Rough Set Approximation Based on Dynamic Granulation

Jiye Liang[1], Yuhua Qian[1], Chengyuan Chu[2], Deyu Li[1], and Junhong Wang[1]

[1] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, People's Republic of China
[2] Institute of Computing Technology, the Chinese Academy of Sciences, Beijing, 100080, People's Republic of China
ljy@sxu.edu.cn, jinchengqyh@126.com

Abstract. In this paper, the concept of a granulation order is proposed in an information system. The positive approximation of a set under a granulation order is defined. Some properties of positive approximation are obtained. For a set of the universe in an information system, 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 the user requirements. An algorithm based on positive approximation is designed for decision rule mining, and its application is illustrated by an example.

1 Introduction

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

In 1979, the problem of fuzzy information granule was introduced by L.A. 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. 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 [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

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3641, pp. 701–708, 2005.

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 general 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 general 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 space. The positive approximation extend classical rough set, enrich rough set theory and its application. The paper is organized as follows: in section 2, the concepts of a granulation order and the positive approximation under it 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 for decision rule mining, The algorithm will be helping for understanding the idea of positive approximation; in section 4, we show how the algorithm MABPA works by the example.

2 Positive Approximation

Let $S = (U, A)$ be an information system, $P, Q \in 2^A$ two attribute subsets. By $IND(P)$ and $IND(Q)$, we denote the indiscernible relation induced by P and Q. we define a partial relation $\preceq$ on 2^A as follows: $P \preceq Q$ ($Q \succeq P$) if and only if, for every $P_i \in U/IND(P)$, there exists $Q_j \in U/IND(Q)$ such that $P_i \subseteq Q_j$, where $U/IND(P) = \{P_1, P_2, ..., P_m\}$ and $U/IND(Q) = \{Q_1, Q_2, ..., Q_n\}$ are partitions induced by $IND(P)$ and $IND(Q)$.

Let $S = (U, A)$ be an information system, X a subset of U and $P \subseteq A$ an attribute set. In rough set theory, X is characterized by $\overline{P}(X)$ and $\underline{P}(X)$, where

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

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

In an information system, a partition $U/IND(R)$ of U induced by the equivalence relation $IND(R)$, $R \in 2^A$, provides a granulation world for describing a concept X. So a sequence of attribute sets $R_i \in 2^A$ ($i = 1, 2, ..., n$) with $R_1 \succeq R_2 \succeq ... \succeq R_n$ can determine a sequence of granulation worlds, from the most rough 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 information system, X a subset of U and $P = \{R_1, R_2, ..., R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq ... \succeq R_n$ ($R_i \in 2^A$), we define P-upper approximation $\overline{P}X$ and P-lower approximation $\underline{P}X$ of X as follows:*

$$\overline{P}X = \overline{R_n}X, \tag{3}$$

$$\underline{P}X = \bigcup_{i=1}^{n} \underline{R_i}X_i, \tag{4}$$

where $X_1 = X$ and $X_i = X - \bigcup_{k=1}^{i-1} \underline{R_k}X_k$, for $i = 2, ..., n$.

$bn_P(X) = \overline{P}X - \underline{P}X$ is called P-boundary region of X, $pos_P(X) = \underline{P}X$ is called P-positive region of X, and $neg_P(X) = U - \overline{P}X$ is called P-negative region of X. Obviously, we have $\overline{P}X = pos_P(X) \cup bn_P(X)$.

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

Theorem 1. *Let $S = (U, A)$ be an information system, X a subset of U and $P = \{R_1, R_2, ..., R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq ... \succeq R_n$ ($R_i \in 2^A$). Let $P_i = \{R_1, R_2, ..., R_i\}$. Then for $\forall P_i$ ($i = 1, 2, ..., n$), we have*

$$\underline{P_i}(X) \subseteq X \subseteq \overline{P_i}(X), \tag{5}$$

$$\underline{P_1}(X) \subseteq \underline{P_2}(X) \subseteq ... \subseteq \underline{P_n}(X). \tag{6}$$

Proof. The proof follows directly from Definition 1.

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

In [14] , the approximation measure $\alpha_R(X)$ was originally introduced by Z. Pawlak for classical lower and upper approximation, where $\alpha_R(X) = \frac{|\underline{R}X|}{|\overline{R}X|}(X \neq)$. Here we introduce the concept to the positive approximation in order to describe the uncertainty of concept under a granulation order.

Definition 2. *Let $S = (U, A)$ be an information system, X a subset of U and $P = \{R_1, R_2, ..., R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq ... \succeq R_n$ ($R_i \in 2^A$). The approximation measure $\alpha_P(X)$ is defined as*

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

where $X \neq$.

Theorem 2. *Let $S = (U, A)$ be an information system, X a subset of U and $P = \{R_1, R_2, ..., R_n\}$ a family of attribute sets with $R_1 \succeq R_2 \succeq ... \succeq R_n$ ($R_i \in 2^A$). Let $P_i = \{R_1, R_2, ..., R_i\}$. Then for $\forall P_i$ ($i = 1, 2, ..., n$), we have*

$$\alpha_{P_1}(X) \leq \alpha_{P_2}(X) \leq ... \leq \alpha_{P_n}(X). \tag{8}$$

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

Theorem 2 states that the approximation measure $\alpha_P(X)$ increases as the granulation order become longer through adding equivalence relation.

3 Application

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 [15, 16].

In this section, we proposed an algorithm for decision rule mining in consistent 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 a consistent decision table [13], where C and D are condition and decision attribute sets respectively, and $C \cap D =$. The positive region of D with respect to C is defined as follows

$$pos_C(D) = \bigcup_{X \in U/D} \underline{C}X. \tag{9}$$

In a decision table $S = (U, C \cup D)$, the significance of $c \in C$ with respect to D is defined as follows [13]:

$$sig^D_{C-\{c\}}(c) = \gamma_C(D) - \gamma_{C-\{c\}}(D), \tag{10}$$

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

$$sig^D_{C'}(c) = \gamma_{C' \cup \{c\}}(D) - \gamma_{C'}(D), \tag{11}$$

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

Algorithm MABPA (mining rules in a consistent decision table)

Input: consistent 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 $core_D(C) \neq$, 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 $U/D = \{Y_1, Y_2, ..., Y_d\}$;
(4) Let $P = \{P_1\}$, $i = 1$, $U^* = U$, $\Gamma = $, $Rule = $;
(5) Compute $U^*/IND(P_i) = \{X_{i1}, X_{i2}, ..., X_{is_i}\}$;
(6) Let $\Gamma^{'} = \{X_k \in U^*/IND(P_i \mid X_k \subseteq Y_j(Y_j \in U/D, j = \{1, 2, ..., d\})\}$. Let $Rule^{'} = $, for $\forall X_k \in \Gamma^{'}$, put $des_{P_i}(X_k) \longrightarrow des_D(Y_j)(Y_j \in U/D, Y_j \supseteq X_k)$ into $Rule^{'}$. Let $Rule = Rule \cup Rule^{'}$, $\Gamma = \Gamma \cup \Gamma^{'}$;
(7) If $\bigcup_{x \in \Gamma} x = U$, go to (8); else, $U^* = U^* - \bigcup_{x \in \Gamma} x$, for $\forall c \in C - P_i$, compute $sig^D_{P_i}(c)$, let $sig^D_{P_i}(c_2) = max\{sig^D_{P_i}(c), c \in C - P_i\}$, $P_{i+1} = P_i \cup \{c_2\}$, let $P = P \cup \{P_{i+1}\}$, $i = i + 1$, 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. By using MABPA algorithm, the time complexity to extract rules 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 $U^*/IND(P_i)$ 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 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 section, we show how the algorithm MABPA works using an example.

4 Case Study

A consistent decision table $S = (U, C \cup D)$ is given by Table 1, where $C = \{a, b, c, d, e\}$ is condition attribute set and $D = \{f\}$ is decision attribute set. By the algorithm MABPA, we can extract decision rules from Table 1. We have:

$$U/C = \{\{1\}, \{2\}, \{3, 11\}, \{4\}, \{5\}, \{6\}, \{7, 12\}, \{8\}, \{9\}, \{10\}\},$$

$$U/D = \{\{1, 2, 3, 4, 5, 11\}, \{6, 7, 8, 9, 10, 12\}\}.$$

According to the formula $sig^D_{C-\{c\}}(c) = \gamma_C(D) - \gamma_{C-\{c\}}(D)$, we have

$$sig^D_{C-\{a\}}(a) = sig^D_{C-\{b\}}(b) = sig^D_{C-\{c\}}(c) = sig^D_{C-\{d\}}(d) = sig^D_{C-\{e\}}(e) = 0$$

So we get $core_D(C) = $.

Table 1. Example of a consistent decision table

U	attributes					
	a	b	c	d	e	f
1	3	2	3	0	2	1
2	2	2	3	0	2	1
3	1	0	2	0	1	1
4	3	1	3	0	2	1
5	2	0	3	0	2	1
6	0	0	1	0	0	0
7	3	2	0	1	1	0
8	1	0	1	0	0	0
9	2	0	2	1	1	0
10	1	1	3	1	0	0
11	1	0	2	0	1	1
12	3	2	0	1	1	0

By the formula $\gamma_{C'}(D) = |pos_{C'}(D)|/|U|(C^{'} \subseteq C)$, we have

$$\gamma_{\{a\}}(D) = 1/12, \gamma_{\{b\}} = 0, \gamma_{\{c\}}(D) = 4/12, \gamma_{\{d\}}(D) = 4/12, \gamma_{\{e\}}(D) = 7/12$$

Hence, $P_1 = \{e\}$ and $P = \{P_1\}$. For

$$U/IND(P_1) = \{\{1,2,4,5\}\}, \{3,7,9,11,12\}, \{6,8,10\}\}$$

we get

$$\Gamma = \{\{1,2,4,5\}, \{6,8,10\}\},$$

and

$$Rule = \{r_1 : des_{\{e\}}(\{1,2,4,5\}) \rightarrow des_D(\{1,2,3,4,5,11\}),\\ r_2 : des_{\{e\}}(6,8,10) \rightarrow des_D(\{6,7,8,9,10,12\})\}.$$

For

$$\bigcup_{x \in \Gamma} x = \{1,2,4,5,6,8,10\} \neq U,$$

we need to compute significance of the rest of attributes a, b, c, d with respect to D. By the formula for $sig^D_{C'}(c)$, we obtain

$$sig^D_{\{a\}\cup\{e\}}(a) = \gamma_{\{a\}\cup\{e\}}(D) - \gamma_{\{e\}}(D) = 5/12,$$

$$sig^D_{\{b\}\cup\{e\}}(b) = \gamma_{\{b\}\cup\{e\}}(D) - \gamma_{\{e\}}(D) = 2/12,$$

$$sig^D_{\{c\}\cup\{e\}}(c) = \gamma_{\{c\}\cup\{e\}}(D) - \gamma_{\{e\}}(D) = 2/12,$$

$$sig^D_{\{d\}\cup\{e\}}(d) = \gamma_{\{d\}\cup\{e\}}(D) - \gamma_{\{e\}}(D) = 5/12.$$

So we can choose a as c_2 (see the step (7) in the algorithm MABPA). Then, we have $P_2 = \{a, e\}, P = \{P_1, P_2\}$ and $U^* = \{3,7,9,11,12\}$. For

$$U^*/IND(P_2) = \{\{3,11\}, \{7,12\}, \{9\}\},$$

we get

$$\Gamma = \{\{1,2,4,5\},\{3,11\},\{6,8,10\},\{7,12\},\{9\}\}$$

and

$$\begin{aligned} Rule = \{r_1 : des_{\{e\}}(\{1,2,4,5\}) \rightarrow des_D(\{1,2,3,4,5,11\}), \\ r_2 : des_{\{e\}}(\{6,8,10\}) \rightarrow des_D(\{6,7,8,9,10,12\}), \\ r_3 : des_{\{a,e\}}(\{3,11\}) \rightarrow des_D(\{1,2,3,4,5,11\}), \\ r_4 : des_{\{a,e\}}(\{7,12\}) \rightarrow des_D(\{6,7,8,9,10,12\}), \\ r_5 : des_{\{a,e\}}(\{9\}) \rightarrow des_D(\{6,7,8,9,10,12\})\}. \end{aligned}$$

It is easy to see $\bigcup_{x\in\Gamma} x = U$. So the algorithm MABPA is ended, and *Rule* is obtained. For intuition, the five decision rules obtained by MABPA from the decision table S are listed in Table 2.

Table 2. Rules obtained for the decision table S

Rule	attributes		
	a	e	f
r_1		2	1
r_2	1	1	1
r_3		0	0
r_4	3	1	0
r_5	2	1	0

This example shows the mechanism of the decision rule mining algorithm based on positive approximation.

5 Conclusions

In this paper, we extend rough set approximation under static granulation to rough set approximation under dynamic granulation, the positive approximation is defined and its some properties are obtained. A target concept can be approached by the change of the positive approximation. An algorithm based on positive approximation for decision rule mining is given, and its application is illustrated by an illustrative example. The results obtained in this paper will play an important role in further research on rough set approximation and granular computing.

Acknowledgements. This work was supported by the national 863 plan project of China (No. 2004AA115460), the national natural science foundation of China (No. 70471003, No. 60275019), the natural science foundation of Shanxi, China (No. 20031036, No. 20041040) and the top scholar foundation of Shanxi, China.

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 (to appear)
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. Zhang, W.X., Wu, W.Z., Liang, J.Y., Li, D.Y.: Theory and method of rough sets (in Chinese). Beijing, Science Press (2001)
14. Pawlak, Z.: Rough sets. Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991)
15. 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.
16. 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.