A novel multi-source information fusion method based on dependency interval

Weihua Xu, Yufei Lin, Na Wang

***Abstract*—With the rapid development of big data era, it is necessary to extract necessary information from a large amount of information. Single-source information systems are often affected by extreme values and outliers, so multi-source information systems are more common and data more reasonable, information fusion is a common method to deal with multi-source information system. Compared with single-valued data, interval-valued data can describe the uncertainty and random change of data more effectively. This article proposes a novel interval-valued multi-source information fusion method: A multi-source information fusion method based on dependency interval. This method needs to construct a dependency function, which takes into account the interval length and the number of data points in the interval, so as to make the obtained data more centralized and eliminate the influence of outliers and extreme values. Due to the unfixed boundary of the dependency interval, a median point within the interval is selected as a bridge to simplify the acquisition of the dependency interval. Furthermore, a multi-source information system fusion algorithm based on dependency intervals was proposed, and experiments were conducted on 9 UCI datasets to compare the classification accuracy and quality of the proposed algorithm with traditional information fusion methods. The experimental results show that this method is more effective than the maximum interval method, quartile interval method, and mean interval method, and the validity of the data has been proven through hypothesis testing.**

***Index Terms*—Interval-valued, Information fusion, Multi-source information system, Dependency interval**

I. Introduction

RECENTLY, information fusion, as an important step of data processing, has been extensively applied to data mining, intelligent computing, and machine learning. According to clear standards, multi-source information fusion method fully utilizes multiple information sources and combines superfluous and reciprocal information which comes from multiple information sources in space or time, and then a consistent interpretation or description of the object under test is obtained. Compared with the system composed of each subset of information system, this method engenders the information system produce better effect Multi-source information fusion in the processing of big data has a wealth of attainments. Multi-source information often has a huge scale, which contains data often is the data of different information sources, or data of different sensors, in the face of so much data, it is of great significance to improve the efficiency of data mining. Multi-source information fusion and proposed in the 1970s, the rest of the application of sensors and other fields, also known as multi-sensor information fusion. In the multi-source information system, the information forms are various, the amount of information is huge, at the same time hope the information processing is prompt, the human brain has already been unable to undertake this heavy duty, therefore, the multi-source sensor system develops rapidly. Now it has been widely used in the fields of rail transit, electric power system, military affairs, ship positioning, aerospace and so on.

In real life, compared with deterministic phenomena, the phenomena people encounter are mainly uncertainty phenomena. At the same time, there are more and more researches on uncertainty. Rough set theory, as an effective tool for dealing with imprecise, inconsistent, incomplete and other incomplete information, has been widely used in the study of uncertainty phenomena. For instance, Xu et al. [1] constructed a two-way learning system model and mechanism based on information granularity in fuzzy data-sets. In reference [9], a new two-way concept-cognitive learning approach to dynamic concept learning in fuzzy context is proposed, which takes less time to learn granule concepts from a given clue and is better at dynamic concept learning. In order to overcome the incompleteness and complexity of concept learning, Xu et al. [56] proposed a new cognitive mechanism, which introduced the movement three-way decision method into two-way cognitive learning, and studied the concept evolution mechanism from the perspective of concept movement. According to the consistency of decision tables in rough set theory, Qian et al. [2] put forward three methods to evaluate the decision performance of decision-rule sets. Huang et al. [5] proposed an extended rough set model for multi-source composite rough sets, which integrates different types of attributes and fuses multiple composite relationships from different information sources. In addition, an incremental algorithm was proposed to update the composite rough approximation. Rough set theory (RST) [21]–[23] is the main tool of information fusion. RST has been proved to be an successful information fusion approach. In the last few decades, a lot of researches have been done. W. H. Xu et al. [40] used internal and external confidence to define the importance of information sources for information fusion. B. Z. Sun et al. [47] established an unused fuzzy rough set method for multi-attribute group decision making. Sang et al. [44] established a multi-grain model for three kinds of decision processes. In the incomplete information system, Zhang et al. [12] proposed a method of information fusion based on information entropy, and proposed four kinds of incremental fusion mechanism characterized by the change of information sources and attributes. Tang et al.

This paper is supported by the National Natural Science Foundation of China (Nos. 61976245).(*Corresponding author: Weihua Xu.*)

W. Xu is with the College of Artificial Intelligence, Southwest University, Chongqing, 400715, P.R. China (E-mail: chxuwh@gmail.com).

Y. Lin is with the College of Artificial Intelligence, Southwest University, Chongqing, 400715, P.R. China (E-mail: yufeilin03@163.com).

[53], inspired by Dempster-Shafer evidence theory, proposed a method of measuring the uncertainty in the negation evidence.

In most cases, single-valued data is not representative and may even be distorted when extreme values occur. Compared with single-valued data, interval-valued data can more effectively describe uncertainty and random changes. In economics, it is often necessary to keep the economy operating within a reasonable range. In statistics, the estimated interval of population parameters constructed from sample statistics is the confidence interval. For data, interval-valued estimation overcomes the shortcoming that point estimation cannot give an accurate interval of the overall parameters. At the same time, it can judge the value range of the estimated value at a certain probability level, so as to understand the aggregation and statistical dispersion of the sample sequence. Interval data has been a research hotspot in recent years. In reference [7], an interval-valued variable regression model was proposed and applied to predict the duration of unemployment. A new classification method was proposed in reference [8], which utilizes rough set methods to classify interval valued data and discover classification rules. D. S. Guru et al. [10] proposed a new similarity measurement method to measure the similarity between two interval-valued data, and applied this method to clustering. Zhang et al. [12] proposed the concept of interval-valued granularity rules in interval-valued decision systems and studied how to extract compact decision rules with specified confidence levels from them. Tien Thanh Nguyen proposed that the information granularity should be processed by constructing the information granularity and using the functions including the length of the strength interval and the end points, so that the reasonable granularity interval can be obtained and converted into the numerical membership degree, and the classifier has a higher prediction accuracy [35].

In the past, single-source information systems were usually used, but single-source information systems are often affected by extreme and abnormal values, which lead to inaccurate data, so multi-source information systems are more common, data is more realistic and reasonable. At the same time, due to the limitations of traditional interval generation methods, such as maximum interval method and minimum interval method can not make the data set centralized, can not rule out the influence of extreme values. Quartile interval method can exclude the influence of outliers, but can not guarantee the centralization of sample data. Therefore, in most cases, single-valued data are not representative and even distorted when extreme values occur. With the development of the research on uncertainty, interval-valued data can describe the uncertainty and random change more effectively than single data. However, the traditional information fusion methods, such as the average method, the maximum method and the minimum method, have great problems and can not include the uncertainty phenomenon. We need an information fusion method that can express uncertain knowledge explicitly. The goal of interval-valued information system is to extract more accurate and uncertain knowledge from a complete information table. Information fusion is an effective method to deal with existing multi-source data tables. It transforms and fuses information from multiple different data sources for information integration. Therefore, in order to describe the uncertainty of information system more accurately and deal with the problems caused by multi-source information, this paper combines multi-source information system with information fusion method based on dependency interval. Therefore, it is necessary to propose an information fusion method based on interval-valued for multi-source information systems. So we propose a new method of multi-source information fusion based on dependency interval, which can get the fusion interval by establishing correlation function. It is a more general, more extensive and more flexible method of uncertainty analysis.

The innovation points of this essay are as follows.

1) As far as we know, this is the first method to model the uncertainty of multi-source information system based on fusion interval.

2) The article defines a novel information fusion method based on dependency interval by dependency function and adjusts the parameters of function to adjust the fusion interval length. The uncertainty of underlying knowledge can be presented more pliable and precisely by interval values.

3) The proposed multi-source information fusion algorithm based on dependency interval is more accurate than the three classical information fusion algorithms mentioned in this paper.

In this paper, a novel multi-source information fusion method based on dependency interval is advanced. First of all, we collect multiple multi-source information tables and represent the data of the same object in the form of a set. Second, the corresponding functional form is constructed and the dependency interval is calculated. Thirdly, For each object under each attribute, the dependency interval is obtained by the same calculation method. Then, compared with other existing fusion methods, our experiments have proved that the multi-source information fusion method based on dependency interval is more productive. This article consists of six parts, as follows. In the second part, we give some essential notions such as tolerance and RST. In the third part, we put forward the concrete information fusion based on dependency interval and set up the corresponding algorithm. In part four, the experimental results are analyzed, in contrast with other fusion rules. At last, the conclusion of this paper is put forward in the fifth part. That is, the process of the article is shown in Fig.1.

II. PRELIMINARIES

In this section, some basic concepts are reviewed, such as RST, multi-source interval-valued decision system, and dependency interval.

A. Rough Sets Theory

An information system can be expressed as $IS = (U, AT, V_{AT}, f_{AT})$, where U represents a set of objects, AT is the finite set of conditional attributes, V_{AT} is the domain of conditional attributes, and $f_{AT} : U \quad AT \; ! \; V_{AT}$ is an information function. Specifically, decision information system can be represented as $DS = IS[fDT, V_{DT}, f_{DT}g$, where IS is an information system, DT is the set of decision attributes,

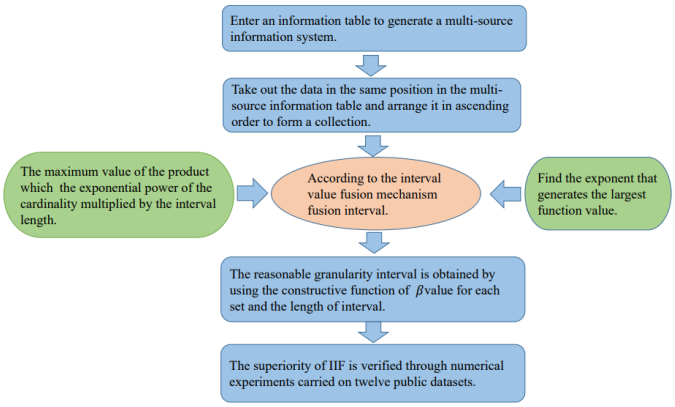

Fig. 1: Block diagram of steps of the proposed approach.

V_{DT} is the domain of DT, and $f_{DT} : U \quad DT \; ! \; V_{DT}$ is an information function. For any $B \quad AT$, an equivalence relation R_B is defined by

$$R_B = f(x,y) \; 2 \; U \quad U \, j f(x,a) = f(y,a), 8a \; 2 \; B g, \quad (1)$$

where $f(x,a)$ and $f(y,a)$ are the values of $f(x,y)$ when x and y are equal to a, respectively. For any $X \quad U$, the lower and upper approximations of X are defined by

$$\underline{R_B}(X) = \quad x \; 2 \; U \; [x]_{R_B} \quad X \quad , \quad (2)$$

$$\overline{R_B}(X) = \quad x \; 2 \; U \; [x]_{R_B} \setminus X \; 6 = ; \quad , \quad (3)$$

where $[x]_{R_B} = f y j (x,y) \; 2 \; R_B g$.

Let $U/D = f Y_1, Y_2, ..., Y_m g$ be the decision partition of U based on DT. For $DS = (U, AT \; [\; DT, V, f)$, the approximation classifier precision(AP) and approximation classifier quality(AQ) of U/D with respect to R_B are defined as following:

$$AP_{R_B}(U/D) = \frac{\sum_{i=1}^{m} \underline{R_B}(Y_i)}{\sum_{i=1}^{m} \overline{R_B}(Y_i)}, \quad (4)$$

$$AQ_{R_B}(U/D) = \frac{\sum_{i=1}^{m} \underline{R_B}(Y_i)}{jUj}. \quad (5)$$

AP and AQ, which were first proposed by Pawlak [39], can be used as indicators to measure the quality of classification. The values of AP and AQ are related to the accuracy and quality of classification. The higher the values of AP and AQ, the higher the precision and quality of approximate classification.

B. Multi-source Interval-valued Decision Information System

Interval-valued data can reflect the uncertainty of data set more accurately than single value data, and interval data can be used to describe random change and imprecise information more effectively, the interval-valued data is defined as follows:

Let $IIS = fU, AT, V, fg$ be an interval-valued information system, where U is a nonempty finite set of objects; AT is a nonempty finite set of condition attributes; $V = \bigcup_{a2AT} V_a$ and V_a is a domain of attribute a; $f : U \quad AT \; ! \; V$ is an information function such that $f(x,a) = f^L(x,a), f^U(x,a) \; 2 \; V_a$ for every $a \; 2 \; AT, x \; 2 \; U$, where $f^L(x,a)$ and $f^U(x,a)$ denote the left and right endpoints of interval, respectively.

In many practical applications, interval data can be obtained from receivers with the same function but in different locations. For example, to measure wind speed and humidity in a city. A multi-source interval-valued information system combines different sources of interval information in the form of IIS, which is called $MIIS$. The formula is as follows:

Given $MIIS = fIIS_i \, j \, IIS_i = (U, AT_i, V_{AT_i}, f_i), i = 1, 2, \ldots, Ng$ be a multi-source interval-valued information system , where IIS_i is the i-th IIS in $MIIS$; U is a nonempty finite set of objects; AT_i is a nonempty finite set of attributes of the i-th IIS_i; V_{AT_i} is the domain of the attribute set AT_i in the i-th subsystem IIS_i, $f_i : U \quad AT_i \; ! \; V_i$ is an information function in the i-th subsystem IIS_i.

A multi-source interval-valued decision system can be represented as $MIDS = fMIIS \; [\; MDg$, where $MIIS$ is a multi-source interval-valued information system and $MD = fD, V_D, f_D g$, where D is the decision attribute set, V_D is the domain of D and $f_D : U \quad D \; ! \; V_D$ is an information function.

C. Dependency Interval

In this script, we get a dependency interval by calculating the data set W with the dependency function, the interval is in the form of $= [u, v]$, where u is the lower bound and v is the upper bound of the interval. The range of this article needs to satisfy the following limitations.

1) Interval Reliability: The data in the data-set should be contained as much as possible by the interval . This interval, which contains more real data, is more reasonable than other generation intervals.

2) Interval Accuracy: This requirement implies a high degree of accuracy of the interval. For example, if the number of data falling into two intervals is the same, interval $= [4, 6]$ is more accurate than $= [3, 7]$, the smaller the interval length, the more accurate.

3) The value of β: For different ranges of values for different attributes, we take different values of β, for example, for attributes with a range of 100, we take a value of β of 1, and for attributes with a range of 10, we take a value of β as 0.1.

A justifiable dependency interval needs to satisfy the above three guidelines, the above three guidelines only for this thesis. Actually, the method of formalizing interval is more demand-oriented, and different rules constitute different intervals. For the above requirements we give a detailed explanation. We use the number of data points that fall into the constructed interval as a measure of interval reliability. In this case, the data set distribution in this paper is a discrete variable,the experimental evidence can be determined by the cardinality of elements in W (denoted by $N f Dg$) falling within the bounds of . We

use $\int_{R} p(x)dx$ as a metric. For the second requirement, the shorter the interval length is, the more accurate the interval is when the reliability is the same.

III. Information fusion based on dependency interval

With the advent of the information age, people want to transform from many information tables to more flexible information systems, which can cover more uncertainty, in order to simplify the data and extract important information. The aim of information fusion is to deal with the data collected by multi-sensors synthetically. What we propose the information fusion method based on dependency interval can be a good solution to the difficulties in data processing. It shows how the information fusion method based on the dependent interval fuses the multi-source information system into the implementation process of the interval-valued information system in Fig.2. In the first step, multi-source information system data is obtained through a single information system, and data points at the same position of each information table are aggregated into a collective form. Step 2, order the data in the collection at each location in ascending order (or descending order, as we have done in this article) , the set is divided into two parts by the median of the set, i. e. , the set less than the median and the set greater than the median, which are called the left set and the right set. In the third step, the dependency functions corresponding to each data point in the left and right collections are calculated respectively. In the fourth step, the data points corresponding to the maximum dependency function in the left and right sets are calculated, and the data points are taken as the lower and upper bounds of the dependency interval. Finally, the interval value information system is composed of the dependent interval.

The interval, which generates by $IFDI$, need be limited. After the request of the granularity interval, the construction function obtains the dependency interval, and finally tests the precision of the interval. The steps are as follows.

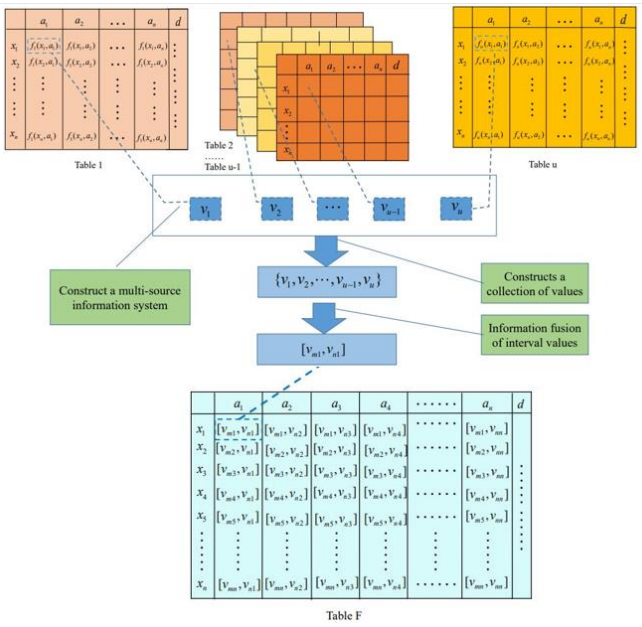

Fig. 2: The information fusion method

A. Dependency interval obtained

According to the requirement of justifiable dependency interval, we construct dependency function. For requirement one, the more data-sets fall into data points within an interval, the higher the reliability of the interval. As to rule two, the smaller the interval length is, the more accurate the dependency interval is, so consider using non-increasing function. For standard three, different β values have different effects on the generation interval, $f(u) = \exp(-\gamma u)(\gamma > 0)$, in which $u = |a - b|$ is the length of the interval , we should choose optimal β. According to the above rules, the function of dependency interval can be set as follows.

Definition 3.1. The definition form of the dependency function is as follows:

$$L(\cdot) = (NfWg) \quad exp(-\delta l), \tag{6}$$

where W is the data set, $NfWg$ is the number of data points in the dependency interval, the coefficient $\delta > 0$, the default value for this article is 1 for δ, and l is the interval length.

Combined with Fig.1, we illustrate the method of multi-source information fusion based on dependency interval. The specific calculation process is as follows: first of all, the attribute values of the same position in the multi-source information system are arranged into a set, which can be arranged in ascending or descending orderthis article is arranged in ascending order by defaultand then we find the median $med(W)$. The boundary of the interval is two unknown parts, and the median of the data set is fixed, and it must be in the generating interval, besides, it is not affected by the extreme value and outliers. Secondly, different β values may affect the precision of the dependency interval. For a data set of 10 or 1000, using the same β will cause the precision of the dependency interval to decrease, so we need to introduce β into the dependency function, and then take the magnitude into account, in order to be an optimal dependency interval. The β value is determined according to the data level. By calculation, we find that the bigger the data value is, the bigger the β value is. Finally, because the median divides the interval into two independent parts, you can use functions to discuss the upper and lower bounds independently, finding the upper and lower bounds based on the following two functions.

Definition 3.2. Calculate dependency interval based on dependency function. Take the numerical point on the left side of the median that satisfies the maximum $L(\cdot_1)$ of the dependency function as the lower boundary point u_{opt} of the dependency interval, and take the numerical point on the right side of the median that satisfies the maximum of the dependency function $L(\cdot_2)$ as the upper boundary point v_{opt} of the dependency interval. The following calculation formula is given:

$$u_{opt} = \arg \max_{u \; med(W);u2W} L(\cdot_1), \tag{7}$$

$$v_{opt} = \arg \max_{med(W) \; v;v2W} L(\cdot_2), \tag{8}$$

where

$$L(\ _1) = (N\ f x_i\ 2\ \mathbf{W}\ j\ u \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ j\text{med}(\mathbf{W}) \quad uj), \tag{9}$$

$$L(\ _2) = (N\ f x_i\ 2\ \mathbf{W}\ j\ v \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ jv \quad \text{med}(\mathbf{W}))j). \tag{10}$$

The final form of the build interval is as follows:

$$\begin{aligned}
W_{a_1}(x_1) &= I^1_{a_1}(x_1), I^2_{a_1}(x_1), \ldots, I^n_{a_1}(x_1) \\
&\quad ! \quad I^i_{a_1}(x_1), I^j_{a_1}(x_1) \ , \\
\mathrm{W}_{a_1}(x_2) &= I^1_{a_1}(x_2), I^2_{a_1}(x_2), \ldots, I^n_{a_1}(x_2) \\
&\quad ! \quad I^i_{a_1}(x_2), I^j_{a_1}(x_2) \ , \\
\mathrm{W}_{a_1}(x_o) &= I^1_{a_1}(x_o), I^2_{a_1}(x_o), \ldots, I^n_{a_1}(x_o) \\
&\quad ! \quad I^i_{a_1}(x_o), I^j_{a_1}(x_o) \ , \\
\mathrm{W}_{a_2}(x_1) &= I^1_{a_2}(x_1), I^1_{a_2}(x_1), \ldots, I^n_{a_2}(x_1) \\
&\quad ! \quad I^i_{a_2}(x_1), I^j_{a_2}(x_1) \ , \\
\mathrm{W}_{a_2}(x_2) &= I^1_{a_2}(x_2), I^2_{a_2}(x_2), \ldots, I^n_{a_2}(x_2) \\
&\quad ! \quad I^i_{a_2}(x_2), I^j_{a_2}(x_2) \ , \\
\mathrm{W}_{a_2}(x_o) &= I^1_{a_2}(x_o), I^2_{a_2}(x_o), \ldots, I^n_{a_2}(x_o) \\
&\quad ! \quad I^i_{a_2}(x_o), I^j_{a_2}(x_o) \ , \\
\mathrm{W}_{a_m}(x_1) &= I^1_{a_m}(x_1), I^2_{a_m}(x_1), \ldots, I^n_{a_m}(x_1) \\
&\quad ! \quad I^i_{a_m}(x_1), I^j_{a_m}(x_1) \ , \\
\mathrm{W}_{a_m}(x_2) &= I^1_{a_m}(x_2), I^2_{a_m}(x_2), \ldots, I^n_{a_m}(x_2) \\
&\quad ! \quad I^i_{a_m}(x_2), I^j_{a_m}(x_2) \ , \\
\mathrm{W}_{a_m}(x_o) &= I^1_{a_m}(x_o), I^2_{a_m}(x_o), \ldots, I^n_{a_m}(x_o) \\
&\quad ! \quad I^i_{a_m}(x_o), I^j_{a_m}(x_o) \ .
\end{aligned} \tag{11}$$

Example 3.1.1. To illustrate the multi-source information fusion method based on dependency interval more clearly, we give the example 3.1.1 as following. Tables I-VII are the source of information system with decision, which denote the results of graduate admissions with three attributes. The attributes a_1-a_3 represent gre, gpa, $rank$ and the d is the decision that means whether the student is admitted.

Let the $V_D = f admitted\ student, unaccepted\ student g$, $U/D = fY_1, Y_2 g$, and by preliminary calculation we get $\beta = [2, 0.5, 0.5]$, where $Y_1 = fx_1, x_2, x_3, x_4, x_5, x_6, x_7 g$, $Y_2 = fx_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15} g$. The value of β is $f2, 1, 0.5g$ respectively.

For the attribute a_2 and the object x_{15}, we have the data-set $W = f1.053030, 3.920000, 3.920000, 4.797893, 4.802906g$. The median of the data-set W is 3.920000. For the lower bound a. When $u = 1.053030$ the function of

$$L(\ _1) = (N\ f x_i\ 2\ \mathbf{W}\ j\ u \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ j\text{med}(\mathbf{W}) \quad uj)$$

$$= 3 \quad exp(\ j1.053030 \quad 3.920000j) = 0.170613,$$

when $u = 3.920000$ the function of

$$L(\ _1) = (N\ f x_i\ 2\ \mathbf{W}\ j\ u \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ j\text{med}(\mathbf{W}) \quad uj)$$

$$= 2 \quad exp(\ j3.920000 \quad 3.920000j) = 2.$$

The maximum of $L(\ _1)$ determines the lower boundary of the interval,

$$u_{opt} = \arg \max_{u \quad med(\mathbf{W}); u2\ \mathbf{W}} L(\ _1) = 3.920000,$$

when $v = 4.797893$ the function of

$$L(\ _2) = (N\ f x_i\ 2\ \mathbf{W}\ j\ v \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ jv \quad \text{med}(\mathbf{W}))j)$$

$$= 2 \quad exp(\ j4.797893 \quad 3.920000j) = 0.831316,$$

when $v = 4.802906$ the function of

$$L(\ _2) = (N\ f x_i\ 2\ \mathbf{W}\ j\ v \quad x_i \quad \text{med}(\mathbf{W}))\ exp(\ jv \quad \text{med}(\mathbf{W}))j)$$

$$= 3 \quad exp(\ j4.802906 \quad 3.920000j) = 1.240738.$$

The best upper boundary is in the same way. Where

$$v_{opt} = \arg \max_{med(\mathbf{W}) \quad v; v2\ \mathbf{W}} L(\ _2) = 1.240738.$$

Therefore, the dependency interval is $[3.920000, 4.802906]$.

After preliminary processing of the multi-source information system, we get data-sets of the product of the number of objects and attributes. For example, the data set of the first attribute $fa_1 g$ and the first uniform $fx_1 g$ is :

$$f380.766, 378.053, 379.856, 377.704, 379.953,$$
$$379.735, 382.205, 379.096, 380.976, 380.000,$$
$$380.000, 380.088, 380.000, 379.622, 380.694,$$
$$379.931, 380.503, 380.000, 380.000, 381.347g.$$

Sort them from small to large,the set is as follows:

$$f377.704, 378.053, 379.096, 379.622, 379.735,$$
$$379.856, 379.931, 379.953, 380.000, 380.000,$$
$$380.000, 380.000, 380.000, 380.088, 380.503,$$
$$380.694, 380.766, 380.976, 381.347, 382.205g.$$

The median of the data set is: 380.000.

According to the multi-source information fusion method based on dependency interval, we can get the dependency interval when the value of β is 1, 10, 100, respectively :

The dependency interval is: $[379.62164, 380.76581]$.

The dependency interval is: $[377.70351, 382.20456]$.

The dependency interval is: $[377.70351, 382.20456]$.

In most cases, we can see that for a uniform data set, the dependency interval results are different for different values of β.Sometimes, for different values of β, the dependency interval results are same.

TABLE I: The source of information system.

U	a_1	a_2	a_3	d
x_1	380	3.61	3	0
x_2	700	3.08	2	0
x_3	520	2.93	4	0
x_4	400	3.08	2	0
x_5	800	4	4	0
x_6	440	3.22	1	0
x_7	540	3.39	3	0
x_8	560	2.98	1	1
x_9	640	3.19	4	1
x_{10}	760	3	2	1
x_{11}	660	3.67	3	1
x_{12}	760	4	1	1
x_{13}	800	4	1	1
x_{14}	700	4	1	1
x_{15}	700	3.92	2	1

TABLE II: The first source of information.

U	a_1	a_2	a_3	d
x_1	380.882906	4.492906	3.882906	0
x_2	700.882906	3.962906	2.882906	0
x_3	521.091283	4.021283	5.091283	0
x_4	401.091283	4.171283	3.091283	0
x_5	801.091283	5.091283	5.091283	0
x_6	440.882906	4.102906	1.882906	0
x_7	541.091283	4.481283	4.091283	0
x_8	561.091283	4.071283	2.091283	1
x_9	640.000000	3.190000	4.000000	1
x_{10}	760.000000	3.000000	2.000000	1
x_{11}	661.091283	4.761283	4.091283	1
x_{12}	760.000000	4.000000	1.000000	1
x_{13}	800.882906	4.882906	1.882906	1
x_{14}	700.882906	4.882906	1.882906	1
x_{15}	700.882906	4.882906	2.882906	1

TABLE III: The second source of information.

U	a_1	a_2	a_3	d
x_1	378.053030	1.663030	1.053030	0
x_2	700.032866	3.112866	2.032866	0
x_3	520.032866	2.962866	4.032866	0
x_4	398.053030	1.133030	0.053030	0
x_5	798.053030	2.053030	2.053030	0
x_6	438.053030	1.273030	-0.946970	0
x_7	540.032866	1.273030	3.032866	0
x_8	560.032866	3.012866	1.032866	1
x_9	640.032866	3.222866	4.032866	1
x_{10}	758.053030	1.053030	0.053030	1
x_{11}	660.000000	3.670000	3.000000	1
x_{12}	758.053030	2.053030	-0.946970	1
x_{13}	800.032866	4.032866	1.032866	1
x_{14}	700.000000	4.000000	1.000000	1
x_{15}	700.000000	3.920000	2.000000	1

TABLE IV: The third source of information.

U	a_1	a_2	a_3	d
x_1	380.428022	4.038022	3.428022	0
x_2	698.613650	1.693650	0.613650	0
x_3	518.613650	1.543650	2.613650	0
x_4	400.000000	3.080000	2.000000	0
x_5	798.613650	2.613650	2.613650	0
x_6	440.428022	3.648022	1.428022	0
x_7	538.613650	2.003650	1.613650	0
x_8	560.000000	2.980000	1.000000	1
x_9	640.428022	3.618022	4.428022	1
x_{10}	760.428022	3.428022	2.428022	1
x_{11}	660.428022	4.098022	3.428022	1
x_{12}	758.613650	2.613650	-0.386350	1
x_{13}	798.613650	2.613650	-0.386350	1
x_{14}	700.428022	4.428022	1.428022	1
x_{15}	700.000000	3.920000	2.000000	1

TABLE V: The forth source of information.

U	a_1	a_2	a_3	d
x_1	377.703508	1.313508	0.703508	0
x_2	700.877893	3.957893	2.877893	0
x_3	520.000000	2.930000	4.000000	0
x_4	397.703508	0.783508	-0.296492	0
x_5	797.703508	1.703508	1.703508	0
x_6	440.755786	4.097893	1.877893	0
x_7	537.703508	1.093508	0.703508	0
x_8	560.877893	3.857893	1.877893	1
x_9	640.877893	4.067893	4.877893	1
x_{10}	757.703508	0.703508	-0.296492	1
x_{11}	660.000000	3.670000	3.000000	1
x_{12}	760.877893	4.877893	1.877893	1
x_{13}	797.703508	1.703508	-1.296492	1
x_{14}	700.000000	4.000000	1.000000	1
x_{15}	700.877893	4.797893	2.877893	1

TABLE VI: The fifth source of information.

U	a_1	a_2	a_3	d
x_1	380.476507	4.086507	2.953014	0
x_2	702.409834	5.489834	4.409834	0
x_3	522.409834	5.339834	6.409834	0
x_4	400.476507	3.556507	2.476507	0
x_5	800.476507	4.476507	4.476507	0
x_6	442.409834	5.629834	3.409834	0
x_7	540.000000	3.390000	3.000000	0
x_8	562.409834	5.389834	3.409834	1
x_9	640.476507	3.666507	4.476507	1
x_{10}	760.000000	3.000000	2.000000	1
x_{11}	662.409834	6.079834	5.409834	1
x_{12}	760.000000	4.000000	1.000000	1
x_{13}	800.476507	4.476507	1.476507	1
x_{14}	700.476507	4.476507	1.476507	1
x_{15}	702.409834	6.329834	4.409834	1

TABLE VII: The dependency interval of information system.

U	a_1	a_2	a_3	d
x_1	[377.70,380.88]	[1.66,4.09]	[1.05,3.48]	0
x_2	[700.03,700.88]	[3.11,3.96]	[2.03,2.882]	0
x_3	[520.00,521.09]	[2.93,4.02]	[4.00,5.091]	0
x_4	[397.70,401.09]	[1.13.56]	[0.05,2.48]	0
x_5	[797.70,801.09]	[2.05,4.48]	[2.05,4.48]	0
x_6	[440.43,440.88]	[3.65,4.10]	[1.43,1.88]	0
x_7	[538.61,540.03]	[2.00,3.42]	[1.61,3.03]	0
x_8	[560.00,561.09]	[2.98,4.07]	[1.00,2.09]	1
x_9	[640.00,640.88]	[3.19,3.67]	[4.00,4.48]	1
x_{10}	[757.70,760.43]	[1.05,3.43]	[0.05,2.43]	1
x_{11}	[660.00,661.09]	[3.67,4.76]	[3.00,4.09]	1
x_{12}	[758.05,760.88]	[2.61,4.88]	[-0.39,1.88]	1
x_{13}	[798.61,800.88]	[2.61,4.48]	[-0.39,1.48]	1
x_{14}	[700.00,700.88]	[4.00,4.48]	[1.00,1.48]	1
x_{15}	[700.00,700.88]	[3.92,4.80]	[2.00,2.88]	1

B. Information fusion method based on dependency interval

With the advent of the big data era, the amount of information has become more and more huge, like sensor data, many of the data from the same information from different sources. Therefore, information fusion is very significant, which can provide a comprehensive and systematic representation, and at the same time simplify the information extraction of important integrated information.

Definition 3.3. Given an $IIS = (U, AT, V_{AT}, f_{AT})$, where $U = \{x_1, x_2, ..., x_n\}$. $\forall x_i, x_j \in U$ and for any $a \in AT$, the distance of any two sample points, which are in U with respect to attribute a is settled as follow:

$$dis_a(x_i, x_j) = \sqrt{\left(f^L(x_i, a) - f^L(x_j, a)\right)^2 + \left(f^U(x_i, a) - f^U(x_j, a)\right)^2}, \quad (12)$$

where $[f^L(x_i, a), f^U(x_i, a)]$ is the value of attribute a and object x_i in the generated interval-valued information system.

Definition 3.4. Given an $IIS = (U, AT, V_{AT}, f_{AT})$, for any $a \in AT$, the tolerance relation T_a is defined as follow:

$$T_a = \left\{ (x_i, x_j) \mid \frac{dis_a(x_i, x_j)}{\max\limits_{y \in U}(dis_a(x_i, y))} \leq \alpha \right\}, \quad (13)$$

where α denotes the threshold. The tolerance class of x_i under a is denoted as $T_a(x_i) = \{x_j \mid (x_i, x_j) \in T_a\}$.

Example 3.1.2. (Continued from Example 3.1.1) According to the method of reference, multi-source information system which contains 20 information tables are constructed, the set of each object and the corresponding attribute values of each attribute are calculated, according to the function of dependency interval, to find the dependency interval, and a new information table of interval values is obtained. Take the attribute a_1 of the first source of information for example. Without loss of generality, let the value of α be 0.5. First, drawing lessons from the distance formula in the preparatory knowledge, we give the distance matrix of samples as follows.

$$Dis = \begin{pmatrix} 0 & 452.21 & 197.91 & \dots & 593.86 & 452.50 & 452.32 \\ 452.22 & 0 & 254.31 & \dots & 141.64 & 0.33 & 0.15 \\ 197.91 & 254.31 & 0 & \dots & 395.94 & 254.59 & 254.41 \\ 27.901 & 424.32 & 170.01 & \dots & 565.96 & 424.60 & 424.42 \\ 593.84 & 141.63 & 395.36 & \dots & 0.02 & 141.34 & 141.53 \\ 84.519 & 367.70 & 113.39 & \dots & 509.34 & 367.98 & 367.80 \\ 226.31 & 225.91 & 28.401 & \dots & 367.55 & 226.19 & 226.01 \\ 254.50 & 197.72 & 56.584 & \dots & 339.37 & 198.01 & 197.83 \\ 367.63 & 84.59 & 169.72 & \dots & 226.23 & 84.88 & 84.69 \\ 536.76 & 84.55 & 338.86 & \dots & 57.10 & 84.26 & 84.44 \\ 395.79 & 56.42 & 197.89 & \dots & 198.06 & 56.71 & 56.53 \\ 536.86 & 84.65 & 338.95 & \dots & 57.00 & 84.36 & 84.54 \\ 593.86 & 141.64 & 395.95 & \dots & 0 & 141.36 & 141.54 \\ 452.50 & 0.33 & 254.59 & \dots & 14.14 & 0 & 0.19 \\ 452.32 & 0.15 & 254.41 & \dots & 141.54 & 0.19 & 0 \end{pmatrix}$$

Divide each element of Dis by the maximum value of all the elements in the corresponding row, we can get the matrix Dis .

$$Dis = \begin{pmatrix} 0 & 0.7615 & 0.3333 & \dots & 1 & 0.7620 & 0.7617 \\ 1 & 0 & 0.5624 & \dots & 0.3132 & 0.7217 & 0.3263 \\ 0.4998 & 0.6423 & 0 & \dots & 1 & 0.6430 & 0.6425 \\ 0.0493 & 0.7497 & 0.3004 & \dots & 1 & 0.7502 & 0.7499 \\ 1 & 0.2385 & 0.6667 & \dots & 0.2944 & 0.2380 & 0.2383 \\ 0.1659 & 0.7219 & 0.2226 & \dots & 1 & 0.7225 & 0.7221 \\ 0.6157 & 0.6146 & 0.0773 & \dots & 1 & 0.6154 & 0.6149 \\ 0.7499 & 0.5826 & 0.1667 & \dots & 1 & 0.5835 & 0.5829 \\ 1 & 0.2301 & 0.4617 & \dots & 0.6154 & 0.2309 & 0.2304 \\ 1 & 0.1575 & 0.6313 & \dots & 0.1064 & 0.1570 & 0.1573 \\ 1 & 0.1426 & 0.5000 & \dots & 0.5004 & 0.1433 & 0.1428 \\ 1 & 1.5767 & 0.6314 & \dots & 0.1062 & 0.1571 & 0.1575 \\ 1 & 0.2385 & 0.6667 & \dots & 0 & 0.2380 & 0.2383 \\ 1 & 0.7213 & 0.5626 & \dots & 0.3124 & 0 & 0.4194 \\ 1 & 0.3263 & 0.5625 & \dots & 0.3129 & 0.4196 & 0 \end{pmatrix}$$

Then the tolerance classes of these ten samples are computed.

$$\begin{aligned}
&T^1_{a_1}(x_1) = \{x_1, x_3, x_4, x_6, x_7, x_8\},\\
&T^1_{a_1}(x_2) = \{x_2, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_3) = \{x_1, x_3, x_4, x_6, x_7, x_8, x_9, x_{11}\},\\
&T^1_{a_1}(x_4) = \{x_1, x_3, x_4, x_6, x_7, x_8\},\\
&T^1_{a_1}(x_5) = \{x_2, x_5, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_6) = \{x_1, x_3, x_4, x_6, x_7, x_8\},\\
&T^1_{a_1}(x_7) = \{x_3, x_6, x_7, x_8, x_9, x_{11}\},\\
&T^1_{a_1}(x_8) = \{x_3, x_7, x_8, x_9, x_{11}\},\\
&T^1_{a_1}(x_9) = \{x_2, x_3, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{10}) = \{x_2, x_5, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{11}) = \{x_2, x_5, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{12}) = \{x_2, x_5, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{13}) = \{x_2, x_5, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{14}) = \{x_2, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\},\\
&T^1_{a_1}(x_{15}) = \{x_2, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}.
\end{aligned}$$

IV. EXPERIMENTAL ANALYSIS

In this section, we use nine data sets to demonstrate the validity of interval fusion of the interval-valued fusion mechanism based on UCI data set, as shown in table IX. We selected the data which fit the condition to carry on the experiment. The steps for conducting the experiment are as follows: the details of the specific environment are shown below. The procedure for conducting the experiment is as follows: the first step is to obtain the multi-source information system according to the original information table. The second part is to fuse the interval-valued information according to the dependency function we constructed, the third stride is to get the fused interval-valued information table. And compared with other methods, the algorithm of this paper is as follows.

Here is the time complexity analysis of algorithm 1. Step 1 generates a multi-source information system, takes the data of two multi-source information systems at the same position and generates a set, the time complexity of this step is $O(n)$. Step 2 is to take out the median, the complexity is $O(n)$. In addition, steps 4 to 10 indicate that the median set is divided into the upper and lower parts, the dependence functions of the two parts are calculated respectively, and the maximum point is selected as the boundary point. The time complexity is $O(n)$, so the time complexity of algorithm 1 is $O(n^3)$.

In the following experiments, Dep is the method proposed in this paper, Max is the maximum interval fusion method, $Quar$ is the quartile interval fusion method, and $Mean$ is the mean interval fusion method. Then we compare the approximate accuracy and approximate quality of the fusion results of different methods to illustrate the superiority of the fusion method proposed in this paper, and use hypothesis testing to illustrate the validity.

TABLE VIII: Operating Ambient

Name	Model	Parameter
CPU	Intel(R)Core(TM) i5-1155G7	2.50GHz
Platform	Python	3.7
System	Windows11	64bit
Memory	DDR4	8GB;1600Mhz
Hard Disk	HTS545050A7E680	500GB

TABLE IX: The description of data sets.

No.	Data sets	Abbreviation	Samples	Attributes	Classes
1	Wine	Wine	178	13	3
2	Ecoli	Ecoli	336	8	8
3	Money	Money	1370	4	2
4	Booknote	Booknote	1300	4	2
5	Seeds	Seeds	140	7	2
6	Transfusion	Transfusion	740	4	2
7	Abalone	Abalone	4177	8	3
8	Wilt	Wilt	4889	6	2
9	Skin Segementation	Skin-St	15000	3	3

Algorithm 1: The fusion algorithm of multi-source information systems based method on dependency interval

Input: $MS = \{(U, AT_i, V_{AT_i}, f_i, W, V_W, f_W), i = 1, 2, ..., N\}$; the decision partition $U/W = \{Y_1, Y_2, ..., Y_m\}$;
Output: A new fusion table

```
for s = 1 : N do
    # N is the number of information sources.
    for each W = {f_i(x_j, u_h)|i = 1, 2..., N} find med(W); do
        for each b ∈ W, v ≠ med(W) do
            compute {O(v) = {N{x_i ∈ W | v ≠ x_i ≠ med(W)}
                f(|med(W) − v|)},
            v_opt = argmax_v (O(v)) ;
        end
        for each u ∈ W, u ≠ med(W) do
            compute {O(u) = {N{x_i ∈ W | u ≠ x_i ≠ med(W)}
                f(|med(W) − u|)},
            u_opt = argmax_u(O(u));
        end
    end
end
return [u_opt, v_opt]
```

A. The analysis of fusion effectiveness

From references [33], it can be seen that AP and AQ can reveal the accuracy and quality of approximate classification respectively. We can use them as a measure of fusion performance. In this paper, three commonly used fusion methods based on AP and AQ are compared with our proposed fusion method. The other three common fusion methods are as follows:

$$1) MaxF_a(x) = \left(\min_{i \in \{1,2,:::,N\}} f_i(x,a), \max_{i \in \{1,2,:::,N\}} f_i(x,a) \right),$$

$$2) QuarF_a(x) = \left(Q^L_{i \in \{1,2,:::,N\}} f_i(x,a), Q^U_{i \in \{1,2,:::,N\}} f_i(x,a) \right),$$

$$3) MeanF_a(x) = \left(\frac{1}{m} \sum_{i=1}^{m} f_i^L(x,a), \frac{1}{m} \sum_{i=1}^{m} f_i^U(x,a) \right).$$

Where $f_i^L(x,a)$ and $f_i^U(x,a)$ express the left and right endpoints in when the information source is with respect to attribute a.

We used four methods to fuse nine data sets with different threshold α, and the results were obtained separately. The above nine tables record AP and AQ of fusion results with the variation of different α and different β. The value of α is set from 0.05 to 0.5 with the step of 0.05, and we set the same

and different value of β. And the β of each table is different, but the interval is optimal.

According to the experimental results, we carry out the following analysis. The deductions are concluded in the $Elico$ data set, the α value was adjusted for analysis by adjusting the value of the parameter β to f0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. When the value of α is in f0.05, 0.10, 0.15, 0.2, 0.25, 0.30, 0.35, 0.40g, the $IFDI$ method strictly outperforms the other three fusion methods which are separately max interval fusion method, min interval fusion method and mean interval fusion method. However, when the threshold of α is 0.45 or 0.5, the values of AP and AQ calculates by four methods are all 0, the intervals calculated by the four methods produce the same effect. For the data set $Transfusion$, the α value was adjusted for analysis by adjusting the value of the parameter β to f0.5, 0.5, 2, 0.5g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. When the value of α is set from 0.15 to 0.35, $IFDI$ rule is totally better than the other three fusion methods, and when the value of α changes from 0.4 to 0.5, all methods have the same value, therefore, they have the same effect. For the $Money$ data set, the α value was adjusted for analysis by adjusting the value of the parameter β to f0.25, 0.25, 0.25, 0.25g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the αvalue was adjusted for analysis. When the value of α is varied from 0.05 to 0.3, the interval-valued fusion method performs better than others perfectly, the four fusion methods is 0, and when the value of α is given from 0.4 to 0.5, so their effect is identical.

In regard to the $Wine$ information system, the α value was adjusted for analysis by adjusting the value of the parameter β to f1, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 100g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. The four methods have the same value when the value of α is in f0.05, 0.10, 0.15, 0.20, 0.25, 0.3 g. Nevertheless, when the value of α is 0.35 and 0.5, the interval-valued fusion method is more preferable than other methods. In the information system of the $Seeds$, the α value was adjusted for analysis by adjusting the value of the parameter β to f1, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. The fusion performance of the $IFDI$ approach is greater than others, when the value of the threshold α is set from 0.35 to 0.5, when the threshold is set from 0.40 to 0.50, the value of α of the interval-valued information fusion method and other fusion methods is 0. In the $Booknote$ data set, the α value was adjusted for analysis by adjusting the value of the parameter β to f0.5, 0.5, 0.5, 0.5g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. The expression of the multi-source information fusion approach based on dependency interval is better than others when the value of the threshold is set from 0.05 to 0.15 , from 0.4 to 0.5 and 0.3, when the threshold changes from 0.45 to 0.5, the value of the optimal approach and other methods are 0.

As to $Abalone$ data, the α value was adjusted for analysis by adjusting the value of the parameter β to f1, 1, 0.5, 5, 1, 1, 1, 10g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. When the value of α is 0.05 or 0.1, the $IFDI$ procedure performs better than others obviously, but when α is 0.15, the Max method is optimal. When the value of α is form 0.2 to 0.5, the Opt method performs better than others strictly, and the value of α is from 0.35 to 0.5, the four fusion methods is 0. For the $Wilt$ data set, the α value was adjusted for analysis by adjusting the value of the parameter β to f1, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 100g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. When α is set from 0.05 and 0.50, the effect of other methods is lower than interval-valued information fusion method. In the $Skin \quad St$ data set, the α value was adjusted for analysis by adjusting the value of the parameter β to f2, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5g when the dependency interval reached the optimal state, and when the dependency interval reached the optimal state, the α value was adjusted for analysis. The fusion performance of the IIF method is better than others, when the value of α is set from 0.05 to 0.5, when the threshold changes from 0.05 to 0.1, the value of α of Opt and other three fusion methods is 1.

In a word, according to the AP and AQ value, in most cases, the dependency interval fusion rule which proposed in this paper is more effective than the other three common multi-source information fusion methods. Therefore, in most situations, $IFDI$ method can be a good choice for multi-source decision information systems to form an interval that contains data uncertainty. Besides, we can conclude that, for the four methods about the nine date-sets, the values of AP and AQ decreases as the threshold value increases. Due to the reduction of α, the number of objects in the upper and lower approximations of the set of objects will change, the lower approximations will become larger, and the upper approximations will become smaller. Consequently, AP and AQ are inversely proportional to α. If we want to get the maximum AP and AQ, we need to set a smaller threshold. But it also leads to the tolerance class of each sample being very small, so the correlation between samples is smaller. Assume in an extreme condition that samples themselves would hardly fall into the tolerance relationship formed by all the samples. There is no denying that it will make the task of data mining arduous. So in an actual application, selecting an suitable threshold to obtain the tolerance class of samples is essential.

Furthermore, in multi-source information systems, we compare the proposed method, the multi-source information fusion method based on dependency interval, with other approaches by calculating the values of AP and AQ, which are shown in Figs3-4. This section illustrates the efficiency of the interval-valued information fusion methods. We compare the values of

TABLE X: AP and AQ of Elico with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.45	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.4	**0.0115**	0.0000	0.0023	0.0023	**0.0227**	0.0000	0.0045	0.0045
0.35	**0.0864**	0.0000	0.0185	0.0185	**0.1591**	0.0000	0.0364	0.0364
0.3	**0.2791**	0.0046	0.0811	0.0973	**0.4363**	0.0091	0.1500	0.1773
0.25	**0.5120**	0.0209	0.1796	0.3333	**0.6773**	0.0409	0.3045	0.5000
0.2	**0.8257**	0.1083	0.3750	0.6730	**0.9045**	0.1955	0.5455	0.8045
0.15	**0.9469**	0.4667	0.7742	0.9383	**0.9727**	0.6364	0.8727	0.9682
0.1	0.9910	0.9298	0.9130	**1.0**	0.9955	0.8160	0.9545	**1.0**
0.05	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0

TABLE XI: AP and AQ of Transfusion with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0007**	0.0007	0.0007	0.0007	**0.0014**	0.0014	0.0014	0.0014
0.45	**0.0007**	0.0007	0.0007	0.0007	**0.0014**	0.0014	0.0014	0.0014
0.4	**0.0007**	0.0007	0.0007	0.0007	**0.0014**	0.0014	0.0014	0.0014
0.35	**0.0020**	0.0014	0.0014	0.0014	**0.0041**	0.0027	0.0027	0.0027
0.3	**0.0034**	0.0034	0.0034	0.0034	**0.0068**	0.0068	0.0068	0.0068
0.25	**0.0054**	0.0054	0.0054	0.0054	**0.0108**	0.0108	0.0108	0.0108
0.2	**0.0089**	0.0075	0.0089	0.0075	**0.0176**	0.0149	0.0176	0.0149
0.15	**0.0137**	0.0137	0.0130	0.0130	**0.0270**	0.0270	0.0257	0.0257
0.1	0.0393	0.0379	0.0393	**0.0400**	0.0756	0.0730	0.0757	**0.0770**
0.05	0.1544	**0.1580**	0.1482	0.1526	0.2676	0.2730	**0.2581**	0.2649

TABLE XII: AP and AQ of Money with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.45	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.4	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.35	0.0051	**0.0058**	0.0051	0.0048	0.0102	**0.0117**	0.0102	0.0095
0.3	**0.0336**	0.0278	0.0285	0.0312	**0.0650**	0.0540	0.0555	0.0606
0.25	**0.1276**	0.0964	0.1089	0.1138	**0.2263**	0.1759	0.1964	0.2044
0.2	**0.4182**	0.3186	0.3673	0.3790	**0.5898**	0.4832	0.5372	0.5496
0.15	**0.7700**	0.6779	0.7364	0.7452	**0.8701**	0.8080	0.8482	0.8540
0.1	**0.9544**	0.9282	0.9433	0.9446	**0.9282**	0.9628	0.9708	0.9715
0.05	**1.0**	0.9985	1.0	1.0	**1.0**	0.9993	1.0	1.0

AP and AQ of nine different information sources and different number of attributes. We can see that the values of the red lines are above those of the other colors, which shows that our classification accuracy and quality are better than those of other methods. We can draw conclusions from the pictures that the proposed method of interval-valued information fusion can improve the values of AP and AQ in fusing multi-source information systems, compared to other methods.

Last but not least, to verify the validity of the experimental results, we performed a hypothesis test on the indicators AP and AQ of each data set. The test results are as follows. By calculation, we get that the p-value of the proposed method is respectively 0.0120, 0.0099, 0.0466 for the other three methods in the first data set, which reject the original hypothesis and is statistically significant, when α is set 0.05. The alternative hypothesis is accepted, so the interval generated by this method is considered to be better. In data-set $Transfusion$, p-value is 0.0256, 0.0134, 0.0749 separately. It can be seen that when the significance level is 0.1, the $IFDI$ method has a significant improvement compared with the Mean method. For the third data-set, p-value is 0.0313, 0.0223 ,0.0290 singly. Therefore, multi-source information fusion based on dependency interval is more advanced than other rules. In data-set $Wine$, p-value is 0.0176, 0.02014, 0.0403. The conclusion is the same as before. For the fifth data-set $Seeds$, p-value is 0.0169, 0.0512, 0.0197 respectively. when the significance level is 0.1, the $IFDI$ method is better than Quartile method. In $Booknote$, p-value is 0.0164, 0.0415, 0.0756 singly. Let the significance level be 0.1, the fusion effect $IFDI$ approach is better than Mean rule. For the seventh data-set, p-value is 0.0433, 0.0878, 0.1318 seperately. When the significance

TABLE XIII: AP and AQ of Wine with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.2683**	0.2093	0.2500	0.2560	**0.4231**	0.3462	0.4000	0.4077
0.45	**0.4943**	0.3000	0.4525	0.4689	**0.6615**	0.4615	0.6231	0.6385
0.4	**0.8056**	0.6456	0.7333	0.7808	**0.8923**	0.7846	0.8462	0.8769
0.35	**0.9697**	0.8978	0.9403	0.9697	**0.9846**	0.9462	0.9692	0.9846
0.3	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.25	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.2	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.15	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.1	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.05	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0

TABLE XIV: AP and AQ of Seeds with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.45	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.4	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.35	**0.0182**	0.0000	0.0072	0.0000	**0.0357**	0.0000	0.0143	0.0000
0.3	**0.0526**	0.0072	0.0108	0.0072	**0.1**	0.0143	0.0214	0.0143
0.25	**0.1155**	0.0145	0.0182	0.0219	**0.2071**	0.0286	0.0357	0.0429
0.2	**0.2963**	0.0256	0.0646	0.1200	**0.4571**	0.0500	0.1214	0.2143
0.15	**0.6374**	0.0687	0.1523	0.4070	**0.7786**	0.1286	0.2643	0.5786
0.1	**0.8792**	0.2227	0.4433	0.8421	**0.9357**	0.3643	0.6143	0.9143
0.05	**1.0**	0.7178	1.0	1.0	**1.0**	0.8357	1.0	1.0

TABLE XV: AP and AQ of Booknote with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.45	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.4	**0.0008**	0.0000	0.0004	0.0000	**0.0015**	0.0000	0.0008	0.0000
0.35	0.0050	0.0054	0.0054	**0.0058**	0.0357	0.0108	0.0108	**0.0115**
0.3	**0.0460**	0.0204	0.0317	0.0297	**0.0092**	0.04	0.0615	0.0577
0.25	0.1349	0.1054	0.1384	**0.1398**	0.2377	0.1908	0.2431	**0.2454**
0.2	0.4231	0.3409	0.4317	**0.4278**	0.5946	0.5084	0.6031	**0.5992**
0.15	**0.7931**	0.7219	0.7747	0.7735	**0.8846**	0.8385	0.8731	0.8723
0.1	**0.9667**	0.9417	0.9564	0.9578	**0.9831**	0.9700	0.9777	0.9785
0.05	**1.0**	1.0	1.0	1.0	**1.0**	0.8357	1.0	1.0

level is 0.05, the $IFDI$ method's information fusion effect is best. As to data-set $Wilt$, p-value is 0.0379, 0.0123, 0.0881 singly. the $IFDI$ method's information fusion effect is first-rate. For the last data-set, p-value is 0.0955, 0.0837, 0.0929 respectively. When we widen the confidence level to 0.1, the information fusion effect of our proposed method is the most effective. Therefore, the hypothesis tests prove the validity and reliability of the method.

Through the analysis of the above results, we can know that in most cases, the fusion method proposed in this paper has better classification accuracy and classification quality than the three common fusion methods. In order to further demonstrate the superiority of the proposed method, we will compare the proposed method with the fusion method in reference [33]. The fusion effect of this method is relatively good at present, and it is also a relatively advanced fusion method. This method first sorts the multi-source information system, and then fuses the multi-source interval-valued data into the form of trapezoidal fuzzy number, which is briefly referred to as Hef. In this paper, two classifiers, classical K-nearest neighbor classification (KNN) and probabilistic neural network classifier (PNN), are used to classify the fusion results. Ten fold cross-validation is used to compare the classification accuracy of the two models. The mean and standard deviation of classification accuracy of fusion results are shown in TABLE XX. From the results in the table, we can conclude that for both the KNN classifier and the PNN classifier, for the data sets $Abalone, Money, Transfusion and Wine$, the mean value of classification accuracy obtained by the proposed method is greater than that obtained by the Hef method Therefore, the fusion effect of the proposed method is better than Hef.

TABLE XVI: AP and AQ of Abalone with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.45	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.4	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.35	**0.0000**	0.0000	0.0000	0.0000	**0.0000**	0.0000	0.0000	0.0000
0.3	**0.0004**	0.0004	0.0000	0.0000	**0.0008**	0.0008	0.0000	0.0000
0.25	**0.0015**	0.0008	0.0004	0.0008	**0.0031**	0.0015	0.0008	0.0015
0.2	**0.0097**	0.0074	0.0081	0.0074	**0.0192**	0.0146	0.0162	0.0146
0.15	0.0281	**0.0338**	0.0277	0.0220	0.0546	**0.0654**	0.0538	0.0431
0.1	**0.1484**	0.1434	0.1125	0.0994	**0.2585**	0.2508	0.2023	0.1808
0.05	**0.5250**	0.3944	0.4620	0.2624	**0.6885**	0.3944	0.4620	0.4157

TABLE XVII: AP and AQ of Wilt with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0081**	0.0081	0.0081	0.0081	**0.0161**	0.0161	0.0161	0.0161
0.45	**0.0088**	0.0088	0.0088	0.0088	**0.0174**	0.0174	0.0174	0.0174
0.4	**0.0091**	0.0091	0.0091	0.0091	**0.0181**	0.0181	0.0181	0.0181
0.35	**0.0101**	0.0101	0.0101	0.0101	**0.02**	0.02	0.02	0.02
0.3	**0.0131**	0.0127	0.0131	0.0127	**0.0258**	0.0252	0.0258	0.0252
0.25	**0.0214**	0.0211	0.0211	0.0214	**0.0419**	0.0413	0.0413	0.0419
0.2	**0.0313**	0.0313	0.0313	0.0313	**0.0606**	0.0606	0.0606	0.0606
0.15	**0.0537**	0.0529	0.0526	0.0533	**0.1019**	0.1006	0.1	0.1013
0.1	**0.0977**	0.0966	0.0969	0.0966	**0.1781**	0.1761	0.1768	0.1761
0.05	**0.2901**	0.2873	0.2889	0.2863	**0.4497**	0.4465	0.4484	0.4452

TABLE XVIII: AP and AQ of Skin-St with the variation of α

	AP				AQ			
α	Dep	Max	Quar	Mean	Dep	Max	Quar	Mean
0.5	**0.0553**	0.0553	0.0494	0.0454	**0.1047**	0.1047	0.0942	0.0868
0.45	**0.1037**	0.1037	0.1033	0.1037	**0.1879**	0.1879	0.1874	0.1879
0.4	**0.1812**	0.1794	0.1808	0.1794	**0.3068**	0.3042	0.3063	0.3042
0.35	**0.2435**	0.2435	0.2426	0.2434	**0.3916**	0.3916	0.3905	0.3921
0.3	**0.3305**	0.3305	0.3305	0.3305	**0.4968**	0.4968	0.4968	0.4968
0.25	**0.4399**	0.4399	0.4383	0.4345	**0.6110**	0.6110	0.6095	0.6058
0.2	**0.7086**	0.7086	0.6957	0.6462	**0.8295**	0.8295	0.8205	0.7851
0.15	**0.9927**	0.7826	0.6939	0.6941	**0.9963**	0.8781	0.8193	0.8195
0.1	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0
0.05	**1.0**	1.0	1.0	1.0	**1.0**	1.0	1.0	1.0

TABLE XIX: P-value hypothesis test.

Data sets	p_1	p_2	p_3	α
Wine	0.0120	0.0099	0.0466	0.05
Ecoli	0.0256	0.0134	0.0749	0.1
Money	0.0313	0.0223	0.0290	0.05
Booknote	0.0176	0.02014	0.0403	0.05
Seeds	0.0169	0.0512	0.0197	0.1
Transfusion	0.0164.	0.0415	0.0756	0.1
abalone	0.0433	0.0878	0.1318	0.15
wilt	0.0379	0.0123	0.0881.	0.1
Skin Segementation	0.0955	0.0837	0.0929	0.1

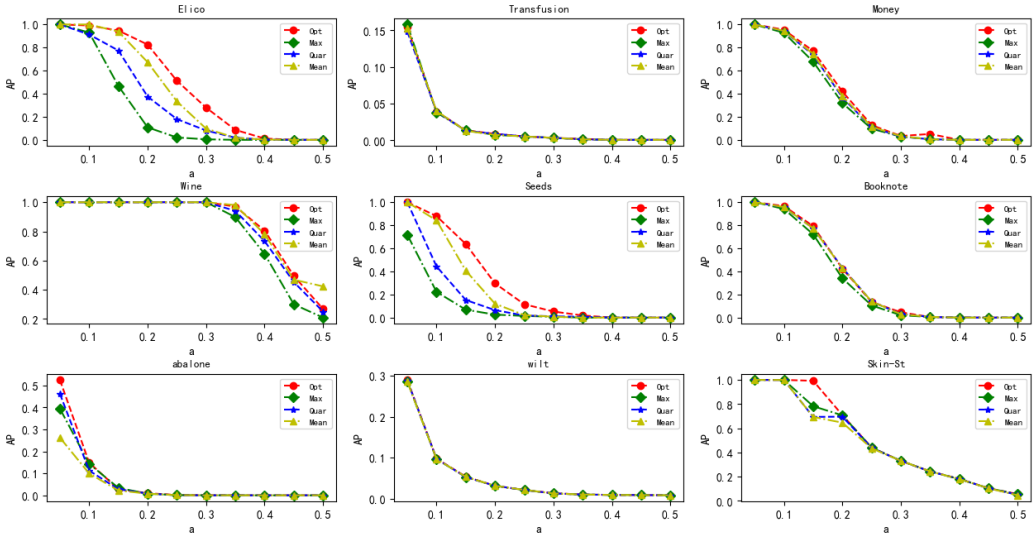

Fig. 3: The AP of information fusion methods

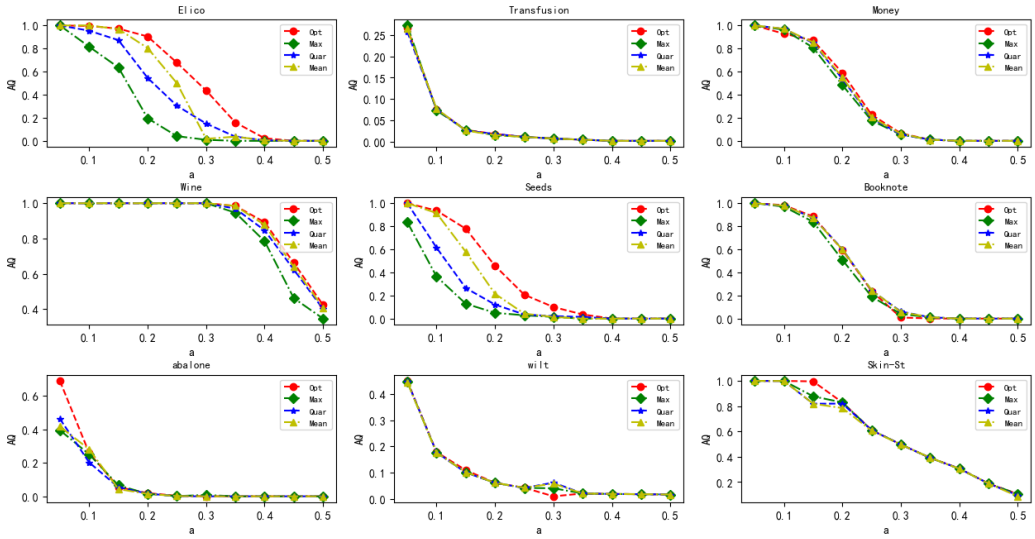

Fig. 4: The AQ of information fusion methods

TABLE XX: Comparison of Dep method and Hef method

	KNN				PNN			
Data sets	Dep		Hef		Dep		Hef	
Wine	**53.7**	**2.8**	52.6	3.4	**54.5**	**3.9**	49.8	4.6
Transfusion	**64.8**	**5.3**	62.7	5.3	**65.2**	**3.8**	64.5	4.4
Money	44.6	2.3	**45.8**	**3.6**	**45.5**	**4.9**	42.7	5.4
Abalone	**36.7**	**2.0**	34.6	1.8	**35.6**	**2.4**	33.8	2.6

V. Conclusions

In this paper, a novel method of information fusion called the multi-source information fusion approach based on interval is advanced, in which several interval-valued tables are fused by constructing reasonable functions. We do not use single-valued method for information fusion, but we use a function to fuse a set of data into an interval value, and then form a new information table of interval values. Firstly, the reasonable function is selected to establish the optimal interval, and the distance between any two intervals is defined and the tolerance grade is determined by the distance. Then, we construct a multi-source information table and calculate the optimal interval according to the rational function. Finally, the experimental results on nine data sets show that the AP and AQ values of the interval-valued fusion method have higher precision in the fusion effect, compared with the mean method, the quartile method and the maximum interval fusion method. In this paper, the β value in the method, we proposed, can only be approximated to the optimal condition of the dependence interval. but it can not reach the best β value, which needs further improvement. In the future, we can apply this method to the multi-source interval-valued information system as one of the interval-valued information fusion methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or persona relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the Associate Editor and the reviewers for their insightful comments and suggestions.

References

[1] W. H. Xu, W. T. Li, Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets. IEEE Transactions on Cybernetics, 2016, 46: 366-379.

[2] Y. H. Qian, J. Y. Liang, D. Y. Li,et al. Measures for evaluating the decision performance of a decision table in rough set theory. Information Sciences, 2008, 178: 181-202.

[3] J. H. Li, C. L. Mei, Y. J. Lv, Knowledge reduction in decision formal contexts. Knowledge-Based Systems, 2011, 24: 709-715.

[4] X. Y. Zhang, W. H. Xu, Lower approximation reduction in ordered information system with fuzzy decision. Applied Mathematics, 2011, 02: 918-921.

[5] Y. Y. Huang, T. R. Li, C. Luo,et al. Dynamic maintenance of rough approximations in multi-source hybrid information systems. Information Sciences, 2020, 530: 108-127.

[6] P. Durso, J. M. Leski, Fuzzy c-ordered medoids clustering for interval-valued data. Pattern Recognition, 2016, 58: 49-67.

[7] S. Dias, P. Brito, Off the beaten track: a new linear model for interval data. European Journal of Operational Research, 2017, 258: 1118-1130.

[8] Y. Leung, M. M. Fischer, W. Z. Wu,et al. A rough set approach for the discovery of classification rules in interval-valued information systems. International Journal of Approximate Reasoning, 2008, 47: 233-246.

[9] W. H. Xu, D. D. Guo, Y. H. Qian, et al. Two-way Concept-cognitive Learning Method: A Fuzzy-based Progressive Learning, IEEE Transactions on Fuzzy Systems. 2022, 31: 1-15.

[10] D.S. Guru, B. B. Kiranagi, P. Nagabhushan, Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. Pattern Recognition Letters, 2004 , 25:1203-1213.

[11] N. Bryson, A. Mobolurin, An action learning evaluation procedure for multiple criteria decision making problems. European Journal of Operational Research,1997, 96: 379-386.

[12] X. Y. Zhang, C. L. Mei, D.G. Chen,et al. Multi-confidence rule acquisition and confidence-preserved attribute reduction in interval-valued decision systems, International Journal of Approximate Reasoning,2014, 55: 1787-1804.

[13] T. T. Nguyen, A. W.-C. Liew, M. T. Tran and M. P. Nguyen, Combining multi classifiers based on a genetic algorithm-A Gaussian mixture model framework. Intelligent Computing Methodologies,2014, 49: 56-67.

[14] C. X. Hu, L. Zhang, S. X. Liu, Incremental approaches to update multigranulation approximations for dynamic information systems. Journal of IntelligentFuzzy Systems Preprint, 2021, 40:4661-4682.

[15] W. Pedrycz, Shadowed sets: Representing and processing fuzzy sets.IEEE Transactions on Systems,1998, 28:103-109.

[16] J. T. Yao, A. V. Vasilakos, W. Pedrycz, Granular computing:Perspectives and challenges. IEEE Transactions on Cybernetics, 2013 , 43:1977-1989.

[17] D. Liu, T. R. Li, J. B. Zhang, Incremental updating approximations in probabilistic rough sets under the variation of attributes. Knowledge-Based Systems, 2015, 73: 81-96.

[18] F. S. Yu, W. Pedrycz, The design of fuzzy information granules: Tradeoffs between specificity and experimental evidence.Applied Soft Computing, 2009, 9:264-273.

[19] T. P. Hong, L. H. Tseng, S. L. Wang, Learning rules from incomplete training examples by rough sets. Expert Systems with Applications, 2002, 22:285-293.

[20] Y. Leung, D. Y. Li, Maximal consistent block technique for rule acquisition in incomplete information systems. Information Sciences, 2003 ,153:85-106.

[21] Z. Pawlak, Rough set theory and its applications to data analysis. Cybernetics and Systems, 1998 ,29: 661-688.

[22] Z. Pawlak, Rough sets and intelligent data analysis. Information Sciences, 2002, 147: 1-12.

[23] Liu. Q, Rough sets and Rough Reasoning. International Journal of General Systems, 2002, 33:569-581.

[24] M. Sokolova, G. Lapalme,A systematic analysis of performance measures for classification tasks. Information Processing & Management, 2009, 45:427-437.

[25] A. Saha, A. Konar, Amit, AK. Nagar, EEG Analysis for Cognitive Failure Detection in Driving Using Type-2 Fuzzy Classifiers, IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1: 437-453.

[26] W. H. Xu, M. M. Li, X. Z. W, Information Fusion Based on Information Entropy in Fuzzy Multi-source Incomplete Information System. International Journal of Fuzzy Systems, 2017 ,19:1200-1216.

[27] J. C. Xu, Y. Wang, H. Y. Mu, et al. Feature Genes Selection Based on Fuzzy Neighborhood Conditional Entropy. Journal of Intelligent and Fuzzy Systems, 2018 , 36: 117-126.

[28] J. H. Dai, W. T. Wang, Q. Xu,et al. Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. Knowledge-Based Systems, 2012 , 27: 443-450.

[29] J. H. Yu, W. H. Xu, Incremental knowledge discovering in interval-valued decision information system with the dynamic data. International Journal of Machine Learning and Cybernetics, 2017 , 8: 849-864 .

[30] J. H. Yu, W. H. Xu, Incremental Computing Approximations with the Dynamic Object Set in Interval-valued Ordered Information System. Fundamenta Informaticae, 2015, 142: 373-397.

[31] J, H. Yu, M. H. Chen, W. H. Xu, Dynamic computing rough approximations approach to time-evolving information granule interval-valued ordered information system. Applied Soft Computing, 2017, 60: 18-29.