Covering rough set-based incremental feature selection for mixed decision system

Yanyan Yang · Degang Chen · Xiao Zhang · Zhenyan Ji

Accepted: 30 November 2021 / Published online: 31 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Covering rough sets conceptualize different types of features with their respective generated coverings. By integrating these coverings into a single covering, covering rough set-based feature selection finds valuable features from a mixed decision system with symbolic, real-valued, missing-valued and set-valued features. Existing approaches to covering rough set-based feature selection, however, are intractable to handle large mixed data. Therefore, an efficient strategy of incremental feature selection is proposed by presenting a mixed data set in sample subsets one after another. Once a new sample subset comes in, the relative discernible relation of each feature is updated to disclose incremental feature selection scheme that decides the strategies of increasing informative features and removing redundant features. The incremental scheme is applied to establish two incremental feature selection algorithms from large or dynamic mixed data sets. The first algorithm updates the feature subset upon the sequent arrival of sample subsets and returns the reduct when no further sample subsets are obtained. The second one merely updates the relative discernible relations and finds the reduct when no subsets are obtained. Extensive experiments demonstrate that the two proposed incremental algorithms, especially the second one speeds up covering rough set-based feature selection without sacrificing too much classification performance.

Keywords Covering rough sets · Feature selection · Incremental learning · Dynamic mixed data

1 Introduction
Rough set theory (Pawlak 1982) is a frequently used mathematical solution for conceptualizing uncertain, insufficient knowledge hidden in data. Its theoretical foundation is the equivalence relation where samples sharing identical feature descriptions are said to be equivalent or indiscernible. As a result, rough set theory is limited to only process data sets with symbolic features rather than other types of features, like real-valued, missing-valued and set-valued ones. It is not uncommon that an actual data set (called mixed decision system or mixed data in this paper) usually includes different types of features. Typically, some of them are redundant or irrelevant to the classifier methods and may worsen the classification performance (Chen and Yang 2014; Zhang et al. 2016). Hence, it is desirable to choose informative features from mixed data to improve the performance of the learning algorithms.

Currently, much effort has been made to extend rough sets to process other types of features. For example, similarity relation (Stefanowski and Tsoukias 1999) was employed to handle missing-valued data sets. To process set-valued data sets, Guan and Wang (2006) and Qian et al. (2009) presented the tolerance relation and the binary dominance relation, respectively. Neighborhood relation was defined in Yao (1998, 2006); Wang et al. (2016) to handle real-valued data sets. Based on the above binary relations, various types of rough set models, such as neighborhood rough set model, were developed in Slowinski and Vanderpooten (2000); Qian et al. (2012), Hu et al. (2008), Wang et al. (2018). By using the fuzzy relation, additionally, fuzzy rough sets were proposed to work on real-valued data sets (Dubois and Prade 1990; Jensen and Shen 2004, 2007; Tsang et al. 2008; Wang et al. 2016). Actually, a covering of a universe is composed
of the family of subsets of the universe generated by every aforesaid relation, where \( \alpha - \) cut sets are used to induce a covering of the universe based on the fuzzy relation (Chen et al. 2007; Du et al. 2011). Thus, these models just fall into the category of covering rough sets introduced by Zakowski (1983). Aggregating the coverings induced by various types of features into a single covering, covering rough set can be viewed as a method of handling mixed data sets.

Feature selection, as one important application of rough sets, has found its way to the domains of data mining and uncertainty reasoning. Compared with the existing feature selection methods, rough set-based feature selection does not require any prior knowledge, like the number of selected features and the probability distribution. Some feature selection algorithms in the field of rough sets were presented to tackle various kinds of data sets. These rough set methods are roughly split into the discernibility-based ones and the uncertainty measure-based ones. The typical representative of the discernibility-based algorithm is the discernibility matrix-based method. For symbolic data sets, Skowron and Rauszer (1992) calculated a reduct from a symbolic data set by introducing the discernibility matrix to construct the discernibility function. To improve this discernibility matrix-based approach, sample pair selection in Chen et al. (2012) was presented to obtain all minimal elements with no need for finding the discernibility matrix such that the searching range and the runtime are narrowed. For real-valued data sets, the discernibility matrix-based method is extended to fuzzy rough sets in Tsang et al. (2008). For mixed data sets, Chen et al. (2007) generalized the discernibility matrix of the classical case to that of covering rough set, where the discernibility function was formed to compute all reducts. Wang et al. (2014) simplified the complex discernibility matrices to more simpler ones and designed a discernibility matrix-based heuristic approach to reducts. A simpler discernibility matrix of covering decision system which essentially has the same form with the one of Wang et al. (2014) was also constructed in Wang et al. (2015).

The uncertainty measure-based algorithm can obtain a proper reduct (Chen et al. 2012). However, it has the complex structures and is not suitable for large data sets. To improve the computation of a reduct, the uncertainty measure-based algorithm uses different uncertainty measures to find a reduct (Slezak 2002; Wang et al. 2002; Liang and Xu 2002). For symbolic data sets, Qian et al. (2010) developed the positive approximation method to accelerate the heuristic process of finding reducts by using three representative information entropy measures. For real-valued data sets, a feature selection method based on fuzzy rough set was presented in Jensen and Shen (2004) to preserve the dependence function. Hu et al. (2006) presented a feature selection method that applies the information entropy to evaluate the significance of a feature subset. For heterogeneous data sets with symbolic and real-valued features, neighborhood rough set model (Hu et al. 2008) was presented to tackle heterogeneous data sets by assigning different thresholds to distinct types of features. Chen and Yang (2014) proposed a heuristic feature selection method by the discernibility relations of heterogeneous features. For mixed data sets with different types of features, Dong et al. (2016) employed the relative discernible relation to calculate minimal elements and developed algorithms for reducts which were experimentally demonstrated to be effective.

The above feature selection algorithms have been extensively studied from the theoretical perspective. However, they have the intensive computation cost when handling large mixed data. An intuitive solution to tackle mixed data is to split it into a family of sample subsets which arrive successively, and then apply the incremental technique to handle sample subsets one by one. Such mixed data sets are sometimes referred to as dynamic mixed data in this paper. As a newly acquired sample subset comes in, the above-mentioned non-incremental feature selection methods require to re-select a valuable feature subset with all historical data samples plus the newly arrived samples (Yang et al. 2014; Li et al. 2007; Chen et al. 2014). As a consequence, these non-incremental methods are uneconomical to deal with dynamic mixed data as older data sets are repeatedly processed. Conversely, incremental feature selection methods update a feature subset by making the best of the previously learned results from the historical data samples, with no need for re-computing the feature subset using the combined data samples of the historical samples and the newly acquired samples (Chen et al. 2015, 2013).

Rough set-based incremental feature selection (Sang et al. 2021, 2020, 2021) has been investigated in a natural scenario where a single sample or a sample group varies. For a single sample arriving, the discernibility matrix was updated to incrementally select a feature subset in Orlowska and Orlowski (1992). However, Hu et al. (2005) experimentally demonstrated the high computational complexity of the method in Orlowska and Orlowski (1992). Thus, Hu et al. (2007) investigated the incremental mechanisms for the positive region to develop an incremental feature selection algorithm. The modified discernibility matrix was incrementally updated in Ming (2007) to compute all reducts. The core was incrementally computed in Feng and Zhang (2012) to expedite the acquisition of a reduct. The minimal elements in variable precision rough sets were updated in Chen et al. (2016) to study the incremental feature selection process. When a sample group dynamically varies, some incremental algorithms were studied based on rough sets. For example, Shu and Shen (2013) incrementally calculated the positive region to update attribute reduction from dynamic incomplete decision systems. Liang et al. (2014) proposed a group incremental feature selection algorithm by incrementally...
finding three representative information entropy. Yang et al. (2017, 2018) investigated incremental feature selection algorithms based on fuzzy rough sets by utilizing the strategies of increasing and removing features. Considering the utility of incoming samples, a novel incremental feature selection method was presented in Yang et al. (2017) by combining the active sample selection process with the incremental feature selection scheme. When adding and deleting samples, the covering element reduction was performed in Lang et al. (2015) by incrementally finding lower and upper approximations of sets. When varying samples, Jing et al. presented incremental feature selection algorithms by knowledge granularity (Jing et al. 2016) and multi-granulation view (Jing et al. 2017). An active incremental feature selection algorithm was presented in Zhang et al. (2020) using fuzzy rough set-based information entropy with incoming representative instances.

The above-mentioned incremental feature selection methods are effective to select a feature subset based on different rough set models. However, they are confined to homogeneous data sets. When processing mixed data sets, the above incremental methods have to preprocess a mixed data set into a homogeneous one. This easily brings about the information loss. Therefore, they cannot effectively meet the challenge that mixed data samples are continuously added in. Inspired by these observations, we investigate incremental feature selection algorithms based on covering rough sets by simulating a mixed data set with a sequence of sample subsets. The major contributions of this paper include as follows.

(i) The relative discernible relation of each feature is calculated to characterize the core and feature selection in covering rough sets.

(ii) Upon a sample subset arriving, we apply the incremental mode to renew the discernible relation of every feature, which is valid for all sequent mixed data samples.

(iii) Using the renewed discernible relations, we investigate incremental feature selection scheme that reveals the strategies of increasing informative features and removing redundant features.

(iv) The incremental scheme is applied to present two incremental algorithms of feature selection: updating the optimal feature subset as sample subsets arrive and finding the optimal feature subset when no sample subsets are obtained.

(v) The comparison results show that the two proposed incremental algorithms, especially the second one expedites the computation of the optimal feature subset from mixed data without sacrificing too much classification performance.

The outline of the paper is given below. We introduce the preliminaries related to our paper in Sect. 2. Section 3 employs the relative discernible relation to present a feature selection method in covering rough sets. Two incremental feature selection algorithms are developed in Sect. 4. Experimental comparisons are conducted in Sect. 5. This paper is summarized in Sect. 6.

2 Preliminaries

This section lists some basic concepts used in the paper.

2.1 Covering rough sets

Definition 1 (Bonikowski et al. 1998) Let \( U \) be a universe, \( C = \{K_1, \ldots, K_t\} \) be a non-empty subset family of \( U \). \( C \) is said to be a covering of \( U \) if \( \bigcup_{i=1}^{t} K_i = U \).

A covering of the universe is apparently its partition. As noted in Chen et al. (2007), Couso and Dubois (2011), Chen et al. (2014), a covering can be generated by various kinds of features, like symbolic features, real-valued features, set-valued features, missing-valued features and so on. In detail, this paper employs the equivalence relation (Pawlak 1982) to generate the covering of a symbolic feature \( a \), and its covering is just \( C_a = \{[x]_a : x \in U \} \) where \([x]_a = \{y \in U : a(x) = a(y)\}\). The neighborhood relation (Yao 1998) is employed to generate the covering of a real-valued feature \( a \), and its covering is \( C_a = \{\delta_a(x) \in U \} \) where \(\delta_a(x) = \{y \in U : \Delta_a(x, y) \leq \delta\}\). Here, \(\delta\) is the neighborhood radius and \(\Delta_a(x, y)\) is the distance between \(x\) and \(y\) with respect to \(a\). Tolerance relation (Guan and Wang 2006) is used to generate the covering of a set-valued feature \( a \), and its covering is computed as \( C_a = \{T_a(x) \in U \} \) where \(T_a(x) = \{y \in U : a(x) \cap a(y) \neq \emptyset\}\). We utilize the characteristic relation (Grzymała-Busse 2004) to induce the covering of a missing-valued feature where ? and * represent the lost value and ‘do not care.’ So, the covering of a missing-valued feature \( a \) is calculated as \( C_a = \{K_a(x) \in U \} \) where \(K_a(x) = \{y \in U : (a(x) \neq ?) \land (a(x) = a(y)) \lor (a(x) = * \lor a(y) = *)\}\). For this reason, the definition of covering information system is introduced.

Definition 2 (Zhang et al. 2013) Given a universe \( U = \{x_1, \ldots, x_n\} \) and a feature set \( A \), \((U, A)\) is said to be a covering information system, if each feature \( c \in A \) can induce a covering of \( U \).

Definition 3 (Chen et al. 2007) Given a covering \( C = \{K_1, K_2, \ldots, K_t\} \) generated by \( c \in A \) and \( x \in U \), let \( \langle x \rangle_c = \bigcap\{K_j \in C : x \in K_j\} \), \( \text{Cov}(C) = \{\langle x \rangle_c : x \in U\} \) is called the induced covering of \( C \).
Definition 4 (Chen et al. 2007) Suppose \((U, A)\) is a covering information system and \(B = \{c_{i_1}, \ldots, c_{i_k}\}\) is a subset of \(A\). \(B = \{C_{i_1}, \ldots, C_{i_k}\}\) is a covering family of \(U\) induced by \(B\). Let \((x)_B = \cap \{ (x)_{c_{i_l}} \in Cov(C_{i_l}) : l = 1, \ldots, k \}\) for \(x \in U\), \(Cov(B) = \{ (x)_B : x \in U \}\) is called the induced covering of \(B\).

According to Definition 4, coverings generated by different types of features can be aggregated into a reasonable covering. Specifically, we assume \(B = \{c_{i_1}, \ldots, c_{i_k}\}\) includes symbolic, real-valued, set-valued and missing-valued features. The covering \(C_{i_j}\) of the feature \(c_{i_j}\) can be generated according to the aforementioned way of coverings induced. Applying Definition 4, the induced covering \(Cov(B) = \{ (x)_B : x \in U \}\) of \(B\) is obtained, where \((x)_B = \cap \{ (x)_{c_{i_j}} \in Cov(C_{i_j}) : l = 1, 2, \ldots, k \}\). The induced covering of the feature set can be used to construct lower and upper approximations of a sample set in covering rough sets. Therefore, it is possible to employ covering rough sets to handle mixed data with different types of features. In this paper, a mixed data set is denoted as a covering decision system defined as follows.

Definition 5 (Zhang et al. 2013) Suppose \((U, A, D)\) is a covering decision system, where \(U = \{x_1, \ldots, x_n\}\) is the universe, \(A = \{c_{i_1}, \ldots, c_{i_r}\}\) is the set of different types of features, in which each feature \(c_i\) induces a covering of \(U\), and \(D\) is the decision feature with symbolic values. The decision partition of \(U\) is denoted by \(U/D = \{[x]_D : x \in U\}\), where \([x]_D = \{y \in U : D(x) = D(y)\}\) is a decision class with the same decision value.

Definition 6 (Chen et al. 2007) Let \((U, A, D)\) be a covering decision system, \(U/D = \{D_1, \ldots, D_l\}\), and \(B \subseteq A\). Lower and upper approximations of \(D_i \in U/D\) with respect to \(B\) are defined as \(\text{B}(D_i) = \{x \in U : (x)_B \subseteq D_i\}\) and \(\text{B}^-(D_i) = \{x \in U : (x)_B \cap D_i \neq \emptyset\}\).

2.2 Covering rough set-based feature selection

Given \((U, A, D)\) and \(B \subseteq A\), the positive region of \(D\) relative to \(B\) is defined as \(pos_B(D) = \cup_{x \in U/D} \text{B}(x)\). By substituting coverings for features, covering rough set is able to be regarded as a mathematical tool to reduce superfluous features. Here, feature selection in covering rough set model is described below.

Definition 7 (Chen et al. 2007) Given a covering decision system \((U, A, D)\), \(P \subseteq A\) is said to be a reduct (also called an optimal feature subset) of \(A\), if the following conditions hold:

1. \(pos_A(D) = pos_P(D)\);
2. \(\forall c \in P, pos_A(D) \neq pos_{P-\{c\}}(D)\).

By Definition 7, the reduct of \((U, A, D)\) is a minimal feature subset preserving the positive region. The discernibility matrix-based method, as a frequently used approach to compute reducts, is given below.

Definition 8 (Wang et al. 2014) Let \(U = \{x_1, \ldots, x_n\}\). The discernibility matrix of \((U, A, D)\) is \(M(U, A, D) = (c_{ij})_{n \times n}\), in which \(c_{ij} = \{c \in A : x_j \notin pos_A(D)\}\) and \(c_{ij} \in pos_A(D)\) and \(D(x_i) \neq D(x_j)\).

The discernibility function of \((U, A, D)\) is denoted by \(f(U, A, D) = \wedge\{ c_{ij} : c_{ij} \neq \emptyset\}\). By distribution and absorption laws, the minimal form of \(f(U, A, D)\) is computed as \(f(U, A, D) = \vee_{k=1}^l B_k\) for \(B_k \subseteq A\). The collection of all the reducts of \((U, A, D)\) is \(\text{Red}(U, A, D) = \{B_1, \ldots, B_l\}\). \(\text{Core}(U, A, D) = \cap \text{Red}(U, A, D)\) is said to be the core of \((U, A, D)\). In reality, one reduct is sufficient for many practical applications. To calculate a reduct, we give the next theorems.

Theorem 1 (Wang et al. 2014) \(\text{Core}(U, A, D) = \{c_k \in A : c_{ij} = \{c_k\}\}\).

Theorem 2 (Wang et al. 2014) \(P \subseteq A\) is a reduct of \((U, A, D)\) iff it satisfies the next requirements:

1. \(\forall c_{ij} \neq \emptyset, P \cap c_{ij} \neq \emptyset\);
2. \(\forall c_k \in P, \text{there exists } c_{ij} \neq \emptyset \) that satisfies \((P - \{c_k\}) \cap c_{ij} = \emptyset\).

Theorem 1 states that the core is made up of the singletons of a discernibility matrix, while Theorem 2 develops an easy method of judging whether a subset of features is an optimal one of the covering decision system. In Theorem 2, the condition (1) shows that a reduct is enough to together discern sample pairs whose discernibility feature sets are non-empty; the condition (2) indicates that every feature in a reduct \(P\) is separatively essential. Hence, a reduct is just a minimum subset distinguishing sample pairs whose discernibility feature sets are non-empty.

3 Feature selection in covering rough sets using relative discernible relations

As mentioned in Sect. 2, covering rough sets can deal with mixed data sets. To describe feature selection of a mixed data set, this section introduces the relative discernible relation (Dong et al. 2016). By keeping the relative discernible relation of the whole feature set, we develop a feature selection algorithm in covering rough sets, in order to calculate a reduct from a mixed data set.
Definition 9 (Dong et al. 2016) Suppose $U = \{x_1, \cdots, x_n\}$. The relative discernible relation of $c \in A$ is defined as $\text{Dis}_D(\{c\}) = \{(x_i, x_j) : x_j \notin \{x_i\}\}$ for $x_i, x_j \in U$ that meet one of the two requirements: (1) $x_i \in pos_A(D)$, $x_j \notin pos_A(D)$; (2) $x_i, x_j \in pos_A(D)$, $D(x_i) \neq D(x_j)$.

The relative discernible relation of $c \in A$ is in fact the collection of sample pairs distinguished by $c$. That is, $(x_i, x_j) \in \text{Dis}_D(\{c\})$ implies that $c$ can distinguish $(x_i, x_j)$. Let $\text{Dis}_D(A) = \cup_{c \in A} \text{Dis}_D(\{c\})$. By Definition 9, $(x_i, x_j) \in \text{Dis}_D(\{c\})$ is equivalent to the fact that $c$ is in fact satisfied for $\forall c_{ij} \neq \emptyset$. The fact indicates $c_{ij} = \{c_k \in A : (x_i, x_j) \in \text{Dis}_D(\{c_k\})\}$ and $\text{Dis}_D(\{c_k\}) = \{(x_i, x_j) : c_k \in c_{ij}\}$. Based on the relative discernible relation, the next two theorems are developed to describe the core and feature selection of a mixed data set, respectively.

Theorem 3 Core($U, A, D$) = $\{c_k \in A : \exists (x_i, x_j) \in \text{Dis}_D(\{c_k\}), s.t., (x_i, x_j) \notin \text{Dis}_D(A - \{c_k\})\}$.

Proof By Theorem 1, $\forall c_k \in \text{Core}(U, A, D)$, $\exists c_{ij}$ s.t. $c_{ij} = \{c_k \in A : (x_i, x_j) \in \text{Dis}_D(\{c_k\})\}$ and $\text{Dis}_D(\{c_k\}) = \{(x_i, x_j) : c_k \in c_{ij}\}$. Based on the relative discernible relation, the next two theorems are developed to describe the core and feature selection of a mixed data set, respectively.

Theorem 4 $P \subseteq A$ is just a reduct of A iff it meets the following requirements:

1. $\text{Dis}_D(A) = \text{Dis}_D(P)$;
2. $\forall c \in P$, $\text{Dis}_D(A) \neq \text{Dis}_D(P - \{c\})$.

Proof (1) $\text{Dis}_D(A) = \text{Dis}_D(P) \iff \forall (x_i, x_j) \in \text{Dis}_D(A)$, $(x_i, x_j) \in \text{Dis}_D(P) \iff \forall (x_i, x_j) \in \text{Dis}_D(A)$, $\exists c \in P$, s.t., $(x_i, x_j) \in \text{Dis}_D(\{c\})$ $\iff \forall c_{ij} \neq \emptyset$, $\exists c \in P$, s.t., $c \in c_{ij} = \forall c_{ij} \neq \emptyset$, $P \cap c_{ij} = \emptyset$.

(2) $\forall c \in P$, $\text{Dis}_D(A) \neq \text{Dis}_D(P - \{c\}) \iff \forall c \in P$, $\exists (x_i, x_j) \in \text{Dis}_D(\{c\})$, s.t., $(x_i, x_j) \notin \text{Dis}_D(P - \{c\})$ $\iff \forall c \in P$, $\exists c_{ij} \neq \emptyset$, s.t., $(P - \{c\}) \cap c_{ij} = \emptyset$.

In light of the two aspects, this theorem is proved according to Theorem 2.

By Theorem 4, a reduct of a mixed data set is a minimum feature subset keeping the relative discernible relation of the whole set of features. Based on Theorems 3 and 4, a feature selection algorithm is designed to calculate a reduct from a mixed data set.

Algorithm 1 employs the relative discernible relation to find a reduct from a mixed data set. Step 1 calculates the induced covering generated by each feature and the induced covering generated by the feature set. Its computational complexity is $O(|U|^2 |A|)$. In Step 2, we calculate the positive region, which has the computational complexity of $O(|U| |A|)$. In Step 3, we calculate the relative discernible relations of every feature and A, which has the computational complexity of $O(|U| |A|)$. In Step 4, we calculate the core of a mixed data set, which has computational complexity of $O(|U|^2 |A|)$. In Steps 6~9, we choose a feature which discerns the most sample pairs at every iteration, which has a computational complexity of $O(|U|^2 |A|)$. In a word, the computational complexity of Algorithm 1 is $O(|U|^2 |A|)$.

As a matter of fact, Algorithm 1 is a non-incremental method of feature selection in covering rough sets, as it fails to effectively apply the previously results learned from the historical data samples as new data samples arrive. When this happens, we need to put historical (old) data set and new data set together and then run a non-incremental algorithm like our proposed Algorithm 1 on the combined data set to obtain a new reduct. This is certainly wasteful in the computation time as the results of the historical data set are effectively thrown away. In Sect. 4, we will concentrate on the incremental scheme of feature selection in covering rough sets while handling dynamic mixed data, as well as large mixed data.

4 Incremental scheme of feature selection in covering rough sets

The section presents the incremental scheme of feature selection in covering rough sets, under the assumption that a mixed data set can be segmented into a collection of sample subsets. Sample subsets successively arrive, such that feature selection can be incrementally performed in covering rough sets. The relative discernible relation of every feature is first updated upon the sequential arrival of sample subsets. The incremental feature selection scheme is further revealed to decide the strategies of increasing informative features and removing redundant features. Based on the incremental scheme, two incremental algorithms are presented to begin...
by the empty set to calculate an optimal feature subset of a mixed data set.

### 4.1 Incremental setup and symbols

In the subsection, we give the descriptions of the incremental setup and the used notations.

To use the incremental technique to obtain an optimal feature subset of \((U, A, D)\), \(U\) is segmented into a sample subset sequence \(\{U_k\}_{k=1}^m\), where every subset is said to be an incoming sample subset. The following requirements obviously hold:

1. \(U = \bigcup_{k=1}^m U_k\);
2. for \(i \neq j\), \(U_i \cap U_j = \emptyset\).

Under the incremental setting, we add sample subsets one at a time. The previously obtained historical samples are stored in a temporary set \(T\) which has the initialization of the empty set. In detail, with the arrival of \(U_1\), the temporary set \(T\) is changed into \(U_1\); upon the arrival of \(U_2\), \(T\) is updated as \(U_1 \cup U_2\); as \(U_m\) arrives, \(T\) is updated as the set of all samples \(U\). In general, \(T\) is updated with the sequential arrival of the first, \(\cdots\), \(m\)th sample subsets.

In \((T, A, D)\), \(pos_A(D)\) is the positive region of \(D\) relative to \(A\), \(Dis_D^{T}(\{c\})\) and \(Dis_D^{U_k}(A)\) are the relative discernible relations of \(c \in A\) and \(A\), and \(red_T\) is a reduct of \((T, A, D)\).

Without the advent of any subset from \(U\), all results of the temporary data set are empty, because \(T\) has an initialization of the empty set. In \((U_k, A, D)\), \(pos_A(U_k)\) is the positive region of \(D\) relative to \(A\), \(Dis_D^{U_k}(\{c\})\) and \(Dis_D^{U_k}(A)\) are the relative discernible relations of \(c \in A\) and \(A\). Upon the arrival of \(U_k\), the temporary data set is updated as \((T \cup U_k, A, D)\).

By the historical results, the incremental manner is used to calculate \(pos_A(D)\), \(Dis_D^{T}(\{c\})\), \(Dis_D^{U_k}(\{c\})\), and \(red_T\) in \((T \cup U_k, A, D)\). A reduct of \((U, A, D)\) is obviously acquired where no further incoming subset is left.

### 4.2 Updating the relative discernible relation

According to Sect. 2, it is necessary to calculate the relative discernible relation of every feature, when selecting a feature subset from a mixed data set. When a sample subset arrives, the incremental mode is thus used to incrementally calculate the relative discernible relation of every feature in this subsection.

As a decision sub-system \((U_k, A, D)\) is put into \((T, A, D)\), some consistent samples in \(T\) and \(U_k\) may be inconsistent in the new data set \((T \cup U_k, A, D)\), which results in the following two facts. One is that some sample pairs of \(Dis_D^{T}(\{c\})\) and \(Dis_D^{U_k}(\{c\})\) may be indiscernible in \((T \cup U_k, A, D)\). The other is sample pairs which are not discerned in \((T, A, D)\) but can be discerned in \((T \cup U_k, A, D)\). By the above two facts, the relative discernible relation of every feature is incrementally calculated.

The following theorem lists sample pairs which cannot be distinguished in \((T \cup U_k, A, D)\).

**Theorem 5** In \((T \cup U_k, A, D)\), the following conclusions hold:

1. for \((x_i, x_j) \in Dis_D^{T}(\{c\})\), if \(x_i \notin pos_A(D)\) and \(x_j \notin pos_A(D)\), then \((x_i, x_j) \notin Dis_D^{T}(\{c\})\);
2. for \((x_i, x_j) \in Dis_D^{U_k}(\{c\})\), if \(x_i, x_j \in pos_A(D)\) and \(D(x_i) = D(x_j)\), then \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);
3. for \((x_i, x_j) \in Dis_D^{U_k}(\{c\})\), if \(x_i \notin pos_A(D)\), then \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);
4. for \((x_i, x_j) \in Dis_D^{U_k}(\{c\})\), if \(x_i, x_j \in pos_A(D)\) and \(D(x_i) = D(x_j)\), then \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);

**Proof**
(1) By the condition (1) of Definition 9, \(x_i \notin pos_A(D)\) implies \((x_i, x_j) \notin Dis_D^{T}(\{c\})\);
(2) By the condition (2) of Definition 9, \(x_i, x_j \in pos_A(D)\) and \(D(x_i) = D(x_j)\) implies \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);
(3) By the condition (1) of Definition 9, \(x_i \notin pos_A(D)\) implies \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);
(4) By the condition (2) of Definition 9, \(x_i, x_j \in pos_A(D)\) and \(D(x_i) = D(x_j)\) implies \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\).

By Theorem 5, we can delete the sample pairs of \(Dis_D^{T}(\{c\})\) and \(Dis_D^{U_k}(\{c\})\) that cannot be discerned in \((T \cup U_k, A, D)\). (\(T - pos_A(D)\)) \(\cap pos_A(D)\) is the set of samples that are inconsistent in \(T\) but consistent in \(T \cup U_k\). \((U_k - pos_A(U_k))\) \(\cap pos_A(D)\) is the set of samples that are inconsistent in \(U_k\) but consistent in \(T \cup U_k\). Hence, we have the following theorem.

**Theorem 6** In \((T \cup U_k, A, D)\), the following conclusions hold:

1. for \(x_i, x_j \in T\), if \(x_i \notin \left( T - pos_A(D) \right) \) \(\cap pos_A(D)\) and \(x_j \notin \left( x_i \right)_c\), then \((x_i, x_j) \notin Dis_D^{T}(\{c\})\);
2. for \(x_i, x_j \in T\), if \(x_i, x_j \in \left( T - pos_A(D) \right) \) \(\cap pos_A(D)\), \(x_j \notin \left( x_i \right)_c\), and \(D(x_i) \neq D(x_j)\), then \((x_i, x_j) \notin Dis_D^{T}(\{c\})\);
3. for \(x_i, x_j \in U_k\), if \(x_i \notin \left( U_k - pos_A(U_k) \right) \) \(\cap pos_A(D)\), \(x_j \notin pos_A(D)\), and \(D(x_i) \neq D(x_j)\), then \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\);
4. for \(x_i, x_j \in U_k\), if \(x_i, x_j \in \left( U_k - pos_A(U_k) \right) \) \(\cap pos_A(D)\), \(x_j \notin \left( x_i \right)_c\), and \(D(x_i) \neq D(x_j)\), then \((x_i, x_j) \notin Dis_D^{U_k}(\{c\})\).

**Proof**
(1) \(x_i \in \left( T - pos_A(D) \right) \) \(\cap pos_A(D)\) implies \(x_i \in pos_A(D)\). By Definition 9, \(x_j \notin \left( x_i \right)_c\) implies \((x_i, x_j) \notin Dis_D^{T}(\{c\})\);
Covering rough set-based incremental feature...

(2) $x_i, x_j \in (T - pos^U_k(D)) \cap pos_A(D)$ implies $x_i, x_j \in pos_A(D)$. By Definition 9, $D(x_i) \neq D(x_j)$ implies $(x_i, x_j) \in Dis_D([c])$;

(3) $x_i \in (U_k - pos^U_k(D)) \cap pos_A(D)$ implies $x_i \in pos_A(D)$. By Definition 9, $x_j \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$;

(4) $x_i, x_j \in (U_k - pos^U_k(D)) \cap pos_A(D)$ implies $x_i, x_j \in pos_A(D)$. By Definition 9, we have $(x_i, x_j) \in Dis_D([c])$.

Furthermore, $pos^U_k(D) \cap (T - pos_A(D))$ is the sample set, in which every sample is inconsistent in $T$ but consistent in $T \cup U_k$. $pos^U_k(D) \cap (U_k - pos_A(D))$ is the set of samples that are inconsistent in $U_k$ but consistent in $T \cup U_k$. Based on the discussions, the next theorem is given.

**Theorem 7** In $(T \cup U_k, A, D)$, the following conclusions hold:

(1) for $x_i, x_j \in T$, if $x_i \notin pos_A(D)$, $x_j \in pos^U_k(D) \cap (T - pos_A(D))$ and $x_j \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$;

(2) for $x_i, x_j \in U$, if $x_i \notin pos_A(D)$, $x_j \in pos^U_k(D) \cap (U_k - pos_A(D))$, and $x_j \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$.

**Proof** (1) $x_i, x_j$ satisfy the condition (2) of Definition 9. By Definition 9, $x_j \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$;

(2) $x_i, x_j$ satisfy the condition (2) of Definition 9. By Definition 9, $x_j \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$.

By Theorems 6 and 7, we can incrementally find the sample pairs that cannot be discerned in $(T, A, D)$ and $(U_k, A, D)$ but can be discerned in $(T \cup U_k, A, D)$. Furthermore, it is necessary to judge whether each sample in $T$ and each sample in $U_k$ are discernible in $(T \cup U_k, A, D)$, which is a key step of incrementally calculating the discernible relations of $V_c \in A$ and $A$.

**Theorem 8** In $(T \cup U_k, A, D)$, the next conclusions hold for $x_i \in T$ and $x_j \in U_k$.

(1) if $x_i \in pos_A(D)$, $x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$;

(2) if $x_i \notin pos_A(D)$, $x_j \in pos_A(D)$ and $x_i \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$;

(3) if $x_i, x_j \in pos_A(D)$, $D(x_i) \neq D(x_j)$ and $x_j \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$;

(4) if $x_i, x_j \in pos_A(D)$, $D(x_i) \neq D(x_j)$ and $x_i \notin (x_i)_c$, then $(x_i, x_j) \in Dis_D([c])$.

**Proof** (1) $x_i, x_j$ satisfy the condition (1) of Definition 9, $x_j \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$;

(2) $x_i, x_j$ satisfy the condition (1) of Definition 9, $x_i \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$;

(3) $x_i, x_j$ satisfy the condition (2) of Definition 9, $x_j \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$.

(4) $x_i, x_j$ satisfy the condition (2) of Definition 9, $x_i \notin (x_i)_c$ implies $(x_i, x_j) \in Dis_D([c])$.

According to Theorem 8, we can obtain sample pairs between $T$ and $U_k$ that can be discerned in $(T \cup U_k, A, D)$. By Theorems 5–8, the incremental algorithm is designed to update the relative discernible relation of every feature.

**Algorithm 2** Updating the relative discernible relations

**Input:** 1) the temporary set $T$; 2) $pos^U_k(D)$; 3) $Dis^T_D([c])$, $Dis^T_D(A)$; 4) $U_k$.

**Output:** $Dis_D([c])$ for each $c \in A$.

1: In $(U_k, A, D)$, find $pos^U_k(D)$, $Dis^U_k([c])$, $Dis^U_k(A)$;
2: In $(T \cup U_k, A, D)$, compute the positive region $pos_A(D)$;
3: for each $c \in A$ do
4: Let $Dis_D([c]) = Dis^T_D([c]) \cup Dis^U_k([c])$;
5: for $\forall(x_i, x_j) \in Dis_D([c])$ do
6: if $x_j \notin pos_A(D)$ then
7: Let $Dis_D([c]) = Dis_D([c]) - (\{x_i, x_j\})$
8: else if $x_i, x_j \notin pos_A(D)$ and $D(x_i) = D(x_j)$ then
9: Let $Dis_D([c]) = Dis_D([c]) - (\{x_i, x_j\})$
10: end if
11: end for
12: for $\forall x_i, x_j \in T$ do
13: if $x_i \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
14: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
15: else if $x_i, x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
16: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
17: else if $x_i \in pos_A(D)$, $x_j \notin pos^U_k(D)$, $x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
18: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
19: end if
20: end for
21: for $\forall x_i, x_j \in U_k$ do
22: if $x_j \notin pos^U_k(D)$, $x_i \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
23: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
24: else if $x_i, x_j \notin pos^U_k(D)$ and $x_i \notin (x_i)_c$ then
25: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
26: else if $x_i \in pos_A(D)$, $x_j \notin pos^U_k(D)$, $x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
27: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
28: end if
29: end for
30: for $\forall x_i \in T$, $x_j \in U_k$ do
31: if $x_j \in pos_A(D)$, $x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
32: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
33: else if $x_i \in pos_A(D)$, $x_j \notin pos_A(D)$ and $x_j \notin (x_i)_c$ then
34: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
35: else if $x_i, x_j \in pos_A(D)$, $D(x_i) \neq D(x_j)$ and $x_j \notin (x_i)_c$ then
36: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
37: else if $x_i, x_j \in pos_A(D)$, $D(x_i) \neq D(x_j)$ and $x_i \notin (x_i)_c$ then
38: Let $Dis_D([c]) = Dis_D([c]) \cup (\{x_i, x_j\})$
39: end if
40: end for
41: end for
42: return $Dis_D([c])$.

In Algorithm 2, Step 1 computes the positive region of $U_k$ and the discernible relation of each feature and $A$ in $(U_k, A \cup D)$, and the time complexity is $O(|U_k|^2 |A|)$. Step 2 calculates the positive region of $U_k \cup T$ in $(T \cup U_k, A \cup D)$, and the time complexity is $O(|T| |U_k||A|)$. Steps 3–41 incrementally calculate the relative discernible relation of every feature, and the time complexity is $O(|T|^2 - |Dis_D([c])||A|)$. In a word, the whole time complexity of Algorithm 2 is $max(O(|T||U_k||A|), O(|T|^2 - |Dis_D([c])||A|))$. Springer
4.3 Incremental feature selection scheme in covering rough sets

With an incoming sample subset arriving, the following two cases will occur: (1) $\text{Dis}_D(A) = \text{Dis}_D(\text{red}_T)$ and (2) $\text{Dis}_D(A) \neq \text{Dis}_D(\text{red}_T)$. In terms of the two cases, incremental feature selection scheme is investigated to disclose the strategies of increasing informative features and removing redundant features.

**Case 1:** $\text{Dis}_D(A) = \text{Dis}_D(\text{red}_T)$.

The following theorem holds for Case 1 by Theorem 4.

**Theorem 9** $\text{Dis}_D(A) = \text{Dis}_D(\text{red}_T)$ indicates that a reduct of $(T \cup U_k, A, D)$ is contained in $\text{red}_T$.

**Proof** $\text{Dis}_D(A) = \text{Dis}_D(\text{red}_T)$ implies sample pairs which are discerned by $A$, are also distinguished by $\text{red}_T$. A reduct of $(T \cup U_k, A, D)$ is contained in $\text{red}_T$ by Theorem 4. $\square$

Theorem 9 indicates that $\text{red}_T$ properly includes an optimal feature subset of $(T \cup U_k, A, D)$, or is just an optimal feature subset of $(T \cup U_k, A, D)$. The strategy of removing redundant features is presented to incrementally calculate a reduct of $(T \cup U_k, A, D)$.

**First feature removal strategy** The feature $c$ can be removed from $\text{red}_T$, if $\text{Dis}_D(\text{red}_T - \{c\}) = \text{Dis}_D(A)$ holds for $c \in \text{red}_T$.

The strategy shows that in the new mixed data, $\text{red}_T - \{c\}$ can distinguish sample pairs which are discerned by $A$. By Theorem 4, the feature $c$ can be removed from $\text{red}_T$. If the strategy is unavailable for $\forall c \in \text{red}_T$, $\text{red}_T$ is just a reduct of $(T \cup U_k, A, D)$ by Theorem 4. In detail, we employ the forward greedy search to delete redundant features from $\text{red}_T$. At each loop, we delete the feature $c$ with $\text{Dis}_D(\text{red}_T - \{c\}) = \text{Dis}_D(A)$ from $\text{red}_T$, and repeat this process until the removal of any remaining feature does not satisfy the First feature removal strategy. Accordingly, an optimal feature subset of the updated temporary data set is calculated incrementally.

**Case 2:** $\text{Dis}_D(A) \neq \text{Dis}_D(\text{red}_T)$.

Since Case 2 does not meet the condition (2) of Theorem 4, $\text{red}_T$ is not an optimal feature subset of the updated mixed data set. So, some informative features $B \subseteq A - \text{red}_T$ should be added into $\text{red}_T$ until $\text{Dis}_D(\text{red}_T \cup B) = \text{Dis}_D(A)$ holds. The following strategy of adding features is proposed by Theorem 4.

**Feature addition strategy** If $B \subseteq A - \text{red}_T$ is a minimum addition subset that satisfies $\text{Dis}_D(\text{red}_T \cup B) = \text{Dis}_D(A)$, $B$ can be put into $\text{red}_T$.

By the feature addition strategy, $\text{red}_T \cup B$ can distinguish sample pairs which are distinguished by $A$. The fact indicates that a reduct of $(T \cup U_k, A, D)$ is contained in $\text{red}_T \cup B$. Some features in $\text{red}_T \cup B$, however, are redundant. Moreover, since $B$ is a minimum addition subset that satisfies $\text{Dis}_D(\text{red}_T \cup B) = \text{Dis}_D(A), \text{Dis}_D(\text{red}_T \cup (B - \{a\})) \neq \text{Dis}_D(A)$ holds for $\forall a \in B$. Therefore, we only need to delete redundant features from $\text{red}_T$. Thus, the following strategy of removing redundant features is presented to calculate an optimal feature subset of $(T \cup U_k, A, D)$.

**Second feature removal strategy** The feature $c$ can be removed from $\text{red}_T$, if $\text{Dis}_D((\text{red}_T - \{c\}) \cup B) = \text{Dis}_D(A)$.

By Theorem 4, when the removal strategy is not satisfied for $\forall c \in \text{red}_T$, $\text{red}_T \cup B$ is just a reduct of $(T \cup U_k, A, D)$. When it is satisfied for $\exists c \in \text{red}_T$, $c$ can be removed from $\text{red}_T$. Similar to the First feature removal strategy, we use the forward greedy search to delete redundant features from $\text{red}_T$. At each loop, a feature $c$ with $\text{Dis}_D((\text{red}_T - \{c\}) \cup B) = \text{Dis}_D(A)$ can be deleted from $\text{red}_T$, and this process is repeated until the Second feature removal strategy is not satisfied. Thus, a reduct of $(T \cup U_k, A, D)$ is obtained by continuously using the second feature removal strategy.

To sum up, the technique details are given below. If a newly acquired sample subset satisfies Case 1, we employ the first feature removal strategy to renew an optimal feature subset of the updated mixed data set. If it satisfies Case 2, we utilize the feature addition strategy to put a feature subset $B \subseteq A - \text{red}_T$ into the current reduct $\text{red}_T$, and apply the second feature removal strategy to remove redundant features from $\text{red}_T$.

4.4 Incremental algorithms for feature selection from mixed data

Two incremental feature selection algorithms are presented to incrementally calculate an optimal feature subset from a mixed data set based on the above strategies of increasing and removing features.

**Incremental-1** Update an optimal feature subset as sample subsets arrive successively.

With the sequential arrival of sample subsets, the incremental process updates a reduct by the updated relative discernible relation of every feature. More specially, if a newly acquired sample subset is in Case 1, the First feature removal strategy is performed; if it is in Case 2, the feature addition strategy is first performed and the second feature removal strategy is then used. A reduct of the whole mixed data set is acquired by repeatedly using the incremental process until no further sample subset is left. By the incremental process, the following incremental algorithm is presented to incrementally obtain an optimal feature subset in covering rough sets. Figure 1 clearly depicts the working principle of Incremental-1.

Algorithm 3 incrementally calculates an optimal feature subset beginning by an empty set. In Step 1, $T$ and the results of $(T, A, D)$ are initialized with empty sets. A new sample subset is obtained in Step 3. Step 4 which has the computational complexity of $\max(O(||T||U_k||A||), O(||T||^2 -$
Algorithm 3 First incremental feature selection algorithm in covering rough sets (denoted by Incremental-1)

**Input:** 1) The sequence of sample subsets $U = \cup_{k=1}^{n} U_k$; 2) the set of features $A$; 3) the decision feature $D$.

1. Initialize: 1) $red = \emptyset$; 2) The set storing all historical samples $T = \emptyset$; 3) $\text{Disc}_T^D(c) = \emptyset$ for every $c \in A$, $\text{Disc}_T^D(A) = \emptyset$; 4) The number of iterations $k = 1$; 5) The addition feature subset $B = \emptyset$.

2. While $U$ is a non-empty set do

3. Put the current obtained subset $U_k$ into $T$.

4. For every $c \in A, D$, calculate $\text{Disc}_D(c)$ for every $c \in A$ and $\text{Disc}_D(A)$ by Algorithm 2.

5. For every $c \in A$, let $\text{Disc}_T^D(c) = \text{Disc}_D(c)$, and $\text{Disc}_T^D(A) = \text{Disc}_D(A)$.

6. While $\text{Disc}_T^D(A) \neq \text{Disc}_T^D(red \cup B)$ do

7. Find $\text{Disc}_T^D(red \cup B \cup \{c\})$ for every $c \in A - red - B$.

8. Select $c_0 \in A - red - B$ s.t. $|\text{Disc}_T^D(red \cup B \cup \{c_0\})| = \max_{c \in A - red - B} |\text{Disc}_T^D(red \cup B \cup \{c\})|$.

9. Let $\text{Disc}_T^D(red \cup B) = \text{Disc}_T^D(red \cup B \cup \{c_0\})$ and $B = B \cup \{c_0\}$.

10. While $\text{Disc}_T^D(A) = \text{Disc}_T^D(red \cup B)$ do

11. Choose $c_0$ s.t. $\text{Disc}_T^D(A) = \text{Disc}_T^D(red - \{c_0\}) \cup B$, and let $red = red - \{c_0\}$.

12. Update the relative discernible relations.

13. Case 1:

14. First feature removal strategy.

15. Case 2:

16. Second feature removal strategy.

17. Update selected feature subset.

18. Return $red$.

Algorithm 4 Second incremental feature selection algorithm in covering rough sets (denoted by Incremental-2)

**Input:** 1) The sequence of sample subsets $U = \cup_{k=1}^{n} U_k$; 2) The set of features $A$; 3) the decision feature $D$.

1. Initialize: 1) $red = \emptyset$; 2) The set storing all historical samples $T = \emptyset$; 3) $\text{Disc}_T^D(c) = \emptyset$ for every $c \in A$, $\text{Disc}_T^D(A) = \emptyset$; 4) The number of iterations $k = 1$; 5) The addition feature subset $B = \emptyset$.

2. While $U$ is a non-empty set do

3. Put the current obtained subset $U_k$ into $T$.

4. For every $c \in A, D$, calculate $\text{Disc}_D(c)$ for every $c \in A$ and $\text{Disc}_D(A)$ by Algorithm 2.

5. For every $c \in A$, let $\text{Disc}_T^D(c) = \text{Disc}_D(c)$, and $\text{Disc}_T^D(A) = \text{Disc}_D(A)$.

6. While $\text{Disc}_T^D(A) \neq \text{Disc}_T^D(red \cup B)$ do

7. Find $\text{Disc}_T^D(red \cup B \cup \{c\})$ for every $c \in A - red - B$.

8. Select $c_0 \in A - red - B$ s.t. $|\text{Disc}_T^D(red \cup B \cup \{c_0\})| = \max_{c \in A - red - B} |\text{Disc}_T^D(red \cup B \cup \{c\})|$.

9. Let $\text{Disc}_T^D(red \cup B) = \text{Disc}_T^D(red \cup B \cup \{c_0\})$ and $B = B \cup \{c_0\}$.

10. While $\text{Disc}_T^D(A) = \text{Disc}_T^D(red \cup B)$ do

11. Choose $c_0$ s.t. $\text{Disc}_T^D(A) = \text{Disc}_T^D(red - \{c_0\}) \cup B$, and let $red = red - \{c_0\}$.

12. Update the relative discernible relations.

13. Case 1:

14. First feature removal strategy.

15. Case 2:

16. Second feature removal strategy.

17. Update selected feature subset.

18. Return $red$.

However, Incremental-1 needs to update the optimal feature subset at each iteration, which is uneconomical for large mixed data sets. In order to expedite Incremental-1, we do not incrementally calculate the selected feature subset with sample subsets arriving successively, which results in the next incremental process.

**Incremental-2** Compute an optimal feature subset when there is no sample subset left.
The second incremental process starts from an empty set to calculate an optimal feature subset from the whole mixed data set where no sample subsets are obtained. With sample subsets arriving continuously, the relative discernible relations of every feature and the feature set are merely updated without incrementally finding the optimal feature subset. As no further sample subset is added, the final discernible relations of every feature and the feature set are obtained on the whole mixed data set. With the obtained relations, the feature addition strategy and the second feature removal strategy are utilized to calculate an optimal feature subset from the whole mixed data set. To clearly depict the working principle of Incremental-2, we show Fig. 2.

By the second incremental process, we present Algorithm 4 to calculate a reduct in covering rough sets. In Step 1 of Algorithm 4, $T$ and the results of $(T, A, D)$ are initialized as empty sets. The relative discernible relations of $\forall c \in A$ and $A$ are updated in Steps 2~4 which has the computational complexity of $\max(\mathcal{O}(|T||U_k||A|), \mathcal{O}(|T|^2 - |\text{Dis}_T^C(c)||A|))$. The feature addition strategy is performed in Step 7~11, which has the computational complexity of $\mathcal{O}(|A|)$. The Second feature removal strategy is performed in Steps 12~16, which has the computational complexity of $\mathcal{O}(|A|)$. Therefore, the whole computational complexity of Algorithm 4 is $\max(\mathcal{O}(|T||U_k||A|), \mathcal{O}(|T|^2 - |\text{Dis}_T^C(c)||A|))$.

**Remark** If new sample subsets with various feature types arrive in a random sequence, these sample subsets will be put into our proposed incremental models one at a time. Specifically, as a new sample subset is randomly added into the incremental model Incremental-1, the optimal feature subset is incrementally updated based on the updated relative discernible relations of each feature and the whole feature set. When the next sample subset arrive randomly, the same way as above is performed using Incremental-1. Furthermore, when a new sample subset is randomly put into Incremental-2, the relative discernible relations of each feature and the whole feature set are updated. After these new sample subsets have been put into Incremental-2, the optimal feature subset is calculated based on the updated relative discernible relations. Thus, the optimal feature subsets obtained by Incremental-1 and Incremental-2 can be used in the classification task.

**5 Experimental results**

This section compares Incremental-2 with Incremental-1, Non-Incremental (Algorithm 1) and consistency-based feature selection method (Dash and Liu 2003) (denoted by Consistency). The main concern is about the comparison of the time efficiency, i.e., the runtime of obtaining an optimal feature subset. The other comparison is the feasibility, i.e., the size and the classification performance of selected features.

**5.1 Experimental setup**

*The hardware environment:* The experiments of Sects. 5.2~5.3 are performed with Windows 7PC and Intel (R) Xeon (R) CPU E5-2620 0 @ 2.00 GHz 2.00 GHz and 80 GB memory. The experiments of Sect. 5.4 is performed with Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz 3.60 GHz and 32.0 GB memory in Windows 10 system.

*The software environment:* MATLAB 2013b.

*Data set:* All fourteen data sets used in our experiments are from UCI Machine Learning Repository, where 'Annealing' in Table 1 includes 798 training samples and 100 test samples. It is suggested that 'Australian,' 'Japan,' 'Mammographic mass' and 'Nursery' are used to further confirm the time efficiency of our proposed incremental algorithms. The detailed information is summarized in Table 1 where 'Missing features' reports the size of missing-valued features. These data sets contain various even-distributed numbers of samples, features and classes, and include symbolic, real-valued and missing-valued features which are denoted by 's,' 'r' and 'm,' respectively.

*Data partitioning:* Before performing feature selection, every data in Table 1 is at random split into 10 equal-sized sample subsets. The 10 sample subsets constitute a sample subset sequence.

*Covering induced by a feature:* The equivalence relation is used to generate the covering of a symbolic feature; the neighborhood relation is used to generate the covering of a
real-valued feature, where the neighborhood radius is taken as 0.15; the tolerance relation is used to generate the covering of a missing-valued feature. By the definition of the induced covering, we can generate the covering of a mixed feature set.

**Classification accuracy:** The classifiers used in the experiments are the k-nearest neighbor function ‘fitcknn’ and the random forest function ‘TreeBagger’ in MATLAB Toolbox, where the number of nearest neighbors is set as 3 and the number of trees is chosen as 500. The tenfold cross-validation is used to compute the classification accuracy. Specifically, the data set before or after feature selection is at random split into 10 parts of equal size, where nine of them is used as the training set and the rest one is the test set. At each round, we use ‘fitcknn’ or ‘TreeBagger’ to train a model on the training set, and compute the classification accuracy of the test set that is the ratio of the test samples classified correctly and all samples in the test set. The average value of classification accuracies in 10 rounds is just reported as the final classification accuracy.

### 5.2 Comparison and analysis of Incremental-2 and covering rough set-based methods

This section compares Incremental-2 with Non-Incremental and Incremental-1 on the fourteen selected data sets. The experimental results are shown in Fig. 3, Tables 2 and 3.

Figure 3 displays that the runtime of Non-Incremental, Incremental-1 and Incremental-2 varies with the successive arrival of sample subsets. In each sub-figure of Fig. 3, the horizontal axis is the index of every incoming sample subset, whereas the vertical axis is the running time of every comparison method. From Fig. 3, we can observe that both Incremental-2 and Incremental-1 are consistently faster than Non-Incremental as each sample subset continuously arrives. The main reason is that Non-Incremental lacks the effective mechanism of dealing with dynamic data sets. With the arrival of a new sample subset, Non-Incremental has to be rerun to calculate an optimal feature subset from the currently arrived sample subsets including the historical samples and the new added-in samples. In result, Non-Incremental costs a large amount of the computation time. Furthermore, we can see that Incremental-2 is more efficient than Incremental-1. This is because with the first nine subsets arriving, Incremental-1 needs to update the optimal feature subset by incrementally updating the discernible relation of each feature, while Incremental-2 merely incrementally calculates the relative discernible relations with no need for updating the optimal feature subset. Therefore, Incremental-2 can save the running time as sample subsets continuously arrive.

Table 2 shows the runtime of the three comparison algorithms, where ‘Total Time’ represents the sum of the runtime of a method on the fourteen data sets. We can observe from Table 2 that Total Time of Incremental-2 (299070.42 seconds) is much less than that of Non-Incremental (427047.18 seconds) and Incremental-1 (337612.82 seconds). Moreover, Incremental-2 is more efficient than Non-Incremental and Incremental-1 on each selected data set including the four suggested data sets. For instance, on ‘Annealing,’ the runtime of Incremental-2 is 115.26 seconds, which is much less than that of Non-Incremental (620.34 seconds) and Incremental-1 (135.19 seconds); on ‘Crowdsourced,’ the runtime of Incremental-2 is 125051.27 seconds, which is much lower than that of Non-Incremental (194859.87 seconds) and Incremental-1 (144789.44 seconds). On the suggested data set ‘Australian,’ the runtime of Incremental-2 is 15.24 seconds, which is less than Non-Incremental (121.91 sec-

### Table 1 Data description

| Data          | Samples | Features | Type | Missing features | Class |
|---------------|---------|----------|------|------------------|-------|
| Dermatology   | 366     | 34       | s,m  | 1                | 6     |
| Credit        | 690     | 15       | s,r,m| 7                | 2     |
| Australian    | 690     | 14       | s,r,m| 7                | 2     |
| Japan         | 690     | 15       | s,r,m| 5                | 2     |
| Annealing     | 898     | 38       | s,r,m| 27               | 6     |
| Mammographic mass | 961     | 5        | s    | 4                | 2     |
| German        | 1000    | 20       | s,r  | 0                | 2     |
| Steel         | 1941    | 27       | r    | 0                | 7     |
| Sick          | 2800    | 28       | s,r,m| 6                | 2     |
| Statlog       | 6435    | 36       | r    | 0                | 7     |
| Anuran calls  | 7195    | 21       | r    | 0                | 4     |
| Mushroom      | 8124    | 22       | s,m  | 1                | 2     |
| Crowdsourced  | 10546   | 29       | r,m  | 1                | 6     |
| Nursery       | 12960   | 8        | s    | 0                | 5     |
with the successive arrival of sample subsets.

In this subsection, we compare Incremental-2 with Consistency (Dash and Liu 2003) on the fourteen selected data sets. In Table 3, 3-NN and RF represent the classification accuracies of selected features with the k-nearest neighbor classifier and the random forest classifier, respectively. From Table 3, we can see that all three comparison algorithms can select an optimal feature subset and can obtain a good generalization. We can see that the 3-NN classification accuracy (0.79) of Incremental-2 is lower than that of 'Raw' (0.81) and the RF classification accuracy (0.89) of Incremental-2 is lower than that of 'Raw' (0.90). This may be because some features removed by Incremental-2 are informative to the 3-NN classifier and the random forest classifier. Furthermore, we find that the difference between 3-NN accuracies of Incremental-2 and 'Raw' is just 0.02 and the difference between RF accuracies of Incremental-2 and 'Raw' is only 0.01. This fact implies that Incremental-2 is feasible with ignoring the small difference of the classification accuracy.

To sum up, in contrast with Incremental-1 and Non-Incremental, our proposed Incremental-2 shows the time efficiency of selecting an optimal feature subset, while sharing the feasibility of obtaining an optimal feature subset with neglecting a small difference of the classification accuracy.

### 5.3 Comparison of incremental-2 and consistency

In this subsection, we compare Incremental-2 with Consistency (Dash and Liu 2003) on the fourteen selected data sets.
Table 2. Runtime of non-incremental, incremental-1 and incremental-2

| Data set            | Non-Incremental/s | Incremental-1/s | Incremental-2/s |
|---------------------|-------------------|-----------------|-----------------|
| Dermatology         | 16.04             | 8.18            | 7.45            |
| Credit              | 26.99             | 18.05           | 17.57           |
| Australian          | 121.91            | 15.61           | 15.24           |
| Japan               | 130.66            | 27.93           | 26.75           |
| Annealing           | 620.34            | 135.19          | 115.26          |
| Mammographic mass   | 71.26             | 13.57           | 13.12           |
| German              | 126.69            | 91.71           | 82.49           |
| Steel               | 2459.79           | 1201.77         | 1075.94         |
| Sick                | 5596.12           | 4766.29         | 4749.38         |
| Statlog             | 70994.85          | 53686.56        | 47630.55        |
| Anuran Calls        | 57923.33          | 52002.33        | 46343.08        |
| Mushroom            | 54146.59          | 46094.01        | 41678.4         |
| Crowdsourced        | 194859.87         | 144789.44       | 125051.27       |
| Nursery             | 39952.74          | 34762.18        | 32263.92        |
| Total Time          | 427047.18         | 337612.82       | 299070.42       |

Table 3. Size and accuracy of features selected by non-incremental, incremental-1 and incremental-2

| Data set            | Raw | Non-Incremental | Incremental-1 | Incremental-2 |
|---------------------|-----|-----------------|---------------|---------------|
|                     | No. | 3-NN RF         | No. | 3-NN RF | No. | 3-NN RF | No. | 3-NN RF |
| Dermatology         | 34  | 0.89 0.97       | 6   | 0.59 0.66 | 4   | 0.33 0.39 | 6   | 0.61 0.69 |
| Credit              | 15  | 0.65 0.86       | 11  | 0.64 0.86 | 5   | 0.66 0.71 | 12  | 0.61 0.88 |
| Australian          | 14  | 0.64 0.86       | 6   | 0.61 0.86 | 3   | 0.59 0.62 | 6   | 0.62 0.86 |
| Japan               | 15  | 0.64 0.85       | 12  | 0.59 0.87 | 6   | 0.64 0.77 | 12  | 0.60 0.87 |
| Annealing           | 38  | 0.99 0.99       | 3   | 0.99 1   | 3   | 0.99 1   | 3   | 1 1 |
| Mammographic mass   | 5   | 0.74 0.80       | 5   | 0.75 0.80 | 2   | 0.72 0.77 | 5   | 0.75 0.81 |
| German              | 20  | 0.69 0.73       | 10  | 0.67 0.72 | 5   | 0.64 0.64 | 11  | 0.68 0.74 |
| Steel               | 27  | 0.41 0.75       | 21  | 0.41 0.78 | 6   | 0.51 0.66 | 7   | 0.59 0.75 |
| Sick                | 29  | 0.96 0.98       | 17  | 0.96 0.98 | 7   | 0.92 0.94 | 20  | 0.96 0.98 |
| Statlog             | 36  | 0.91 0.91       | 36  | 0.91 0.92 | 4   | 0.83 0.84 | 11  | 0.89 0.92 |
| Anuran Calls        | 21  | 0.99 0.98       | 12  | 0.98 0.98 | 3   | 0.93 0.93 | 8   | 0.98 0.97 |
| Mushroom            | 22  | 1 1             | 4   | 1 0.99   | 4   | 1 1      | 4   | 1 1 |
| Crowdsourced        | 28  | 0.97 0.94       | 19  | 0.96 0.94 | 4   | 0.81 0.84 | 13  | 0.94 0.93 |
| Nursery             | 8   | 0.84 0.98       | 8   | 0.83 0.99 | 8   | 0.83 0.99 | 8   | 0.84 0.99 |
| Average             | 22.29 | 0.81 0.90       | 12.14 | 0.78 0.88 | 4.57 | 0.74 0.79 | 9   | 0.79 0.89 |

Before performing this experiment, we first clarify the executing process of Consistency, by which we can obtain the results of consistency-based feature selection algorithm. As well known, a feature selection method consists of the subset generation, subset evaluation and stopping criterion. In our experiment, we start from the empty set of features and put one feature that makes the increment of consistency maximal into the selected feature subset at every loop. This is the step of subset generation. The subset evaluation is embedded into the step by maximizing the increment of consistency. The consistency-based feature selection algorithm does not stop until the consistency of the original feature set is kept. Consistency is a consistency-based feature selection approach of seeking for the minimum subset which classifies samples as consistently as the full set under the best first search strategy. It is parallel to rough set-based feature selection methods. So, it is indispensable to compare it with Incremental-2. However, Consistency can only handle data sets with symbolic features. When facing with real-valued data sets, it requires to discretize real-valued features; when dealing with missing-valued data sets, it has to fill up these missing feature values. Such preprocessing methods can easily lead to the information loss, which may worsen the classification accuracy of the learning algorithms. Fur-
Table 4 Comparison results of incremental-2 and consistency

| Data set            | Incremental-2 | Consistency |
|---------------------|---------------|-------------|
|                     | No. | 3-NN | RF | No. | 3-NN | RF |
| Dermatology         | 6   | 0.61 | 0.69 | 6   | 0.48 | 0.65 |
| Credit              | 12  | 0.61 | 0.88 | 11  | 0.63 | 0.86 |
| Australian          | 6   | 0.62 | 0.86 | 6   | 0.62 | 0.86 |
| Japan               | 12  | 0.60 | 0.87 | 11  | 0.63 | 0.86 |
| Annealing           | 3   | 1    | 1   | 3   | 1    | 1   |
| Mammographic mass   | 5   | 0.75 | 0.81 | 5   | 0.74 | 0.81 |
| German              | 11  | 0.68 | 0.74 | 10  | 0.69 | 0.73 |
| Steel               | 7   | 0.59 | 0.75 | 20  | 0.49 | 0.72 |
| Sick                | 20  | 0.96 | 0.98 | 16  | 0.96 | 0.98 |
| Statlog             | 11  | 0.89 | 0.92 | 31  | 0.91 | 0.90 |
| Anuran Calls        | 8   | 0.98 | 0.97 | 11  | 0.98 | 0.96 |
| Mushroom            | 4   | 1    | 1   | 4   | 0.99 | 1   |
| Crowdsouced         | 13  | 0.94 | 0.93 | 16  | 0.95 | 0.92 |
| Nursery             | 8   | 0.84 | 0.99 | 8   | 0.83 | 0.98 |
| Average             | 9   | 0.79 | 0.89 | 11.29 | 0.78 | 0.87 |

Moreover, due to different mechanisms, we are only devoted to the comparisons of the size and the classification accuracy of selected features rather than the comparison of the runtime. The experimental results are summarized in Table 4.

Table 4 reports the number of features chosen by Incremental-2 and Consistency. From Table 4, we can easily see that the average size of features chosen by Incremental-2 (9) is smaller than that of Consistency (11.29). In detail, on 'Steel,' the size of features selected by Incremental-2 is only 7, while the size of features selected by Consistency is 20. On 'Statlog,' the number of features chosen by Incremental-2 is just 11, whereas the size of features selected by Consistency is 31. These facts imply that our proposed method Incremental-2 can delete more redundant features than Consistency on most of selected data sets.

Moreover, we can see from Table 4 that 3-NN accuracy of Incremental-2 (0.79) is higher than that of Consistency (0.78), and the RF accuracy of Incremental-2 (0.89) is better than that of Consistency (0.87). For example, under the condition of 6 selected features on 'Dermatology,' the 3-NN accuracy of Incremental-2 is 0.61, while the 3-NN accuracy of Consistency is just 0.48; the RF accuracy of Incremental-2 is 0.69, whereas the RF accuracy of Consistency is 0.65. Furthermore, despite of less features selected by Incremental-2 on some selected data sets, the accuracy of Incremental-2 is better than that of Consistency. For instance, on 'Steel,' the 3-NN and RF accuracies of Incremental-2 are 0.59 and 0.75, while the 3-NN and RF accuracies of Consistency are 0.49 and 0.72. These above facts indicate that Incremental-2 has a higher classification accuracy in comparison with Consistency.

5.4 The effect of the neighborhood radius on Incremental-1 and Incremental-2

In this section, we show the effect of changing the neighborhood radius on Incremental-1 and Incremental-2 for the first nine data sets of Table 1. Here, the neighborhood radius is adjusted to vary from zero to one with a step of 0.05. The experimental results are shown in Figs. 4, 5 and 6.

Figure 4 depicts the time changes with varying the neighborhood radius. From Fig. 4, we can see that the proposed Incremental-2 is faster than Incremental-1 nearly at each neighborhood radius. This demonstrates the time efficiency of Incremental-2 that has been concluded in Sect. 5.2. The main reason is that at the arrival of each sample subset, Incremental-2 updates the relative discernible relations without updating the optimal feature subset. Thus, Incremental-2 can save the runtime of searching for an optimal feature subset each time a sample subset arrives. On the data sets 'Dermatology' and 'Mammographic mass,' since they have no real-valued features, the neighborhood relations are not generated. So, the runtime of Incremental-1 and Incremental-2 fluctuates a little with changing the neighborhood radius. Furthermore, on the rest seven data sets, the runtime of both Incremental-1 and Incremental-2 changes with varying the neighborhood radius. Specifically, on 'Credit,' 'Australian,' 'Japan,' 'German' and 'Steel,' the overall trend is that the runtime of Incremental-1 and Incremental-2 gets slow as the neighborhood radius increases. On 'Annealing' and 'Sick,' the overall trend is that the runtime of Incremental-1 and Incremental-2 gets fast with the increasing of the neighborhood radius. Therefore, for data sets with real-valued features, the changing of the neighborhood radius has an
influence on the runtime of Incremental-1 and Incremental-2. Meanwhile, Incremental-2 is faster than Incremental-1 nearly at each neighborhood radius.

Figure 5 show the changing of the number of features selected by Incremental-1 and Incremental-2 when varying the neighborhood radius. We can see from Fig. 5 that the size of features selected by Incremental-2 is more than that of features selected by Incremental-1 at each neighborhood radius. This is because in contrast with Incremental-2, Incremental-1 can remove more redundant features. Moreover, on two symbolic data sets 'Dermatology' and 'Mammographic mass,' the size of features selected by Incremental-1 and Incremental-2 basically keeps invariant at each neighborhood radius. The reason is that the two symbolic data sets do not require to generate the neighborhood relations of real-valued features.

Figure 6 shows the variations of the 3-NN classification accuracies with changing the neighborhood radius.
Fig. 5  Size of features selected by Incremental-1 and Incremental-2 with changing neighborhood radius

Fig. 6, we can observe that the 3-NN classification accuracy of Incremental-2 is higher than that of Incremental-1 nearly at each neighborhood radius. This may be because Incremental-1 may remove some redundant features that are informative to the 3-NN classifier. Furthermore, on the two symbolic data sets 'Dermatology' and 'Mammographic mass,' the 3-NN classification accuracies of Incremental-1 and Incremental-2 fluctuates a little. This is because the two symbolic data sets do not need to generate the neighborhood relations of real-valued features.

In a word, the neighborhood radius has an effect on the performance of Incremental-1 and Incremental-2 for data sets with real-valued features. In contrast with Incremental-1, Incremental-2 can obtain a satisfactory feature subset in a shorter time nearly at each neighborhood radius. It is difficult to determine a proper neighborhood radius.
pared with Incremental-1, Non-Incremental and Consistency, Incremental-2 can efficiently obtain an optimal feature subset from mixed data sets without sacrificing too much classification accuracy.

6 Conclusion
In covering rough set model, incremental feature selection scheme is investigated to accelerate the computation of a reduct from large mixed data with the segmentation of a sample subset sequence. The incremental scheme is determined

Fig. 6 3-NN accuracy of features selected by Incremental-1 and Incremental-2 with changing the neighborhood radius
to reveal the strategies of increasing informative features and removing redundant features by incrementally updating the relative discernible relation of every feature. Two incremental feature selection approaches are presented by the strategies of increasing and removing features. The first one incrementally calculates the relative discernible relation and the optimal feature subset upon sample subset arriving continuously and returns the final reduct when all subsets have been added. The second one merely renews the relative discernible relations upon sample subsets arriving successively, and then calculates the reduct when all subsets have arrived. Our results show the two facts: (1) the presented incremental methods can accelerate the acquirement of the optimal feature subset in covering rough sets without sacrificing too much classification accuracy; (2) our proposed incremental-2 is more efficient than Incremental-1.

On the basis of the above theoretical and experimental results, further studies are given below:

(1) We will apply the presented incremental methods to some practical applications;
(2) We will investigate the effect of the number of sample subsets on the runtime and the classification accuracy of the proposed incremental methods;
(3) To improve the classification performance of the learning algorithms, we will investigate how to establish novel incremental feature selection method in covering rough sets.

Acknowledgements The authors declare that they have no conflict of interest. This paper is funded by the National Natural Science Foundation of China (grant numbers 61806108, 12171388, 12071131 and 52175493) and the Fundamental Research Funds for the Central Universities (grant number 2019RC055).

References

Bonikowski Z, Bryniarski E, Wybraniec-Skardowska U (1998) Extensions and intentions in the rough set theory. Inf Sci 107(1):149–167
Chen DG, Li WL, Zhang X, Kwong S (2014) Evidence-theory-based numerical algorithms of attribute reduction with neighborhood-covering rough sets. Int J Approx Reason 55(3):908–923
Chen HM, Li TR, Luo C, Horng SJ, Wang GY (2015) A decision-theoretic rough set approach for dynamic data mining. IEEE Trans Fuzzy Syst 23(6):1958–1970
Chen DG, Yang YY, Dong Z (2016) An incremental algorithm for attribute reduction with variable-precision rough sets. Appl Soft Comput 45:129–149
Cousio I, Dubois D (2011) Rough sets, coverings and incomplete information. Fundamenta Informaticae 108(3–4):223–247
Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151:155–176
Dong Z, Sun M, Yang YY (2016) Fast algorithms of attribute reduction for covering decision systems with minimal elements in discernibility matrix. Int J Mach Learn Cybern 7(2):297–310
Du Y, Hu QH, Zhu PF, Ma PJ (2011) Rule learning for classification based on neighborhood covering reduction. Inf Sci 181(24):5457–5467
Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gener Syst 17(2–3):191–209
Feng SR, Zhang DZ (2012) Increment algorithm for attribute reduction based on improvement of discernibility matrix. J Shenzhen Univ Sci Eng 29(5):405–411
Grzymla-Busse JW (2004) Characteristic relations for incomplete data: A generalization of the indiscernibility relation, In Rough Sets and Current Trends in Computing, Springer, 244-253
Guan YY, Wang HK (2006) Set-valued information systems. Inf Sci 176(17):2507–2525
Hu QH, Yu DR, Xie ZX (2006) Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recogn Lett 27(5):414–423
Hu F, Dai J, Gy Wang (2007) Incremental algorithms for attribute reduction in decision table. Control Decis 22(3):268–272
Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178(18):3577–3594
Hu F, Wang GY, Huang H, Wu Y (2005) Incremental attribute reduction based on elementary sets, In Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Springer, 185-193
Jensen R, Shen Q (2004) Fuzzy rough attribute reduction with application to web categorization. Fuzzy Sets Syst 141(3):469–485
Jensen R, Shen Q (2007) Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst 15(1):73–89
Jing YG, Li TR, Luo C, Horng SJ, Wang GY, Yu Z (2016) An incremental approach for attribute reduction based on knowledge granularity. Knowl-Based Syst 104:24–38
Jing YG, Li TR, Fujita H, Yu Z, Wang B (2017) An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view. Inf Sci 411:23–38
Kira K, Rendell LA (1992) A practical approach to feature selection. Mach Learn Proc 48(1):249–256
Koller D, Sahami M (1996) Toward optimal feature selection, In 13th International Conference on Machine Learning, 284-292
Lang GM, Li QG, Cai MJ, Yang T (2015) Characteristic matrices-based knowledge reduction in dynamic covering decision information systems. Knowl-Based Syst 85:1–26
Li TR, Ruan D, Geert W, Song J, Xu Y (2007) A rough sets based characteristic relation approach for dynamic attribute generalization in data mining. Knowl-Based Syst 20(5):485–494
Liang JY, Xu ZB (2002) The algorithm on knowledge reduction in incomplete information systems. Int J Uncertain, Fuzziness Knowl-Based Syst 10(1):95–103
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

Springer
Ming Y (2007) An incremental updating algorithm for attribute reduction based on improved discernibility matrix. Ch J Comput 30(5):815–822

Orlowska ME, Orlowski MW (1992) Maintenance of knowledge in dynamic information systems, I: Intelligent Decision Support. Springer, 315–329

Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356

Qian YH, Dang CY, Liang JY, Tang DW (2009) Set-valued ordered information systems. Inf Sci 179(16):2809–2832

Qian YH, Liang JY, Pedrycz W, Dang CY (2010) Positive approximation: an accelerator for attribute reduction in rough set theory. Artif Intell 174(9):597–618

Qian YH, Liang JY, Song P, Dang CY, Wei W (2012) Evaluation of the decision performance of the decision rule set from an ordered decision table. Knowl-Based Syst 36:39–50

Robnik M, Ikonja I (2003) Kononenko, Theoretical and empirical analysis of relief and relieff. Mach Learn 53(1–2):23–69

Sang BB, Chen HM, Yang L, Li TR, Xu WH Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Dominance Neighborhood Rough Sets. IEEE Trans Fuzzy Syst

Sang BB, Chen HM, Li TR, Xu WH, Yu H (2020) Incremental approaches for heterogeneous feature selection in dynamic ordered data. Inf Sci 541:475–501

Sang BB, Chen HM, Yang L, Zhou DP, Li TR, Xu WH (2021) Incremental attribute reduction approaches for ordered data with time-evolving objects. Knowl-Based Syst 212:1–18

Shu WH, Shen H (2013) A rough-set based incremental approach for updating attribute reduction under dynamic incomplete decision systems. In Fuzzy Systems (FUZZ), 2013 IEEE International Conference on, IEEE, 1–7

Skowron A, Rauszer C (1992) The discernibility matrices and functions in information systems. In: Slowinski R (ed) Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory. Kluwer, Dordrecht, The Netherlands, pp 331–362

Slezak D (2002) Approximate entropy reduces. Fundamenta Informatice 53(3):365–390

Slowinski R, Vanderpooten D (2000) A generalized definition of rough approximations based on similarity. IEEE Trans Knowl Data Eng 12(2):331–336

Stefanowski J, Tsoukias A (1999) On the extension of rough sets under incomplete information, in: New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, Springer, 73-81

Tsang EC, Chen DG, Yeung DS, Wang XZ, Lee JW (2008) Attributes reduction using fuzzy rough sets. IEEE Trans Fuzzy Syst 16(5):1130–1141

Wang GY, Yu H, Yang DC (2002) Decision table reduction based on conditional information entropy. Ch J Comput 25(7):759–766

Wang CZ, He Q, Chen DG, Hu QH (2014) A novel method for attribute reduction of covering decision systems. Inf Sci 254:181–196

Wang CZ, Shao MW, Sun BZ, Hu QH (2015) An improved attribute reduction scheme with covering based rough sets. Appl Soft Comput 26:235–243

Wang CZ, Qi YL, Shao MW et al (2016) A fitting model for feature selection with fuzzy rough sets. IEEE Trans Fuzzy Syst 25(4):741–753

Wang CZ, Shao MW, He Q, Qian YH, Qi YL (2016) Feature subset selection based on fuzzy neighborhood rough sets. Knowl-Based Syst 111(1):173–179

Wang CZ, Hu QH, Wang XZ, Chen DG, Qian YH (2018) Feature selection based on neighborhood discrimination index. IEEE Trans Neural Netw Learn Syst 29(7):2986–2999

Yang XB, Qi Y, Yu HL, Song XN, Yang JY (2014) Updating multigranulation rough approximations with increasing of granular structures. Knowl-Based Syst 64:59–69

Yang YY, Chen DG, Wang H, Tsang ECC, Zhang DL (2017) Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving. Fuzzy Sets Syst 312:66–86

Yang YY, Chen DG, Wang H (2017) Active sample selection based incremental algorithm for attribute reduction with rough sets. IEEE Trans Fuzzy Syst 25(4):825–838

Yang YY, Chen DG, Wang H, Wang XZ (2018) Incremental Perspective for Feature Selection Based on Fuzzy Rough Sets. IEEE Trans Fuzzy Syst 26(3):1257–1273

Yao YY (1998) Relational interpretations of neighborhood operators and rough set approximation operators. Inf Sci 111(1):239–259

Yao YY (2006) Neighborhood systems and approximate retrieval. Inf Sci 176(23):3431–3452

Zakowski W (1983) Approximations in the space \((u, \pi)\). Demonstr Mathematica 16(3):761–769

Zhang X, Mei CL, Chen DG, Li JH (2013) Multi-confidence rule acquisition oriented attribute reduction of covering decision systems via combinatorial optimization. Knowl-Based Syst 50:187–197

Zhang X, Mei CL, Chen DG, Li JH (2016) Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy. Pattern Recognit 56:1-15

Zhang X, Mei CL, Chen DG, Yang YY, Li JH (2020) Active incremental feature selection using a Fuzzy-Rough-Set-based information entropy. IEEE Trans Fuzzy Syst 28(5):901–915

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

2669