Information Sciences 666 (2024) 120450

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A method of data analysis based on division-mining-fusion strategy

Qingzhao Kong [a], Wanting Wang [a], Weihua Xu [b,*], Conghao Yan [a]

[a] Department of Science, Jimei University, Xiamen, 361021, PR China
[b] College of Artificial Intelligence, Southwest University, Chongqing, 400715, PR China

ARTICLE INFO

Keywords:
Division-mining-fusion strategy
Feature selection
Information table
Rough set theory

ABSTRACT

With the advancement of data technology and storage services, the scale and complexity of data are rapidly growing. Consequently, promptly analyzing data and deriving precise insights have become urgent. Nevertheless, traditional methods struggle to balance the speed and accuracy of data mining. This paper proposes a data analysis technique called the Division-Mining-Fusion (DMF) strategy to tackle this challenge. Specifically, we divide a large-scale and complex dataset into multiple small-scale and simple sub-datasets. Then, we extract the knowledge embedded within each sub-dataset. Finally, we combine the extracted knowledge from each sub-dataset to accomplish learning tasks. To demonstrate the superior performance of the DMF strategy, we apply it to two fields: rough set theory and feature selection. The DMF strategy can accelerate the speed of data mining, enhance the accuracy of data analysis, and reduce the dimensionality of data. These advantages suggest that the DMF strategy outperforms traditional methods in processing data more efficiently. In addition, the number of sub-datasets is a crucial parameter of the DMF strategy. As the number of sub-datasets increases, the ability of the DMF strategy to analyze data continuously improves.

1. Introduction

Data is a type of physical symbol that can be recognized, which reflects the essence, condition, and mutual connections of objective things. As internet technology, data storage services, and communication methods continue to develop, data is becoming increasingly extensive and complex. Data, like air, pervades every aspect of production and life. Extracting useful and reliable knowledge from vast and intricate datasets is a pivotal concern today. Numerous technologies and methods analyze data and acquire meaningful information or objective laws to accomplish learning tasks [1–5].

1.1. Overview of related works

Probability theory is a highly effective approach to data analysis. However, when employing the probabilistic method to process data, it must have the probability distribution and associated parameters in advance. Fuzzy set theory is a valuable tool for studying fuzzy data. Nevertheless, one should determine the membership degree of each datum or sample before utilizing fuzzy set theory for fuzzy analysis and inference. Rough set models (RSMs) perform well with data problems related to uncertainty reasoning and

[*] Corresponding author.
E-mail addresses: kongqingzhao@163.com (Q. Kong), 957329054@qq.com (W. Wang), chxuwh@gmail.com (W. Xu), 840223328@qq.com (C. Yan).

https://doi.org/10.1016/j.ins.2024.120450
Received 2 July 2023; Received in revised form 27 February 2024; Accepted 5 March 2024

Available online 12 March 2024

uncertainty management. Unlike other theories, one of the advantages of RSM is that all parameters are available in the sample data [6]. Based on various learning tasks and data characteristics, two approximation operators are constructed in RSMs. These two operators can approximately describe any concept. Researchers have applied RSM to uncertainty analysis, granular computing, and machine learning [7–9]. So far, they have developed many RSMs. For instance, the local RSM, proposed by Y.H. Qian, can significantly boost the efficiency of data mining by focusing on the data related to the target concept. However, it cannot significantly enhance the accuracy of knowledge classification [10]. Q.Z. Kong introduces the variable universe RSM, which emphasizes data highly relevant to the learning tasks [11]. This model enables rapid analysis of data. Parallel computing utilizes many resources to solve large and complex computational problems, efficiently enhancing computing speed. Numerous scholars have adopted parallel computing to tackle the challenges of rough set theory. For example, based on parallel algorithms, indiscernible relation [12] can rapidly compute the approximations of a rough set. The method of parallel matrix quickly obtains the approximations in the dominance-based rough set [13]. Under an incomplete information table, seeking approximations of rough sets through parallel computing algorithms is also actively explored [14]. These methods mentioned focus on the computational efficiency of mining data. In addition, inspired by the idea of dividing all samples into three disjoint sample subsets in rough set theory, Y. Y. Yao proposes an important data processing method called three-way decision [15]. The three-way decision idea involves artificial intelligence, network security, conflict analysis, and other fields [16,17]. A three-way decision acts as a transitional state of a two-way decision. In fact, we usually need to make a three-way decision multiple times to approximate the two-way decision. Therefore, three-way decisions can effectively improve decision-making accuracy but reduce decision-making efficiency.

Feature selection is also called attribute selection or attribute reduction. It aims to select representative attributes to optimize the data information system or select some crucial attributes to reduce the dimensionality of the data [18–21]. Feature selection significantly improves the efficiency of learning algorithms. It has vital and in-depth applications in many fields, such as pattern recognition, data mining, and machine learning [22–24]. Ignoring a dataset′s useless or unimportant attributes will not result in information loss. At this point, we only need to select those critical attributes. This process brings three benefits. The first advantage is that it will complete learning tasks more easily, simplify the model, and make it easier to understand. The second one is to save storage space and time consumption. The third one is to reduce the disaster of data dimension and the risk of overfitting. Scholars have defined and studied feature selection from many perspectives. There are many methods for selecting attributes or features. For example, researchers widely use concept-cognitive learning and information entropy for feature selection [25–28]. Parallel computing also commonly handles attribute reduction in rough set theory [29,30]. Chen et al. researched the attribute reduction of a dominance-based neighborhood rough set using parrel computation [31]. The parallel feature selection algorithms based on a rough hypercuboid approach address the rapidly expanding data [32]. The parallel multi-reduction algorithm exacts more knowledge from complex information systems [33]. These methods can significantly improve the accuracy of feature selection. In addition, researchers often employ rough set theory and granular ball theory for feature selection. The advantage of these methods is that they can quickly select features, while the disadvantage is that they do not significantly improve the accuracy [34,35].

1.2. Our work

There are many measures for evaluating the quality of a data mining method. Undoubtedly, efficiency and effectiveness are the two most important. The above analysis shows that some methods can significantly accelerate speed but cannot improve accuracy. Others are good at increasing accuracy but cannot quickly analyze data. Therefore, it is challenging for traditional methods to balance the efficiency and effectiveness of data mining. This study addresses this issue by proposing the Division-Mining-Fusion (DMF) strategy to analyze data and achieve the following three goals:

- Reduce consumption time and improve the efficiency of data mining;
- Reduce errors in data analysis and improve the effectiveness of data mining;
- Reduce the data dimension and relieve dimension disasters or over-fitting problems.

The DMF strategy involves three data processing steps. First, it divides a dataset (an information table) into multiple sub-datasets (information subtables). The first effect of splitting a dataset is that it significantly reduces the size of each sub-dataset, which helps to improve the speed of data analysis. The second effect is that the label in each sub-dataset is relatively simple, reducing the complexity of the data and improving the accuracy of data mining. This step provides the necessary preparation for quickly and accurately mining data. Second, it mines the knowledge hidden in each sub-dataset. Finally, it fuses all the knowledge inferred from each sub-dataset to complete learning tasks. This third step ensures that the process fully considers all information from the data and that the obtained knowledge is true and reliable.

The difference between the traditional method and the DMF strategy appears in Figs. 1 and 2 below.

The DMF strategy does not only fully consider all the data in the dataset but also effectively reduces the interference of the data size and complexity. The research framework or main content of this paper appears in Fig. 3.

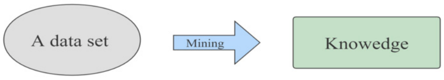

Fig. 1. Traditional method for data analysis.

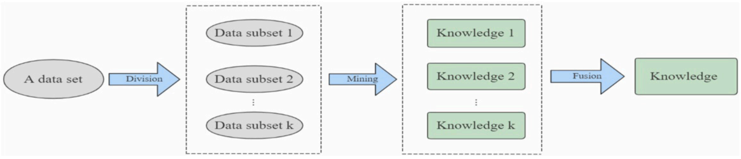

Fig. 2. DMF strategy for data analysis.

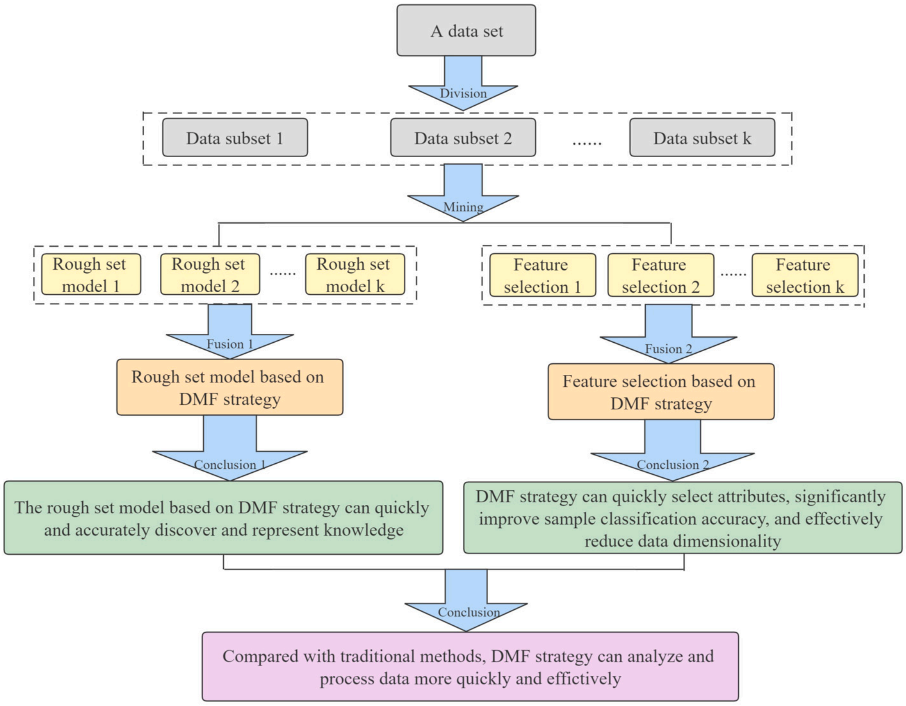
A data set
Division
Data subset 1
Data subset 2
......
Data subset k
Mining
Rough set model 1
Rough set model 2
......
Rough set model k
Feature selection 1
Feature selection 2
......
Feature selection k
Fusion 1
Fusion 2
Rough set model based on DMF strategy
Feature selection based on DMF strategy
Conclusion 1
Conclusion 2
The rough set model based on DMF strategy can quickly and accurately discover and represent knowledge
DMF strategy can quickly select attributes, significantly improve sample classification accuracy, and effectively reduce data dimensionality
Conclusion
Compared with traditional methods, DMF strategy can analyze and process data more quickly and effictively

Fig. 3. The research framework of this paper.

Next, we introduce the main contents of this study. Section 2 lists several concepts related to Pawlak RSM and attribute reduction are listed. Section 3 proposes the information subtable family of an information table and explains it with two examples. Finally, some properties of an information table and its information subtable family are studied. Section 4 develops the RSM based on the DMF strategy. RSM based on the DMF strategy has more advantages in knowledge discovery than Pawlak RSM. Section 5 defines and studies three types of attribute reductions related to DMF strategy. Section 6 designs four algorithms for computing RSM and attribute reductions based on the DMF strategy. Compared with the traditional algorithms, the time complexities of these algorithms based on DMF strategy are notably lower. Section 7, through the detailed numerical experiments, confirms that the DMF strategy can better complete learning tasks than traditional methods. Section 8 briefly summarizes the main contents of this paper and clarifies the works that need further study.

2. Preliminaries

The information table that displays the data visually is an important information system. Data mining often gathers the collected samples and their labels to form an information table and then analyzes the data in the information table. Therefore, this paper regards a dataset and an information table as equivalent, with no difference.

Usually, an information table can be marked by

$$I = (OB, AT, \{V_a | a \in AT\}, \{f_a | a \in AT\}) \tag{1}$$

where OB, AT, V_a and f_a represent the universe, attribute set, attribute value of attribute a, and information function about attribute a, respectively [36].

For each $A \subseteq AT$ and each $x \in OB$, the equivalence relation E_A, the equivalence class of x, and a partition of the universe induced from E_A can be written as [37]:

$$xE_A y \Leftrightarrow \forall a \in A\ (f_a(x) = f_a(y));$$

$$[x]_A = \{y \in OB | xE_A y\};$$
$$OB/E_A = \{[x]_A | x \in OB\}.$$

2.1. Rough set theory

Z. Pawlak first proposed and deeply studied RSM using the partition of the universe. RSM is increasingly becoming a suitable method for intelligent data processing [38–42]. Based on Pawlak's idea, any subset of the universe can be approximately described by two approximation sets as follows [6].

Definition 2.1. In the information table described by Eq. (1), for each $X \subseteq OB$, we call

$$\underline{apr}_A(X) = \{x \in OB \mid [x]_A \subseteq X\},$$
$$\overline{apr}_A(X) = \{x \in OB \mid [x]_A \cap X \neq \emptyset\}$$

the lower and upper approximations of X, respectively.

Knowledge representation represents information as a pattern consistent with machine processing. Researchers solve complex tasks in artificial intelligence by using it to simulate human understanding and reasoning of the world. The essence of knowledge representation is a description of knowledge. Due to the complexity, multi-modality, and noise of data, the description of knowledge is often approximate. In rough set theory, any concept is approximately described by two approximation sets. From Definition 2.1, for each $X \subseteq OB$, the relation $\underline{apr}_A(X) \subseteq X \subseteq \overline{apr}_A(X)$ holds, that is, X is approximately described by $\underline{apr}_A(X)$ and $\overline{apr}_A(X)$. The greater the difference between $\overline{apr}_A(X)$ and $\underline{apr}_A(X)$ is, the coarser the description is. Therefore, Z. Pawlak introduced the following definition to illustrate the accuracy of the description [6,37].

Definition 2.2. In the information table described by Eq. (1), for each $X \subseteq OB$, we call

$$\alpha_A(X) = \frac{|\underline{apr}_A(X)|}{|\overline{apr}_A(X)|} \tag{2}$$

the accuracy of the approximate description of the concept X.

From Definition 2.1, for any concept $X \subseteq OB$, all samples fall into three disjoint subsets, namely positive, negative, and boundary regions.

$$Pos_A(X) = \underline{apr}_A(X);$$
$$Neg_A(X) = OB - \overline{apr}_A(X);$$
$$Bou_A(X) = \overline{apr}_A(X) - \underline{apr}_A(X).$$

According to Definition 2.1, there are three basic facts:

(1) For any $x \in Pos_A(X)$, x must be the sample in X,

(2) For any $x \in Neg_A(X)$, x must not belong to X,

(3) For any $x \in Bou_A(X)$, it cannot be determined whether x belongs to X.

Through the above analysis, for any sample $x \in OB$, if $x \in (Pos_A(X) \cup Neg_A(X))$, then x can be accurately classified; while if $x \overline{\in} (Pos_A(X) \cup Neg_A(X))$ or $x \in Bou_A(X)$, then x cannot be accurately classified. So, the following formula can measure the accuracy of knowledge classification [37].

Definition 2.3. In the information table described by Eq. (1), for each $X \subseteq OB$,

$$\beta_A(X) = \frac{|Pos_A(X)| + |Neg_A(X)|}{|OB|} \tag{3}$$

is called the accuracy of knowledge classification with respect to X.

The core issue of many learning tasks is to make reasonable and scientific decisions or predictions based on the results of data analysis. A decision information table, which contains conditional attributes and decision attributes, is a specialized and critical information expression system. In many cases, we study various decision problems using decision information tables.

A decision information table is usually described by a quintuple [43]:

$$DI = (OB, AT \cup \{d\}, \{V_a | a \in AT\} \cup \{V_d\}, \{f_a | a \in AT\} \cup \{f_d\}) \tag{4}$$

where AT is the set of condition attributes, d is a decision attribute, V_d and f_d represent the attribute value and information function about attribute d, respectively. Here, the meanings of OB, V_a, f_a are the same as those in the information table in Eq. (1). At the same time, according to $E_A, OB/E_A$ in the information table in Eq. (1), $E_{\{d\}}, OB/E_{\{d\}}$ can also be similarly defined in the decision information table.

The focus is on whether the decision is correct when conducting decision analysis. Because RSMs have deep and extensive practical applications in decision reasoning, we must evaluate the decision accuracy based on the Pawlak model in the decision information table. Then, a specific formula for calculating the decision accuracy follows [43].

Definition 2.4. In the decision information table described by Eq. (4), let $OB/E_{\{d\}} = \{D_1, D_2, \cdots, D_s\}$ be a partition on universe OB, then we call

$$\gamma_A(d) = \frac{|\cup_{j=1}^{s} \underline{apr_A}(D_j)|}{|OB|} \tag{5}$$

the decision accuracy with respect to the decision attribute d.

The decision accuracy $\gamma_A(d)$ represents the ratio of the number of samples that can be accurately classified into the equivalent classes in $OB/E_{\{d\}}$ to the number of all samples in the universe. Then, decision accuracy describes the degree of dependence of decision attributes on conditional attributes. Meanwhile, we can regard every sample in the decision information table as a decision rule. The decision information table is actually a collection of many decision rules. Furthermore, $\gamma_A(d) = 0$ means that all the rules obtained from the decision table are uncertain. And $\gamma_A(d) = 1$ means all rules are completely reliable. Therefore, the accuracy of a decision rule can reflect the proportion of deterministic decisions, and the larger the value of $\gamma_A(d)$ is, the more likely it is to obtain reliable rules.

2.2. Feature selection

Feature selection refers to choosing some crucial features from existing features. Its purpose is to reduce the dimension of the dataset, improve learning efficiency, and simplify or optimize specific tasks. RSM has a wide range of applications in feature selection. In rough set theory, feature selection is called attribute reduction. Attribute reduction is a core content in rough set theory and granular computing theory. Many scholars conduct long-term and extensive research on attribute reduction and obtain many meaningful conclusions [6,12,44–46].

Based on various learning tasks, researchers introduce many reductions and study them in depth. Generally, all reductions fall into two categories: one aims to maintain knowledge classification ability unchanged; another ensures that the knowledge representation ability does not decrease. Three common and important reductions follow:

Definition 2.5. In the information table described by Eq. (1), if $A' \subseteq A$ satisfies:

(1) $OB/E_{A'} = OB/E_A$,

(2) For any $a \in A'$, $OB/E_{A'-\{a\}} \neq OB/E_A$,

then A' is called the reduction of A. And we use $Reduct(A)_R$ to denote A', i.e., $Reduct(A)_R = A'$. Because the reduction of A is often not unique, the set of all reductions is represented by $REDUCT(A)_R$.

Definition 2.6. In the information table described by Eq. (1), for any $X \subseteq U$, if $A' \subseteq A$ satisfies:

(1) $\underline{apr}_{A'}(X) = \underline{apr}_A(X)$,

(2) For any $a \in A'$, $\underline{apr}_{A'-\{a\}}(X) \neq \underline{apr}_A(X)$,

then A' is called the reduction of A with respect to lower approximation set $\underline{apr}_A(X)$. And we use $Reduct(A)_L$ to denote A', i.e., $Reduct(A)_L = A'$. Similarly, the set of all reductions is marked by $REDUCT(A)_L$.

Definition 2.7. In the decision information table described by Eq. (4), $OB/E_{\{d\}} = \{D_1, D_2, \cdots, D_s\}$ is a partition on universe OB. If $A' \subseteq A$ satisfies:

(1) $Pos_{A'}(\{d\}) = Pos_A(\{d\})$,

(2) For any $a \in A'$, $Pos_{A'-\{a\}}(\{d\}) \neq Pos_A(\{d\})$,

then A' is called the reduction of A with respect to $Pos_A(\{d\})$. And we use $Reduct(A)_P$ to denote A', i.e., $Reduct(A)_P = A'$. And the set of all reductions is denoted by $REDUCT(A)_P$.

3. A dataset (information table) and its sub-datasets (information subtables)

We describe the datasets conveniently and intuitively using information tables to represent the datasets. The first step of the DMF strategy is to divide the dataset into many sub-datasets, that is, to divide an information table into multiple information subtables.

Table 1
An information table about 24 students.

OB	a_1	a_2	a_3	a_4	a_5	a_6
x_1	1	1	1	0	1	0
x_2	1	1	0	0	1	0
x_3	0	0	0	0	1	0
x_4	1	0	0	0	0	0
x_5	0	0	0	1	1	0
x_6	1	1	1	0	1	0
x_7	1	0	0	0	0	0
x_8	0	0	0	0	0	1
x_9	1	0	0	0	1	0
x_{10}	0	1	1	0	1	0
x_{11}	0	0	0	0	0	0
x_{12}	1	1	0	1	0	0
x_{13}	1	1	1	0	1	0
x_{14}	0	1	1	0	1	0
x_{15}	0	0	0	1	1	0
x_{16}	1	0	0	0	0	0
x_{17}	0	0	0	1	1	0
x_{18}	0	0	0	0	0	0
x_{19}	1	0	0	0	1	0
x_{20}	1	0	0	1	0	0
x_{21}	1	1	1	0	0	0
x_{22}	0	0	0	1	1	0
x_{23}	1	0	0	1	0	0
x_{24}	0	0	0	0	0	1

Based on the above analysis,

$$\mathcal{I} = \{I_h = (U_h, AT, \{V_{a,h}|a \in AT\}, \{f_{a,h}|a \in AT\})|h = 1, 2, \cdots, k\} \tag{6}$$

is called the information subtable family of the information table described by Eq. (1), where $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$ is a partition on OB.

For any $h \in \{1, 2, \cdots, k\}$, we call

$$I_h = (U_h, AT, \{V_{a,h}|a \in AT\}, \{f_{a,h}|a \in AT\})$$

the hth information subtable of the information table described by Eq. (1). Meanwhile, $V_{a,h}$ and $f_{a,h}$ respectively represent the attribute value and information function for attribute a (where all objects are from U_h).

Similarly, in the hth information subtable $I_h = (U_h, AT, \{V_{a,h}|a \in AT\}, \{f_{a,h}|a \in AT\})$, $E_{A,h}, [x]_{A,h}$ and $U_h/E_{A,h}$ can be respectively induced as follows:

$$xE_{A,h}y \Leftrightarrow \forall a \in A\ (f_a(x) = f_a(y));$$

$$[x]_{A,h} = \{y \in U_h|xE_{A,h}y\};$$

$$U_h/E_{A,h} = \{[x]_{A,h}|x \in U_h\}.$$

Next, two specific examples further explain the information subtable family of an information table.

Example 3.1. In the information table described by Eq. (1), where $OB = \{x_1, x_2, \cdots, x_{24}\}$ is a set of 24 students, $AT = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ is a set of 6 attributes. a_1, a_2, a_3, a_4, a_5 and a_6 respectively represent male, excellent results, volunteer experience, poor student, social practice, and cheating record. In addition, if x has attribute a, it is represented by $f_a(x) = 1$. Otherwise, write it as $f_a(x) = 0$. More details are in Table 1.

Let $A = AT$, one can find that

$$OB/E_A = \{\{x_1, x_6, x_{13}\}, \{x_2\}, \{x_3\}, \{x_4, x_7, x_{16}\}, \{x_5, x_{15}, x_{17}, x_{22}\}, \{x_8, x_{24}\}, \{x_9, x_{19}\}, \{x_{10}, x_{14}\}, \{x_{11}, x_{18}\}, \{x_{12}\}, \{x_{20}, x_{23}\}, \{x_{21}\}\}.$$

Suppose the 24 students in Table 1 are from three different grades. That is, the first eight students are from grade 1; the middle eight ones are from grade 2; the rest are from grade 3. And it is required to select several students from three grades as scholarship winners.

This method can obtain three information subtables. Thus,

$$\mathcal{I} = \{I_h = (U_h, AT, \{V_{a,h}|a \in AT\}, \{f_{a,h}|a \in AT\})|h = 1, 2, 3\}$$

Table 2
An information subtable about grade 1.

U_1	a_1	a_2	a_3	a_4	a_5	a_6
x_1	1	1	1	0	1	0
x_2	1	1	0	0	1	0
x_3	0	0	0	0	1	0
x_4	1	0	0	0	0	0
x_5	0	0	0	1	1	0
x_6	1	1	1	0	1	0
x_7	1	0	0	0	0	0
x_8	0	0	0	0	0	1

Table 3
An information subtable about grade 2.

U_2	a_1	a_2	a_3	a_4	a_5	a_6
x_9	1	0	0	0	1	0
x_{10}	0	1	1	0	1	0
x_{11}	0	0	0	0	0	0
x_{12}	1	1	0	1	0	0
x_{13}	1	1	1	0	1	0
x_{14}	0	1	1	0	1	0
x_{15}	0	0	0	1	1	0
x_{16}	1	0	0	0	0	0

Table 4
An information subtable about grade 3.

U_3	a_1	a_2	a_3	a_4	a_5	a_6
x_{17}	0	0	0	1	1	0
x_{18}	0	0	0	0	0	0
x_{19}	1	0	0	0	1	0
x_{20}	1	0	0	1	0	0
x_{21}	1	1	1	0	0	0
x_{22}	0	0	0	1	1	0
x_{23}	1	0	0	1	0	0
x_{24}	0	0	0	0	0	1

is the information subtable family, where $U_1 = \{x_1, x_2, \cdots, x_8\}$, $U_2 = \{x_9, x_{10}, \cdots, x_{16}\}$, $U_3 = \{x_{17}, x_{18}, \cdots, x_{24}\}$. And the three information subtables I_1, I_2 and I_3 are represented by Tables 2, 3, and 4, respectively.

According to Tables 2–4, we have

$$U_1/E_{A,1} = \{\{x_1, x_6\}, \{x_2\}, \{x_3\}, \{x_4, x_7\}, \{x_5\}, \{x_8\}\};$$

$$U_2/E_{A,2} = \{\{x_9\}, \{x_{10}, x_{14}\}, \{x_{11}\}, \{x_{12}\}, \{x_{13}\}, \{x_{15}\}, \{x_{16}\}\};$$

$$U_3/E_{A,3} = \{\{x_{17}, x_{22}\}, \{x_{18}\}, \{x_{19}\}, \{x_{20}, x_{23}\}, \{x_{21}\}, \{x_{24}\}\}.$$

Example 3.2 *(Continued from Example 3.1).* Here, we consider another question: for 24 students in Table 1, we need to select several students from boys and girls to participate in the men's and women's debate competitions, respectively. Based on gender, two information subtables can be obtained, that is,

$$\mathcal{I} = \{I_h = (U_h, AT, \{V_{a,h} | a \in AT\}, \{f_{a,h} | a \in AT\}) | h = 1, 2\}$$

is the information subtable family of the information table described in Table 1. And the two information subtables I_1 and I_2 are in Tables 5 and 6, respectively.

From Examples 3.1 and 3.2, sometimes we need to divide an information table into multiple information subtables for data mining.

4. Rough set theory based on DMF strategy

Rough set is a granular computing model selected to address data problems in many fields, such as intelligent analysis, decision-making, and prediction. However, it is simple to find that all existing rough set models are based on an information table. We must propose an appropriate RSM to extract useful information from multiple information tables.

Table 5
An information subtable about male students.

U_1	a_1	a_2	a_3	a_4	a_5	a_6
x_1	1	1	1	0	1	0
x_2	1	1	0	0	1	0
x_4	1	0	0	0	0	0
x_6	1	1	1	0	1	0
x_7	1	0	0	0	0	0
x_9	1	0	0	0	1	0
x_{12}	1	1	0	1	0	0
x_{13}	1	1	1	0	1	0
x_{16}	1	0	0	0	0	0
x_{19}	1	0	0	0	1	0
x_{20}	1	0	0	1	0	0
x_{21}	1	1	1	0	0	0
x_{23}	1	0	0	1	0	0

Table 6
An information subtable about female students.

U_2	a_1	a_2	a_3	a_4	a_5	a_6
x_3	0	0	0	0	1	0
x_5	0	0	0	1	1	0
x_8	0	0	0	0	0	1
x_{10}	0	1	1	0	1	0
x_{11}	0	0	0	0	0	0
x_{14}	0	1	1	0	1	0
x_{15}	0	0	0	1	1	0
x_{17}	0	0	0	1	1	0
x_{18}	0	0	0	0	0	0
x_{22}	0	0	0	1	1	0
x_{24}	0	0	0	0	0	1

4.1. Rough set model based on the DMF strategy

Here, using the three steps of division, mining, and fusion, we introduce how to construct a novel RSM based on the DMF strategy.

Step 1 (Division): First, by the strategy in Section 3, we divide an information table into k information subtables.

Step 2 (Mining): Second, suppose that $\mathcal{I}=\{I_h=(U_h,AT,\{V_{a,h}|a\in AT\},\{f_{a,h}|a\in AT\})|h=1,2,\cdots,k\}$ is the information subtable family of the information table described by Eq. (1), $A\subseteq AT$ is a subset of attributes, $X\subseteq OB$ is a subset of the universe OB. For the hth information subtable, according to Definition 2.1, a Pawlak RSM can be induced as follows:

$$\underline{apr}_{A,h}(X\cap U_h)=\{x\in U_h\mid [x]_{A,h}\subseteq(X\cap U_h)\};$$
$$\overline{apr}_{A,h}(X\cap U_h)=\{x\in U_h\mid [x]_{A,h}\cap(X\cap U_h)\neq\emptyset\}.$$

In this way, based on k information subtables, we can obtain k RSMs.

Step 3 (Fusion): Third, the k RSMs are fused employing the union of sets, and we develop a new rough set model as follows.

$$\underline{apr}_{A,DMF}(X)=\cup_{h=1}^{k}\underline{apr}_{A,h}(X\cap U_h);$$
$$\overline{apr}_{A,DMF}(X)=\cup_{h=1}^{k}\overline{apr}_{A,h}(X\cap U_h).$$

Therefore, after the three steps of division, mining, and fusion, we propose a novel RSM based on multiple information tables as follows.

Definition 4.1. In the information subtable family described by Eq. (6), for each $X\subseteq OB$, we call

$$\begin{aligned}\underline{apr}_{A,DMF}(X)&=\cup_{h=1}^{k}\underline{apr}_{A,h}(X\cap U_h)\\&=\cup_{h=1}^{k}\{x\in U_h\mid [x]_{A,h}\subseteq(X\cap U_h)\};\\\overline{apr}_{A,DMF}(X)&=\cup_{h=1}^{k}\overline{apr}_{A,h}(X\cap U_h)\\&=\cup_{h=1}^{k}\{x\in U_h\mid [x]_{A,h}\cap(X\cap U_h)\neq\emptyset\}\end{aligned}$$

the lower and upper approximations of X, respectively.

By Definitions 2.1 and 4.1, we know that if $k = 1$, the model proposed in Definition 4.1 degenerates into the Pawlak model, that is, if $k = 1$, the following two equations hold.

$$\underline{apr}_{A,DMF}(X) = \underline{apr}_{A}(X),\ \overline{apr}_{A,DMF}(X) = \overline{apr}_{A}(X).$$

Example 4.1 *(Continued from Example 3.1).* Suppose $X = \{x_1, x_4, x_5, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}\}$, based on Definition 4.1, we have

$$\begin{aligned}
&\underline{apr}_{A,1}(X \cap U_1) = \{x_5\}, && \overline{apr}_{A,1}(X \cap U_1) = \{x_1, x_4, x_5, x_6, x_7\};\\
&\underline{apr}_{A,2}(X \cap U_2) = \{x_{11}, x_{13}, x_{15}\}, && \overline{apr}_{A,2}(X \cap U_2) = \{x_{11}, x_{13}, x_{15}\};\\
&\underline{apr}_{A,3}(X \cap U_3) = \{x_{18}, x_{19}\}, && \overline{apr}_{A,3}(X \cap U_3) = \{x_{17}, x_{18}, x_{19}, x_{22}\}.
\end{aligned}$$

Hence, one can find that

$$\begin{aligned}
&\underline{apr}_{A,DMF}(X) = \{x_5, x_{11}, x_{13}, x_{15}, x_{18}, x_{19}\};\\
&\overline{apr}_{A,DMF}(X) = \{x_1, x_4, x_5, x_6, x_7, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}, x_{22}\}.
\end{aligned}$$

4.2. *The properties of rough set model based on the DMF strategy*

In this part, we discuss the properties of an RSM based on the DMF strategy. To better understand this new RSM, we give an equivalent description of Definition 4.1 as follows.

Proposition 4.1. *In the information subtable family described by Eq. (6), for each $X \subseteq OB$,*

$$\begin{aligned}
\underline{apr}_{A,DMF}(X) &= \cup_{h=1}^{k} \underline{apr}_{A,h}(X)\\
&= \cup_{h=1}^{k} \{x \in U_h \mid [x]_{A,h} \subseteq X\};\\
\overline{apr}_{A,DMF}(X) &= \cup_{h=1}^{k} \overline{apr}_{A,h}(X)\\
&= \cup_{h=1}^{k} \{x \in U_h \mid [x]_{A,h} \cap X \neq \emptyset\}.
\end{aligned}$$

Many factors affect the quality of the collected data. For example, the accuracy of the instrument will limit it. Transmission delay, smoothness, and other factors in the network transmission process will also affect quality. Sometimes, data will be missing or have considerable noise. These facts indicate that the data to be analyzed is complex and multi-modal. The knowledge obtained will have errors to some extent. We must evaluate the reliability or accuracy of the acquired knowledge. Therefore, it is necessary to design reasonable measures to solve this problem. This section will propose three measures related to RSM based on DMF strategy from the perspectives of knowledge representation, knowledge classification, and decision rules.

For the RSM based on the DMF strategy, any sample subset is between its lower and upper approximation sets. Thus, it is approximately represented by these two approximation sets. Then, the accuracy of knowledge representation can be estimated as follows.

Definition 4.2. In the information subtable family described by Eq. (6), for each $X \subseteq OB$, we call

$$\alpha_{A,DMF}(X) = \frac{|\underline{apr}_{A,DMF}(X)|}{|\overline{apr}_{A,DMF}(X)|} \tag{7}$$

the accuracy of approximation description of X.

Example 4.2 *(Continued from Examples 3.1 and 4.1).* For $X = \{x_1, x_4, x_5, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}\}$, we can get

$$\begin{aligned}
&\underline{apr}_{A,DMF}(X) = \{x_5, x_{11}, x_{13}, x_{15}, x_{18}, x_{19}\};\\
&\overline{apr}_{A,DMF}(X) = \{x_1, x_4, x_5, x_6, x_7, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}, x_{22}\}.
\end{aligned}$$

Then

$$\alpha_{A,DMF}(X) = 0.5$$

Meanwhile, we can list the two approximation sets of the Pawlak model as follows.

$$\underline{apr}_A(X) = \{x_{11}, x_{18}\};$$
$$\overline{apr}_A(X) = \{x_1, x_4, x_5, x_6, x_7, x_9, x_{11}, x_{13}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{22}\}.$$

Then we have

$$\alpha_A(X) = 0.14.$$

Proposition 4.2. *In the information subtable family described by Eq. (6), for any $X \subseteq OB$, we have*

$$\alpha_A(X) \le \alpha_{A,DMF}(X).$$

Proposition 4.2 and Example 4.2 show that an RSM based on the DMF strategy has a stronger knowledge description ability than the Pawlak model.

Like the Pawlak model, for any subset $X \subseteq OB$, the RSM based on the DMF strategy can divide all samples into three disjoint sample subsets as follows.

$$Pos_{A,DMF}(X) = \underline{apr}_{A,DMF}(X);$$
$$Neg_{A,DMF}(X) = OB - \overline{apr}_{A,DMF}(X);$$
$$Bou_{A,DMF}(X) = \overline{apr}_{A,DMF}(X) - \underline{apr}_{A,DMF}(X).$$

For any sample $x \in (Pos_{A,DMF}(X) \cup Neg_{A,DMF}(X))$, we can clearly infer that x belongs to X or does not belong to X. But for any sample $x \in Bou_{A,DMF}(X)$, we cannot determine whether x belongs to X. Therefore, we define the accuracy of sample classification based on the DMF strategy as follows.

Definition 4.3. In the information subtable family described by Eq. (6), $A \subseteq AT$ is a subset of attributes. For each $X \subseteq OB$,

$$\beta_{A,DMF}(X) = \frac{|Pos_{A,DMF}(X)| + |Neg_{A,DMF}(X)|}{|OB|} \tag{8}$$

is called the accuracy of knowledge classification related to concept X.

Proposition 4.3. *In the information subtable family described by Eq. (6), $A \subseteq AT$ is a subset of attributes. For each $X \subseteq OB$, we have*

$$\beta_{A,DMF}(X) = 1 - \frac{|Bou_{A,DMF}(X)|}{|OB|}.$$

Example 4.3 *(Continued from Example 4.1).* For $X = \{x_1, x_4, x_5, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}\}$, we have

$$\underline{apr}_{A,DMF}(X) = \{x_5, x_{11}, x_{13}, x_{15}, x_{18}, x_{19}\},$$
$$\overline{apr}_{A,DMF}(X) = \{x_1, x_4, x_5, x_6, x_7, x_{11}, x_{13}, x_{15}, x_{17}, x_{18}, x_{19}, x_{22}\}.$$

Then

$$\beta_{A,DMF}(X) = 0.75.$$

In addition, one can find that

$$\underline{apr}_A(X) = \{x_{11}, x_{18}\},$$
$$\underline{apr}_A(X) = \{x_1, x_4, x_5, x_6, x_7, x_9, x_{11}, x_{13}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{22}\}.$$

Then

$$\beta_A(X) = 0.5.$$

Proposition 4.4. *In the information subtable family described by Eq. (6), $A \subseteq AT$ is a subset of attributes. For each $X \subseteq OB$,*

$$\beta_A(X) \le \beta_{A,DMF}(X).$$

Proposition 4.4 and Example 4.3 show that more samples can be accurately classified based on the DMF strategy.

Table 7
An decision information table about 24 students.

OB	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	1	1	1	0	1	0	1
x_2	1	1	0	0	1	0	2
x_3	0	0	0	0	1	0	3
x_4	1	0	0	0	0	0	4
x_5	0	0	0	1	1	0	2
x_6	1	1	1	0	1	0	1
x_7	1	0	0	0	0	0	4
x_8	0	0	0	0	0	1	4
x_9	1	0	0	0	1	0	3
x_{10}	0	1	1	0	1	0	3
x_{11}	0	0	0	0	0	0	4
x_{12}	1	1	0	1	0	0	3
x_{13}	1	1	1	0	1	0	1
x_{14}	0	1	1	0	1	0	2
x_{15}	0	0	0	1	1	0	3
x_{16}	1	0	0	0	0	0	4
x_{17}	0	0	0	1	1	0	3
x_{18}	0	0	0	0	0	0	4
x_{19}	1	0	0	0	1	0	2
x_{20}	1	0	0	1	0	0	4
x_{21}	1	1	1	0	0	0	2
x_{22}	0	0	0	1	1	0	3
x_{23}	1	0	0	1	0	0	4
x_{24}	0	0	0	0	0	1	4

Table 8
An decision information subtable about grade 1.

U^1	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	1	1	1	0	1	0	1
x_2	1	1	0	0	1	0	2
x_3	0	0	0	0	1	0	3
x_4	1	0	0	0	0	0	4
x_5	0	0	0	1	1	0	2
x_6	1	1	1	0	1	0	1
x_7	1	0	0	0	0	0	4
x_8	0	0	0	0	0	1	4

For the decision information table described by Eq. (4),

$$\mathcal{DI} = \{DI_h = (U_h, AT \cup \{d\}, \{V_{a,h}|a \in AT\} \cup \{V_{d,h}\}, \{f_{a,h}|a \in AT\} \cup \{f_{d,h}\})|h = 1,2,\cdots,k\} \tag{9}$$

is called the decision information subtable family of the decision information table, where $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$ is a partition on OB.

Definition 4.4. In the decision information subtable family described by Eq. (9), $OB/E_{\{d\}} = \{D_1, D_2, \cdots, D_s\}$ is a partition on universe OB. Then

$$\gamma_{A,DMF}(d) = \frac{|\cup_{j=1}^{s} \underline{apr}_{A,DMF}(D_j)|}{|OB|} \tag{10}$$

is called the accuracy of the decision rule related to the decision attribute d.

Next, we will explain the accuracy of the decision rule proposed in Definition 4.4 through a specific example.

Example 4.4 *(Continued from Example 3.1)*. By adding a decision attribute d to Tables 1–4, we can obtain four decision information tables, shown in Tables 7–10. Here, attribute d represents “scholarship level”. And $V_d = \{1,2,3,4\}$, where $f_d(x) = i$ indicates that student x has won the i-class scholarship, $i = 1,2,3$. $f_d(x) = 4$ means student x did not receive any scholarships.

Let $D_j = \{x \in OB | f_d(x) = j\}, j = 1,2,3,4$. Then we have

$$D_1 = \{x_1, x_6, x_{13}\},$$

$$D_2 = \{x_2, x_5, x_{14}, x_{19}, x_{21}\},$$

Table 9
An decision information subtable about grade 2.

U^2	a_1	a_2	a_3	a_4	a_5	a_6	d
x_9	1	0	0	0	1	0	3
x_{10}	0	1	1	0	1	0	3
x_{11}	0	0	0	0	0	0	4
x_{12}	1	1	0	1	0	0	3
x_{13}	1	1	1	0	1	0	1
x_{14}	0	1	1	0	1	0	2
x_{15}	0	0	0	1	1	0	3
x_{16}	1	0	0	0	0	0	4

Table 10
An decision information subtable about grade 3.

U^3	a_1	a_2	a_3	a_4	a_5	a_6	d
x_{17}	0	0	0	1	1	0	3
x_{18}	0	0	0	0	0	0	4
x_{19}	1	0	0	0	1	0	2
x_{20}	1	0	0	1	0	0	4
x_{21}	1	1	1	0	0	0	2
x_{22}	0	0	0	1	1	0	3
x_{23}	1	0	0	1	0	0	4
x_{24}	0	0	0	0	0	1	4

$$D_3 = \{x_3, x_9, x_{10}, x_{12}, x_{15}, x_{17}, x_{22}\},$$

$$D_4 = \{x_4, x_7, x_8, x_{11}, x_{16}, x_{18}, x_{20}, x_{23}, x_{24}\}.$$

Suppose that $A = \{a_2, a_3, a_4, a_5\}$, We can obtain the deterministic decision rules based on the decision information Table 7 as follows:

$$\begin{cases} r_1 : (a_2,1) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,1) \Rightarrow (d,2) & \text{related to } x_2, \\ r_2 : (a_2,1) \wedge (a_3,0) \wedge (a_4,1) \wedge (a_5,0) \Rightarrow (d,3) & \text{related to } x_{12}, \\ r_3 : (a_2,1) \wedge (a_3,1) \wedge (a_4,0) \wedge (a_5,0) \Rightarrow (d,2) & \text{related to } x_{21}, \\ r_4 : (a_2,0) \wedge (a_3,0) \wedge (a_4,1) \wedge (a_5,0) \Rightarrow (d,4) & \text{related to } x_{20}, x_{23}, \\ r_5 : (a_2,0) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,0) \Rightarrow (d,2) & \text{related to } x_4, x_7, x_8, x_{11}, x_{16}, x_{18}, x_{24}. \end{cases}$$

Then, according to Definition 2.3, we have

$$\gamma_A(d) = \frac{|\cup_{i=1}^{4} \underline{apr_A}(D_i)|}{|OB|} = 0.5.$$

Next, we can get three groups of deterministic decision rules $\{r_1^1, r_2^1, r_3^1, r_4^1, r_5^1\}$, $\{r_1^2, r_2^2, r_3^2, r_4^2\}$ and $\{r_1^3, r_2^3, r_3^3, r_4^3, r_5^3\}$, which are induced from Tables 8–10, respectively.

$$\begin{cases} r_1^1 : (a_2,1) \wedge (a_3,1) \wedge (a_4,0) \wedge (a_5,1) \Rightarrow (d,1) & \text{related to } x_1, x_6, \\ r_2^1 : (a_2,1) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,1) \Rightarrow (d,2) & \text{related to } x_2, \\ r_3^1 : (a_2,0) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,1) \Rightarrow (d,3) & \text{related to } x_3, \\ r_4^1 : (a_2,0) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,0) \Rightarrow (d,4) & \text{related to } x_4, x_7, x_8, \\ r_5^1 : (a_2,0) \wedge (a_3,0) \wedge (a_4,1) \wedge (a_5,1) \Rightarrow (d,2) & \text{related to } x_5. \end{cases}$$

$$\begin{cases} r_1^2 : (a_2,0) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,1) \Rightarrow (d,3) & \text{related to } x_9, \\ r_2^2 : (a_2,0) \wedge (a_3,0) \wedge (a_4,0) \wedge (a_5,0) \Rightarrow (d,4) & \text{related to } x_{11}, x_{16}, \\ r_3^2 : (a_2,1) \wedge (a_3,0) \wedge (a_4,1) \wedge (a_5,0) \Rightarrow (d,3) & \text{related to } x_{12}, \\ r_4^2 : (a_2,0) \wedge (a_3,0) \wedge (a_4,1) \wedge (a_5,1) \Rightarrow (d,3) & \text{related to } x_{15}. \end{cases}$$

$$\begin{cases} r_1^3: (a_2,0)\wedge(a_3,0)\wedge(a_4,1)\wedge(a_5,1)\Rightarrow(d,3) & \text{related to } x_{17}, x_{22},\\ r_2^3: (a_2,0)\wedge(a_3,0)\wedge(a_4,0)\wedge(a_5,0)\Rightarrow(d,4) & \text{related to } x_{18}, x_{24},\\ r_3^3: (a_2,0)\wedge(a_3,0)\wedge(a_4,0)\wedge(a_5,1)\Rightarrow(d,2) & \text{related to } x_{19},\\ r_4^3: (a_2,0)\wedge(a_3,0)\wedge(a_4,1)\wedge(a_5,0)\Rightarrow(d,4) & \text{related to } x_{20}, x_{23},\\ r_5^3: (a_2,1)\wedge(a_3,1)\wedge(a_4,0)\wedge(a_5,0)\Rightarrow(d,2) & \text{related to } x_{21}. \end{cases}$$

Based on Definition 4.4, one can find

$$\gamma_{A,DMF}(d)=\frac{|\cup_{j=1}^{4}\underline{apr}_{A,DMF}(D_j)|}{|OB|}=0.875.$$

For the accuracy of decision-making, the following important conclusion results from Definitions 2.5, 4.4 and Example 4.4.

Proposition 4.5. *In the decision information subtable family described by Eq. (9), $A\subseteq AT$ is a subset of attributes. Then we have*

$$\gamma_A(d)\le\gamma_{A,DMF}(d).$$

Proposition 4.5 and Example 4.4 show that we can significantly improve the reliability of decisions using the DMF strategy.

5. Feature selection based on DMF strategy

Feature selection is widely studied and applied in machine learning, data mining, decision analysis, and other fields. It reduces the dimension and complexity of data and avoids over-fitting. In an information table, we can regard the attributes of sample data as the features. So, feature selection is often called attribute reduction.

5.1. Attribute reduction based on DMF strategy

This part, based on the DMF strategy, develops three types of attribute reductions. Here, we only explain how to use the DMF strategy to construct attribute reduction concerning the information subtable family. The other two reductions will appear directly.

Step 1 (Division): First, an information table described by Eq. (1) contains multiple information subtables and provides the information subtable family shown by Eq. (6).

Step 2 (Mining): Second, in the hth information subtable $I_h=(U_h,AT,\{V_{a,h}|a\in AT\},\{f_{a,h}|a\in AT\})$, $A'_h\subseteq A$ must satisfy:

(1) $U_h/E_{A'_h}=U_h/E_A$;

(2) For any $a\in A'_h$, $U_h/E_{A'_h-\{a\}}\neq U_h/E_A$.

Then, A'_h is a reduction of A with respect to subtable I_h. Based on the k subtables, we can get k reductions of the attribute set A.

Step 3 (Fusion): Third, we fuse the k reductions obtained in Step 2. That is, we need to find an attribute subset $A'\subseteq A$ that satisfies:

(1) For any $U\in\mathcal{P}$, $U/E_{A'}=U/E_A$;

(2) For any $a\in A'$ and at least one subtable I_h such that $U_h/E_{A'-\{a\}}\neq U_h/E_A$.

Then, based on the DMF strategy, the reduction concerning the information subtable family will be defined as follows.

Definition 5.1. In the information subtable family described by Eq. (6), if $A'\subseteq A$ satisfies:

(1) For any $U\in\mathcal{P}$, $U/E_{A'}=U/E_A$,

(2) For any $A''\subset A'$, there exists $U\in\mathcal{P}$, such that $U/E_{A''}\neq U/E_A$,

then A' is called the reduction of A with respect to the information subtable family. And we use $Reduct(A)_{R,DMF}$ to denote A', i.e., $Reduct(A)_{R,DMF}=A'$. Meanwhile, the set of all the reductions of A with respect to the information subtable family is denoted by $REDUCT(A)_{R,DMF}$.

Definition 5.2. In the information subtable family described by Eq. (6), X is a subset of the universe OB. If $A'\subseteq A$ satisfies:

(1) $\underline{apr}_{A',DMF}(X)=\underline{apr}_{A,DMF}(X)$,

(2) For any $A''\subset A'$, $\underline{apr}_{A'',DMF}(X)\neq\underline{apr}_{A,DMF}(X)$,

then A' is called the reduction of A with respect to lower approximation set $\underline{apr}_{A,DMF}(X)$. We use $Reduct(A)_{L,DMF}$ to denote A', i.e., $Reduct(A)_{L,DMF} = A'$. And the set of all the reductions of A with respect to the lower approximation set $\underline{apr}_{A,DMF}(X)$ is denoted by $REDUCT(A)_{L,DMF}$.

Definition 5.3. In the decision information subtable family described by Eq. (9), $OB/E_{\{d\}} = \{D_1, D_2, \cdots, D_s\}$ is a partition on universe OB. If $A' \subseteq A$ satisfies:

(1) $Pos_{A',DMF}(d) = Pos_{A,DMF}(d)$,

(2) For any $A'' \subset A'$, $Pos_{A'',DMF}(d) \neq Pos_{A,DMF}(d)$,

then A' is called the reduction of A with respect to $Pos_{A,DMF}(d)$. We use $Reduct(A)_{P,DMF}$ to denote A', i.e., $Reduct(A)_{P,DMF} = A'$. In addition, the set of all the reductions of A with respect to $Pos_{A,DMF}(d)$ is denoted by $REDUCT(A)_{P,DMF}$, where $Pos_{A,DMF}(d) = \cup_{j=1}^{s} \underline{apr}_{A,DMF}(D_j)$.

5.2. The properties of attribute reduction based on DMF strategy

Here, by employing a few examples, we will deeply explore the attribute reductions based on DMF strategy and obtain some important results.

Example 5.1 *(Continued from Example 3.1).* Suppose that $A = AT$, based on Definition 5.1,

$$REDUCT(A)_{R,DMF} = \{\{a_1, a_2, a_3, a_4, a_6\}, \{a_1, a_3, a_4, a_5, a_6\}\}.$$

In addition, from Definition 2.5, we have

$$REDUCT(A)_R = \{AT\}.$$

Proposition 5.1. *In the information subtable family described by Eq. (6), $A \subseteq AT$ is an attribute subset. Then we have $Reduct(A)_{R,DF} \in REDUCT(A)_{R,DF}$ and $Reduct(A)_R \in REDUCT(A)_R$ such that*

$$Reduct(A)_{R,DMF} \subseteq Reduct(A)_R.$$

Example 5.2 *(Continued from Example 3.1).* Let $X = \{x_1, x_4, x_6, x_7, x_9, x_{10}\}$, then

$$\underline{apr}_{A,DMF} = \{x_4, x_6, x_7, x_9, x_{10}\}.$$

Based on Definition 5.2,

$$REDUCT(A)_{L,DMF} = \{\{a_1, a_2, a_3, a_4, a_6\}, \{a_1, a_3, a_4, a_5, a_6\}\}.$$

In addition, according to Definition 2.6, we have

$$REDUCT(A)_L = \{AT\}.$$

Proposition 5.2. *In the information subtable family described by Eq. (6), $A \subseteq AT$ is an attribute subset, and $X \subseteq OB$ is an object subset. Then, there exist $Reduct(A)_{L,DMF} \in REDUCT(A)_{L,DMF}$ and $Reduct(A)_L \in REDUCT(A)_L$ such that*

$$Reduct(A)_{L,DMF} \subseteq Reduct(A)_L.$$

Example 5.3 *(Continued from Example 4.4).* Let $A = AT$, based on Definition 5.3,

$$REDUCT(A)_{P,DMF} = \{\{a_1, a_3, a_4, a_5, a_6\}\}.$$

In addition, according to Definition 2.7, we have

$$REDUCT(A)_P = \{AT\}.$$

Proposition 5.3. *In the decision information subtable family described by Eq. (9), $A \subseteq AT$ is an attribute subset. Then, there exist $Reduct(A)_{P,DMF} \in REDUCT(A)_{P,DMF}$ and $Reduct(A)_P \in REDUCT(A)_P$ such that*

$$Reduct(A)_{P,DMF} \subseteq Reduct(A)_P.$$

6. Algorithms

We have proven that the DMF strategy has more advantages than traditional methods. Next, we design four algorithms to calculate RSM and attribute reduction based on the DMF strategy. Then, we will compare the time complexities of these algorithms with those of the corresponding traditional algorithms.

Definition 2.1 details the Pawlak RSM. Once proposed, this model has attracted considerable attention. In Section 4, we design an RSM based on the DMF strategy. Here, we develop Algorithm 1 to calculate the approximations of this RSM.

Algorithm 1: An algorithm for computing $\underline{apr}_{A,DMF}(X)$ and $\overline{apr}_{A,DMF}(X)$.

```
Input  : An information table described by Eq. (1), a partition $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$, an attribute subset $A$, and an object subset $X \subseteq OB$;
Output : Two approximation sets $\underline{apr}_{A,DMF}(X)$ and $\overline{apr}_{A,DMF}(X)$.
begin
    $\underline{apr}_{A,DMF}(X) \leftarrow \emptyset$, $\overline{apr}_{A,DMF}(X) \leftarrow \emptyset$;
    for $h = 1 : k; h <= k; h++$ do
        Computing $U_h/E_{A,h} = \{[x]_{A,h} | x \in U_h\}$;  //where $[x]_{A,h} = [x]_A \cap U_h$
        Computing $\underline{apr}_{A,h}(X)$,  $\overline{apr}_{A,h}(X)$;
        $\underline{apr}_{A,DMF}(X) \leftarrow \underline{apr}_{A,DMF}(X) \cup \underline{apr}_{A,h}(X)$,  $\overline{apr}_{A,DMF}(X) \leftarrow \overline{apr}_{A,DMF}(X) \cup \overline{apr}_{A,h}(X)$;
    end
end
```

Section 2 reviews three attribute reductions in Pawlak's rough set theory. Researchers have deeply studied these three attribute reductions and have widely applied them to various data problems. Section 5 generalizes these classical attribute reductions and obtains three new ones based on the DMF strategy. Next, we will develop three algorithms to calculate these reductions based on the DMF strategy.

Algorithm 2: An algorithm for computing $Reduct(A)_{R,DMF}$.

```
Input  : An information table described by Eq. (1), a partition $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$ and an attribute subset $A = \{a_1, a_2, \cdots, a_l\} \subseteq AT$;
Output : The reduction $Reduct(A)_{R,DMF}$.
begin
    $A \leftarrow Reduct(A)_{R,DMF}$;
    for $i = 1 : l; i <= l; i++$ do
        If $(U_1/E_{Reduct(A)_{R,DMF}} = U_1/E_{Reduct(A)_{R,DMF}/\{a_i\}}) \wedge (U_2/E_{Reduct(A)_{R,DMF}} = U_2/E_{Reduct(A)_{R,DMF}/\{a_i\}}) \wedge \cdots \wedge (U_k/E_{Reduct(A)_{R,DMF}} = U_k/E_{Reduct(A)_{R,DMF}/\{a_i\}})$,
        then $Reduct(A)_{R,DMF} \leftarrow Reduct(A)_{R,DMF}/\{a_i\}$;
        Otherwise $i \leftarrow i + 1$;
    end
end
```

Algorithm 3: An algorithm for computing $Reduct(A)_{L,DMF}$.

```
Input  : An information table $\mathcal{DI} = (OB, AT \cup \{d\}, \{V_a | a \in AT\} \cup \{V_{d,h}\}, \{f_a | a \in AT\} \cup \{f_{d,h}\})$, a partition $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$, an attribute subset
         $A = \{a_1, a_2, \cdots, a_l\} \subseteq AT$ and an object subset $X$;
Output : The reduction $Reduct(A)_{L,DMF}$.
begin
    $A \leftarrow Reduct(A)_{L,DMF}$;
    for $i = 1 : l; i <= l; i++$ do
        If $\underline{apr}_{Reduct(A)_{L,DMF}/\{a_i\}}(X) = \underline{apr}_{Reduct(A)_{L,DMF}}(X)$
        then $Reduct(A)_{L,DMF} \leftarrow Reduct(A)_{L,DMF}/\{a_i\}$;
        Otherwise $i \leftarrow i + 1$;
    end
end
```

Algorithm 4: An algorithm for computing $Reduct(A)_{P,DMF}$.

```
Input  : An information table $I = (OB, AT, \{V_a | a \in AT\}, \{f_a | a \in AT\})$, a partition $\mathcal{P} = \{U_1, U_2, \cdots, U_k\}$, an attribute subset $A = \{a_1, a_2, \cdots, a_l\} \subseteq AT$, a
         decision attribute $d$;
Output : The reduction $Reduct(A)_{P,DMF}$.
begin
    $A \leftarrow Reduct(A)_{P,DMF}$;
    for $i = 1 : l; i <= l; i++$ do
        Computing $OB/E_{\{d\}} = \{D_1, D_2, \cdots, D_s\}$
        Computing $pos_{A,DMF}(d)$
        If $pos_{Reduct(A)_{P,DMF}/\{a_i\}}(d) = pos_{Reduct(A)_{P,DMF}}(d)$
        then $Reduct(A)_{P,DMF} \leftarrow Reduct(A)_{P,DMF}/\{a_i\}$;
        Otherwise $i \leftarrow i + 1$;
    end
end
```

Table 11
The time complexity of Algorithms.

Algorithms	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4
DMF strategy	$O\left(\|OB\|^2/k\right)+O(\|X\|\|OB\|)$	$O\left(l\times\|OB\|^2/k\right)+O(l\times\|OB\|)$	$O\left(l\times\|OB\|^2/k\right)+O(l\times\|X\|\|OB\|)$	$O\left(l\times\|OB\|^2/k\right)+O\left(l\times\|OB\|^2\right)$
Traditional methods	$O\left(\|OB\|^2\right)+O(\|X\|\|OB\|)$	$O\left(l\times\|OB\|^2\right)+O(l\times\|OB\|)$	$O\left(l\times\|OB\|^2\right)+O(l\times\|X\|\|OB\|)$	$O\left(l\times\|OB\|^2\right)+O\left(l\times\|OB\|^2\right)$

Table 12
Specific information about the data sets.

No.s	Datasets	Objects	Attributes	Classes
1	Molecular Biology (Promoter Gene Sequences)	106	57	2
2	Autism Screening Adult	704	21	4
3	Statlog (German Credit Data)	1000	24	2
4	Semeion Handwritten Digit	1593	256	2
5	OPPORTUNITY Activity Recognition	2511	242	7
6	TTC-3600	3600	3209	6
7	Turkiye-student-evaluation-generic	5820	33	2
8	Gisette	6000	5000	2
9	Mushroom	8124	23	4

Table 13
Specific information about the operating environment.

Name	Model	Parameter
CPU	AMD Ryzen 7 4800H with Radeon Graphics	2.90 GHz
Platform	Python	3.9
System	Windows10	64 bit
Memory	SAMSUNG DDR4	16 GB; 2666 MHz
Hard Disk	Intel SSDPEKNW	512 GB

We now analyze the time complexity of these four algorithms designed based on the DMF strategy and compare it with the traditional algorithms in Pawlak rough set theory. See Table 11 for details.

Here, the letter k represents the number of information subtables. It is a crucial parameter in DMF strategy. Table 11 shows that the time complexity of algorithms related to the DMF strategy is significantly lower than that of traditional algorithms. Moreover, the complexity of Algorithms 1-4 decreases with the increase of parameter k.

7. Experimental analysis

The previous sections proved that the DMF strategy can analyze data better than traditional methods in RSM and feature selection. Here, numerical experiments verify the excellent data processing ability of the DMF strategy. We use nine data sets on UCI (http://archive.ics.uci.edu/ml/datasets.html). The details of these datasets are in Table 12. A private computer completed the experimental processes and results. Table 13 shows the experimental operating environment, including relevant parameters. As we all know, the datasets of the information subtable family are difficult to find on the Internet, so we randomly divide the information table into multiple subtables to obtain the information subtable family.

In this section, from the experimental perspective, we attempt to validate the advantages of the DMF strategy from three aspects: efficiency, accuracy, and dimensionality.

The number of information subtables has a crucial impact on the effectiveness of the DMF strategy. Therefore, we will conduct an experimental comparison under five cases: $k=1, k=5, k=10, k=15$, and $k=20$. Here, $k=1$ indicates that numerical experiments are conducted based on one information table. Thus, $k=1$ means all experimental results are obtained using traditional methods.

In addition, we select two statistical tests, namely the Wilcoxon and Friedman tests, to detect the performance of the DMF strategy. First, we use the Wilcoxon test for significance analysis and choose the significance level as 0.05. Then, compared with the traditional methods at $k=1$, we analyze the experimental results of the DMF strategy at $k=5,10,15$ and 20. Second, using the Friedman test with a significance level of 0.05, we study the significance of the experimental results at $k=1,5,10,15$ and 20.

7.1. *Efficiency of DMF strategy in analyzing data*

In this part, we will test that the DMF strategy can rapidly analyze data. We will conduct numerical experiments to verify the higher efficiency of the DMF strategy from the perspective of calculating approximations, $Reduct(A)_{R,DMF}$, $Reduct(A)_{L,DMF}$, and $Reduct(A)_{P,DMF}$ of RSM. The time consumptions of computing approximations, $Reduct(A)_{R,DMF}$, $Reduct(A)_{L,DMF}$, $Reduct(A)_{P,DMF}$ of RSM are denoted as TC AP, TC R_R, TC R_L, and TC R_P, respectively.

Table 14
Time consumption changes as parameter k increases.

TC	k	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	TC	k	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9
AP	1	0.1368	0.1845	0.1707	1.0220	1.5993	26.3656	0.5280	337.2549	0.9542	R_L	1	0.2297	0.6938	0.9824	3.1541	4.4392	365.6978	1.0757	3015.0409	1.9033
	5	0.1179	0.1650	0.1552	0.9035	1.3707	22.1383	0.5032	252.9411	0.9025		5	0.1663	0.6334	0.9209	2.6632	3.9625	272.6095	0.8822	1934.1578	1.5171
	10	0.1104	0.1586	0.1535	0.8731	1.2828	19.1383	0.4966	198.0275	0.8537		10	0.1554	0.5979	0.7533	2.3227	3.6246	213.6243	0.8335	1135.6826	1.3342
	15	0.1082	0.1543	0.1506	0.8409	1.1352	14.7706	0.4837	151.3023	0.7794		15	0.1226	0.5433	0.6216	2.1272	3.3358	158.9261	0.7797	651.5687	1.1370
	20	0.1057	0.1492	0.1495	0.8326	1.0471	11.7385	0.4785	113.4188	0.7438		20	0.1082	0.5184	0.5683	1.9669	3.1244	115.7070	0.7062	392.5054	1.0036
R_R	1	0.2126	0.6772	0.9824	3.0604	4.2563	419.6223	0.9117	4362.5567	1.6958	R_P	1	0.2353	0.6816	1.0035	3.2277	4.4548	387.5186	0.9256	3618.8491	1.9047
	5	0.1615	0.6146	0.9209	2.4227	3.8571	336.6789	0.8540	2763.0856	1.4861		5	0.1741	0.6299	0.9623	2.7705	3.9222	289.1821	0.8623	2259.5736	1.5214
	10	0.1486	0.5695	0.7533	2.1092	3.5126	273.8774	0.8067	1654.4826	1.3015		10	0.1599	0.5774	0.8455	2.4695	3.6558	225.0465	0.8233	1249.2508	1.3670
	15	0.1145	0.5131	0.6216	1.9852	3.2288	208.7672	0.7349	837.2549	0.9964		15	0.1370	0.5351	0.6454	2.2463	3.3343	165.4994	0.7558	709.304	1.0394
	20	0.1033	0.4958	0.5683	1.8377	3.0844	144.6338	0.6771	576.0784	0.9276		20	0.1158	0.5078	0.5227	2.0018	3.1269	101.7679	0.6801	471.0065	0.9342

Table I
P value of the Wilcoxon test.

Data	$(k=1,\,k=5)$	$(k=1,\,k=10)$	$(k=1,\,k=15)$	$(k=1,\,k=20)$
TC AP	<0.01	<0.01	<0.01	<0.01
TC R_R	<0.01	<0.01	<0.01	<0.01
TC R_L	<0.01	<0.01	<0.01	<0.01
TC R_P	<0.01	<0.01	<0.01	<0.01

Table II
Result of the Friedman test.

Data	Friedman value	χ_F^2	P value
TC AP	18.06	28.89	8.23×10^{-6}
TC R_R	22.5	36	2.89×10^{-7}
TC R_L	22.5	36	2.89×10^{-7}
TC R_P	22.5	36	2.89×10^{-7}

All data of time consumption for computing TC AP, TC R_R, TC R_L, and TC R_P are in Table 14. From Table 14, the time consumptions of the Algorithms 1-4 at $k=5{,}10{,}15$, and 20 are much lower than those of these algorithms at $k=1$. As the parameter k continues to increase, the time consumption of these algorithms monotonically decreases.

To distinguish the time consumption, we divide each of the nine datasets into ten equally sized parts, denoted as $U_1', U_2', \ldots, U_{10}'$. And $U_1, U_2, \ldots, U_{10}$ satisfy the following equations:

$$U_i = U_1' \cup U_2' \cup \cdots \cup U_i', i = 1, 2, \ldots, 10.$$

Without loss of generality, we randomly selected three of the nine datasets for experimental analysis. Fig. 4 shows that as the size of the datasets increases, all the time consumptions related to a different parameter k increase. But when $k=1$, the consumption time is always the longest. As the parameter k increases, the time consumption gradually decreases.

Next, we perform the Wilcoxon and Friedman tests for statistical significance analysis to confirm the higher efficiency of the DMF strategy. First, we conduct the Wilcoxon test on the data from Table 14. All P values shown in Table I are less than 0.01, far below the significance level of 0.05. It indicates that the efficiency of the DMF strategy in processing data is significantly higher than traditional methods. Second, using the Friedman test, we conduct a significance analysis of the data in Table 14. One can find that all P values in Table II are also lower than the significance level of 0.05. This result means there is a significant difference in the efficiency of the DMF strategy under different parameter values. That is, as the parameter k increases, the efficiency of the DMF strategy significantly improves.

7.2. Accuracy of DMF strategy in analyzing data

This section analyzes the accuracy of the DMF strategy in mining data. We will consider two important facts. First, the RSM based on the DMF strategy ($k>1$) has higher accuracy of knowledge mining than the classic Pawlak RSM ($k=1$). Second, the classification accuracies of classifiers induced by attribute reductions of algorithms for the DMF strategy are higher than those of classifiers of traditional algorithms.

(1) Accuracy of three measures based on the DMF strategy in RSM

Many fields, such as data mining and knowledge representation, require measures to objectively and scientifically evaluate the accuracy of acquired knowledge or decisions. In rough set theory, approximate description accuracy ($\alpha_A(X)$), knowledge classification accuracy ($\beta_A(X)$), and decision accuracy ($\gamma_A(X)$), shown in Definitions 2.5, 2.6, and 2.7, are three traditional measures, which characterize the performance of an RSM in knowledge representation, knowledge classification, and knowledge decision-making, respectively. This paper generalizes these traditional measures using the DMF strategy, and we propose three new measures

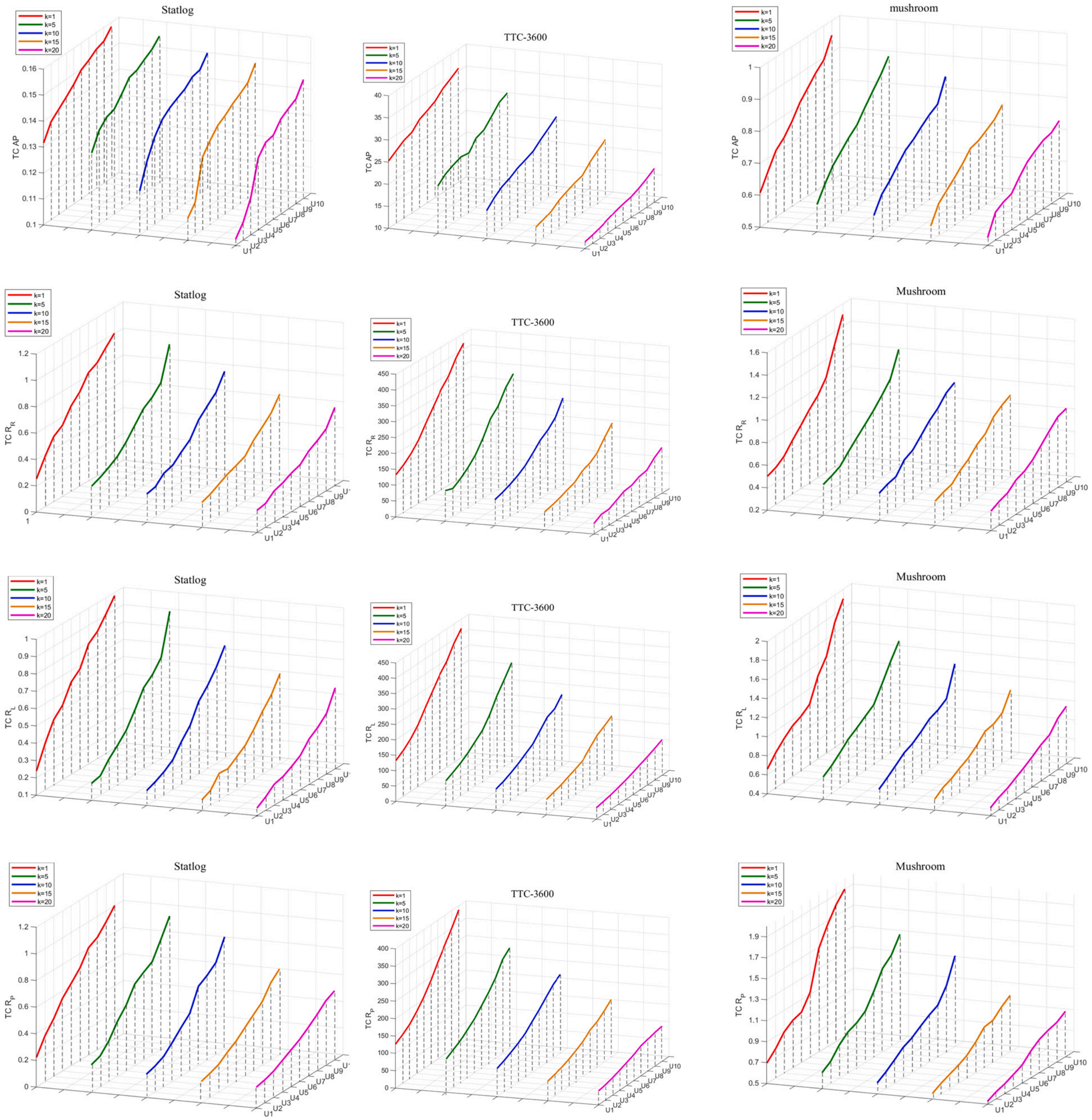

Fig. 4. Time consumption changes with the increase of the size of dataset.

$(\alpha_{A,DMF}(X), \beta_{A,DMF}(X), \gamma_{A,DMF}(X))$. We will compare these new measures with the traditional measures. The DMF strategy leads to more accurate results in knowledge mining.

Table 15 shows that three accuracies based on the DMF strategy are much higher than traditional accuracies. Moreover, with the increase of parameter k (the number of information subtables is increasing), these accuracies based on the DMF strategy rapidly increase.

Fig. 5 indicates that, as the number of attributes continues to increase, both accuracies based on the DMF strategy and traditional accuracies increase, but the ones based on the DMF strategy increase faster than traditional accuracies.

In addition, to detect the higher knowledge discovery ability of the DMF strategy, we select the Wilcoxon and Friedman tests for statistical significance analysis. First, the Wilcoxon test compares the data from Table 15. All P values shown in Table III are less than the significance level of 0.05. We infer that the knowledge discovery ability of the DMF strategy is significantly higher than that of traditional methods. Second, significance analysis is conducted on the data in Table 15 by employing the Friedman test. One can find that all P values in Table IV are also lower than the significance level of 0.05. This result shows a significant difference

Table 15
Accuracy changes as parameter k increases.

Accuracy	k	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9
$\alpha_{A,DMF}(X)$	1	0.3586	0.5478	0.8242	0.2859	0.3248	0.1748	0.2449	0.1684	0.2657
	5	0.7000	0.8141	0.9287	0.5979	0.4438	0.2874	0.2640	0.2451	0.3938
	10	0.8181	0.8637	0.9607	0.7366	0.5274	0.3756	0.2782	0.2997	0.5382
	15	0.8750	0.9354	0.9742	0.8610	0.6146	0.4197	0.2909	0.4071	0.6168
	20	0.9354	0.9377	0.9867	0.9147	0.6967	0.4739	0.3272	0.5844	0.7178
$\beta_{A,DMF}(X)$	1	0.4433	0.7585	0.8330	0.3716	0.6022	0.1758	0.6711	0.1698	0.6363
	5	0.8000	0.9171	0.9340	0.6522	0.7141	0.3103	0.6810	0.2126	0.7327
	10	0.8800	0.9414	0.9640	0.8049	0.7893	0.3772	0.7049	0.2674	0.8172
	15	0.9238	0.9739	0.9767	0.8820	0.8207	0.4258	0.7123	0.4659	0.8573
	20	0.9600	0.9742	0.9880	0.9303	0.8948	0.4783	0.7463	0.6455	0.9014
$\gamma_{A,DMF}(X)$	1	0.2735	0.5994	0.3250	0.4564	0.5069	0.1033	0.7855	0.2535	0.4830
	5	0.8000	0.8285	0.5990	0.6911	0.6559	0.5417	0.8097	0.3367	0.7018
	10	0.9200	0.9042	0.7090	0.7916	0.7284	0.7181	0.8408	0.4636	0.7241
	15	0.9248	0.9405	0.8010	0.8864	0.7682	0.8569	0.8544	0.5728	0.7449
	20	0.9800	0.9571	0.8240	0.9366	0.8439	0.8614	0.8771	0.7351	0.7545

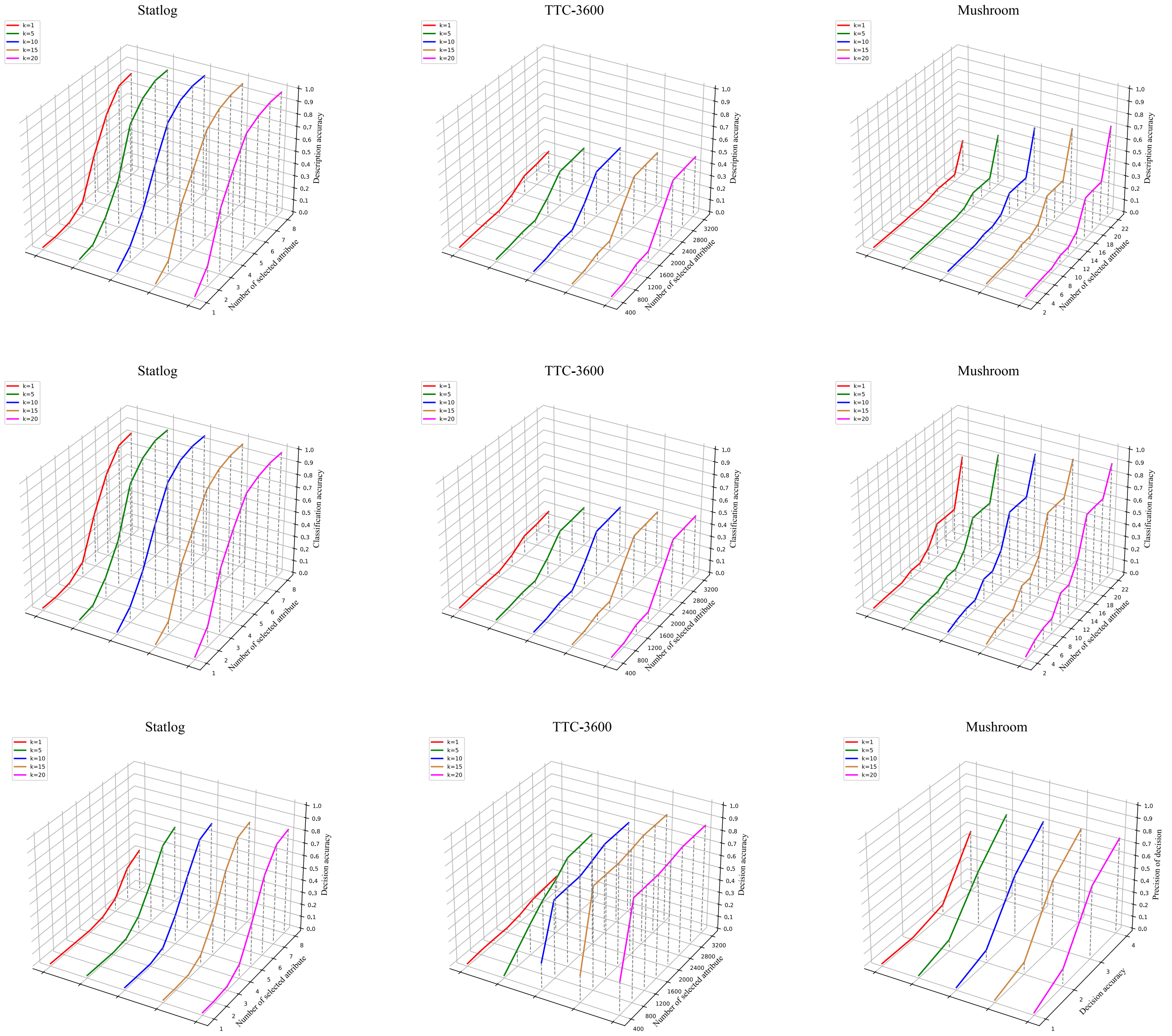

Fig. 5. Accuracies change as the number of attribute increases.

Table III
P value of the Wilcoxon test.

Data	$(k=1, k=5)$	$(k=1, k=10)$	$(k=1, k=15)$	$(k=1, k=20)$
$\alpha_{A,DMF}(X)$	< 0.01	< 0.01	< 0.01	< 0.01
$\beta_{A,DMF}(X)$	< 0.01	< 0.01	< 0.01	< 0.01
$\gamma_{A,DMF}(X)$	< 0.01	< 0.01	< 0.01	< 0.01

Table IV
Result of the Friedman test.

Data	Friedman value	χ_F^2	P value
$\alpha_{A,DMF}(X)$	22.06	35.29	4.05×10^{-7}
$\beta_{A,DMF}(X)$	22.06	35.29	4.05×10^{-7}
$\gamma_{A,DMF}(X)$	22.26	35.82	3.15×10^{-7}

Table 16
Classification accuracy of three classifiers.

Reductions	Classifiers	k	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9
$Reduct(A)_{R,DMF}$	SVM	1	0.5893 ± 0.1275	0.7228 ± 0.0637	0.6667 ± 0.0571	0.8025 ± 0.1764	0.7809 ± 0.1405	0.5847 ± 0.1792	0.7944 ± 0.2367	0.6383 ± 0.1356	0.5125 ± 0.1767
		5	0.7371 ± 0.2358	0.7985 ± 0.0743	0.6952 ± 0.1224	0.8563 ± 0.0987	0.8240 ± 0.2563	0.6433 ± 0.0879	0.8286 ± 0.2109	0.7196 ± 0.1752	0.6346 ± 0.0835
		10	0.8002 ± 0.2164	0.8247 ± 0.0982	0.7538 ± 0.1868	0.9183 ± 0.2638	0.8639 ± 0.2444	0.7081 ± 0.1138	0.8524 ± 0.2774	0.7932 ± 0.1437	0.6595 ± 0.1322
		15	0.8456 ± 0.2367	0.8576 ± 0.0774	0.7689 ± 0.1536	0.9607 ± 0.2249	0.9201 ± 0.1912	0.7861 ± 0.0643	0.8964 ± 0.1755	0.8559 ± 0.1005	0.7069 ± 0.0910
		20	0.9012 ± 0.2081	0.8714 ± 0.0546	0.8269 ± 0.1125	0.9881 ± 0.2708	0.9675 ± 0.1387	0.8787 ± 0.1322	0.9347 ± 0.1469	0.9145 ± 0.1178	0.8034 ± 0.1271
	KNN	1	0.6187 ± 0.1562	0.8088 ± 0.2318	0.7471 ± 0.1923	0.8088 ± 0.2318	0.7471 ± 0.1923	0.5571 ± 0.1241	0.8099 ± 0.0844	0.6185 ± 0.1292	0.5599 ± 0.2769
		5	0.7421 ± 0.1790	0.8382 ± 0.1535	0.7970 ± 0.2212	0.8382 ± 0.1535	0.7970 ± 0.2212	0.6239 ± 0.1087	0.8407 ± 0.1121	0.6981 ± 0.1608	0.6494 ± 0.1876
		10	0.7917 ± 0.1107	0.9030 ± 0.1444	0.8758 ± 0.2575	0.9030 ± 0.1444	0.8758 ± 0.2575	0.6825 ± 0.1381	0.8700 ± 0.1678	0.7677 ± 0.1199	0.6829 ± 0.2442
		15	0.8337 ± 0.0941	0.9334 ± 0.2118	0.9132 ± 0.1636	0.9334 ± 0.2118	0.9132 ± 0.1636	0.7628 ± 0.1169	0.9013 ± 0.1533	0.8351 ± 0.0911	0.7347 ± 0.3318
		20	0.8798 ± 0.1495	0.9684 ± 0.1712	0.9581 ± 0.1522	0.9684 ± 0.1712	0.9581 ± 0.1522	0.8651 ± 0.1437	0.9417 ± 0.1248	0.9064 ± 0.1285	0.8349 ± 0.1931
	C4.5	1	0.6165 ± 0.1462	0.6723 ± 0.1851	0.6752 ± 0.2076	0.7918 ± 0.2944	0.8249 ± 0.0739	0.5766 ± 0.1274	0.8236 ± 0.2696	0.6030 ± 0.1327	0.5975 ± 0.3273
		5	0.7405 ± 0.0979	0.8113 ± 0.2154	0.6911 ± 0.2387	0.8205 ± 0.2583	0.8577 ± 0.1226	0.6846 ± 0.1448	0.8595 ± 0.1779	0.6973 ± 0.1213	0.6639 ± 0.1688
		10	0.7815 ± 0.1570	0.8623 ± 0.1626	0.7453 ± 0.1823	0.8976 ± 0.2021	0.9173 ± 0.1417	0.7335 ± 0.1013	0.8743 ± 0.0988	0.7654 ± 0.1627	0.6944 ± 0.2294
		15	0.8431 ± 0.1062	0.8767 ± 0.2011	0.7785 ± 0.1930	0.9405 ± 0.1623	0.9411 ± 0.0695	0.7913 ± 0.1269	0.9059 ± 0.1167	0.8332 ± 0.1104	0.7466 ± 0.0752
		20	0.8796 ± 0.1425	0.9022 ± 0.1327	0.8259 ± 0.0923	0.9704 ± 0.1855	0.9761 ± 0.0531	0.9087 ± 0.1161	0.9480 ± 0.0822	0.8926 ± 0.1219	0.8365 ± 0.1365
$Reduct(A)_{L,DMF}$	SVM	1	0.6323 ± 0.1739	0.6763 ± 0.2141	0.6651 ± 0.1159	0.7939 ± 0.0982	0.8144 ± 0.1574	0.6138 ± 0.1423	0.8258 ± 0.2096	0.6485 ± 0.1327	0.5873 ± 0.2573
		5	0.7593 ± 0.2688	0.8061 ± 0.1336	0.6862 ± 0.2071	0.8318 ± 0.1677	0.8445 ± 0.2458	0.6979 ± 0.1448	0.8633 ± 0.1368	0.7231 ± 0.1608	0.6688 ± 0.1615
		10	0.7972 ± 0.1466	0.8578 ± 0.1579	0.7624 ± 0.1255	0.9021 ± 0.1394	0.9126 ± 0.1531	0.7615 ± 0.1089	0.8839 ± 0.1005	0.7973 ± 0.1910	0.6895 ± 0.0828
		15	0.8322 ± 0.1542	0.8939 ± 0.0755	0.7911 ± 0.1430	0.9439 ± 0.0587	0.9388 ± 0.0826	0.8153 ± 0.1305	0.9117 ± 0.0658	0.8505 ± 0.1724	0.7529 ± 0.1366
		20	0.8562 ± 0.0912	0.9255 ± 0.1076	0.8313 ± 0.1947	0.9667 ± 0.0631	0.9732 ± 0.1275	0.9018 ± 0.0827	0.9466 ± 0.0527	0.9389 ± 0.1285	0.8453 ± 0.0632
	KNN	1	0.6652 ± 0.2279	0.6537 ± 0.0967	0.6845 ± 0.2433	0.8018 ± 0.2326	0.7496 ± 0.1452	0.6289 ± 0.1174	0.8430 ± 0.1956	0.6421 ± 0.1357	0.6395 ± 0.3755
		5	0.7661 ± 0.1965	0.8031 ± 0.1582	0.7138 ± 0.2562	0.8370 ± 0.1654	0.7899 ± 0.0528	0.7055 ± 0.1436	0.8315 ± 0.0871	0.7123 ± 0.1263	0.7130 ± 0.2533
		10	0.7940 ± 0.0899	0.8678 ± 0.2239	0.7639 ± 0.1567	0.9010 ± 0.1661	0.8825 ± 0.0390	0.7628 ± 0.1092	0.8794 ± 0.1363	0.8066 ± 0.1575	0.7461 ± 0.2235
		15	0.8256 ± 0.1556	0.8950 ± 0.1167	0.7968 ± 0.0916	0.9362 ± 0.1734	0.9222 ± 0.0853	0.8349 ± 0.1247	0.9088 ± 0.0289	0.8747 ± 0.1006	0.7822 ± 0.1674
		20	0.8899 ± 0.0939	0.9165 ± 0.1604	0.8297 ± 0.1222	0.9658 ± 0.0889	0.9604 ± 0.0465	0.9189 ± 0.1375	0.9396 ± 0.0322	0.9516 ± 0.1493	0.8507 ± 0.1285
	C4.5	1	0.6475 ± 0.1536	0.6601 ± 0.1897	0.6871 ± 0.1359	0.8018 ± 0.2344	0.7500 ± 0.1704	0.6485 ± 0.1327	0.8448 ± 0.0868	0.6562 ± 0.1546	0.6573 ± 0.3422
		5	0.7658 ± 0.0879	0.8113 ± 0.0512	0.7174 ± 0.1783	0.8370 ± 0.1687	0.7937 ± 0.0824	0.7231 ± 0.1608	0.8295 ± 0.0752	0.7483 ± 0.1305	0.7261 ± 0.2681
		10	0.7964 ± 0.1059	0.8738 ± 0.1567	0.7655 ± 0.0736	0.9010 ± 0.1158	0.8966 ± 0.1273	0.7973 ± 0.1910	0.8802 ± 0.1344	0.8152 ± 0.1349	0.7545 ± 0.1838
		15	0.8167 ± 0.1211	0.8904 ± 0.1346	0.7882 ± 0.1488	0.9362 ± 0.1703	0.9278 ± 0.0977	0.8505 ± 0.1724	0.9065 ± 0.1179	0.8849 ± 0.1372	0.7918 ± 0.0835
		20	0.8881 ± 0.0533	0.9195 ± 0.0912	0.8311 ± 0.1141	0.9658 ± 0.0897	0.9677 ± 0.0764	0.9146 ± 0.1083	0.9359 ± 0.0471	0.9453 ± 0.1051	0.8543 ± 0.1396
$Reduct(A)_{P,DMF}$	SVM	1	0.6316 ± 0.1043	0.6568 ± 0.1315	0.6771 ± 0.1680	0.8037 ± 0.2072	0.7544 ± 0.1481	0.6443 ± 0.1878	0.8456 ± 0.0949	0.6692 ± 0.1675	0.6597 ± 0.1725
		5	0.7532 ± 0.1888	0.8138 ± 0.0966	0.7165 ± 0.1349	0.8398 ± 0.0695	0.7980 ± 0.1891	0.7169 ± 0.1513	0.8260 ± 0.1588	0.7486 ± 0.1358	0.7341 ± 0.2184
		10	0.7624 ± 0.0912	0.8746 ± 0.1001	0.7678 ± 0.1511	0.9077 ± 0.1370	0.9005 ± 0.0673	0.7966 ± 0.1046	0.8833 ± 0.1122	0.8273 ± 0.0934	0.7692 ± 0.1687
		15	0.8054 ± 0.1424	0.8957 ± 0.0379	0.7897 ± 0.0204	0.9394 ± 0.1484	0.9311 ± 0.1774	0.8447 ± 0.1362	0.9052 ± 0.0976	0.8823 ± 0.1411	0.8066 ± 0.1290
		20	0.8705 ± 0.0695	0.9140 ± 0.1067	0.8585 ± 0.0885	0.9642 ± 0.1282	0.9688 ± 0.1256	0.9393 ± 0.1178	0.9423 ± 0.0624	0.9657 ± 0.1096	0.8596 ± 0.0842
	KNN	1	0.6287 ± 0.0687	0.6690 ± 0.0937	0.6813 ± 0.1390	0.8061 ± 0.1255	0.7634 ± 0.0736	0.6389 ± 0.1391	0.8437 ± 0.1231	0.6517 ± 0.1872	0.6546 ± 0.1474
		5	0.7476 ± 0.1226	0.8173 ± 0.1106	0.7220 ± 0.0924	0.8383 ± 0.1620	0.8012 ± 0.0983	0.7037 ± 0.1164	0.8295 ± 0.1501	0.7368 ± 0.1676	0.7352 ± 0.1300
		10	0.7831 ± 0.1831	0.8804 ± 0.1338	0.7649 ± 0.0554	0.9111 ± 0.0997	0.8983 ± 0.1164	0.7948 ± 0.1533	0.8859 ± 0.0776	0.8127 ± 0.0894	0.7677 ± 0.1584
		15	0.8121 ± 0.1618	0.8988 ± 0.1036	0.7972 ± 0.1146	0.9369 ± 0.1250	0.9294 ± 0.1425	0.8228 ± 0.1653	0.9043 ± 0.1362	0.8656 ± 0.1378	0.8109 ± 0.0805
		20	0.8675 ± 0.0967	0.9195 ± 0.1240	0.8522 ± 0.1472	0.9601 ± 0.0535	0.9437 ± 0.0802	0.9234 ± 0.1438	0.9442 ± 0.1187	0.9591 ± 0.1353	0.8553 ± 0.1057
	C4.5	1	0.6334 ± 0.1216	0.6599 ± 0.0709	0.6762 ± 0.0976	0.8115 ± 0.1376	0.7589 ± 0.1060	0.6408 ± 0.1274	0.8524 ± 0.1613	0.6621 ± 0.1547	0.6483 ± 0.1198
		5	0.7488 ± 0.0532	0.8224 ± 0.1938	0.7179 ± 0.1124	0.8349 ± 0.1087	0.7954 ± 0.0971	0.7056 ± 0.1639	0.8277 ± 0.1367	0.7464 ± 0.1826	0.7273 ± 0.1597
		10	0.7772 ± 0.0824	0.8879 ± 0.1500	0.7611 ± 0.1363	0.9151 ± 0.0942	0.8936 ± 0.0630	0.8064 ± 0.1165	0.8873 ± 0.0231	0.8391 ± 0.1114	0.7581 ± 0.0952
		15	0.8124 ± 0.0254	0.9001 ± 0.1236	0.7904 ± 0.0597	0.9331 ± 0.0795	0.9305 ± 0.1870	0.8476 ± 0.1258	0.9107 ± 0.0708	0.8923 ± 0.1315	0.8075 ± 0.0390
		20	0.8705 ± 0.1166	0.9241 ± 0.1440	0.8467 ± 0.0623	0.9582 ± 0.1385	0.9414 ± 0.1559	0.9425 ± 0.1487	0.9485 ± 0.1293	0.9783 ± 0.1604	0.8489 ± 0.0989

in the knowledge discovery ability of the DMF strategy under different parameter values. Thus, as the parameter k increases, the knowledge discovery ability of the DMF strategy significantly improves.

(2) Classification accuracy of classifiers induced from feature selection based on DMF strategy

This paper explores three types of feature selections based on the DMF strategy. It studies the classification accuracies of three classifiers, namely support vector machine (SVM), K-Nearest Neighbor (KNN, K = 3), and decision tree ($C_{4.5}$). Traditionally, all samples are considered an information table, and the classification accuracy of the classifier is obtained based on a single information table. However, according to the DMF strategy, the information table contains multiple information subtables. Each information subtable can provide a classification accuracy. Therefore, the average of the classification accuracies for all information subtables defines the classification accuracy of the classifier induced by the DMF strategy. As shown in Table 16, we obtain each data using the 10-fold cross-validation method.

Table V
P value of the Wilcoxon test.

Data	$(k=1, k=5)$	$(k=1, k=10)$	$(k=1, k=15)$	$(k=1, k=20)$
$(Reduct(A)_{R,DMF}, SVM)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{R,DMF}, KNN)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{R,DMF}, C4.5)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{L,DMF}, SVM)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{L,DMF}, KNN)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{L,DMF}, C4.5)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{P,DMF}, SVM)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{P,DMF}, KNN)$	<0.01	<0.01	<0.01	<0.01
$(Reduct(A)_{P,DMF}, C4.5)$	<0.01	<0.01	<0.01	<0.01

Table VI
Result of the Friedman test.

Data	Friedman value	χ_F^2	P value
$(Reduct(A)_{R,DMF}, SVM)$	22.5	36	2.89×10^{-7}
$(Reduct(A)_{R,DMF}, KNN)$	22.06	35.29	4.05×10^{-7}
$(Reduct(A)_{R,DMF}, C4.5)$	22.26	35.82	3.15×10^{-7}
$(Reduct(A)_{L,DMF}, SVM)$	22.5	36	2.89×10^{-7}
$(Reduct(A)_{L,DMF}, KNN)$	22.5	36	2.89×10^{-7}
$(Reduct(A)_{L,DMF}, C4.5)$	22.06	35.29	4.05×10^{-7}
$(Reduct(A)_{P,DMF}, SVM)$	22.26	35.82	3.15×10^{-7}
$(Reduct(A)_{P,DMF}, KNN)$	22.5	36	2.89×10^{-7}
$(Reduct(A)_{P,DMF}, C4.5)$	22.5	36	2.89×10^{-7}

Table 17
The number of attributes.

Accuracy	k	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9
$Reduct(A)_{R,DMF}$	1	39	15	16	72	81	1279	28	2309	21
	5	32	13	14	50	75	820	24	1547	18
	10	25	12	13	44	61	532	21	1034	16
	15	22	12	13	39	57	488	19	765	15
	20	19	11	13	37	55	363	16	528	14
$Reduct(A)_{L,DMF}$	1	36	14	15	68	77	1053	25	1751	19
	5	29	13	13	46	68	685	20	1160	15
	10	22	11	11	38	54	464	17	753	13
	15	20	10	11	36	49	341	15	536	12
	20	17	10	10	33	46	282	13	422	12
$Reduct(A)_{P,DMF}$	1	33	16	17	65	72	1156	28	1978	20
	5	26	14	14	44	63	722	22	1276	17
	10	20	12	10	42	58	485	19	827	14
	15	18	11	10	38	54	389	17	593	13
	20	15	10	9	35	51	314	14	479	13

According to Table 16, all classification accuracies of the three classifiers under the DMF strategy are significantly higher than those of the classifiers under traditional methods. As the number of information subtables continues to increase, the classification accuracies of the three classifiers are also constantly improving.

Here, we select the Wilcoxon and Friedman tests for statistical significance analysis to prove the higher classification accuracy of the DMF strategy. First, we conduct the Wilcoxon test on the results from Table 16. All P values shown in Table V are less than the significance level of 0.05. Thus, the classification accuracy of the DMF strategy is significantly higher than that of traditional methods. Second, we analyze the results in Table 16 for significance through the Friedman test. It shows that all P values in Table VI are lower than the significance level of 0.05. Therefore, it confirms a significant difference in the classification accuracy of the DMF strategy under different parameter values. As the parameter k increases, the classification accuracy of the DMF strategy significantly improves.

7.3. Dimension of data based on DMF strategy

Here, we will analyze the dimension of samples in three types of attribute reduction based on the DMF strategy. Table 17 shows that, as the number of information subtables increases, the numbers of attributes in $Reduct(A)_{R,DMF}$, $Reduct(A)_{L,DMF}$ and $Reduct(A)_{P,DMF}$ fluctuates, but in general, there is a downward trend. We can derive the following two facts:

Table VII
P value of the Wilcoxon test.

Data	$(k=1, k=5)$	$(k=1, k=10)$	$(k=1, k=15)$	$(k=1, k=20)$
$Reduct(A)_{R,DMF}$	< 0.01	< 0.01	< 0.01	< 0.01
$Reduct(A)_{L,DMF}$	< 0.01	< 0.01	< 0.01	< 0.01
$Reduct(A)_{P,DMF}$	< 0.01	< 0.01	< 0.01	< 0.01

Table VIII
Result of the Friedman test.

Data	Friedman value	χ^2_F	P value
$Reduct(A)_{R,DMF}$	20.85	34.71	5.34×10^{-7}
$Reduct(A)_{L,DMF}$	21.53	35.23	4.17×10^{-7}
$Reduct(A)_{P,DMF}$	22.01	35.62	3.47×10^{-7}

Table 18
Conclusion on DMF strategy.

Experimental analysis	Conclusion
Efficiency	DMF strategy has higher efficiency in analyzing data than traditional methods.
Effectiveness	The accuracy of data analysis based on DMF strategy is higher than that of data analysis based on traditional methods.
Dimensions	In feature selection, DMF strategy can reduce the dimension of sample data more effectively than traditional methods.
Efficiency	When the parameter k increases, the efficiency of using DMF strategy to analyze data increases.
Effectiveness	When the parameter k increases, the accuracy of using DMF strategy to analyze data continuously improves.
Dimensions	As the parameter k increases, the dimension of the sample data in attribute reduction based on DMF strategy gradually decreases.

(1) Compared with traditional reduction methods, the DMF strategy can effectively delete redundant attributes and reduce the dimension and complexity of data.

(2) As the number of information subtables (data subsets) increases, the number of deleted attributes increases, and the dimension of the data decreases. Thus, the DMF strategy performs excellently in reducing data complexity and avoiding over-fitting problems.

Next, statistical significance analysis uses the Wilcoxon and Friedman tests to detect the lower data dimension. First, we conduct the Wilcoxon test on the results from Table 17 in sequence. All P values shown in Table VII are less than the significance level of 0.05. Thus, the data dimension based on the DMF strategy is significantly lower than that of traditional methods.

Second, we conduct a significance analysis of the results in Table 17 employing the Friedman test. All P values recorded in Table VIII are also lower than the significance level of 0.05. We conclude that there is a significant difference in the data dimension based on the DMF strategy under different parameter values. Thus, the data dimension based on the DMF strategy will significantly decrease with the increase of parameter k.

8. Conclusion and future work

Data has gradually replaced technology, capital, and manpower as a crucial factor of production. People increasingly rely on data for life, production, and decision-making. Due to the constraints of data sources, collection technologies, human cognition, and other factors, the forms of data are often complex and diverse. How to mine massive complex data to obtain useful knowledge to complete the established learning task is a research hot spot.

8.1. Conclusion

This article proposes a data analysis method called DMF strategy to reduce data complexity and improve data processing speed. It first develops an RSM based on the DMF strategy to verify the advantages of the DMF strategy. The RSM based on the DMF strategy performs better in efficiency and effectiveness than traditional RSMs. In addition, the DMF strategy is applied to feature selection theory. Compared with traditional methods, the DMF strategy has at least three advantages: First, one can select suitable attributes more quickly. Second, the classification accuracy of the classifiers will be higher. And third, the dimensionality of the data will be lower. Table 18 summarizes the results of all experimental analyses to understand the performance of the DMF strategy in processing data.

Of course, any data mining method has limitations and cannot effectively solve all data problems. The most critical technique of the DMF strategy is to divide a large-scale complex dataset into multiple small-scale and simpler sub-datasets. If the amount of data in a dataset is relatively small or the form of data labels is relatively simple, the performance of the DMF strategy will be severely compromised. Additionally, it may be impossible to divide a dataset into multiple sub-datasets for data mining. In this case, the DMF strategy becomes ineffective.

8.2. Future work

This paper develops an RSM from the DMF strategy for the first time. It fully confirms the advantages of this model from theoretical and experimental perspectives. Therefore, we can develop other RSMs based on the DMF strategy. For example, concerning the DMF strategy, we can construct and study probabilistic RSMs, multi-granulation RSMs, and covering RSMs. Furthermore, we can combine the DMF strategy with fuzzy set theory, three-way decision, and formal concept analysis to deal with complex data problems. We showed that the larger the parameter k, the better the performance of the DMF strategy in analyzing data. However, in practical circumstances, we cannot infinitely increase the parameter k. Therefore, when using the DMF strategy to handle practical data problems, it is necessary to study the optimal value of parameter k.

CRediT authorship contribution statement

Qingzhao Kong: Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation. **Wanting Wang:** Validation, Software, Investigation. **Weihua Xu:** Supervision, Project administration, Funding acquisition, Conceptualization. **Conghao Yan:** Validation, Software, Investigation.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (No. 62376229) and Natural Science Foundation of Chongqing (No. CSTB2023NSCQ-LZX0027).

References

[1] E. Eirola, A. Lendasse, V. Vandewalle, et al., Mixture of Gaussians for distance estimation with missing data, Neurocomputing 131 (2014) 32–42.
[2] Q. Yu, Y. Miche, E. Eirola, et al., Regularized extreme learning machine for regression with missing data, Neurocomputing 102 (2013) 45–51.
[3] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[4] B. Kovalerchuk, E. Triantaphyllou, J.F. Ruizet, Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lobulation, Artif. Intell. Med. 11 (1997) 75–85.
[5] W.T. Li, S.C. Zhai, W.H. Xu, et al., Feature selection approach based on improved fuzzy C-means with principle of refined justifiable granularity, IEEE Trans. Fuzzy Syst. 31 (2023) 2112–2126.
[6] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[7] W.H. Xu, M. Huang, Z.Y. Jiang, et al., Graph-based unsupervised feature selection for interval-valued information system, IEEE Trans. Neural Netw. Learn. Syst. (2023), https://doi.org/10.1109/TNNLS.2023.3263684.
[8] W.H. Xu, Z.T. Yuan, Z. Liu, Feature selection for unbalanced distribution hybrid data based on k-nearest neighborhood rough set, IEEE Trans. Artif. Intell. (2023), https://doi.org/10.1109/TAI.2023.3237203.
[9] Q.Z. Kong, W.T. Wang, D.X. Zhang, et al., Two kinds of average approximation accuracy, CAAI Trans. Intell. Technol. (2023), https://doi.org/10.1049/cit2.12222.
[10] Y.H. Qian, X.Y. Liang, Q. Wang, et al., Local rough set: a solution to rough data analysis in big data, Int. J. Approx. Reason. 97 (2018) 38–63.
[11] Q.Z. Kong, X.E. Chang, Rough set model based on variable universe, CAAI Trans. Intell. Technol. 7 (2022) 503–511.
[12] J.B. Zhang, T.R. Li, D. Ruan, et al., A parallel method for computing rough set approximations, Inf. Sci. 194 (2012) 209–223.
[13] S.Y. Li, T.R. Li, A parallel matrix-based approach for computing approximations in dominance-based rough sets approach, in: 9th International Conference on Rough Sets and Knowledge Technology (RSKT), Shanghai, China, 2014.
[14] J.B. Zhang, J.S. Wong, Y. Pan, et al., A parallel matrix-based method for computing approximations in incomplete information systems, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 326–339.
[15] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341–353.
[16] Q.Z. Kong, X.W. Zhang, W.H. Xu, et al., A novel granular computing model based on three-way decision, Int. J. Approx. Reason. 144 (2022) 92–112.
[17] Y.Y. Yao, Three-way conflict analysis: reformulations and extensions of the Pawlak model, Knowl.-Based Syst. 180 (2019) 26–37.
[18] V. Srishti, S. Seba, Sentiment cognition from words shortlisted by fuzzy entropy, IEEE Trans. Cogn. Dev. Syst. 12 (2020) 541–550.
[19] M. Hu, Y.T. Chen, D.G. Chen, et al., Attribute reduction based on neighborhood constrained fuzzy rough sets, Knowl.-Based Syst. 274 (2023) 110632.
[20] P. Wang, J.L. He, Z.W. Li, Attribute reduction for hybrid data based on fuzzy rough iterative computation model, Inf. Sci. 632 (2023) 555–575.
[21] W.B. Qian, S.D. Yu, J. Yang, et al., Multi-label feature selection based on information entropy fusion in multi-source decision system, Evol. Intell. 13 (2020) 255–268.

[22] O.O. Aremu, R.A. Cody, D. Hyland-Wood, et al., A relative entropy based feature selection framework for asset data in predictive maintenance, Comput. Ind. Eng. 145 (2020) 106536.

[23] L. Sun, et al., Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy data, IEEE Trans. Fuzzy Syst. 29 (2021) 19–33.

[24] W.H. Xu, D.D. Guo, J.S. Mi, et al., Two-way concept-cognitive learning via concept movement viewpoint, IEEE Trans. Neural Netw. Learn. Syst. 34 (10) (2023) 6798–6812, https://doi.org/10.1109/TNNLS.2023.3235800.

[25] W.H. Xu, D.D. Guo, Y.H. Qian, et al., Two-way concept-cognitive learning method: a fuzzy-based progressive learning, IEEE Trans. Fuzzy Syst. 31 (2023) 1885–1899.

[26] W.H. Xu, Y.Z. Pan, X.W. Chen, et al., A novel dynamic fusion approach using information entropy for interval-valued ordered datasets, IEEE Trans. Big Data 9 (2023) 845–859.

[27] W.H. Xu, K.H. Yuan, W.T. Li, et al., An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution, IEEE Trans. Emerg. Top. Comput. Intell. 7 (2022) 76–88.

[28] B.B. Sang, H.M. Chen, L. Yang, et al., Incremental feature selection using a conditional entropy based on fuzzy dominance neighborhood rough sets, IEEE Trans. Fuzzy Syst. 30 (2022) 1683–1697.

[29] W.P. Ding, T.Z. Qin, X.J. Shen, et al., Parallel incremental efficient attribute reduction algorithm based on attribute tree, Inf. Sci. 610 (2022) 1102–1121.

[30] Y. Yang, Z.R. Chen, Z. Liang, et al., Attribute reduction for massive data based on rough set theory and MapReduce, in: 5th International Conference on Rough Set and Knowledge Technology (RSKT), Beijing, China, 2010.

[31] H.M. Chen, T.R. Li, Y. Cai, et al., Parallel attribute reduction in dominance-based neighborhood rough set, Inf. Sci. 373 (2016) 351–368.

[32] C. Luo, S.Z. Wang, T.R. Li, et al., Spark rough hypercuboid approach for scalable feature selection, IEEE Trans. Knowl. Data Eng. 35 (2023) 3130–3144.

[33] G.Y. Dai, T.B. Jiang, Y.L. Mu, et al., A novel rough sets positive region based parallel multi-reduction algorithm, in: 4th International Conference on Advanced Intelligent Systems and Informatics (AISI), Cairo, Egypt, 2018.

[34] X.Y. Zhang, J.L. Hou, A novel rough set method based on adjustable-perspective dominance relations in intuitionistic fuzzy ordered decision tables, Int. J. Approx. Reason. 154 (2023) 218–241.

[35] S.Y. Xia, Y.S. Liu, X. Ding, et al., Granular ball computing classifiers for efficient, scalable and dobust learning, Inf. Sci. 483 (2019) 136–152.

[36] Z. Pawlak, Information systems theoretical foundations, Inf. Syst. 6 (1981) 205–218.

[37] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Boston, 1991.

[38] C.Z. Wang, Y. Wang, M.W. Shao, et al., Fuzzy rough attribute reduction for categorical data, IEEE Trans. Fuzzy Syst. 28 (2020) 818–830.

[39] W.H. Xu, Y.T. Guo, Generalized multigranulation double-quantitative decision-theoretic rough set, Knowl.-Based Syst. 105 (2016) 190–205.

[40] Q.Z. Kong, W.H. Xu, D.X. Zhang, A comparative study of different granular structures induced from the information systems, Soft Comput. 26 (2022) 105–122.

[41] X.L. Yang, H.M. Chen, H. Wang, et al., Feature selection with local density-based fuzzy rough set model for noisy data, IEEE Trans. Fuzzy Syst. 31 (2023) 1614–1627.

[42] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC Press, 2013.

[43] W.X. Zhang, W.Z. Wu, J.Y. Liang, et al., Rough Set Theory and Method, Science Press, Beijing, 2001.

[44] Z.H. Wang, H.M. Chen, X.L. Yang, et al., Fuzzy rough dimensionality reduction: a feature set partition-based approach, Inf. Sci. (2023), https://doi.org/10.1016/j.ins.2023.119266.

[45] Z.J. Guo, Y. Shen, T. Yang, et al., Semi-supervised feature selection based on fuzzy related family, Inf. Sci. 652 (2023), https://doi.org/10.1016/j.ins.2023.119660.

[46] H.Q. Zhang, X. Yu, T.R. Li, et al., Noise-aware and correlation analysis-based for fuzzy-rough feature selection, Inf. Sci. 659 (2024) 120047.