IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 February 2024, accepted 5 April 2024, date of publication 12 April 2024, date of current version 19 April 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3387978*

RESEARCH ARTICLE

# Maximum Information Coefficient Feature Selection Method for Interval-Valued Data

**XIAOBO QI[1,2], JINYU SONG[3], HUI QI[1,2], AND YING SHI[1,4]**
[1]School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China
[2]Shanxi Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong 030619, China
[3]School of Mathematics and Statistics, Taiyuan Normal University, Jinzhong 030619, China
[4]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

Corresponding author: Xiaobo Qi (xbqi@tynu.edu.cn)

This work was supported in part by Shanxi Patent Transformation Special Programs under Grant 202302009 and Grant 202302012, in part by the Basic Research Program (Free Exploration) of Shanxi Province under Grant 20210302123334, and in part by Taiyuan Normal University Achievement Transformation and Technology Transfer Base under Grant 2023P003.

**ABSTRACT** The feature selection for interval-valued data (IVD) aims to identify representative features from a large set of features, which can reduce the model complexity, minimize the training time, and enhance the generalization ability of the model. Addressing the inter-feature correlations in IVD, we propose a feature selection method called the maximum information coefficient for interval-valued data (IVD_MIC). First, the method balances the relationship between the midpoint and radius of IVD with an adjustment factor, constructing the interval-valued data unified representation frame (URF). Based on the URF, the method measures the degree of correlation between two features by calculating the maximum information coefficient, and obtains the maximum information coefficient matrix for IVD. Then the features with strong correlation are progressively removed from three perspectives(row, column, and both row and column), generating a series of corresponding candidate feature subsets. Finally, IVD_MIC is validated on candidate feature subsets to obtain the final classification accuracy and optimal feature subset. The experiment results on synthetic and real-world datasets with different classifiers demonstrate that the overall performance of IVD_MIC surpasses other methods. The average accuracy of IVD_MIC is higher, improving by 0.23%, 0.53% and 0.45% compared to the second-best method on LIBSVM, CART Tree and KNN, respectively.

**INDEX TERMS** Feature selection, interval-valued data, maximum information coefficient, optimal feature subset, unified representation frame.

## I. INTRODUCTION

With the onset of big data era, there has been a significant surge in data volume, accompanied by a constant enrichment of data types. IVD represents a common type of quantitative symbolic data characterized by each attribute feature being not a single numerical value but an interval range. The unique structure of IVD poses a challenge to traditional analysis methods, which are not directly applicable. Currently, researchers often resort to using the midpoint, upper and lower bounds, or median radius of IVD to represent it. Subsequently, these representations are explored in the context of principal component analysis [1], [2], [3], discriminant analysis [4], [5], [6], regression analysis [7], [8], [9], and cluster analysis [10], [11]. However, these traditional research methods for IVD have limitations, as they may either sacrifice location information, size information, or result in an increased number of features.

To address the challenges mentioned above, it is crucial to preserve key information within IVD while minimizing a significant increase in the number of features. Feature selection plays a vital role in achieving this objective by

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

eliminating redundant or irrelevant features, thereby reducing the overall feature count. However, the unique structure of IVD necessitates a different method for feature selection compared to traditional methods. Reference [12] introduced a method that defined a similarity boundary to measure the similarity between features and label. This similarity boundary was then utilized to formulate a boundary-based target function, optimizing the objective function to evaluate feature importance. Subsequently, feature selection was performed. C. H. Guo and Y. C. Liu employed the median-radius distance to measure the similarity of IVD. They proposed a maximum similarity optimization model between sample points and sample label centers to estimate the feature weight of IVD, followed by feature selection [13]. Reference [14] used the midpoint of IVD for numerical representation, employed KL divergence to measure the correlation between two interval-valued feature data, and introduced a cost-sensitive interval-valued feature selection method. It was noteworthy that these methods relied on converting IVD into numerical representations before feature selection, potentially resulting in information loss or increased data analysis complexity, which might subsequently impact classifier classification performance.

To address the challenges of feature selection for IVD and the potential loss or multiplication of features, a novel method called maximum information coefficient feature selection for IVD is proposed. The method begins by using an adjustment factor to balance the relationship between the midpoint and radius of IVD, constructing IVD under the URF. Within this URF, the degree of correlation between features is measured by calculating the maximum information coefficient.

The experimental analysis is conducted from various perspectives of the maximum information coefficient matrix, resulting in a series of candidate feature subsets. Subsequently, these candidate feature subsets are experimentally validated using different classifiers to obtain classification accuracy and identify the optimal feature subset.

On synthetic and real-world datasets, IVD_MIC demonstrates superior accuracy compared to comparison methods, particularly excelling on real-world datasets. Specifically, on LIBSVM, CART Tree, and KNN, the average unit feature accuracy of IVD_MIC is the highest, followed by Spearman, with Pearson ranking the lowest. The average unit feature accuracy of IVD_MIC surpasses the Spearman method by 1.43%, 1.46%, and 1.49%, and outperforms the Pearson method by 3.07%, 3.04%, and 3.08%. These results indicate the superiority of IVD_MIC over other methods.

The main contributions of this paper are as follows:

1. By adjusting the factor to balance the relationship between the midpoint and radius in IVD, constructing the URF for IVD effectively balances the feature correlations within the data.

2. Utilizing the maximum information coefficient to measure inter-feature correlations and progressively eliminating highly correlated features to generate an optimal feature subset enhances the model's classification accuracy and generalization ability.

3. A feature selection method named interval-valued data maximum information coefficient(IVD_MIC) is proposed, demonstrating superior average accuracy over other methods across various classifiers and datasets.

The rest of this paper is organized as follows:

Introduction of IVD_MIC under the IVD and URF (Section II): This section provides an overview of IVD within the URF. It specifically elucidates the key concepts and principles behind the proposed maximum information coefficient feature selection method for IVD.

Experimental datasets and settings (Section III): Section III outlines the experimental datasets used in the study and provides detailed information on the experimental settings. Additionally, it presents the results of the experiments along with a thorough analysis.

Conclusion and future work (Section IV): The final section of the paper offers the conclusion drawn from the study and outlines plans for future research and improvements.

## II. MAXIMUM INFORMATION COEFFICIENT FEATURE SELECTION METHOD FOR INTERVAL-VALUED DATA

This section introduces IVD and URF. Then describes the theoretical foundations of IVD_MIC and the main steps of IVD_MIC in detail.

### A. DEFINITION OF INTERVAL-VALUED DATA

The relevant definitions of IVD under the URF are as follows:

*Definition 1 (Interval-Valued Data Unit [15]):* Let $u = [u^-, u^+]$ be an interval-valued data unit, where $u^-, u^+ \in R$ and $u^- \leq u^+$, $u^-$ and $u^+$ are called the lower and upper boundary respectively. If $u^- = u^+$, $u$ becomes a general single value, that is, $u = u^- = u^+$.

*Definition 2 (Interval-Valued Matrix [15]):* Denote $U = [u_{ij}]$ as an $n \times p$ interval-valued matrix $U$, *i.e.*,

$$U = (U_1, U_2, \cdots, U_p) = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{np} \end{pmatrix}, \quad (1)$$

where $U_j = ([u_{1j}^-, u_{1j}^+], [u_{2j}^-, u_{2j}^+], \cdots, [u_{nj}^-, u_{nj}^+])^T$ represents the $j$th feature vectors with all samples, where $u_{ij} = [u_{ij}^-, u_{ij}^+]$ as an interval-valued data unit.

*Definition 3 (Midpoint and Radius of Interval-Valued Data Unit [15]):* Let $u^m$ and $u^r$ be the midpoint and radius of interval-valued data unit $u$, defined as

$$u^m = \frac{u^- + u^+}{2}. \quad (2)$$

$$u^r = \frac{u^+ - u^-}{2}. \quad (3)$$

According to the above definitions, let $u^{mr}$ be the midpoint-radius value, it can be represented as:

$$u^{mr} = \alpha u^m + (1 - \alpha) u^r, \quad (4)$$

where $\alpha \in [0, 1]$, $\alpha$ can be regarded as the adjustment factor of IVD unit, which is used to balance the relationship betwe-en the midpoint and radius of the IVD unit. The midpoint-radius matrix is constructed as:

$$U^{\mathrm{mr}} = (U_1^{mr}, U_2^{mr}, \cdots, U_p^{mr}) = \begin{pmatrix} u_{11}^{mr} & u_{12}^{mr} & \cdots & u_{1p}^{mr} \\ u_{21}^{mr} & u_{22}^{mr} & \cdots & u_{2p}^{mr} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1}^{mr} & u_{n2}^{mr} & \cdots & u_{np}^{mr} \end{pmatrix}. \quad (5)$$

where $U_j^{mr} = (u_{1j}^{mr}, u_{2j}^{mr}, \cdots, u_{nj}^{mr})^T$ represents the $j$th feature vectors with all samples under the URF.

### B. CONSTRUCTION OF THE IVD_MIC METHOD

The relevant definitions of the IVD _ MIC method are constructed as follows:

*Definition 4 (Mutual Information):* The mutual information of two discrete random variables $X$ and $Y$ is defined as:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \quad (6)$$

where $p(x, y)$ is the joint probability of $X$ and $Y$, $p(x)$ and $p(y)$ are the edge probabilities of $X$ and $Y$, respectively. A larger mutual information value indicates a stronger correlation between two variables. when two variables are independent from each other, $p(x, y) = p(x)p(y)$, the mutual information is 0, that is, there is no same information between two variables.

*Definition 5 (Maximum Information Coefficient of Interval-Valued Data under the URF):* The maximum information coefficient of two random interval-valued feature vectors under the URF is constructed as:

$$IVD\_MIC(U_i^{mr}, U_j^{mr}) = \max_{a*b<B(n)} \frac{\max_{a*b} I(U_i^{mr}; U_j^{mr})}{\log_2 \min(a, b)}, \quad (7)$$

Among it,

$$\max_{a*b} I(U_i^{mr}; U_j^{mr}) = \sum_{u_i^{mr} \in U_i^{mr}} \sum_{u_{\mathrm{j}}^{\mathrm{mr}} \in U_j^{mr}} p(u_i^{mr}, u_{\mathrm{j}}^{\mathrm{mr}}) \log \frac{p(u_i^{mr}, u_{\mathrm{j}}^{\mathrm{mr}})}{p(u_i^{mr})p(u_{\mathrm{j}}^{\mathrm{mr}})}. \quad (8)$$

where $i$, $j$ represent the $i$th, $j$th feature vectors, $i = 1, 2, \ldots, p$. $j$=1, 2, …, $p$. $n$ is the number of samples. $a$, $b$ are the number of partition grids in the $x$, $y$ direction. Extensive experiments show that the method has a good performance when we choose $B(n) = n^{0.6}$ [16]. Eq. (7) normalizes with $\log_2 \min(a, b)$ so that the range is [0, 1].

A simple proof procedure is given as follows:

The strength of correlation between two interval-valued feature vectors under the URF is measured by calculating the maximum information coefficient. Due to the high complexity of mutual information calculation, the two feature vectors are dispersed in a two-dimensional space and represented by a scatter diagram. The current two-dimensional space is divided into certain intervals in the $x$, and $y$ direction, and then the situation of the scatter diagram in each region is viewed to solve the calculation problem in mutual information. For example, under the URF, for a finite set $D \in R^2$ and partition grids $G$, assuming that the partition grids in the direction of $x$ are $a_i$($i$=1,2, …, $x$) respectively, assuming that the partition grids in the direction of $y$ are $b_j$( $j$=1,2, …, $y$), put $D$ in a two-dimensional space, according to the calculation equation of the mutual information in the probability theory, $p(a_i)$ and $p(b_j)$ are probabilities that the dots fall in column $i$ and row $j$ in 2-dimensional space respectively, $p(a_i, b_j)$ denotes probability that is overlapped dots of the row $j$ and column $i$ in 2-dimensional space.

mutual information:

$$\begin{aligned} I(D|G) &= \sum_{i=1}^{x} \sum_{j=1}^{y} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)} \\ &= \sum_{i=1}^{x} \sum_{j=1}^{y} p(a_i, b_j) \log \frac{p(a_i|b_j)}{p(a_i)} \\ &= \sum_{i=1}^{x} \sum_{j=1}^{y} p(a_i, b_j) \log \frac{1}{p(a_i)} \\ &\quad - \sum_{i=1}^{x} \sum_{j=1}^{y} p(a_i, b_j) \log \frac{1}{p(a_i|b_j)} \\ &\leq \sum_{i=1}^{x} \sum_{j=1}^{y} p(a_i, b_j) \log \frac{1}{p(a_i)} \\ &= \sum_{i=1}^{x} p(a_i) \log \frac{1}{p(a_i)}, \end{aligned} \quad (9)$$

Therefore, when the logarithmic function is bottomed by 2, and $a$ represents the number of partition grids in the $x$ direction, there is:

$$\begin{aligned} & I(D|G) - \sum_{i=1}^{x} p(a_i) \log a \\ &\leq \sum_{i=1}^{x} p(a_i) \log \frac{1}{p(a_i)} - \sum_{i=1}^{x} p(a_i) \log a \\ &= \sum_{i=1}^{x} p(a_i) \log \frac{1}{a \cdot p(a_i)} \\ &= \sum_{i=1}^{x} p(a_i) \frac{\ln \frac{1}{a \cdot p(a_i)}}{\ln 2} \\ &= \sum_{i=1}^{x} p(a_i) \ln \frac{1}{a \cdot p(a_i)} \cdot \frac{\ln e}{\ln 2} \\ &= \sum_{i=1}^{x} p(a_i) \ln \frac{1}{a \cdot p(a_i)} \cdot \log e \\ &\leq \sum_{i=1}^{x} p(a_i) \left( \frac{1}{a \cdot p(a_i)} - 1 \right) \cdot \log e \end{aligned}$$

$$= \sum_{i=1}^{x} (\frac{1}{a} - p(a_i)) \cdot \log e$$
$$= (\sum_{i=1}^{x} \frac{1}{a} - \sum_{i=1}^{x} p(a_i)) \cdot \log e$$
$$= 0. \quad (10)$$

From the above Eq. (9) and Eq. (10) to obtain:

$$I(D|G) \leq \log a,$$

and

$$I(D|G) \leq \log b. \quad (11)$$

Therefore:

$$I(D|G) \leq \log \min\{a, b\}. \quad (12)$$

The final proof:

$$0 \leq \max_{a*b<B(n)} \frac{\max_{a*b} I(U_i^{mr}; U_j^{mr})}{\log_2 \min(a, b)} \leq 1. \quad (13)$$

*Nature 1: (Symmetry):* IVD_MIC value of two interval-valued feature variables $U_i^{mr}$ and $U_j^{mr}$ satisfies symmetryis, and is constructed as:

$$IVD\_MIC(U_i^{mr}, U_j^{mr}) = IVD\_MIC(U_j^{mr}, U_i^{mr}) \quad (14)$$

*Definition 6 (Maximum Information Coefficient Matrix of IVD):* Contained information between the interval-valued data features is presented in the form of a matrix, recorded as an IVD_MIC matrix. The maximum information coefficient matrix is constructed as:

$$IVD\_MIC = \begin{bmatrix} U_{11}^{mr} & U_{12}^{mr} & \cdots & U_{1k}^{mr} \\ U_{21}^{mr} & U_{22}^{mr} & \cdots & U_{2k}^{mr} \\ \vdots & \vdots & \ddots & \vdots \\ U_{l1}^{mr} & U_{l2}^{mr} & \cdots & U_{lk}^{mr} \end{bmatrix} \quad (15)$$

where $U_{lk}^{mr}$ is matrix element, $U_{lk}^{mr}$ represents the degree of correlation between the $l$th and $k$th features, that is $IVD\_MIC(U_l^{mr}, U_k^{mr})$, the larger element value indicates the stronger correlation between two features. From *Nature 1*, the IVD _ MIC matrix is symmetric.

### C. IVD_MIC

The main ideas of IVD_MIC are as follows: First, the method uses an adjustment factor to balance the relationship between the midpoint and radius of the IVD unit, constructing IVD under the URF. Within this URF, the IVD_MIC matrix of different datasets is calculated, a larger matrix element value indicates a stronger correlation between two features. Since the IVD_MIC matrix is symmetric, the analysis is performed only in the upper triangle (excluding the main diagonal line) of the matrix. Then from three perspectives of the matrix, those are, row(r), column(c), and both row and column (rc), we progressively remove the features with strong correlation, obtaining a series of corresponding candidate feature subsets. When a feature is deleted, the candidate feature subset obtained from perspective of both row and column is the same as the subset obtained from one of the other two perspectives, so the two features involved are directly deleted from perspective of both row and column. Finally, the candidate feature subsets are learned and verified on different classifiers to select the optimal feature subset. The main steps of the IVD_MIC method are summarized as Algorithm 1.

**Algorithm 1** IVD_MIC Method

**Input:** input an interval-valued dataset $U$, label $Y$, and adjustment factor $\alpha$.

**Output:** output the classification accuracy *acc,* candidate feature subset $A$, and optimal feature subset $E$.

1: Initialize: $A = \emptyset, E = \emptyset$.
2: Convert IVD to midpoint and radius with Eq.(2) and (3), then construct URF for IVD according to Eq.(4).
3: Calculate IVD_MIC values between features with Eq.(7) and form the IVD _ MIC matrix according to Eq. (15).
4: From three perspectives (row, column, and both row and column) of the IVD_MIC matrix, gradually delete the features with strong correlation and put the remaining features into the empty candidate feature subset $A$. Obtain a series of corresponding candidate feature subsets from different perspectives.
5: Candidate feature subsets are learned and classified in the classifiers, obtain corresponding *acc*.
6: Until the stop condition is reached(accuracy drops), the subset with high accuracy and a few number of features is output as the optimal feature subset $E$.

We give an example from the meteorological data labeled by Harbin and Taiyuan to clearly explain the basic process of the proposed IVD_MIC, as shown in Table 1. $U = \{u_1, u_2, \cdots, u_{20}\}$ represents 20 samples, F={temperature, atmospheric pressure, humidity, horizontal visibility, dew-point temperature} represents the feature set.

First of all, we convert IVD to midpoint and radius, and select $\alpha = 0.5$ as the optimal adjustment factor, constructing URF for IVD. Then the degree of correlation between features is measured according to the element value of the IVD_MIC matrix that is shown in Table 2. The element values are ranked as 1.0000, 0.9341, 0.8623, 0.5283, 0.5073, 0.4799, and 0.3674 from large to small in the upper triangle matrix. From three perspectives, features with strong correlation are progressively removed to obtain a series of candidate feature subsets. The removal process is as follows:

For the row perspective of IVD_MIC matrix, T, AP, H, and HV are deleted in turn.

For the column perspective, DPT, AP, H, and HV are deleted in turn.

For the both row and column perspective, first, T and DPT are directly deleted, and then we delete AP and H in turn.

The candidate feature subsets are learned and validated with the classifier to obtain the classification accuracies corresponding to different candidate feature subsets, as shown in

**TABLE 1.** Interval-valued dataset *U* and label *Y*.

| $U$ | Temperature(T) | Atmospheric pressure(AP) | Humidity(H) | Horizontal visibility(HV) | Dew-point temperature(DPT) | $Y$ |
|---|---|---|---|---|---|---|
| $u_1$ | [-23.5,-12.6] | [1029.7,1031.9] | [54,83] | [10.0,12.0] | [-25.8,-20.0] | Harbin |
| $u_2$ | [-24.1,-14.7] | [1029.4,1030.3] | [50,72] | [12.0,20.0] | [-27.7,-22.3] | Harbin |
| $u_3$ | [-26.2,-17.8] | [1030.1,1033.3] | [55,71] | [15.0,25.0] | [-30.1,-24.3] | Harbin |
| $u_4$ | [-29.3,-21.6] | [1020.2,1033.9] | [57,79] | [10.0,20.0] | [-32.3,-26.1] | Harbin |
| $u_5$ | [-28.3,-16.6] | [1018.3,1022.5] | [72,84] | [6.0,12.0] | [-31.1,-20.0] | Harbin |
| $u_6$ | [-22.9,-14.4] | [1023.5,1029.0] | [75,87] | [5.0,12.0] | [-25.3,-17.5] | Harbin |
| $u_7$ | [-19.5,-8.6] | [1021.7,1028.1] | [81,93] | [6.0,8.0] | [-21.2,-9.8] | Harbin |
| $u_8$ | [-18.7,-9.2] | [1024.1,1032.6] | [80,90] | [3.0,8.0] | [-20.2,-10.8] | Harbin |
| $u_9$ | [-13.4,-8.7] | [1020.9,1031.4] | [87,93] | [2.0,7.0] | [-14.6,-10.0] | Harbin |
| $u_{10}$ | [-8.8,-2.6] | [1021.1,1027.7] | [62,90] | [7.0,12.0] | [-11.4,-8.5] | Harbin |
| $u_{11}$ | [3.7,17.4] | [1009.8,1018.1] | [30,68] | [10.0,12.0] | [-4.7,3.5] | Taiyuan |
| $u_{12}$ | [0.2,14.9] | [1013.3,1020.6] | [22,52] | [8.0,12.0] | [-9.4,-6.1] | Taiyuan |
| $u_{13}$ | [2.0,13.3] | [1013.3,1022.1] | [18,48] | [10.0,15.0] | [-11.8,-8.1] | Taiyuan |
| $u_{14}$ | [2.2,20.6] | [1006.2,1017.4] | [10,67] | [6.0,15.0] | [-15.5,-0.2] | Taiyuan |
| $u_{15}$ | [4.1,15.6] | [1008.5,1018.1] | [24,56] | [10.0,15.0] | [-7.8,1.9] | Taiyuan |
| $u_{16}$ | [1.2,19.1] | [1001.7,1013.8] | [26,69] | [10.0,12.0] | [-4.8,2.0] | Taiyuan |
| $u_{17}$ | [-1.9,15.5] | [1016.1,1024.2] | [14,68] | [10.0,12.0] | [-12.3,-6.1] | Taiyuan |
| $u_{18}$ | [1.9,22.7] | [999.3,1016.2] | [13,53] | [12.0,18.0] | [-8.1,-4.2] | Taiyuan |
| $u_{19}$ | [7.5,17.3] | [1002.9,1005.3] | [14,43] | [10.0,15.0] | [-11.0,-4.3] | Taiyuan |
| $u_{20}$ | [5.3,19.5] | [1004.5,1017.7] | [15,91] | [5.0,15.0] | [-10.3,4.3] | Taiyuan |

**TABLE 2.** IVD _ MIC value between features.

| Feature / IVD_MIC value / Feature | T | AP | H | HV | DPT |
|---|---|---|---|---|---|
| T | 1.0000 | 0.9341 | 0.8623 | 0.4799 | 1.0000 |
| AP | 0.9341 | 1.0000 | 0.5238 | 0.4799 | 0.9341 |
| H | 0.8623 | 0.5238 | 1.0000 | 0.3674 | 0.8623 |
| HV | 0.4799 | 0.4799 | 0.3674 | 0.9928 | 0.5073 |
| DPT | 1.0000 | 0.9341 | 0.8623 | 0.5073 | 1.0000 |

Table 3. In Table 3, from perspective of column, the feature subset composed of T retains the fewest features, and has the highest accuracy, reaching 100%. Therefore, considering three perspectives, the feature subset composed of T is finally selected as the optimal feature subset.

For the $n \times p$ interval-valued dataset, $n$ is the number of samples and$p$ is the feature dimension. Since $\alpha$ adjusts the midpoint and radius of the interval-valued data unit, the feature dimension does not increase, so the time complexity of IVD _ MIC is $O(p^2)$.

## III. EXPERIMENTAL RESULTS AND ANALYSIS

IVD_MIC emerges as a potent solution for the feature selection of IVD. In this section, we validate the effectiveness of IVD_MIC by conducting a comparative analysis of experimental results. The performance of IVD_MIC is benchmarked against other feature selection methods using both synthetic and real-world datasets. This comparative evaluation aims to demonstrate the superiority and efficacy of IVD_MIC in handling feature selection challenges posed by IVD.

### A. DATA DESCRIPTION

The experiment utilizes datasets and the best $a$ mentioned in the [17], and detailed information about the datasets is provided in Table 4. It includes 8 synthetic datasets and 4 real-world datasets, categorized into two types: low dimensional and high dimensional for synthetic datasets. The construction

**TABLE 3.** Accuracies varied with removed features.

| Perspectives | Feature set（$A$） | Accuracy（%） |
|---|---|---|
| Universal set(U) | {T, AP, H, HV, DPT} | 80 |
| Row（r） | {AP, H, HV, DPT} | 85 |
| | {H, HV, DPT} | 95 |
| | {HV, DPT} | 95 |
| | {DPT} | **95** |
| Column（c） | {T, AP, H, HV} | 85 |
| | {T, H, HV} | 85 |
| | {T, HV} | **100** |
| | {T} | **100** |
| Both row and column（rc） | {AP, H, HV} | 85 |
| | {H, HV} | **95** |
| | {HV} | 60 |

**TABLE 4.** Datasets used in experiments.

| Order number | Data form | Datasets | Num. of samples | Num. of feature | Num. of classes |
|---|---|---|---|---|---|
| 1 | Low dimensional synthetic datasets | Ds1 | 3000 | 4 | 2 |
| 2 | | Ds2 | 4500 | 4 | 3 |
| 3 | | Ds3 | 1493 | 4 | 3 |
| 4 | | Ds4 | 2000 | 4 | 4 |
| 5 | High dimensional synthetic datasets | Set1 | 1200 | 100 | 2 |
| 6 | | Set2 | 900 | 100 | 3 |
| 7 | | Set3 | 800 | 100 | 4 |
| 8 | | Set4 | 1000 | 100 | 4 |
| 9 | Real-world datasets | HS_Ds | 7302 | 5 | 2 |
| 10 | | TB_Ds | 7302 | 5 | 2 |
| 11 | | HSTB_Ds | 14604 | 5 | 4 |
| 12 | | Water | 316 | 48 | 2 |

**TABLE 5.** Four comparison algorithms.

| Order number | Method | Type |
|---|---|---|
| 1 | SU | Mutual information |
| 2 | Pearson | Covariance and standard deviation |
| 3 | Spearman | Location |
| 4 | Kendall | Pairs |

method aligns with the method outlined in [17]. Additionally, irrelevant features have been removed in Ds1 and Ds2. Among the 4 real-world datasets, the first three are meteorological data sourced from the "Reliable Prognosis" site [18], the Water dataset comprises 30-minute flow records spanning 1 year in the Barcelona water distribution network (from June 1, 2003 to May 31, 2004) [19].

To evaluate the effectiveness of IVD_MIC, four benchmark methods are employed for comparison. The specific settings are outlined in Table 5. Among them, SU (symmetrical uncertainty) is a traditional method based on mutual information [17]. The Pearson correlation coefficient measures the linear correlation between two variables based on covariance and standard deviation. The Spearman correlation coefficient is a rank correlation coefficient obtained based on the variable locations in the data. The Kendall correlation coefficient, another rank correlation coefficient, measures the monotonic relationship between two ordered variables using "pairs" and "pairs" with consistent and disagreement pairs, indicating the consistency and inconsistency of the values of the two variables, respectively.

### B. PARAMETER SETTING

In the experiment, the step size for the low-dimensional synthetic and the 5-dimensional real-world datasets is set to 1. The high-dimensional synthetic and Water datasets have more feature dimensions, so the stepwise shrinkage method is used with preset step sizes of 10, 5, and 1. When the step size is 10, the data set is preliminarily screened until the accuracy drops. Taking the feature subset with high accuracy and fewer features after preliminary screening as the center, neighboring subsets of features are selected, and then the screening is continued with 5 as the step size. The above process is repeated until the step size is 1 and the accuracy drops, and the optimal feature subset can be selected.

In order to avoid randomness, the experiment adopts the method of 10-fold cross validation. The samples on the data set are randomly divided into 10 parts of the same size. Each part serves as the test set in turn, while the remaining 9 parts form the training set. Finally, the experiments are repeated 10 times on each dataset, and the average value is taken as the experimental result. Furthermore, LIBSVM, CART Tree, and KNN classifiers are used to verify the performance of 5 methods, SVM is implemented using the LIBSVM toolkit, utilizing the RBF kernel function with the parameter $\gamma = 0.2$ and a penalty factor of $C = 1$. CART Tree adopts the Gini

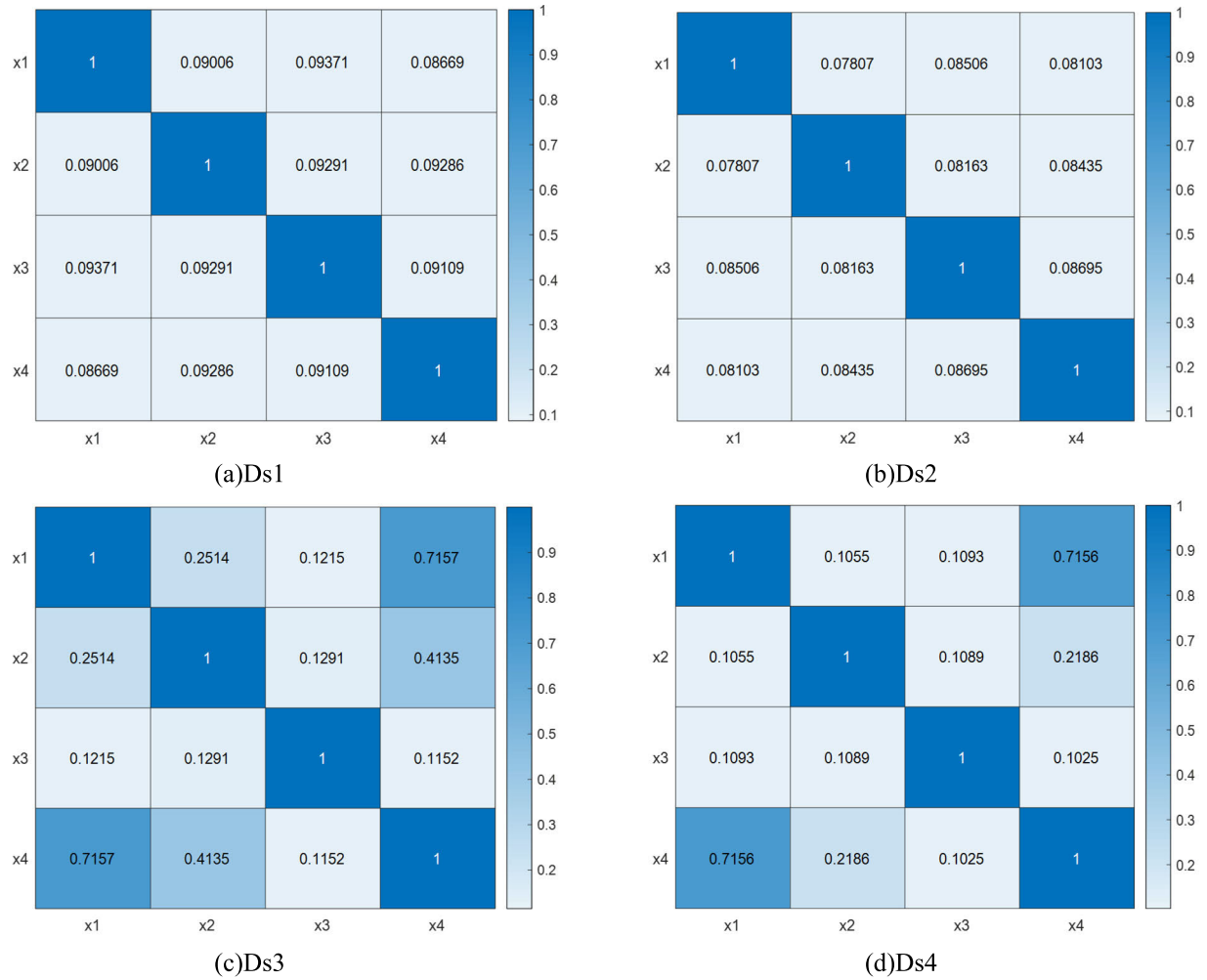



**FIGURE 1.** IVD _ MIC value between features.

coefficient with no parameters. For KNN, euclidean distance is selected as the nearest neighbor with the parameter $K = 4$.

### C. ANALYSIS OF THE EXPERIMENTAL RESULTS

#### 1) CORRELATION ANALYSIS BETWEEN FEATURES FOR IVD_MIC

The strength of the relationship between features is measured based on the element values of the IVD_MIC matrix, which is presented as a thermal diagram of the correlation coefficient matrix in Fig.1. Here, x1, x2, x3, and x4 represent four different features. The darker color of the small square where the element value is located indicates the larger element value and the stronger relationship between two features.

In Fig.1, for Ds1 and Ds2, the element values are generally small. For Ds3 and Ds4, the IVD_MIC value of x1 and x4 is the largest, and the IVD_MIC value of x2 and x4 is second. The upper triangle element values of the matrix are ranked from large to small, and features are removed from perspectives of row, column, and both row and column of the IVD_MIC matrix to obtain a series of candidate feature subsets. The results are shown in Table 6, where "$\surd$" indicates a retained feature, and "$\times$" indicates a removed feature in the feature subset.

The experiment results in Table 6 are obtained by removing features step by step according to the element values in Fig.1. Let's take Ds3 as an example. The element values in the upper triangle matrix, ranked from large to small, are 0.7157, 0.4135, 0.2514, 0.1291, 0.1215, and 0.1152. The removal process is as follows:

From perspective of row:
Remove x1 (0.7157).
Remove x2 (0.4135).
Remove x3 (0.1152).
From perspective of column:
Remove x4 (0.7157).
Remove x2 (0.2514).
Remove x3 (0.1291).
From perspective of both row and column:
Remove x1 and x4 directly(0.7157).
Remove x2(0.4135).

**TABLE 6.** Candidate feature subsets varied with removed features.

| Datasets | Three perspectives | Number of retained features for the three perspectives | Retained features in the feature subset | | | |
|---|---|---|---|---|---|---|
| | | | x1 | x2 | x3 | x4 |
| Ds1 | r | r3 | × | √ | √ | √ |
| | | r2 | × | × | √ | √ |
| | | r1 | × | × | × | √ |
| | c | c3 | √ | √ | × | √ |
| | | c2 | √ | √ | × | × |
| | | c1 | √ | × | × | × |
| | rc | rc2 | × | √ | × | √ |
| | | rc1 | × | × | × | √ |
| Ds2 | r | r3 | √ | √ | × | √ |
| | | r2 | × | √ | × | √ |
| | | r1 | × | × | × | √ |
| | c | c3 | √ | √ | √ | × |
| | | c2 | √ | √ | × | × |
| | | c1 | √ | × | × | × |
| | rc | rc2 | √ | √ | × | × |
| | | rc1 | × | √ | × | × |
| Ds3 | r | r3 | × | √ | √ | √ |
| | | r2 | × | × | √ | √ |
| | | r1 | × | × | × | √ |
| | c | c3 | √ | √ | √ | × |
| | | c2 | √ | × | √ | × |
| | | c1 | √ | × | × | × |
| | rc | rc2 | × | √ | √ | × |
| | | rc1 | × | × | √ | × |
| Ds4 | r | r3 | × | √ | √ | √ |
| | | r2 | × | × | √ | √ |
| | | r1 | × | × | × | √ |
| | c | c3 | √ | √ | √ | × |
| | | c2 | √ | √ | × | × |
| | | c1 | √ | × | × | × |
| | rc | rc2 | × | √ | √ | × |
| | | rc1 | × | × | √ | × |

**TABLE 7.** Candidate feature subsets varied with removed features.

| Datasets | Three perspectives | Number of retained features for the three perspectives | Retained features in the feature subset | | | | |
|---|---|---|---|---|---|---|---|
| | | | T | AP | H | HV | DPT |
| HS_Ds | r | r4 | × | √ | √ | √ | √ |
| | | r3 | × | × | √ | √ | √ |
| | | r2 | × | × | √ | × | √ |
| | | r1 | × | × | × | × | √ |
| | c | c4 | √ | √ | √ | √ | × |
| | | c3 | √ | × | √ | √ | × |
| | | c2 | √ | × | √ | × | × |
| | | c1 | √ | × | × | × | × |
| | rc | rc3 | × | √ | √ | √ | × |
| | | rc2 | × | × | √ | √ | × |
| | | rc1 | × | × | √ | × | × |
| TB_Ds | r | r4 | × | √ | √ | √ | √ |
| | | r3 | × | × | √ | √ | √ |
| | | r2 | × | × | × | √ | √ |
| | | r1 | × | × | × | × | √ |
| | c | c4 | √ | × | √ | √ | √ |
| | | c3 | √ | × | √ | √ | × |
| | | c2 | √ | × | √ | × | × |
| | | c1 | √ | × | × | × | × |
| | rc | rc3 | × | × | √ | √ | √ |
| | | rc2 | × | × | √ | √ | × |
| | | rc1 | × | × | × | √ | × |
| HSTB_Ds | r | r4 | × | √ | √ | √ | √ |
| | | r3 | × | × | √ | √ | √ |
| | | r2 | × | × | × | √ | √ |
| | | r1 | × | × | × | × | √ |
| | c | c4 | √ | √ | √ | √ | × |
| | | c3 | √ | × | √ | √ | × |
| | | c2 | √ | × | × | √ | × |
| | | c1 | √ | × | × | × | × |
| | rc | rc3 | × | √ | √ | √ | × |
| | | rc2 | × | × | √ | √ | × |
| | | rc1 | × | × | × | √ | × |

Similar processes are followed for the other datasets, and features are removed in turn from three perspectives to obtain a series of feature subsets. For example, for Ds1 and Ds4, x1, x2, and x3 are deleted in turn from perspective of row. For Ds2, x3, x1, and x2 are deleted. For the column perspective, x3, x4, and x2 are deleted in turn for Ds1, and for Ds2 and Ds4, x4, x3, and x2 are removed in turn. In the both row and column perspective, for Ds1, x1 and x3 are deleted directly, then x2 is deleted. Similarly, for Ds2, x3 and x4 are deleted, then x1 is deleted. For Ds4, x1 and x4 are deleted, then x2 is deleted.

Due to the large number of features in high-dimensional datasets, thermal diagrams of correlation coefficient matric and feature subsets are not provided. However, for Set1-Set3, the elemental values of the correlation coefficient are distributed between [0, 0.2]. For Set4, the distribution is as follows:

The first-row elemental values of the matrix are distributed between [0.2, 0.3].

The element values of the second row are distributed between [0.6, 0.7].

The element values for the remaining row are 0.7219.

Similarly, matrix element values on each dataset are arranged in descending order, and features are gradually removed from three perspectives using the stepwise shrinkage method to obtain a series of feature subsets.

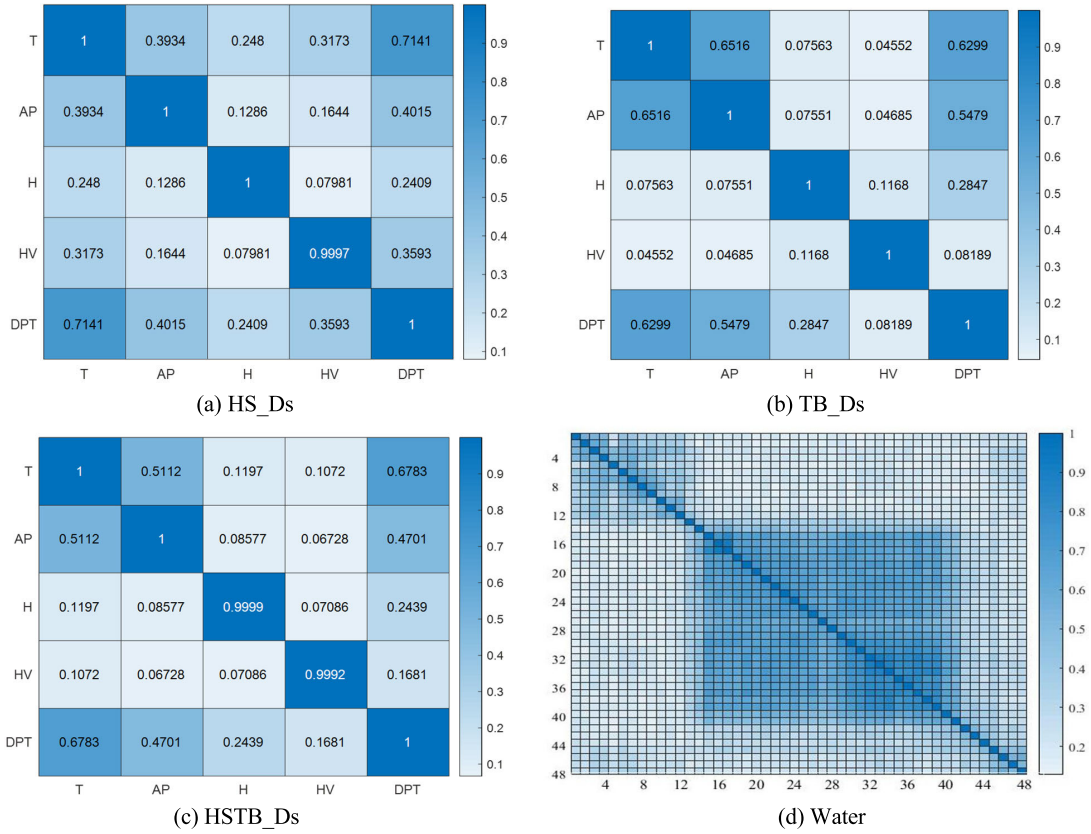
Combining Fig.2 and Table 7, we can observe the features in the obtained subsets. For the three five-dimensional datasets (T, AP, H, HV, DPT), where T represents temperature, AP is atmospheric pressure, H is humidity, HV is horizontal visibility, and DPT is dew-point temperature, the upper triangle (excluding the main diagonal line) of the thermal diagram shows that element values are distributed between 0 and 1. Arranging them from large to small, we delete features from three perspectives and obtain feature subsets. Taking HS_Ds as an example, the upper triangle element values order is 0.7141, 0.4015, 0.3934, 0.3593, 0.3173, 0.2480, 0.2409, 0.1644, 0.1286 and 0.0798.

For the row perspective of IVD_MIC matrix, we delete T, AP, HV, and H in turn.

For the column perspective, we delete DPT, AP, HV, and H in turn.

For the both row and column perspective, first, T and DPT are directly deleted, then we delete AP and HV in turn.

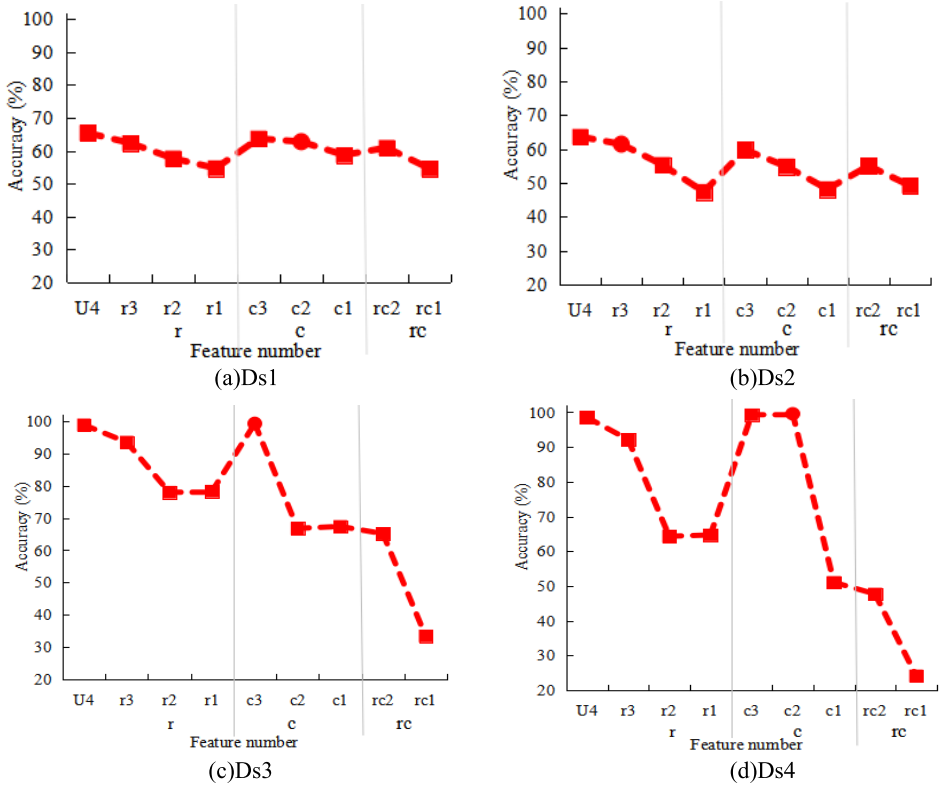
Similar processes are applied to the other two five-dimensional datasets, where T, AP, H and HV are deleted in turn from perspective of row. For the column perspective, AP, DPT, HV and H are deleted in turn on TB_Ds. On HSTB_Ds, DPT, AP, H and HV are deleted successively. For the both row and column perspective, T and AP are directly deleted,




(a) HS_Ds (b) TB_Ds

(c) HSTB_Ds (d) Water

**FIGURE 2.** IVD _ MIC value between features.




(a)Ds1 (b)Ds2

(c)Ds3 (d)Ds4

**FIGURE 3.** Accuracy trend of feature subsets on low-dimensional synthetic datasets.

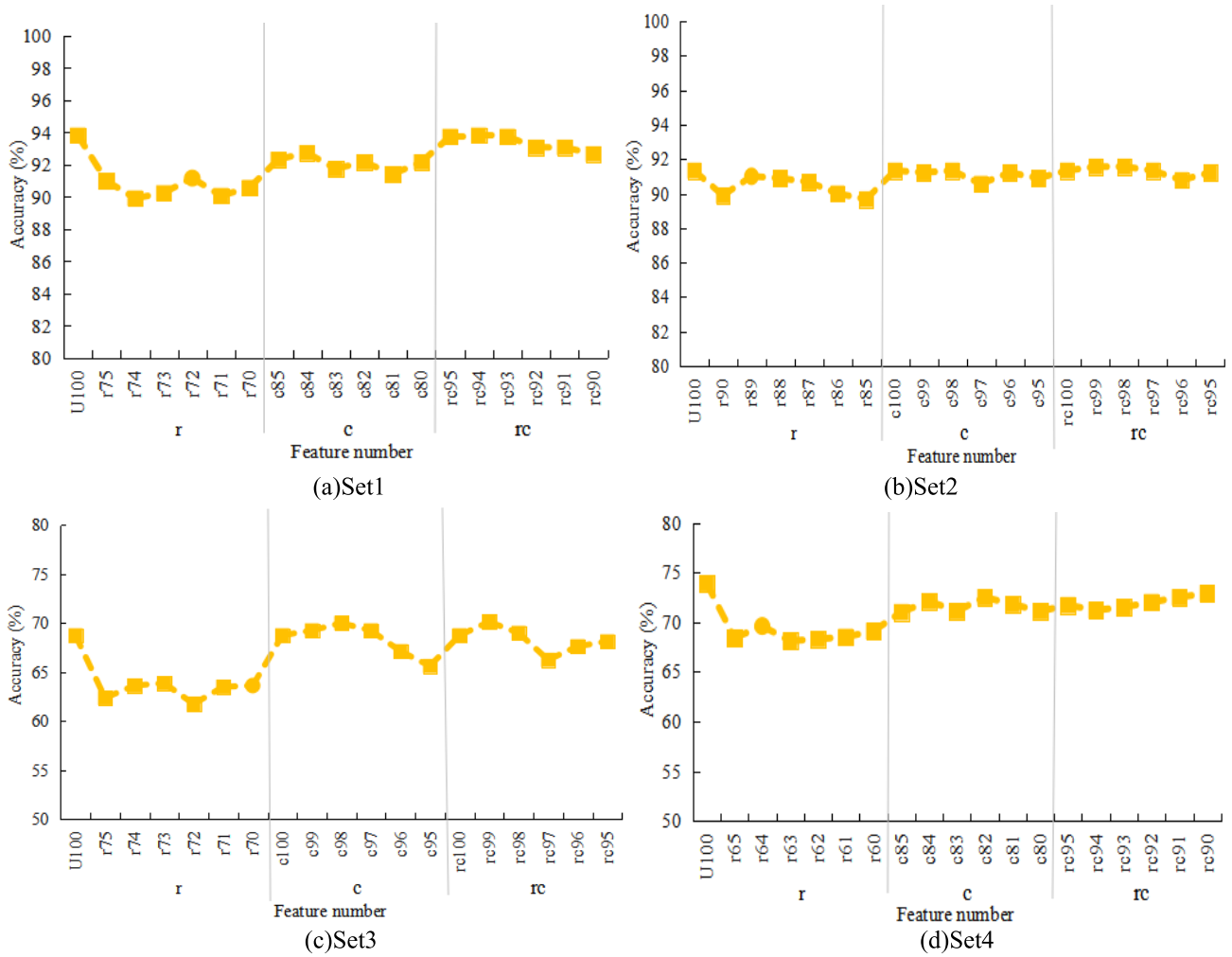



**FIGURE 4.** Accuracy trend of feature subsets on high-dimensional synthetic datasets.

followed by the deletion of DPT and H in turn on TB_Ds. T and DPT are directly deleted, followed by the deletion of AP and H in turn on HSTB_Ds.

Although IVD_MIC values and feature subsets of the Water dataset are not detailed, we can see from the thermal diagram that features with strong correlation are concentrated in the middle and upper-left corner of the figure. Features are removed using the stepwise shrinkage method according to the idea applied to the five-dimensional datasets to obtain a series of feature subsets.

### 2) SELECTION OF OPTIMAL FEATURE SUBSET FOR IVD _ MIC

IVD_MIC obtains a series of candidate feature subsets from three perspectives of the IVD_MIC matrix to select the optimal feature subset. Fig.3-Fig.5 show the accuracy trend of IVD_MIC on synthetic datasets and real-world datasets. In the figures, the y-axis represents the classification accuracy, and the x-axis represents the number of features corresponding to the feature subsets under different perspectives of the complete set (U), row (r), column (rc), and both row and column (rc). The small squares represent the number of features of different candidate feature subsets and the corresponding accuracies, and the circle dot represents the final selected optimal feature subset. Each subgraph is divided into three regions from left to right according to three different perspectives by two vertical lines.

From Fig.3, it can be clearly seen that the accuracies corresponding to each perspective on both Ds1 and Ds2 datasets are decreasing. When the number of features is 3 and 2 on the column perspective for Ds1, the accuracies of the corresponding feature subsets vary by less than 1the accuracy of the subset containing 2 features in the column perspective is higher than the corresponding accuracy of the subset containing 3 features in the row perspective. Therefore, the feature subset corresponding to c2 is selected as the optimal subset for Ds1, and r1 is the optimal subset for Ds2. Similarly, for Ds3 and Ds4, after comparison from three perspectives, the feature subsets corresponding to c3 and c2 are finally selected as the optimal feature subsets, respectively.

In Fig.4, the classification accuracies of the subsets fluctuate continuously with decreasing in the number of selected features. In order to select the optimal feature subset quickly,

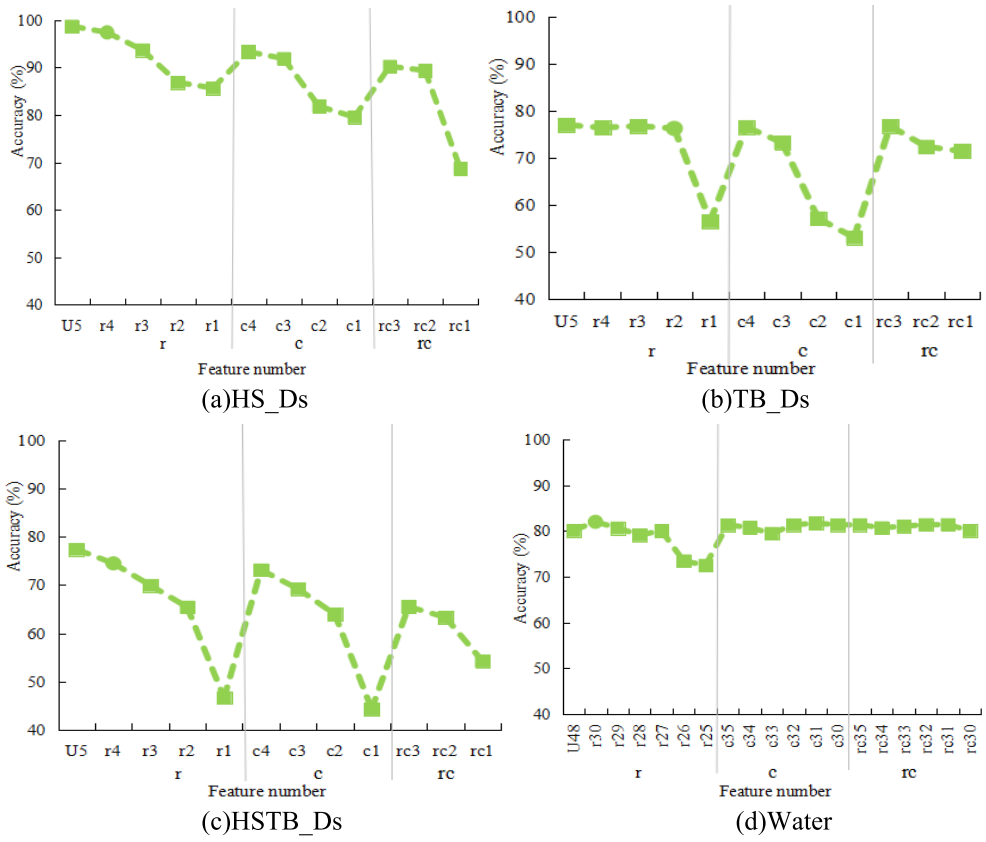



**FIGURE 5.** Accuracy trend of feature subsets on real-world datasets.

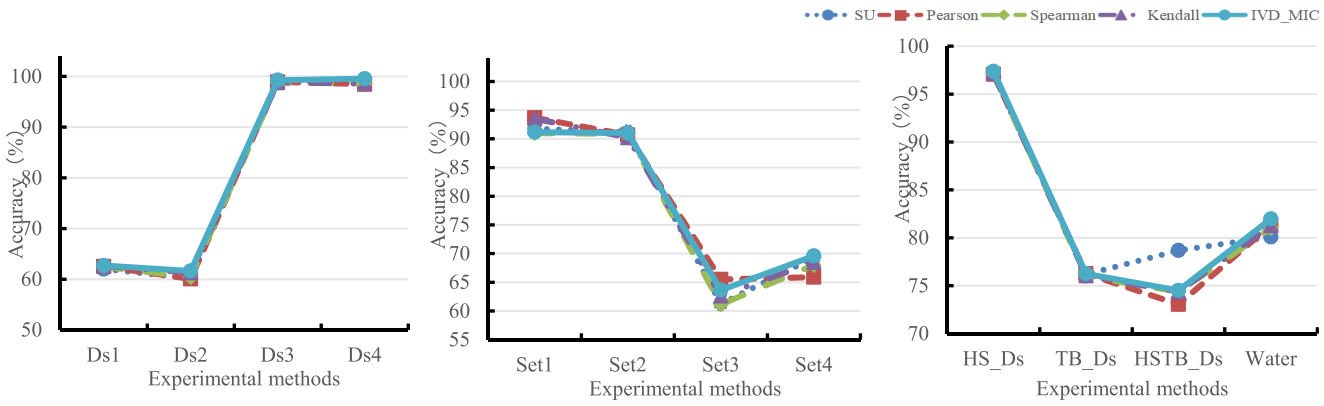
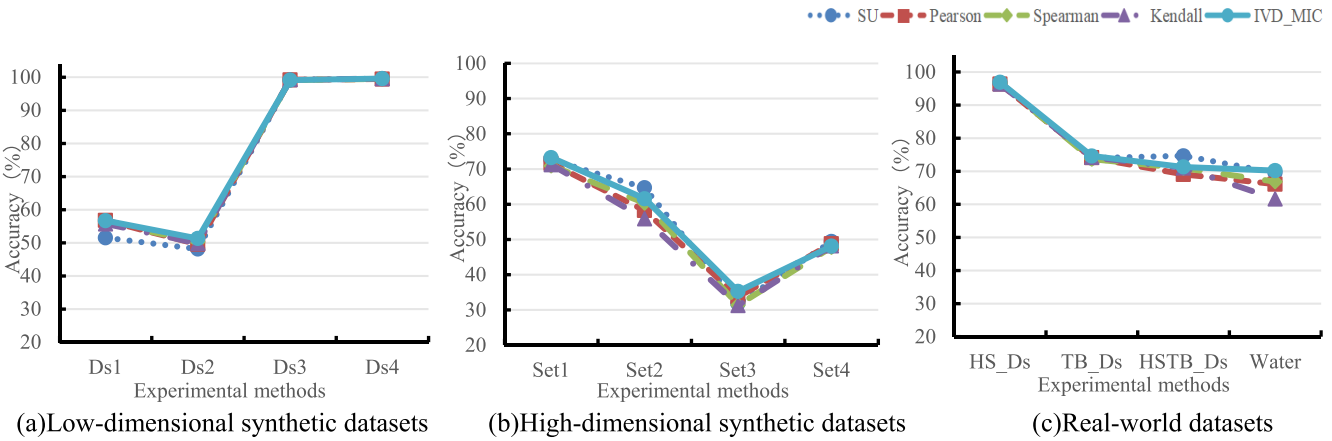



**FIGURE 6.** Comparison of classification accuracy among five methods on LIBSVM.

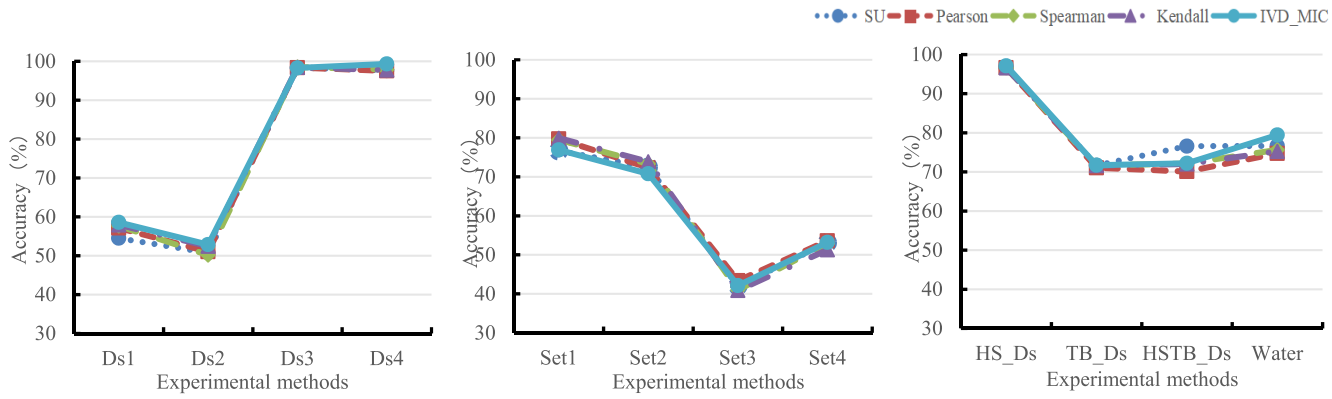
we can first select a better feature subset from three perspectives, and then compare to obtain the optimal feature subset. For Set1, in the three perspectives, the number of features of the selected better feature subsets is 72, 84, and 93 respectively, whose corresponding accuracies are 91.17%, 92.75%, and 93.75% respectively. The maximum value of accuracies difference among the three is 2.58%, while the difference of the corresponding number of features is 21, and the corresponding average unit feature accuracies difference improves by 0.27%. Therefore, the feature subset corresponding to r72 is selected as the optimal feature subset. In like manner, for Set2, Set3, and Set4, the feature subsets corresponding to r89, r70, and r64 are selected as the optimal subset after a comparison from three perspectives.

Fig.5 shows the changing trend of the accuracies corresponding to different subsets on the real-world datasets.




(a)Low-dimensional synthetic datasets (b)High-dimensional synthetic datasets (c)Real-world datasets

**FIGURE 7.** Comparison of classification accuracy among five methods on CART Tree.




(a)Low-dimensional synthetic datasets (b)High-dimensional synthetic datasets (c)Real-world datasets

**FIGURE 8.** Comparison of classification accuracy among five methods on KNN.

As for both HS_Ds and HSTB_Ds, the accuracies show a downward trend, therefore, the subset with the highest accuracy is selected as the optimal feature subset from three perspectives, that is the subset corresponding to r4. For TB_Ds, in the row perspective, when the number of retained features is 4, 3, and 2, the corresponding accuracies are similar and relatively high. In two perspectives of column, and both row and column, when the number of features is 4 and 3, respectively, the corresponding accuracies are relatively high, and the accuracies are similar to the subset with the retained 2 features at the row perspective. So r2 is selected as the optimal feature subset. For the Water, in the three perspectives, the selected better subsets have 30, 31, and 31 features respectively, whose corresponding accuracies are 81.98%, 81.70%, and 81.36%. It is obvious that the subset with 30 features has the highest accuracy and the least number of features, so r30 is selected as the optimal feature subset.

### 3) CLASSIFICATION PERFORMANCE OF IVD_MIC

A good feature selection method should not only select fewer features but also choose the most representative features that can improve the classification performance. To evaluate the performance of IVD_MIC, five methods are validated on LIBSVM, CART Tree, and KNN. The performance of IVD_MIC is reflected through different evaluation indicators which are classification accuracy, average accuracy, MRR, Win/Tie/Loss, average ranking, and average unit feature accuracy.

Fig.6-Fig.8 show a comparison of classification accuracy among five methods on three classifiers. For LIBSVM and CART Tree classifiers, IVD_MIC achieves the highest classification accuracy in 7 out of the 12 datasets. For KNN, IVD_MIC achieves the highest classification accuracy in 6 datasets. Additionally, compared to other methods, IVD_MIC performs better on the real-world datasets. In conclusion, IVD_MIC demonstrates more outstanding classification performance.

We compare the accuracies of five methods on three classifiers, and the results on 12 datasets are listed in Table 8. The best results are indicated with bold values. Average results, MRR, and Win/Tie/Loss of each method are also listed in the last three row of each classifier. MRR is the average

**TABLE 8.** Comparison of classification accuracy on different classifiers.

| Classifier | Dataset | SU | Pearson | Spearman | Kendall | IVD_MIC |
|---|---|---|---|---|---|---|
| LIBSVM | Ds1 | 62.00 (5) | 62.53 (3) | 62.47 (4) | 62.63 (2) | **62.70 (1)** |
| | Ds2 | 60.20 (4) | 60.07 (5) | 60.40 (3) | 61.30 (2) | **61.64 (1)** |
| | Ds3 | 99.20 (2) | 98.93 (3) | 98.79 (5) | 98.80 (4) | **99.26 (1)** |
| | Ds4 | 98.71 (2) | 98.45 (4) | 98.60 (3) | 98.60 (3) | **99.56 (1)** |
| | Set1 | 91.75 (3) | **93.67 (1)** | 91.00 (5) | 93.50 (2) | 91.17 (4) |
| | Set2 | **91.22 (1)** | 90.67 (4) | 90.88 (3) | 90.22 (5) | 91.00 (2) |
| | Set3 | 61.25 (4) | **65.50 (1)** | 61.13 (5) | 62.63 (3) | 63.63 (2) |
| | Set4 | 69.30 (2) | 65.90 (5) | 68.00 (4) | 68.50 (3) | **69.60 (1)** |
| | HS_Ds | 97.25 (2) | 97.07 (5) | 97.15 (3) | 97.14 (4) | **97.37 (1)** |
| | TB_Ds | 76.18 (3) | **76.27 (1)** | 76.01 (5) | 76.05 (4) | 76.24 (2) |
| | HSTB_Ds | **78.68 (1)** | 73.05 (4) | 74.38 (3) | 74.38 (3) | 74.51 (2) |
| | Water | 80.09 (5) | 81.32 (2) | 81.00 (4) | 81.30 (3) | **81.98 (1)** |
| | Average | 80.49 | 80.29 | 79.98 | 80.42 | **80.72** |
| | MRR | 0.46 | 0.46 | 0.27 | 0.34 | **0.77** |
| | Win/Tie/Loss | 9/0/3 | 9/0/3 | 12/0/0 | 11/0/1 | — |
| CART Tree | Ds1 | 51.53 (5) | **56.77 (1)** | 56.27 (3) | 55.73 (4) | 56.70 (2) |
| | Ds2 | 48.13 (5) | 49.56 (4) | 50.31 (2) | 49.84 (3) | **51.31 (1)** |
| | Ds3 | 99.20 (2) | 99.20 (2) | 99.13 (3) | **99.26 (1)** | 99.12 (4) |
| | Ds4 | 99.56 (2) | 99.45 (5) | 99.50 (4) | 99.55 (3) | **99.59 (1)** |
| | Set1 | 72.58 (2) | 71.58 (3) | 70.92 (5) | 71.25 (4) | **73.25 (1)** |
| | Set2 | **64.67 (1)** | 58.22 (4) | 60.33 (3) | 55.89 (5) | 61.56 (2) |
| | Set3 | 31.75 (3) | 33.75 (2) | 31.38 (4) | 31.25 (5) | **35.25 (1)** |
| | Set4 | **49.40 (1)** | 48.80 (2) | 47.80 (5) | 48.20 (3) | 48.10 (4) |
| | HS_Ds | 96.62 (3) | 96.37 (4) | 96.80 (2) | 96.37 (4) | **96.92 (1)** |
| | TB_Ds | 73.99 (4) | 74.14 (3) | 73.68 (5) | 74.23 (2) | **74.62 (1)** |
| | HSTB_Ds | **74.65 (1)** | 69.04 (5) | 70.94 (4) | 71.03 (3) | 71.31 (2) |
| | Water | 69.55 (2) | 66.10 (4) | 66.84 (3) | 61.65 (5) | **70.20 (1)** |
| | Average | 69.30 | 68.58 | 68.66 | 67.85 | **69.83** |
| | MRR | 0.53 | 0.38 | 0.31 | 0.35 | **0.75** |
| | Win/Tie/Loss | 8/0/4 | 9/0/3 | 11/0/1 | 10/0/2 | — |
| KNN | Ds1 | 54.50 (5) | 57.13 (4) | 57.67 (3) | 57.93 (2) | **58.57 (1)** |
| | Ds2 | 50.78 (4) | 51.03 (3) | 50.18 (5) | 52.44 (2) | **52.84 (1)** |
| | Ds3 | 98.39 (3) | 98.39 (3) | **98.53 (1)** | 98.46 (2) | 98.31 (4) |
| | Ds4 | 97.61 (4) | 97.55 (5) | 97.90 (2) | 97.75 (3) | **99.33 (1)** |
| | Set1 | 76.67 (5) | 79.75 (2) | 79.50 (3) | **79.92 (1)** | 76.92 (4) |
| | Set2 | 72.44 (3) | 72.33 (4) | 73.33 (2) | **73.89 (1)** | 70.78 (5) |
| | Set3 | 42.00 (3) | **43.38 (1)** | 41.00 (4) | 40.88 (5) | 42.13 (2) |
| | Set4 | 53.10 (3) | **53.60 (1)** | 52.90 (4) | 51.30 (5) | 53.20 (2) |
| | HS_Ds | 96.73 (2) | 96.64 (4) | 96.67 (3) | 96.63 (5) | **97.10 (1)** |
| | TB_Ds | 71.65 (3) | 71.06 (5) | 71.69 (2) | 71.63 (4) | **71.71 (1)** |
| | HSTB_Ds | **76.57 (1)** | 70.09 (5) | 72.14 (3) | 72.03 (4) | 72.21 (2) |
| | Water | 76.64 (2) | 74.72 (5) | 75.65 (3) | 75.28 (4) | **79.45 (1)** |
| | Average | 72.26 | 72.14 | 72.26 | 72.35 | **72.71** |
| | MRR | 0.38 | 0.39 | 0.41 | 0.43 | **0.68** |
| | Win/Tie/Loss | 9/0/3 | 7/0/5 | 9/0/3 | 9/0/3 | — |

reciprocal rank [20], used to measure the comprehensive ranking of a method, the larger MRR value indicates the better ranking. For example, IVD_MIC is ranked on LIBSVM as follows: 1, 1, 1, 1, 4, 2, 2, 1, 1, 2, 2, 1. The MRR of IVD_MIC is calculated as $(1 + 1 + 1 + 1 + 1 / 4 + 1 / 2 + 1 / 2 + 1 + 1 + 1 + 1 / 2 + 1 / 2 + 1) / 12 = 0.7708$. Win/Tie/Loss [21] indicates that IVD_MIC outperforms the current method (Win), has the same performance (Tie), or is not outperformed by the current method (Loss). From Table 6, it is clear that IVD_MIC outperformed other methods for the average on the three classifiers, and IVD_MIC has the best MRR value with about 0.77, 0.75, and 0.68, respectively, on the three classifiers. Besides, Win/Tie/Loss results are significant, with the number of Ties being 0 and the number of Wins being more than Losses on three classifiers. This indicates that the accuracy of IVD_MIC performs the best over the current methods in most datasets. In conclusion, the comprehensive performances of IVD_MIC are optimal, indicating that IVD_MIC performs better in feature selection and has good robustness, especially on LIBSVM.

Fig.9 shows the average rankings of each method on different classifiers. As seen in Fig.9, IVD_MIC not only ranks first but also has a small standard deviation on three classifiers. For LIBSVM and CART Tree, IVD_MIC ranks significantly better than other methods, and the standard deviation is similar to Spearman. For KNN, IVD_MIC is ranked optimal with a small standard deviation (1.44). This further reinforces the conclusion that IVD_MIC performs exceptionally well in comparison to other methods across different classifiers.

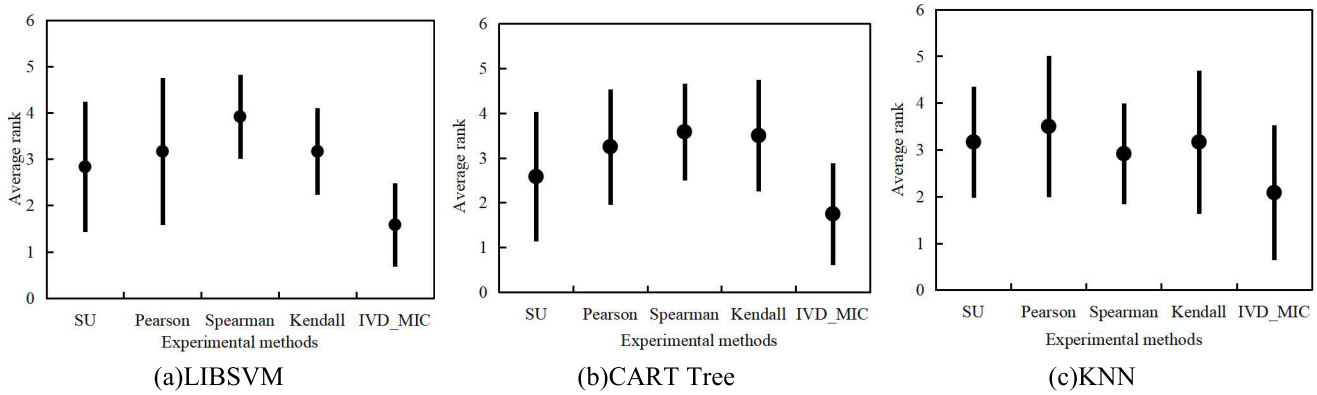



**FIGURE 9.** Average ranking of the experimental methods (average ± standard deviation).

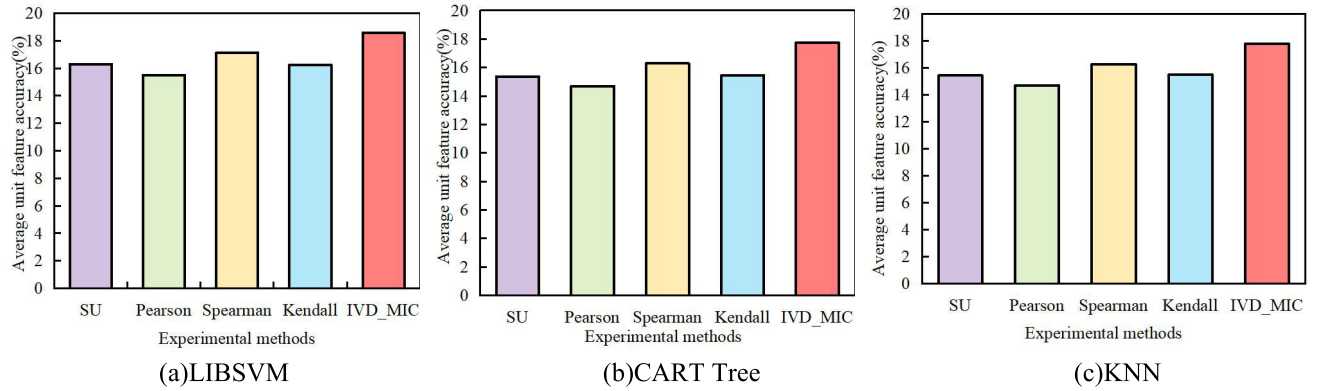



**FIGURE 10.** Average unit feature accuracy of the experimental methods.

The average unit feature accuracies of the five methods are compared to verify the effectiveness of IVD_MIC, and Fig.10 shows the comparison results of average unit feature accuracy for each method on different classifiers. In Fig.10, on the three classifiers, the average unit feature accuracy of IVD_MIC is the highest, Spearman is second, and Pearson is last. On LIBSVM, CART Tree, and KNN, the average unit feature accuracies of IVD_MIC are 18.57%, 17.75%, and 17.79%, respectively. They are 1.43%, 1.46%, and 1.49% higher than Spearman, and 3.07%, 3.04%, and 3.08% higher than Pearson. The experimental results demonstrate that IVD_MIC is the most effective method for feature selection of IVD.

## IV. CONCLUSION

In conclusion, the traditional feature selection methods for IVD face some challenges, such as losing key information (location or size), and increasing the number of features. To address these issues, we propose the IVD_MIC. We construct URF with the best adjustment factor, and then calculate the IVD_MIC matrix. Subsequently, we remove strongly correlated features from three perspectives, and obtain the optimal feature subset. Experimental results demonstrate that IVD_MIC outperforms other methods in both comprehensive performance and average unit feature accuracy. Since we concentrate on resolving redundancy issues among features, there's a possibility of eliminating features more closely associated with the category or retaining those less relevant. In subsequent endeavors, we aim to experiment with a two-step feature selection approach to tackle the redundant features of IVD. This method will preserve features strongly linked to the category in the initial step and then eliminate redundant features from the selected set in the subsequent step. Alternatively, we seek to explore a feature selection method that combines the above two steps into one, seeking maximum correlation and minimum redundancy.

## REFERENCES

[1] L. Billard and J. Le-Rademacher, "Principal component analysis for interval data," *WIREs Comput. Statist.*, vol. 4, no. 6, pp. 535–540, Nov. 2012.

[2] Q. X. Liu, *Principal Component Analysis and Effectiveness Study of Interval-Type Symbol Data*. Meghalaya, India: Univ. Sci. Technol., 2019.

[3] D. Deng, *Evaluation of the Effectiveness of Principal Component Analysis and Cluster Analysis for Interval-Type Symbol Data*. Tianjin, China: Tianjin University, 2010.

[4] G. R. Jahanshahloo, F. H. Lotfi, F. R. Balf, and H. Z. Rezai, "Discriminant analysis of interval data using Monte Carlo method in assessment of overlap," *Appl. Math. Comput.*, vol. 191, no. 2, pp. 521–532, Aug. 2007.

[5] A. B. Ramos-Guajardo and P. Grzegorzewski, "Distance-based linear discriminant analysis for interval-valued data," *Inf. Sci.*, vol. 372, pp. 591–607, Dec. 2016.
[6] Y. N. Huang, *Research on the Fisher Discrimination Analysis Model and Algorithm of Triangular Fuzzy Number and Interval Number Based on Distance*. Ningxia, China: Ningxia University, 2018.
[7] J. P. Guo, R. Zhao, and W. H. Li, "Regression analysis of interval-type symbolic data considering internal scatter," *J. Manage. Sci.*, vol. 21, no. 4, pp. 114–126, 2018.
[8] P. Qiao, *Research on the Nonlinear Regression Method Based on Interval-Type Data and Its Application*. Xiamen, China: Xiamen University, 2019.
[9] Y. Wang, S. L. Yan, and F. He, "A CCRM interval regression method considering the degree of interval coincidence," *Statist. Decis.-Making*, vol. 38, no. 5, pp. 11–16, 2022.
[10] N. Lethikim and T. Vovan, "Fuzzy cluster analysis for interval data based on the overlap distance," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 30, no. 4, pp. 625–648, Aug. 2022.
[11] Y. Chen and L. Billard, "A study of divisive clustering with Hausdorff distances for interval data," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106969.
[12] L. Hedjazi, J. Aguilar-Martin, and M.-V. Le Lann, "Similarity-margin based feature selection for symbolic interval data," *Pattern Recognit. Lett.*, vol. 32, no. 4, pp. 578–585, Mar. 2011.
[13] C. H. Guo and Y. C. Liu, "Feature selection method for interval-type symbol data," *Oper. Res. Manage.*, vol. 24, no. 1, pp. 67–74, 2015.
[14] Q. Liu, J. H. Dai, and J. L. Chen, "Cost-sensitive feature selection for the interval-value data," *J. Nanjing Univ.*, vol. 57, no. 1, pp. 121–129, 2021.
[15] X. Qi, H. Guo, Z. Artem, and W. Wang, "An interval-valued data classification method based on the unified representation frame," *IEEE Access*, vol. 8, pp. 17002–17012, 2020.
[16] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.
[17] X. B. Qi, *Representation and Classification of the Interval-Type Data*. Taiyuan, China: Shanxi University, 2021.
[18] (2016). *The Weather of 243 Countries in the World*. [Online]. Available: https://rp5.ru/
[19] (2016). *Water Dataset, Barcelona Water Distribution Network*. [Online]. Available: http://lhedjazi.jimdo.com/useful-links
[20] S. Goel, R. Kumar, M. Kumar, and V. Chopra, "An efficient page ranking approach based on vector norms using sNorm(p) algorithm," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 1053–1066, May 2019.
[21] Q. Liu and A. L. Jiang, "A two-stage feature selection algorithm based on interaction information," *Comput. Eng. Design*, vol. 44, no. 1, pp. 125–132, 2023.

**XIAOBO QI** received the Ph.D. degree in computer science from Shanxi University, in 2021. She is currently an Associate Professor with Taiyuan Normal University. Her research interests include image processing, distributed computing, and data analysis.

**JINYU SONG** received the bachelor's degree in mathematics and applied mathematics from Xinzhou Teachers University, in 2021. She is currently pursuing the master's degree with Taiyuan Normal University. Her research interests include machine learning and data modeling.

**HUI QI** received the master's degree in computer science from Shanxi University, in 2009. She is currently a Professor with Taiyuan Normal University. Her research interests include machine learning, data mining, and image processing.

**YING SHI** received the master's degree in computer science from Shanxi University, in 2015. She is currently a Lecturer with Taiyuan Normal University. Her research interests include machine learning and image processing.