Attribute Reduction in Incomplete Information Systems

Shibao Sun[1,2,*], Jianhui Duan[1], and Dandan Wanyan[1]

[1] Electronic Information Engineering College,
Henan University of Science and Technology, Luoyang Henan 471003, China
sunshibao@126.com
[2] National Laboratory Of Software Development Environment,
Beijing University of Aeronautics & Astronautics, Beijing, 100191, China

Abstract. Through changing the equivalence relation of objects to reflexive and symmetric binary relation in the incomplete information system, a cumulative variable precision rough set model is proposed. The basic properties of β lower and β upper cumulative approximation operators are investigated. β upper, and β lower distribution consistent set are explored for defining β upper, and β lower distribution cumulative reduction. Finally, two attribute reduction approaches, such as β upper (β lower) distribution cumulative reduction, in the incomplete information system are given through discernible matrix and function. The example proves that the cumulative variable precision rough set model can effectively deal with information and fully maintain knowledge in incomplete information systems.

Keywords: Variable precision rough set, Incomplete information system, Decision table, Distribution cumulative reduction.

1 Introduction

Rough set theory (RST) has been proposed by Pawlak [1] as a tool to conceptualize, organize and analyze various types of data in knowledge discovery. This method is especially useful for dealing with uncertain and vague knowledge in information systems. Many applications of the RST method to process control, economics, environmental science, chemistry, psychology, conflict analysis and other fields can be found in [2, 3]. However, the classical RST is based on an equivalence relation and cannot be applied in many real situations. Therefore, many extended RST models, e.g., binary relation based rough sets [4], covering based rough sets [5, 6], and fuzzy rough sets [7, 8] have been proposed. In order to solve classification problems with uncertain data and no functional relationship between attributes and relax the rigid boundary definition of the classical RST model to improve the model suitability, the variable precision rough set (VPRS) model was firstly proposed by Ziarko [9] in 1993. It is an effective mathematical tool with an error-tolerance capability to handle uncertainty problem. Basically, the VPRS is an extension of classical RST [1-3], allowing for partial classification. By setting a confidence threshold,

* Corresponding author.

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 587–594, 2011.

$\beta(\beta \in [0,0.5))$, the VPRS can allow data noise or remove data errors [10]. Recently the VPRS model has been widely applied in many fields [11].

The key issues of the VPRS model mainly concentrate on generalizing models and reduction approaches under the equivalence relation. Reduction approaches, such as β-reduct [12], β lower (upper) distribution reduction [13] and reduction based on structure [14], etc, are under the equivalence relation. But in many practical problems, we need to generalize the VPRS model because the equivalence relation of objects is difficult to construct, or the equivalence relation of objects essentially does not exist. The ideas of generalization are from two aspects. One is to generalize approximated objects from crisp set to fuzzy set [15]; The other is to generalize the relation on the universe from the equivalence relation to fuzzy relation [15], binary relation [16], and covering relation [17, 18], etc. Mieszkowicz-Rolka and Rolka devoted to introduce the idea of the VPRS to fuzzy rough set to inquire into theory and application of fuzzy rough set [15]. Gong, et al., generalized the equivalence relation to the binary relation R on universe U in VPRS model, so that generalized variable precision rough set model was obtained [16]. Covering rough set model was obtained when the equivalence relation on the universe was generalized to cover on the universe in rough set model [19]. Zhang, et al., generalized the equivalence relation to cover on universe U in VPRS model such that we obtain two kinds of variable precision covering rough set models [17, 18]. Mieszkowicz-Rolka and Rolka gave the definition of variable precision rough fuzzy set model under the equivalence relation [20].

The classical rough set approach requires the data table to be complete, i.e., without missing values. In practice, however, the data table is often incomplete. To deal with these cases, Greco, et al., proposed an extension of the rough set methodology to the analysis of incomplete data tables [21]. The extended indiscernible relation between two objects is considered as a directional statement where a subject is compared to a referent object. It requires that the referent object has no missing values. The extended rough set approach maintains all good characteristics of its original version. It also boils down to the original approach when there is no missing value. The rules induced from the rough approximations defined according to the extended relation verify a suitable property: they are robust in a sense that each rule is supported by at least one object with no missing value on the condition attributes represented in the rule. Sun, et al. [22] have proposed a new VPRS model and a cumulative VPRS model based on transitive binary relation in the incomplete information system, the cumulative VPRS model maintains the monotonic property of the lower and upper approximation operators. But they only gave the definitions of the lower and upper approximation operators, and did not discuss the approaches of attributes reduction. Obviously, it will bring some difficulties for using these ideas to process information in fact. So, in this paper we will explore the approaches of attributes reduction of the cumulative VPRS model in the incomplete information system.

The paper is organized as follows. In Section 2, a general view of the VPRS approach and incomplete information systems are given. In Section 3, we explore two approaches of attribute reduction approaches in the incomplete information system. In Section 4, we present an illustrative example which is intended to explain the concepts introduced in Section 3.

2 Paper Preparation

Definition 1 [1]**.** An information system is the 4-tuple $S=(U,Q,V,f)$, where U is a non-empty finite set of objects (universe), $Q=\{q_1,q_2,\cdots,q_m\}$ is a finite set of attributes, V_q is the domain of the attribute q, $V=\bigcup_{q\in Q}V_q$ and $f:U\times Q\to V$ is a total function such that $f(x,q)\in V_q$ for each $q\in Q$, $x\in U$, called an information function. If $Q=C\bigcup\{d\}$ and $C\bigcap\{d\}=\varnothing$, then $S=(U,C\bigcup\{d\},V,f)$ is called a decision table, where d is a decision attribute.

To every (non-empty) subset of attributes $P\subseteq C$ is associated an indiscernible relation on U, denoted by R_P:

$$R_P=\{(x,y)\in U\times U: f(x,q)=f(y,q),\ \forall q\in P\}. \tag{1}$$

If $(x,y)\in R_P$, it is said that the objects x and y are P-indiscernible. Clearly, the indiscernible relation thus defined is an equivalence relation.

Similarly, decision attribute d is associated an indiscernible relation R_d:

$$R_d=\{(x,y)\in U\times U: f(x,d)=f(y,d)\} \tag{2}$$

Definition 2 [9]**.** Let X and Y be subsets of non-empty finite universe U, if every $e\in X$ then $e\in Y$, we call Y contain X, denoted as $Y\supseteq X$. Let

$$c(X,Y)=\begin{cases}1-|X\bigcap Y|/|X|, & |X|>0,\\ 0, & |X|=0,\end{cases} \tag{3}$$

Where $|X|$ is cardinality of set X.

Definition 3 [9]**.** Let S be a decision table, X is a nonempty subset of U, $0\le\beta<0.5$ and $\varnothing\ne P\subseteq C$. The β lower approximation and the β upper approximation of X in S are defined, respectively, by:

$$\underline{P}_\beta(X)=\{x\in U: c([x]_P,X)\le\beta\}. \tag{4}$$

$$\overline{P}_\beta(X)=\{x\in U: c([x]_P,X)<1-\beta\}. \tag{5}$$

Definition 4 [21]**.** An information system is called an incomplete information system if there exists $x\in U$ and $a\in C$ that satisfy that the value $f(x,a)$ is unknown, denoted as "*". It assumes here that at least one of the states of x in terms of P is

certain where $P \subseteq C$, i.e. $\exists a \in P$ such that $f(x,a)$ is known. Thus, $V = V_C \bigcup V_d \bigcup \{*\}$.

Definition 5 [21]**.** For each $x, y \in U$ and for each $P \subseteq C$, $yI_P^* x$ means that $f(x,q) = f(y,q)$ *or* $f(x,q) = *$ and/or $f(y,q) = *$ for every $q \in P$. Let $I_P^*(x) = \{y \in U : yI_P^* x\}$ for each $x \in U$ and for each $P \subseteq C$. I_P^* is reflexive and symmetric but not transitive. We can define cumulative β lower and β upper approximation of X :

$$\underline{P}_\beta^*(X) = \{x \in U_P^* : c(I_P^*(x), X) \le \beta. \tag{6}$$

$$\overline{P}_\beta^*(X) = \{x \in U_P^* : c(I_P^*(x), X) < 1 - \beta\}. \tag{7}$$

where $U_P^* = \{x \in U : f(x,q) \neq * \ for$ at least one $q \in P\}$.

$\underline{P}_\beta^*(X)$ and $\overline{P}_\beta^*(X)$ satisfy the following properties:

① For each $X \subseteq U$ and for each $P \subseteq C$: $\underline{P}_\beta^*(X) \subseteq \overline{P}_\beta^*(X)$;

② For each $X \subseteq U$ and for each $P \subseteq C$: $\underline{P}_\beta^*(X) = U_P^* - \overline{P}_\beta^*(U - X)$;

③ For each $X \subseteq U$ and for each $P, R \subseteq C$, if $P \subseteq R$, then $\underline{P}_\beta^*(X) \subseteq \overline{R}_\beta^*(X)$. Furthermore, if $U_P^* = U_R^*$, then $\overline{P}_\beta^*(X) \supseteq \overline{R}_\beta^*(X)$.

3 Attribute Reduction in Incomplete Information Systems

Corollary 1. When $\beta = 0$, variable precision rough set model defined in formula (6) and (7) is equivalent to rough set model in incomplete information system [21].

Proof. In formula (7), $c(I_P^*(x), X) \le \beta$ is equivalent to $c(I_P^*(x), X) \le 0$, so $1 \le |I_P^*(x) \bigcap X| / |I_P^*(x)|$, such that $I_P^*(x) \subseteq X$, that is to say, $\underline{P}_\beta^*(X)$ is equivalent to $\underline{P}^*(X)$. Analogously, $\overline{P}_\beta^*(X)$ is equivalent to $\overline{P}^*(X)$.

Corollary 2. If an information system is complete, variable precision rough set model defined in formula (6) and (7) is equivalent to the classical variable precision rough set model.

Proof. $\forall x, y \in U$, $P \subseteq C$, $yI_P^* x$, then for every $q \in P$, we have (1) $f(y,q) = *$, (2) $f(x,q) = f(y,q)$ *or* $f(x,q) = *$. In a complete information

system, we have $f(x,q)=f(y,q)$, so $I_P^*(x)$ is equal to $[x]$, such that formula (6) is equivalent to formula (4) and formula (7) is equivalent to formula (5).

Definition 6. Suppose that $S=(U,C\bigcup\{d\},V,f)$ is an incomplete information system, for any $A\subseteq C$, $U/R_d=\{d_1,d_2,\cdots,d_r\}$, $j=1,\cdots,r$.

A is called β upper distribution consistent set iff $\overline{A}_\beta^*(d_j)=\overline{C}_\beta^*(d_j)$;

A is called β lower distribution consistent set iff $\underline{A}_\beta^*(d_j)=\underline{C}_\beta^*(d_j)$.

Theorem 1. Let S be an incomplete information system, for each $A\subseteq C$, let

$$M_A^\beta(x)=\{d_j: x\in \overline{A}_\beta^*(d_j)\},(x\in U_A^*). \tag{8}$$

$$G_A^\beta(x)=\{d_j: x\in \underline{A}_\beta^*(d_j)\},(x\in U_A^*). \tag{9}$$

Then

(1) A is called β upper distribution consistent set iff $M_A^\beta(x)=M_C^\beta(x)$;

(2) A is called β lower distribution consistent set iff $G_A^\beta(x)=G_C^\beta(x)$.

Proof. (1) For each $x\in \overline{A}_\beta^*(d_j)$ iff $d_j\in M_A^\beta(x)$, and $x\in \overline{C}_\beta^*(d_j)$ iff $d_j\in M_C^\beta(x)$.

(2) Similar to (1).

Definition 7. Suppose that $S=(U,C\bigcup\{d\},V,f)$ is an incomplete information system, $U/R_C=\{I_c^*(x):x\in U_c^*\}=\{C_i:i=1,\cdots,t\}$

$$D_1^{*\beta}=\{(I_c^*(x),I_c^*(y)):M_C^\beta(x)\neq M_C^\beta(y)\},\ x,y\in U_c^* \tag{10}$$

$$D_2^{*\beta}=\{(I_c^*(x),I_c^*(y)):G_C^\beta(x)\neq G_C^\beta(y)\},\ x,y\in U_c^* \tag{11}$$

Using $f_k(C_i)$ describe $f(a_k,C_i)$. Let

$$D_l^\beta(C_i,C_j)=\begin{cases}\{a_k\in A: f_k(C_i)\neq f_k(C_j)\}, & (C_i,C_j)\in D_l^{*\beta},\\ A, & (C_i,C_j)\notin D_l^{*\beta}.\end{cases}\ (l=1,2) \tag{12}$$

Then $D_1^\beta(C_i,C_j)$,and $D_2^\beta(C_i,C_j)$ are called β upper (β lower) distribution discernible attribute set of C_i and C_j. $\mathrm{D}_1^\beta=(D_1^\beta(C_i,C_j),i,j\le t)$,and $\mathrm{D}_2^\beta=(D_2^\beta(C_i,C_j),i,j\le t)$ are called β upper (β lower) distribution discernible attribute matrix of incomplete information system.

Theorem 2. Suppose that $S=(U,C\bigcup\{d\},V,f)$ be an incomplete information system, for each $A\subseteq C$, then

(1) A is called β upper distribution consistent set iif for each $(C_i,C_j)\in D_1^{*\beta}$, $A\bigcap D_1^{\beta}(C_i,C_j)\neq\varnothing$;

(2) A is called β lower distribution consistent set iif for each $(C_i,C_j)\in D_2^{*\beta}$, $A\bigcap D_2^{\beta}(C_i,C_j)\neq\varnothing$.

Proof. (1) A is called β upper distribution consistent set. For $\forall(C_i,C_j)\in D_1^{*\beta}$, $x,y\in U_C$, let $C_i=I_c^*(x)$ and $C_j=I_c^*(y)$, then $M_C^{\beta}(x)\neq M_C^{\beta}(y)$. For $I_A(x)\bigcap I_A(y)=\varnothing$, then $a_k\in A$, $f_k(x)\neq f_k(y)$, that is $f_k(C_i)\neq f_k(C_j)$. So $a_k\in D_1^{\beta}(C_i,C_j)$.that is $A\bigcap D_1^{\beta}(C_i,C_j)\neq\varnothing$.

On the contrary, if exists $(C_i,C_j)\in D_1^{*\beta}$ and $A\bigcap D_1^{\beta}(C_i,C_j)=\varnothing$, then $\forall x,y\in U_C$, $C_i=I_c^*(x)$ and $C_j=I_c^*(y)$. On the one hand, $(C_i,C_j)\in D_1^{*\beta}$, $M_C^{\beta}(x)\neq M_C^{\beta}(y)$; On the other hand, $\forall a_k\in A$, $a_k\notin D_1^{\beta}(C_i,C_j)$. So $f_k(C_i)=f_k(C_j)$, that is $f_k(x)=f_k(y)$. Then $I_c^*(x)=I_c^*(y)$. That is A is not β upper distribution consistent set.

(2) Similar to (1).

Definition 8. $\mathrm{D}_1^{\beta}=(D_1^{\beta}(C_i,C_j))$, and $\mathrm{D}_2^{\beta}=(D_2^{\beta}(C_i,C_j))$ $(i,j\leq t)$ are β upper (β lower) distribution discernible attribute matrix of incomplete information system. let

$$\begin{aligned}\mathrm{M}_l^{\beta}&=\wedge\{\vee\{a_k:a_k\in D_l^{\beta}(C_i,C_j)\}:i,j\leq t\}\\&=\wedge\{\vee\{a_k:a_k\in D_l^{\beta}(C_i,C_j)\}:(C_i,C_j)\in D_l^{*\beta}\}\end{aligned}\quad(l=1,2)\tag{13}$$

Then M_1^{β} ,and M_2^{β} are called β upper (β lower) distribution cumulative discernible formulas respectively.

Theorem 3. Suppose that $S=(U,C\bigcup\{d\},V,f)$ be an incomplete information system. Minimal disjunctive formulas of M_l^{β} $(l=1,2)$ is defined as:

$$\mathrm{M}_l^{\beta}=\bigvee_{k=1}^{p}(\bigwedge_{s=1}^{q_k}a_s)\ (l=1,2)\tag{14}$$

Let $B_{lk}=\{a_{l_s}:s=1,2,\cdots,q_k\}$, then $\{B_{lk}:k=1,2,\cdots,r\}$ is set of β upper (β lower) distribution cumulative reduction.

Proof. For any $k \le p$ and $(C_i, C_j) \in D_l^{*\beta}$, we have $B_{lk} \bigcap D_l^{\beta}(C_i, C_j) \neq \varnothing$. B_{lk} is β upper(β lower) distribution consistent set. For M_l^{β}, if $B_{lk}' = B_{lk} - a_{l_s}$, *where* $s = 1, 2, \cdots, q_k$, then there exists $(C_i, C_j) \in D_l^{*\beta}$ such that $B_{lk}' \bigcap D_l^{\beta}(C_i, C_j) = \varnothing$. So B_{lk}' is not β upper(β lower) distribution consistent set. That is to say B_{lk}' is not β upper (β lower) distribution reduction.

β upper (β lower) distribution discernible formulas include all $D_l^{\beta}(C_i, C_j)$, so there does not exist β upper (β lower) distribution cumulative reduction.

4 An Example

The director wants to make an evaluation to students based on the level in Mathematics, Physics and Literature. However, there are some missing values as shown in Table 1.

Table 1. Student evaluations with missing values

Student	Mathematics	Physics	Literature	Global evaluation
1	medium	bad	bad	bad
2	good	medium	*	good
3	medium	*	medium	bad
4	*	medium	medium	good
5	*	good	bad	bad
6	good	medium	bad	good

For $\beta = 0.3$, Let $C = \{Mathematics, Physics, Literature\}$ be condition attributes and $\{Global\ evaluation\}$ be decision attribute. Let $bad = \{1,3,5\}$ and $good = \{2,4,6\}$, $U_C^* = \{1,2,3,4,5,6\}$, $\underline{C}_{\beta}^*(bad) = \{1,5\}$, $\overline{C}_{\beta}^*(bad) = \{1,3,4,5\}$, $\underline{C}_{\beta}^*(good) = \{2,6\}$, $\overline{C}_{\beta}^*(good) = \{2,3,4,6\}$. $\mathrm{M}_1^{\beta} = P$. P is β lower distribution cumulative reductions of attribute set C.

From this example, we can see that the cumulative VPRS model in incomplete information system can effectively reduce database, roundly wide application range of the VPRS, and fully maintain information for database.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China (60873108), Application Technology Research and Development Foundation of Luoyang(1101018A), Natural Science Research Foundation of Henan University of Science and Technology (09001172,13440072, 2009Y-016).

References

1. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
2. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Information Sciences 177(1), 28–40 (2007)
3. Pawlak, Z., Skowron, A.: Rough sets and boolean reasoning. Information Sciences 177(1), 41–73 (2007)
4. Zhu, W., Wang, F.-Y.: Binary relation based rough sets. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 276–285. Springer, Heidelberg (2006)
5. Zhu, W.: Topological approaches to covering rough sets. Information Sciences 177(6), 1499–1508 (2007)
6. Zhu, W., Wang, F.-Y.: On three types of covering rough sets. IEEE Transactions on Knowledge and Data Engineering 19(8), 10–41 (2007)
7. Qin, K.Y., Pei, Z.: On the topological properties of fuzzy rough sets. Fuzzy Sets and Systems 151(3), 601–613 (2005)
8. Wu, W.Z., Zhang, W.X.: Constructive and axiomatic approaches of fuzzy approximation operators. Information Sciences 159(3-4), 233–254 (2004)
9. Ziarko, W.: Variable precision rough set model. Journal of Computer System Science 46(1), 39–59 (1993)
10. Slezak, D., Ziarko, W.: The investigation of the Bayesian rough set model. International Journal of Approximation Reason 40, 81–91 (2005)
11. Tao, Z., Xu, B.D., Wang, D.W., Li, R.: Rough Rules Mining Approach Based on Variable Precision Rough Set Theory. Information and Control 33(1), 18–22 (2004)
12. Beynon, M.: Reducts within the variable precision rough sets model: A further investigation. European Journal of Operational Research 134, 592–605 (2001)
13. Zhang, W.X., Liang, Y., Wu, W.Z.: Information System and Knowledge Discovery. Science Press, Beijing (2003) (in Chinese)
14. Inuiguchi, M.: Structure-Based Approaches to Attribute Reduction in Variable Precision Rough Set Models. In: Proceeding of IEEE ICGC 2005, pp. 34–39 (2005)
15. Mieszkowicz-Rolka, A., Rolka, L.: Remarks on approximation quality in variable precision fuzzy rough sets model. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 402–411. Springer, Heidelberg (2004)
16. Gong, Z.T., Sun, B.Z., Shao, Y.B., Chen, D.G.: Variable precision rough set model based on general relations. Journal of Lanzhou University (Natural Sciences) 41(6), 110–114 (2005) (in Chinese)
17. Zhang, Y.J., Wang, Y.P.: Covering rough set model based on variable precision. Journal of Liaoning Institute of Technology 26(4), 274–276 (2006) (in Chinese)
18. Sun, S.B., Liu, R.X., Qin, K.Y.: Comparison of Variable Precision Covering Rough Set Models. Computer engineering 34(7), 10–13 (2008) (in Chinese)
19. Zhu, W., Wang, F.Y.: Reduction and axiomization of covering generalized rough sets. Information Sciences 152, 217–230 (2003)
20. Mieszkowicz-Rolka, A., Rolka, L.: Fuzziness in Information Systems. Electronic Notes in Theoretical Computer Science 82(4), 1–10 (2003)
21. Yang, X.B., Yang, J.Y., Wu, C., Yu, D.J.: Dominance-based rough set approach and knowledge reductions in incomplete ordered information system. Information Sciences 178, 1219–1234 (2008)
22. Sun, S.B., Zheng, R.J., Wu, T.T., Li, T.R.: VPRS-Based Knowledge Discovery Approach in Incomplete Information System. Journal of Computers 5(1), 110–116 (2010)