Approaches to Knowledge Reduction of Decision Systems based on Conditional Rough Entropy

Lin Sun¹,*, Jiucheng Xu¹, Lingjun Zhang¹
¹College of Computer & Information Technology, Henan Normal University, Henan 453007, China, E-mail: linsunok@gmail.com

Abstract

Knowledge reduction in rough set theory is an important feature selection method. Since it is an NP-hard problem, it is necessary to investigate fast and effective approximate algorithms. In this paper, to address this issue, by introducing rough entropy in information systems, the novel measures of conditional rough entropy with distinguishing consistent objects from inconsistent objects are presented for both consistent and inconsistent decision systems. Thus, many important propositions, properties, and conclusions for reduct are drawn, and by using decomposition, radix sorting, hash, and input sequence techniques, we construct a forward greedy algorithm for knowledge reduction. Finally, through analyzing the given example, compared with some standard UCI datasets and other knowledge reduction algorithms, the proposed technique is effective and suitable for both consistent and inconsistent decision systems. Thus, it establishes the theoretical basis for seeking efficient algorithm of knowledge acquisition in decision systems.

Keywords: Rough Set Theory, Knowledge Reduction, Decision System, Rough Entropy, Conditional Rough Entropy

1. Introduction

Knowledge discovery in database attempts to identify valid, novel, potentially useful, and ultimately understandable patterns in data. One of the key problems of knowledge discovery is knowledge reduction. Rough set theory [1-3] is perhaps the most recent one making significant contribution to the field. In recent years, it has been demonstrated to be useful in fields such as pattern recognition, machine learning, automated knowledge acquisition, and so on [3-5]. The advantage of the rough set data analysis is that it uses only internal knowledge, avoids external parameters, and does not rely on prior model assumptions. Among other things, attribute reduction is a fundamental aspect of rough sets theory, which involves the search for a minimal subset of condition attributes such that the reduced set provides the same classification obtained by using the original condition attribute set. Knowledge reduction is performed in information systems by means of the notion of a reduct based on a specialization of the general notion of independence due to Marczewski [6]. Many types of knowledge reduction have been proposed in the analysis of information systems or decision systems [1-4, 6-9]. Each of them aims at some basic requirements. Various approaches have also been developed to perform knowledge reduction and obtain optimal true, certain, and possible decision rules in decision systems. For example, possible rules and reducts have been proposed to deal with inconsistency in inconsistent decision systems [10].

Attribute reduction has been carried out in consistent decision systems over the years [1, 6-9]. In the real world, however, most decision systems are inconsistent because of noise in data, compact representation, limited prediction capability, etc. More attention has been paid to knowledge reduction in inconsistent systems in rough set research [10]. In view of decision makers’ preferences and data decomposition [11], Slezak proposed new kinds of reducts which seem to be more suitable for finding decision rules than those defined by crisp indiscernibility. He also showed that knowledge reduction preserving the membership distribution is equivalent to knowledge reduction preserving the value of the generalized inference measure function. Recently, by eliminating the rigorous conditions required by distribution reduct, maximum distribution reduct was introduced [12]. Ma et al. [13] proposed notions of the distribution
reduct and maximum distribution reduct, and discussed the relationships among the maximum distribution reduct, the distribution reduct, and the possible reduct. Unlike possible reduct [14], maximum distribution reduct can derive decision rules that are compatible with the original systems. In addition, in traditional rough set theory, MIBARK-algorithm [15] cannot ensure the reduct is the minimal attribute subset, which keeps the decision rule invariant in inconsistent decision systems. Although, many methods above are efficient, they are still unsuitable for the reduct of voluminous data, and only suitable for consistent or inconsistent decision systems. Therefore, proposing an efficient and effective attribute evaluation approach to knowledge reduction, which is suitable for not only information systems but also consistent and inconsistent decision systems, is very desirable. This paper focuses on creating such a solution.

In this paper, the problems of knowledge representation and knowledge reduction are addressed in both consistent and inconsistent decision systems. Then, introducing the classical rough entropy, we propose the conditional rough entropy in consistent decision systems to measure the roughness of knowledge, and that of rough set in inconsistent decision systems. The conclusion shows that the conditional rough entropy of knowledge monotonically decreases as the information granule becomes small through finer classification in both consistent and inconsistent decision systems. By using decomposition, radix sorting, hash, and input sequence techniques, we have succeeded in creating an efficient knowledge reduction algorithm for decision systems. The algorithm does not change the size of the original system, and its complexity is cut down to approximately $O(|C|^2|U|)$. Hence, it is relatively efficient method and suitable for information systems, consistent decision systems, and inconsistent decision systems.

2. Some basic concepts and related works

In this section, we review briefly some notions and results related to information systems and decision systems. Detailed description and formal definitions of the theory can be found in [1, 2].

An information system is a pair $S = (U, A)$, where $U = \{u_1, u_2, \ldots, u_n\}$ is a non-empty finite set of objects; $A$ is a non-empty finite set of attributes; for every $a \in A$, there is a mapping $f: U \rightarrow V_a$, where $V_a$ is called the value set of $a$.

Each subset of attributes $P \subseteq A$ determines a binary indistinguishable relation $IND(P)$, given by $IND(P) = \{(u, v) \in U \times U \mid f(a, u) = f(a, v), \forall a \in P\}$. Obviously, $IND(P) = \bigcap_{a \in P} IND(\{a\})$. It can be shown that $IND(P)$ is an equivalence relation on $U$. For any $P \subseteq A$, the relation $IND(P)$ constitutes a partition of $U$, which is denoted by $U/IND(P)$, or just $U/P$. That is, $U/P = \{[u]_P \mid u \in U\}$ is called an information on $U$, where $[u]_P = \bigcup \{u \in U \mid (u, u) \in IND(P)\} = \bigcup \{u \in U \mid f(a, u) = f(a, u), \forall a \in P\}$ is called an equivalence block (equivalence class) of $u$, with reference to $P$.

In particular, if $U/P = \{X \mid X = \{u\}, u \in U\} = \delta$, it is called an identity relation, and if $U/P = \{X \mid X = \{\}\} = \emptyset$, it is called a universal relation.

Now, we define a partial order on all partition sets of $U$. Let $U/P$ and $U/Q$ be two partitions of a finite set $U$. If $Q \subseteq P$, then we define that the partition $U/Q$ is coarser than the partition $U/P$ (or the partition $U/P$ is finer than the partition $U/Q$), denoted by $P \leq Q$, between partitions by $P \leq Q \iff \forall P_i \in U/P, \exists Q_i \subseteq U/Q \rightarrow P_i \subseteq Q_i$. Hence, in an information system $S = (U, A)$, we have $P \leq Q$ for any $Q \subseteq P \subseteq A$.

Let $S = (U, A = C \cup D)$ be a decision system, where $C$ is a condition attribute set, and $D$ is a decision attribute set with $C \cap D = \emptyset$. Thus, for any $P \subseteq C$, the positive region of $P$ to $D$ is denoted by $POS_P(D) = \bigcup \{PX \mid X \in U/D\}$, where $PX = \bigcup_{u \in U} [\{u\}_P \subseteq X]$ is called the $P$ lower approximation of set $X$.

In [2], if $POS_C(D) = U$, then this decision system is called a consistent one, otherwise an inconsistent one. If $S = (U, C, D)$ is an inconsistent decision system, then the set $POS_C(D)$ is called the consistent objects set of $S$, and the set $U/POS_C(D)$ is called the inconsistent objects set of $S$.

Let $U$ be a given universe and $P, Q \subseteq C \cup D$, $U/P = \{X_1, X_2, \ldots, X_n\}$, $U/Q = \{Y_1, Y_2, \ldots, Y_m\}$, then the conditional information entropy of knowledge $Q$ with reference to $P$ is denoted by...
\[
H(Q \mid P) = -\sum_{i \in I} \frac{|X_i|}{|U|} \sum_{j \in J} \frac{|Y_j \cap X_i|}{|X_i|} \log \frac{|Y_j \cap X_i|}{|X_i|}.
\]

Thus, let \( S = (U, C, D) \) be a decision system and for any \( a \in P \subseteq C \). If \( H(D|P) = H(D|P-\{a\}) \), then \( a \) is unnecessary for \( D \) in \( P \), else necessary. For any \( P \subseteq C \), if every element in \( P \) is necessary for \( D \), then \( P \) is independent to \( D \).

Hence, in a decision system \( S = (U, C, D) \), if \( A_2 \subseteq A_1 \subseteq C \), then one has that \( H(D|A_1) \leq H(D|A_2) \). The necessary and sufficient condition of equation is that for any \( X_i, X_j \subseteq U/A_1, X_i \neq X_j \), if \( X_i \cup X_j \subseteq U/A_2 \), then \( \frac{|X_i \cap D_i|}{|X_i|} = \frac{|X_j \cap D_j|}{|X_j|} \) always holds, where \( D_i \subseteq U/D \).

### 3. Conditional rough entropy and knowledge reduction algorithm

In this section, we introduce the classical rough entropy of knowledge in information systems [16], propose conditional rough entropy of knowledge in both consistent and inconsistent decision systems, and derive some important propositions, properties, and conclusions for an effective reduce way of decision systems.

#### 3.1 Interpretations of rough entropy and conditional rough entropy

Let \( S = (U, A) \) be an information system and \( P \subseteq A \), \( U/P = \{X_1, X_2, \ldots, X_n\} \). Then the rough entropy of knowledge \( P \) is denoted by

\[
RE(P) = -\sum_{i \in I} \frac{|X_i|}{|U|} \log_2 \frac{1}{|X_i|},
\]

where \( \frac{|X_i|}{|U|} \) represents the probability of equivalence block \( X_i \) within the universe \( U \), and \( \frac{1}{|X_i|} \) denotes the probability of one of the values in equivalence block \( X_i \).

Thus, from the above formulation of rough entropy, it is easy to obtain the following properties in an information system.

**Property 1.** If \( U/P = \omega \), then the rough entropy of \( P \) achieves its minimum value 0.

**Property 2.** If \( U/P = \delta \), then the rough entropy of \( P \) achieves its maximum value \( \log_2|U| \).

Obviously, when \( IND(A) \) is an identity relation on \( U \), or a universal relation on \( U \), we find that \( 0 \leq RE(A) \leq \log_2|U| \).

**Proposition 1.** Given an information system \( S = (U, A) \), if \( P, Q \subseteq A \) and \( P \prec Q \), then \( RE(P) < RE(Q) \).

**Proof.** Let \( U \) be a given universe and \( P, Q \subseteq A \), \( U/P = \{X_1, X_2, \ldots, X_m\} \), \( U/Q = \{Y_1, Y_2, \ldots, Y_n\} \). Since \( P \prec Q \), it shows that \( n > m \), and then there exists a partition \( \{I_1, I_2, \ldots, I_m\} \) of \( \{1, 2, \ldots, n\} \) such that \( Y_j = \bigcup \{X_i \mid i \in I_j, j = 1, 2, \ldots, m\} \). Hence, we find that

\[
RE(Q) = -\sum_{j=1}^{m} \frac{|Y_j|}{|U|} \log_2 \frac{1}{|Y_j|} = -\frac{1}{|U|} \sum_{j=1}^{m} \frac{|Y_j|}{|U|} \sum_{i \in I_j} X_i \log_2 \frac{1}{|X_i|} = -\frac{1}{|U|} \sum_{j=1}^{m} \left( \sum_{i \in I_j} X_i \log_2 \frac{1}{|X_i|} \right).
\]

Since \( n > m \), it follows that there exists \( I_{h} \in \{I_1, I_2, \ldots, I_m\} \) such that \( |I_{h}| > 1 \). Thus, one has that

\[
\sum_{i \in I_{h}} X_i \log_2 \sum_{i \in I_{h}} |X_i| < \sum_{i \in I_{h}} |X_i| \log_2 \frac{1}{|X_i|}, \quad \text{and} \quad \sum_{i \in I_{j}, j \neq h} X_i \log_2 \sum_{i \in I_{j}, j \neq h} |X_i| \leq \sum_{i \in I_{j}, j \neq h} |X_i| \log_2 \frac{1}{|X_i|}.
\]

Therefore, it obviously follows that

\[
RE(Q) = -\frac{1}{|U|} \sum_{j=1}^{m} \left( \sum_{i \in I_j} X_i \log_2 \sum_{i \in I_j} |X_i| \right) > -\frac{1}{|U|} \sum_{j=1}^{m} \left( \sum_{i \in I_j} |X_i| \log_2 \frac{1}{|X_i|} \right) = -\frac{1}{|U|} \sum_{i \in I_j} |X_i| \log_2 \frac{1}{|X_i|} = RE(P).
\]

Hence, obviously \( RE(P) < RE(Q) \) and the proposition holds.

Thus, it can be seen easily from Proposition 1 that the rough entropy of knowledge decreases as the equivalence blocks become smaller through finer partitioning.
Proposition 2. Given an information system $S = ( U, A )$, if $P, Q \subseteq A$, then one has that 
$IND(P) \cap IND(Q) = IND(P \cup Q)$

Proof. Let $IND(P) = \bigcap_{a \in P} IND(\{ a \})$ and $IND(Q) = \bigcap_{a \in Q} IND(\{ a \})$, then we find that 
$IND(P) \cap IND(Q)$
$= (\bigcap_{a \in P} IND(\{ a \})) \cap (\bigcap_{a \in Q} IND(\{ a \}))$
$= \bigcap_{a \in P \cap Q} IND(\{ a \})$
$= \bigcap_{a \in P \cup Q} IND(\{ a \})$
$= IND(P \cup Q)$. Hence, obviously $IND(P) \cap IND(Q) = IND(P \cup Q)$ and the proposition holds.

Proposition 3. Let $U/P$ and $U/Q$ be two classifications with the respective indistinguishable relations $P$ and $Q$ on $U$. The intersection $\cap$ between two classifications $U/P$ and $U/Q$ is denoted as follows: $U(P \cap Q) = U((P \cup Q)) = U(P \cup Q)$ (also called classification $U/P$ AND $U/Q$).

Proof. It is straightforward from Proposition 2.

Thus, from Proposition 3, it is easy to obtain the following property in decision systems.

Property 3. Let $U$ be a given universe and $P \subseteq C$, $U/P = \{ X_1, X_2, \ldots, X_n \}$, for any $a \in C - P$, then $U((P \cup \{ a \}) = \bigcup \{ X/\{a \} \} | i = 1, 2, \ldots, n \}$, i.e. $U(P \cup \{ a \}) = \bigcup \{ X/\{a \} | X \in U/P \}$.

Definition 1. Let $S = ( U, C, D )$ be a decision system and $P, Q \subseteq C \cup D$, $U/P = \{ X_1, X_2, \ldots, X_n \}$, $U/Q = \{ Y_1, Y_2, \ldots, Y_m \}$, and $U/(P \cup Q) = U(Q \cap U) = \{ Y_1 \cap X_1, Y_1 \cap X_2, \ldots, Y_m \cap X_1, Y_m \cap X_m \}$, then the rough entropy of knowledge $Q \cup P$ is defined as 
$RE(Q \cup P) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{| Y_i \cap X_j |}{| U |} \log_2 \frac{1}{| Y_i \cap X_j |}$. (3)

In [1], let $U$ be a given universe and $P \subseteq C$, then $H(D|P) = H(D|U) - H(P)$, where $H(D|P)$ is the conditional information entropy of $D$ with reference to $P$, and $H(D|U)$ and $H(P)$ represent the information entropy of $D$ and $P$, respectively.

Based on these representations and from Proposition 1, it shows that $D \cup P \leq P$, then $RE(D \cup P) \leq RE(P)$. Hence, we give the definition of conditional rough entropy in decision systems as follows.

Definition 2. Let $S = ( U, C, D )$ be a decision system and $P \subseteq C$, $U/P = \{ X_1, X_2, \ldots, X_n \}$, $U/D = \{ Y_1, Y_2, \ldots, Y_m \}$. Then, let $RE(D|P)$ denote the conditional rough entropy of $D$ with reference to $P$ of $S$ as follows 
$RE(D|P) = RE(P) - RE(D \cup P)$
$= -\sum_{i=1}^{n} \frac{| X_i |}{| U |} \log_2 \frac{1}{| X_i |} + \sum_{i=1}^{n} \frac{| X_i |}{| U |} \sum_{j=1}^{m} \frac{| Y_j \cap X_i |}{| X_i |} \log_2 \frac{| X_i |}{| Y_j \cap X_i |}$
$= \sum_{i=1}^{n} \frac{| X_i |}{| U |} \log_2 | X_i | (1 - \sum_{j=1}^{m} \frac{| Y_j \cap X_i |}{| X_i |}) + \sum_{i=1}^{n} \frac{| X_i |}{| U |} \sum_{j=1}^{m} \frac{| Y_j \cap X_i |}{| X_i |} \log_2 \frac{| X_i |}{| Y_j \cap X_i |}$
$= \sum_{i=1}^{n} \frac{| X_i |}{| U |} \sum_{j=1}^{m} \frac{| Y_j \cap X_i |}{| X_i |} \log_2 \frac{| X_i |}{| Y_j \cap X_i |}$.

Proposition 4. Let $U$ be a given universe and $P \subseteq C$, $U/P = \{ X_1, X_2, \ldots, X_n \}$, $U/D = \{ Y_1, Y_2, \ldots, Y_m \}$. Assume that $U/(P-\{ a \}) = \{ X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_q, X_{q+1}, \ldots, X_r, X_r \cup X_q \}$ is another partition generated through combining equivalence blocks $X_p$ and $X_q$, with $X_r \cup X_q$ and $X_r$ are two equivalence blocks randomly selected from $U/P$, and $a \in P$. Then one has that $RE(D|P) \leq RE(D|P-\{ a \})$.

Proof. From Definition 2, we can obtain that 
$RE(D|P-\{ a \}) = RE(D|P) - \sum_{i=1}^{n} \frac{| X_i |}{| U |} \log_2 \frac{1}{| X_i |} - \sum_{j=1}^{m} \frac{| Y_j \cap Y_p \cup Y_q |}{| Y_j \cap Y_q |} \log_2 \frac{| Y_j \cap Y_p \cup Y_q |}{| Y_j \cap Y_q |}$
$+ \sum_{j=1}^{m} \frac{| Y_j \cap (X_p \cup X_q) |}{| X_p \cup X_q |} \log_2 \frac{| X_p \cup X_q |}{| Y_j \cap (X_p \cup X_q) |}$
$= \sum_{i=1}^{n} \frac{| X_i |}{| U |} \sum_{j=1}^{m} \frac{| Y_j \cap X_i |}{| X_i |} \log_2 \frac{| X_i |}{| Y_j \cap X_i |}$
$+ \sum_{j=1}^{m} \frac{| Y_j \cap (X_p \cup X_q) |}{| X_p \cup X_q |} \log_2 \frac{| X_p \cup X_q |}{| Y_j \cap (X_p \cup X_q) |}$.
Let \( |X_a| = x, |X_b| = y, |X_c Y| = ax, \) and \( |X_c \cap Y| = by. \) There must be \( x > 0, y > 0, 0 \leq a \leq 1, \) and \( 0 \leq b \leq 1 \) such that one has that

\[
RE_\Delta = RE(D|P) - RE(D|P - \{a\}) = \frac{1}{|U|} \sum_{j=1}^{m} (ax \log_2 \frac{ax + by}{ax + ay} + by \log_2 \frac{ax + by}{bx + by}).
\]

Assume that a function \( f_j = ax \log_2 \frac{ax + by}{ax + ay} + by \log_2 \frac{ax + by}{bx + by}, \) where \( j = 1, 2, \ldots, m. \) Hence, if \( a = 0 \) or \( b = 0, i.e. X_c Y = \emptyset \) or \( X_c \cap Y = \emptyset, \) then \( f_j < 0, \) and if \( a = b = 0, i.e. X_c Y = X_c \cap Y = \emptyset, \) then \( f_j = 0. \)

Thus, we consider the cases of \( 0 < a \leq 1 \) and \( 0 < b \leq 1 \) in the following proof.

Suppose that \( ax = \lambda \) and \( by = \beta, \) there must be \( \lambda > 0 \) and \( \beta > 0, \) then we find that

\[
f_j = \lambda \log_2 \frac{\lambda + \beta}{\lambda + \beta} + \beta \log_2 \frac{\lambda + \beta}{\lambda + \beta}.\]

Thus, we calculate the derivative of \( f_j \) with reference to \( \varphi \) as follows:

\[
\frac{d(f_j)}{d(\varphi)} = -\lambda \frac{\beta}{(\lambda + \beta)} \ln 2 + \beta \frac{1}{(\lambda + \beta) \ln 2} \frac{\lambda}{(\lambda + \beta) \ln 2} = \frac{\lambda \beta (1 - \varphi)}{\varphi (\lambda + \beta) \ln 2}.
\]

Therefore, if \( 0 < \varphi < 1, \) then \( \frac{d(f_j)}{d(\varphi)} > 0. \) If \( \varphi = 1, \) then \( \frac{d(f_j)}{d(\varphi)} = 0. \) If \( \varphi > 1, \) then \( \frac{d(f_j)}{d(\varphi)} < 0. \) Thus, \( f_j \) has its maximal value \( 0, i.e. f_j = 0, \) when \( \varphi = 1. \) Above all, if \( a = b, i.e. |Y_j \cap X_p| = |Y_j \cap X_q|, \)

then one has that \( RE_\Delta = RE(D|P) - RE(D|P - \{a\}) = 0. \) If \( a \neq b, \) then one has that \( RE_\Delta = RE(D|P) - RE(D|P - \{a\}) < 0. \) Hence, obviously \( RE(D|P) \leq RE(D|P - \{a\}) \) and the proposition holds.

Therefore, according to Proposition 4, we know that the combination of blocks induced by condition attributes will increase the conditional rough entropy monotonously, and the conditional rough entropy will remain unchanged only if the memberships of the combined blocks for all decision blocks are the same. Thus, the memberships of all equivalence blocks, induced by condition attributes for all decision blocks, will remain unchanged after the combination. Hence, it is easy to obtain the following properties from Proposition 4 and Theorem 4.1 in [1].

\textbf{Property 4.} Let \( U \) be a given universe and for any \( a \in P \subseteq C, \) if \( RE(D|P) = RE(D|P - \{a\}), \) then \( POS(D) = POS_{P - \{a\}}(D). \)

\textbf{Property 5.} Let \( U \) be a given universe and for any \( a \in P \subseteq C, \) if \( H(D|P) = H(D|P - \{a\}), \) then \( POS(D) = POS_{P - \{a\}}(D). \)

\textbf{Property 6.} Let \( U \) be a given universe and for any \( a \in P \subseteq C, \) then \( RE(D|P) = RE(D|P - \{a\}) \iff H(D|P) = H(D|P - \{a\}). \)

\textbf{Property 7.} Let \( U \) be a given universe and \( P \subseteq Q \subseteq C, \) then \( RE(D|Q) \leq RE(D|P). \)

\textbf{Property 8.} Let \( U \) be a given universe and \( P \subseteq C, \) then \( RE(D|C) \leq RE(D|P). \)

For a general information system, the definition of reducts in the algebra view is equivalent to its definition in the information view [1]. Thus, from these representations above, we find that the definition of the relative reducts of a consistent decision system (i.e., there are no conflicts or inconsistent objects in the decision system) in the algebra view is also equivalent to its definition, based on the conditional rough entropy. However, inconsistent decision systems occur often in real life, and then we need to calculate the reducts of inconsistent decision systems. Hence, the conditional rough entropy of a relative reduct defined above may not remain unchanged. On the other hand, if an attribute cannot provide any additional information for an existing attribute set to make a decision system, then it is reducible. That is, the definition of relative reducts of a decision system, based on the conditional rough entropy, includes its definition in the algebra view. Any relative reduct of a decision
system, based on the conditional rough entropy, must be its relative reduct in the algebra view. Thus, this paper focuses on creating such a heuristic algorithm developed further using this result.

In [15], let $S = (U, C, D)$ be a decision system and $P \subseteq C$. If $H(D|P) = H(D|C)$, and $P$ is independent to $D$, then $P$ is called the attribute reduct of $C$ with reference to $D$. Thus, we introduce the idea of reduct above to construct the following reduct based on the conditional rough entropy.

**Definition 3.** Let $S = (U, C, D)$ be a decision system and $P \subseteq C$. If $RE(D|P) = RE(D|C)$, and $P$ is independent to $D$, then $P$ is called the attribute reduction of $C$ with reference to $D$.

From the analyses above, in the algebra view, the significance measure [8] is regarded as the quantitative computation of radix for the objects of positive region, merely describes the subsets of certain blocks in $U$. Moreover, in the information view, the significance measure [15, 17] is only obtained by detaching objects in different decision blocks from equivalent blocks generated by the condition attribute subset. Since there exist inconsistent objects, however, these current measures to rough set still lack of dividing $U$ into consistent object sets and inconsistent object sets in inconsistent decision systems. Therefore, the heuristic reduct measures will not be equivalent in the representation of concepts and operations for inconsistent decision systems. It is necessary to seek for a new kind of measure to knowledge roughness for searching the precise reduct effectively in decision systems.

**Proposition 5.** Given a decision system $S = (U, C, D)$, if $A, B \subseteq C$, $U/D = \{D_1, D_2, \ldots, D_m\}$, then $POS_c(D) = POS_d(D)$ if and only if $AD_i = BD_i$, where $i = 1, 2, \ldots, m$.

**Proof.** The proof is similar to that of Theorem 1 in [17].

Thus, obviously, Proposition 5 shows that when $A \subseteq C$, then $POS_c(D) = POS_c(D)$ if and only if $AD_1 = CD_1$, where $i = 1, 2, \ldots, m$.

In a decision system $S = (U, C, D)$, suppose that $D_0 = U – POS_c(D)$, from the definition of positive region, one has that $CD_0 = D_0$.

Therefore, when $A \subseteq C$, suppose that any set of $\{AD_0, AD_1, AD_2, \ldots, AD_m\}$ isn’t empty, then the sets must be also a decision partition of $U$. If there exists an empty decision block $AD_i$, then the $AD_i$ is called a redundant block of the partition. When the redundant blocks are taken out, it makes no difference to the decision partition. If the subset $A$ is a reduct of $C$, thus the partition $\{AD_0, AD_1, AD_2, \ldots, AD_m\}$ can divide $U$ into consistent object sets and inconsistent object sets, and then all inconsistent objects detached form the unattached set. Also, one has another partition $\{CD_0, CD_1, CD_2, \ldots, CD_m\}$ of $C$ on $U$, and then we construct a new equivalent relation, denoted by $R_D$ generated by the new partition above. Hence, it easily follows that there exists $UR_D = \{CD_0, CD_1, CD_2, \ldots, CD_m\}$.

Accordingly, it can be seen that the presented decision partition $UR_D$ has not only detached consistent objects from different decision blocks in $U$, but also distinguished consistent objects from inconsistent objects, while $U/D$ is gained through detaching objects from different decision blocks.

Thereby, distinguishing consistent objects from inconsistent objects will help us to get minimal or optimal reducts of knowledge. According to the analyses above, to compensate for the disadvantages of classical rough reduction algorithms, we propose the new definition of conditional rough entropy, which has effects not only on the subsets of certain blocks but also on the subsets of the uncertain blocks in $U$.

**Definition 4.** Let $S = (U, C, D)$ be a decision system and $P \subseteq C$. If $U/P = \{X_1, X_2, \ldots, X_n\}$, $D = \{d\}$, $U/D = \{Y_1, Y_2, \ldots, Y_m\}$, and $U/R_D = \{CY_0, CY_1, CY_2, \ldots, CY_m\}$, then let $RE(R_D|P)$ denote the conditional rough entropy of $D$ with reference to $P$ of $S$ as follows

$$RE(R_D|P) = \sum_{i=1}^{n} \frac{|X_i|}{|U|} \sum_{j=0}^{m} \frac{|CY_j \cap X_i|}{|X_i|} \log_2 \left( \frac{|X_i|}{|CY_j \cap X_i|} \right). \quad (4)$$

**Proposition 6.** Given a decision system $S = (U, C, D)$, if $C \subseteq P$, then $RE(R_D|C) \leq RE(R_D|P)$.

**Proof.** The proof is similar to that of Proposition 4.

In the following, we show the performance of Proposition 6 in a decision system through an illustrative example.

**Example 1.** Given a decision system $S = (U, C, D)$ in [1], shown in Table 1, where $U = \{x_1, x_2, \ldots, x_9\}$, $C = \{a, b, c, e\}$, and $D = \{d\}$. From Table 1, it can be obtained easily that $U/(\{a, b\}) = \{\{1\}, \{2\}, \{3, 4, 5, 6, 7, 8, 9\}\}$, $U/C = \{\{1\}, \{2\}, \{3\}, \{4, 5, 6\}, \{7, 9\}, \{8\}\}$, $U/D = \{\{1, 3, 8\}, \{2, 4, 5, 6, 7, 9\}\}$, and $UR_D = U/D$. Therefore, since $\{a, b\} \subseteq C$, that is, $C \prec \{a, b\}$, then we find that...
\[ RE(R_o \mid \{a, b\}) = \frac{7}{9} \left(\frac{2}{7} \log_2 \frac{7}{2} + \frac{5}{7} \log_2 \frac{7}{5}\right) = 0.671 > RE(R_o \mid C) = 0. \]

<table>
<thead>
<tr>
<th>( U )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( x_5 )</th>
<th>( x_6 )</th>
<th>( x_7 )</th>
<th>( x_8 )</th>
<th>( x_9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( b )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( c )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( e )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( d )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. A decision system

For convenience, the monotonicity of conditional rough entropy value, induced by the partial relation between \( P \) and \( C \), is called granulation monotonicity in a decision system. Hence, we have the conclusion that the conditional rough entropy of knowledge monotonically decreases as the information granule becomes small through finer classification in both consistent and inconsistent decision systems.

Thus, we define the conditional rough entropy above to measure the uncertainty of both consistent and inconsistent decision systems, and then may apply the entropy to reduce redundant attributes. Hence, the corresponding significance measures are listed as follows.

**Definition 5.** Let \( S = (U, C, D) \) be a decision system and \( P \subseteq C \), for any \( a \in P \), then the significance measure of \( a \) in \( P \) with reference to \( D \) is defined as
\[
\text{Sig}_{\text{outer}}(a, P, D) = RE(R_o \mid P \setminus \{a\}) - RE(R_o \mid P). \quad (5)
\]

**Definition 6.** Let \( S = (U, C, D) \) be a decision system and \( P \subseteq C \), for any \( a \in C \setminus P \), then the significance measure of \( a \) in \( P \) with reference to \( D \) is defined as
\[
\text{Sig}_{\text{outer}}(a, P, D) = RE(R_o \mid P) - RE(R_o \mid P \cup \{a\}). \quad (6)
\]

Notice that when \( P = C \), \( \text{Sig}_{\text{outer}}(\cdot, C, D) = 0 \). Obviously, \( 0 \leq \text{Sig}_{\text{outer}}(a, P, D) \leq \log_2 |U| \). If \( \text{Sig}_{\text{outer}}(a, P, D) = 0 \), then the attribute \( a \) is a current redundant attribute.

From Definition 6, it shows that the significance measure \( \text{Sig}_{\text{outer}}(a, P, D) \) indicates the importance of attribute \( a \) added to \( P \) with reference to \( D \) in decision systems, offering the powerful reference to the decision. The bigger the significance measure of attribute is, the higher its position in the decision system is, and otherwise the lower its position is. Therefore, if each of the significance measure of attribute is calculated, then the attribute with the zero or lower significance measure is removed, the knowledge reduction can be finished.

Thus, we know that if the novel proposed \( \text{Sig}_{\text{outer}}(a, P, D) = 0 \), then the significance measures of attribute, based on the positive region and the conditional information entropy, are also zero. On the other hand, if the objects radix of positive region fills out after adding any attributes, then the significance measure of attribute based on the positive region isn’t zero, accordingly we also have \( \text{Sig}_{\text{outer}}(a, P, D) \neq 0 \). So the new proposed \( \text{Sig}_{\text{outer}}(a, P, D) \) can not only include more information than that based on positive region, but also compensate for the limitations of algebra and information views.

Hence, all the definitions above are used as a heuristic algorithm of knowledge reduction to select a reduct from consistent or inconsistent data sets. For a given decision system, the intersection of all attribute reducts is said to be indispensable and is called the core. Each attribute in the core must be in every attribute reduction of the decision system. The core may be an empty set. The significance measures above can be used to find the core attributes. The following propositions are of interest with this regard.

**Proposition 7.** Given a decision system \( S = (U, C, D) \) and for any \( a \in C \), if \( \text{Sig}_{\text{outer}}(a, C, D) > 0 \), then \( a \) is a core attribute of \( S \).

In a heuristic algorithm of knowledge reduction, based on the above propositions, one can find a reduct by gradually adding selected attributes to the core attributes.

**Proposition 8.** Given a decision system \( S = (U, C, D) \) and \( P \subseteq C \), then any \( a \in P \) is dispensable in \( P \) with reference to \( D \) if and only if \( RE(R_o \mid P \setminus \{a\}) = RE(R_o \mid P) \).

**Definition 7.** Let \( S = (U, C, D) \) be a decision system and \( P \subseteq C \). The attribute set \( P \) is a reduct of \( C \) with reference to \( D \) if and only if for any \( a \in P \), \( RE(R_o \mid P) \neq RE(R_o \mid P \setminus \{a\}) \) and \( RE(R_o \mid P) = RE(R_o \mid C) \).

Thus, it easily shows from Definition 6 and Definition 7 that the significance measure and the relative reduct, based on the conditional rough entropy, not only is equivalent to that in the
algebra view and the information view in consistent decision systems, but also illustrates the performance in inconsistent decision systems. Therefore, the proposed reduction method is suitable for both consistent and inconsistent decision systems. Specially, after wiping off the decision attribute, it is also available in information systems.

3.2 Knowledge reduction algorithm

In the following, we will not only consider how to discretize numerical attributes and construct a heuristic function for knowledge reduction, but also focus on how to improve computational efficiency of a heuristic reduction algorithm in the context of large data sets. To this end, we construct the input sequence by sorting the attributes in order of increasing significance measure. The relative reduct can be found by repeatedly deleting the head node of the input sequence. Then we introduce the idea of partitioning. Thus, all of the policies will help to reduce the quantity of computation and the time-space of search.

**Algorithm 1.** For calculating $\text{Sig}_{\text{outer}}(c, C, D)$ and $\text{RE}(R_0 | P \cup \{a\})$

Input: $S = (U, C, D)$, where $U = \{u_1, u_2, \ldots, u_n\}$, $P \subseteq C$, $a \in C - P$, and $c \in C$.

Output: $U/C, U/D, U/R_0, U/P, U/(P \cup \{a\}), \text{RE}(R_0 | P), \text{RE}(R_0 | P \cup \{a\})$, and $\text{Sig}_{\text{inner}}(c, C, D)$.

1. Calculate $U/C$ and $U/D$ to obtain $\text{POS}_C(D)$ and $U - \text{POS}_C(D)$, by radix sorting and hash, thus get the partition $U/R_0$.
2. Calculate $U/R_0 \cup C = U/(C - \{c\}) \cup U/(R_0 \cup (C - \{c\}))$, by radix sorting and Property 3, thus get $\text{RE}(R_0 | C) = \text{RE}(C - \text{RE}(R_0 \cup C)$ and $\text{RE}(R_0 | C - \{c\}) = \text{RE}(C - \{c\}) - \text{RE}(R_0 \cup (C - \{c\}))$, to obtain $\text{Sig}_{\text{inner}}(c, C, D)$.
3. Calculate $U/P, U/(R_0 \cup P), U/(P \cup \{a\})$, and $U/(R_0 \cup P \cup \{a\})$, by radix sorting and Property 3, thus get $\text{RE}(R_0 | P) = \text{RE}(P) - \text{RE}(R_0 \cup P)$ and $\text{RE}(R_0 | P \cup \{a\}) = \text{RE}(P \cup \{a\}) - \text{RE}(R_0 \cup P \cup \{a\})$.

Thus, through computing partitions and positive region from Algorithm 1 based on radix sorting and hash, costing the time complexity $O(|U|)$, and then we make it easy to construct the efficient method of computing the significance measure in a decision system.

**Algorithm 2.** For calculating a reduct of a decision system

Input: $S = (U, C, D)$, where $C = \{c_1, c_2, \ldots, c_p\}, D = \{d\}$ and $p = |C|$.

Output: $\text{reduct}$, a reduct of $S$.

1. Let $P = \emptyset$.
2. Calculate $U/C, U/D$, and $U/R_0$ to obtain $\text{RE}(R_0 | C)$ with Algorithm 1.
3. Construct an ascending input sequence ordered by $\text{RE}(R_0 | \{a_i\})$ with Algorithm 1, where $a_i \subseteq C$, and denote the result by $<a_1, a_2, \ldots, a_p>$, and let $T = \{a_1, a_2, \ldots, a_p\}$.
4. Calculate $\text{Sig}_{\text{inner}}(a_i, C, D)$ with Algorithm 1, where $a_i \subseteq T$.
5. Put $a_i \subseteq T$ into $P$, where $\text{Sig}_{\text{inner}}(a_i, C, D) > 0$.
6. Let $\text{core} = P$, then go to (8);
7. Select $a_i$ with $\min\{|\text{RE}(R_0 | P \cup \{a_i\})|\}$ to put $a_i$ into $H$, where for any $a_i \subseteq C - P$ and $a_i \subseteq T$;
    - If $|H| \neq 1$, then select $a_i \subseteq H$ with $\min\{|U/(P \cup \{a_i\})|\}$;
    - If this selected $a_i$ is still not only, then select the front; $P = P \cup \{a_i\}$.
8. If $\text{RE}(R_0 | P) \neq \text{RE}(R_0 | C)$, then go to (7), else
   - Let $P = P - \text{core}$;
   - Construct an ascending input sequence, denoted by $P$, and let $P = T \cap P = \{a_1, a_2, \ldots, a_p\}$;
   - $t = |P|$;
   - For ($i = 1; i \leq t; i++)$
\[ \{a_i \in P; \]
\[ P = P \setminus \{a_i\}; \]
\[ \text{If } RE(R_P[|\text{core} \cup P|) \neq RE(R_P[C]), \text{ then } P = P \cup \{a_i\}; \} \]

(9) \text{ reduct } = P \cup \text{ core}.

It can be seen easily from (8) that the completeness for the minimal reduct of the above method is proved obviously. That is, none of the attributes in P can be eliminated again without decreasing its discriminating capability, whereas a great many reduction algorithms are incomplete, which can’t ensure that the final reduct will be obtained [8]. Then through analyzing, it is known that these reduction algorithms in [8, 17] are also complete, while those algorithms in [15] are not. Making full use of the feasible measures of computation regarding partitions, positive region and core attributes, we can easily see that the time complexity of the algorithm may be decreased to approximately \(O(|C|^2|U|)\), which is less than that of [8, 15, 17].

4. Experimental analysis

In the following, we give an example to explain the validity of the proposed knowledge reduction algorithm, and then apply the proposed approach and other knowledge reduction approaches to several data sets from the UCI Machine Learning Repository (These data sets can be downloaded at [http://www.ics.uci.edu](http://www.ics.uci.edu)), to evaluate the proposed approach.

**Example 2.** In Table 2, \(S = (U, C, D)\) is an inconsistent decision system, where \(U = \{x_1, x_2, \ldots, x_{16}\}\), \(C = \{a, b, c, d, e\}\).

**Table 2.** An inconsistent decision system

<table>
<thead>
<tr>
<th>(U)</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(x_4)</th>
<th>(x_5)</th>
<th>(x_6)</th>
<th>(x_7)</th>
<th>(x_8)</th>
<th>(x_9)</th>
<th>(x_{10})</th>
<th>(x_{11})</th>
<th>(x_{12})</th>
<th>(x_{13})</th>
<th>(x_{14})</th>
<th>(x_{15})</th>
<th>(x_{16})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(b)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(c)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(d)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(e)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(D)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

According to Table 2, based on the above measures of significance and the method of core attributes, we can calculate the significance and relative conditional significance of single attribute, and select the attributes with maximal significance and relative conditional significance one by one, and then we have the core attribute \(\{e\}\) and the minimal reduct \(\{a, b, e\}\). Thus, the significance measures of other condition attributes with respect to \(\{e\}\) are all shown in Figure 1. Then their advantages and disadvantages can be found easily through comparing roundly the Algorithm 4 in [8] and the Algorithm CEBARKCC in [15] with the proposed Algorithm 2, shortly denoted by Alg_a, Alg_b, and Alg_c, respectively.

![Figure 1](image-url) Comparisons of significance measure

According to Figure 1, it is known that \(SGF(b, \{e\}, D)\) in [1, 8, 15] is relative minimum. The searching results of the heuristic reduction algorithms in [8, 15], which are based on the positive region and the conditional information entropy, are \(\{a, c, d, e\}\), rather than the minimal reduct \(\{a, b, e\}\). The searching results of the heuristic reduction algorithms, which are based on the new conditional
information entropy [17] and the novel conditional rough entropy of knowledge, are the minimal reduct \( \{a, b, e\} \). In particular, \( \text{Sig}_{\text{outer}}(b, \{e\}, D) \) is relative maximum. Therefore, the relative significance measures of the attributes \( a \) and \( b \) are nicely represented by the proposed \( \text{Sig}_{\text{outer}} \), and then the proposed heuristic algorithm will be greatly efficient in searching the minimal or optimal reduct.

Here we choose six discrete data sets from the UCI Machine Learning Repository, and compare roundly the Algorithm CEBARKCC in [15], the Algorithm 4 in [8], the Algorithm 2 in [17] with the proposed Algorithm 2 above, shortly denoted by Alg_a, Alg_b, Alg_c, and Alg_d, respectively. The experimental hardware environment is as follows: Inter(R) Pentium(R) D CPU 3.4 GHz, 2 GB memory, Windows XP. What’s more, we employ \( m \) and \( n \) to denote the radix of the primal attribute set and the minimal reduct after reduction, respectively. Then, the comparison results are outlined in Table 3.

Table 3. Comparison results of knowledge reduction

<table>
<thead>
<tr>
<th>Data set</th>
<th>Consistent or not</th>
<th>( m )</th>
<th>( n )</th>
<th>( n )</th>
<th>( n )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver-disorders</td>
<td>Yes</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Zoo</td>
<td>No</td>
<td>17</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Vehicle</td>
<td>Yes</td>
<td>20</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mushroom</td>
<td>Yes</td>
<td>23</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Voting-records</td>
<td>Yes</td>
<td>17</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Breast cancer wisconsin</td>
<td>Yes</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

5. Conclusions

Nowadays, data analysis, dependency analysis, and learning are some of the most important applications of rough set theory. As an important concept of rough set theory, an attribute reduct is a subset of attributes that are jointly sufficient and individually necessary for preserving a particular property. In this paper, the problems of knowledge representation and knowledge reduction in both consistent and inconsistent decision systems are addressed, and then by introducing the rough entropy, we propose the novel conditional rough entropy. Thus, some important propositions, properties, and conclusions for reduct are drawn, and the efficient algorithms for decision systems are proposed. Numerical experiments indicate that the proposed algorithm is also efficient and of practical value in engineering. The experiment results are consistent with our theoretical analysis. In sum, the proposed method is an effective means of knowledge reduction for not only information systems but also both consistent and inconsistent decision systems. More generally, our work establishes the theoretical basis for seeking efficient algorithm of knowledge representation and knowledge acquisition from the rough entropy view of rough set theory in decision systems. Thus, based on these works, exploring efficient approaches of knowledge acquisition for incomplete information (decision) systems is our next task.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant (No. 60873104, No. 61040037), the Science and Technology Key Project of Henan Province of China (No. 112102210194), and the Natural Science Foundation of Educational Department of Henan Province of China (No. 2008B520019).

7. References

Approaches to Knowledge Reduction of Decision Systems based on Conditional Rough Entropy
Lin Sun, Jiucheng Xu, Lingjun Zhang


