



# An incremental attribute reduction approach based on knowledge granularity for incomplete decision systems

Chuai Zhang<sup>1</sup> · Jianhua Dai<sup>1</sup>

Received: 16 January 2019 / Accepted: 6 May 2019  
© Springer Nature Switzerland AG 2019

## Abstract

Attribute reduction is a core issue in rough set theory. In recent years, with the fast development of data processing tools, information systems may increase quickly in objects over time. How to update attribute reducts efficiently becomes more and more important. Although some approaches have been proposed, they are used for complete decision systems. There are relatively few studies on incremental attribute reduction for incomplete decision systems. We introduce knowledge granularity, that can be obtained by the tolerance classes, to measure the uncertainty in incomplete decision systems. Furthermore, we propose incremental attribute reduction algorithms for incomplete decision systems when adding multiple objects and when deleting multiple objects, respectively. Finally, experimental results show that the proposed incremental approach is effective and efficient to update attribute reducts with the variation of objects in incomplete decision systems.

**Keywords** Incremental attribute reduction · Knowledge granularity · Incomplete decision system · Rough sets

## 1 Introduction

Rough set theory, proposed by Pawlak (1982), is a powerful mathematical tool to deal with uncertainty, granularity, and incompleteness of knowledge in information systems. It has been applied successfully in many fields including machine learning, intelligent data analysis, decision making, knowledge engineering, disease diagnosis, and so on (Chen and Tanuwijaya 2011; Derrac et al. 2012; Chen and Chang 2011; Lin et al. 2011; Formica 2012; Wafo Soh et al. 2018; Min et al. 2011; Li et al. 2012, 2016; Wang et al. 2019a; Chen et al. 2013; Liu et al. 2018; D'Eer et al. 2016; Liao et al. 2018; Jothi and Hannah 2016; Zhan et al. 2017; Koley et al. 2016; Afridi et al. 2018; Xu et al. 2017). Since rough set theory can achieve a subset of all attributes which preserves the discernible ability of original features, using the data only with no additional information, it has been widely applied in attribute reduction (also called attribute selection or feature

selection) (Dai et al. 2017a; Raza and Qamar 2016; Pacheco et al. 2017; Wang et al. 2016, 2018, 2019b; Cheng et al. 2016; Min and Xu 2016; Raza and Qamar 2017; Li et al. 2017; Das et al. 2017; Tiwari et al. 2018; Lin et al. 2018; Yao and Zhang 2017; Dai et al. 2018).

As we know, attribute reduction plays an important role in data mining and knowledge discovery. Attribute reduction methods can be classified into non-incremental methods and incremental methods according to whether the computation of attribute reduction is from scratch or not when the data vary dynamically. Non-incremental attribute reduction, also called classic attribute reduction or static attribute reduction, has been fully studied and has yielded many important results (Pawlak 1991; Qian and Liang 2008; Xu and Yu 2017; Dai and Tian 2013; Dai and Xu 2013; Liang et al. 2014; Raza and Qamar 2017; Yao and Zhang 2017). In practice, real data may change dynamically nowadays. Non-incremental methods are often infeasible, since they need to compute repeatedly and consume a large amount of computational time. Incremental methods are considered as effective approaches to deal with dynamic data, because they can directly update the results using the previous results from the original decision system. In dynamic data environments, data changes take on three basic forms and the attribute

✉ Jianhua Dai  
jhdai@hunnu.edu.cn

<sup>1</sup> Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China

reduction problem has three issues correspondingly: variation of object sets, variation of attribute sets, and variation of attribute values (Xie and Qin 2018). Some achievements have been made to solve these problems. For example, Jing et al. (2016) considered the situation of variation of the attribute sets in complete decision systems, and introduced incremental mechanisms to calculate the new knowledge granularity and presented the corresponding incremental algorithms for attribute reduction based on the calculated knowledge granularity when multiple attributes are added to the decision system. Chen et al. (2016) presented an incremental algorithm for attribute reduction with variable precision rough sets for the same purpose, by introducing two Boolean row vectors to characterize the discernibility matrix and reduct and employing an incremental manner to update minimal elements in the discernibility matrix at the arrival of an incremental sample. Xie and Qin (2018) considered the situation of variation of the attribute values, introduced the concept of an inconsistency degree in an incomplete decision system, and proposed a framework of the incremental attribute reduction algorithm based on three update strategies of inconsistency degree for dynamic incomplete decision systems. Wei et al. (2018) proposed a discernibility matrix based incremental attribute reduction algorithm to incrementally acquire all reducts of dynamic data and another incremental attribute reduction algorithm of more efficiency.

The above methods are suitable only for complete information systems or complete decision systems and cannot be applied to incomplete situations. Thus, further studies on uncertainty measures for incomplete decision systems have been developed. Dai et al. (2013) proposed a new form of conditional entropy and obtained some important properties, which can be used as a reasonable uncertainty measure for incomplete decision systems. Dai et al. (2017b) investigated the uncertainty measures in incomplete interval-valued information systems based on an  $\alpha$ -weak similarity relation and defined accuracy, roughness, and approximation accuracy to evaluate the uncertainty based on a rough set model constructed. Liu et al. (2016) provided a novel three-way decision model and corresponding algorithm based on incomplete information system by defining a new relation to describe the similarity degree of incomplete information and utilizing interval number to acquire the loss function. Du and Hu (2016) investigated an approach on the basis of the discernibility matrix and the discernibility function to compute all the reducts in incomplete ordered information systems by introducing the characteristic-based dominance relation, and designed a heuristic algorithm with polynomial time complexity for finding a unique reduct using the inner and outer significance measures of each criterion candidate.

Based on the aforementioned survey, we can see that the methods mentioned above have not investigated the incremental attribute reduction mechanism for information decision systems of character as incomplete and dynamic simultaneously. To our best knowledge, there are only a few work on the incremental attribute reduction mechanism for incomplete information decision systems. Shu and Shen (2013) introduced a simpler way of computing tolerance classes than the classical method and presented an incremental attribute reduction algorithm to compute an attribute reduct for a dynamically increasing incomplete decision system. Yang et al. (2017) presented an efficient incremental algorithm including active sample selection process and incremental attribute reduction process from dynamic data sets with increasing samples. Shu and Shen (2014b) proposed an positive region-based attribute reduction algorithm to solve the attribute reduction problem efficiently in incomplete decision systems with dynamically varying attribute sets. Shu and Shen (2014a) employed an incremental manner to compute the new positive region and developed two efficient incremental feature selection algorithms, respectively, for single object and multiple objects with varying feature values. From above, it appears that some of them mainly focused on the variation of attribute sets or attribute values, and others only considered the case of increasing objects dynamically. Thus, being inspired by Dai and Tian (2013), we introduce the tolerance class to measure knowledge granularity for the proposed incremental mechanism and develop two incremental attribute reduction algorithms for incomplete information decision systems in the cases of increasing and decreasing objects respectively.

The remainder of this paper is organized as follows. Section 2 reviews some basic concepts in rough set theory and introduces a tolerance class based presentation of the knowledge granularity. Incremental mechanisms to calculate knowledge granularity, relative knowledge granularity, and significance measurements of attributes for incomplete decision systems when objects vary dynamically and their corresponding attribute reduction algorithms are investigated in Sect. 3. In Sect. 4, experiments and comparisons are conducted. Section 5 concludes the whole paper.

## 2 Preliminary knowledge

In this section, we first review some basic concepts in rough set theory, which can also be referred to Pawlak (1991), Kryszkiewicz (1998), and Pawlak and Skowron (2007). Furthermore, we recall the concepts of incomplete information systems and decision systems. At last, the tolerance relation is reviewed.

## 2.1 Basic concepts in rough set theory

An information system is a quadruple  $IIS = \langle U, A, V, f \rangle$ , where  $U$  denotes a non-empty finite set of objects, which is called the universe;  $A$  denotes a non-empty finite set of conditional attributes;  $V$  is the union of attribute domains,  $V = \bigcup_{a \in A} V_a$ , where  $V_a$  is the value set of attribute  $a$ , called the domain of  $a$ ;  $f : U \times A \rightarrow V$  is an information function which assigns particular values from domains of attribute to objects such as  $a \in A, x \in U, f(a, x) \in V_a$ , where  $f(a, x)$  denotes the value of attribute  $a$  for object  $x$ . Each attribute subset  $B \subseteq A$  determines a binary indiscernible relation as follows:

$$IND(B) = \{(u_i, u_j) \in U^2 | \forall a \in B, a(u_i) = a(u_j)\}. \quad (1)$$

By the relation  $IND(B)$ , we obtain the partition of  $U$  denoted by  $U/IND(B)$  or  $U/B$ . For  $B \subseteq A$  and  $X \subseteq U$ , the lower approximation and the upper approximation of  $X$  can be defined as follows:

$$\begin{aligned} \underline{B}X &= \{u_i \in U | [u_i] \subseteq X\} \\ \overline{B}X &= \{u_i \in U | [u_i] \cap X \neq \emptyset\}, \end{aligned}$$

where  $\underline{B}X$  is a set of objects that belong to  $X$  with certainty, while  $\overline{B}X$  is a set of objects that possibly belong to  $X$ . If  $\underline{B}X = \overline{B}X$ ,  $X$  is named B-definable. Otherwise,  $X$  is named B-rough. Based on  $\underline{B}X$  and  $\overline{B}X$ , the B-positive region, B-negative region, and B-borderline region of  $X$  are defined, respectively, as follows:

$$\begin{aligned} POS_B(X) &= \underline{B}X \\ NEG_B(X) &= U - \overline{B}X \\ BN_B &= \overline{B}X - \underline{B}X. \end{aligned}$$

## 2.2 Incomplete decision systems and tolerance relation

An information system is a quadruple  $IS = \langle U, A, V, f \rangle$ . If there exists  $x \in U$  and  $a \in A$ , such that  $f(a, x)$  is equal to a missing value (a null or unknown value, denoted as “\*”), i.e.,  $* \in V_a$ , then the information system is called an incomplete information system (IIS). Thus, the IIS can be denoted as:  $IIS = \langle U, A, V, f \rangle$ , where  $* \in V_A$ .

A decision system is a quadruple  $IDS = \langle U, C \cup D, V, f \rangle$ , where  $D$  is the decision attribute set,  $C$  is the conditional attribute set, and  $C \cap D = \emptyset$ ;  $V$  is the union of attribute domain, i.e.,  $V = V_C \cup V_D$ . In general, we assume that  $D = \{d\}$ . If there exists an  $a \in A, x \in U$ , such that  $f(a, x)$  is equal to a missing value, then the decision system is called an incomplete decision system (IDS). Thus, the IDS can be denoted as:  $IDS = \langle U, C \cup D, V, f \rangle$ , where  $* \in V_C, * \notin V_D$ .

**Definition 1** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , for any subset of attributes  $B \subseteq C$ , let  $T(B)$  denote the binary tolerance relation between objects that are possibly indiscernible in terms of values of attributes in  $B$ .  $T(B)$  is defined as

$$T(B) = \{(x, y) | \forall a \in B, f(a, x) = f(a, y) \vee f(a, x) = * \vee f(a, y) = *\}, \quad (2)$$

where  $T(B)$  is reflexive and symmetric, but not necessarily transitive.

**Definition 2** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $x \in U$  and  $B \subseteq C$ , the tolerance class of the object  $x$  with respect to attribute set  $B$  is defined by

$$T_B(x) = \{y | (x, y) \in T(B)\} \quad (3)$$

## 2.3 Knowledge granulation in incomplete decision systems

**Definition 3** (Dai and Tian 2013) Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $T_C(x)$  is the tolerance class of object  $x$  with respect to attribute set  $C$ . Based on the tolerance class, the knowledge granularity of  $C$  on  $U$  is defined as follows:

$$GK_U(C) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |T_C(u_i)|, \quad (4)$$

where  $|U|$  stands for the number of objects in  $U$ .

**Example 1** Example for computing of the knowledge granularity.

Table 1 shows an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , where  $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\}$ ,  $C = \{a_1, a_2, a_3, a_4, a_5\}$  and  $D = \{0, 1\}$ . According to Definition 2, we have  $T_C(u_1) = \{u_1, u_6\}$ ,  $T_C(u_2) = \{u_2, u_4, u_8\}$ ,  $T_C(u_3) = \{u_3, u_5\}$ ,  $T_C(u_4) = \{u_2, u_4, u_8\}$ ,  $T_C(u_5) = \{u_3, u_5\}$ ,  $T_C(u_6) = \{u_1, u_6, u_7\}$ ,  $T_C(u_7) = \{u_6, u_7\}$ ,  $T_C(u_8) = \{u_2, u_4, u_8, u_9\}$ , and  $T_C(u_9) = \{u_8, u_9\}$ . According to Definition 3, we have

**Table 1** Incomplete decision system 1

$U$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$d$
$u_1$	1	*	0	0	1	0
$u_2$	0	0	1	*	1	0
$u_3$	0	1	*	1	0	1
$u_4$	*	0	1	0	1	0
$u_5$	0	1	1	1	0	1
$u_6$	1	0	0	0	*	1
$u_7$	1	0	0	*	0	1
$u_8$	0	*	1	0	1	0
$u_9$	*	1	1	0	1	1

$$GK_U(C) = \frac{1}{9^2} (2 + 3 + 2 + 3 + 2 + 3 + 2 + 4 + 2) = \frac{23}{81}.$$

Similarly, we can get  $GK_U(C \cup D) = \frac{19}{81}$ .

**Proposition 1** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $P, Q \subseteq C$ . IF  $P \subseteq Q$ , then  $GK_U(P) \geq GK_U(Q)$ .

**Proof** According to Definition 2, we have  $\forall u_i \in U$ ,  $T_P(u_i) \supseteq T_Q(u_i)$ . Then, we get  $|T_P(u_i)| \geq |T_Q(u_i)|$ .

Since  $GK_U(P) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |T_P(u_i)|$  and  $GK_U(Q) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} |T_Q(u_i)|$ , we have  $\frac{1}{|U|^2} \sum_{i=1}^{|U|} |T_P(u_i)| \geq \frac{1}{|U|^2} \sum_{i=1}^{|U|} |T_Q(u_i)|$ .

It is obvious that  $GK_U(P) \geq GK_U(Q)$ .  $\square$

Let  $IDS = \langle U, C \cup D, V, f \rangle$  be an incomplete system,  $P, Q \subseteq C$ . If  $P \subseteq Q$ , we can see that  $GK_U(P) \geq GK_U(Q)$ . In other word, the knowledge granularity of attribute set declines with the increase of number of attributes. Thus, the measure of knowledge granularity has the monotonicity with respect to attributes and is reasonable to be used as a uncertainty measure in rough set theory.

**Definition 4** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $T(C)$  and  $T(C \cup D)$  are the tolerance relation for attribute set  $C$  and  $C \cup D$ , respectively. The knowledge granularity of  $C$  relative to  $D$  on  $U$  is defined as follows:

$$GK_U(D|C) = GK_U(C) - GK_U(C \cup D) \\ = \frac{1}{|U|^2} \sum_{i=1}^{|U|} (|T_C(u_i)| - |T_{C \cup D}(u_i)|). \quad (5)$$

**Definition 5** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $T_C$ ,  $T_{C-\{a\}}$ ,  $T_{C \cup D}$ , and  $T_{(C-\{a\}) \cup D}$  are the tolerance relation for  $C$ ,  $C - \{a\}$ ,  $C \cup D$ , and  $(C - \{a\}) \cup D$ , respectively. The significance measure (inner significance) of  $a$  in  $C$  on  $U$  is defined as follows:

$$\text{Sig}_U^{\text{inner}}(a, C, D) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} [|T_{C-\{a\}}(u_i)| \\ - |T_{(C-\{a\}) \cup D}(u_i)| - (|T_C(u_i)| \\ - |T_{C \cup D}(u_i)|)], \quad (6)$$

where  $a$  denotes any one attribute in  $C$  and the following are the same.

**Definition 6** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , the core of  $IDS$  is defined as follows:

$$\text{Core}_C = \{a \in C | \text{Sig}_U^{\text{inner}}(a, C, D) > 0\}. \quad (7)$$

If  $\forall a \in C$ ,  $\text{Sig}_U^{\text{inner}}(a, C, D) = 0$ , then  $\text{Core}_C = \emptyset$ .

**Example 2** (Continued from Example 1) According to Definitions 3, 4, and 5, we have

$$\text{Sig}_U^{\text{inner}}(a_1, C, D) = \frac{1}{81} [(23 - 19) - (23 - 19)] = 0$$

$$\text{Sig}_U^{\text{inner}}(a_2, C, D) = \frac{1}{81} [(27 - 19) - (23 - 19)] = \frac{4}{81}$$

$$\text{Sig}_U^{\text{inner}}(a_3, C, D) = \frac{1}{81} [(29 - 21) - (23 - 19)] = \frac{4}{81}$$

$$\text{Sig}_U^{\text{inner}}(a_4, C, D) = \frac{1}{81} [(23 - 19) - (23 - 19)] = 0$$

$$\text{Sig}_U^{\text{inner}}(a_5, C, D) = \frac{1}{81} [(25 - 19) - (23 - 19)] = \frac{2}{81}$$

Then, we have  $\text{Core}_C = \{a_2, a_3, a_5\}$ .

**Definition 7** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$  and  $B \subseteq C$ ,  $T_B$ ,  $T_{B \cup D}$ ,  $T_{B \cup \{a\}}$  and  $T_{(B \cup \{a\}) \cup D}$  are the tolerance relation for  $B$ ,  $B \cup D$ ,  $B \cup \{a\}$  and  $(B \cup \{a\}) \cup D$ , respectively. Then,  $\forall a \in (C - B)$ , the significance measure (outer significance) of  $a$  in  $B$  on  $U$  is defined as follows:

$$\text{Sig}_U^{\text{outer}}(a, B, D) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} [|T_B(u_i)| - |T_{B \cup D}(u_i)| \\ - (|T_{B \cup \{a\}}(u_i)| - |T_{(B \cup \{a\}) \cup D}(u_i)|)]. \quad (8)$$

**Definition 8** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$  and  $B \subseteq C$ , then  $B$  is a relative reduct based on the knowledge granularity of  $IDS$  if

- (1)  $GK_U(D|B) = GK_U(D|C)$ .
- (2)  $\forall a \in B$ ,  $GK_U(D|(B - \{a\})) \neq GK_U(D|B)$ .

### 3 Incremental attribute reduction algorithms for incomplete decision systems when objects vary dynamically

After having investigated an incremental mechanism to compute knowledge granularity for incomplete decision systems to which multiple objects are added one by one, this section introduces an incremental attribute reduction algorithm for the addition of multiple objects based on knowledge granularity.

#### 3.1 An incremental mechanism to calculate knowledge granularity for IDS when adding an object

This section investigates changes of tolerance class, relative knowledge granularity, inner significance, and outer significance, and then introduces the incremental

mechanism for calculating knowledge granularity when an object is added to an incomplete decision system.

**Proposition 2** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , where  $U = \{u_1, u_2, \dots, u_n\}$  denotes a non-empty finite set containing  $n$  objects.  $u_{(n+1)}$  is the incremental object that will be added to  $IDS$ , and  $T'_C$  is the tolerance relation on  $U^+ = (U \cup \{u_{(n+1)}\})$ . The knowledge granularity of  $C$  on  $U^+$  is

$$GK_{U^+}(C) = \frac{1}{(n+1)^2} (n^2 GK_U(C) + 2|T'_C(u_{(n+1)})| - 1), \quad (9)$$

where  $T'_C(u_{(n+1)}) = \{u_j | (u_{(n+1)}, u_j) \in T'_C, 1 \leq j \leq n+1\}$ .

**Proof** After  $u_{(n+1)}$  is adding to  $U$ , the tolerance class of  $u_i$  is

$$T'_C(u_i) = \begin{cases} T_C(u_i) \cup \{u_{(n+1)}\}, & (u_i, u_{(n+1)}) \in T'_C \\ T_C(u_i), & (u_i, u_{(n+1)}) \notin T'_C \end{cases}, \quad 1 \leq i \leq n$$

Suppose

$$\Delta|T'_C(u_i)| = \begin{cases} 1, & (u_i, u_{(n+1)}) \in T'_C \\ 0, & (u_i, u_{(n+1)}) \notin T'_C \end{cases}, \quad 1 \leq i \leq n$$

then

$$|T'_C(u_i)| = |T_C(u_i)| + \Delta|T'_C(u_i)|, \quad 1 \leq i \leq n.$$

Because tolerance relation is symmetric, if  $(u_i, u_{(n+1)}) \in T'_C$ , then  $(u_{(n+1)}, u_i) \in T'_C, 1 \leq i \leq n$ . In other words, if  $u_i \in T'_C(u_{(n+1)})$ , then  $u_{(n+1)} \in T'_C(u_i), 1 \leq i \leq n$ . Obviously, the number of objects that has tolerance relation with  $u_{(n+1)}$  is equal to those whose tolerance class contains  $u_{(n+1)}$  except  $u_{(n+1)}$  itself after  $u_{(n+1)}$  is added to  $IDS$ . Then, we can get

$$\sum_{i=1}^{|U|} \Delta|T'_C(u_i)| = |T'_C(u_{(n+1)})| - 1.$$

According to Definition 3, the knowledge granularity of  $C$  on  $U^+$  is described as follows:

$$\begin{aligned} GK_{U^+}(C) &= \frac{1}{|U^+|^2} \sum_{i=1}^{|U^+|} |T'_C(u_i)| \\ &= \frac{1}{(n+1)^2} \left( \sum_{i=1}^{|U|} |T'_C(u_i)| + |T'_C(u_{(n+1)})| \right) \\ &= \frac{1}{(n+1)^2} \left( \sum_{i=1}^{|U|} |T_C(u_i)| + \sum_{i=1}^{|U|} \Delta|T'_C(u_i)| + |T'_C(u_{(n+1)})| \right) \\ &= \frac{1}{(n+1)^2} (n^2 GK_U(C) + 2|T'_C(u_{(n+1)})| - 1). \end{aligned}$$

□

**Proposition 3** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ .  $u_{(n+1)}$  is the incremental object.  $T'_C(u_{(n+1)})$  and  $T'_{C \cup D}(u_{(n+1)})$  are the tolerance classes of  $u_{(n+1)}$  with respect to attribute set  $C$  and  $C \cup D$  on  $U^+$ , respectively. The knowledge granularity of  $C$  relative to  $D$  on  $U^+$  is

$$GK_{U^+}(D|C) = \frac{1}{(n+1)^2} (n^2 GK_U(D|C) + 2(|T'_C(u_{(n+1)})| - |T'_{C \cup D}(u_{(n+1)})|)), \quad (10)$$

where  $| \cdot |$  denotes the cardinality of a set.

**Proof** According to Definition 4 and Proposition 2, we have

$$\begin{aligned} GK_{U^+}(D|C) &= GK_{U^+}(C) - GK_{U^+}(C \cup D) \\ &= \frac{1}{(n+1)^2} (n^2 GK_U(C) + 2|T'_C(u_{(n+1)})| - 1) \\ &\quad - \frac{1}{(n+1)^2} (n^2 GK_U(C \cup D) + 2|T'_{C \cup D}(u_{(n+1)})| - 1) \\ &= \frac{1}{(n+1)^2} (n^2 (GK_U(C) - GK_U(C \cup D)) \\ &\quad + 2|T'_C(u_{(n+1)})| - 2|T'_{C \cup D}(u_{(n+1)})|) \\ &= \frac{1}{(n+1)^2} (n^2 GK_U(D|C) + 2(|T'_C(u_{(n+1)})| - |T'_{C \cup D}(u_{(n+1)})|)). \end{aligned}$$

□

**Proposition 4** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $u_{(n+1)}$  is the incremental object.  $T'_{C-\{a\}}(u_{(n+1)})$ ,  $T'_{(C-\{a\}) \cup D}(u_{(n+1)})$ ,  $T'_C(u_{(n+1)})$  and  $T'_{C \cup D}(u_{(n+1)})$  are the tolerance classes of  $u_{(n+1)}$  with respect to attribute set  $C - \{a\}$ ,  $(C - \{a\}) \cup D$ ,  $C$  and  $C \cup D$  on  $U^+$ , respectively. Then  $\forall a \in C$ , the inner significance of  $a$  in  $C$  on  $U^+$  is

$$\begin{aligned} \text{Sig}_{U^+}^{\text{inner}}(a, C, D) &= \frac{1}{(n+1)^2} (n^2 \text{Sig}_U^{\text{inner}}(a, C, D) \\ &\quad + 2(|T'_{C-\{a\}}(u_{(n+1)})| - |T'_C(u_{(n+1)})| \\ &\quad + |T'_{C \cup D}(u_{(n+1)})| - |T'_{(C-\{a\}) \cup D}(u_{(n+1)})|)) \end{aligned} \quad (11)$$

**Proof** According to Definition 5 and Proposition 3, we can get



$$\begin{aligned}
 \text{Sig}_{U^+}^{\text{inner}}(a, C, D) &= GK_{U^+}(D|C - \{a\}) - GK_{U^+}(D|C) \\
 &= \frac{1}{(n+1)^2} (n^2 GK_U(D|C - \{a\}) + 2(|T'_{C-\{a\}}(u_{(n+1)})| \\
 &\quad - |T'_{(C-\{a\}) \cup D}(u_{(n+1)})|)) - \frac{1}{(n+1)^2} (n^2 GK_U(D|C) \\
 &\quad + 2(|T'_C(u_{(n+1)})| - |T'_{C \cup D}(u_{(n+1)})|)) \\
 &= \frac{1}{(n+1)^2} (n^2 (GK_U(D|C - \{a\}) - GK_U(D|C)) \\
 &\quad + 2(|T'_{C-\{a\}}(u_{(n+1)})| - |T'_C(u_{(n+1)})| \\
 &\quad + |T'_{C \cup D}(u_{(n+1)})| - |T'_{(C-\{a\}) \cup D}(u_{(n+1)})|)) \\
 &= \frac{1}{(n+1)^2} (n^2 \text{Sig}_U^{\text{inner}}(a, C, D) + 2(|T'_{C-\{a\}}(u_{(n+1)})| \\
 &\quad - |T'_C(u_{(n+1)})| + |T'_{C \cup D}(u_{(n+1)})| - |T'_{(C-\{a\}) \cup D}(u_{(n+1)})|))
 \end{aligned}$$

□

**Proposition 5** Given an incomplete decision system  $\text{IDS} = \langle U, C \cup D, V, f \rangle$ ,  $u_{(n+1)}$  is the incremental object.  $|T'_B(u_{(n+1)})|$ ,  $|T'_{B \cup D}(u_{(n+1)})|$ ,  $|T'_{B \cup \{a\}}(u_{(n+1)})|$  and  $|T'_{B \cup \{a\} \cup D}(u_{(n+1)})|$  are the tolerance classes of  $u_{(n+1)}$  with respect to attribute set  $B$ ,  $B \cup D$ ,  $B \cup \{a\}$  and  $B \cup \{a\} \cup D$  on  $U^+$ , respectively. Then  $\forall a \in (C - B)$ , the outer significance of  $a$  in  $B$  on  $U^+$  is

$$\begin{aligned}
 \text{Sig}_{U^+}^{\text{outer}}(a, B, D) &= \frac{1}{(n+1)^2} (n^2 \text{Sig}_U^{\text{outer}}(a, B, D) \\
 &\quad + 2(|T'_B(u_{(n+1)})| - |T'_{B \cup D}(u_{(n+1)})| \\
 &\quad - |T'_{B \cup \{a\}}(u_{(n+1)})| + |T'_{B \cup \{a\} \cup D}(u_{(n+1)})|))
 \end{aligned} \quad (12)$$

**Proof** According to Definition 7 and Proposition 3, we can get

$$\begin{aligned}
 \text{Sig}_{U^+}^{\text{outer}}(a, B, D) &= GK_{U^+}(D|B) - GK_{U^+}(D|(B \cup \{a\})) \\
 &= \frac{1}{(n+1)^2} (n^2 GK_U(D|B) + 2(|T'_B(u_{(n+1)})| \\
 &\quad - |T'_{B \cup D}(u_{(n+1)})|)) - \frac{1}{(n+1)^2} (n^2 GK_U(D|B \cup \{a\}) \\
 &\quad + 2(|T'_{B \cup \{a\}}(u_{(n+1)})| - |T'_{B \cup \{a\} \cup D}(u_{(n+1)})|)) \\
 &= \frac{1}{(n+1)^2} (n^2 (GK_U(D|B) - GK_U(D|B \cup \{a\})) \\
 &\quad + 2(|T'_B(u_{(n+1)})| - |T'_{B \cup D}(u_{(n+1)})| \\
 &\quad - |T'_{B \cup \{a\}}(u_{(n+1)})| + |T'_{B \cup \{a\} \cup D}(u_{(n+1)})|)) \\
 &= \frac{1}{(n+1)^2} (n^2 \text{Sig}_U^{\text{outer}}(a, B, D) + 2(|T'_B(u_{(n+1)})| \\
 &\quad - |T'_{B \cup D}(u_{(n+1)})| - |T'_{B \cup \{a\}}(u_{(n+1)})| \\
 &\quad + |T'_{B \cup \{a\} \cup D}(u_{(n+1)})|))
 \end{aligned}$$

□

### 3.2 An incremental reduction algorithm for IDS when adding one object

First, a traditional heuristic attribute reduction algorithm for decision systems is introduced in Algorithm 1 (Pawlak 1991; Wang et al. 2013; Liang et al. 2014). Based on the incremental mechanism of knowledge granularity above, this subsection introduces an incremental attribute reduction algorithm (see Algorithm 2) under knowledge granularity when adding an object to the decision system. At last, a brief comparison of time complexity between incremental reduction algorithm and traditional heuristic reduction algorithm is given.

The detailed execution process of Algorithm 1 is as follows. In Step 1, an empty set is assigned to  $RED_U$ . Steps 2–7 are actually used to get the core of the incomplete decision system according to Definition 6, which constitute a loop with  $|C|$  times.  $|C|$  denotes the number of all conditional attributes in the incomplete decision system. According to Definition 5, the inner significance of an certain  $a_j \in C$ , that is  $\text{Sig}_U^{\text{inner}}(a_j, C, D)$ , is calculated in Step 3. In Step 4,  $\text{Sig}_U^{\text{inner}}(a_j, C, D)$  is compared with zero if  $\text{Sig}_U^{\text{inner}}(a_j, C, D) > 0$ , then  $a_j$  is added to  $RED_U$ , which denotes  $a_j$  is indispensable and should be added to the core of the incomplete decision system. In Step 8,  $RED_U$  is assigned to  $B$ . Steps 9–15 are actually used to find an attribute set that satisfies the first condition in Definition 8 and a loop stopping until  $GP_U(D|B)$  is equal to  $GP_U(D|C)$ . Steps 10–12 constitute a loop used to calculate the outer significance of an certain  $(a_i) \in (C - B)$ , that is  $\text{Sig}_U^{\text{outer}}(a_i, B, D)$  according to Definition 7. In Step 13,  $a_0$  is used to store the attribute that has the maximum value of outer significance. In Step 14,  $a_0$  is added to  $B$ . Steps 16–20 are actually used to delete attributes that satisfy the second condition in Definition 8 and a loop with  $|B|$  times. In Step 17,  $GP_U(D|(B - \{a\}))$  is compared with  $GP_U(D|C)$  if they are equal, then  $a_i$  is deleted from  $B$ , which indicates that  $a_i$  is redundant and cannot be in the reduct of the incomplete decision system. In Step 21,  $B$  is assigned to  $RED_U$ . In Step 22,  $RED_U$  is returned as the result of Algorithm 1.

**Algorithm 1** Traditional Heuristic attribute reduction Algorithm based on knowledge granularity for decision systems (THA)

**Input:** A decision system  $IDS = \langle U, C \cup D, V, f \rangle$

**Output:** A reduct  $RED_U$  on  $U$

```

1:  $RED_U \leftarrow \emptyset$ ;
2: for  $1 \leq j \leq |C|$  do
3:   Calculate  $Sig_U^{inner}(a_j, C, D)$ ;
4:   if  $Sig_U^{inner}(a_j, C, D) > 0$  then
5:      $RED_U \leftarrow (RED_U \cup \{a_j\})$ ;
6:   end if
7: end for
8:  $B \leftarrow RED_U$ ;
9: while  $GP_U(D|B) \neq GP_U(D|C)$  do
10:  for each  $(a_i) \in (C - B)$  do
11:    Compute  $Sig_U^{outer}(a_i, B, D)$ ;
12:  end for
13:   $a_0 = \max\{Sig_U^{outer}(a_i, B, D), a_i \in (C - B)\}$ ;
14:   $B \leftarrow (B \cup \{a_0\})$ ;
15: end while
16: for each  $(a_i) \in B$  do
17:  if  $GP_U(D|(B - \{a_i\})) = GP_U(D|C)$  then
18:     $B \leftarrow (B - \{a_i\})$ ;
19:  end if
20: end for
21:  $RED_U \leftarrow B$ ;
22: return  $RED_U$ ;

```

The time complexity of Algorithm 1 is  $O(|U|^2|C|^2)$ . According to Definition 2, we first compute  $T_C(u_i)$ ,  $1 \leq i \leq n$  with a time complexity being  $O(|U||C|)$ . Then, we can get  $GK_U(C)$  with a time complexity being  $O(|U|^2|C|)$ . Thus, according to Definition 5, the time of calculating  $Sig_U^{inner}(a_j, C, D)$  is  $O(4|U|^2|C|) \approx O(|U|^2|C|)$ . Therefore, the time complexity of Steps 2–7 is  $O(|U|^2|C|^2)$ . Similarly, the time complexity of Steps 9–15 is  $O(|U|^2|C|^2)$ , and the time complexity of Steps 16–20 is also  $O(|U|^2|C|^2)$ . Hence, the time complexity of Algorithm 1 is  $O(|U|^2|C|^2)$  based on the foregoing analysis.

In Algorithm 2,  $U^+$  denotes the new decision system after adding the incremental object  $u_{(n+1)}$  to the original decision system  $U$ . The detailed execution process of Algorithm 2 is as follows. In Step 1, the reduct  $RED_U$  on  $U$  is assigned to  $B$ . In Step 2, four tolerance classes of the incremental object  $u_{(n+1)}$  in  $U^+$ , which are  $T'_C(u_{(n+1)})$ ,  $T'_{C \cup D}(u_{(n+1)})$ ,  $T'_B(u_{(n+1)})$  and  $T'_{B \cup D}(u_{(n+1)})$ , are calculated respectively according to Definition 2. In Step 3, the knowledge granularity of  $B$  relative to  $D$  on  $U^+(GK_{U^+}(D|B))$  and the knowledge granularity of  $C$  relative to  $D$  on  $U^+(GK_{U^+}(D|C))$  are calculated, respectively, according to Proposition 3. In Step 4,  $GK_{U^+}(D|B)$  is compared with  $GK_{U^+}(D|C)$  if they are equal, then algorithm flow skips to step 19, which indicates that the reducts of  $U^+$  and  $U$  are identical. Otherwise, the loop consisting of Steps 7–13 executes. Steps 8–10 constitute a loop used to

calculate the outer significance of an certain  $(a_i) \in (C - B)$ , that is,  $Sig_{U^+}^{outer}(a_i, B, D)$ , according to Proposition 5. In Step 11,  $a_0$  is used to store the attribute that has the maximum value of outer significance. In Step 12,  $a_0$  is added to  $B$ . Steps 14–18 are actually used to delete attributes that satisfy the second condition in Definition 8 and a loop with  $|B|$  times. In Step 15,  $GK_{U^+}(D|(B - \{a_i\}))$  is compared with  $GK_{U^+}(D|C)$  if they are equal, then  $a_i$  is deleted from  $B$ , which indicates that  $a_i$  is redundant and cannot be in the reduct of  $U^+$ . In Step 19,  $B$  is assigned to  $RED_{U^+}$ . In Step 20,  $RED_{U^+}$  is returned as the result of Algorithm 2.

**Algorithm 2** An Knowledge Granularity based Incremental Reduction algorithm when Adding an object (KGIRA)

**Input:** An incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , the reduct  $RED_U$  on  $U$ , the incremental object  $u_{n+1}$ .

**Output:** A new reduct  $RED_{U^+}$  on  $(U^+)$  after adding  $u_{n+1}$  to  $IDS$

```

1:  $B \leftarrow RED_U$ ;
2: Compute  $T'_C(u_{(n+1)})$ ,  $T'_{C \cup D}(u_{(n+1)})$ ,  $T'_B(u_{(n+1)})$ ,  $T'_{B \cup D}(u_{(n+1)})$ ;
3: Compute  $GK_{U^+}(D|B)$  and  $GK_{U^+}(D|C)$  (according to Proposition 3);
4: if  $GK_{U^+}(D|B) = GK_{U^+}(D|C)$  then
5:   go to 19;
6: end if
7: while  $GK_{U^+}(D|B) \neq GK_{U^+}(D|C)$  do
8:   for each  $(a_i) \in (C - B)$  do
9:     Compute  $Sig_{U^+}^{outer}(a_i, B, D)$  (according to Proposition 5);
10:   end for
11:    $a_0 = \max\{Sig_{U^+}^{outer}(a_i, B, D), a_i \in (C - B)\}$ ;
12:    $B \leftarrow (B \cup \{a_0\})$ ;
13: end while
14: for each  $(a_i) \in B$  do
15:   if  $GK_{U^+}(D|(B - \{a_i\})) = GK_{U^+}(D|C)$  then
16:      $B \leftarrow (B - \{a_i\})$ ;
17:   end if
18: end for
19:  $RED_{U^+} \leftarrow B$ ;
20: return  $RED_{U^+}$ ;

```

The time complexity of Algorithm 2 is  $O(|U||C|^2)$ . According to Definition 2, the time complexity of computing  $T'_C(u_{(n+1)})$  is  $O(|U||C|)$ . Then, according to Proposition 3, the time complexity of Steps 2–3 is  $O(|U||C| + |U|(|C| + 1) + |U||B| + |U|(|B| + 1)) \approx O(|U||C|)$ . According to Proposition 4, the time complexity of Steps 7–13 is  $O(|U||C|^2)$ . Similarly, the time complexity of Steps 14–18 is  $O(|U||C|^2)$ . Hence, the time complexity of Algorithm 2 is  $O(|U||C| + |U||C|^2 + |U||C|^2) \approx O(|U||C|^2)$  based on the foregoing analysis.

Since the number of objects is  $|U| + 1$ , the time complexity of THA is actually  $O(|C|^2(|U| + 1)^2)$  when adding

the object  $u_{n+1}$  to the object set  $U$ . Because  $|U||C|^2 \leq (|U| + 1)^2|C|^2$ ,  $O(|U||C|^2)$  is much lower than  $O((|U| + 1)^2|C|^2)$ . Hence, KGIRA spends much less time than THA.

### 3.3 An incremental mechanism to calculate knowledge granularity for IDS when deleting one object

For the sake of convenience, given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ .  $U = \{u_1, u_2, \dots, u_n\}$  denotes the original object set and  $u_n$  denotes the object to be deleted from  $U$ . For simplicity,  $U^- = U - \{u_n\} = \{u_1, u_2, \dots, u_{n-1}\}$  is written as  $U^-$  in the following.

**Proposition 6** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , where  $U = \{u_1, u_2, \dots, u_n\}$  denotes a non-empty finite set containing  $n$  objects.  $u_n$  is the object that will be deleted from IDS, and  $T_C$  is the tolerance relation on  $U$ . The knowledge granularity of  $C$  on  $U^-$  is

$$GK_{U^-}(C) = \frac{1}{(n-1)^2} [n^2 GK_U(C) - (2|T_C(u_n)| - 1)] \quad (13)$$

**Proof** After  $u_n$  is deleting to  $U$ , the tolerance class of  $u_i$  is

$$T_C''(u_i) = \begin{cases} T_C(u_i) - \{u_n\}, & (u_i, u_n) \in T_C \\ T_C(u_i), & (u_i, u_n) \notin T_C \end{cases} \quad 1 \leq i \leq n-1.$$

Suppose

$$\Delta|T_C(u_i)| = \begin{cases} 1, & (u_i, u_n) \in T_C \\ 0, & (u_i, u_n) \notin T_C \end{cases} \quad 1 \leq i \leq n-1,$$

then  $|T_C''(u_i)| = |T_C(u_i)| - \Delta|T_C(u_i)|$ ,  $1 \leq i \leq n-1$ . Because tolerance relation is symmetric, if  $(u_i, u_n) \in T_C$ , then  $(u_n, u_i) \in T_C$ ,  $1 \leq i \leq n-1$ . In other words, if  $u_i \in T_C(u_n)$ , then  $u_n \in T_C(u_i)$ ,  $1 \leq i \leq n-1$ . Obviously, the number of objects that has tolerance relation with  $u_n$  is equal to those whose tolerance class contains  $u_n$  except  $u_n$  itself on  $U$ . Then, we can get

$$\sum_{i=1}^{n-1} \Delta|T_C(u_i)| = |T_C(u_n)| - 1.$$

According to Definition 3, the knowledge granularity of  $C$  on  $U^-$  is described as follows:

$$\begin{aligned} GK_{U^-}(C) &= \frac{1}{|U^-|^2} \sum_{i=1}^{|U^-|} |T_C''(u_i)| \\ &= \frac{1}{(n-1)^2} \left( \sum_{i=1}^n |T_C(u_i)| - \sum_{i=1}^{n-1} \Delta|T_C(u_i)| - |T_C(u_n)| \right) \\ &= \frac{1}{(n-1)^2} \left[ \sum_{i=1}^n |T_C(u_i)| - (|T_C(u_n)| - 1) - |T_C(u_n)| \right] \\ &= \frac{1}{(n-1)^2} [n^2 GK_U(C) - (2|T_C(u_n)| - 1)]. \end{aligned}$$

□

**Proposition 7** Given an incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ ,  $u_n$  is the deleting object.  $T_C(u_n)$  and  $T_{C \cup D}(u_n)$  are the tolerance classes of  $u_n$  with respect to attribute set  $C$  and  $C \cup D$  on  $U$ , respectively. The knowledge granularity of  $C$  relative to  $D$  on  $U^-$  is

$$GK_{U^-}(D|C) = \frac{1}{(n-1)^2} [n^2 GK_U(D|C) - 2(|T_C(u_n)| - |T_{C \cup D}(u_n)|)]. \quad (14)$$

**Proof** According to Definition 4 and Proposition 6, we have

$$\begin{aligned} GK_{U^-}(D|C) &= GK_{U^-}(C) - GK_{U^-}(C \cup D) \\ &= \frac{1}{(n-1)^2} [n^2 GK_U(C) - (2|T_C(u_n)| - 1)] \\ &\quad - \frac{1}{(n-1)^2} [n^2 GK_U(C \cup D) - (2|T_{C \cup D}(u_n)| - 1)] \\ &= \frac{1}{(n-1)^2} [n^2 (GK_U(C) - GK_U(C \cup D)) - 2(|T_C(u_n)| - |T_{C \cup D}(u_n)|)] \\ &= \frac{1}{(n-1)^2} [n^2 GK_U(D|C) - 2(|T_C(u_n)| - |T_{C \cup D}(u_n)|)] \end{aligned}$$

□

### 3.4 An incremental reduction algorithm for IDS when deleting an object

Based on the incremental mechanism of knowledge granularity above, this subsection introduces an incremental attribute reduction algorithm (see Algorithm 3) when deleting multiple objects from the decision system.



**Algorithm 3** An Knowledge Granularity based Incremental Reduction algorithm when Deleting an object (KGIRD)

**Input:** An incomplete decision system  $IDS = \langle U, C \cup D, V, f \rangle$ , the reduct  $RED_U$  on  $U$ .

**Output:** A new reduct  $RED_{U^-}$  on  $(U^-)$  after deleting  $u_n$  from  $IDS$ .

```

1:  $B \leftarrow RED_U$ ;
2: Compute  $GP_{U^-}(D|C)$  (according to Proposition 7);
3: for each( $a_i \in B$ ) do
4:   if  $GP_{U^-}(D|(B - \{a_i\})) = GP_{U^-}(D|C)$  then
5:      $B \leftarrow (B - \{a_i\})$ ;
6:   end if
7: end for
8:  $RED_{U^-} \leftarrow B$ ;
9: return  $RED_{U^-}$ ;

```

In Algorithm 3,  $U^-$  denotes the new decision system after deleting an object from the original decision system  $U$ . The detailed execution process of Algorithm 3 is as follows. In Step 1, the reduct  $RED_U$  on  $U$  is assigned to  $B$ . In Step 2, the knowledge granularity of  $C$  relative to  $D$  on  $U^-$  ( $GP_{U^-}(D|C)$ ) is calculated according to Proposition 7. Steps 3–7 are actually used to delete attributes that satisfies the second condition in Definition 8 and a loop with  $|B|$  times. In Step 4,  $GP_{U^-}(D|(B - \{a_i\}))$  is compared with  $GP_{U^-}(D|C)$  if they are equal, then  $a_i$  is deleted from  $B$ , which indicates that  $a_i$  is redundant and cannot be in the reduct of  $U^-$ . In Step 8,  $B$  is assigned to  $RED_{U^-}$ . In Step 9,  $RED_{U^-}$  is returned as the result of Algorithm 3.

The time complexity of Algorithm 3 is  $O(|C|^2(n-1))$ . According to Definition 2 and Proposition 7, the time complexity of Step 2 is  $O((|C|+1)(|U|-1))$ . Then, similarly, the time complexity of Steps 3–9 is  $O(|C|^2(|U|-1))$ . Hence, the time complexity of Algorithm 3 is  $O((|C|+1)(|U|-1) + |C|^2(|U|-1)) \approx O(|C|^2(|U|-1))$  based on the foregoing analysis.

Since the number of objects is  $|U| - 1$ , the time complexity of THA is actually  $O(|C|^2(|U|-1)^2)$  when deleting the object  $u_n$  from the object set  $U$ . Because  $|C|^2(|U|-1) \leq |C|^2(|U|-1)^2$ ,  $O(|C|^2(|U|-1))$  is lower than  $O(|C|^2(|U|-1)^2)$ . Hence, KGIRD spends much less time than THA.

## 4 Empirical experiments

### 4.1 A description of data sets and experimental environment

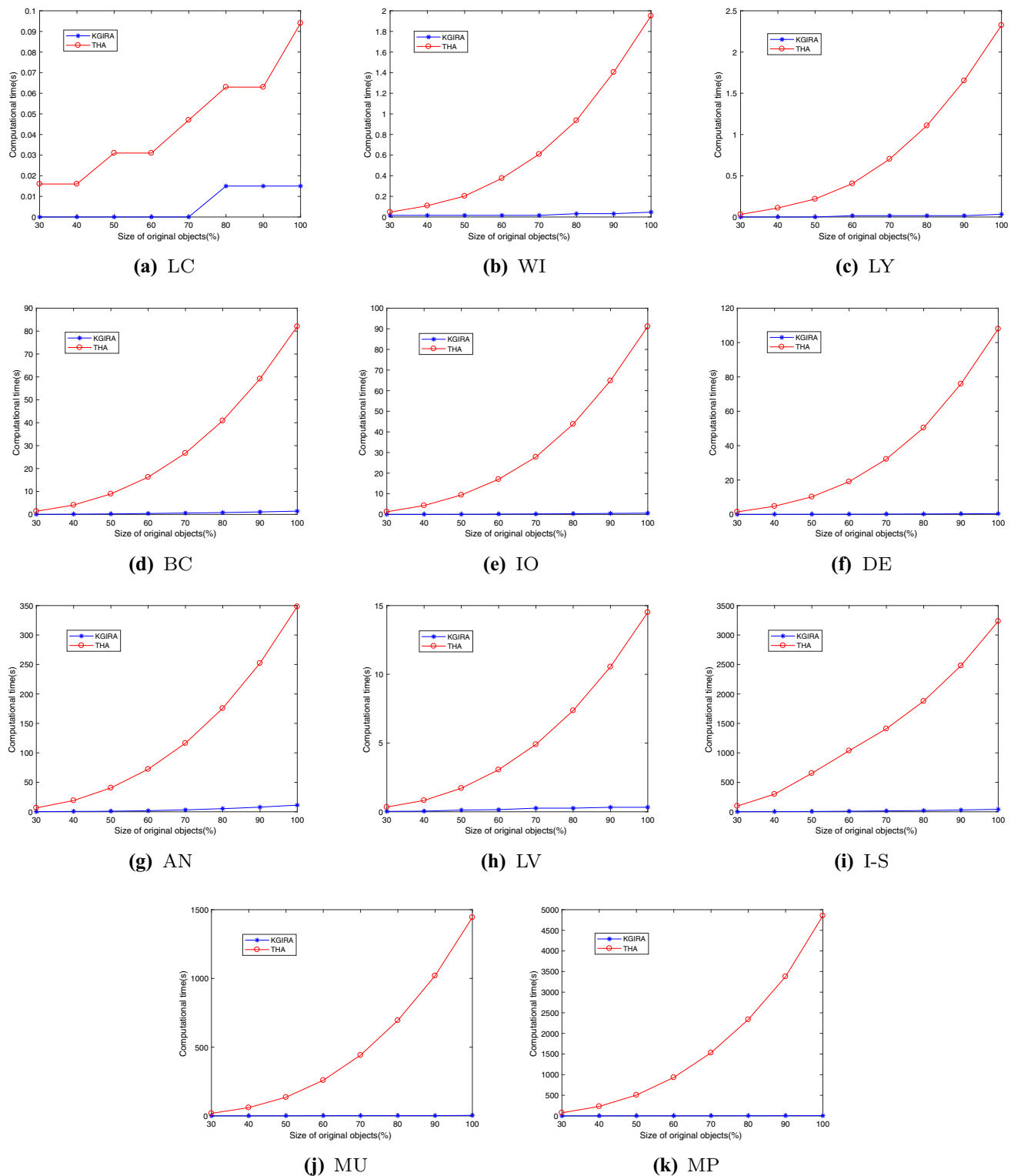
In this section, the proposed incremental attribute reduction approach is tested on several real-life data sets available from the University of California Irvine (UCI) Repository of Machine Learning Database (Dua and Taniskidou 2017). The characteristics of the data sets are summarized in Table 2. For a complete data set, 5% of the attribute values randomly selected from the data set are converted into missing values, which makes them into incomplete decision systems. The proposed algorithms are coded in Eclipse IDE for Java Developers with Neon.3 Release (4.6.3) version using JDK 1.8.0\_111 version and they have been carried out on a personal computer with the following specification: Intel Core i5-3570 3.4 GHz CPU, 4.0 GB of memory, and 64-bit Win7.

### 4.2 Performance comparison between algorithm KGIRA and algorithm THA

In the experiments, each original data set is divided into ten parts averagely according to the number of objects. 20% of each data set is used for basic data set and 80% remaining objects is used for incremental data set added one by one in

**Table 2** Description of data sets

	Data sets	Data type	Row	Attribute	Class
1	Lung cancer (LC)	Incomplete	32	56	3
2	Wine (WI)	Complete	178	13	3
3	Lymphography (LY)	Complete	148	18	4
4	Breast Cancer Wisconsin (BC)	Incomplete	699	9	2
5	Ionosphere (IO)	Complete	351	34	2
6	Dermatology (DE)	Complete	366	33	6
7	Anneal (AN)	Incomplete	798	38	6
8	LSVT voice (LV)	Complete	126	310	2
9	Image segmentation (I-S)	Incomplete	2310	19	7
10	Musk1 (MU)	Complete	476	166	2
11	Mice protein (MP)	Incomplete	1080	80	8
12	Splice-junction gene (SP)	Complete	3190	60	3
13	Multiple features (MF)	Complete	2000	216	10



**Fig. 1** Results of execution for KGIRA and THA on data sets from UCI

subsequent steps. Objects in incremental data set are added to basic data set one by one. Once the number of objects added reaches 10% of the original data set, the time spent is recorded, until all objects in incremental data set are added

to basic data set. In each subfigure of Fig. 1, the  $x$ -axis is the percent of objects that exist in basic data set and the  $y$ -axis is the value of computational time. Circle marked lines are

**Table 3** Comparison of reducts between algorithm THA and algorithm KGIRA

Data sets		KGIRA (Alg. 2)		THA (Alg. 1)	
		LR	Reduct	LR	Reduct
1	LC	6	2, 1, 5, 4, 3, 7	4	5, 2, 11, 14
2	WI	5	1, 3, 0, 2, 6	5	10, 5, 0, 4, 1
3	LY	8	17, 13, 12, 14, 0, 1, 7, 5	7	13, 17, 12, 1, 14, 15, 0
4	BC	5	5, 0, 2, 3, 1	5	5, 1, 0, 3, 2
5	IO	10	21, 18, 6, 7, 16, 19, 5, 2, 11, 13	8	32, 23, 4, 33, 18, 17, 6, 13
6	DE	13	15, 3, 18, 2, 14, 4, 6, 1, 9, 5, 12, 17, 8	10	8, 3, 15, 18, 2, 31, 27, 1, 4, 13
7	AN	13	2, 3, 4, 7, 8, 11, 32, 33, 34, 31, 6, 16, 36	13	2, 3, 4, 6, 7, 8, 11, 16, 31, 32, 33, 34, 36
8	LV	7	82, 3, 1, 13, 6, 34, 8	3	86, 300, 51
9	I-S	8	0, 1, 17, 13, 15, 11, 5, 14	8	0, 1, 5, 13, 14, 15, 17, 10
10	MU	14	1, 13, 130, 3, 5, 7, 10, 2, 14, 0, 12, 8, 22, 16	11	91, 5, 87, 107, 13, 100, 93, 150, 14, 157, 1
11	MP	8	46, 76, 30, 48, 2, 1, 3, 4	7	46, 76, 49, 21, 58, 64, 1
12	SP	13	24, 33, 52, 43, 2, 1, 3, 0, 7, 4, 8, 5, 6	12	34, 21, 2, 17, 5, 7, 0, 18, 25, 9, 26, 1
13	MF	10	92, 151, 12, 179, 2, 4, 1, 5, 0, 3	8	54, 9, 56, 151, 10, 71, 130, 4

**Table 4** Comparison of THA and KGIRA on classification accuracy

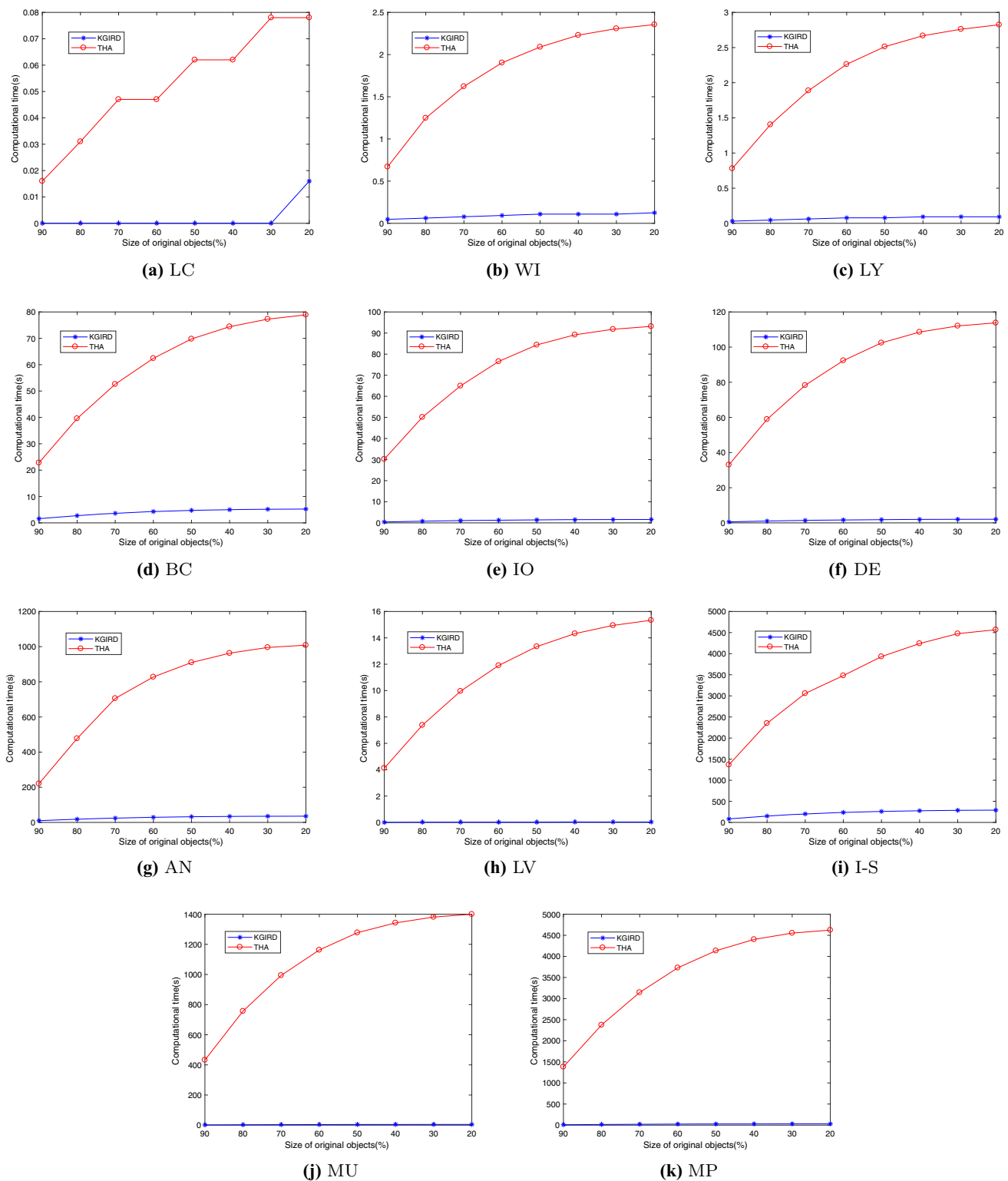
Data sets		NB		REPTree	
		KGIRA (Alg. 2)	THA (Alg. 1)	KGIRA (Alg. 2)	THA (Alg. 1)
1	LC	46.25 ± 4.15	48.75 ± 4.00	43.13 ± 5.38	43.13 ± 5.73
2	WI	92.25 ± 0.75	89.44 ± 0.83	84.04 ± 1.51	74.61 ± 2.94
3	LY	83.04 ± 1.11	80.68 ± 0.69	72.16 ± 1.83	71.69 ± 2.52
4	BC	96.34 ± 0.15	96.29 ± 0.15	94.59 ± 0.39	94.62 ± 0.37
5	IO	86.50 ± 0.57	84.56 ± 0.21	84.99 ± 1.26	86.98 ± 0.86
6	DE	93.52 ± 0.32	84.95 ± 0.41	80.08 ± 1.54	63.69 ± 1.83
7	AN	77.05 ± 0.30	77.14 ± 0.40	77.99 ± 0.44	77.80 ± 0.39
8	LV	72.54 ± 1.51	75.32 ± 2.11	69.92 ± 2.64	70.56 ± 1.64
9	I-S	88.26 ± 0.25	89.54 ± 0.07	91.02 ± 0.39	92.27 ± 0.22
10	MU	69.85 ± 0.82	72.16 ± 0.82	68.95 ± 1.59	71.28 ± 1.37
11	MP	66.47 ± 0.56	64.17 ± 0.58	64.19 ± 0.76	62.94 ± 1.33
12	SP	59.21 ± 0.17	67.09 ± 0.17	56.48 ± 0.80	62.86 ± 0.56
13	MF	83.35 ± 0.20	78.84 ± 0.21	64.43 ± 0.60	58.28 ± 0.82
Average		78.05	77.61	73.23	71.59

computational time of algorithm THA and asterisk marked lines are computational time of algorithm KGIRA.

From Fig. 1, for each data set, the computational time of algorithm KGIRA and algorithm THA increase when the number of objects added to decision system increases. The computational time of algorithm KGIRA is much smaller than that of the algorithm THA. Moreover, the computational time of algorithm KGIRA on the last two data sets, SP and MF, is 174.954 s and 57.221 s, respectively. The computational time of algorithm THA on the same two data sets are both more than 8 h. Therefore, the comparison of the computational time of algorithm KGIRA and algorithm THA on data sets SP and MF is not shown in the graph.

In Table 3, we can find that some reducts obtained by algorithm KGIRA are not the same as those obtained by algorithm THA. Moreover, the number of selected attributes in reduct, also called the length of reduct (LR), which obtained by algorithm KGIRA is a little more than that obtained by algorithm THA on most of data sets. It is because that algorithm KGIRA gets the reduct based on the previous result.

In Table 4, the precision of classification is calculated on selecting the reducts obtained by the algorithms THA and KGIRA. The classification accuracy is calculated by Naive Bayes Classifier (NB) and Decision Tree algorithm (REPTree) with fivefold cross validation. From Table 4, it is clear that the average classification accuracy of the



**Fig. 2** Results of execution for algorithm KGIRD and algorithm THA on data sets from UCI

reducts found by incremental algorithm KGIRA is better than those by algorithm THA. The experimental results show that the incremental algorithm KGIRA can discover a better attribute reduct compared with algorithm THA from

the viewpoint of average classification performance. Moreover, incremental algorithm KGIRA can discover a feasible attribute reduct within a rather shorter calculation time.

**Table 5** Comparison of reducts between algorithm THA and algorithm KGIRD

Data sets		KGIRD (Alg.3)		THA (Alg.1)	
		LR	Reduct	LR	Reduct
1	LC	2	5, 2	2	33, 2
2	WI	3	10, 1, 0	3	9, 5, 0
3	LY	5	17, 12, 1, 14, 0	4	17, 13, 12, 14
4	BC	3	5, 1, 0	3	1, 5, 0
5	IO	5	33, 18, 6, 17, 7	5	22, 17, 15, 25, 21
6	DE	7	8, 3, 15, 18, 27, 1, 4	6	15, 3, 18, 2, 17, 14
7	AN	9	2, 3, 4, 7, 8, 11, 32, 33, 34	9	2, 3, 4, 7, 8, 11, 32, 33, 34
8	LV	2	86, 51	2	82, 4
9	I-S	5	0, 1, 14, 15, 17	6	0, 1, 17, 13, 15, 11
10	MU	7	91, 5, 87, 13, 93, 157, 1	5	96, 1, 13, 149, 130
11	MP	6	46, 76, 49, 21, 58, 64	5	46, 76, 30, 48, 0
12	SP	10	34, 21, 2, 7, 18, 9, 26, 3, 10, 28	9	21, 24, 7, 3, 33, 52, 42, 51, 0
13	MF	7	54, 9, 151, 10, 71, 130, 4	6	92, 151, 12, 198, 179, 2

### 4.3 Performance comparison between algorithm KGIRD and algorithm THA

In the experiment, each original data set is used for basic data set and divided into ten parts averagely according to the number of objects. Objects are deleted from the basic data set one by one until 20% of the original data set remains. Once the number of objects deleted reaches 10% of the original data set, the time spent is recorded. In each subfigure of Fig. 1, the  $x$ -axis is the size of basic data set and the  $y$ -axis is the value of computational time. Circle marked lines are computational time of algorithm THA and asterisk marked lines are computational time of algorithm KGIRD.

From Fig. 2, for each data set, the computational time of algorithm KGIRD and algorithm THA increase when the

number of objects deleted to decision system increases. The computational time of algorithm KGIRD is much smaller than that of the algorithm THA. Moreover, the computational time of algorithm KGIRD on the last two data sets, SP and MF, is 1777.764 s and 217.447 s, respectively. The computational time of algorithm THA on the same two data sets is both more than eight hours. Therefore, the comparison of the computational time of algorithm KGIRD and algorithm THA on data sets SP and MF is not shown in the graph.

In Table 5, we can find that nearly, half of reducts obtained by algorithm KGIRD are not the same as those obtained by algorithm THA. It is because algorithm KGIRD gets the reduct based on the previous result and algorithm THA computes the result from the beginning. Furthermore, reducts obtained by algorithm KGIRD from

**Table 6** Comparison of THA and KGIRD on classification accuracy

Data sets		NB		REPTree	
		KGIRD (Alg. 3)	THA (Alg. 1)	KGIRD (Alg. 3)	THA (Alg. 1)
1	LC	50.00 $\pm$ 0.00	66.67 $\pm$ 0.00	25.00 $\pm$ 15.37	26.67 $\pm$ 15.28
2	WI	66.86 $\pm$ 3.88	79.43 $\pm$ 2.80	42.57 $\pm$ 4.51	44.29 $\pm$ 6.29
3	LY	87.59 $\pm$ 3.52	79.31 $\pm$ 5.34	63.79 $\pm$ 6.02	61.03 $\pm$ 6.18
4	BC	97.48 $\pm$ 0.36	97.27 $\pm$ 0.29	93.60 $\pm$ 0.82	93.45 $\pm$ 1.22
5	IO	75.14 $\pm$ 1.83	79.00 $\pm$ 3.26	73.71 $\pm$ 3.14	76.43 $\pm$ 3.01
6	DE	78.22 $\pm$ 2.41	63.97 $\pm$ 2.04	49.32 $\pm$ 4.58	45.89 $\pm$ 5.24
7	AN	81.07 $\pm$ 0.44	81.07 $\pm$ 0.91	80.75 $\pm$ 0.42	80.44 $\pm$ 1.11
8	LV	71.60 $\pm$ 3.77	87.60 $\pm$ 3.32	70.40 $\pm$ 4.80	86.00 $\pm$ 3.22
9	I-S	80.67 $\pm$ 0.81	86.26 $\pm$ 0.86	68.48 $\pm$ 1.03	73.48 $\pm$ 2.02
10	MU1	65.26 $\pm$ 2.70	68.21 $\pm$ 0.92	57.26 $\pm$ 3.27	57.37 $\pm$ 3.06
11	MP	47.18 $\pm$ 2.45	59.63 $\pm$ 1.97	37.87 $\pm$ 2.53	37.73 $\pm$ 2.11
12	SP	73.93 $\pm$ 0.55	59.51 $\pm$ 1.12	70.72 $\pm$ 1.46	56.60 $\pm$ 1.10
13	MF	66.38 $\pm$ 1.64	64.20 $\pm$ 1.00	42.13 $\pm$ 1.66	37.10 $\pm$ 2.43
Average		72.41	74.78	59.66	59.73



Table 5 are included in that obtained by algorithm THA from Table 3 on most data sets. This is because the original reducts are obtained by algorithm THA on the original data sets, and then, algorithm KGIRD gets the reducts based on the previous result.

In Table 6, the classification performance is calculated on the reducts obtained by algorithms THA and KGIRD. The results of classification accuracy are calculated by Naive Bayes Classifier (NB) and Decision Tree algorithm (REPTree) with fivefold cross validation. From Table 6, it is clear that the average classification accuracy of the reducts found by incremental algorithm KGIRD are little lower than those by algorithm THA, due to the bad performance on the first data set LC. The experimental results show that the incremental algorithm KGIRD can discover a satisfactory attribute reduct from the viewpoint of classification performance. Moreover, incremental algorithm KGIRD can discover a feasible attribute reduct within a rather shorter calculation time.

## 5 Conclusion

In this paper, we use knowledge granularity to measure the uncertainty and the importance of attributes in incomplete decision systems. Based on this measure, incremental reduction algorithms for incomplete decision systems are constructed when adding multiple objects and deleting multiple objects one by one, respectively. To test the effectiveness of the constructed attribute selection approaches based on knowledge granularity, experiments on several real-life data from UCI data sets are conducted. Results show that the proposed approaches are effective to reduce the numbers of attributes obviously and incremental approaches are more efficient to update attribute reducts when the objects vary in incomplete decision systems than the non-incremental approach. Compared with existing methods (Jing et al. 2016; Yang et al. 2017), our approaches can deal with incomplete decision systems, which are more complicated. We plan to study knowledge granularity-based incremental attribute reduction solutions for incomplete decision systems under the situation of adding or deleting multiple objects at a time. Furthermore, it is worth of future research to use knowledge granularity to deal with the attribute reduction problem in fuzzy rough sets.

**Acknowledgements** This work was partially supported by the National Natural Science Foundation of China (Nos. 61473259, 61070074, and 60703038), and the Hunan Provincial Science and Technology Project Foundation (2018TP1018 and 2018RS3065).

## References

- Afridi MK, Azam N, Yao JT, Alanazi E (2018) A three-way clustering approach for handling missing data using gtrs. *Int J Approx Reason* 98:11–24
- Chen SM, Chang YC (2011) Weighted fuzzy rule interpolation based on ga-based weight-learning techniques. *IEEE Trans Fuzzy Syst* 19(4):729–744
- Chen SM, Tanuwijaya K (2011) Fuzzy forecasting based on high-order fuzzy logical relationships and automatic clustering techniques. *Expert Syst Appl* 38(12):15425–15437
- Chen SM, Manalu GM, Pan JS, Liu HC (2013) Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and particle swarm optimization techniques. *IEEE Trans Cybern* 43(3):1102–1117
- Chen DG, Yang YY, Dong Z (2016) An incremental algorithm for attribute reduction with variable precision rough sets. *Appl Soft Comput* 45:129–149
- Cheng SH, Chen SM, Jian WS (2016) Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. *Inf Sci* 327:272–287
- Dai JH, Tian HW (2013) Entropy measures and granularity measures for set-valued information systems. *Inf Sci* 240(11):72–82
- Dai J, Xu Q (2013) Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Appl Soft Comput* 13(1):211–221
- Dai JH, Wang WT, Xu Q (2013) An uncertainty measure for incomplete decision tables and its applications. *IEEE Trans Cybern* 43(4):1277–1289
- Dai JH, Hu QH, Zhang JH, Hu H, Zheng NG (2017a) Attribute selection for partially labeled categorical data by roughset approach. *IEEE Trans Cybern* 47(9):2460–2471
- Dai JH, Wei BJ, Zhang XH, Zhang QL (2017b) Uncertainty measurement for incomplete interval-valued information systems based on  $\alpha$ -weak similarity. *Knowl Based Syst* 136:159–171
- Dai JH, Hu H, Wu WZ, Qian YH, Huang DB (2018) Maximal discernibility pairs based approach to attribute reduction in fuzzy rough sets. *IEEE Trans Fuzzy Syst* 26(4):2174–2187
- Das AK, Das S, Ghosh A (2017) Ensemble feature selection using bi-objective genetic algorithm. *Knowl Based Syst* 123:116–127
- D'Eer L, Cornelis C, Yao YY (2016) A semantically sound approach to pawlak rough sets and covering-based rough sets. *Int J Approx Reason* 78:62–72
- Derrac J, Cornelis C, García S, Herrera F (2012) Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection. *Inf Sci* 186:73–92
- Du WS, Hu BQ (2016) Dominance-based rough set approach to incomplete ordered information systems. *Inf Sci* 346:106–129
- Dua D, Taniskidou EK (2017) UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Formica A (2012) Semantic web search based on rough sets and fuzzy formal concept analysis. *Knowl Based Syst* 26(9):40–47
- Jing YG, Li TR, Huang JF, Zhang YY (2016) An incremental attribute reduction approach based on knowledge granularity under the attribute generalization. *Int J Approx Reason* 76:80–95
- Jothi G, Hannah IH (2016) Hybrid tolerance rough set-firefly based supervised feature selection for mri brain tumor image classification. *Appl Soft Comput* 46:639–651
- Koley S, Sadhu AK, Mitra P, Chakraborty B, Chakraborty C (2016) Delineation and diagnosis of brain tumors from post contrast T1-weighted MR images using rough granular computing and random forest. *Appl Soft Comput* 41:453–465
- Kryszkiewicz M (1998) Rough set approach to incomplete information systems. *Inf Sci* 112(1):39–49

- Li YL, Tang JF, Kwaisang C, Han Y, Luo XG (2012) A rough set approach for estimating correlation measures in quality function deployment. *Inf Sci* 189(7):126–142
- Li H, Li DY, Zhai YH, Wang SG, Zhang J (2016) A novel attribute reduction approach for multi-label data based on rough set theory. *Inf Sci* 367–368:827–847
- Li FC, Yang JN, Jin CX, Guo CM (2017) A new effect-based roughness measure for attribute reduction in information system. *Inf Sci* 378:348–362
- Liang JY, Wang F, Dang CY, Qian YH (2014) A group incremental approach to feature selection applying rough set technique. *IEEE Trans Knowl Data Eng* 26(2):294–308
- Liao SJ, Zhu QX, Qian YH, Lin GP (2018) Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs. *Knowl Based Syst* 158:25–42
- Lin F, Li TR, Da R, Gou SR (2011) A vague-rough set approach for uncertain knowledge acquisition. *Knowl Based Syst* 24(6):837–843
- Lin YJ, Li YW, Wang CX, Chen JK (2018) Attribute reduction for multi-label learning with fuzzy rough set. *Knowl Based Syst* 152:51–61
- Liu D, Liang DC, Wang CC (2016) A novel three-way decision model based on incomplete information system. *Knowl Based Syst* 91:32–45
- Liu H, Cocea M, Ding WL (2018) Multi-task learning for intelligent data processing in granular computing context. *Granul Comput* 3(3):257–273
- Min F, Xu J (2016) Semi-greedy heuristics for feature selection with test cost constraints. *Granul Comput* 1(3):199–211
- Min F, He HP, Qian YH, Zhu W (2011) Test-cost-sensitive attribute reduction. *Inf Sci* 181(22):4928–4942
- Pacheco F, Cerrada M, Sanchez RV, Cabrera D, Li C, Oliveira JVD (2017) Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery. *Expert Syst Appl* 71:69–86
- Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11(5):341–356
- Pawlak Z (1991) Rough sets: theoretical aspect of reasoning about data. Kluwer Academic Publishers, Dordrecht
- Pawlak Z, Skowron A (2007) Rudiments of rough sets. *Inf Sci* 177(1):3–27
- Qian YH, Liang JY (2008) Combination entropy and combination granulation in rough set theory. *Int J Uncert Fuzziness Knowl Based Syst* 16(2):179–193
- Raza MS, Qamar U (2016) An incremental dependency calculation technique for feature selection using rough sets. *Inf Sci* 343–344:41–65
- Raza MS, Qamar U (2017) Feature selection using rough set-based direct dependency calculation by avoiding the positive region. *Int J Approx Reason* 92:175–197
- Shu WH, Shen H (2013) A rough-set based incremental approach for updating attribute reduction under dynamic incomplete decision systems. In: *Proceedings of 2013 IEEE international conference on fuzzy systems*, IEEE, pp 1–7
- Shu WH, Shen H (2014a) Incremental feature selection based on rough set in dynamic incomplete data. *Patt Recognit* 47(12):3890–3906
- Shu WH, Shen H (2014b) Updating attribute reduction in incomplete decision systems with the variation of attribute set. *Int J Approx Reason* 55(3):867–884
- Tiwari AK, Shreevastava S, Som T, Shukla KK (2018) Tolerance-based intuitionistic fuzzy-rough set approach for attribute reduction. *Expert Syst Appl* 101:205–212
- Wafu Soh C, Njilla LL, Kwiat KK, Kamhoua CA (2018) Learning quasi-identifiers for privacy-preserving exchanges: a rough set theory approach. *Granul Comput*. <https://doi.org/10.1007/s41066-018-0127-0>
- Wang CZ, Huang Y, Shao MW, Chen DG (2019a) Uncertainty measures for general fuzzy relations. *Fuzzy Sets Syst* 360:82–96
- Wang CZ, Huang Y, Shao MW, Fan XD (2019b) Fuzzy rough set-based attribute reduction using distance measures. 164:205–212
- Wang F, Liang JY, Dang CY (2013) Attribute reduction for dynamic data. *Appl Soft Comput* 13(1):676–689
- Wang CZ, Shao MW, He Q, Qian YH, Qi YL (2016) Feature subset selection based on fuzzy neighborhood rough sets. *Knowl Based Syst* 111:173–179
- Wang CZ, He Q, Shao MW, Hu QH (2018) Feature selection based on maximal neighborhood discernibility. *Int J Mach Learn Cybern* 9(11):1929–1940
- Wei W, Wu XY, Liang JY, Cui JB, Sun YJ (2018) Discernibility matrix based incremental attribute reduction for dynamic data. *Knowl Based Syst* 140:142–157
- Xie XJ, Qin XL (2018) A novel incremental attribute reduction approach for dynamic incomplete decision systems. *Int J Approx Reason* 93:443–462
- Xu WH, Yu JH (2017) A novel approach to information fusion in multi-source datasets: a granular computing viewpoint. *Inf Sci* 378:410–423
- Xu WH, Li WT, Zhang XT (2017) Generalized multigranulation rough sets and optimal granularity selection. *Granul Comput* 2(4):271–288
- Yang YY, Chen DG, Hui W (2017) Active sample selection based incremental algorithm for attribute reduction with rough sets. *IEEE Trans Fuzzy Syst* 25(4):825–838
- Yao YY, Zhang XY (2017) Class-specific attribute reducts in rough set theory. *Inf Sci* 418:601–618
- Zhan JM, Liu Q, Herawan T (2017) A novel soft rough set: soft rough hemirings and its multicriteria group decision making. *Appl Soft Comput* 54:393–402

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.