NEW ROUGH SET APPROACH TO KNOWLEDGE REDUCTION IN
DECISION TABLE

JIAN-MEI XIAO, TENG-FEI ZHANG

Department of Electrical and Automation, Shanghai Maritime University, Shanghai 200135, China
E-MAIL: jmxiao@cen.shmtu.edu.cn

Abstract:
The core and knowledge reduction of a decision table are
the key points of many information process procedures. It has
been proved that computing all the reductions and the optimal
reduction of a decision table is a NP-complete problem. In this
paper, the algorithms for finding relative core and relative
knowledge reduction are presented, which are based on the
positive region in rough set theory. The effectiveness of the
algorithms is demonstrated by some typical examples.

Keywords:
rough set theory; decision table; positive region; core;
knowledge reduction

1. Introduction

Rough set theory was developed by Z. Pawlak in the
early 1980s[1, 2]. It is a valid mathematical tool to deal with
imprecise, incomplete and inconsistent data. After almost
20 years of pursuing and development, rough set theory has
been successfully applied to solving many real-life
problems in medicine, pharmacology, engineering, banking,
financial and market analysis, etc.

Rough set theory is based on the indiscernibility
relation. We can describe a set by the concepts of upper and
lower approximation set. Algorithm for core and reduction
in information system is a heat field on rough set theory
research and application. In practice, the rough set theory
mostly focuses on decision information systems, which are
categorized as decision tables. Therefore, it is very
important for finding relative core and reduction in decision
tables. But the knowledge reduction is a NP-complete
problem[3], and there isn’t a high-efficient method to
calculate the optimal reduction. But in the practical
applications, it can find a sub-optimal attribute reduction.
For this reason, lots of algorithms for core and reduction
have been proposed. A general method is adopting the
attribute tolerance of importance as the heuristic
information to seek reduction. Some of them are based on
such as discernibility matrix method[3], or based on positive
region[4], or based on information entropy[5], and so on.

In this paper, the rough set theory is deeply
investigated. Some properties of positive region are
discovered. Then, we present an algorithm for relative core
and two algorithms for relative reduction based on the
heuristic information of positive region.

2. Rough set theory

In this section, we illustrate the main concepts of
rough set theory, which are necessary for our further
formulation.

Knowledge representation in rough set theory is done
via information system. An information system can be
characterized as a decision table, where the information is
stored in a table. Each row in the table represents an
individual record. Each column represents some an attribute
of the records or a field.

Definition 1. An information system

\[ S = \langle U, R, V, f \rangle \]

\( U \) is a finite set of objects, \( R \) is a finite set of attributes
which are further classified into disjoint sets of condition
attributes \( C \) and decision attributes \( D \). \( R = C \cup D \),
\( C \cap D = \emptyset \). \( V \) is a set of attributes values. \( f : U \times R \rightarrow V \)
is an information function[] which appoint the value of each
object \( x \) in \( U \).

An equivalence relation is a reflexive, symmetric, and
transitive binary relation on \( U \). With each subset \( P \subseteq R \)
we associate an equivalence relation \( IND(P) \) on \( U \) by
setting

\[ IND(P) = \{ (x, y) \in U \times U : \forall a \in P, a(x) = a(y) \} \]

The partition associated with \( IND(P) \) is denoted as
\( U / IND(P) \). We denote by the symbol \([x]_{IND(P)}\) the
equivalence class in \( U / IND(P) \) which contains \( x \)

\[ [x]_{IND(P)} = \{ y : y \in U, xIND(P)y \} \]
**Definition 2.** Given the information system 
\[ S = \langle U, R, V, f \rangle \]
Each subset \( X \subseteq U \) and indiscernibility relation \( R \), the lower and upper approximation set can be separately defined as:
\[ RX = \bigcup \{ y \in U / R \mid Y \subseteq X \} \]
\[ \overline{RX} = \bigcup \{ y \in U / R \mid Y \cap X \neq \Phi \} \].
The lower approximation set \( RX \) is called as the positive region over \( X \), which denote as: \( POS(R)(X) \).

**Definition 3.** Let \( P \) and \( Q \) be equivalence relations over \( U \), then the set
\[ POS_P(Q) = \bigcup_{X \subseteq U / Q} PX \]
is called the \( P \)-positive region of \( Q \), which is the set of all objects of the universe \( U \) that can be properly classified to classes of \( U / Q \) employing knowledge \( U / P \).

It is well known that an information system or a decision system may usually have irrelevant or superfluous knowledge (attributes), which is inconvenient for us to get concise and meaningful decision. Since then, the reduction of attributes is demanded greatly.

In rough set theory, the reduction is defined as a minimal set of attributes that enables the same classification of attributes is demanded greatly. Usually a set of attributes may have more than one reduction. Intersection of all reductions is called the core.

**Definition 4.** Let \( P \) and \( Q \) be equivalence relations over \( U \), we say that \( r \in P \) is \( Q \)-dispensable in \( P \), if
\[ POS_P(Q) = POS_{P \rightarrow \{r\}}(Q) \]
Otherwise \( r \) is \( Q \)-independent in \( P \). If every \( r \in P \) is \( Q \)-independent, \( P \) is \( Q \)-indispensable.

The attributes subset \( R \subseteq P \) will be called a \( Q \)-reduction of \( P \), if and only if \( R \) is the \( Q \)-independent subset of \( P \) and \( POS_R(Q) = POS_P(Q) \).

Intersection of all \( Q \)-reductions of \( P \) is called the \( Q \)-core of \( P \), denoted as \( \text{Core}_Q(P) \).

In particular, the core may be empty.

### 3. Algorithm for finding core

In rough set theory, the definition of relative core and relative reduction are based on positive region. Given the information system \( S = \langle U, R, V, f \rangle \), in which \( U \) is the objects, \( R = C \cup D \) , suppose \( a_i \in C \), according to the definition 1, we can know that indiscernibility relation \( C \) can classify \( U \) into more particular than relation \( (C - \{a_i\}) \), associating with the definition 4 further, the following property of positive region is obvious.

**Property 1.** \( POS_{(C - \{a_i\})}(D) \subseteq POS_C(D) \).

**Property 2.** \( a_i \) is a attribute in \( C \), if satisfy the equation
\[ POS_{(C - \{a_i\})}(D) = POS_C(D) \]
then \( a_i \) is not the \( D \)-core attribute of \( C \).

**Proof:** According to the definition 4, we can see that if the above equation is satisfied, then \( a_i \) is \( D \)-dispensable in \( C \), in other words, it can be reduced. Obviously \( a_i \) is not the core attribute.

**Property 3.** An attribute \( a_i \) in \( C \) is a \( D \)-core attribute of \( C \), if and only if
\[ POS_{(C - \{a_i\})}(D) \neq POS_C(D) \]

**Proof:**
(\( \Leftarrow \): if \( POS_{(C - \{a_i\})}(D) \neq POS_C(D) \), according to the definition 4, \( a_i \) is \( D \)-dispensable in \( C \), can not be reduced. In order to explain that the attribute \( a_i \) is the core attribute, here we adopt the proof by contradiction. Suppose the attribute is not the core attribute, according to the definition of relative core, there is a reduction \( P \) at least, which not include the attribute \( a_i \). according to property 1 we can find that
\[ POS_P(D) \subseteq POS_{(C - \{a_i\})}(D) \subseteq POS_C(D) \]
Furthermore, according to the definition of relative reduction, we can know that there must have
\[ POS_P(D) = POS_C(D) \]
Therefore:
\[ POS_{(C - \{a_i\})}(D) = POS_C(D) \]
It is incompatible with the given condition.
(\( \Rightarrow \): if \( POS_{(C - \{a_i\})}(D) = POS_C(D) \), according to the Property 1, \( a_i \) must be not a core attribute, therefore, if \( a_i \) is a core attribute, must satisfy the condition:
\[ POS_{(C - \{a_i\})}(D) \neq POS_C(D) \]

Based on the above analysis, we propose an algorithm for relative core based on positive region directly.

**Algorithm 1.** Find relative core.

**Input:** Information system \( S = \langle U, R, V, f \rangle \), where
$R = C \cup D$ and $C \cap D = \emptyset$, $C$ and $D$ are condition attributes and decision attributes respectively.

**Output:** The set of relative core, denoted as Core.

**Step 1:** calculate $POS_C(D)$.

**Step 2:** $Core = \emptyset$.

**Step 3:** for $i=1:m$

if $POS_{C \setminus \{a_i\}}(D) \neq POS_C(D)$

then $Core = Core \cup \{a_i\}$.

end.

**Step 4:** Return the set Core.

Remark: $m$ is the count of condition attributes; $a_i$ is the $i$th attribute in $C$.

### 4. Algorithm for knowledge reduction

In the decision system, each attribute may be has unequal significance. In order to find the significance of each attribute, we often get rid of the attribute and look into the change of partition without the attribute. If there is big change, illuminate that the attribute is very important; whereas, it is less important.

Here we use the positive region as the heuristic information of the attribute tolerance of importance, where we use the value of

$$pos = POS_C(D) - POS_{C \setminus \{a_i\}}(D)$$

as the estimate criterion of the attribute tolerance of importance. In the following, we give two algorithms for reduction based on positive region. The first one is based on the core, add the more important attribute to the subset $Reduct$ step by step until the condition

$$POS_{Reduct}(D) = POS_C(D)$$

is satisfied. The other one regard the whole condition attribute set $C$ as a initial subset $Reduct$, then get rid of the dispensable attribute step by step by the heuristic information, but satisfying the equation that is just mentioned above. Thus the last subset $Reduct$ is the finding reduction.

**Algorithm 2.** For relative reduction based on Core.

**Input:** Information system $S = \langle U, R, V, f \rangle$, where $R = C \cup D$ and $C \cap D = \emptyset$, $C$ and $D$ are condition attributes and decision attributes respectively.

**Output:** the set of relative reduction, denoted as $Reduct$.

**Step 1:** calculate $POS_C(D)$.

**Step 2:** for each attribute $a_i$, calculate the value

$$pos = POS_C(D) - POS_{C \setminus \{a_i\}}(D)$$

and array each attribute $r$ by descending order in light of the above value;

**Step 3:** order $Reduct = C$; for each attribute $a_i$, execute the following operation in light of the pos by descending order:

If $POS_{Reduct \setminus \{a_i\}}(D) = POS_C(D)$, the attribute $a_i$ should be reduced, and then $Reduct = Reduct \setminus \{a_i\}$; else it can’t be reduced and the set of $Reduct$ doesn’t change.

**Step 4:** the set $Reduct$ is a seeking reduction.

As we all know, the relative core of any information, in the rough set theory, is exclusive and is included in all the reductions. Algorithm 2 use the core as the initial set, add the attribute that has the strong partition ability to the subset step by step until satisfy the condition of reduction. So that, the method can finds the optimal reduction by and large. In another way, algorithm 3 find the reduction via eliminate the small partition ability attribute step by step without losing essential information about the original data in information system. Algorithm 3 doesn’t computer the
5. Experiments

In order to verify the validity of above-mentioned algorithms, we use the example presented in reference [2], denoted as decision table 1, and the other three decision tables in UCI machine learning database[6], the experimental results is shown in the Table 1.

The experimental results show that algorithm 2 and 3 can effectively find the reduction, and can get the optimal reduction in most cases. Only for decision table 1, algorithm 3 does not find the optimal reduction, but gained the sub-optimal reduction. That we use the positive region as the heuristic information of attribute reduction can only measure the attribute tolerance of importance approximately in terms of the partition ability on equivalence relation. We can’t distinguish the important degree of each attribute strictly. Therefore, how measures the attribute tolerance of importance in the rough set should be research further.

Table 1. The experimental results

<table>
<thead>
<tr>
<th>Decision system</th>
<th>Card of $U$</th>
<th>Card of $C$</th>
<th>Algorithm 1</th>
<th>Algorithm 2</th>
<th>Algorithm 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision table 1</td>
<td>21</td>
<td>9</td>
<td>{D,I}</td>
<td>{A,D,E,I}</td>
<td>{B,D,F,G,I}</td>
</tr>
<tr>
<td>BUPA liver disorders</td>
<td>345</td>
<td>6</td>
<td>{}</td>
<td>{A,B,E}</td>
<td>{C,D,E}</td>
</tr>
<tr>
<td>Glass Identification</td>
<td>214</td>
<td>10</td>
<td>{}</td>
<td>{A}</td>
<td>{A}</td>
</tr>
<tr>
<td>Ionosphere Database</td>
<td>351</td>
<td>34</td>
<td>{}</td>
<td>{4,18,24}</td>
<td>{30,33,34}</td>
</tr>
</tbody>
</table>

6. Conclusions

The search for relative core and the minimum size of reduction is the hot field in rough set theory. There exists a variety of methods presently. Rough set theory is deeply investigated in this paper and the contribution of this article is discovering some useful properties of the positive region. In terms of the properties, this article proposes a algorithm for relative core directly and gives two algorithm for relative reduction using the information of positive region. The experimental results show that these algorithms are valid.

Because of the heuristic information is not very mature, in order to seek the optimal reduction more valid, measurement of the attribute tolerance of importance needs research further.

Acknowledgements

This work was supported by Science Foundation of Shanghai Education (03IK09) and Key Subject Construction Foundation (No.4) of Shanghai Municipal Education Commission.

References