Relation between Significance of Attribute Set and Single Attribute

Xiuqin Ma, Norrozila Binti Sulaiman, Hongwu Qin

Abstract—In the research field of Rough Set, few papers concern the significance of attribute set. However, there is important relation between the significance of single attribute and that of attribute set, which should not be ignored. In this paper, we draw conclusions by case analysis that (1) the attribute set including single attributes with high significance is certainly significant, while, (2) the attribute set which consists of single attributes with low significance possibly has high significance. We validate the conclusions on discernibility matrix and the results demonstrate the contribution of our conclusions.

Keywords—relation; attribute set; single attribute; rough set; significance

I. INTRODUCTION

THIS Rough set theory, proposed by Z. Pawlak in 1982 [1], can be considered as a new mathematical tool for dealing with uncertainties and vagueness [2]. It has been applied to many other theories. Despite this overlap, rough set theory may be considered as an independent discipline in its own right. The main advantage of rough set theory is that it does not need any preliminary or additional information about data like probability distributions in statistics, basic probability assignments in Dempster–Shafer theory, a grade of membership or the value of possibility in fuzzy set theory [3].

It is typically assumed that we have a finite set of objects described by a finite set of attributes. An information system [3] is a data table containing rows labeled by objects of interest, columns labeled by attributes and entries of the table are attribute values. Attribute values can be also numerical. In data analysis the basic problem we are interested in is to find patterns in data, i.e., to find a relationship between some set of attributes. Decision tables are one type of information tables with a decision attribute that gives the decision classes for all objects. Attribute reduction is an important problem of rough set theory [4]. The objective of reduct construction is to reduce the number of attributes, and at the same time, preserve a certain property that we want [5]. A reduct is a minimum subset of attributes that provides the same descriptive or classification ability as the entire set of attributes [1]. In other words, attributes in a reduct are jointly sufficient and individually necessary.

Conditional attributes have the different significances in decision making and data classification. Single attribute, with its significance equal to 0, is omitted in attribute reduction. It is regarded as uselessness for making decision. As a consequence, useful knowledge will be possibly omitted and then data mining is affected. Hence it is necessary for us to research relation between significance of attribute set and single attribute, especially on the attribute set consisting of low single significances.

The document [6] mentioned dependability of attribute set, which focused on the single attribute dependability in fact. The document [7] discussed algorithm of attribute set dependability related to decision attribute. Up to the present, few documents have focused on relation between significance of attribute set and single attribute. In this paper, we analyze the relation between the significance of attribute set and that of single attribute, and draw conclusions that it is not certain that attribute set which consists of single attributes with low significance is not significant, while, attribute set including single attributes with high significance has certainly high significance. As a result, the significance of attribute set is more authentic compared with the significance of single attribute.

The rest of the paper is organized as follows. Section II reviews the basic notations of the rough set theory. Section III gives the algorithm for solution to single significance and attribute set significance. Section IV provides an example. Section V gives validation. Section VI concludes the paper.

II. BASIC NOTIONS [1, 8, 9, 10, 11]

A. Definition 1 Information Systems

In the Rough Set Theory, information systems are used to represent knowledge. An information system $S = (\mathcal{U}, \mathcal{A}, \mathcal{V}, f)$ consists of: $\mathcal{U}$ - a nonempty, finite set named universe, which is a set of objects, $\mathcal{A} = \{x_1, x_2, ..., x_m\}$; $\mathcal{A}$ - a nonempty, finite set of attributes, $\mathcal{A} = \mathcal{C} \cup \mathcal{D}$, in which $\mathcal{C}$ is the set of condition attributes, and $\mathcal{D}$ is the set of decision attributes; $\mathcal{V} = \bigcup_{a \in \mathcal{A}} \mathcal{V}_a$ is the domain of a; $f : \mathcal{U} \times \mathcal{A} \rightarrow \mathcal{V}$ - an information function. For each $a \in \mathcal{A}$ and $x \in \mathcal{U}$, an information function $f(x, a) \in \mathcal{V}_a$ is defined, which means that for each object $x$ in $\mathcal{U}$, $f$ specify its attribute value.

B. Definition 2 Lower and Upper Approximation

Let $A = \langle U, R \rangle$ be an approximation space and let $X$ be any subset of $U$. The $R$-lower approximations of $X$,
denoted $R(X)$ and $R$-upper approximations of $X$, $\overline{R}(X)$ respectively, are defined by
\[
\overline{R}(X) = \bigcup \{ x \in U \mid R : [x] R \subseteq X \} \quad (1)
\]
and
\[
\overline{R}(X) = \bigcup \{ x \in U \mid R : [x] R \cap X \neq \emptyset \} \quad (2)
\]

C. \textbf{Definition 3 Dependability}

Suppose $S=(U, A, V, f)$ is a decision table. The dependability between Condition attributes $C$ and Decision attributes $D$ is defined as:
\[
k = \gamma_c(D) = \frac{\text{card}(\text{POS}_c(D))}{\text{card}(U)},
\]
where, card () represents the cardinal number of set.

D. \textbf{Definition 4 Significance of Single Attribute and Attribute Sets}

In the above decision table, significance of condition attribute subset $C' (C' \subseteq C)$ related to $D$ is defined as:
\[
\sigma_{CD}(C') = \gamma_c(D) - \gamma_{C-C'}(D). \quad (4)
\]
Especially, $C' = \{a\}$, significance of single attribute $a \in C$ related to $D$ is defined as:
\[
\sigma_{CD}(a) = \gamma_c(D) - \gamma_{C-C}(D). \quad (5)
\]

III. \textbf{ALGORITHM FOR SOLUTION TO SINGLE SIGNIFICANCE AND ATTRIBUTE SET SIGNIFICANCE}

E. \textbf{Solution to Single Significance}

Suppose that condition attribute set $C=\{C_1, C_2, \ldots, C_n\}$, decision attribute $D$, the algorithm for solution to single significance of $C_n$ as follows:
1) Get $U/\text{ind}(C)$, which denotes the family of all equivalence classes of $C$, written $U/C$ for short.
2) Get $U/D$, which denotes the equivalence classes of $D$.
3) Get $\text{pos}_c(D)$.
4) Compute $\gamma_c(D)$, which is the dependability of decision attribute $D$ for condition attribute set $C$.
5) Get $U/\{C - \{C_n\}\}$.
6) Get $\text{pos}_{C - \{C_n\}}(D)$.
7) Compute $\gamma_{C - \{C_n\}}(D)$.
8) Compute $\sigma_{CD}(C_n)$.

F. \textbf{Solution to Significance of Attribute Set}

In the above information system, condition attribute subset $C' \subseteq C$, the algorithm for solution to significance of attribute subset $C'$ is as follows:
1) Compute $\gamma_c(D)$.
2) Get $U/(C - C')$.
3) Get $\text{pos}_{C - C'}(D)$.
4) Compute $\gamma_{C-C}(D)$.
5) Compute $\sigma_{CD}(C')$.

IV. \textbf{EXAMPLE}

We construct a decision table. Let $U=\{u_1, u_2, \ldots, u_{10}\}$ be the set of objects, the condition attributes set $C=\{C_1, C_2, C_3, C_4\}$, and the decision attributes set $D=\{D\}$. They are illustrated in the TABLE I.

<table>
<thead>
<tr>
<th>Objects (u)</th>
<th>Condition attributes (C)</th>
<th>Decision Attributes (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(C_1)     (C_2)     (C_3)     (C_4)</td>
<td></td>
</tr>
<tr>
<td>(u_1)</td>
<td>1      2      1     0</td>
<td>1</td>
</tr>
<tr>
<td>(u_2)</td>
<td>0      0      2     1</td>
<td>0</td>
</tr>
<tr>
<td>(u_3)</td>
<td>0      2      1     2</td>
<td>1</td>
</tr>
<tr>
<td>(u_4)</td>
<td>0      1      0     1</td>
<td>0</td>
</tr>
<tr>
<td>(u_5)</td>
<td>1      1      2     2</td>
<td>3</td>
</tr>
<tr>
<td>(u_6)</td>
<td>1      1      1     2</td>
<td>3</td>
</tr>
<tr>
<td>(u_7)</td>
<td>1      2      0     2</td>
<td>1</td>
</tr>
<tr>
<td>(u_8)</td>
<td>1      2      0     1</td>
<td>2</td>
</tr>
<tr>
<td>(u_9)</td>
<td>0      1      2     1</td>
<td>3</td>
</tr>
<tr>
<td>(u_{10})</td>
<td>0      1      2     1</td>
<td>0</td>
</tr>
</tbody>
</table>

A. \textbf{Significance of Single Attribute}

Based on the above steps of solution to single significance, we can get the significance of every single attribute, illustrated in the TABLE II.

<table>
<thead>
<tr>
<th>Condition attributes</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>0</td>
</tr>
<tr>
<td>(C_2)</td>
<td>0</td>
</tr>
<tr>
<td>(C_3)</td>
<td>0</td>
</tr>
<tr>
<td>(C_4)</td>
<td>2/10</td>
</tr>
</tbody>
</table>

B. \textbf{Significance of Attribute Set Consisting of Two Attributes}

We get the significance of attribute set consisting of two attributes, illustrated in the TABLE III.

<table>
<thead>
<tr>
<th>Condition attribute sets</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1,C_2)</td>
<td>5/10</td>
</tr>
<tr>
<td>(C_1,C_3)</td>
<td>0</td>
</tr>
<tr>
<td>(C_1,C_4)</td>
<td>4/10</td>
</tr>
<tr>
<td>(C_2,C_3)</td>
<td>4/10</td>
</tr>
<tr>
<td>(C_2,C_4)</td>
<td>5/10</td>
</tr>
<tr>
<td>(C_3,C_4)</td>
<td>3/10</td>
</tr>
</tbody>
</table>
C. Significance of Attribute Set Consisting of Three Attributes

We get the significance of attribute set consisting of three attributes, illustrated in the TABLE IV.

<table>
<thead>
<tr>
<th>Condition attribute sets</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁,C₂,C₃</td>
<td>9/10</td>
</tr>
<tr>
<td>C₁,C₂,C₄</td>
<td>1</td>
</tr>
<tr>
<td>C₁,C₃,C₄</td>
<td>1</td>
</tr>
<tr>
<td>C₂,C₃,C₄</td>
<td>1</td>
</tr>
</tbody>
</table>

We can draw conclusions from the above example:

1) It is seen from TABLE II that significance of single attribute C₄ is the greatest, while significance of single attribute C₁, C₂, C₃ is equal to 0.

2) It is seen from TABLE III that significance of attribute set consisting of C₁ and C₂ is the greatest among sets which are composed of two attributes, though significance of single attribute C₁ and C₂ are equal to 0.

3) It is seen from TABLE III and TABLE IV that attribute sets including the greatest significance of single attribute C₄ have a high significance.

V. VALIDATION

The discernibility matrix was proposed by A. Skowron in 1991[12]. We make use of discernibility matrix to get discernibility function and then get the reduction of the decision table.

The discernibility function of TABLE I is:

\[ \delta_{\delta}(C_4, C_6, C_7) = C_4C_6C_7 \]

From the result, we can deduce that C₆, C₇ and C₄ cannot be ignored.

VI. CONCLUSIONS

In the present paper, we reach conclusions:

1) It is not certain that attribute sets which consist of low single significances are not significant.

2) Attribute sets including high single significances have certainly high significance.

Consequently, single attribute which possesses zero or low significance can not easily be discarded in the decision table. Attribute set significance is more authentic compared with single significance.

Besides the decision table constructed in the section IV, we also experimented on some other decision tables with larger amount of data and drew the same conclusion. So the conclusion can be generalized.

REFERENCES