International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:12, 2010

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 machine learning, intelligent systems, inductive reasoning, pattern recognition, expert systems, data analysis, data mining and knowledge discovery. Rough set theory overlaps with 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 in data analysis 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 $S= (U, A, V, f)$ consists of: U -a nonempty, finite set named universe, which is a set of objects, $U=\{x_1, x_2, ..., x_m\}$; A -a nonempty, finite set of attributes, $A=C\cup D$, in which C is the set of condition attributes, and D is the set of decision attributes ;$V=\bigcup_{a\in A} V_a$, V is the domain of a ; $f: U\times A\rightarrow V$ -an information function. For each $a\in A$ and $x\in U$, an information function $f(x, a) \in Va$ is defined, which means that for each object x in U, f specify its attribute value.

B. Definition 2 Lower and Upper Approximation

Let $A=(U,R)$ be an approximation space and let X be any subset of U. The R-lower approximations of X,

Xiuqin Ma is with Faculty of Computer Systems and Software Engineering Universiti Malaysia Pahang,Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia (corresponding author to provide phone: 0169605749; e-mail: xueener@yahoo.com.cn).

Norrozila Binti Sulaiman is with Faculty of Computer Systems and Software Engineering Universiti Malaysia Pahang, Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia (e-mail: norrozila@ump.edu.my).

Hongwu Qin is with Faculty of Computer Systems and Software Engineering Universiti Malaysia Pahang, Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia (e-mail: qhwump@gmail.com).

denoted $\underline{R}(X)$ and *R*-upper approximations of *X*, $\overline{R}(X)$ respectively, are defined by

$$\underline{R}(X) = \cup\{[x]_R \in U/R : [x]_R \subseteq X\} \quad (1)$$

and

$$\overline{R}(X) = \cup\{[x]_R \in U/R : [x]_R \cap X \neq \phi\} \quad (2)$$

C. *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:

$$k = \gamma_C(D) = \frac{card(POS_C(D))}{card(U)}\ , \quad (3)$$

Where, card () represents the cardinal number of set.

D. *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:

$$\sigma_{CD}(a) = \gamma_C(D) - \gamma_{C-\{a\}}(D)\,. \quad (5)$$

III. Algorithm for Solution to Single Significance and Attribute Set Significance

E. *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_m as follows:

1) Get *U/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 $pos_c(D)$.
4) Compute $\gamma_C(D)$,which is the dependability of decision attribute *D* for condition attribute set *C*.
5) Get $U/\{C\text{-}\{C_m\}\}$.
6) Get $pos_{(C\text{-}\{Cm\})}(D)$.
7) Compute $\gamma_{C\text{-}\{Cm\}}(D)$.
8) Compute $\sigma_{CD}(C_m)$.

F. *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 $U/\{C\text{-}\ C'\}$.
3) Get $pos_{(C\text{-}\ C')}(D)$.
4) Compute $\gamma_{C\text{-}\ C'}(D)$.
5) Compute $\sigma_{CD}(C')$.

IV. 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 I
A DECISION TABLE

Objects	**Condition attributes(C)**				**Decision Attributes(D)**
u	**C_1**	**C_2**	**C_3**	**C_4**	**D**
u_1	1	2	1	0	1
u_2	0	0	2	1	0
u_3	0	2	1	2	1
u_4	0	1	0	1	0
u_5	1	1	2	2	3
u_6	1	1	1	2	3
u_7	1	2	0	2	1
u_8	1	2	0	1	2
u_9	1	0	1	2	3
u_{10}	0	1	2	1	0

A. *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 II
SIGNIFICANCE OF EVERY SINGLE ATTRIBUTE

Condition attributes	**significance**
C_1	0
C_2	0
C_3	0
C_4	2/10

B. *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 III
SIGNIFICANCE OF ATTRIBUTE SET CONSISTING OF TWO ATTRIBUTES

Condition attribute sets	**significance**
C_1,C_2	5/10
C_1,C_3	0
C_2,C_3	4/10
C_1,C_4	4/10
C_2,C_4	5/10
C_3,C_4	3/10

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 IV
SIGNIFICANCE OF ATTRIBUTE SET CONSISTING OF THREE ATTRIBUTES

Condition attribute sets	**significance**
C_1,C_2,C_3	9/10
C_1,C_2,C_4	1
C_1,C_3,C_4	1
C_2,C_3,C_4	1

We can draw conclusions from the above example:

1）It is seen from TABLE II that significance of single attribute C_4 is the greatest, while significance of single attribute C_1 ,C_2, C_3 is equal to 0.

2) It is seen from TABLE III that significance of attribute set consisting of C_1 and C_2 is the greatest among sets which are composed of two attributes, though significance of single attribute C_1 and C_2 are equal to 0.

3) It is seen from TABLE III and TABLE IV that attribute sets including the greatest significance of single attribute C_4 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:

$$f_{M(S)}(C_1, C_2, C_3, C_4) = C_2C_3C_4$$

From the result, we can deduce that C_2, C_3 and C_4 can not 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

[1] Pawlak Z. Rough Sets [J]. International Journal of Computer and Information Science, 1982, 11:341 .

[2] Pawlak Z. Rough Sets and Intelligent Data Analysis[J]. Information Sciences,2002, 147(1-4): 1-12.hu

[3] Zdzislaw Pawlak, Andrzej Skowron. Rudiments of rough sets[J]. Information Sciences, 177(2007) 3-27

[4] D.Q. Miao a, Y. Zhao b, Y.Y. Yao b, H.X. Li b,c, F.F. Xu a,b. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model[J]. Information Sciences, 179 (2009) 4140–4150.

[5] Yiyu Yao, Yan Zhao, Attribute reduction in decision-theoretic rough set models. Information Sciences 178 (2008) 3356–3373

[6] Zhu Hong. Research of Representation Formula for the Dependence Degree Among Attributes [J]. COMPUTER ENGINEERING, 2005, 31(1): 174-175, 211.

[7] Meng Qingquan，Mei canhua. New dependability of attribute sets [J]. JOURANL OF COMPUTER APPLICATIONS，2007,27(7):1748-1750

[8] Zhang Wenxiu, Wu Weizhi, Liang Jijie, Li Deyu. Rough set theory and method[M]. Bei Jing: Science express.2001:1.

[9] Jia wei Han, Micheline Kamber, Data Mining Concepts and Techniques, Publishing House of Mechanical Industry,2001.8.

[10] LEUNG Y, WU W Z, ZHANG W X. Knowledge acquisition in incomplete information systems: a rough set approach [J]. European Joumal of Operational Research. 2006, 168(1): 164-180.

[11] Jiang Yun; Li Zhanhuai; Wang Yong; Zhang Longbo, A Better Classifier Based on Rough Set and Neural Network for Medical Images. Data Mining Workshops, 2006. ICDM Workshops 2006. Page(s):853 – 857.

[12] A.Skowron. Rough Sets in KDD. Special Invited Speaking, WCC 2000 in Beijing, Aug.2000.