On New Concept in Computation of Reduct in Rough Sets Theory

Faizah Shaari[1,*], Azuraliza Abu Bakar[2], and Abd Razak Hamdan[2]

[1] Polytechnic Sultan Salahuddin Abdul Aziz Shah, Persiaran Usahawan, Seksyen U1, 40150 Shah Alam, Selangor DE., Malaysia
sfaizah@psa.edu.my
[2] Center of Artificial Intelligence Technology, Faculty of Technology and Information Science, University of National Malaysia (UKM), 43600 Bangi Selangor DE, Malaysia
{aab,arh}@ukm.my

Abstract. A new concept of Reduct computation is proposed. This set is referred as *NewReduct*. It is discovered by defining the Indiscernibility matrix modulo (iDMM *D*) and Indiscernibility function modulo(iDFM *D*). Reduct is known as interesting and important set of attributes that able to represent the IS, in adverse the *NewReduct* is set of superfluous, redundant and non-interesting attributes. The computation of attributes defines sets of dispensable attributes and the partitioning of the objects based on indiscernibility relations shaped the information of a new knowledge. It is assumed that the sets of *NewReduct* attributes are able to uncover hidden knowledge that lies under a hidden pattern. One important knowledge that may be discovered from the IS or DS is the outliers knowledge.

Keywords: *NewReduct*, non-interesting, redundant, not-important, dispensable, rare.

1 Rough Sets Theory (RST)

RST is a mathematical tool introduced by Pawlak [1] as early as in the 1980's. It concerns with the analysis and modeling of classification and decision problems involving with vagueness, imprecise and uncertain or incomplete information. RST invokes the concept of approximation reasoning, hence has fundamental importance to Artificial Intelligence and Cognitive Sciences especially in areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from database, expert system, decision support systems, inductive reasoning and pattern recognition. Among many important applications found to be developed effectively applying RST are in the field of medicine, pharmacology, business analysis, banking, meteorology and security systems.

Rough sets offers two different kinds of knowledge representations called information system and decision system. An information system(IS) is the most basic kind

* Corresponding Author.

P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 136–143, 2009.

of knowledge that consist of a set of objects where each object is a collection of attributes values. An IS is an ordered pair $\mathcal{A}=(U, A)$, where U is a nonempty finite set call universe and A is nonempty finite set of attributes. Each attribute $a \in A$ is a total function $a : U \rightarrow V_a$ where V_a is the set of values of a, called range of a. The element in universe is referred as *objects*. A DS, $\mathcal{A}=(U, A, \{d\})$ is an IS for which the attributes are separated into disjoint sets of condition attributes A and decision attributes $d(A \cap \{ d \} = \phi)$.

Let $\mathcal{A}=(U, A)$ be an IS and every subset of attributes $B \subseteq A$ defines an equivalence relation, called an indiscernibility relation $IND_{\mathrm{A}}(B)$, defined as $IND_{\mathrm{A}}(B) = \{(x_i, x_j) \in U^2 \mid \forall a \in B, (a(x_i) = a(x_j))\}$. Hence, objects that are indiscernible to attributes of $B \subseteq A$ that is defined as $[x]_B = \{ y \in U \mid (x, y) \in IND(B)\}$ is referred as an equivalence class. In a decision system, the properties of the decision classes are the particular interest. Therefore, it is not necessary to be able to distinguish between classes that are mapped into the same decision class X_i.

Given a DS, $\mathcal{A}=(U, A, \{d\})$ and a subset of attributes $B \subseteq A$. The discernibility matrix modulo D (DMM D) of $\mathcal{A}$, M_B^d, is defined as follows where $m_B^d(i, j)$ is the set of attributes that discerns between objects x_i and x_j and also discerns the decision attributes δ where $\delta(x_i) \neq \delta(x_j)$ where $1 < i, j < n = | U / IND(B) |$, defined as

$m_B^d(i, j) = \{ a \in B : a(xi) \neq a(xj) \}$.

In DS, the discernibility function matrix modulo f (DFM D) can be calculated from the discernibility matrix modulo, M_B^d. The discernibility functions consist of *object related* and *full* discernibility functions. Let $\mathcal{A} =(U, A, \{d\})$ be an IS, $B \subseteq A$ a set of attributes, and M_B^d, the discernibility matrix modulo decision for that DS. For an object $x_i \in U$, the *object related* discernibility function modulo decision f is defined as $f_B^d(x_i) = \wedge\vee_{x_j \in U} m_B^d(i, j)$, where $m_B^d(i, j)$ is the element matrix of row i and column j [2, 3, 4, 5]. The full discernibility function modulo decision of $\mathcal{A}$, using attributes in B is the conjunctions of $f_B^d(x)$ defined as $g_B^d(A) = \wedge_{x \in U} f_B^d(x)$ [5]. *Reduct* is an important part of an IS which can discern all objects that are discernible by the original IS. The core contains the attributes that are dispensable to the discrimination of objects that is the attributes that are contained in all *Reducts*. Given $\mathcal{A}=(U, A)$ be an IS, let $B \subseteq A$, an attribute a is said to be *dispensable* in $B \subseteq A$ if $IND(B) = IND(B - \{ a \})$ otherwise the attribute is indispensable in B. A *Reduct* of B is a set of attributes $B' \subseteq B$ such that all attributes $a \in B - B'$ are dispensable and $IND(B) = IND(B')$ [2,3,6].

2 The New Concept in Computation of Reduct

Reduct is determined from the set of prime implicants of the discernibility function. *Reduct* does not contain redundant attributes but is usually interesting and important attributes. The computation of *Reduct*, can be used to represent Information System(IS) or Decision System(DS).

In this research, the set of attributes which is referred as superflous and redundant is of prime interest. Although this set of attributes is referred as not interesting, it is presumed that these attributes are important and able to discover hidden knowledge in datasets. In the following sub-sections, the new concept of *Reduct* computation is introduced by defining new definitions and the process during the computation of new *Reduct* is demonstrated using an example of a DS. The results obtained are recorded.

The new concept of *Reduct* computation introduced in this section is originated from the current concept of *Reduct* computation as found in [2,3,4]. Let refer the new set of *Reduct* as set of *NewReduct*.

As mentioned in section 2, the discernibility function *f* which determined *Reduct* is computed from the process of DMM *D* and DFM *D*. In similar, the *NewReduct* can be computed using a new formulation of iDMM *D* and iDFM *D*. The following subsection 3.1 explains the new creation of the Indiscernibility matrix modulo decision (iDMM *D*), and subsection 3.2 and 3.3 describe on the Indiscernibility function matrix modulo(iDFM) and *NewReduct* respectively.

2.1 Indiscernibility Matrix Modulo Decision (iDMM *D*)

In RST, a DS is similar to an IS, but a distinction is made between the condition and the decision attributes. In an IS, the information is not interpreted but in a DS, each object of the domain is assigned with a value of an expert classification attribute. A simple DS with distribution of equivalence classes is as shown in Table 1 [4].

Table 1. Example of Equivalence Class of a Decision System(DS)

Class	a	b	c	Decision	Num of Objects
E1	1	2	3	1	50
E2	1	2	1	2	5
E3	2	2	3	2	30
E4	2	3	3	2	10
E5,1	3	5	1	3	4
E5,2	3	5	1	4	1

In decision system, the properties of the decision classes are the particular interest. Therefore, to compute *NewReduct*, the calculation of iDMM *D* is to find a set of attributes from every pair of equivalence classes which are indiscern in attribute values as well as indiscern in the decision attributes from the matrix. These attributes are represented in Conjunctive Normal Form(CNF). In obtaining the set of attributes as mentioned above, iDMM *D* is hereby defined as below :

Definition 1

Given a DS, $\mathcal{A}$=(*U, A,* {*d*}) and a subset of attributes $B \subseteq A$. An indiscernibility relation by attribute *B*, *IND(B)* allow objects to be classified into set of equivalence classes, where $n = |U/IND(B)|$. The *indiscernibility matrix modulo D* of $\mathcal{A}$, M'^{d}_{B}, is defined as follows where $m'^{d}_{B}(i, j)$ is the set of attributes that indiscerns between objects x_i and x_j and also indiscerns the decision attributes δ where $\delta(x_i) = \delta(x_j)$ where $1 < i, j < n = |U / IND(B)|$ as shown in Eq.(1) below.

$$m'^{d}_{B}(i, j) = \{ a \in B : a(xi) = a(xj) \} \tag{1}$$

Table 2 below illustrates the iDMM *D* from a decision system $\mathcal{A}$. The simplification of the disjunction and conjunction of the matrix gives the Indiscernibility function modulo *D* (iDFM), f' as shown in the rightmost column in the table. In next section 3.2, the new iDFM *D* is described.

Table 2. Indiscernibility Matrix Modulo(iDMM) from Decision System, $\mathcal{A}$

	E1	E2	E3	E4	E5	*f'*
E1	{}	{}	{}	{}	{}	-
E2	{}	{}	{b}	{}	{}	b
E3	{}	{b}	{}	{a,c}	{}	(a,b)(a,c)
E4	{}	{}	{a,c}	{}	{}	(a,c)
E5	{}	{}	{}	{}	{}	-

2.2 Indiscerniblity Function Modulo *D*(iDFM *D*)

In analogy to the definition of DFM *D* as in the section 2, iDFM *D* is defined as in Definition 2 below:

Definition 2

Let $\mathcal{A}$=(*U*, *A*, {*d*}) be an IS, $B \subseteq A$ a set of attributes, M'^{d}_{B} is the indiscernibility matrix modulo decision for that DS. For an object $x_i \in U$, the *object related* indiscernibility function modulo decision *f'* is defined as in Eq(2) below:

$$f'^{d}_{B}(x_i) = \underset{x_j \in U}{\wedge\vee} m'^{d}_{B}(i, j) \tag{2}$$

where $m'^{d}_{B}(i,j)$ is the element matrix of row *i* and column *j*. The full indiscernibility function modulo decision of $\mathcal{A}$, using attributes in *B* is the conjunctions of $f'^{d}_{B}(x)$ defined in Eq(3) below:

$$g'^{d}_{B}(A) = \bigwedge_{x \in U} f'^{d}_{B}(x) \tag{3}$$

In next section 3.3, the definition of *NewReduct* is discussed.

2.3 Definition of *NewReduct*

Reduct is used, which is defined as follows: Given $\mathcal{A} = (U,A)$, let $B \subseteq A$, a Reduct of *B* is a set of attributes $B' \subseteq B$ such that all attributes $a \in B-B'$ are dispensable, and $IND(B') = IND(B)$. The set of *Reduct* of *B* is denoted *Red(B)*. [2,3, 6]. Let us define the very intuitive definition of *NewReduct* as follows:

Definition 3

(*NewReduct*). Given $\mathcal{A} = (U,A)$, let $B \subseteq A$, let *Reduct* of *B* is a set of attributes $B' \subseteq B$ such that all attributes $a \in B-B'$ are dispensable, and $IND(B') = IND(B)$. A *NewReduct* of *B* is defined as, there exist a set of attributes $B-B' \subseteq B$, such that all attributes $a \in B'$ are indispensable, and $IND(B-B') \neq IND(B)$. The set of *NewReduct* of *B* is denoted *NewRed(B)*.

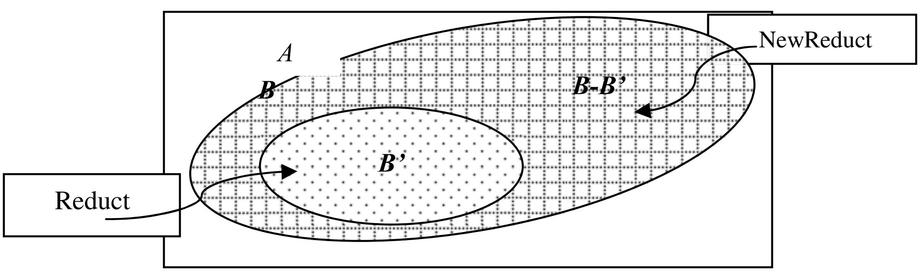

Fig. 1. illustrates the *Reduct* and *New Reduct* of a DS

The intuition behind the notion of *NewRed(B)* is that the computation of attributes *B-B'* defines sets of dispensable attributes and the partitioning of the objects based on indiscernibility relations shaped the information of a new knowledge. It is assumed that the sets of *NewReduct* attributes are able to uncover hidden knowledge that lies under a hidden pattern. One important knowledge that may be discovered from the IS or DS is the outliers knowledge.

2.4 Observation: *Reducts* vs. *NewReducts*

Reduct computation as defined by Bakar[5]; Mollestad & Komoroswki[3]; Polkowski[7] can be translated to computing the prime implicants of a boolean function. Table 3 depicts the computation of *Reduct* and *NewReduct* based on a DS. The first column in the table lists the equivalence classes from E1 to E5 of five different classes, each of which contains a number of objects from universe that are indiscernible by attributes *a* through *c*. A set of *Reducts* is as shown in the third leftmost column in the table. Each of the *Reduct* is the prime implicants (*f*) of the CNF as shown in the second leftmost column in the table.

Correspondingly, *NewReduct* can be translated to computing the prime implicants of a boolean function. Hence, a prime implicant which is determined from Boolean function is the iDFM *D* defined from Definition 2 which is obtained from the iDMM *D* as defined in Definition 1. The rightmost column in the table shows the *NewReducts*. Each of the *NewReduct* is the prime implicants (*f'*) of the CNF as shown in the second rightmost column in the table.

Table 3. *Reducts* and *NewReducts* of a DS

Equivalence Class	Prime Implicants (f)	*Reducts*	Prime Implicants (f')	*NewReducts*
E1	$c \wedge a$	{a,c}	-	-
E2	$c\ (a \vee b)$	{a,c}{b,c}	b	{b}
E3	a	{a}	$b\ (a \vee c)$	{a,b}{b ,c}
E4	$a \vee b$	**{a}**{b}	$a \vee c$	**{a}**{c}
E5	$a \vee b$	{a}{b}	-	-

The following propositions describe the properties of *NewRed(B).*

Proposition 3.1. Given A = *(U,A), let* $B \subseteq A$, The set of dispensable attributes *B-B'* from *NewReduct of B* union with set of indispensable attributes *B'* from *Reduct of B*, hence both sets of attributes belong to set of *B*. Then,

$$Red(B) \cup NewRed(B) \subseteq B$$

Proof

Since $B' \subseteq B$, $B\text{-}B' \subseteq B$ thus $B' \cup (B\text{-}B') \subseteq B$. Therefore, $Red(B) \cup NewRed(B) \subseteq B$.

Proposition 3.2. Given A = *(U,A), let* $B \subseteq A$, assuming $B^* = B\text{-}B'$, the *NewReduct* of *B* defines that if there exist intersection *B'* and *B** but the intersection is not inclusion of *B'* or *B**, hence the intersection between *Reduct(B)* and *NewReduct(B)* contains zero set. Below is the second property of *NewReduct of B.*

$$Red(B) \cap NewRed(B) = \phi \quad \textit{if} \ (Red(B) \cap NewRed(B) \not\subset B') \ or \ (Red(B) \cap NewRed(B) \not\subset B^*)$$

Proof

For $Red(B) \cap NewRed(B) \not\subset B'$:

Supposed $\forall x$, *if* $x \in B' \cap B^*$ *then* $x \in B^*$;

In a simple statement it can be explained that; supposed x is an element in $B' \cap B^*$, where x is in B' and B^* but x is not of inclusion in B'; *then,* in particular x is in B^*.

For $Red(B) \cap NewRed(B) \not\subset B^*$:

$$\text{Supposed } \forall x, \text{ if } x \in B' \cap B^* \text{ then } x \in B' ;$$

In a simple statement it can be explained that; supposed x is an element in $B' \cap B^*$, where x is in B' and B^* but x is not of inclusion in B^*; *then,* in particular x is in B'.

Thus, the exclusive of x either in B' or B^* shows that the intersection of B' and B^* contains zero set.

Therefore: $Red(B) \cap NewRed(B) = \phi$

Proposition 3.3. Given $A = (U,A)$, let $B \subseteq A$, assuming $B^* = B\text{-}B'$, In special and small cases, there exist set of intersection between *Reduct of B* and *NewReduct of B* that is inclusion of B' and B^*, resulting the intersection between *Reduct(B)* and *Non-Reduct(B)* not equal zero set.

Then,

$$Red(B) \cap NewRed(B) \neq \phi \text{ if } (Red(B) \cap NewRed(B) \subseteq B') \text{ and } (Red(B) \cap NewRed(B) \subseteq B^*)$$

Proof

For $Red(B) \cap NewRed(B) \subseteq B'$:

$$\text{Supposed } \forall x, \text{ if } x \in B' \cap B^* \text{ then } x \in B' ;$$

In a simple statement it can be explained that; supposed x is an element in $B' \cap B^*$, where x is in B' and B^* and the intersection is inclusion of B' ; *then,* in particular x is in B'.

For $Red(B) \cap NewRed(B) \subseteq B^*$:

$$\text{Suppose } \forall x, \text{ if } x \in B' \cap B^* \text{ then } x \in B^*;$$

In a simple statement it can be explained that; supposed x is an element in $B' \cap B^*$, where x is in B' and B^* ; and the intersection is inclusion of B^*; *then,* in particular x is in B^*.

Thus, shows that there is an element of x in B' and B^* ;

Therefore: $Red(B) \cap NewRed(B) \neq \phi$.

3 Conclusion

In this paper, a new concept in computation of *Reduct* is proposed. The *Reduct* is denoted as *NewReduct*. *Reduct* is a set of interesting attributes that is capable of representing the knowledge in a DS, therefore *NewReduct* can be defined as a non-interesting set of attributes which is presumed to contain outliers knowledge. The outliers knowledge has its importance to many data mining applications like network intrusion system, fraud detection and medical diagnosis.

References

1. Pawlak, Z.: Rough Sets. Int. Journal of Computer and Information Sciences 11(5), 341–356 (1982)
2. Mollestad, T.: A Rough Set Aproach to Data Mining: Extracting a Logic of Default Rules from Data, The Norwegian University of Science and Technology (1997)
3. Mollestad, T., Komorowski, J.: A Rough Set Framework of Prepositional for Mining Default Rules. Springer, Singapore (1998)
4. Bakar, A.A., Sulaiman, M.N., Othman, M., Selamat, M.H.: IP Algorithms in Compact Rough Classification Modeling. Intelligent Data Analysis 5(5), 419–429 (2001)
5. Polkowski, L.: Advances in Soft Computing, Rough Sets Mathematical Foundations. Springer, Heidelberg (2002)
6. Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W.: Rough Sets. Communication of the ACM [Electronic Version] 38 (1995)
7. Polkowski, L.: Advances in Soft Computing, Rough Sets Mathematical Foundations. Springer, Heidelberg (2002)