*Available online www.jocpr.com*

**Journal of Chemical and Pharmaceutical Research, 2014, 6(7):310-315**

**Research Article**

**ISSN : 0975-7384**
**CODEN(USA) : JCPRC5**

# An improved attribute reduction algorithm based on mutual information with rough sets theory

**Yang Su-min[1,2], Meng Jie[1], Liu Qi-ming[2] and Chen Jing[2]**

[1]*State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, Henan, China*
[2]*Department of Information Engineering, Ordnance Engineering College, Shijiazhuang, China*

---

**ABSTRACT**

*There aren't core attributes in some information system, but the core attributes are the basis of attribute reduction algorithm based on mutual information, in order to solve the problem of a new attribute importance degree one new method on the basis of mutual information is proposed in the paper, which consists of its own information entropy and mutual information. Then the corresponding heuristic reduction algorithm is proposed. Experimental results show that the algorithm can solve non-core information system attribute reduction, but also can get attribute reduction faster, and the reduction number is also relatively small.*

**Key words:** Rough set; attribute reduction; mutual information; attribute importance degree

---

## INTRODUCTION

At present there are various kinds of evaluation attributes in the command and information system, to ensure efficiency and effectiveness of evaluation it is essential to constitute the most concise attribute system. Attribute reduction is to remove unnecessary attribute without changing classification result of information system. Rough sets theory is a valid mathematical theory developed in recent years, which can analyze and deal with imprecise, incomplete and inconsistent information effectively, and can dig out the connotative knowledge, and reveal potential rules. By rough sets theory, we can usually obtain a few reduction results for an information system, so we always hope to find the minimal reduction. The core and reduction of attributes are two important topics in the research on rough sets theory, but researchers have proven that it is NP-hard problem to look for all reduction or minimum reduction of an information system. But the researchers have found that an efficient attribute reduction algorithm can be obtained on the condition that the relationship between knowledge and information system established with information entropy. Skowron[1] put forward one attribute reduction algorithm based on discernibility matrices, Articles [2-6] present some improved attribute reduction algorithm based on discernibility matrices, which have lower computational complexity and storing capacity. On the basis of conditional information entropy, articles [7-8] study the computation of a core and attribute reduction in distributed environment. Qian[9] analyzed the relationship between attribute reduction and conditional information quantity and gave one new conditional information quantity which cut down the number of attributes and time complexity. Teng[10] presented a new reduced definition which integrates the complete and incomplete information systems into the corresponding reduced algorithm. Liang etc.[11-13] studied the incomplete information systems. Miao [14] proposed the knowledge reduction algorithm which is based on the mutual information between the conditional attributes and decision attributes. Jia [15] proposed one attribute reduction algorithm based on mutual information gain. Articles [16-20] proposed rough sets attribute reduction algorithm based on mutual information, which can make use of heuristic information to reduce the search space, and can shorten the search time as far as possible, and can finally get an optimal or approximate optimal solution. But the attribute reduction algorithm based on mutual information is the bottom-up approach, whose starting point is from the relative core attribute of decision table, then the most important attributes selected from the other attributes are added to the relative core, and the computing processing is ended when the core mutual information is equal with the conditional attribute. In actual information system, there will be a lack of core attributes. When the core attributes are empty, we must calculate mutual information after choosing one attribute, so the computational complexity increases significantly.

In this paper, one new attribute importance degree method is proposed, which depends on its own information entropy and mutual information，and the author gives the corresponding heuristic reduction algorithm. The experimental results show that the proposed algorithm not only can solve the problem of attribute reduction for the non-core information system, but also can obtain the attribute reduction results faster, and the number of reduction attribute is relatively small.

**THE BASIC CONCEPTS OF ROUGH SETS THEORY**

**Definition 1:** $S=(U,A,V,f)$ is set to an information system. Among them, $U=\{U_1,U_2,\cdots,U_{[U]}\}$ is non empty finite sets which is called the domain space, $A=\{a_1,a_2,\cdots,a_{[A]}\}$ is non empty finite attribute set, which is called the attribute set, $V=\cup A_a$ , $a\in A$ , $V_a$ is attribute's domain range, $f:U\times A\to V_a$ is the information function. When $x$ is $a$ , $x$ has unique value in $V_a$ .On the side, for sequence $C(c_1(x),c_2(x),\cdots,c_n(x)$ and sequence $D(d_1(x),d_2(x),\cdots,d_n(x)$ , $A=C\cup D, C\cap D=\phi$ , $S=(U,A,V,f)$ is called as decision table of the information system. $c_1(x),c_2(x),\cdots,c_n(x)$ is called as the condition attribute set.

**Definition 2:** For the given knowledge representation system $S=(U,A,V,f)$ , the in-discernable relationship of any attribute is as follows:

$$IND(B)=:\{(x,y)\in U\times U:\forall a\in B(f(x,a)=f(y,a)\}\qquad(1)$$

**Definition 3:** For the given the knowledge representation system $S=(U,A,V,f)$ , $P\subseteq A, X\subseteq U, x\in U$, the lower and upper approximation set for $X$ with regard to $IND(B)$ is as below respectively:

$$\underline{R(X)}=\cup\{x\in U:IND(C)\subseteq X\};\qquad(2)$$

$$\overline{R(X)}=\cup\{x\in U:IND(C)\cap X\neq\phi\}\qquad(3)$$

**Definition 4:** For the given knowledge representation system $S=(U,A,V,f)$ , if $P,Q\subseteq A$ , the positive domain $POS_p(Q)$ is defined as:

$$POS_p(Q)=\bigcup_{X\subseteq U/P}\underline{R(X)}$$

Among them, $\underline{R(X)}$ is the lower approximation of $X$ .

**Definition 5:** $U$ is a domain set, $P$ and $U$ is two equivalent relation of domain $U$ (knowledge), $U/ind(P)=\{x_1,x_2,\cdots,x_n\}$ , $U/ind(Q)=\{y_1,y_2,\cdots,y_n\}$ , then the probability distribution that $P$ and $Q$ effect on the U is defined as follows:

$$[X:p]=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\\ p(x_1) & p(x_2) & \cdots & p(x_n)\end{bmatrix};[Y:p]=\begin{bmatrix}y_1 & y_2 & \cdots & y_m\\ p(y_1) & p(y_2) & \cdots & p(y_m)\end{bmatrix}\qquad(4)$$

Among them, $p(x_i)=\frac{|x_i|}{U}, i=1,2,\cdots,n;;$ $p(y_i)=\frac{|y_j|}{U}, j=1,2,\cdots,m$ , the symbol $|E|$ is the base of $E$ .

**Definition 6:** According to the information theory, the information entropy of knowledge $P$ is

$H(P) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$, the conditional entropy $H(Q|P)$ of the knowledge $P$ relative to $Q$ is :

$$H(Q|P) = -\sum_{i=1}^{n} p(x_i)\sum_{j=1}^{m} p(y_j|x_i)\log p(y_j|x_i) \qquad (5)$$

The mutual information $I(P;Q)$ of the knowledge $P$ relative to $Q$ is:

$$I(P;Q) = H(Q) - H(Q|P); \qquad (6)$$

**THE ATTRIBUTE IMPORTANCE BASED ON THE MUTUAL INFORMATION**

In the process of decision, we pay attention which condition attribute is the most important for the last decision, so we must consider the mutual information between condition attribute and decision attribute. The article [17] proposed the method that obtained the attribute importance by the increasing amount of mutual information with adding one attribute. It is defined as follows:

$$SGF(a,R,Q) = I(Q;R\cup\{a\}) - I(Q;R) = H(Q|R) - H(Q|R\cup\{a\}) \qquad (7)$$

Based on the above formula, the chosen attributes are that there is more number in the domain, but from the information theory, it is to select the one which is chaotic, but the selected attributes are not maybe useful for the decision.

In view of the above problems, the article [19] has made the improvement to the importance of attributes, which is defined as follows:

$$SGF_{old}(a,R,Q) = I(Q|R\cup\{a\}) - I(Q|R)/H(Q|a) = (H(Q|R) - H(Q|R\cup\{a\}))/H(Q|a) \qquad (8)$$

But the above calculation processing depends on the core attributes, in order to overcome no core problem, the author improves the formula (8), the result is as follows:

$$SGF_{new}((a,R,Q)) = (I(C-\{c\};D) - I(C;D) + I(c;D))/H(D|c) = (H(D|C) - H(D|C-\{c\}) + H(D) - H(D|c))/H(D|c)$$
$$= (H(D|C) - H(D|C-\{c\}) + H(D))/H(D|c) - 1 \qquad (9)$$

The improved method not only considers the increment of mutual information after adding the attribute, but also considers its own information entropy. When the mutual information increment is equal, the smaller $H(Q|a)$ is, the higher attribute importance degree is. The equation (9) can be in agreement with the actual situation, and also solve the attribute reduction with no nuclear attribute information system.

**AN IMPROVED ATTRIBUTE REDUCTION ALGORITHM BASED ON MUTUAL INFORMATION**

On the basis of the formula (9) in the section 2, in this paper the author proposes a new attribute reduction algorithm, which does not need to calculate core attribute, whether one attribute is added into attribute reduction set or not is decided by the increment between the mutual information and conditional entropy. The specific algorithm description is as follows:

**The Input:** A compatible decision table system, $C$ is condition attributes set, $D$ is the decision attribute, $U$ is the domain.

**The Output:** One reduction attribute set;

(1) The mutual information is calculated between condition attribute and decision attribute set;

(2)let $R = \phi$, the proceed is performed on the attribute set $R' = C - R, C' = C - R$ as follows:

①For each attribute $c_i \in C'$, we calculate $I(c_i, D)/H(D, c_i)$, and select the one that has maximum value $a$, if there is the same value for multiple attributes, we choose one which comes the earliest , then $R = R \cup \{a\}$

②then we judge whether $I(C;D)$ and $I(R;D)$ is equal, if they are the same, then the next step goes to

(3), otherwise goes to ①.
(3) $R$ is a reduction result, and we output it.

**THE SIMULATION EXPERIMENT**
In order to verify the effectiveness of the above algorithm, we take it to compute the reduction set of a command and information system, as shown in table 1. From it, we can see that the system has 4 attributes, 14 experts give those evaluation results, the condition attributes are $\{c_1, c_2, c_3, c_4\}$, decision attributes are $\{d\}$, the value of set $C$ is set as V={0,1,2}, which corresponds good, general, poor state respectively. The value of set $D$ is set as {0,1}, which corresponds good and bad for operation effect, after pre-treatment on the original data, we get the decision table shown in table1.

The below is the processing in accordance with the algorithm of section 3:

We consider table 1 as an information system

$U = \{x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14\}$,condition attribute set is
$C = \{c_1, c_2, c_3, c_4\}$, decision attribute set is $D = \{d\}$, then
$IND(D) = \{\{x1, x2, x6, x8, x14\}, \{x3, x4, x5, x7, x9, x10, x11, x12, x13\}$

**TABLE 1: Decision table of information system**

| samples | Communication quality $c_1$ | System exchange quality $c_2$ | Safety measure $c_3$ | Personnel diathesis $c_4$ | decision result $d$ |
|---|---|---|---|---|---|
| x1 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 0 | 0 | 1 | 0 |
| x3 | 1 | 0 | 0 | 0 | 1 |
| x4 | 2 | 1 | 0 | 0 | 1 |
| x5 | 2 | 2 | 1 | 0 | 1 |
| x6 | 2 | 2 | 1 | 1 | 0 |
| x7 | 1 | 2 | 1 | 1 | 1 |
| x8 | 0 | 1 | 0 | 0 | 0 |
| x9 | 0 | 2 | 1 | 0 | 1 |
| x10 | 2 | 1 | 1 | 0 | 1 |
| x11 | 0 | 1 | 1 | 1 | 1 |
| x12 | 1 | 1 | 0 | 1 | 1 |
| x13 | 1 | 0 | 1 | 0 | 1 |
| x14 | 2 | 1 | 0 | 1 | 0 |

According to the description the algorithm of section 3, the steps are as follows:
(1) Firstly we calculate the mutual information between set $C$ and the set $D$;
(2) We calculate the importance degree of each attribute according to the formula (9), the table 2 lists the result, from which we can see that the attribute $c_2$ has maximum value, so it was chosen as the reduction elements, $R = \{c_2\}, C^{'} = C - R$.
(3) Then we calculate $I(R; D) = 0.261$, we can see $I(R; D) \neq I(C; D)$
(4) According to the algorithm of section 3, we need to select another attribute to join reduction set from the remaining attributes, from table 2 we can find that the $c_1$ attribute has the most importance degree among them, so we add it into the reduction set, $R = \{c_2, c_1\}$.
(5) Then we calculate $I(R; D) = 0.4605$,but $I(R; D) \neq I(C; D)$;
(6) Then we select another attribute from table 1 to join $R$, according to the important degree, $c_3$ are selected, because the equivalence classes of $\{c_1, c_2, c_4\}$ is
$\{\{x_1,\}, \{x_2\}, \{x_3, x_{13}\}, \{x_4, x_{10}\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}, \{x_9\}, \{x_{11}\}, \{x_{12}\}, \{x_{14}\}\}$ and the equivalence classes of $\{c_1, c_3, c_4\}$ is
$\{\{x_1, x_8\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5, x_{10}\}, \{x_6\}, \{x_7\}, \{x_9\}, \{x_{11}\}, \{x_{12}\}, \{x_{13}\}, \{x_{14}\}\}$ .Because their kind

and number of attribute combinations is the same. At present $c_2$ is already in the reduction set, therefore we select $c_4$ to join the R set without $c_3$. We get the result as follows:

$$R = \{c_2, c_1, c_4\}, I(R; D) = 0.9403 .$$

(7) Due to $I(R; D) = I(C; D)$, so we terminate the processing of the algorithm. $R = \{c_2, c_1, c4\}$ is one reduction set of the original information system.

(8)

**TABLE 2 Attribute important degree**

| Attribute | $H(D)$ | $H(D\|\{C-c\})$ | $H(D\|c)$ | $SGF_{old}(c)$ | $SGF'_{new}(c)$ |
|---|---|---|---|---|---|
| $c_1$ | 0.9403 | 0.5714 | 0.6935 | 0.5714 | 1.1798 |
| $c_2$ | 0.9403 | 0 | 0.6793 | 0 | 0.3842 |
| $c_3$ | 0.9403 | 0 | 0.7885 | 0 | 0.1925 |
| $c_4$ | 0.9403 | 0.2857 | 0.8922 | 0.2857 | 0.3741 |

From table 2, we can see that the attribute importance degree of $c_1, c_2, c_3, c_4$ obtained by formula (8) were 0.5714,0,0, 0.2857 respectively, which is inconsistent with the actual situation. But the attribute importance degree of $c_1, c_2, c_3, c_4$ is obtained according to the algorithm proposed in this paper is 1.1798, 0.3842, 0.1925, 0.3741 respectively, the attribute importance degree of $c_2$ is 0.3842 which is consistent with the actual system.

## CONCLUSION

Because some information systems may have no core attributes, but the core attributes is the foundation of the present attribute reduction algorithm based on mutual information, in order to solve the problem the author puts forward a new method to measure the importance degree of attribute and construct the corresponding heuristic reduction algorithm. This proposed algorithm takes into account the increment of mutual information after adding a attribute, but also its own information entropy, which can significantly decrease the ratio that the important attribute is taken as redundant attribute to remove. The experimental results show that the algorithm can not only solve the attribute reduction of non-core information system, but also be able to get reduction attribute faster and the reduction number is less than the present algorithms.

**Acknowledgments**

This research is supported by State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System under Grants No. CEMEE2014K0368B.

## REFERENCES

[1] Skowron A, Rauszer C. The discernibility matrices and functions in information systems[C]. In: Slowinski R(Eds.): Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory, *Kluwer Academic Publishers*. London, pp: 331-362,**1992.**

[2] Zhang Teng-fei, XIAO Jian-mei,WANG Xi-huai. *Acta Electronica Sinica*, Vol.11, no. 33, pp:2080-2083, **2005**.

[3] YE Dong yi. *Acta Electronica Sinica*, Vol. 28, no. 12,pp:80-82,**2000**.

[4] Cheng Jing, Zhu Jing, Zhang Fan. *Journal of Hunan University (Natural Sciences)*, Vol.36, no.4, pp:86-88,**2009**.

[5] Jiang Y, Wang X, YE Z. *Journal of System Simulation*, Vol. 20, no. 14, pp : 3718- 3721, **2008**.

[6] Chen Z, Zhang Q H. Study on attribute reduction algorithm based on fuzzy rough set[D]. *Huaqiao University College of Computer Science and Technology*,**2006**.

[7] Yang ming, Yang Ping. *Control and Decision*, Vol. 23, no.10, pp: 1103-1107, **2008**.

[8] Zhang WeiXu, Zhangyan, Wang Xiaoyu. *Computer Applications and Software*, Vol30, no. 7, pp:43-46,**2013**.

[9] QIAN Jin , YE Feiyue , MENG Xiang Ping, LIU Da You. *Engineering and Electronics*, Vol.12, no. 29, pp:2154-2157,**2007**.

[10] TENG Shu hua, ZHOU Shi lin, SUN Ji xiang, Li Zhi yong. *Journal Of National University Of Defense Technology*, Vol.32, no. 1, pp:90-94,**2010**.

[11] Liang J Y, Xu Z B. *International Journal of Uncertainty, Fuzziness and Knowledge based Systems*, , Vol.10, no. 1, pp: 95- 103,**2002**.

[12] Yee L, Ma J.M, Zhang W.X, Li T.J. *International Journal of Approximate Reasoning*, Vol.3,no.1, pp:623-630,

**2008.**
[13] Pei X B, Wang Y Z. Novel approaches to knowledge reduction in inconsistent decision systems, Natural Language Processing and knowledge engineering. *IEEE NLP-KE'05 Proceedings of 2005 IEEE Internatonal Conference*, pp:746-749, January ,**2005**.
[14] Miao Duoqian,Fan Shidong. *System engineering theory & practice*, pp:48- 56, January ,**2002**.
[15] Jia Ping, Dai Jianhua, Pan Yunhe, et al . *Journal of ZheJiang University ( Engineering Science)* , Vol.40,no.6, pp: 1041 – 1044, **2006**.
[16] Wang guoKuang. Rough theory and knowledge acquisition[M] . *Xi'an jiao tong university press*, **2001**.
[17] Wang Hongkai, Yao Bingxue,Hu haiqing. The method of ascertaining weight based on rough sets theory [J].*Computer engineering and application*, Vol.36, pp: 20- 21,**2003**.
[18] Tan zongfeng, Xu zhangyan, Wangshuai. *Computer engineering and application*, Vol.48, no.18, pp: 115-118,**2008**.
[19] Yanyan,Yang huizhong. *tsinghua science and technology*, Vol.47,no.S2, pp:1903-1906, **2007**.
[20] Bao xinzhong,Liu cheng. *Journal of management*, Vol. 6,no.6,pp:729-732,**2009**.
[21] Sun limin,Jin xiangju. *Journal of Guangdong University of Petrochemical technology,* Vol. 22, no.6, pp:66-68,**2009**.