Rough Set Approach Under Dynamic Granulation in Incomplete Information Systems

Yuhua Qian[1,2,3], Jiye Liang[1,2], Xia Zhang[1,2], and Chuangyin Dang[3]

[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
[3] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong
jinchengqyh@126.com, ljy@sxu.edu.cn, zhangxia@sxu.edu.cn, mecdang@cityu.edu.hk

Abstract. In this paper, the concept of a granulation order is proposed in an incomplete information system. Positive approximation of a set under a granulation order is defined and its some useful properties are investigated. Unlike classical rough set, this approach focuses on 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 under a granulation order. This means that a proper family of granulations can be chosen for a target-concept approximation according to user requirements.

Keywords: Information systems, granular computing, dynamic granulation, partial relation.

1 Introduction

Granular computing is a new active area of current research in artificial intelligence, and is a new concept and computing formula for information processing. It has been widely applied to many branches of artificial intelligence such as problem solving, knowledge discovery, image processing, semantic Web services.

In 1979, the problem of fuzzy information granules was introduced by L.A. Zadeh in [1]. Then, in [2-4] he introduced a 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. Z. 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 Y.Y. Yao in

A. Gelbukh and A.F. Kuri Morales (Eds.): MICAI 2007, LNAI 4827, pp. 1–8, 2007.

[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 T.Y. Lin in [7]. Quotient space theory was extended to fuzzy quotient space theory based on fuzzy equivalence relation by L. Zhang and B. 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 G.J. 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 J.Y. Liang in [10, 11]. Decision rule granules and a granular language for logical reasoning based on rough set theory were studied by Q. Liu in [12].

In the view of granular computing, a concept described by a set is always characterized via the so-called upper and lower approximations under static granulation in rough set theory, and a static boundary region of the concept is induced by the upper and lower approximations. However a concept described by using the 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 [13]. 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.

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.

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 [14, 15]. Further on, we will denote the null value by $*$.

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 [16, 17]. 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 $i = 2, \cdots, n$.

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

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*

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

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

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 [18] , 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*

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

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 the 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 **P**) in incomplete information systems, we can re-define **P**-positive approximation of X by using some tolerance classes on U.

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

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

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

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 the 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 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 $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$. Let $\boldsymbol{P}_i = \{P_1, P_2, \cdots, P_i\}$, then $[\underline{SIM(P_i)}(X)] \sqsubseteq [\underline{\boldsymbol{P}_i}(X)]$.*

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 building 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 $A_i = \{a_1, a_2, \cdots, a_i\}$ $(i \leq n)$, then $\boldsymbol{P} = \{A_1, A_2, \cdots, A_n\}$ is a granulation order from rough to fine.*

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 $[bn_{\mathbf{P}}\Gamma] = \{[\overline{\mathbf{P}}(D_i)] - [\underline{\mathbf{P}}(D_i)]\} : i \leq r)$ $\mathbf{P}$-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 $\boldsymbol{P} = \{P_1, P_2, \cdots, P_n\}$ a family of attribute sets with $P_1 \succeq P_2 \succeq \cdots \succeq P_n$ $(P_i \in 2^C)$. The approximation measure $\alpha_{\boldsymbol{P}}(\Gamma)$ 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*

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

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

3 Conclusions

In this paper, we have extended rough set approximation under static granulation to rough set approximation under dynamic granulation in the context of incomplete information systems, and its some properties have been investigated. A target concept can be approached by the change of the positive approximation. The results obtained in this paper will be helpful for further research on rough set theory and its practical applications.

Acknowledgements. This work was supported by the national high technology research and development program (No. 2007AA01Z165), the national natural science foundation of China (No. 70471003, No. 60573074), the foundation of doctoral program research of the ministry of education of China (No. 20050108004), key project of science and technology research of the ministry of education of China (No. 206017) and the graduate student innovation foundation of Shanxi.

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, pp. 3–18. North-Holland, Amsterdam (1979)
2. Zadeh, L.A.: Fuzzy logic=computing with words. IEEE Transactions on Fuzzy Systems 4(1), 103–111 (1996)
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, 111–127 (1997)
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), 23–25 (1998)
5. Pawlak, Z.: Granularity of knowledge, indiscernibility and rough sets. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems, pp. 106–110. IEEE Computer Society Press, Los Alamitos (1998)
6. Yao, Y.Y.: Granular computing: basic issues and possible solutions. In: Proceedings of the Fifth International Conference on Computing and Information, vol. I, pp. 186–189 (2000)
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, pp. 107–140. Physica, Heidelberg (1998)
8. Zhang, L., Zhang, B.: Theory of fuzzy quotient space (methods of fuzzy granular computing). Journal of Software (in Chinese) 14(4), 770–776 (2003)
9. Klir, G.J.: Basic issues of computing with granular computing. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems, pp. 101–105. IEEE Computer Society Press, Los Alamitos (1998)
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), 37–46 (2004)
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), 641–654 (2006)
12. Liu, Q.: Granules and applications of granular computing in logical reasoning. Journal of Computer Research and Development (in Chinese) 41(4), 546–551 (2004)
13. Liang, J.Y., Qian, Y.H, Chu, C.Y., Li, D.Y., Wang, J.H.: Rough set approximation based on dynamic granulation. In: Ślezak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 701–708. Springer, Heidelberg (2005)
14. Kryszkiewicz, M.: Rough set approach to incomplete information system. Information Sciences 112, 39–49 (1998)

15. Kryszkiewicz, M.: Rule in incomplete information systems. Information Sciences 113, 271–292 (1999)
16. Kryszkiewicz, M.: Comparative study of alternative type of knowledge reduction in incomsistent systems. International Journal of Intelligent systems 16, 105–120 (2001)
17. Leung, Y., Li, D.Y.: Maximal consistent block technique for rule acquisition in incomplete information systems. Information Sciences 153, 85–106 (2003)
18. Pawlak, Z.: Rough sets. Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991)
19. Liang, J.Y., Li, D.Y.: Uncertainty and knowledge acquisition in information systems (in Chinese). Science Press, Beijing (2005)