -0.2917, \quad \rho_{2,5} = \rho_{5,2} = 0.2440\\
&\rho_{2,6} = \rho_{6,2} = 0.6190, \quad \rho_{3,4} = \rho_{4,3} = -0.1250\\
&\rho_{3,5} = \rho_{5,3} = 0.4464, \quad \rho_{3,6} = \rho_{6,3} = 0.2202\\
&\rho_{4,5} = \rho_{5,4} = 0.3333, \quad \rho_{4,6} = \rho_{6,4} = -0.2440\\
&\rho_{5,6} = \rho_{6,5} = 0.3333
\end{aligned}
$$

and the adjacency matrix B is

$$
\begin{pmatrix}
0.0273 & 0.0291 & 0.6269 & 0.5063 & 0.9429 & 0.6090\\
0.0291 & 0.0000 & 0.6198 & 0.4992 & 0.9268 & 0.6126\\
0.6269 & 0.6198 & 0.5198 & 0.7823 & 0.8661 & 0.7537\\
0.5063 & 0.4992 & 0.7823 & 0.2867 & 0.9000 & 0.7251\\
0.9429 & 0.9268 & 0.8661 & 0.9000 & 0.7000 & 0.9000\\
0.6090 & 0.6126 & 0.7537 & 0.7251 & 0.9000 & 0.4983
\end{pmatrix}.
$$

Then, we can get the eigenvalue of matrix B and the following computation:

$$
\begin{aligned}
&\{\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5\}\\
&\quad = \{3.9337, -0.0150, -0.2036,\\
&\qquad -0.2585, -0.3945, -1.0300\}\\
&\varsigma = 0.9/\delta(B) = 0.9/3.933693 = 0.2287927\\
&\breve{\Theta} = (\boldsymbol{E} - \varsigma B)^{-1} - \boldsymbol{E}\\
&\quad = \begin{pmatrix}
0.8336 & 0.8203 & 1.2501 & 1.1281 & 1.5520 & 1.2334\\
0.8203 & 0.8002 & 1.2294 & 1.1091 & 1.5252 & 1.2152\\
1.2501 & 1.2294 & 1.6409 & 1.5593 & 2.0471 & 1.6729\\
1.1281 & 1.1091 & 1.5593 & 1.3314 & 1.8846 & 1.5309\\
1.5520 & 1.5252 & 2.0471 & 1.8846 & 2.4174 & 2.0309\\
1.2334 & 1.2152 & 1.6729 & 1.5309 & 2.0309 & 1.5992
\end{pmatrix}\\
&\breve{\varrho}(i) = \left[\breve{\Theta} I\right]_i\\
&\quad = (6.8174, 6.6994, 9.3997, 8.5434, 11.4572, 9.2826)_i^T.
\end{aligned}
$$

Finally, we can get the feature ranking h_5, h_3, h_6, h_4, h_1, and h_2, and the top-ranked features should be selected first as a reduction set.

IV. Experiments and Results

In this section, we carry out a series of experiments to verify the efficiency and effectiveness of the proposed unsupervised feature selection algorithm for IVIS.

The configuration of the computer used for experiments is given as follows. CPU is AMD R7-5800H. The processor base frequency is 3.2 GHz, and the acceleration frequency is 4.4 GHz. The memory capacity is 16 GB. The operation system is 64-bit Windows 11. The algorithms are coded in Python and run on the PyCharm platform. We downloaded 14 datasets, as summarized in Table II, from the UCI machine learning repository and Kaggle. The feature values of the 14 datasets are real values, so these datasets are processed into interval value datasets before experiments. The preprocessing way is given as follows.

Let (U, C, F) be a real-valued information system. Then, for all $a \in C$ and $x \in U$, let $a(x)^l = a(x) - 2\text{std}$ and $a(x)^r = a(x) + 2\text{std}$, where std is the standard deviation of the corresponding information values of objects belonging to the same class as object x under the attribute [40], [41].

TABLE I
IVIS

(U, A)	h_1	h_2	h_3	h_4	h_5	h_6
x_1	$(0.22, 0.29)$	$(0.25, 0.30)$	$(0.53, 0.72)$	$(0.28, 0.30)$	$(0.33, 0.40)$	$(0.54, 0.66)$
x_2	$(0.34, 0.48)$	$(0.34, 0.49)$	$(0.72, 1.05)$	$(0.27, 0.29)$	$(0.49, 0.60)$	$(0.36, 0.44)$
x_3	$(0.18, 0.27)$	$(0.18, 0.30)$	$(0.72, 1.03)$	$(0.40, 0.43)$	$(0.41, 0.50)$	$(0.27, 0.33)$
x_4	$(0.14, 0.21)$	$(0.14, 0.21)$	$(0.26, 0.39)$	$(0.41, 0.50)$	$(0.08, 0.10)$	$(0.20, 0.24)$
x_5	$(0.35, 0.54)$	$(0.34, 0.51)$	$(0.64, 1.03)$	$(0.42, 0.44)$	$(0.16, 0.20)$	$(0.41, 0.50)$
x_6	$(0.22, 0.31)$	$(0.24, 0.33)$	$(0.44, 0.71)$	$(0.55, 0.60)$	$(0.82, 1.00)$	$(0.72, 0.88)$
x_7	$(0.25, 0.40)$	$(0.25, 0.41)$	$(0.71, 1.13)$	$(0.78, 0.81)$	$(0.65, 0.80)$	$(0.36, 0.44)$
x_8	$(0.10, 0.17)$	$(0.11, 0.18)$	$(0.36, 0.57)$	$(0.53, 0.65)$	$(0.41, 0.50)$	$(0.08, 0.10)$

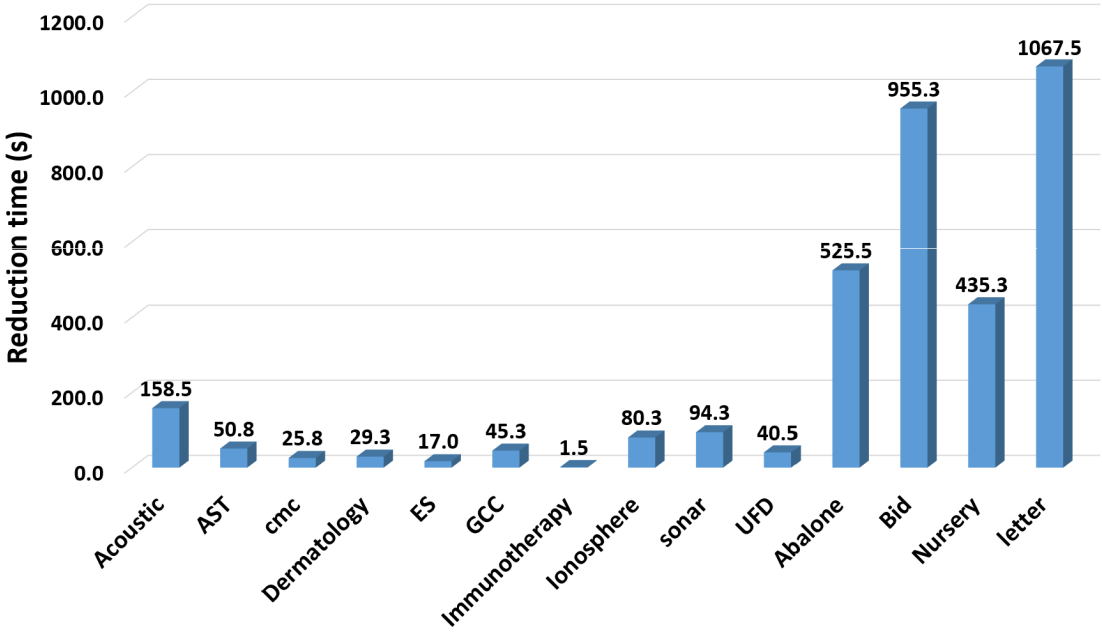

Fig. 2. Reduction time of IGUFS.

TABLE II
SUMMARY OF THE EXPERIMENTAL DATASETS

Datasets	Abbreviation	Objects	Attributes
Turkish Music Emotion	Acoustic	400	51
Autism screening data for toddlers	AST	1054	19
Contraceptive method choice	cmc	1473	9
Dermatology	Dermatology	358	35
Email spam classification	ES	1114	11
German credit card	GCC	1000	17
Immunotherapy	Immunotherapy	90	8
Ionosphere	Ionosphere	351	34
Connectionist Bench	sonar	208	61
Ultrasonic flowmeter diagostics	UFD	180	44
Abalone	Abalone	4177	9
Shill Bidding	Bid	6321	10
Nursery	Nursery	12960	9
Letter-recognition	letter	20000	17

Four classification algorithms, including KNN, SVM [42], Bayes, and Random Forest, are used to evaluate the clustering effects of our three algorithms and other eight comparative algorithms. We use the results obtained by classifiers to compare the classification results of the original dataset to obtain the classification accuracy of various algorithms. In order to avoid significant differences between the experimental classification results and the actual situation due to random parameters in clustering algorithms, the experiment is repeated five times. Eventually, the final results are compared with the average of the results of five repeated experiments.

A. Compared Algorithms

Seven feature selection (attribute reduction) algorithms are adopted as comparison algorithms, as shown in the following.

1) *Original Attributes (Original Data):* All conditional attributes in the original dataset are used to classify.
2) *Unsupervised Attribute Reduction Based on α-Approximate Equal Relation (AERAR) [23]:* The algorithm AERAR uses α-approximate equal relation to generate unsupervised attribute reduction.
3) *Infinite Feature Selection (Inf-FS) [33]:* The filtering feature selection framework considers subsets of features as paths in a graph, where a node is a feature and an edge indicates pairwise (customizable) relations among features, dealing with relevance and redundancy principles.
4) *θ-Rough Degree-Based Method (UM) [19]:* This algorithm UM uses θ-rough degree to measure uncertainty to perform unsupervised attribute reduction.

TABLE III
AVERAGE NUMBER OF SELECTED ATTRIBUTES

Data sets	Original data	HKCMI	UM	MNS3WD	AVDP	Inf-UFS	AERAR	BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
Acoustic	50	1.4	1	1	1	22	17.6	23	22	32	42
AST	18	1	1	1	1	8	18	8	8	11	15
cmc	8	6.2	1	1	1	3	8	4	3	5	6
Dermatology	34	5	1	1	1	15	15.6	15	15	22	28
ES	10	9	1	1	1	4	10	5	4	6	8
GCC	16	2	1	1	1	7	15	7	7	10	13
Immunotherapy	7	2.6	1	1	1	3	7	3	3	4	5
Ionosphere	33	4	1	1	1	14	28.8	15	14	21	28
sonar	60	1	1	1	1	21	46.2	27	21	39	51
UFD	43	2	1	1	1	19	40.4	19	19	27	36
Average	25.5	3.1	1	1	1	10.2	18.6	11.2	10.2	15.6	20.5
Abalone	8	2	1	1	1	3	–	4	3	5	6
Bid	9	2	1	1	1	4	–	4	4	5	7
Nursery	8	–	–	–	–	3	–	4	3	5	6
letter	16	–	–	–	–	3	–	7	7	10	13

5) *Hybrid-Kernel-Based Fuzzy Complementary Mutual Information Method (HKCMI) [25]:* The fuzzy complementary entropy, the corresponding complementary conditional entropy, and the fuzzy complementary mutual information are used to construct the algorithm of feature selection.
6) *Multilevel Neighborhood-Based Sequential Three-Way Decision (MNS3WD) [26]:* This is an integrative multigranularity approach to the sequential three-way decision in a neighborhood system by the evolution mechanism of data and parameters. We only consider the horizontal granularity based on γ and set the decision thresholds (α, β) in $(0.15, 0.3)$ and $(0, 0.15)$ in the following experiments.
7) *Distance Measure-Based Fuzzy Rough Set (AVDP) [27]:* This method first constructs a fuzzy rough set model based on distance measure with a fixed parameter. Then, the fixed distance parameter is replaced by a variable one to better characterize attribute reduction with fuzzy rough sets. The parameter γ is set to 0.005 in the following experiments.
8) *Neural Networks (MIV-BPNN):* The mean impact value (MIV) was selected as an indicator to evaluate the importance of each variable on the dependent variable. MIV is an indicator used to determine the magnitude of the effect of input neurons on output neurons, and its absolute size represents the relative importance of the impact.
9) The interval-valued feature selection based on the graph theory (IGUFS) method is proposed in this article.

In general, AERAR and UM are unsupervised feature selection algorithms for IVISs, and HKCMI and Inf-FS are unsupervised feature selection algorithms for single value information systems, while MNS3WD, AVDP, and MIV-BPNN are supervised feature selection algorithms for the single-valued information system. For these feature selection algorithms suitable for single-valued information systems, we convert the original IVIS into a single-valued information system by taking the midpoint value of the interval when using these algorithms.

B. Experimental Design

First, the reduction time of our algorithm IGUFS on 14 different datasets is given, and the number of reduction sets from different algorithms mentioned before is compared and analyzed. Second, we compare the classification accuracy on the original dataset, the classification accuracy of the seven comparative algorithms, and the classification accuracy obtained by the three algorithms proposed in this article. In addition, two hypothesis tests are designed to further analyze the differences between the classification accuracy of our algorithm and the comparison algorithms. Ultimately, the influence of different parameter combinations on the final clustering effect of our algorithm is explored.

There are two parameters α and β when building the algorithm IGUFS. The parameter α is introduced to control the proportion between relevance and nonredundancy principles, and the parameter β is used as a certain degree of deviation should be allowed when ordering sample values. From the definition, different combinations of α and β in algorithm IGUFS could lead to eventual unequal clustering accuracy. As a result, we set the variation range of the parameter α in the experiment from 0 to 1, where the step size of each variation is 0.1. Simultaneously, the parameter β value was set at 0.001, 0.01, 0.1, and 0, respectively. The final parameter combination of α and β is chosen based on the final classification results of the fivefold cross-validation on the training set. To facilitate the comparison with algorithm IGUFS, the percentage of selected attributes in IGUFS was set to 45%, 65%, and 85% in advance. In addition, all other comparison algorithms also use fivefold cross-validation to determine the number of features selected and classification accuracy.

C. Experimental Results

When the reduction time of the algorithm on the dataset exceeds 24 h, the experiment will be terminated due to the excessive time cost of the algorithm. The reduction time difference of our algorithm IGUFS on 14 different datasets is shown in Fig. 2. For the dataset letter with the maximum sample size of 20 000 and 16 features, the reduction time of

TABLE IV

CLASSIFICATION ACCURACY OF REDUCED DATA BASED ON KNN (%)

Data sets	Original data	HKCMI	UM	MNS3WD	AVDP	AERAR	Inf-UFS	BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
Acoustic	67.75 ± 5.88	30.25 ± 3.39	25.75 ± 2.69	25.75 ± 2.69	25.75 ± 4.51	59.25 ± 5.10	45.00 ± 4.40	46.75 ± 2.81	62.50 ± 2.96	$\mathbf{68.50 \pm 4.72}$	$\mathbf{69.00 \pm 3.20}$
AST	$\mathbf{95.16 \pm 1.18}$	63.57 ± 2.46	63.57 ± 2.46	63.57 ± 2.46	68.12 ± 3.97	$\mathbf{95.16 \pm 1.18}$	72.77 ± 2.54	68.22 ± 2.27	92.98 ± 1.09	95.06 ± 1.61	95.07 ± 1.91
cmc	57.77 ± 2.27	63.88 ± 2.08	60.96 ± 2.74	60.96 ± 2.74	62.32 ± 3.28	57.77 ± 2.27	$\mathbf{71.08 \pm 4.62}$	65.11 ± 2.83	$\mathbf{71.08 \pm 4.62}$	$\mathbf{78.14 \pm 1.58}$	$\mathbf{74.54 \pm 2.29}$
Dermatology	97.76 ± 1.13	59.51 ± 6.72	29.64 ± 5.36	29.64 ± 5.36	48.06 ± 8.14	95.81 ± 3.05	86.60 ± 1.84	71.77 ± 6.00	96.37 ± 2.86	97.49 ± 2.39	$\mathbf{98.04 \pm 2.08}$
ES	93.36 ± 1.54	93.36 ± 1.11	74.52 ± 30.54	74.52 ± 30.54	74.52 ± 30.54	93.36 ± 1.54	32.54 ± 27.22	91.83 ± 1.82	91.20 ± 0.79	$\mathbf{93.45 \pm 1.15}$	92.19 ± 1.58
GCC	96.10 ± 1.11	95.50 ± 0.89	90.40 ± 11.98	90.40 ± 11.98	90.40 ± 11.98	95.90 ± 1.11	95.40 ± 1.28	95.90 ± 1.36	95.90 ± 1.36	96.00 ± 0.84	$\mathbf{96.40 \pm 1.32}$
Immunotherapy	80.00 ± 6.67	77.78 ± 7.03	21.11 ± 8.89	21.11 ± 8.89	77.78 ± 9.30	80.00 ± 6.67	75.56 ± 10.30	72.22 ± 7.03	75.56 ± 10.30	$\mathbf{81.11 \pm 7.54}$	78.89 ± 7.54
Ionosphere	91.16 ± 3.05	78.03 ± 8.20	35.88 ± 7.85	35.88 ± 7.85	35.88 ± 7.85	91.18 ± 1.90	86.89 ± 3.77	87.74 ± 4.12	91.16 ± 2.93	$\mathbf{91.73 \pm 2.46}$	91.17 ± 2.92
sonar	79.34 ± 5.72	56.86 ± 17.33	53.38 ± 2.98	53.38 ± 2.98	53.38 ± 2.98	80.34 ± 7.67	82.71 ± 4.64	80.77 ± 4.28	$\mathbf{83.18 \pm 4.25}$	$\mathbf{85.57 \pm 4.28}$	$\mathbf{83.68 \pm 2.74}$
UFD	88.89 ± 3.93	46.67 ± 13.43	44.44 ± 5.83	44.44 ± 5.83	57.22 ± 9.56	88.33 ± 3.69	84.44 ± 5.44	71.67 ± 5.67	$\mathbf{90.00 \pm 4.16}$	$\mathbf{90.56 \pm 3.93}$	87.78 ± 3.33
Average	84.73 ± 3.25	66.54 ± 6.26	49.97 ± 8.13	49.97 ± 8.13	59.54 ± 9.21	83.71 ± 3.42	73.30 ± 6.61	75.20 ± 3.82	$\mathbf{84.99 \pm 3.53}$	$\mathbf{87.76 \pm 3.05}$	$\mathbf{86.68 \pm 2.89}$
Abalone	49.01 ± 1.87	46.23 ± 1.28	$\mathbf{57.41 \pm 0.65}$	$\mathbf{57.41 \pm 0.65}$	53.22 ± 1.58	–	49.65 ± 1.57	49.37 ± 1.11	50.42 ± 1.72	50.37 ± 2.02	50.90 ± 0.83
Bid	98.69 ± 0.24	89.32 ± 0.93	89.26 ± 0.79	89.26 ± 0.79	20.89 ± 5.03	–	97.68 ± 1.22	89.86 ± 0.73	$\mathbf{98.83 \pm 0.21}$	$\mathbf{99.30 \pm 0.23}$	$\mathbf{99.30 \pm 0.10}$
Nursery	97.08 ± 0.50	–	–	–	71.55 ± 8.37	–	87.12 ± 12.42	86.88 ± 0.86	$\mathbf{97.68 \pm 1.22}$	$\mathbf{98.02 \pm 0.26}$	$\mathbf{98.32 \pm 0.22}$
letter	93.27 ± 0.61	–	–	–	–	–	83.50 ± 0.74	70.99 ± 0.93	86.30 ± 0.41	91.52 ± 0.30	$\mathbf{93.27 \pm 0.37}$

TABLE V

CLASSIFICATION ACCURACY OF REDUCED DATA BASED ON SVM (%)

Data sets	Original data	HKCMI	UM	MNS3WD	AVDP	AERAR	Inf-UFS	BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
Acoustic	$\mathbf{78.00 \pm 3.67}$	30.00 ± 6.52	29.75 ± 6.77	29.75 ± 6.77	35.00 ± 4.51	74.75 ± 4.70	53.25 ± 1.87	52.00 ± 2.57	70.75 ± 1.27	76.50 ± 3.10	77.75 ± 4.14
AST	98.86 ± 1.33	68.31 ± 4.19	68.31 ± 4.19	68.31 ± 4.19	69.07 ± 3.97	98.86 ± 1.33	74.76 ± 1.63	69.07 ± 1.66	93.93 ± 2.54	$\mathbf{98.96 \pm 0.76}$	$\mathbf{98.86 \pm 0.88}$
cmc	63.34 ± 4.13	63.68 ± 4.72	60.89 ± 3.30	60.89 ± 3.30	59.95 ± 3.28	63.34 ± 4.13	62.66 ± 2.95	$\mathbf{68.09 \pm 1.98}$	62.66 ± 2.95	62.93 ± 2.95	62.73 ± 3.61
Dermatology	97.21 ± 2.49	68.71 ± 5.48	34.93 ± 5.30	34.93 ± 5.30	48.63 ± 8.14	94.69 ± 3.69	85.19 ± 3.01	50.56 ± 4.07	96.10 ± 3.09	$\mathbf{97.49 \pm 2.39}$	$\mathbf{97.21 \pm 3.17}$
ES	93.18 ± 1.14	92.55 ± 1.04	89.68 ± 0.82	89.68 ± 0.82	89.68 ± 0.82	93.18 ± 1.14	88.15 ± 1.33	91.47 ± 1.37	91.38 ± 1.04	$\mathbf{93.54 \pm 1.19}$	$\mathbf{93.18 \pm 0.99}$
GCC	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$
Immunotherapy	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 9.30}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{80.00 \pm 9.03}$	$\mathbf{78.89 \pm 8.89}$	$\mathbf{80.00 \pm 9.02}$
Ionosphere	94.00 ± 2.78	85.74 ± 7.74	74.94 ± 3.83	74.94 ± 3.83	74.94 ± 7.85	$\mathbf{94.29 \pm 2.86}$	91.17 ± 3.17	91.16 ± 2.93	93.44 ± 3.34	$\mathbf{94.29 \pm 2.86}$	$\mathbf{94.29 \pm 2.72}$
sonar	84.19 ± 5.25	71.21 ± 8.37	59.58 ± 8.08	59.58 ± 8.08	59.58 ± 8.08	83.19 ± 3.92	78.85 ± 2.74	82.22 ± 2.83	81.78 ± 6.09	$\mathbf{84.67 \pm 5.70}$	$\mathbf{86.12 \pm 5.63}$
UFD	63.89 ± 4.97	58.89 ± 6.43	55.00 ± 8.13	55.00 ± 8.13	49.44 ± 9.56	63.87 ± 4.97	62.78 ± 6.48	21.11 ± 4.51	$\mathbf{65.56 \pm 4.84}$	$\mathbf{66.11 \pm 4.84}$	$\mathbf{66.11 \pm 5.39}$
Average	84.79 ± 3.58	71.43 ± 5.45	64.83 ± 5.04	64.83 ± 5.04	66.15 ± 5.66	84.14 ± 3.68	77.20 ± 3.32	70.09 ± 3.19	83.19 ± 3.53	$\mathbf{84.97 \pm 3.38}$	$\mathbf{85.26 \pm 3.67}$
Abalone	55.40 ± 2.28	54.32 ± 2.15	50.95 ± 2.05	50.95 ± 2.05	54.13 ± 1.77	–	53.58 ± 1.85	51.33 ± 1.50	54.66 ± 2.12	$\mathbf{55.54 \pm 1.98}$	$\mathbf{55.61 \pm 2.32}$
Bid	98.20 ± 0.44	89.32 ± 0.93	89.50 ± 0.84	89.50 ± 0.84	89.50 ± 0.84	–	90.14 ± 0.94	90.21 ± 0.33	$\mathbf{98.77 \pm 0.43}$	$\mathbf{98.67 \pm 0.27}$	$\mathbf{98.26 \pm 0.39}$
Nursery	$\mathbf{97.98 \pm 0.40}$	–	–	–	64.54 ± 0.90	–	69.28 ± 3.41	90.14 ± 0.84	74.44 ± 1.58	78.64 ± 0.67	80.49 ± 4.53
letter	$\mathbf{92.24 \pm 0.48}$	–	–	–	–	–	76.10 ± 0.42	64.51 ± 0.73	80.25 ± 0.78	87.63 ± 0.60	91.41 ± 0.55

TABLE VI

CLASSIFICATION ACCURACY OF REDUCED DATA BASED ON BAYES (%)

Data sets	Original data	HKCMI	UM	MNS3WD	AVDP	AERAR	Inf-UFS	BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
Acoustic	75.00 ± 4.26	33.50 ± 3.42	31.00 ± 6.68	31.00 ± 6.68	35.75 ± 4.91	71.25 ± 6.71	55.75 ± 2.81	66.75 ± 2.81	$\mathbf{78.25 \pm 3.76}$	$\mathbf{76.50 \pm 3.66}$	$\mathbf{78.25 \pm 3.76}$
AST	96.96 ± 2.48	69.07 ± 4.98	69.07 ± 4.98	69.07 ± 4.98	69.07 ± 4.98	96.95 ± 2.48	72.68 ± 2.50	88.60 ± 4.70	93.26 ± 2.29	94.88 ± 1.41	$\mathbf{98.01 \pm 1.10}$
cmc	63.34 ± 3.69	$\mathbf{64.01 \pm 3.74}$	62.18 ± 4.23	62.18 ± 4.23	59.95 ± 2.99	63.00 ± 1.58	61.17 ± 2.52	63.21 ± 1.54	61.91 ± 3.78	63.20 ± 3.79	63.34 ± 4.26
Dermatology	$\mathbf{87.16 \pm 3.41}$	64.24 ± 7.13	35.49 ± 5.30	42.74 ± 9.12	70.67 ± 2.05	50.82 ± 8.56	78.17 ± 8.30	64.24 ± 7.13	82.97 ± 1.57	85.20 ± 1.60	86.88 ± 3.08
ES	92.64 ± 1.88	77.93 ± 2.86	89.68 ± 0.82	89.68 ± 0.82	87.70 ± 1.90	92.64 ± 1.88	76.61 ± 1.88	91.47 ± 1.95	91.02 ± 1.21	$\mathbf{93.09 \pm 1.98}$	92.73 ± 1.78
GCC	93.40 ± 6.25	96.20 ± 1.29	$\mathbf{96.30 \pm 1.12}$	$\mathbf{96.30 \pm 1.12}$	96.10 ± 1.46	93.50 ± 3.03	94.60 ± 0.97	95.50 ± 0.89	96.00 ± 1.22	95.20 ± 1.25	94.20 ± 1.44
Immunotherapy	80.00 ± 8.31	78.89 ± 8.39	78.89 ± 8.39	78.89 ± 8.39	78.89 ± 8.39	80.00 ± 8.31	78.89 ± 8.39	70.00 ± 8.31	$\mathbf{81.11 \pm 9.67}$	$\mathbf{82.22 \pm 9.56}$	$\mathbf{84.44 \pm 4.16}$
Ionosphere	88.02 ± 3.71	74.10 ± 7.80	74.94 ± 3.83	74.94 ± 3.83	74.94 ± 3.83	82.05 ± 3.47	81.49 ± 2.94	88.60 ± 4.70	87.46 ± 3.43	$\mathbf{88.60 \pm 2.74}$	$\mathbf{89.17 \pm 2.67}$
sonar	66.34 ± 4.33	64.52 ± 8.56	58.12 ± 5.38	58.12 ± 5.38	58.18 ± 5.60	70.17 ± 5.73	71.63 ± 4.91	73.61 ± 5.63	71.63 ± 4.67	73.58 ± 3.23	$\mathbf{71.65 \pm 2.61}$
UFD	86.67 ± 6.43	70.56 ± 4.16	66.11 ± 4.08	66.11 ± 4.08	62.22 ± 8.16	86.67 ± 6.43	$\mathbf{87.22 \pm 6.71}$	76.67 ± 9.40	$\mathbf{87.22 \pm 6.71}$	$\mathbf{88.33 \pm 6.43}$	$\mathbf{87.22 \pm 6.71}$
Average	82.95 ± 4.48	69.30 ± 5.23	66.18 ± 4.48	66.18 ± 4.48	69.35 ± 4.43	78.71 ± 4.82	73.14 ± 4.27	79.26 ± 4.82	$\mathbf{83.08 \pm 3.83}$	$\mathbf{84.08 \pm 3.57}$	$\mathbf{84.59 \pm 3.16}$
Abalone	51.83 ± 1.19	51.86 ± 1.61	49.32 ± 1.88	49.32 ± 1.88	53.72 ± 1.80	–	53.65 ± 1.64	49.37 ± 1.11	51.33 ± 1.39	$\mathbf{54.54 \pm 1.66}$	52.86 ± 1.40
Bid	97.06 ± 0.31	88.55 ± 0.65	88.39 ± 0.64	88.39 ± 0.64	89.32 ± 0.93	–	87.60 ± 0.76	89.86 ± 0.73	$\mathbf{97.18 \pm 0.71}$	$\mathbf{97.39 \pm 0.56}$	$\mathbf{98.02 \pm 0.48}$
Nursery	$\mathbf{99.27 \pm 0.23}$	–	–	–	64.54 ± 0.90	–	56.87 ± 4.19	86.88 ± 0.86	53.67 ± 1.19	67.35 ± 1.36	70.76 ± 1.15
letter	95.52 ± 0.45	–	–	–	–	–	48.65 ± 0.75	$\mathbf{70.99 \pm 0.93}$	49.56 ± 0.45	55.81 ± 0.61	58.70 ± 0.40

TABLE VII

CLASSIFICATION ACCURACY OF REDUCED DATA BASED ON RANDOM FOREST (%)

Data sets	Original data	HKCMI	UM	MNS3WD	AVDP	AERAR	Inf-UFS	BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
Acoustic	77.75 ± 5.33	34.50 ± 5.28	27.00 ± 4.08	27.00 ± 4.08	29.25 ± 5.84	78.25 ± 4.23	53.75 ± 2.62	74.25 ± 5.79	$\mathbf{78.75 \pm 3.71}$	$\mathbf{80.00 \pm 3.79}$	$\mathbf{79.00 \pm 4.57}$
AST	$\mathbf{100.00 \pm 0.00}$	63.47 ± 1.56	63.57 ± 1.72	63.57 ± 1.72	69.07 ± 4.98	$\mathbf{100.00 \pm 0.00}$	74.00 ± 2.77	91.46 ± 0.67	92.88 ± 1.56	96.68 ± 1.44	$\mathbf{100.00 \pm 0.00}$
cmc	64.22 ± 3.10	62.45 ± 3.37	62.18 ± 4.23	62.18 ± 4.23	59.95 ± 2.99	64.16 ± 1.95	65.38 ± 3.49	65.44 ± 1.00	$\mathbf{65.78 \pm 3.80}$	$\mathbf{65.78 \pm 3.85}$	65.10 ± 3.49
Dermatology	97.21 ± 2.49	67.03 ± 6.93	34.93 ± 5.30	34.93 ± 5.30	48.63 ± 8.40	91.63 ± 5.19	84.07 ± 6.20	90.23 ± 3.60	95.54 ± 3.55	$\mathbf{97.49 \pm 1.62}$	$\mathbf{97.49 \pm 2.23}$
ES	93.00 ± 1.51	92.46 ± 1.04	89.68 ± 0.82	89.68 ± 0.82	87.70 ± 1.90	92.82 ± 1.09	88.15 ± 1.33	91.83 ± 1.55	91.56 ± 1.24	$\mathbf{93.45 \pm 1.08}$	$\mathbf{93.54 \pm 1.19}$
GCC	96.30 ± 0.75	95.60 ± 1.16	96.30 ± 1.12	96.30 ± 1.12	93.50 ± 1.14	$\mathbf{96.50 \pm 0.89}$	95.20 ± 1.21	95.90 ± 1.43	96.10 ± 1.24	95.80 ± 1.03	96.30 ± 1.12
Immunotherapy	$\mathbf{87.78 \pm 9.56}$	74.44 ± 8.31	78.89 ± 8.89	78.89 ± 8.89	66.67 ± 9.94	84.44 ± 8.16	76.67 ± 8.89	71.11 ± 9.56	80.00 ± 8.31	$\mathbf{87.78 \pm 8.89}$	87.78 ± 9.56
Ionosphere	93.44 ± 3.46	85.75 ± 5.72	74.94 ± 3.83	74.94 ± 3.83	74.94 ± 3.83	92.87 ± 3.01	90.61 ± 3.04	92.87 ± 2.40	$\mathbf{94.02 \pm 3.18}$	$\mathbf{94.30 \pm 3.73}$	$\mathbf{94.87 \pm 4.20}$
sonar	79.85 ± 6.39	57.84 ± 13.24	50.51 ± 3.45	50.51 ± 3.45	53.37 ± 7.76	77.44 ± 8.59	83.62 ± 4.24	86.49 ± 4.60	86.04 ± 3.58	$\mathbf{86.57 \pm 4.39}$	84.20 ± 7.08
UFD	88.89 ± 7.86	66.11 ± 9.20	61.67 ± 7.74	61.67 ± 7.74	63.33 ± 7.74	$\mathbf{88.89 \pm 6.09}$	85.00 ± 5.44	80.00 ± 5.67	$\mathbf{88.89 \pm 6.09}$	$\mathbf{92.78 \pm 3.77}$	$\mathbf{91.11 \pm 5.67}$
Average	87.84 ± 4.05	69.97 ± 5.58	63.97 ± 4.12	63.97 ± 4.12	64.64 ± 5.45	86.70 ± 3.92	79.65 ± 3.92	83.96 ± 3.63	86.96 ± 3.63	$\mathbf{89.06 \pm 3.36}$	$\mathbf{88.94 \pm 3.91}$
Abalone	54.27 ± 1.23	50.04 ± 1.47	49.32 ± 1.11	49.32 ± 1.11	47.19 ± 1.21	–	50.73 ± 0.50	51.33 ± 1.50	53.77 ± 2.30	53.05 ± 1.08	$\mathbf{54.68 \pm 1.77}$
Bid	99.45 ± 0.26	88.85 ± 0.35	91.22 ± 0.69	91.22 ± 0.69	83.18 ± 0.43	–	88.50 ± 0.74	90.21 ± 0.33	98.80 ± 0.19	$\mathbf{99.56 \pm 0.19}$	$\mathbf{99.68 \pm 0.19}$
Nursery	$\mathbf{99.27 \pm 0.23}$	–	–	–	57.48 ± 18.41	–	69.14 ± 3.62	90.14 ± 0.84	77.61 ± 0.56	78.42 ± 0.55	84.29 ± 0.93
letter	$\mathbf{95.52 \pm 0.45}$	–	–	–	–	–	85.86 ± 0.52	64.51 ± 0.73	87.67 ± 0.34	93.30 ± 0.41	94.62 ± 0.26

IGUFS is 1067.5 s or about 18 min. However, there are five comparison algorithms with a reduction time of more than 24 h on this dataset.

The average numbers of selected features under comparative algorithms and our algorithms are given in Table III, where we can know that these feature selection methods could reduce attributes effectively on different levels. However, the advantage of IGUFS over other algorithms is that the number or proportion of selected features can be easily changed according to actual needs. For example, if the goal of higher classification accuracy for the dataset is pursued, for some datasets with relatively few attributes, such as Immunotherapy, the proportion of selected attributes can be appropriately increased like 65% and 85%. For datasets with a large number of attributes, such as sonar and Acoustic, the proportion of selected attributes can be appropriately decreased like 45%. If it is to maximize the efficiency of reduction, that is, a large number of attributes are reduced, but the reduction of classification accuracy is small, the proportion of selected attributes can be appropriately decreased like 45% or lower to determine the final option. For instance, 45% seems a better choice than 85% for dataset Dermatology on account of that the classification accuracy is only reduced from 98.04% to 96.37% on KNN when the number of attributes is reduced by 13 (or the ratio of attribute reduction is reduced by 40% of the original dataset). However, for most feature selection methods, such as AERAR and UM, they can only give a unique reduction set result and cannot be adjusted according to actual needs.

Table IV–VII show the classification accuracy of the original data and the reduced datasets based on these seven algorithms and our proposed three kinds of algorithms on KNN, SVM, Bayes, and Random Forest, respectively. The bolded numbers in each row from Table IV–VII are the highest classification accuracy obtained by a total of nine algorithms under the corresponding dataset. We regard these datasets (Acoustic, AST, cmc, Dermatology, ES, GCC, Immunotherapy, Ionosphere, sonar, and UFD) as normal datasets and the remaining four datasets (Abalone, Bid, Nursery, and letter) as large sample size datasets. Then, among 40 records on normal datasets in Tables IV–VII, the algorithm IGUFS (45%) has 13 records to achieve the best clustering accuracy, and the algorithm IGUFS (65%) has 28 records to achieve the best classification accuracy, while the algorithm IGUFS (85%) has 25 records to achieve the best classification accuracy. However, for algorithms HKCMI, UM, MNS3WD, AVDP, AERAR, Inf-UFS, and BPNN, only three, three, three, two, seven, four, and three records achieve the best accuracy. What is more, the average classification accuracy of algorithm IGUFS (85%) and the average classification accuracy of algorithm IGUFS (65%) on ten normal datasets tend to have a better performance against selected comparative algorithms. For the four large sample datasets selected in the experiment, among 16 records (ignoring the algorithms without classification accuracy, comparing the algorithms with classification accuracy), the algorithm IGUFS (45%) has four records to achieve the best clustering accuracy, and the algorithm IGUFS (65%) has seven records to achieve the best classification accuracy, while the algorithm IGUFS (85%) has eight records to achieve the best classification accuracy. The original dataset has only five records to achieve the best classification accuracy. Most obviously, our proposed algorithm IGUFS can improve or maintain the classification accuracy of the original data efficiently from Table IV to VII. As a result, algorithm IGUFS has a degree of superiority for feature selection of IVISs on classification tasks.

D. Statistical Testing

In order to further compare the experimental results of different algorithms, two statistical test methods, the Friedman test [43] and the Wilcoxon test, are selected to verify the validity of the algorithm comparison.

The null hypothesis of the Friedman test, as a nonparametric statistical test method, is that all experimental algorithms have equivalent classification performance. The formula is defined as

$$F_F = \frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2}$$

$$\chi_F^2 = \frac{12T}{s(s+1)}\left(\sum_{i=1}^{s} R_i^2 - \frac{s(s+1)^2}{4}\right)$$

where T and s are the numbers of experimental datasets and experimental algorithms, respectively, and R_i represents the average ranking value of the classification accuracy results of algorithm i on different classifiers. Since some comparison algorithms fail to generate results on large sample datasets within 24 h, subsequent analyses in this part are performed on normal datasets mentioned above.

Table VIII represents the average ranking of the classification accuracy results of the algorithms AERAR, Inf-UFS, UM, HKCMI, MNS3WD, AVDP, MIV-BPNN, and IGUFS on classifier KNN, SVM, Bayes, and Random Forest, respectively. In the comparison experiments where $T = 10$ and $s = 10$, the Friedman values of the algorithms classification performance can be obtained from the average ranking results in Table IX.

As shown in Table IX, when the significance level is 0.05, if the P value is greater than 0.05, it means that the null hypothesis is not rejected. If the P value is less than 0.05, it means that the null hypothesis is rejected. From Table IX, we know that the P values of Friedman tests are all less than the significance level of 0.05, so the null hypothesis can be rejected. In addition, the Wilcoxon test can further analyze the relative performance and differences of all the compared algorithms. Three sets of hypothesis tests are performed as follows, corresponding to the three selection ratios of the reduction algorithm proposed in this article. For each hypothesis test, the null hypothesis is that the classification accuracy obtained by the method in this article is less than or equal to the other seven selected comparative methods, and the alternative hypothesis is that the classification accuracy obtained by the method in this article is greater than the other seven methods. The P values are shown in Table X–XIII.

When the null hypothesis is rejected, it indicates that the classification accuracy of our model is higher than other comparison models. It can be seen from the table that, when

TABLE VIII
AVERAGE RANKING OF CLASSIFICATION ACCURACIES OF ALGORITHMS

	HKCMI	UM	MNS3WD	AVDP	Inf-UFS	AERAR	MIV-BPNN	IGUFS(45%)	IGUFS(65%)	IGUFS(85%)
KNN	4.75	2.10	2.10	3.35	5.50	7.35	5.70	**7.45**	**10.20**	**9.40**
SVM	5.15	2.95	2.95	2.75	5.20	7.80	5.50	7.05	**9.10**	**9.20**
Bayes	4.55	3.60	3.60	3.25	4.65	6.50	6.65	**7.40**	**8.95**	**9.40**
Random Forest	3.80	3.05	3.05	2.20	4.90	7.65	6.15	**7.85**	**9.60**	**9.60**

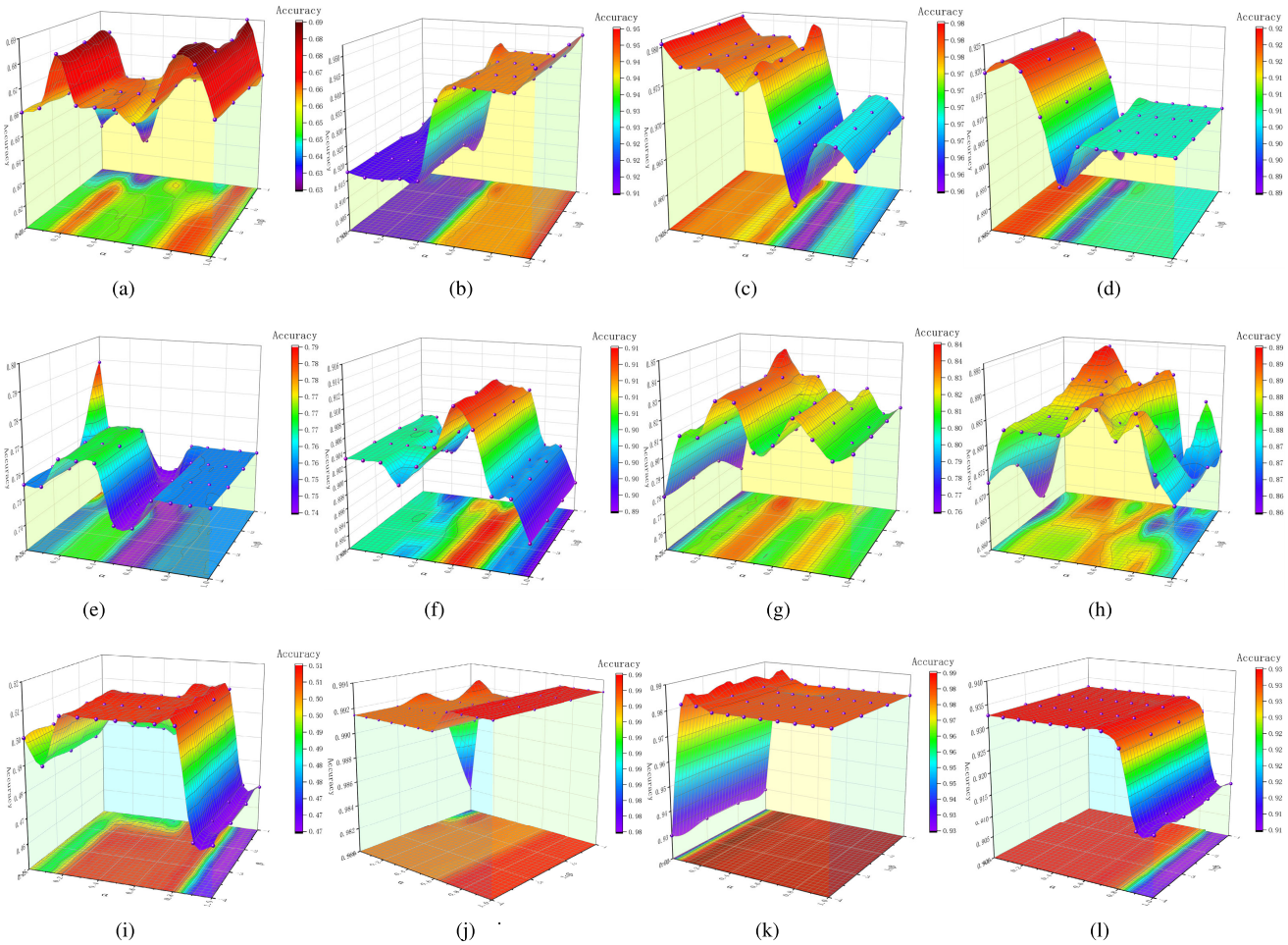

Fig. 3. (a) Acoustic-KNN. (b) AST-KNN. (c) Dermatology-KNN. (d) ES-KNN. (e) Immunotherapy-KNN. (f) Ionosphere-KNN. (g) sonar-KNN. (h) UFD-KNN. (i) Abalone-KNN. (j) Bid-KNN. (k) Nursery-KNN. (l) letter-KNN.

TABLE IX
RESULT OF THE FRIEDMAN TEST

	Friedman Value	χ_F^2	P Value
KNN	34.28	71.28	2.50×10^{-11}
SVM	22.71	64.46	5.14×10^{-10}
Bayes	8.89	44.72	2.44×10^{-6}
Random Forest	29.54	68.98	6.97×10^{-11}

TABLE X
P VALUE OF THE WILCOXON TEST ON KNN

	45%	65%	85%
HKCMI	< 0.01	< 0.01	< 0.01
UM	< 0.01	< 0.01	< 0.01
MNS3WD	< 0.01	< 0.01	< 0.01
AVDP	< 0.01	< 0.01	< 0.01
AERAR	0.20	< 0.01	0.07
Inf-UFS	< 0.01	< 0.01	< 0.01
MIV-BPNN	< 0.01	< 0.01	< 0.01

the selection ratio of feature selection of the model proposed in this article is determined to be 65%, whether it is the KNN classifier or the SVM classifier or the Bayes classifier or the Random tree classifier, the test result is to reject the null hypothesis, indicating that the classification accuracy of our model is significantly higher than models from other comparison algorithms. When the proportion of selected features is 45% and 85%, only the hypothesis test with the comparison algorithm AERAR and Inf-UFS fails, which may be caused by the small sample size.

TABLE XI
P VALUE OF THE WILCOXON TEST ON SVM

	45%	65%	85%
HKCMI	0.02	0.01	< 0.01
UM	< 0.01	< 0.01	< 0.01
MNS3WD	< 0.01	< 0.01	< 0.01
AVDP	< 0.01	< 0.01	< 0.01
AERAR	0.87	0.04	0.03
Inf-UFS	< 0.01	< 0.01	< 0.01
MIV-BPNN	0.04	0.03	0.02

TABLE XII
P VALUE OF THE WILCOXON TEST ON BAYES

	45%	65%	85%
HKCMI	< 0.01	< 0.01	< 0.01
UM	< 0.01	< 0.01	< 0.01
MNS3WD	< 0.01	< 0.01	< 0.01
AVDP	< 0.01	< 0.01	< 0.01
AERAR	< 0.01	< 0.01	< 0.01
Inf-UFS	0.09	< 0.01	< 0.01
MIV-BPNN	0.08	0.03	0.03

TABLE XIII
P VALUE OF THE WILCOXON TEST ON RANDOM FOREST

	45%	65%	85%
HKCMI	< 0.01	< 0.01	< 0.01
UM	< 0.01	< 0.01	< 0.01
MNS3WD	< 0.01	< 0.01	< 0.01
AVDP	< 0.01	< 0.01	< 0.01
AERAR	< 0.01	< 0.01	< 0.01
Inf-UFS	0.41	0.02	< 0.01
MIV-BPNN	0.02	< 0.01	0.01

E. Experimental Parameter

For the purpose of analyzing the influence of different parameter combinations of α and β on the classification accuracy of the algorithm IGUFS proposed in this article, we draw the classification accuracy results obtained by combining the previous parameter combinations of α and β for the algorithm IGUFS on the classifier KNN on 12 datasets, as shown in Fig. 3. It can be seen from the figures that it is meaningful to set different combinations of parameters of α and β because different parameter combinations will lead to significant differences in the final classification accuracy results obtained. In addition, the parameter combinations to obtain the optimal classification results tend to be inconsistent for different datasets. For example, we can see that, for datasets Acoustic, AST, and Nursery, IGUFS needs to choose larger α to achieve better performance, while, for datasets Dermatology, ES, and Immunotherapy, IGUFS needs to choose smaller α. When β approaches 0.1, IGUFS would achieve better performance for most datasets. For most datasets, the optimal classification accuracy of the algorithm can be achieved by different parameter combinations rather than just one parameter combination. For each dataset, we can choose the appropriate combination of parameters α and β to achieve relatively optimal performance. For instance, on the dataset sonar with the combination of $\alpha = 0.3$ and $\beta = 0.1$, the proposed algorithm IGUFS could achieve relatively optimal performance. In conclusion, different parameter combinations could cause significant differences in the eventual classification accuracy results obtained. Hence, we must choose the appropriate parameter combination of α and β to achieve the relatively optimal classification accuracy results of the algorithm IGUFS.

In summary, the algorithm IGUFS is efficient and effective for the feature selection for IVISs under four clustering algorithms of KNN, SVM, Bayes, and Random Forest.

V. CONCLUSION

This article proposes a new feature ranking procedure with the principles of relevancy and nonredundancy by using the graph theory and the algebra of matrix for IVISs, and we set three different selected proportions to construct our feature selection algorithms. The experiments are performed on 14 public datasets with seven comparative algorithms. The results of the experiments verify that the proposed method is an effective and efficient scheme for reducing interval-valued attributes.

This work studies feature selection algorithms for static IVISs, while, for the dynamic interval-valued datasets, it is our present research direction to get an efficient attribute reduction mechanism for dynamic IVISs based on the research of this article. In addition, there is no automatic mechanism to decide the number of features to choose when we pick the top-ranked features to be the reduction subset. Therefore, studying how to determine the number of features more perfectly also belongs to future work.

REFERENCES

[1] D. Guru, B. Kiranagi, and P. Nagabhushan, "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns," *Pattern Recognit. Lett.*, vol. 25, no. 10, pp. 1203–1213, 2004.

[2] P. D'Urso and J. Leski, "Fuzzy *c*-ordered medoids clustering for interval-valued data," *Pattern Recognit.*, vol. 58, pp. 49–67, 2016.

[3] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 551–577, Dec. 2017.

[4] H. Zeng and Y.-M. Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[5] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[6] A. Bommert et al., "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Statist. Data Anal.*, vol. 143, Mar. 2020, Art. no. 106839.

[7] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," *Eng. Appl. Artif. Intell.*, vol. 32, pp. 112–123, Jun. 2014.

[8] M. A. Anitha and K. K. Sherly, "A novel forward filter feature selection algorithm based on maximum dual interaction and maximum feature Relevance(MDIMFR) for machine learning," in *Proc. Int. Conf. Adv. Comput. Commun. (ICACC)*, Oct. 2021, pp. 1–7.

[9] A. Wang, H. Liu, J. Liu, H. Ding, J. Yang, and G. Chen, "Stable and accurate feature selection from microarray data with ensembled fast correlation based filter," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2996–2998.

[10] M. Labani et al., "An evolutionary based multi-objective filter approach for feature selection," in *Proc. World Congr. Comput. Commun. Technol.*, 2017, pp. 151–154.

[11] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, Jan. 2018.
[12] G. Manikandan, E. Susi, and S. Abirami, "Feature selection on high dimensional data using wrapper based subset selection," in *Proc. 2nd Int. Conf. Recent Trends Challenges Comput. Models (ICRTCCM)*, Feb. 2017, pp. 320–325.
[13] R. Chakraborty and N. R. Pal, "Feature selection using a neural framework with controlled redundancy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 35–50, Jan. 2015.
[14] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 703–715, May 2019.
[15] F. Nie et al., "Unsupervised feature selection with constrained $\ell_{2,0}$ and optimized graph," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1702–1713, Sep. 2022.
[16] X. Du, F. Nie, W. Wang, Y. Yang, and X. Zhou, "Exploiting combination effect for unsupervised feature selection by $\ell_{2,0}$ norm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 201–214, Jun. 2018.
[17] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.
[18] H. Zapata et al., "Interval-valued implications and interval-valued strong equality index with admissible orders," *Int. J. Approx. Reasoning*, vol. 88, pp. 91–109, Sep. 2017.
[19] J. Dai, W. Wang, and J. Mi, "Uncertainty measurement for interval-valued information systems," *Inf. Sci.*, vol. 251, pp. 63–78, Dec. 2013.
[20] X. Zhang, C. Mei, D. Chen, and J. Li, "Multi-confidence rule acquisition and confidence-preserved attribute reduction in interval-valued decision systems," *Int. J. Approx. Reasoning*, vol. 55, no. 8, pp. 1787–1804, Nov. 2014.
[21] J. Dai et al., "Attribute reduction in interval-valued information systems based on information entropies," *Frontiers Inf. Technol. Electron. Eng.*, vol. 17, no. 9, pp. 919–928, 2016.
[22] W. Shu, W. Qian, Y. Xie, and Z. Tang, "An efficient uncertainty measure-based attribute reduction approach for interval-valued data with missing values," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 27, no. 6, pp. 931–947, Dec. 2019.
[23] X. Liu, J. Dai, J. Chen, and C. Zhang, "Unsupervised attribute reduction based on α-approximate equal relation in interval-valued information systems," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 9, pp. 2021–2038, Sep. 2020.
[24] J. Dai, Y. Liu, J. Chen, and X. Liu, "Fast feature selection for interval-valued data through kernel density estimation entropy," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 12, pp. 2607–2624, Dec. 2020.
[25] Z. Yuan et al., "Fuzzy complementary entropy using hybrid-kernel function and its unsupervised attribute reduction," *Knowl.-Based Syst.*, vol. 231, Nov. 2021, Art. no. 107398.
[26] X. Yang et al., "A multilevel neighborhood sequential decision approach of three-way granular computing," *Inf. Sci.*, vol. 538, pp. 119–141, Oct. 2020.
[27] C. Wang et al., "Fuzzy rough set-based attribute reduction using distance measures," *Knowl.-Based Syst.*, vol. 164, pp. 205–212, Jan. 2019.
[28] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[29] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2013.
[30] C. Hou, F. Nie, D. Yi, and D. Tao, "Discriminative embedded clustering: A framework for grouping high-dimensional data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1287–1299, Jun. 2014.
[31] L. Zhang, L. Qiao, and S. Chen, "Graph-optimized locality preserving projections," *Pattern Recognit.*, vol. 43, no. 6, pp. 1993–2002, 2010.
[32] Y. Qian, J. Liang, and C. Dang, "Interval ordered information systems," *Comput. Math. Appl.*, vol. 56, no. 8, pp. 1994–2009, 2008.
[33] G. Roffo et al., "Infinite feature selection: A graph-based feature filtering approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4396–4410, Jun. 2021.
[34] G. Bierman, "Power-series evaluation of transition and covariance matrices," *IEEE Trans. Autom. Control*, vol. AC-17, no. 2, pp. 228–232, Apr. 1972.
[35] N. Young, "The rate of convergence of a matrix power series," *Linear Algebra Appl.*, vol. 35, pp. 261–278, Feb. 1981.
[36] E. Fieller, H. Hartley, and E. Pearson, "Tests for rank correlation coefficients. I," *Biometrika*, vol. 44, pp. 470–481, Dec. 1957.
[37] R. Graham, D. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*. Reading, MA, USA: Addison-Wesley, 1994.
[38] E. Bergshoeff, "Ten physical applications of spectral zeta functions," *Classical Quantum Gravity*, vol. 13, no. 7, p. 034, 1997.
[39] J. Powers and M. Sen, *Mathematical Methods in Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[40] Y. Leung et al., "A rough set approach for the discovery of classifcation rules in interval-valued information systems," *Int. J. Approx. Reasoning*, vol. 47, no. 2, pp. 233–246, 2008.
[41] Y. Zhang et al., "Incremental updating of rough approximations in interval-valued information systems under attribute generalization," *Inf. Sci.*, vol. 373, pp. 461–475, Dec. 2016.
[42] N. Gerhardt et al., "Quality assessment of olive oils based on temperature-ramped HS-GC-IMS and sensory evaluation: Comparison of different processing approaches by LDA, kNN, and SVM," *Food Chem.*, vol. 286, pp. 307–308, Apr. 2019.
[43] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, pp. 86–92, Mar. 1940.

Weihua Xu received the M.Sc. degree from the School of Mathematics and Information Sciences, Guangxi University, Nanning, China, in 2004, and the Ph.D. degree from the School of Sciences, Xi'an Jiaotong University, Xi'an, China, in 2007.

He is currently a Professor with the College of Artificial Intelligence, Southwest University, Chongqing, China. He has published four monographs and over 180 articles in international journals. His current research interests include granular computing, approximate reasoning, fuzzy sets, rough sets, concept lattices, cognitive computing, and evolutionary computing.

Dr. Xu serves on the editorial boards of several international journals.

Man Huang received the B.Sc. degree from the Zhongnan University of Economics and Law, Wuhan, China, in 2021. She is currently pursuing the M.Sc. degree with the School of Artificial Intelligence, Southwest University, Chongqing, China.

Her research interests cover feature selection and granular computing.

Zongying Jiang is currently pursuing the B.Sc. degree with the College of Artificial Intelligence, Southwest University, Chongqing, China.

His current research interests include fuzzy sets and granular computing.

Yuhua Qian (Member, IEEE) received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently the Director of the Institute of Big Data Science and Industry, Shanxi University, where he is also a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education. He is best known for artificial intelligence, machine learning, and machine vision. He has authored over 100 articles on these topics in international journals.