

Received December 24, 2019, accepted January 13, 2020, date of publication January 20, 2020, date of current version January 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967780

An Interval-Valued Data Classification Method Based on the Unified Representation Frame

XIAOBO QI¹, HUSHENG GUO¹, ZADOROZHNYI ARTEM¹, AND WENJIAN WANG^{1,2}

¹School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

²Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China

Corresponding author: Wenjian Wang (wjwang@sxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673249, Grant U1805263, Grant 61503229, and Grant 61703252, in part by the Key Research and Development Program of Shanxi Province (International Cooperation, 201903D421050), and in part by the Natural Science Foundation of Shanxi Province under Grant 201901D111033.

ABSTRACT Interval-valued data (IVD) is a kind of data where each feature is an interval. The midpoint and boundary are the two commonly used methods for representing IVD. However, their structure information (such as location, size) may be incomplete because only midpoint or endpoint is adopted which will lead to poor results of data processing. To better depict the structural information of IVD, a unified representation frame (URF) for IVD is proposed. It not only takes into account the size and location information, but the relationship between them as well. This frame can also represent the midpoint and boundary methods in a unified way. Besides, symmetrical uncertainty (SU) is adopted to measure the relationship between features and classes quantitatively, and irrelevant features will be eliminated based on SU. The proposed URF_ SU is applied in some traditional classifiers like LIBSVM, CART Tree and KNN. The experimental results on synthetic and real-world datasets demonstrate that the proposed approach is more effective than other representation methods of IVD in classification tasks.

INDEX TERMS Interval-valued data, unified representation frame, symmetrical uncertainty, feature selection.

I. INTRODUCTION

In many real situations, inaccuracy, uncertainty or variability may be in some important available information. Classical data are not able to describe the nuances, and other kinds of data, such as interval-valued data, are required. For instance, daily temperatures measured at meteorological stations can be considered as interval-valued data. The temperature values are measured hourly, but during this hour, they change continuously. Point data can only depict the temperature at a certain time; however, IVD can better describe daily temperatures variation. IVD is usually resulted from the limited range of the IVD themselves, the error caused by the repeated measurements, and data missing due to the incomplete information, etc. Compared with point data, IVD can express uncertainty and variability of data. Therefore, IVD's good representation and processing method is of great significance in decision-making process.

The associate editor coordinating the review of this manuscript and approving it for publication was Victor S. Sheng.

In distinct from point data, the difficulty of IVD lies in their representation. Roughly speaking, there are three main representation methods for IVD. (1) Midpoint method takes the midpoint as a special value of IVD, and uses traditional methods to deal with it [1]–[3]. This method only considers the internal condition of IVD, but loses the size information. (2) Boundary value method lets upper and lower boundary values replace IVD, and then deals with IVD in a general way [4]–[6]. In this method, the upper and lower values are regarded as two features, but it ignores the internal distribution of IVD. (3) Midpoint and radius method takes location information into account on the basis of boundary values [7], [8]. Reference [7] used the traditional regression method to generate regression equations for the midpoint and radius respectively, then predicted the upper and lower bounds of IVD with generated equations. Reference [8] represented IVD by the midpoint and radius, then predicted these two independent variables by symmetric linear regression model. The method considers both internal condition and size information, but there is no correlation between these two elements.

The researches on IVD in recent years are mainly focusing on clustering analysis [9], regression analysis [10], principal component analysis [11]–[14] and discriminant analysis [15]–[23], less on classification tasks. Typical classification methods cannot fit for processing IVD directly because they do not address the inherent uncertainty of IVD. Reference [24] proposed a novel feature selection approach for supervised interval valued features, which can achieve good results for interval-valued data classification. The existing representation methods lose either size information or location information, but do not notice the relationship between them. Besides, they may even have twice the number of features of the original IVD. In this paper, a united representation frame, which only contains the same number of features as the original IVD and considers the relationship between midpoint and radius, is proposed. Irrelevant features are a major obstacle in classification problems, symmetrical uncertainty measures the relationship between features and classes quantitatively. Therefore, an interval-valued data classification method based on URF, namely, URF_SU, is proved to be able to select related features and obtain a good classification performance.

The rest of this paper is organized as follows. In Section II, we introduce the preliminaries about IVD, and explain the classification method in detail. In Section III, first, the experimental datasets are carefully depicted, then experimental results and analysis are presented. Finally, some conclusions and plans for future work are given in Section IV.

II. AN INTERVAL-VALUED DATA CLASSIFICATION METHOD BASED ON THE UNIFIED REPRESENTATION FRAME

To illustrate the unified representation frame clearly, a brief introduction to some basic concepts, such as the definitions about IVD unit and interval-valued matrix, is given. Then the proposed classification method will be explained in depth.

A. THE UNIFIED REPRESENTATION FRAME FOR IVD

The so-called IVD unit refers to the value for a certain range, and can be expressed as an interval. The relevant definitions are as follows.

Definition 1: (Interval-Valued Data Unit): 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): Denote $U = [u_{ij}]$ as an $n \times p$ interval-valued matrix U , i.e.,

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

where $U_j = ([u_{1j}^-, u_{1j}^+], [u_{2j}^-, u_{2j}^+], \dots, [u_{nj}^-, u_{nj}^+])^T$ 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): 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 $\alpha \in [0, 1]$, α can be regarded as the adjustment factor of IVD unit, which is used to balance the relationship between the midpoint and radius of the IVD unit. The midpoint-radius matrix is constructed as:

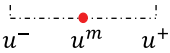
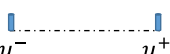
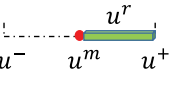
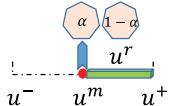
$$U^{mr} = (U_1^{mr}, U_2^{mr}, \dots, U_p^{mr}) = \begin{pmatrix} u_{11}^{mr} & u_{12}^{mr} & \dots & u_{1p}^{mr} \\ u_{21}^{mr} & u_{22}^{mr} & \dots & u_{2p}^{mr} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1}^{mr} & u_{n2}^{mr} & \dots & u_{np}^{mr} \end{pmatrix}. \quad (5)$$

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

The existing methods of IVD representation can be incorporated into the unified representation frame. Table 1 shows that the URF includes entirety and only has one feature. When $\alpha = 1$, it can be treated as midpoint method with one feature. When $\alpha = 0$, it is expressed by radius that contains the size information of u^- and u^+ , similar to boundary value method with two features. When $\alpha \in (0, 1)$, it contains both factors, i.e., u^m and u^r , like midpoint and radius method. Meanwhile, it measures the relationship between them, but MR method does not. And more, URF only has one feature, which is half of the feature number of MR method. Therefore, different from MR method, URF contains more information (midpoint, radius and their relationship).

Midpoint and radius are derived from interval-valued data, which represents the location or size information. Although the midpoint and radius can be considered at same time, they are often regarded as independent features, which means that the relationship between them will not be applied. For example, it is impossible to determine which couple of midpoint and radius are from the same IVD when there are many features (they may be listed disorderly). Therefore, the corresponding relationship between the midpoint and radius should not be omitted when the IVD is converted to discrete values. The URF is a unified frame to represent interval-valued data. It can not only represent the location and size information, but also the corresponding relationship between them.

TABLE 1. Representation for IVD.

Method	Abbr.	Description	Representation	Num. of features
Midpoint method	M		u^m	1
Boundary value method	BV		u^-, u^+	2
Midpoint and radius method	MR		u^m, u^r	2
Unified representation frame	URF		$\alpha u^m + (1 - \alpha)u^r$	1

B. FEATURE SELECTION BASED ON THE UNITED REPRESENTATION FRAME FOR IVD

It is unavoidable to have irrelevant features in some datasets, but these features may cause more computational costs and lead to the over-fitting of an algorithm. For interval-valued data, it is hard to select effective features directly by conventional methods, meanwhile, the existing IVD's feature selection algorithms rank, weigh and analyze features only using midpoint or endpoint [13], [14], [25]–[27]. Symmetrical uncertainty measures the correlation degree between features and classes quantitatively, so this indicator is adopted to select the relevant features for IVD based on URF. In this method, we need to calculate the SU value between each feature and class first, then select the features with the greater SU values.

Assuming that the feature set for IVD is U^{mr} , and the class is $Y = \{y_1, y_2, \dots, y_m\}$. SU value is calculated for each feature as follows

$$SU_j = 2 \left[\frac{IG(Y|U_j^{mr})}{H(Y) + H(U_j^{mr})} \right], \tag{6}$$

where $H(Y)$ and $H(U_j^{mr})$ are information entropy, indicating the information quantity of the features. $IG(Y|U_j^{mr})$ is information gain, representing the amount of information shared between two variables.

Usually information gain is calculated by information entropy and shown below

$$IG(Y|U_j^{mr}) = H(Y) - H(Y|U_j^{mr}), \tag{7}$$

where

$$H(Y) = - \sum_{y_t \in Y} p(y_t) \log_2(p(y_t)), \tag{8}$$

$$H(Y|U_j^{mr}) = - \sum_{u_{ij}^{mr} \in U_j^{mr}} p(u_{ij}^{mr}) \sum_{y_t \in Y} p(y_t|u_{ij}^{mr}) \log_2(p(y_t|u_{ij}^{mr})). \tag{9}$$

Here, $p(\cdot)$ is the probability, and $p(\cdot|*)$ is the conditional probability. A simple proof is given as follows:

information entropy:

$$H(Y) = - \sum_{y \in Y} p(y) \log_2 p(y), \tag{10}$$

conditional entropy:

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x), \tag{11}$$

information gain:

$$\begin{aligned} IG(Y|X) &= \sum_{x \in X, y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x \in X, y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)} \\ &\quad - \sum_{x \in X, y \in Y} p(x, y) \log_2 p(y) \\ &= \sum_{x \in X, y \in Y} p(x)p(y|x) \log_2 p(y|x) \\ &\quad - \sum_{x \in X, y \in Y} p(x, y) \log_2 p(y) \\ &= - \sum_{x \in X} p(x)H(Y|X = x) - \sum_{y \in Y} p(y) \log_2 p(y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{12}$$

C. THE URF_SU CLASSIFICATION ALGORITHM

Given a data set $\mathbf{T} = \{(U, Y)\}$, where U is the interval-valued data set, and Y is the classification identification. The main idea of proposed URF_SU method is to calculate the SU values and get the feature matrix M arranged in descending order of SU values first. Then the classifier learns and verifies each new feature $m_i(m_i \in M)$ when it's added to the empty feature subset E in turn, until the corresponding stopping condition is reached. The stopping condition here is that the classification accuracy begins to decrease or the number of feature subset has reached the defined maximum θ .

The main steps of the proposed URF_SU method are summarized as *Algorithm 1*.

TABLE 2. Interval-valued dataset U and label Y .

U	temperature	atmospheric pressure	humidity	horizontal visibility	dew-point temperature	Y
u_1	[-9.0,4.3]	[1031.3,1037.8]	[28,65]	[15.0,20.0]	[-15.5,-10.4]	taiyuan
u_2	[-5.8,5.2]	[1032.1,1036.7]	[20,52]	[125.0,25.0]	[-15.9,-11.1]	taiyuan
u_3	[-9.5,4.7]	[1029.3,1035.4]	[36,89]	[5.0,10.0]	[-11.0,-4.8]	taiyuan
u_4	[-10.1,1.4]	[1035.7,1041.0]	[54,90]	[5.0,6.0]	[-11.4,-6.4]	taiyuan
u_5	[-6.5,0.0]	[1031.7,1039.3]	[52,85]	[8.0,10.0]	[-8.9,-7.8]	taiyuan
u_6	[-8.1,9.0]	[1019.8,1030.9]	[19,76]	[125.0,15.0]	[-13.6,-7.8]	taiyuan
u_7	[-4.6,7.5]	[1020.3,1028.9]	[23,62]	[18.0,20.0]	[-12.4,-9.4]	taiyuan
u_8	[-6.6,3.4]	[1028.6,1032.3]	[32,90]	[3.0,20.0]	[-11.7,-5.1]	taiyuan
u_9	[-7.7,3.9]	[1026.2,1030.5]	[47,90]	[6.0,7.0]	[-9.1,-4.2]	taiyuan
u_{10}	[-4.6,2.7]	[1025.7,1029.5]	[54,85]	[7.0,7.0]	[-6.7,-5.1]	taiyuan
u_{11}	[-5.1,2.7]	[1022.9,1027.0]	[39,73]	[7.0,15.0]	[-14.4,-5.3]	beijing
u_{12}	[-6.5,4.1]	[1021.5,1024.5]	[18,45]	[10.0,15.0]	[-18.1,-16.1]	beijing
u_{13}	[-6.1,3.0]	[1023.7,1028.5]	[15,40]	[10.0,15.0]	[-21.1,-16.3]	beijing
u_{14}	[-4.3,2.7]	[1028.7,1032.1]	[14,27]	[20.0,30.0]	[-22.6,-20.0]	beijing
u_{15}	[-6.3,1.1]	[1030.7,1033.7]	[12,21]	[25.0,30.0]	[-26.7,-22.1]	beijing
u_{16}	[-9.9,-3.4]	[1034.9,1040.9]	[15,22]	[30.0,30.0]	[-27.6,-23.6]	beijing
u_{17}	[-6.2,-4.2]	[1034.7,1040.5]	[17,85]	[18.0,20.0]	[-27.2,-6.3]	beijing
u_{18}	[-5.3,0.3]	[1026.3,1032.3]	[53,87]	[6.0,20.0]	[-9.1,-7.1]	beijing
u_{19}	[-4.4,4.1]	[1025.0,1026.9]	[40,87]	[6.0,20.0]	[-12.3,-6.2]	beijing
u_{20}	[-5.8,5.4]	[1023.7,1027.9]	[42,88]	[4.0,10.0]	[-7.5,-5.1]	beijing

Algorithm 1 URF_SU**Input:** input an interval-valued dataset U and label Y **Output:** output the final classification accuracy acc and optimal feature subset E

- 1: Initialize: $E = \emptyset$.
- 2: Convert IVD to midpoint and radius with Eqs.(2) and (3), then construct unified representation frame for IVD according to Eq.(4).
- 3: Tune the parameter α , then select the best α and obtain the optimal URF.
- 4: For each attribute, calculate the SU value between each feature and its class with Eq.(6).
- 5: Rank the features in descending order of SU values, and get the sorted feature matrix M .
- 6: Add each feature $m_i(m_i \in M)$ sequentially to the empty feature subset E , then learn and classify in each cycle.
- 7: Go to Step5 until the stop condition is reached (accuracy drops or θ reaches the maximum).

We give the following example for illustration of the proposed algorithm.

Example1: Consider the example is from meteorological data labeled by Taiyuan and Beijing. Let $U = \{u_1, u_2, \dots, u_{20}\}$ be a set of 20 days of weather, $F = \{\text{temperature, atmospheric pressure, humidity, horizontal visibility, dewpoint temperature}\}$ be the feature set. The dataset is shown in Table 2.

Step 1: Input the interval-valued dataset U and label Y as in Table 2.

Step 2: Initialize: $E = \emptyset$.

Step 3: Convert IVD to midpoint and radius with Eqs.(2) and (3), then construct the unified representation frame for IVD according to Eq.(4).

Step 4: Let $\alpha = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 1$. Then select the best $\alpha = 0.5$, and obtain the optimal URF for IVD as in Table 3.

Step 5: Calculate the SU value with Eq.(6), and get the sorted feature matrix M by descending order of SU values as in Table 4.

Step 6: Add each feature sequentially to the empty feature subset E , and calculate the acc in each cycle as in Table 5.

From Table 5, we can see that the accuracy is highest except for the horizontal visibility, so we select the first four features for classification. Because of the small number of samples, the overall accuracy is not very good.

For the $n \times p$ interval-valued dataset U , n is the number of samples, p represents the feature dimension.

In the symmetrical uncertainty estimation stage, the time complexity is $O(p \cdot n^2)$. In the feature sorting stage, the time complexity is $O(p \log p)$. In the iterative computation process, the features are added gradually and the worst time complexity is $O(p \cdot n^2)$. Therefore, in the worst case, the time complexity of URF_SU is $O(p \log p + 2p \cdot n^2)$. The M_SU has the same time cost with URF_SU, but the time complexity of BV_SU is $O(2p \log 2p + 4p \cdot n^2)$ as the number of features is doubled.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, some synthetic datasets and comparison methods are prepared for the experiments. Then the experiments on the synthetic and real-world datasets are carried out to evaluate the effectiveness of the proposed approach.

A. EXPERIMENTAL DATA AND COMPARISON METHODS

In order to test the validity of proposed approach, eight datasets, containing four synthetic datasets and four

TABLE 3. Midpoint-radius matrix (URF).

U	temperature	atmospheric pressure	humidity	horizontal visibility	dew-point temperature	Y
u_1	2.15	518.9	32.5	10	-5.2	taiyuan
u_2	2.6	518.35	26	12.5	-5.55	taiyuan
u_3	2.35	517.7	44.5	5	-2.4	taiyuan
u_4	0.7	520.5	45	3	-3.2	taiyuan
u_5	0	519.65	42.5	5	-3.9	taiyuan
u_6	4.5	515.45	38	7.5	-3.9	taiyuan
u_7	3.75	514.45	31	10	-4.7	taiyuan
u_8	1.7	516.15	45	10	-2.55	taiyuan
u_9	1.95	515.25	45	3.5	-2.1	taiyuan
u_{10}	1.35	514.75	42.5	3.5	-2.55	taiyuan
u_{11}	1.35	513.5	36.5	7.5	-2.65	beijing
u_{12}	2.05	512.25	22.5	7.5	-8.05	beijing
u_{13}	1.5	514.25	20	7.5	-8.15	beijing
u_{14}	1.35	516.05	13.5	15	-10	beijing
u_{15}	0.55	516.85	10.5	15	-11.05	beijing
u_{16}	-1.7	520.45	11	15	-11.8	beijing
u_{17}	-2.1	520.25	42.5	10	-3.15	beijing
u_{18}	0.15	516.15	43.5	10	-3.55	beijing
u_{19}	2.05	513.45	43.5	10	-3.1	beijing
u_{20}	2.7	513.95	44	5	-2.55	beijing

TABLE 4. Feature rank and SU value.

Rank	1	2	3	4	5
Sorted features (M)	atmospheric pressure (AP)	humidity (H)	dew-point temperature (DPT)	temperature (T)	horizontal visibility (HV)
SU value	0.3758	0.3633	0.3514	0.3460	0.2240

TABLE 5. Accuracy varied with added features.

Feature set (E)	{AP}	{AP, H}	{AP, H, DPT}	{AP, H, DPT, T}	{AP, H, DPT, T, HV}
Accuracy	34.0	41.33	51.67	57.33	26.67

TABLE 6. Datasets used in experiments.

Dataset	Num. of samples	Num. of features	Num. of classes
Ds1	3000	6	2
Ds2	4500	6	3
Ds3	1493	4	3
Ds4	2000	4	4
HS_ Ds	7302	5	2
TB_ Ds	7302	5	2
HSTB_ Ds	14604	5	4
Water	316	48	2

real-world datasets, are used in the experiments. They are listed in Table 6.

The IVD in Ds1 and Ds2 are constructed from the seed data and the formula is $[z - r, z + r]$. z is the seed data, generated according to the normal distribution; r is the width, drawn from the uniform distribution. The classes of Ds1 and Ds2 are mainly separated by location and specific configurations are shown in Table 7, where r_5, r_6 are the irrelevant values.

Ds1 is composed of class 1 and 2, while Ds2 consists of all the three classes. Ds3 and Ds4 are the same simulation generation method as [28], and the label is determined by the midpoint symbols of first two features. So in four synthetic

datasets, the classes are basically separated by location. Fig. 1 shows the distribution of four synthetic datasets.

The first three real-world datasets contains meteorological data of ten years (from January 1, 2006 to December 31, 2015) provided by the ‘Reliable Prognosis’ site [29]. There are two classes of equal sample sizes of 3651 in HS_ Ds and TB_ Ds, and four classes in HSTB_ Ds, labeled by different cities: Harbin, Sanya, Taiyuan and Beijing. Each sample is described by 5 interval attributes {temperature, atmospheric pressure, humidity, horizontal visibility and dew-point temperature (the temperature of water vapor in the air when it becomes dewdrop)}.

The last real-world water dataset concerns 30-min flow records of 1 year (from June 1,2003 to May 31,2004) of Barcelona water distribution network [30]. It contains only 316 days of data, each day is characterized by 48 30-min interval features where each feature characterizes the maximal and minimal variation observed based on three consecutive 10-min flow measure. Each day is labeled by one of two classes according to the type of the day: weekends (Saturday, Sunday, Holidays) or workdays.

In the experiments, the proposed SU-based method is used for feature selection. LDA and PCA are adopted as

TABLE 7. Specific configurations in Ds1 and Ds2.

Conf.	z	r_1	r_2	r_3	r_4	r_5	r_6
Class 1	$\mu = 20, \sigma = 2$	$U(1, 5)$	$U(6, 10)$	$U(7, 11)$	$U(2, 7.5)$	$U(0, 1)$	$U(0, 1)$
Class 2	$\mu = 20.8, \sigma = 2$	$U(1.5, 5)$	$U(6, 10.5)$	$U(6.5, 11)$	$U(2, 7)$	$U(0, 1)$	$U(0, 1)$
Class 3	$\mu = 20, \sigma = 1$	$U(2, 5)$	$U(7, 9)$	$U(6, 11)$	$U(1, 6)$	$U(0, 1)$	$U(0, 1)$

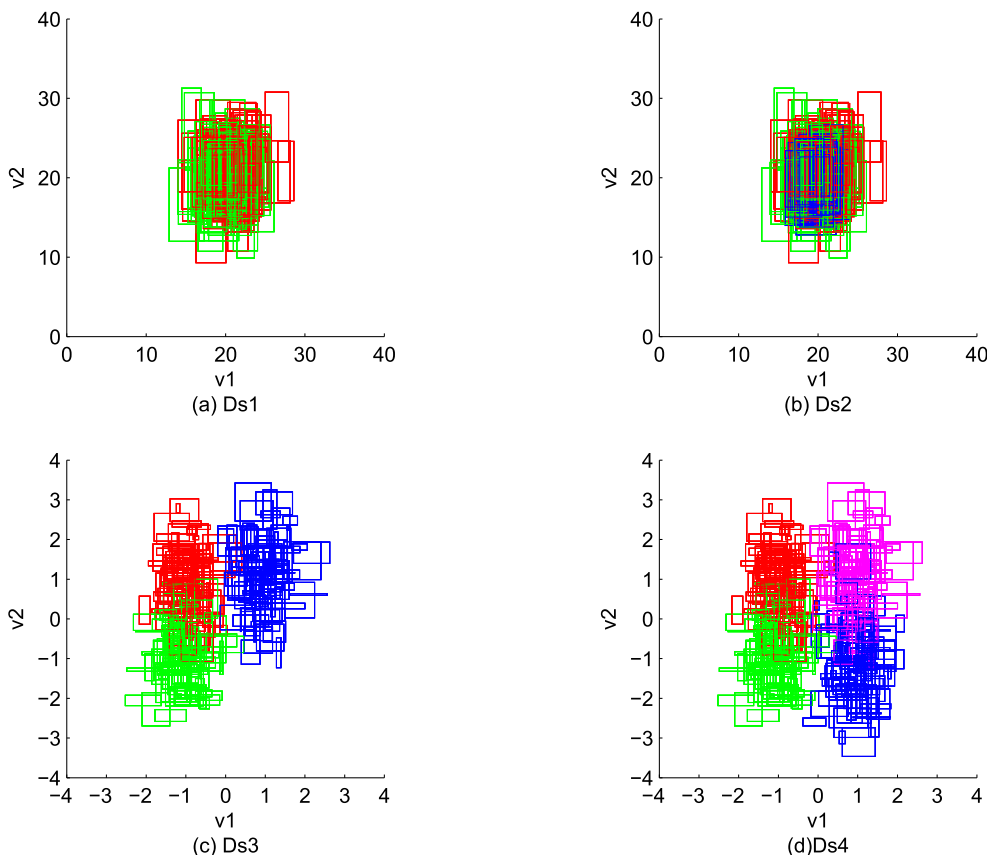


FIGURE 1. Distribution of synthetic datasets.

other two compared feature selection methods. In order to get the most reliable original information, PCA methods choose the features whose cumulative contribution rate is not less than 95%, while LDA methods choose at most $|Num. of classes - 1|$ features. On the representation of interval-valued data, the proposed URF is also compared with existing M-based and BV-based methods. Therefore, 9 models for feature selection are testified on 4 synthetic and 4 real-world datasets, respectively. In addition, three general classifiers (LIBSVM, CART Tree and KNN) are used to verify the performance of 9 methods. All methods are carried out on the same dataset. Therefore, the description of reference methods is shown in Table 8.

B. TUNING PARAMETER α

URF is a unified representation frame, and α is used to balance the tendency of midpoint and radius. Generally, α needs to be set firstly, and it can take different values for

TABLE 8. Reference methods in experiments.

Method abbr.	Representation			Feature selection		
	BV	M	URF	LDA	PCA	SU
BV_LDA	✓			✓		
BV_PCA	✓				✓	
BV_SU	✓					✓
M_LDA		✓		✓		
M_PCA		✓			✓	
M_SU		✓				✓
URF_LDA			✓	✓		
URF_PCA			✓		✓	
URF_SU			✓			✓

different IVD values. Of course, it may require additional computation to select a suitable α for each value, so α is set the same for a dataset here for simplicity. The algorithms URF_LDA, URF_PCA and URF_SU all need to select suitable parameter and we choose it through experiments. For each dataset, let $\alpha = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 1$. Those α

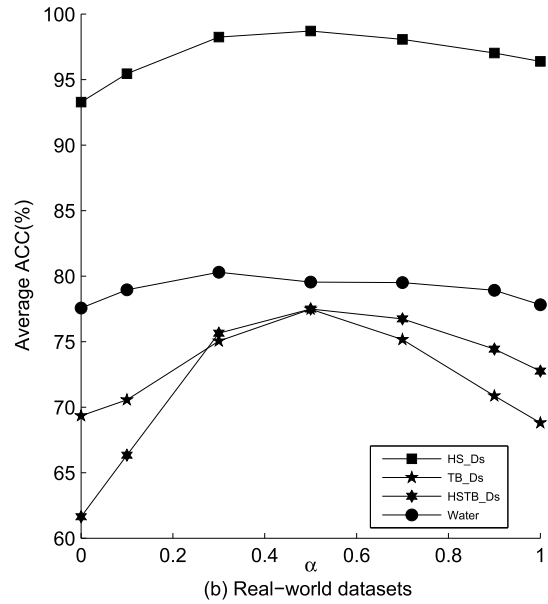
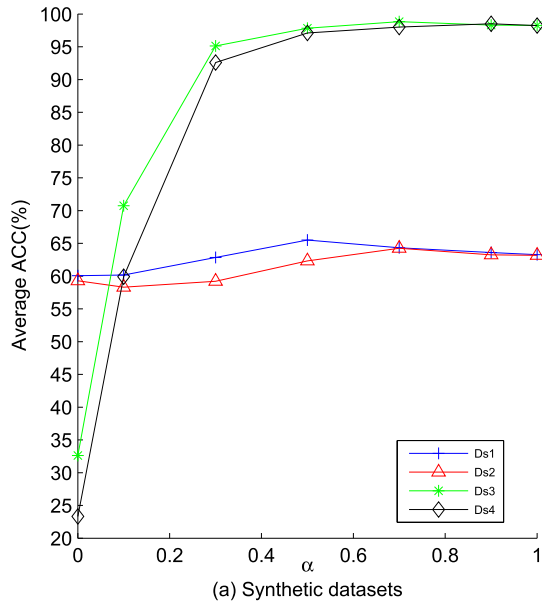


FIGURE 2. Selection of α .

with the best result will be selected. Fig.2 gives the selection results of parameter α and the ‘Average ACC’ refers to the average accuracy of ten times experiments for each α .

From Fig.2, it can be clearly seen that the trends of different datasets vary greatly. In Fig.2(a), the curves of Ds1 and Ds2 are relatively gentle, while the curves of Ds3 and Ds4 are basically on the rise. When α reaches 0.3, the average accuracies of Ds3 and Ds4 are much better than those of Ds1 and Ds2. When $\alpha = 0$, the precision of Ds3 and Ds4 decreases significantly because URF only has radius. It indicates that only size information may not represent IVD well. In Fig.2(b), all trends increase first and then decrease. The results of HS_Ds changes smoothly and is much better than that of other three datasets. The trend of Water is also gentle and the average accuracy is better than that of TB_Ds and HSTB_Ds, but the trends of TB_Ds and HSTB_Ds change greatly. When $\alpha = 0$ or 1, the accuracies reduces obviously because the interval-valued data is only represented by radius or midpoint. It is verified again that only size or location information is not comprehensive enough to represent interval-valued data. Therefore, both midpoint and radius play an important role in the representation of IVD. On each dataset, the influence of location and size information may be different. Hence, the α on different datasets are set with different values and listed in Table 9.

C. RESULTS OF FEATURE SELECTION

In this section, the feature selection results are shown in Fig.3. It indicates the feature ratio of each method on synthetic and real-world datasets.

In Fig.3, the rows (different colors) and columns represent different datasets and different comparison methods, respectively. In Fig.3(a) (synthetic datasets), the ratio of irrelevant features is 25 to 30 percent of total features.

TABLE 9. Appropriate α on different datasets.

Dataset	Appropriate α
Ds1	0.5
Ds2	0.7
Ds3	0.7
Ds4	0.9
HS_Ds	0.5
TB_Ds	0.5
HSTB_Ds	0.5
Water	0.3

In Fig.3(b), the ratio of the irrelevant features account is 25 percent on 3 meteorological datasets and 50 percent on Water dataset. It can be observed that 5 methods (BV_SU, M_PCA, M_SU, URF_PCA and URF_SU) can eliminate all irrelevant features on Ds1, Ds2 and Ds3. Another 5 methods (M_PCA, M_SU, URF_LDA, URF_PCA and URF_SU) remove all irrelevant features on Ds4. Although other methods obtain less features, they delate some relevant features which lead to poor classification results (see Table 10 in next section). Overall, the BV-based and LDA-based methods do not work well in feature selection on synthetic datasets. For the 4 real-world datasets, only the proposed URF_SU selects all relevant features. Other methods delete some relevant features in varying degrees except for irrelevant features, especially the LDA-based methods (BV_LDA, M_LDA and URF_LDA) eliminate too much relevant features on Water dataset.

Experiments both on synthetic and real-world datasets demonstrate that the proposed URF_SU can always select all relevant features. Certainly, the good result of features selection may not mean less features but the proper features which are contribute to improve classification accuracy. It is supported by the following classification performance comparison experiments, and it also shows the superiority of

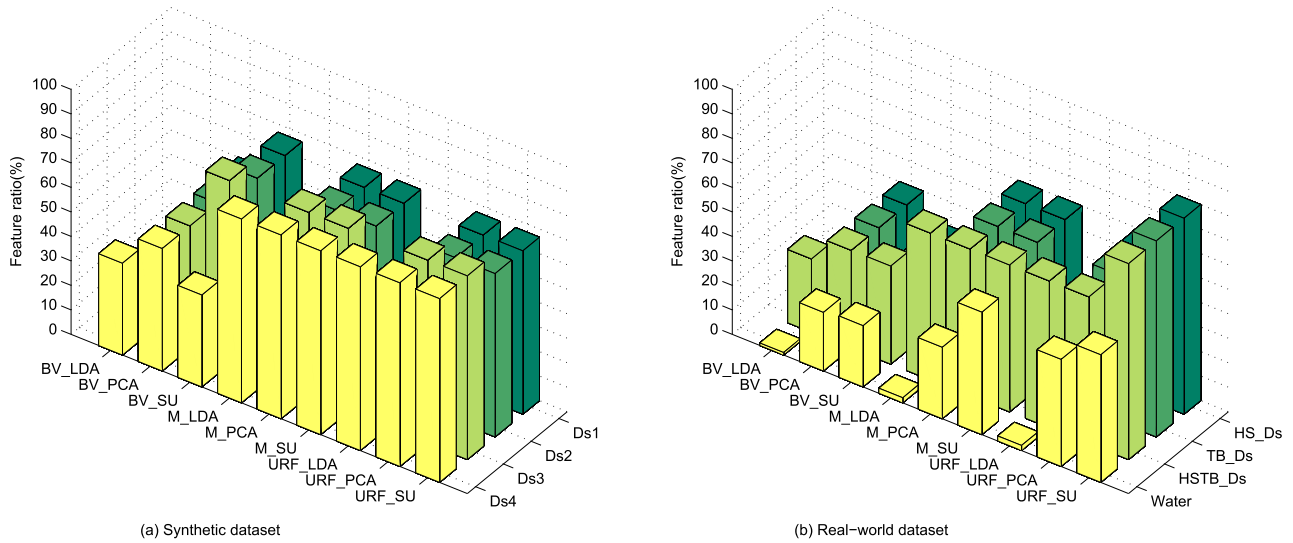


FIGURE 3. Selected feature ratio on different datasets.

TABLE 10. Classification accuracies of the methods (mean(st.dev)).

Dataset	Boundary matrix			Midpoint matrix			Midpoint-radius matrix		
	BV_LDA	BV_PCA	BV_SU	M_LDA	M_PCA	M_SU	URF_LDA	URF_PCA	URF_SU
LIBSVM									
Ds1	65.11(3.17)	61.72(3.05)	62.64(2.66)	64.13(2.55)	63.00(2.59)	63.40(2.93)	65.28(3.36)	65.31(1.95)	65.72(3.21)
Ds2	59.90(1.15)	64.31(1.87)	67.50(2.83)	55.67(1.75)	63.01(2.70)	63.36(1.82)	56.02(1.91)	63.80(2.02)	64.36(2.89)
Ds3	96.99(0.69)	99.00(0.86)	99.08(0.53)	96.77(1.53)	98.66(1.14)	98.94(1.03)	96.01(2.19)	98.54(0.94)	99.28(0.70)
Ds4	97.55(1.00)	98.44(0.76)	99.30(0.67)	97.50(1.12)	97.92(1.05)	98.69(1.03)	97.58(0.93)	97.84(1.67)	98.71(0.85)
HS_Ds	98.74(0.48)	92.22(0.88)	95.50(0.63)	95.87(0.60)	97.27(0.50)	94.64(0.68)	97.98(0.59)	94.44(0.57)	97.18(0.43)
TB_Ds	73.55(1.99)	65.64(1.44)	76.49(1.20)	68.51(2.07)	67.87(1.96)	72.47(1.82)	73.39(1.43)	73.21(1.84)	77.81(1.34)
HSTB_Ds	78.51(1.57)	60.19(0.92)	74.33(0.97)	75.12(0.77)	66.05(1.18)	74.27(1.01)	75.09(0.97)	68.55(1.28)	78.68(0.86)
Water	70.73(5.90)	75.59(8.30)	78.01(4.94)	70.21(7.65)	79.47(5.17)	79.63(7.53)	70.49(7.62)	80.01(6.96)	80.82(9.02)
Average	80.14	77.14	81.61	77.97	79.16	80.68	78.98	80.21	82.82
CART Tree									
Ds1	56.68(2.87)	59.22(2.38)	58.95(1.34)	56.82(2.46)	56.26(3.25)	56.66(3.85)	56.54(2.39)	56.76(3.75)	57.04(2.77)
Ds2	50.43(1.97)	58.95(3.03)	59.28(2.16)	47.08(2.78)	52.84(2.46)	53.65(2.16)	48.58(2.24)	53.21(1.95)	54.26(3.57)
Ds3	96.69(0.81)	96.90(1.68)	99.16(0.74)	97.51(1.45)	97.14(1.10)	99.81(0.32)	97.19(1.28)	96.87(1.51)	99.22(0.64)
Ds4	97.08(1.32)	96.05(1.99)	98.72(0.71)	97.92(0.83)	96.79(1.29)	99.91(0.20)	97.61(1.45)	96.47(1.94)	99.56(0.55)
HS_Ds	98.07(0.41)	97.59(0.45)	95.27(0.91)	94.24(0.74)	96.89(0.43)	93.90(0.48)	97.19(0.51)	92.61(0.69)	96.44(0.51)
TB_Ds	65.97(1.36)	67.19(1.21)	74.60(1.14)	61.11(2.18)	63.18(2.52)	68.82(1.31)	65.44(2.10)	68.66(2.05)	74.76(1.55)
HSTB_Ds	72.76(0.90)	63.42(1.14)	75.16(1.26)	69.62(0.92)	62.36(0.98)	70.66(0.87)	68.93(1.35)	62.60(1.42)	74.65(1.27)
Water	63.40(7.20)	69.49(6.98)	71.71(7.35)	68.98(10.61)	71.31(6.20)	71.66(6.50)	77.88(8.48)	70.94(7.05)	72.66(6.20)
Average	75.14	76.10	79.11	74.16	74.60	76.88	76.17	74.77	78.57
KNN									
Ds1	58.16(2.10)	57.76(3.31)	58.15(2.85)	58.78(3.93)	56.91(2.83)	58.25(2.27)	58.62(1.98)	57.61(2.04)	58.98(2.90)
Ds2	52.46(2.31)	59.55(2.13)	60.87(2.31)	48.38(1.59)	55.78(2.06)	56.42(2.74)	48.99(2.44)	55.95(2.60)	57.13(2.18)
Ds3	97.62(0.96)	98.41(0.85)	98.44(1.25)	97.87(1.07)	97.99(1.14)	98.66(0.83)	97.04(1.53)	97.59(1.17)	98.70(0.95)
Ds4	97.48(0.77)	97.30(1.04)	98.79(0.94)	98.16(0.84)	97.20(1.48)	98.23(0.84)	97.78(1.17)	97.26(0.91)	97.61(1.16)
HS_Ds	98.55(0.53)	98.37(0.46)	94.93(0.53)	95.25(0.61)	97.45(0.76)	93.87(0.78)	97.50(0.64)	93.96(0.61)	96.67(0.55)
TB_Ds	67.01(3.27)	69.74(1.58)	72.28(0.90)	61.59(1.75)	64.49(1.19)	67.12(2.55)	66.35(2.72)	68.34(2.00)	72.94(1.19)
HSTB_Ds	75.99(1.46)	66.36(1.27)	75.69(1.03)	72.33(1.43)	64.36(1.12)	70.91(1.43)	71.94(0.97)	65.54(1.29)	76.57(1.68)
Water	60.38(8.20)	76.14(8.38)	76.72(7.33)	67.50(6.43)	75.09(5.19)	76.53(7.93)	79.29(6.33)	75.56(7.74)	77.71(9.55)
Average	75.96	77.95	79.48	74.98	76.16	77.50	77.19	76.48	79.54

the URF_SU than other reference methods under the same feature selection ratios.

D. PERFORMANCE

In this section, the experiments results are analyzed. Then we compare the stability and applicability of the approaches.

1) CLASSIFICATION RESULTS

In order to verify the performance of the proposed URF_SU methods on classification tasks, we choose three traditional classifiers LIBSVM, CART Tree and KNN. For LIBSVM, RBF kernel is adopted with $\gamma = 0.2$, To reduce the experimental error, the penalty factor is set to the default ($C = 1$)

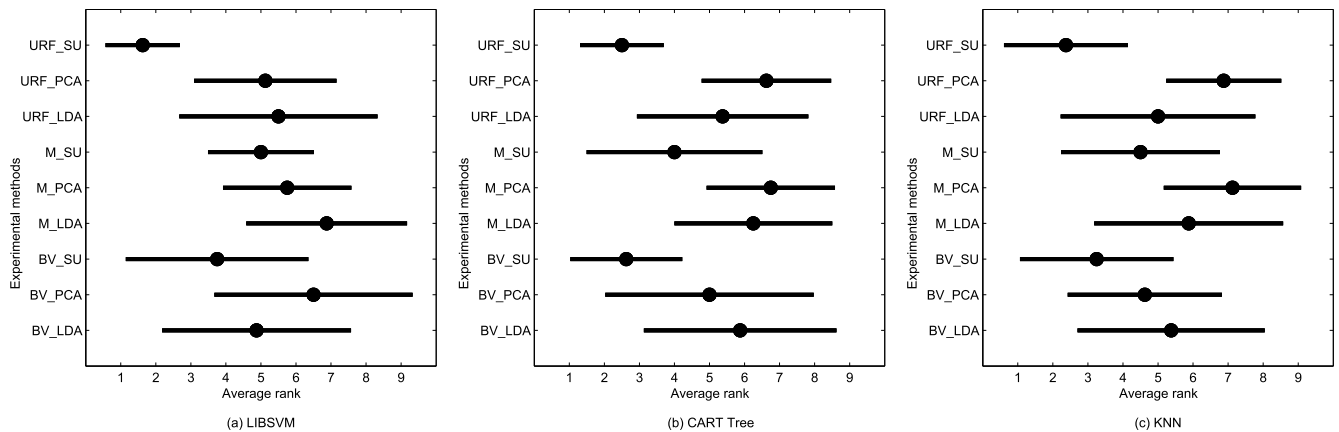


FIGURE 4. Ranks of experimental methods (average ± standard deviation).

directly. CART Tree is divided by Gini coefficient and there is no parameter for it. For KNN, Euclidean distance is adopted to choose nearest neighbors, and K is generally related to task in KNN. High accuracy can be reached when $K = 4$. All these parameters are obtained through experiments.

To avoid randomness, we repeat experiments ten times for each dataset. Each dataset is randomly divided into ten parts with same size, and each part is used as test data and the other nine parts are used as train data. The average of ten experiment results will be the final results listed in Table 10.

From Table 10, it can be perceived that URF_SU achieve more better results on three classifiers. For LIBSVM classifier, there is only one maximum value on BV_LDA method, others are on SU-based methods, especially on URF_SU method. In addition, the accuracies of URF_SU on TB_Ds, HSTB_Ds and Water are especially prominent. The lowest increased percentage is 1.01% on Water compared to URF_PCA, while the highest increased percentage reaches up to 30.72% on HSTB_Ds compared to BV_PCA. For CART Tree classifier, the maximum values are more dispersed, and only five optimal values are distributed on the SU-based methods. For TB_Ds, the lowest increased percentage is 2.14% compared to BV_SU, while the highest increased percentage reaches up to 22.34% compared to M_LDA. For KNN classifier, there are only two best values on LDA-based methods, and most on URF_SU. The lowest increased percentage is 0.34% on Ds1 compared to M_LDA, while the highest increased percentage reaches up to 18.97% on HSTB_Ds compared to M_PCA. Therefore, the results on the different classifiers demonstrate that the SU-based methods can achieve better results. Moreover, BV-based methods performs better than M-based methods, but BV-based method has more features and high time complexity. Above experiments support that the proposed URF_SU performs more prominently, especially on the real-world datasets.

Besides, literatures [24]–[26] reported some effective feature selection methods and provided experiment results on some real-world datasets. Hence, the comparison of the proposed methods with methods presented in above literatures on the same Water dataset is given in Table 11.

TABLE 11. Comparison of proposed feature selection method with existing methods on Water dataset.

Methods	Feature subset	Accuracy (%)	
Proposed method (LIBSVM)	URF_SU	25	80.82
	M_SU	24	79.63
	BV_SU	24	78.01
Proposed method (CART Tree)	URF_SU	25	72.66
	M_SU	24	71.66
	BV_SU	24	71.71
Proposed method (KNN)	URF_SU	25	77.72
	M_SU	24	76.53
	BV_SU	24	76.72
D.S. et al. [24]	C-1	2	76.34
	C-2	10	79.57
Lyamine et al. [25]	LAMDA	14	77
Chih-ching et al. [26]	LAMDA	11	78.66

Note: the experimental results of C-1, C-2 and LAMDA originated from [24].

From Table 11, it is clearly that the proposed methods on LIBSVM outperforms on CART Tree and KNN. And the proposed URF_SU and M_SU on LIBSVM are always superior to other referenced methods including that mentioned literature in terms of accuracy. Other SU-based are comparable with referenced methods. For the results of feature selection, the proposed approaches do not work well, and the reason may be that some redundant features are not eliminated. In the future, we will pursue to handle redundant features so as to achieve better results.

2) STABILITY AND APPLICABILITY

To illustrate the advantages and disadvantages of each method, Fig.4 shows the average accuracy ranks and the standard deviations of each approach on different classifiers.

Fig.4 indicates that the URF_SU on the three classifiers is not only ranked first, but also the standard deviation is small. Other conclusions: (1) For LIBSVM classifier, the URF_SU is clearly ranked top priority and the standard deviations is low (1.06), while the others are lagging behind. (2) For CART Tree classifier, the rank of BV_SU is very close to the URF_SU, but the standard deviation of URF_SU is smaller (1.20);

(3) For KNN classifier, the URF_SU behaves best and also its standard deviation is the smallest (1.77). For these three classifiers, URF-based methods always rank in the top 3, which means that they are more effective than other representation methods. In general, the smaller standard deviation of URF_SU indicates that this method has higher stability. Hence, we can conclude that the robustness of URF_SU is the best especially on LIBSVM from the above experimental results.

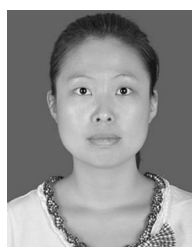
IV. CONCLUSION

The unified representation frame is proposed to solve IVD's representation problem. It can incorporate the existing representation methods and make the midpoint and radius reach a good compromise by adjustment factor. Although the URF is verified to be suitable for classification, it is also feasible for other tasks, like clustering, regression and other issues for IVD. Therefore, the URF is a generalization representation method for IVD, which provides a basis for the processing and analysis of IVD.

Although the proposed model can effectively delete irrelevant features, there may be still some redundant features in all selected relevant features. They are not guaranteed to be eliminated in the proposed method. In future, we will consider to handle the redundant features so as to achieve better results. Besides, the parameter α of the proposed method is the same on a dataset for simplicity, and we will also consider how to set suitable parameter α for each datum in future. The IVD in the paper are generally uniformly distributed. In the future, we will do some research to address unbalanced or missing interval-valued data.

REFERENCES

- [1] L. Billard and E. Diday, "Regression analysis for interval-valued data," in *Data Analysis, Classification, and Related Methods* (Studies in Classification, Data Analysis, and Knowledge Organization), H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, and M. Schader, Eds. Berlin, Germany: Springer, 2000.
- [2] H. H. Bock and E. Diday, "Analysis of symbolic data, exploratory methods for extracting statistical information from complex data," *J. Classification*, vol. 18, no. 2, pp. 291–294, 2000.
- [3] A. Chouakria, P. Cazes, and E. Diday, "Symbolic principal component analysis," in *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, H.-H. Bock and E. Diday, Eds. Berlin, Germany: Springer, 2000, pp. 200–212.
- [4] L. Billard and E. Diday, "Symbolic regression analysis," *Classification, Clustering, Data Anal. Stud. Classification, Data Anal., Knowl. Org.*, vol. 37, no. 24, pp. 6317–6328, 2002.
- [5] G. Cabanes, Y. Bennani, R. Destenay, and A. Hardy, "A new topological clustering algorithm for interval data," *Pattern Recognit.*, vol. 46, no. 11, pp. 3030–3039, Nov. 2013.
- [6] F. A. T. De Carvalho, "Fuzzy clustering algorithms for symbolic interval data based on adaptive and non-adaptive Euclidean distances," in *Proc. 9th Brazilian Symp. Neural Netw. (SBRN)*, Oct. 2006, pp. 60–65.
- [7] E. D. A. Lima Neto and F. D. A. De Carvalho, "Centre and range method for fitting a linear regression model to symbolic interval data," *Comput. Statist. Data Anal.*, vol. 52, no. 3, pp. 1500–1515, Jan. 2008.
- [8] M. A. Domingues, R. M. De Souza, and F. J. A. Cysneiros, "A robust method for linear regression of symbolic interval data," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 1991–1996, Oct. 2010.
- [9] Z. Lv, H. Jin, P. Yuan, and D. Zou, "A fuzzy clustering algorithm for interval-valued data based on gauss distribution functions," *Acta Electronica Sinica*, vol. 38, no. 2, pp. 295–300, 2010.
- [10] P. Hao and J. Guo, "Constrained center and range joint model for interval-valued symbolic data regression," *Comput. Statist. Data Anal.*, vol. 116, pp. 106–138, Dec. 2017.
- [11] L. Billard and J. Le-Rademacher, "Principal component analysis for interval data," *Wiley Interdiscipl. Rev. Comput. Statist.*, vol. 4, no. 6, pp. 535–540, Nov. 2012.
- [12] H. Wang, Y. Li, and R. Guan, "A comparison study of two methods for principal component analysis of interval data," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 24, no. 4, pp. 86–89, 2011.
- [13] H. Wang, R. Guan, and J. Wu, "CIPCA: Complete-information-based principal component analysis for interval-valued data," *Neurocomputing*, vol. 86, pp. 158–169, Jun. 2012.
- [14] C.-C. He and J.-T. Jeng, "Feature selection of weather data with interval principal component analysis," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jul. 2016, pp. 1–4.
- [15] C. Angulo, D. Anguita, L. Gonzalez-Abril, and J. Ortega, "Support vector machines for interval discriminant analysis," *Neurocomputing*, vol. 71, nos. 7–9, pp. 1220–1229, Mar. 2008.
- [16] A. Appice, C. D'amato, F. Esposito, and D. Malerba, "Classification of symbolic objects: A lazy learning approach," *Intell. Data Anal.*, vol. 10, no. 4, pp. 301–324, Jun. 2006.
- [17] A. P. D. Silva and P. Brito, "Linear discriminant analysis for interval data," *Comput. Statist.*, vol. 21, no. 2, pp. 289–308, Jun. 2006.
- [18] A. P. D. Silva and P. Brito, "Discriminant analysis of interval data: An assessment of parametric and distance-based approaches," *J. Classification*, vol. 32, no. 3, pp. 516–541, Oct. 2015.
- [19] H. Ishibuchi, H. Tanaka, and N. Iwamoto, "Discriminant analysis of multi-dimensional interval data and its application to smell sensing," *Int. J. General Syst.*, vol. 16, no. 4, pp. 311–329, 2009.
- [20] G. Jahanshahloo, F. H. Lotfi, F. R. Balf, and H. Z. Rezaei, "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.
- [21] N. C. Lauro, R. Verde, and F. Palumbo, "Factorial discriminant analysis on symbolic objects," in *Symbolic Data Analysis and the SODAS Software*, H.-H. Bock and E. Diday Eds. Heidelberg, Germany: Springer, 2000, pp. 212–233.
- [22] N. C. Lauro, R. Verde, and A. Iripino, "Factorial discriminant analysis," in *Symbolic Data Analysis and the SODAS Software*, E. Diday and M. Noirhomme-Fraiture, Eds. Chichester, U.K.: Wiley, 2008, pp. 341–358.
- [23] 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.
- [24] D. S. Guru and N. V. Kumar, "Class specific feature selection for interval valued data through interval K-means clustering," in *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Singapore: Springer, 2016, pp. 228–239.
- [25] L. Hedjazi, J. Aguilar-Martin, and M. V. L. Lann, "Similarity-margin based feature selection for symbolic interval data," *Pattern Recognit. Lett.*, vol. 32, no. 4, pp. 578–585, 2011.
- [26] C.-C. Hsiao, C.-C. Chuang, and S.-F. Su, "Robust Gaussian kernel based approach for feature selection," in *Advanced Intelligent Systems*. Cham, Switzerland: Springer, 2014, pp. 25–33.
- [27] D. S. Guru and N. V. Kumar, "Novel feature ranking criteria for interval valued feature selection," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 149–155.
- [28] C. Guo and Y. Liu, "A feature selection method for symbolic interval data," *Oper. Res. Manage. Sci.*, vol. 24, no. 1, pp. 67–74, 2015.
- [29] *The Weather of 243 Countries in the World*. Accessed: May 2016. [Online]. Available: <http://rp5.ru/>
- [30] *HEDJAZI Lyamine*. Accessed: Jun. 2016. [Online]. Available: <http://lhedjazi.jimdo.com/useful-links>



XIAOBO QI received the B.S. degree from the School of Computer and Information Technology, Shanxi University, in 2015, where she is currently pursuing the Ph.D. degree with the School of Computer and Information Technology. Her research interests include data mining and machine learning.



HUSHENG GUO received the Ph.D. degree from the School of Computer and Information Technology, Shanxi University, in 2014, where he is currently an Associate Professor with the School of Computer and Information Technology. His research interests include support vector machine, kernel methods, and machine learning. He is a member of the CCF.



ZADOROZHNYI ARTEM received the M.S. degree from the Automation and Information Technology Department, Samara State Technical University, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Shanxi University. His research interests include pattern recognition and data mining.



WENJIAN WANG received the Ph.D. degree from Xi'an Jiaotong University, in 2004. She is currently a full-time Professor and a Ph.D. supervisor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education, Shanxi University. She has published more than 170 academic papers. Her current research interests include machine learning, data mining, and intelligent computing.

...