Positive Approximation and Rule Extracting in Incomplete Information Systems

Yuhua Qian, Jiye Liang

[1] School of Computer and Information Technology, Shanxi University
Taiyuan, 030006, Shanxi, China
[2] Key Laboratory of Computational Intelligence and Chinese Information Processing
of Ministry of Education, Taiyuan, 030006, Shanxi, China
jinchengqyh@126.com, ljy@sxu.edu.cn

Abstract. Set approximation is a kernel concept in rough set theory. In this paper, by introducing a notion of granulation order, positive approximation of a target set under a granulation order is defined in an incomplete information system and its some useful properties are investigated. Unlike classical rough set theory, this approximation deals with how to describe the structure of a rough set in incomplete information systems. For a subset of the universe, its approximation accuracy is monotonously increasing with a granulation order becoming longer. This means that a proper family of granulations can be chosen for a target concept approximation according to user requirements. Furthermore, an algorithm based on the positive approximation, called MABPA II, is designed for decision-rule extracting and a practical example is employed to illustrate its mechanism.

Keywords: Information systems, granular computing, dynamic granulation, rule extracting.

1 Introduction

Granulation is originally a physics concept, used to denote "average measure of granules". Physics granulation is fine partition of physics objects, while information granulation is fine partition of information or knowledge in information systems. Granular computing is an active area of current research in artificial intelligence, and is a new concept and computing pattern for information processing. It has been widely applied to many branches in artificial intelligence field, such as problem solving, knowledge discovery, image processing, semantic Web services.

As follows, for our further development, we briefly review some existing results about granular computing. In 1979, the problem of fuzzy information granule was introduced by Zadeh in [1]. Then, in [2-4] he introduced the concept of granular computing, as a term with many meanings, covering all the research of theory, methods, techniques and tools related to granulation. A general model based on fuzzy set theory was proposed, and granules were defined and constructed basing on the concept of generalized constraints in [3]. Relationships

among granules were represented in terms of fuzzy graphs or fuzzy if-then rules. Pawlak [5] proposed that each equivalence class may be viewed as a granule consisting of indistinguishable elements, also referred to as to an equivalence granule. Some basic problems and methods such as logic framework, concept approximation, and consistent classification for granular computing were outlined by Yao in [6]. The structure, modeling, and applications of granular computing under some binary relations were discussed, and the granular computing methods based on fuzzy sets and rough sets were proposed by Lin in [7]. Quotient space theory was extended to fuzzy quotient space theory based on fuzzy equivalence relation by Zhang and Zhang in [8], providing a powerful mathematical model and tools for granular computing. By using similarity between granules, some basic issues on granular computing were discussed by Klir in [9]. Several measures in information systems closely associated with granular computing, such as granulation measure, information and rough entropy, as well as knowledge granulation, were discussed by Liang in [10, 11]. Decision rule granules and a granular language for logical reasoning based on rough set theory were studied by Liu in [12]. In [13-15], Qian and Liang extended classical rough set model to multi-granulations rough set model, which can overcome the rigorous condition that any two attributes must be independent in rough set theory.

Rough set theory is an important model to research on granular computing. Rough set theory, proposed by Pawlak [16], has become well established as a mechanism for uncertainty management in a wide variety of applications related to artificial intelligence. In recent years, several extensions of rough set model have been proposed, such as variable precision rough set (VPRS) model [17], rough set model based on tolerance relation [18], Bayesian rough set model [19], fuzzy rough set model and rough fuzzy set model [20]. In the view of granular computing, in these rough set models, a concept described by a set is always characterized via the so-called upper and lower approximations under a static granulation, and a static boundary region of the concept is induced by the upper and lower approximations. However a concept described by using positive approximation is characterized via the variational upper and lower approximations under dynamic granulation, which is an aspect of people's comprehensive solving ability at some different granulation spaces [21]. The positive approximation extends classical rough set, and enriches rough set theory and its application. This paper aims to extend this approach to the rough set approximation under dynamic granulation in incomplete information systems.

The rest of this paper is organized as follows: in Section 2, the concepts of a granulation order and the positive approximation under dynamic granulation in an incomplete information system are proposed. For any general concept of the universe, its boundary region is changeable and the approximation accuracy measure is monotonously increasing under a granulation order. This means that a proper family of granulations can be chosen for a target concept approximation according to the requirements of users; in Section 3, an algorithm based on positive approximation is designed to extract decision rules, which will be helping for understanding the mechanism of positive approximation. An illus-

trate example is employed to show how the algorithm MABPA II works as well. Finally, Section 4 concludes whole paper.

2 Positive Approximation in Incomplete Information Systems

In this section, we review some basic concepts such as incomplete information systems, tolerance relation and partial relation of knowledge, introduce the notion of positive approximation to describe the structure of a set approximation in incomplete information systems, and investigate its some useful properties as well.

An information system is a pair $S = (U, A)$, where,

(1) U is a non-empty finite set of objects;

(2) A is a non-empty finite set of attributes;

(3) for every $a \in A$, there is a mapping a, $a : U \to V_a$, where V_a is called the value set of a.

It may happen that some of the attribute values for an object are missing. For example, in medical information systems there may exist a group of patients for which it is impossible to perform all the required tests. These missing values can be represented by the set of all possible values for the attribute or equivalence by the domain of the attribute. To indicate such a situation, a distinguished value, a so-called null value is usually assigned to those attributes.

If V_a contains a null value for at least one attribute $a \in A$, then S is called an incomplete information system, otherwise it is complete [18, 22]. Further on, we will denote the null value by $*$. For example, Table 1 is an incomplete information system.

Table 1. An incomplete information system about car

Car	Price	Mileage	Size	Max-Speed
u_1	high	low	full	low
u_2	low	$*$	full	low
u_3	$*$	$*$	compact	low
u_4	high	$*$	full	high
u_5	$*$	$*$	full	high
u_6	low	high	full	$*$

Let $S = (U, A)$ be an information system, $P \subseteq A$ an attribute set. We define a binary relation on U as follows

$$SIM(P) = \{(u, v) \in U \times U | \ \forall a \in P, a(u) = a(v) \ \text{ or } \ a(u) = * \ \text{ or } \ a(v) = *\}.$$

In fact, $SIM(P)$ is a tolerance relation on U, the concept of a tolerance relation has a wide variety of applications in classification [23, 24].

It can be easily shown that $SIM(P) = \bigcap_{a \in P} SIM(\{a\})$.

Let $S_P(u)$ denote the set $\{v \in U | (u, v) \in SIM(P)\}$. $S_P(u)$ is the maximal set of objects which are possibly indistinguishable by P with u.

Let $U/SIM(P)$ denote the family sets $\{S_P(u) | u \in U\}$, the classification or the knowledge induced by P. A member $S_P(u)$ from $U/SIM(P)$ will be called a tolerance class or a granule of information. It should be noticed that the tolerance classes in $U/SIM(P)$ do not constitute a partition of U in general. They constitute a cover of U, i.e., $S_P(u) \neq \emptyset$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u) = U$.

Let $S = (U, A)$ be an incomplete information system, we define a partial relation $\preceq$ (or $\succeq$) on 2^A as follows: we say that Q is coarser than P (or P is finer than Q), denoted by $P \preceq Q$ (or $Q \succeq P$), if and only if $S_P(u_i) \subseteq S_Q(u_i)$ for $i \in \{1, 2, \cdots, |U|\}$. If $P \preceq Q$ and $P \neq Q$, we say that Q is strictly coarser than P (or P is strictly finer than Q) and denoted by $P \prec Q$ (or $Q \succ P$).

In fact, $P \prec Q \Leftrightarrow$ for $i \in \{1, 2, \cdots, |U|\}$, we have that $S_P(u_i) \subseteq S_Q(u_i)$, and $\exists j \in \{1, 2, \cdots, |U|\}$, such that $S_P(u_j) \subset S_Q(u_j)$.

Let $S = (U, A)$ be an incomplete information system, X a subset of U and $P \subseteq A$ an attribute set. In the rough set model based on tolerance relation [14], X is characterized by $\overline{SIM(P)}(X)$ and $\underline{SIM(P)}(X)$, where

$$\underline{SIM(P)}(X) = \bigcup \{Y \in U/SIM(P) | Y \subseteq X\}, \tag{1}$$

$$\overline{SIM(P)}(X) = \bigcup \{Y \in U/SIM(P) | Y \bigcap X \neq \emptyset\}. \tag{2}$$

In an incomplete information system, a cover $U/SIM(P)$ of U induced by the tolerance relation $SIM(P)$, $P \in 2^A$, provides a granulation world for describing a concept X. So a sequence of attribute sets $P_i \in 2^A$ ($i = 1, 2, \cdots, n$) with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ can determine a sequence of granulation worlds, from the most rough one to the most fine one. We define the upper and lower approximations of a concept under a granulation order.

Definition 1. *Let $S = (U, A)$ be an incomplete information system, X a subset of U and $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ ($P_i \in 2^A$), we define $\boldsymbol{P}$-upper approximation $\overline{\boldsymbol{P}}X$ and $\boldsymbol{P}$-lower approximation $\underline{\boldsymbol{P}}X$ of X as follows:*

$$\overline{\boldsymbol{P}}(X) = \overline{SIM(P_n)}(X), \tag{3}$$

$$\underline{\boldsymbol{P}}(X) = \bigcup_{i=1}^{n} \underline{SIM(P_i)}(X_i), \tag{4}$$

where $X_1 = X$ and $X_i = X - \bigcup_{k=1}^{i-1} \underline{SIM(P_k)}(X_k)$ for $i = 2, \cdots, n$.

$bn_{\boldsymbol{P}}(X) = \overline{\boldsymbol{P}}(X) - \underline{\boldsymbol{P}}(X)$ is called $\boldsymbol{P}$-boundary region of X, $pos_{\boldsymbol{P}}(X) = \underline{\boldsymbol{P}}(X)$ is called $\boldsymbol{P}$-positive region of X, and $neg_{\boldsymbol{P}}(X) = U - \overline{\boldsymbol{P}}(X)$ is called $\boldsymbol{P}$-negative region of X. Obviously, we have $\overline{\boldsymbol{P}}(X) = pos_{\boldsymbol{P}}(X) \cup bn_{\boldsymbol{P}}(X)$.

Definition 1 shows that a target concept is approached by the change of the lower approximation $\underline{\mathbf{P}}(X)$ and the upper approximation $\overline{\mathbf{P}}(X)$.

From this definition, we have the following theorem.

Theorem 1. *Let $S = (U, A)$ be an incomplete information system, X a subset of U and $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ $(P_i \in 2^A)$. Let $\boldsymbol{P}_i = \{P_1, P_2, \cdots, P_i\}$. Then for $\boldsymbol{P}_i$ $(i = 1, 2, \cdots, n)$, we have that*

$$\underline{\boldsymbol{P}_i}(X) \subseteq X \subseteq \overline{\boldsymbol{P}_i}(X), \tag{5}$$

$$\underline{\boldsymbol{P}_1}(X) \subseteq \underline{\boldsymbol{P}_2}(X) \subseteq \cdots \subseteq \underline{\boldsymbol{P}_n}(X). \tag{6}$$

Proof. The proof follows directly from Definition 1.

Theorem 1 states that the lower approximation enlarges as the granulation order become longer through adding attribute subsets, which help to describe exactly a target concept.

In [25], the approximation measure $\alpha_R(X)$ was originally introduced by Z. Pawlak for classical lower and upper approximation, where $\alpha_R(X) = \frac{|\underline{R}X|}{|\overline{R}X|}(X \neq \emptyset)$. Here we introduce the concept to the positive approximation in order to describe the uncertainty of a target concept under a granulation order.

Definition 2. *Let $S = (U, A)$ be an incomplete information system, X a subset of U and $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ $(P_i \in 2^A)$. The approximation measure $\alpha_{\boldsymbol{P}}(X)$ is defined as*

$$\alpha_{\boldsymbol{P}}(X) = \frac{|\underline{\boldsymbol{P}}(X)|}{|\overline{\boldsymbol{P}}(X)|}, \tag{7}$$

where $X \neq \emptyset$.

Theorem 2. *Let $S = (U, A)$ be an incomplete information system, X a subset of U and $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ $(P_i \in 2^A)$. Let $\boldsymbol{P}_i = \{P_1, P_2, \cdots, P_i\}$, then*

$$\alpha_{\boldsymbol{P}_1}(X) \leq \alpha_{\boldsymbol{P}_2}(X) \leq \cdots \leq \alpha_{\boldsymbol{P}_n}(X). \tag{8}$$

Proof. The proof follows directly from Theorem 1 and Definition 2.

Theorem 2 states that the approximation measure $\alpha_{\mathbf{P}}(X)$ increases as the granulation order become longer through adding attribute subsets.

In order to illustrate the essence that positive approximation is mainly concentrated on the change of the construction of the target concept X (tolerance classes in lower approximation of X with respect to $\mathbf{P}$) in incomplete information systems, we can re-define $\mathbf{P}$-positive approximation of X by using some tolerance classes on U.

Therefore, the structure of $\mathbf{P}$-upper approximation $\overline{\mathbf{P}}(X)$ and $\mathbf{P}$-lower approximation $\underline{\mathbf{P}}(X)$ of $\mathbf{P}$-positive approximation of X can be represented as follows

$$[\overline{\mathbf{P}}(X)] = \{S_{P_n}(u) \mid S_{P_n}(u) \cap X \neq \varnothing, u \in U\}, \tag{9}$$

$$[\underline{\mathbf{P}}(X)] = \{S_{P_i}(u) \mid S_{P_i}(u) \subseteq X_i, i \leq n, u \in U\}, \tag{10}$$

where $X_1 = X$, $X_i = X - \bigcup_{k=1}^{i-1} \underline{SIM(P_k)(X_k)}$ for $i = 2, \cdots, n$, and $[\cdot]$ denotes the structure of a rough approximation.

In the following, we show how positive approximation in an incomplete information system works by an illustrate example.

Example 1. Support $S = (U, A)$ be an incomplete information system, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, $P, Q \subseteq A$ two attribute sets, $X = \{u_1, u_2, u_3, u_5, u_6\}$, $SIM(P) = \{\{u_1, u_2\}, \{u_1, u_2\}, \{u_2, u_3\}, \{u_3, u_4, u_5\}, \{u_4, u_5, u_6\}, \{u_4, u_5, u_6\}\}$, $SIM(Q) = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4, u_5\}, \{u_4, u_5\}, \{u_5, u_6\}\}$.

Obviously, $P \succeq Q$ holds. Hence, we can construct a granulation order (a family of tolerance relations) $\mathbf{P} = \{P, Q\}$, where $\mathbf{P}_1 = \{P\}$, $\mathbf{P}_2 = \{P, Q\}$.

By computing the positive approximation of X with respect to $\mathbf{P}$, we obtain easily that

$$[\underline{\mathbf{P}_1}(X)] = \{\{u_1, u_2\}, \{u_1, u_2\}, \{u_2, u_3\}\}$$
$$[\overline{\mathbf{P}_1}(X)] = \{\{u_1, u_2\}, \{u_1, u_2\}, \{u_2, u_3\}, \{u_3, u_4, u_5\}, \{u_4, u_5, u_6\}, \{u_4, u_5, u_6\}\},$$
$$[\underline{\mathbf{P}_2}(X)] = \{\{u_1, u_2\}, \{u_1, u_2\}, \{u_2, u_3\}, \{u_5, u_6\}\},$$
$$[\overline{\mathbf{P}_2}(X)] = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4, u_5\}, \{u_4, u_5\}, \{u_5, u_6\}\}.$$

Where $\{u_1, u_2\}, \{u_2, u_3\}$ in $[\underline{\mathbf{P}_2}(X)]$ are not induced by the tolerance relation $SIM(Q)$ but $SIM(P)$, and $[\overline{\mathbf{P}_2}(X)]$ is induced by the tolerance relation $SIM(Q)$. In other words, the target concept X is described by using the granulation order $\mathbf{P} = \{P, Q\}$.

In order to reveal the properties of positive approximation based on dynamic granulation in incomplete information systems, we here introduce the notion of $\sqsubseteq$.

Assume A, B be two families of tolerance classes sets, where $A = \{A_1, A_2, \cdots, A_m\}$, $B = \{B_1, B_2, \cdots, B_n\}$. We say $A \sqsubseteq B$, if and only if, for $A_i \in A$, there exists $B_j \in B$ such that $A_i \subseteq B_j$ $(i \leq m, j \leq n)$.

Theorem 3. *Let $S = (U, A)$ be an incomplete information system, $X \subseteq U$ and $\mathbf{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$. Let $\mathbf{P}_i = \{P_1, P_2, \cdots, P_i\}$, then $[\underline{SIM(P_i)}(X)] \sqsubseteq [\underline{\mathbf{P}_i}(X)]$.*

Proof. It follows from Definition 1 and Theorem 1 that $\underline{SIM(P_i)}(X) = \underline{\mathbf{P}_i}(X)$. And, from Definition 1, we know that $\underline{SIM(P_i)}(X)$ can be regarded as the union of $\underline{\mathbf{P}_{i-1}}(X)$ and $\underline{SIM(P_i)}(X - \underline{\mathbf{P}_{i-1}}(\overline{X}))$.

Hence, for arbitrary $S_{P_i}(u) \in [\underline{SIM(P_i)}(X)]$, if $S_{P_i}(u) \in [\underline{SIM(P_i)}(X - \underline{\mathbf{P}_{i-1}}(X))]$, then $S_{P_i}(u) \in [\underline{\mathbf{P}_i}(X)]$ holds; if $S_{P_i}(u) \in [\underline{SIM(P_i)}(\underline{\mathbf{P}_{i-1}}(X))]$, then there exist k such that $S_{P_i}(u) \subseteq S_{P_k}(u)$ $(1 \leq k < i)$ holds. In other words, for arbitrary $u \in \underline{SIM(P_i)}(X)$, there exist some k $(1 \leq k \leq i)$ such that $S_{P_i}(u) \subseteq S_{P_k}(u)$, where $\overline{S_{P_k}(u)} \in [\underline{\mathbf{P}_i}(X)]$. Therefore, $[\underline{SIM(P_i)}(X)] \sqsubseteq [\underline{\mathbf{P}_i}(X)]$ holds.

This completes the proof.

Remark. Theorem 3 states that there is an inclusion relationship between the structure of the classical lower approximation $\underline{SIM(P_i)}(X)$ and the structure of this new lower approximation $\underline{\mathbf{P}_i}(X)$ based on a granulation order. In fact, for approximating a target concept, this mechanism establishes a family of tolerance classes with a hierarchy nature from rough to fine on the basis of keeping the approximation measure. Hence, in a board sense, the positive approximation will be helpful for extracting decision rules with hierarchy nature according to user requirements in incomplete information systems.

In the following, we introduce an approach to build a granulation order in an incomplete information system. As we know, the tolerance classes induced by an attribute set are finer than those of induced by any attribute subset in general. This idea can be used to build a granulation order from rough to fine on attribute power set. It can be understood by the below theorem.

Theorem 4. *Let $S = (U, A)$ be an incomplete information system, where $A = \{a_1, a_2, \cdots, a_n\}$. Denote by $A_i = \{a_1, a_2, \cdots, a_i\}$ $(i \le n)$, then $\mathbf{P} = \{A_1, A_2, \cdots, A_n\}$ is a granulation order from rough to fine.*

Proof. It is straightforward.

In practical issues, a granulation order on attribute set can be appointed by user or experts, or be built according to the significance of each attribute. In particular, in an incomplete decision table (i.e., an incomplete information system with a decision attribute), some certain/uncertain decision rules can be extracted through constructing the positive approximation of a target decision.

Let $S = (U, C \cup D)$ be an incomplete decision table, $\mathbf{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$. $\Gamma = U/D = \{D_1, D_2, \cdots, D_r\}$ be a decision (partition) on U, a lower approximation and a upper approximation of Γ related to $\mathbf{P}$ are defined by

$$[\underline{\mathbf{P}\Gamma}] = \{[\underline{\mathbf{P}}(D_1)], [\underline{\mathbf{P}}(D_2)], \cdots, [\underline{\mathbf{P}}(D_r)]\},$$
$$[\overline{\mathbf{P}\Gamma}] = \{[\overline{\mathbf{P}}(D_1)], [\overline{\mathbf{P}}(D_2)], \cdots, [\overline{\mathbf{P}}(D_r)]\}.$$

In addition, we call $[bn_{\mathbf{P}}\Gamma] = \{[\overline{\mathbf{P}}(D_i)] - [\underline{\mathbf{P}}(D_i)]\} : i \le r)$ $\mathbf{P}$-boundary region of Γ. Note that tolerance classes in $[\underline{\mathbf{P}\Gamma}]$ can induce certain decision rules, while those in $[bn_{\mathbf{P}}\Gamma]$ can extract uncertain decision rules from an incomplete decision table.

Similar to the formula (7), in the following, we give the notion of approximation measure of a target decision under a granulation order in an incomplete decision table.

Definition 3. *Let $S = (U, C \cup D)$ be an incomplete decision table, $\Gamma = U/D = \{D_1, D_2, \cdots, D_r\}$ and $\mathbf{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ $(P_i \in 2^C)$. The approximation measure $\alpha_{\mathbf{P}}(\Gamma)$ is defined as*

$$\alpha_{\boldsymbol{P}}(\Gamma) = \sum_{k=1}^{r} \frac{|D_r|}{|U|} \frac{|\underline{\boldsymbol{P}}(D_k)|}{|\overline{\boldsymbol{P}}(D_k)|}. \tag{11}$$

Theorem 5. *Let $S = (U, C \cup D)$ be an incomplete decision table, $\Gamma = U/D = \{D_1, D_2, \cdots, D_r\}$ and $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ ($P_i \in 2^C$). Let $\boldsymbol{P}_i = \{P_1, P_2, \cdots, P_i\}$, then*

$$\alpha_{\boldsymbol{P}_1}(\Gamma) \leq \alpha_{\boldsymbol{P}_2}(\Gamma) \leq \cdots \leq \alpha_{\boldsymbol{P}_n}(\Gamma). \tag{12}$$

Proof. Suppose $1 \leq i \leq j \leq r$.

From Theorem 2 and Definition 3, we have that

$$\alpha_{\mathbf{P}_i}(\Gamma) = \sum_{k=1}^{r} \frac{|D_r|}{|U|} \frac{|\underline{\mathbf{P}_i}(D_k)|}{|\overline{\mathbf{P}_i}(D_k)|} \leq \sum_{k=1}^{r} \frac{|D_r|}{|U|} \frac{|\underline{\mathbf{P}_j}(D_k)|}{|\overline{\mathbf{P}_j}(D_k)|} = \alpha_{\mathbf{P}_j}(\Gamma).$$

Hence, it follows that $\alpha_{\mathbf{P}_1}(\Gamma) \leq \alpha_{\mathbf{P}_2}(\Gamma) \leq \cdots \leq \alpha_{\mathbf{P}_n}(\Gamma)$.

This completes the proof.

Theorem 5 states that the approximation measure $\alpha_{\mathbf{P}}(\Gamma)$ increases as the granulation order become longer through adding attribute subsets.

3   An application for decision rule extracting

We apply rough set methods for decision rule mining from decision tables. It is not always possible to extract general laws from experimental data by computing first all reducts of a decision table and next decision rules on the basis of these reducts [26, 27].

In this section, we proposed an algorithm for decision rule mining in incomplete decision tables by using positive approximation. The application will be helping for understanding the idea of positive approximation proposed in the paper.

Let $S = (U, C \cup D)$ be an incomplete decision table, where C and D are condition and decision attribute sets respectively, and $C \cap D = \emptyset$. The positive region of D with respect to C is defined as follows

$$pos_C(D) = \bigcup_{X \in U/D} \underline{SIM(C)}(X). \tag{13}$$

In a decision table $S = (U, C \cup D)$, the significance of $c \in C$ with respect to D is defined as follows [21]:

$$sig_{C-\{c\}}^D(c) = \gamma_C(D) - \gamma_{C-\{c\}}(D), \tag{14}$$

where $\gamma_C(D) = \frac{|pos_C(D)|}{|U|}$.

In a decision table $S = (U, C \cup D)$, the significance of $c \in C - C'$ ($C' \subseteq C$) with respect to D is defined as follows

$$sig_{C'}^D(c) = \gamma_{C' \cup \{c\}}(D) - \gamma_{C'}(D), \tag{15}$$

where $\gamma_C(D) = \frac{|pos_{C'}(D)|}{|U|}$.

Based on the algorithm MABPA for extracting hierarchy decision rules from a complete decision table [21], we here extend it to mine certain decision rules from an incomplete decision table in the context of a granulation order.

Algorithm MABPA II (mining rules in an incomplete decision table)

Input: an incomplete decision table $S = (U, C \cup D)$;

Output: decision rules $Rule$.

(1) For $\forall c \in C$, compute the significance and relative core

$$core_D(C) = \{c \in C | sig^D_{C-c}(c) > 0\};$$

(2) If $core_D(C) \neq \emptyset$, let $P_1 = core_D(C)$; else, for $\forall c \in C$, compute the dependence $\gamma_c(D)$ of D to c; let $\gamma_{c_1}(D) = \max\{\gamma_c(D) | c \in C\}$ and $P_1 = c_1$;

(3) Compute $U/D = \{Y_1, Y_2, \cdots, Y_d\}$;

(4) Let $\mathbf{P} = \{P_1\}$, $i = 1$, $U^* = U$, $\Omega = \emptyset$, $Rule = \emptyset$;

(5) Compute $U^*/SIM(P_i) = \{S_{P_i}(u) : u \in U^*\}$;

(6) Let $\Omega' = \{S_{P_i}(u) \in U^*/SIM(P_i) \mid S_{P_i}(u) \subseteq Y_j (Y_j \in U/D, j = \{1, 2, \cdots, d\})\}$. Let $Rule' = \emptyset$, for arbitrary $S_{P_i}(u) \in \Omega'$, put $des_{P_i}(S_{P_i}(u)) \longrightarrow des_D(Y_j)(Y_j \in U/D, S_{P_i}(u) \subseteq Y_j$ into $Rule'$. Let $Rule = Rule \cup Rule'$, $\Omega = \Omega \cup \Omega'$;

(7) If $\bigcup_{Z \in \Omega} Z = U$, go to (8); else, $U^* = U^* - \bigcup_{Z \in \Omega} Z$, for $\forall c \in C - P_i$, compute $sig^D_{P_i}(c)$, let $sig^D_{P_i}(c_2) = max\{sig^D_{P_i}(c), c \in C - P_i\}$, if $sig^D_{P_i}(c_2) = 0$ then go to (8), otherwise $P_{i+1} = P_i \cup \{c_2\}$, let $\mathbf{P} = \mathbf{P} \cup \{P_{i+1}\}$, $i = i + 1$, go to (5);

(8) Output $Rule$.

Obviously, generation of decision rules is not based on a reduct of a decision table, but P (a granulation order) and U^* in the MABPA II.

By using MABPA II algorithm, the time complexity to extract decision rules from an incomplete decision table is polynomial.

At the first step, we need to compute $core_D(C)$, i.e., compute $sig^D_{C-c}(c)$ for all $c \in C$. The time complexity for computing $core_D(C)$ is $O(|C||U|^2)$.

At step 3, the time complexity for computing U/D is $O(|U|^2)$.

At step 5, the time complexity for computing $U^*/SIM(P_i)$ is $O(|U|^2)$.

At step 7, the time complexity for computing all $sig^D_{P_i}(c)$ is $O(|C - P_i||C||U|^2)$; the time complexity to choose maximum for significance of attribute is $|C - P_i|$.

From step 5 to step 7, $|C| - 1$ is the maximum value for the circle times. Therefore, the time complexity is

$$\sum_{i=1}^{|C|-1}(O(|U|^2) + O(|C - P_i||C||U|^2) + O(|C - P_i|)) = O(|C|^3|U|^2).$$

Other steps will not be considered because that their time complexity are all const.

Thus the time complexity of the algorithm MABPA II is as follows

$$O(|C||U|^2) + O(|U|^2) + O(|U|^2) + O(|C|^3|U|^2) = O(|C|^3|U|^2).$$

In next part, we show how the algorithm MABPA II works by the following example.

A consistent decision table $S = (U, C \cup D)$ is given by Table 2, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, $C = \{a_1, a_2, a_3, a_4\}$ with a_1-Price, a_2-Mileage, a_3-Size, a_4-Max-Speed, and $D = \{d\}$. By the algorithm MABPA II, we can extract decision rules from Table 2.

Table 2. The incomplete decision table about car [28]

Car	Price	Mileage	Size	Max-Speed	d	∂_d
u_1	high	low	full	low	good	{good}
u_2	low	*	full	low	good	{good}
u_3	*	*	compact	low	poor	{poor}
u_4	high	*	full	high	good	{good, excellent}
u_5	*	*	full	high	excellent	{good, excellent}
u_6	low	high	full	*	good	{good, excellent}

By computing, it follows that $U/SIM(C) = \{S_C(u_1), S_C(u_2), S_C(u_3), S_C(u_4), S_C(u_5), S_C(u_6)\}$ and $U/D = \{\{1, 2, 4, 6\}, \{3\}, \{5\}\}$, where $S_C(u_1) = \{u_1\}$, $S_C(u_2) = \{u_2, u_6\}$, $S_C(u_3) = \{u_3\}$, $S_C(u_4) = \{u_4, u_5\}$, $S_C(u_5) = \{u_4, u_5, u_6\}$, $S_C(u_6) = \{u_2, u_5, u_6\}$.

According to the formula $sig_{C-\{c\}}^D(c) = \gamma_C(D) - \gamma_{C-\{c\}}(D)$, we have that $sig_{C-\{a_1\}}^D(a_1) = 0$, $sig_{C-\{a_2\}}^D(a_2) = 0$, $sig_{C-\{a_3\}}^D(a_3) = \frac{2}{3}$, $sig_{C-\{a_4\}}^D(a_4) = \frac{1}{6}$. So we get $core_D(C) = \{a_3, a_4\}$.

Hence, $P_1 = \{a_3, a_4\}$ and $\mathbf{P} = \{P_1\}$.

By computing, we know $U/SIM(P_1) = \{\{u_1, u_2, u_6\}, \{u_1, u_2, u_6\}, \{u_3\}, \{u_4, u_5, u_6\}, \{u_4, u_5, u_6\}, \{u_1, u_2, u_4, u_5, u_6\}\}$ and $\Omega = \Omega' = \{\{u_1, u_2, u_6\}, \{u_1, u_2, u_6\}, \{u_3\}\}$. Thus, we get two certain decision rules as follows

$$Rule = \{r_1 : des_{\{P_1\}}(\{u_1, u_2, u_6\}) \rightarrow des_D(\{u_1, u_2, u_4, u_6\}),$$
$$r_2 : des_{\{P_1\}}(u_3) \rightarrow des_D(\{u_3\})\}.$$

For $\bigcup_{Z \in \Omega} Z = \{u_1, u_2, u_3, u_6\} \neq U$, we need to compute significance of the rest of attributes a_1, a_2 with respect to D. By the formula $sig_{C'}^D(c) = \gamma_{C' \cup \{c\}}(D) - \gamma_{C'}(D)$, we can obtain

$$sig_{P_1 \cup \{a_1\}}^D(a_1) = \gamma_{P_1 \cup \{a_1\}}(D) - \gamma_{P_1}(D) = 0,$$
$$sig_{P_1 \cup \{a_2\}}^D(a_2) = \gamma_{P_1 \cup \{a_2\}}(D) - \gamma_{P_1}(D) = 0.$$

Since $sig_{P_1 \cup \{a_1\}}^D(a_1) = sig_{P_1 \cup \{a_2\}}^D(a_2) = 0$, the algorithm MABPA II is ended, and $Rule$ is obtained.

For intuition, the two certain decision rules obtained by MABPA II from the decision table S are listed in Table 3.

This example shows the mechanism of the decision rule mining algorithm based on positive approximation. In fact, this algorithm can be also used to extract uncertain decision rules from the boundary of positive approximation of a target decision in an incomplete decision table.

Table 3. Rules obtained for the decision table S.

Rule	attrubutes		
	Size	Max-Speed	d
r_1	full	low	good
r_2	compact	low	poor

4   Conclusions

In this paper, through using a concept of granulation order, we have extended the rough set approximation under static granulation to the rough set approximation under dynamic granulation in the context of incomplete information systems and have presented the concept of positive approximation with granulation from rough to fine. Some of its important properties have been investigated. Note that a target concept can be approached by the change of the positive approximation. An algorithm based on the positive approximation for decision rule mining has been given and its application has been illustrated by an illustrative example as well. The results developed in this paper will be helpful for further research on rough set theory and its practical applications.

References

1. Zadeh, L.A.: Fuzzy Sets and Information Granularity. In: Gupta, M., Ragade, R., Yager, R. (Eds.): Advances in Fuzzy Set Theory and Application, North-Holland, Amsterdam (1979) 3-18
2. Zadeh, L.A.: Fuzzy logic=computing with words. IEEE Transactions on Fuzzy Systems, 4(1) (1996) 103-111
3. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 90 (1997) 111-127
4. Zadeh, L.A.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information / intelligent systems. Soft Computing, 2(1) (1998) 23-25
5. Pawlak, Z.: Granularity of knowledge, indiscernibility and rough sets. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems (1998) 106-110
6. Yao, Y.Y.: Granular computing: basic issues and possible solutions. In: Proceedings of the Fifth International Conference on Computing and Information, I (2000) 186-189

7. Lin, T.Y.: Granular computing on binary relations I: Data mining and neighborhood systems, II: Rough sets representations and belief functions. In: Polkowski, L., Skowron, A. (Eds.): Rough Sets in Knowledge Discovery 1. Physica-Verlag, Heidelberg (1998) 107-140

8. Zhang, L., Zhang, B.: Theory of fuzzy quotient space (methods of fuzzy granular computing). Journal of Software (in Chinese), 14(4) (2003) 770-776

9. Klir, G.J.: Basic issues of computing with granular computing. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems (1998) 101-105

10. Liang, J.Y., Shi, Z.Z.: The information entropy, rough entropy and knowledge granulation in rough set theory. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 12(1) (2004) 37-46

11. Liang, J.Y., Shi, Z.Z., Li, D.Y.: The information entropy, rough entropy and knowledge granulation in incomplete information systems. International Journal of General Systems, 35(6) (2006) 641-654

12. Liu, Q.: Granules and applications of granular computing in logical reasoning. Journal of Computer Research and Development (in Chinese), 41(4) (2004) 546-551

13. Qian, Y.H., Liang J.Y.: Rough set method based on multi-granulations. Proceedings of 5th IEEE Conference on Cognitive Informatics, Vol. I (2006) 297-304

14. Qian, Y.H., Dang C.Y., Liang J.Y.: Multi-granulations rough set approach and rule extracting. Artificial Intelligence (In review)

15. Qian, Y.H., Liang J.Y.: MGRS in incomplete information systems. The Proceedings of 2007 IEEE International Conference on Granular Computing (In press)

16. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science, 11 (1982) 341-356

17. Ziarko, W.: Variable precision rough sets model. Journal of Computer System Science, 46(1) (1993) 39-59

18. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Information Sciences, 112 (1998) 39-49

19. Ślezak, D, Ziarko, W.: The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, 40 (2005) 81-91

20. Dübois, D, Prade, H.: Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems, 17 (1990) 191-209

21. Liang, J.Y., Qian, Y.H, Chu, C.Y., li, D.Y., Wang, J.H.: Rough set approximation based on dynamic granulation. Lecture Notes in Artificial Intelligence, 3641 (2005) 701-708

22. Kryszkiewicz, M.: Rule in incomplete information systems. Information Sciences, 113 (1999) 271-292

23. Kryszkiewicz, M.: Comparative study of alternative type of knowledge reduction in incomsistent systems. International Journal of Intelligent systems, 16 (2001) 105-120

24. Leung, Y., Li, D.Y.: Maximal consistent block technique for rule acquisition in incomplete information systems. Information Sciences, 153 (2003) 85-106

25. Pawlak, Z.: Rough sets. Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991)

26. Bazan, J., Nguyen, H. S., Skowron, A.: Rough set methods in approximation of hierarchical concepts. In: S. Tsumoto et al. (Eds.): Rough Sets and Current Trends in Computing RSCTC 2004, Lecture Notes in Computer Science 3066 (2004) 342-351.

27. Bazan, J., Nguyen, H. S., Skowron, A., Szczuka, M.: A view on rough set concept approximations, In: Wang, G., Liu, Q., Yao, Y.Y., Skowron, A.: Proceedings of the Ninth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing RSFDGrC (2003), Chongqing, China, May 26-29, 2003, Lecture Notes in Artificial Intelligence, Springer-Verlag, Heidelberg, 2639 (2003) 181-188.
28. Liang, J.Y., Li, D.Y.: Uncertainty and knowledge acquisition in information systems (in Chinese). Beijing, Science Press (2005)