

变精度多粒度粗糙集近似集更新的矩阵算法

郑文彬^{1,2*}, 李进金³, 于佩秋³, 林艺东⁴

(1.闽南师范大学 计算机学院, 福建 漳州 363000; 2.福建省粒计算及其应用重点实验室(闽南师范大学), 福建 漳州 363000;

3.闽南师范大学 数学与统计学院, 福建 漳州 363000; 4.厦门大学 数学科学学院, 福建 厦门 361005)

(*通信作者电子邮箱 zznxzwb@126.com)

摘要: 随着信息大爆炸时代的到来,数据集的巨大化、数据集结构的复杂化已经成为近似计算中一个不能忽视的问题。动态计算是解决这些问题的一种行之有效的途径。本文对现有的应用于经典多粒度粗糙集动态近似集更新方法进行了改进,提出了应用于变精度多粒度粗糙集的向量矩阵近似计算与更新方法。首先,提出了一种基于向量矩阵的变精度多粒度粗糙集的近似集静态计算算法;其次,重新考虑了变精度多粒度粗糙集近似集更新时的搜索区域,并根据变精度多粒度粗糙集的性质缩小了该区域,这能有效地提升近似集更新算法的时间效率;再次,根据新的搜索区域,在变精度多粒度粗糙集近似集静态计算算法的基础上,提出了一种新的变精度多粒度粗糙集近似集更新的向量矩阵算法;最后,进行实验验证了本文提出的算法的有效性。

关键词: 动态计算;近似集更新;变精度多粒度粗糙集;矩阵算法;数据挖掘**中图分类号:** TP18**文献标志码:** A

Matrix-based algorithm for updating approximations in variable precision multigranulation rough sets

ZHENG Wenbin^{1,2*}, LI Jinjin³, YU Peiqiu³, LIN Yidong⁴

(1.School of Computer Science Minnan Normal University, Zhangzhou Fujian 363000, China;

2.Lab of Granular Computing Minnan Normal University, Zhangzhou Fujian 363000, China;

3.School of Mathematics and Statistics Minnan Normal University, Zhangzhou Fujian 363000, China;

4.School of Mathematical Sciences Xiamen University, Xiamen Fujian 361005, China)

Abstract: In an information explosion era, data sets containing information could be huge and complex that sometimes are difficult to handle. Dynamic computing is an efficient approach to solve these problems. The purpose of this study is to propose approaches in updating approximations in variable precision multigranulation rough sets (VPMGRS) while adding and deleting attributes. First, static algorithm for computing approximations based on matrix for VPMGRS is presented in this paper. Second, searching regions in data sets for updating approximations in VPMGRS are shrunk. Third, matrix-based approaches for updating approximations in VPMGRS are proposed. The incremental algorithms for updating approximations in VPMGRS are then designed. Finally, the efficiency and validity of the designed algorithms are verified by experiments.

Keywords: dynamic computing; updating approximations; variable precision multi-granulation rough sets; matrix algorithm; data mining

0 引言

自粗糙集^[1]模型于 1982 年被波兰学者 Pawlak 提出以来,粗糙集模型被广泛应用于各种领域,例如模式识别^[2]、机器学习^[3]、图像处理^[4]、数据挖掘^[5]等。为了拓展粗糙集模

型的应用范围,学者提出了许多粗糙集模型的拓展模型,例如覆盖粗糙集^[6]、变精度粗糙集^[7]、概率粗糙集^[8]、多粒度粗糙集^[9]、变精度多粒度粗糙集^[10]等。

变精度多粒度粗糙集模型是窦慧莉等人于 2012 年提出的,该模型进一步拓展了多粒度粗糙集模型的应用前景。变

收稿日期: 2019-05-06; 修回日期: 2019-06-05; 录用日期: 2019-06-17。

基金项目: 国家自然科学基金资助项目(61379021, 11871259); 国家自然科学基金青年项目(11701258)。

作者简介: 郑文彬(1971—),男,福建仙游人,高级讲师,硕士,主要研究方向:粗糙集、粒计算、数据挖掘、人工智能;李进金(1960—),男,福建晋江人,博士,教授,主要研究方向:人工智能、粒计算、拓扑学;于佩秋(1991—),男,内蒙古赤峰人,硕士研究生,主要研究方向:粗糙集、粒计算、数据挖掘和人工智能;林艺东(1989—),男,福建漳州人,博士研究生,主要研究方向:不确定性理论。

精度多粒度粗糙集是一种强大的数学工具,在许多多粒度视角下的现实生活场景中具有非常广泛的应用和非常强大的表示能力.然而自变精度多粒度粗糙集被提出以来,较少有学者进行变精度多粒度粗糙集近似集计算与近似集更新方面的研究.不论我们应用何种粗糙集模型,包括且不限于变精度多粒度粗糙集模型,近似集计算都是应用中必要且重要的一环;例如在属性约简时必须计算正域,在规则提取时必须计算下近似集和上近似集.在数据大爆炸时代,近似集计算开始变得越来越困难,数据集的巨大化使得有时我们不可能在整个数据集上进行计算,数据集的复杂化使得数据的结构也经常变化,属性增减,属性值改变,样本增多与减少在近似集计算时变得非常普遍,针对这些问题,在各种改变的场景区下动态计算近似集是行之有效的解决方案.计算和更新多粒度粗糙集及其拓展模型的上下近似集引起众多研究者的广泛关注,这些研究通常分为四类,即如何在增减属性时更新近似集^[11],如何在改变属性值时更新上下近似集^[12],如何在决策属性改变时更新上下近似集^[13],如何在论域中对象变化时更新上下近似集^[14].

本文尝试提出一种在属性增加和减少时计算和更新变精度多粒度粗糙集上下近似集的新方法.首先提出了一种应用于变精度多粒度粗糙集的向量矩阵近似集静态计算方法;其次,根据变精度多粒度粗糙集的性质,缩小了更新上下近似集的搜索区域,更小的搜索区域意味着更少的计算时间.再次,根据新的搜索区域,再提出了在属性增减时基于矩阵向量的变精度多粒度粗糙集近似集更新算法.最后,采用了几个UCI (<http://archive.ics.uci.edu/ml/index.php>)数据集,进行实验验证了所提出的算法的有效性.

1 预备知识

在本节,列举所有与本文工作相关的定义与定理.

定义 1^[15] 设集合 A 与 B 是论域 U 的非空子集,定义集合 A 相对于集合 B 的相对正确分类率为:

$$P(A, B) = \frac{|A \cap B|}{|A|}.$$

其中 $|A|$ 代表集合 A 的基数.

定义 2^[9] 令 $S = (U, AT, V, f)$ 为一个信息系统,其中 $U = \{x_1, x_2, \dots, x_n\}$ 为非空有限论域, $AT = \{A_1, A_2, \dots, A_m\}$ 为非空有限属性集族, $A_k \in AT$ 是一个属性集. $V = \bigcup_{A \in AT} V_A$ 为属性值的值域, V_A 为属性集 A 的值域. $f: U \times AT \rightarrow V$ 为一个决策函数使得 $f(x, A) \in V_A$ 对任意的 $A \in AT, x \in U$ 都成立.

定义 3^[10] (变精度多粒度乐观粗糙集, OVPMGRS) 令 $S = (U, AT, V, f)$ 为一个信息系统.对于任意的 $X \subseteq U$, X 的可变精度多粒度乐观下近似集与乐观上近似集分别表示为:

$$\sum_{k=1}^m A_k^\beta (X) = \left\{ \begin{array}{l} x \in U \mid P([x]_{A_1}, X) \geq \beta \vee P([x]_{A_2}, X) \\ \geq \beta \vee \dots \vee P([x]_{A_m}, X) \geq \beta \end{array} \right\},$$

$$\overline{\sum_{k=1}^m A_k^\beta (X)} = \sim \sum_{k=1}^m A_k^\beta (\sim X)$$

其中 $\sim X$ 表示 X 的补集, $\beta \in (0.5, 1]$.

由对偶性容易得到以下定理:

定理 1^[10] 令 $S = (U, AT, V, f)$ 为一个信息系统.对于任意的 $X \subseteq U$, X 的可变精度多粒度乐观上近似集表示为

$$\overline{\sum_{k=1}^m A_k^\beta (X)}$$

则有:

$$\sum_{k=1}^m A_k^\beta (X) = \left\{ \begin{array}{l} x \in U \mid P([x]_{A_1}, X) > 1 - \beta \wedge P([x]_{A_2}, X) \\ 1 - \beta \wedge \dots \wedge P([x]_{A_m}, X) > 1 - \beta \end{array} \right\}.$$

定义 4^[10] (变精度多粒度悲观粗糙集 PVPMGRS) 令 $S = (U, AT, V, f)$ 为一个信息系统.对于任意的 $X \subseteq U$, X 的可变精度悲观下近似集与上近似集分别表示为:

$$\sum_{k=1}^m A_k^\beta (X) = \left\{ \begin{array}{l} x \in U \mid P([x]_{A_1}, X) \geq \beta \wedge P([x]_{A_2}, X) \\ \geq \beta \wedge \dots \wedge P([x]_{A_m}, X) \geq \beta \end{array} \right\},$$

$$\overline{\sum_{k=1}^m A_k^\beta (X)} = \sim \sum_{k=1}^m A_k^\beta (\sim X).$$

由对偶性容易得到以下定理:

定理 2^[10] 令 $S = (U, AT, V, f)$ 为一个信息系统. 对于任意的 $X \subseteq U$, X 的可变精度多粒度悲观上近似集表示为

$$\overline{\sum_{k=1}^m A_k^\beta} (X), \text{ 则有:}$$

$$\overline{\sum_{k=1}^m A_k^\beta} (X) = \left\{ \begin{array}{l} x \in U \mid P([x]_{A_1}, X) > 1 - \beta \vee P([x]_{A_2}, X) \\ > 1 - \beta \vee \dots \vee P([x]_{A_m}, X) > 1 - \beta \end{array} \right\}.$$

定义 5^[6] 令 $S = (U, AT, V, f)$ 为一个信息系统. 对于任意的 $X \subseteq U$, X 的向量矩阵表示为 $G(X) =$

$$[g_1(X), g_2(X), \dots, g_{|U|}(X)]^T, \text{ 其中}$$

$$g_i(X) = \begin{cases} 1, & x_i \in X, \\ 0 & x_i \notin X, \end{cases}$$

“ T ”表示矩阵的转置。

定理 3 对于任意的集合 $X, Y \subseteq U$, X 相对于 Y 的相对正确分类率为:

$$P(X, Y) = \frac{[G(X)]^T \cdot G(Y)}{\sum_{i=1}^{|U|} g_i(X)},$$

其中“ \cdot ”为矩阵乘法。

证明:

$$\begin{aligned} \forall x_i \in X \cap Y &\Leftrightarrow g_i(X) = 1 \wedge g_i(Y) = 1 \Leftrightarrow |X \cap Y| \\ &= [G(X)]^T \cdot G(Y), \text{ 综上所述, } P(X, Y) = \frac{[G(X)]^T \cdot G(Y)}{\sum_{i=1}^{|U|} g_i(X)}. \end{aligned}$$

2 基于矩阵的近似集静态计算算法

为了对变精度多粒度粗糙集近似集更新方法进行研究, 本文首先对其近似集计算方法进行研究, 给出静态的变精度多粒度粗糙集上下近似集计算算法。

定义 6 令 $S = (U, AT, V, f)$ 为一个信息系统. 对于任意的 $X \subseteq U$, X 在属性集 A_k 上的变精度多粒度近似集特征向量定义为 $H^{A_k}(X) = [h_1^{A_k}(X), h_2^{A_k}(X), \dots, h_{|U|}^{A_k}(X)]^T$, 其中:

$$h_i^{A_k}(X) = P([x_i]_{A_k}, X), \forall k \in \{1, 2, \dots, m\}.$$

显然, 当 $[x_i]_{A_k} \cap X = \emptyset$ 时, 有 $h_i^{A_k}(X) = 0$, 即如果能

够确保 $[x_i]_{A_k} \cap X = \emptyset$, 则可以直接令 $h_i^{A_k}(X) = 0$ 。

定义 7 令 $S = (U, AT, V, f)$ 为一个信息系统. 对于任

意的 $X \subseteq U$, 定义对于向量 $H^{A_k}(X)$ 两个截向量

$$H_{\geq \alpha}^{A_k}(X) = [h_1^{A_k}(X), h_2^{A_k}(X), \dots, h_{|U|}^{A_k}(X)]^T \quad \text{与}$$

$$H_{> \alpha}^{A_k}(X) = [\lambda_1^{A_k}(X), \lambda_2^{A_k}(X), \dots, \lambda_{|U|}^{A_k}(X)]^T, \text{ 其中}$$

$$h_i^{A_k}(X) = \begin{cases} 0 & h_i^{A_k} < \alpha, \\ 1 & h_i^{A_k} \geq \alpha, \end{cases}$$

$$\lambda_i^{A_k}(X) = \begin{cases} 0 & h_i^{A_k} \leq \alpha, \\ 1 & h_i^{A_k} > \alpha, \end{cases}$$

$\forall k \in \{1, 2, \dots, m\}, \forall i \in \{1, 2, \dots, |U|\}, \alpha \in R$ 为任一实

数。

定理 4 令 $S = (U, AT, V, f)$ 为一个信息系统. 对于任意的 $X \subseteq U$, X 的变精度多粒度乐观下近似集与乐观上近似

集分别表示为 $\underline{\sum_{k=1}^m A_k^\beta} (X)$ 和 $\overline{\sum_{k=1}^m A_k^\beta} (X)$, 则有:

$$(1) \underline{\sum_{k=1}^m A_k^\beta} (X) = \bigvee_{k=1}^m H_{\geq \beta}^{A_k}(X),$$

$$(2) \overline{\sum_{k=1}^m A_k^\beta} (X) = \bigwedge_{k=1}^m H_{> 1-\beta}^{A_k}(X).$$

证明: (1)

$$\begin{aligned} \forall x_i \in \underline{\sum_{k=1}^m A_k^\beta} (X) &\Leftrightarrow g_i(\underline{\sum_{k=1}^m A_k^\beta} (X)) = 1 \Leftrightarrow \exists k \in \\ &\{1, 2, \dots, m\}, \text{ s.t. } P([x]_{A_k}, X) \geq \beta \Leftrightarrow \bigvee_{k=1}^m h_i^{A_k}(X) \geq \\ &\beta \Leftrightarrow \bigvee_{k=1}^m h_i^{A_k}(X) = 1. \end{aligned}$$

综上所述, $\underline{\sum_{k=1}^m A_k^\beta} (X) = \bigvee_{k=1}^m H_{\geq \beta}^{A_k}(X)$ 。

(2)证明过程类似于(1)。

算法 1 基于矩阵的变精度多粒度粗糙集上下近似集算法

输入: (1) $S = (U, AT, V, f)$, (2) $G([x_i]_{A_k}), \forall k \in \{1, 2, \dots, m\}, \forall i \in \{1, 2, \dots, |U|\}$, (3) $G(X)$ (4) β 。

输出: $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 和 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$, $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 和 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$.

```

1   $V_{A_k} \leftarrow [0, 0, \dots, 0]^T$ 
2   $n \leftarrow |U|$ 
3  For  $i = 1 \rightarrow n$ 
4    For  $j = 1 \rightarrow n$ 
5      For  $k = 1 \rightarrow m$  #  $m = |AT|$ 
6        If  $g_i(X) = 1 \wedge g_i([x_j]_{A_k}) = 1$ 
           #保证  $[x_j]_{A_k} \cap X \neq \emptyset$ 
7           $h_j^{A_k}(X) \leftarrow \frac{[G([x_j]_{A_k})]^T \cdot G(X)}{\sum_{i=1}^{|U|} g_i([x_j]_{A_k})}$ 
8        End If
9      End For
10     End For
11   End For
12    $G(\sum_{k=1}^m \overline{A_k^\beta}^O(X)) \leftarrow \vee_{k=1}^m H_{\geq \beta}^{A_k}(X)$ 
13    $G(\sum_{k=1}^m \overline{A_k^\beta}^O(X)) \leftarrow \wedge_{k=1}^m H_{> 1-\beta}^{A_k}(X)$ 
14    $G(\sum_{k=1}^m \underline{A_k^\beta}^P(X)) \leftarrow \wedge_{k=1}^m H_{\geq \beta}^{A_k}(X)$ 
15    $G(\sum_{k=1}^m \underline{A_k^\beta}^P(X)) \leftarrow \vee_{k=1}^m H_{> 1-\beta}^{A_k}(X)$ 
16   Return  $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$  和  $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ ,  $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$  和  $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 

```

定理 5 令 $S = (U, AT, V, f)$ 为一个信息系统。对于任意的 $X \subseteq U$, X 的变精度多粒度悲观下近似集与上近似集分

别表示为 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 和 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$, 则有:

$$(1) G(\sum_{k=1}^m \underline{A_k^\beta}^P(X)) = \wedge_{k=1}^m H_{\geq \beta}^{A_k}(X),$$

$$(2) G(\sum_{k=1}^m \overline{A_k^\beta}^O(X)) = \vee_{k=1}^m H_{> 1-\beta}^{A_k}(X)。$$

证明: 证明过程类似于定理 4。

由定理 4 与定理 5 可以得到算法 1, 即为基于矩阵的变精度多粒度粗糙集上下近似集算法, 若

$$R_{A_k}(X) = \{x | [x]_{A_k} \cap X \neq \emptyset\},$$
 则该算法的时间复

杂度为 $O(\sum_{k=1}^m |R_{A_k}(X)| |U|)$ 。3-11 步是用来计算

$H^{A_k}(X)$ 的, 时间复杂度为 $O(\sum_{k=1}^m |R_{A_k}(X)| |U|)$;

12-15 步是用来计算 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 、 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 、

$\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 、 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 的, 时间复杂度为 $O(m|U|)$ 。

3 基于矩阵的近似集更新方法

基于本文给出的变精度多粒度粗糙集近似集静态计算方法, 提出了基于矩阵的变精度多粒度粗糙集近似集更新算法。

3.1 添加属性时的近似集更新算法

在本节中, 用 $S^t = (U, AT, V, f)$ 表示一个 t 时刻的信息系统, 其中 $AT = \{A_1, A_2, \dots, A_m\}$ 为非空有限属性集族, 用

$S^{t+1} = (U, AT, \mathcal{V}, \mathcal{F})$ 表示一个 $t+1$ 时刻的信息系统, 其中

$AT = \{A_1, A_2, \dots, A_m\}$ 为非空有限属性集族。用 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$

和 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 表示 t 时刻 OVPMGRS 的上下近似集, 用

$\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 和 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 表示 t 时刻 PVPMGRS 的上下近

似集, A_k 是一个属性集, $\beta \in (0.5, 1]$ 。用 $\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 和

$\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 表示 $t+1$ 时刻 OVPMGRS 的上下近似集, 用

$\sum_{k=1}^m \overline{A_k^\beta}^O(X)$ 和 $\sum_{k=1}^m \underline{A_k^\beta}^P(X)$ 表示 $t+1$ 时刻 PVPMGRS 的上

下近似集, \mathcal{A}_k 为属性集且满足 $\forall \mathcal{A}_k \in AT, \exists \mathcal{A}_k \in \mathcal{AT}$ 使得

$A_k \subseteq \mathcal{A}_k$, 即从 t 时刻到 $t+1$ 时刻属性增加了。

定理 6 令 $S^t = (U, AT, V, f)$ 为一个 t 时刻的信息系统, $S^{t+1} = (U, \mathcal{AT}, \mathcal{V}, \mathcal{F})$ 为一个 $t+1$ 时刻的信息系统。

$H^{A_k}(X)$ 为 X 在属性集 A_k 上的变精度多粒度近似特征向量, $H^{\mathcal{A}_k}(X)$ 为 X 在属性集 \mathcal{A}_k 上的变精度多粒度近似集特征向量, 则对于任意的 $k \in \{1, 2, \dots, m\}$, X 有以下性质成立:

$$(1) h_i^{A_k}(X) = 1 \Rightarrow h_i^{\mathcal{A}_k}(X) = 1,$$

$$(2) h_i^{A_k}(X) = 0 \Rightarrow h_i^{\mathcal{A}_k}(X) = 0.$$

证明:

$$(1) h_i^{A_k}(X) = 1 \Leftrightarrow P([x_i]_{A_k}, X) = 1 \Leftrightarrow [x_i]_{A_k} \subseteq X$$

$$\Rightarrow [x_i]_{\mathcal{A}_k} \subseteq X \Leftrightarrow h_i^{\mathcal{A}_k}(X) = 1.$$

(2) 证明类似于 (1)

定理 6 说明, 在向属性集中添加属性时, 计算 $H^{A_k}(X)$

可以不计算 $H^{A_k}(X)$ 中 0 和为 1 的那些位, 因为由 $[x]_{A_k} \subseteq [x]_{\mathcal{A}_k}$, $\forall x \in U$ 都成立, 有 $H^{A_k}(X)$ 中为 0 和 1 的位在 $H^{\mathcal{A}_k}(X)$ 中也不变。

算法 2 基于矩阵的变精度多粒度粗糙集上下近似集更新算法

输入: (1) $S = (U, AT, V, f)$, (2) $G([x_i]_{A_k}), \forall k \in \{1, 2, \dots, m\}, \forall i \in \{1, 2, \dots, |U|\}$,

(3) $G(X)$, (4) β , (5) $H^A(X)$.

输出: $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 和 $\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$, $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 和 $\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$.

```

1   $H^A(X) \leftarrow H^A(X)$ 
2   $n \leftarrow |U|$ 
3  For  $i = 1 \rightarrow n$ 
4      For  $k = 1 \rightarrow m$  #  $m = |AT|$ 
5          If  $h_i^{A_k}(X) \in (0, 1)$ 

```

$$6 \quad h_i^{A_k}(X) \leftarrow \frac{[G([x_i]_{A_k})]^T \cdot G(X)}{\sum_{i=1}^{|U|} g_i([x_i]_{A_k})}$$

7 **End If**

8 **End For**

9 **End For**

$$10 \quad G(\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)) \leftarrow \vee_{k=1}^m H_{\geq \beta}^{A_k}(X)$$

$$11 \quad G(\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)) \leftarrow \wedge_{k=1}^m H_{> 1-\beta}^{A_k}(X)$$

$$12 \quad G(\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)) \leftarrow \wedge_{k=1}^m H_{\geq \beta}^{A_k}(X)$$

$$13 \quad G(\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)) \leftarrow \vee_{k=1}^m H_{> 1-\beta}^{A_k}(X)$$

14 **Return** $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 和 $\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$, $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 和 $\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$

由定理 6 我们可以得到变精度多粒度粗糙集上下近似集

更新算法, 因为只对 $H^{A_k}(X)$ 中处于 $(0, 1)$ 区间内的元素进行更新, 若 $R_{A_k}^{\beta}(X) = \{x | [x]_{A_k} \cap X \neq \emptyset \wedge [x]_{A_k} \not\subseteq X\}$, 则

算法 2 的时间复杂度为 $O(\sum_{k=1}^m |R_{A_k}^{\beta}(X)| |U|)$ 。3-9 步是用来

更新 $H^{A_k}(X)$ 的, 时间复杂度为 $O(\sum_{k=1}^m |R_{A_k}^{\beta}(X)| |U|)$; 10-13

步是用来更新 $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 、 $\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$ 、 $\sum_{k=1}^m \overline{\mathcal{A}_k^{\beta}}(X)$ 、

$\sum_{k=1}^m \underline{\mathcal{A}_k^{\beta}}(X)$ 的, 时间复杂度为 $O(m|U|)$ 。显然有

$R_{A_k}^{\beta}(X) \subseteq R_{A_k}(X)$ 即最坏的情况下动态更新算法的时间复杂度才会等于静态计算算法。

3.2 减少属性时的近似集更新

下面将举出反例说明属性减少时, 定理 6 将不再成立。

例 1 一个多粒度信息系统如表 1 所示: 令

$$X = \{x_2, x_5, x_6\}, A_1 = \{a_1, a_2, a_3\}, A_2 = \{a_4, a_3, a_6\},$$

$$A_3 = \{a_7, a_8, a_9\}, A_4 = \{a_2\}, A_5 = \{a_4\}, A_6 = \{a_9\},$$

由定义 6

知

$h_2^{A_1}(X)=1$, $h_2^{A_2}(X)=1/2 < 1$; $h_1^{A_3}(X)=0$, $h_1^{A_4}(X)=1/2 \neq 0$. 所以当属性减少时定理 6 不再成立。

例 1 说明了当属性减少时不能使用本文所提出的近似集更新方法进行上下近似集更新, 当本文的动态算法不能用于属性减少时, 可以用本文提出的静态算法来计算变精度多粒度粗糙集上下近似集。

表 1 一个多粒度信息系统

Tab. 1 A multigranulation information system

	A_1			A_2			A_3		
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
x_1	1	2	3	1	1	2	2	1	1
x_2	3	2	1	1	1	2	1	3	1
x_3	1	3	1	2	3	1	2	1	1
x_4	1	2	3	2	3	1	1	1	1
x_5	3	2	1	1	3	2	2	2	1
x_6	1	3	1	1	1	1	1	3	1

4 UCI 数据实验分析

为了验证算法的有效性, 本文选取了 8 个 UCI 数据来验证近似集静态计算算法与近似集更新算法的有效性。实验所使用的 UCI 数据集如表 2 所示。可以看到样本数从 194 到 1000, 属性数目从 5 到 59。所有试验均在系统为 64-bit windows 10, CPU 为 Inter(R) Core(TM) i7 6700HQ CPU @2.60GHz, 16GB 内存的个人计算机上进行。所使用的编程语言是 Matlab r2015b。

4.1 不同大小数据集的计算时间对比

表 2 实验使用的 UCI 数据集

Tab. 2 Details of data sets

	数据集	实例数	属性数
1	Blood Trans Fusion	748	5
2	Dermatology	366	20
3	Extention of ZAlizadehsani	303	59
4	Facebookmetrics	500	19
5	Flags	194	30
6	German Credit Data	1000	21
7	Soybean Large	307	36
8	Student	396	33

在本节中, 使 $\beta=0.9$, 将数据集等分成十份

U_1, U_2, \dots, U_{10} , 然后以每一份数据集为增加的步长, 逐步地

使数据集增大, 即 $dataset1=U_1$, $dataset2=U_1 \cup U_2$,

$dataset3=U_1 \cup U_2 \cup U_3 \dots$ 并使 X 大致为每个临时数据集的

0.85 倍大小, X 中元素从每个临时数据集中随机选择。模拟

属性增加时, 对每个数据集, 首先随机选出一个属性 \hat{a} , 然

后将剩余的属性分为大致相等的三组 A'_1, A'_2, A'_3 , 首先用算

法 1 计算 VPMGRS 上下近似集, 然后令

$$A_1^{t+1} = A'_1 \cup \hat{a}, A_2^{t+1} = A'_2 \cup \hat{a}, A_3^{t+1} = A'_3 \cup \hat{a},$$

分别使用算

法 1 (静态算法) 与算法 2 (动态算法) 及对比算法基于关系

矩阵的多粒度粗糙集近似更新算法^[17]与基于关系矩阵的多

粒度粗糙集近似计算算法^[17]分别计算 VPMGRS 上下近似集

并对比计算时间, 得出如图 1 的实验结果。

4.2 不同大小目标概念的计算时间对比

在本节中, 使 $\beta=0.9$, 将数据集等分成十份

U_1, U_2, \dots, U_{10} , 然后以每一份数据集为增加的步长, 逐步地

使 X 增大, 即 $X=U_1$, $X=U_1 \cup U_2$, $X=U_1 \cup U_2 \cup U_3 \dots$

构造和模拟属性增加的过程与 4.1 一致。使用本文所提的两

种算法及两种对比算法分别计算 VPMGRS 上下近似集并对

比计算时间, 得出如图 2 的实验结果。

由图 2 可以得到当数据集大小不变而目标概念的大小增大时，本文的静态算法与动态算法的运行时间波动不大，由图 2 依然可以看出本文的动态算法与静态算法的运行时间具有显著的差异，且动态算法的运行时间波动不大，这是因为动态算法只计算特征向量中不为 0 和 1 的那些位而与目标概念的大小没有直接关系。图 2 显示的效果符合本文的预期，动态算法明显地优于静态算法。

4.3 不同大小 β 的计算时间对比

在本节中，使 β 增大，并使 X 为 U 中随机取出 90% 的元素。构造与模拟属性增加的过程与 4.1 一致。使用所有算法分别计算 VPMGRS 上下近似集并对比计算时间，得出如图 3 的实验结果。

由图 3 可以得到各算法的运行时间几乎不受 β 大小的影响。由图 3 依然有本文的算法运行时间最短，效率最高，且在提供 $H^A(X)$ 时，本文所提的动态更新算法时间效率远高于静态计算算法。

5 结束语

本文首先提出了一种基于矩阵点乘的变精度多粒度粗糙集静态的近似集计算方法，而后基于静态计算方法提出了动态计算的方法。最后对本文所提出的算法进行了对比试验。实验结果表明，文所提出的动态近似集更新算法的时间效率高于本文所提出的静态计算算法与基于关系矩阵的计算、更新算法。

近似集计算与近似集更新是属性约简的重要内容，基于本文所提的算法，可以研究变精度多粒度粗糙集属性约简的方法等。

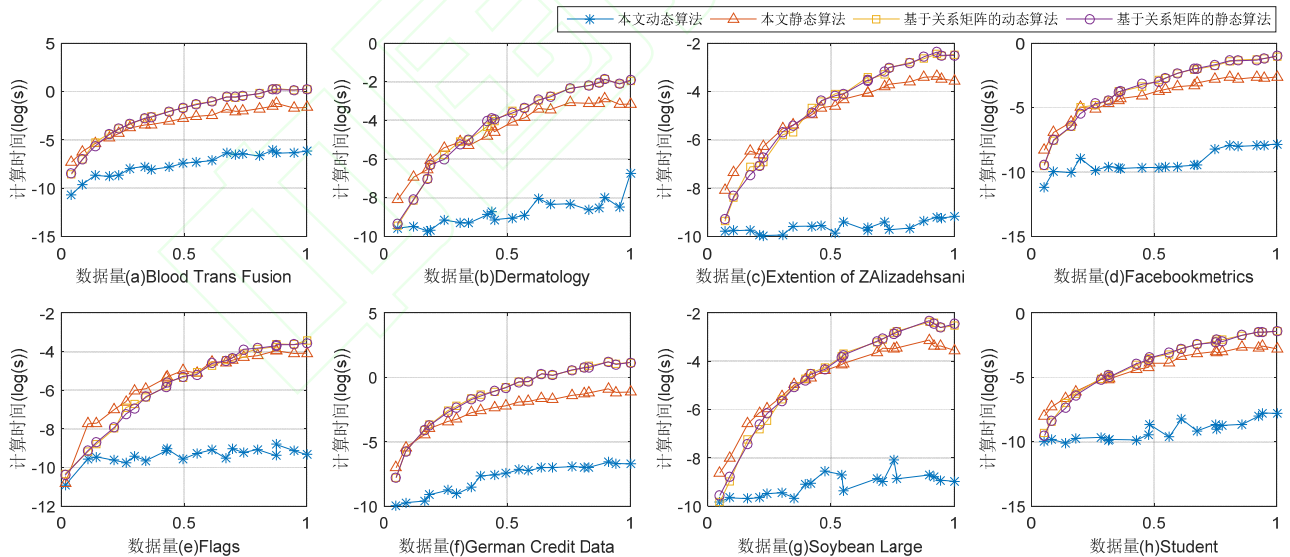


图 1 数据集增大时运行时间对比

Fig. 1 Comparison of computational time while datasets increasing

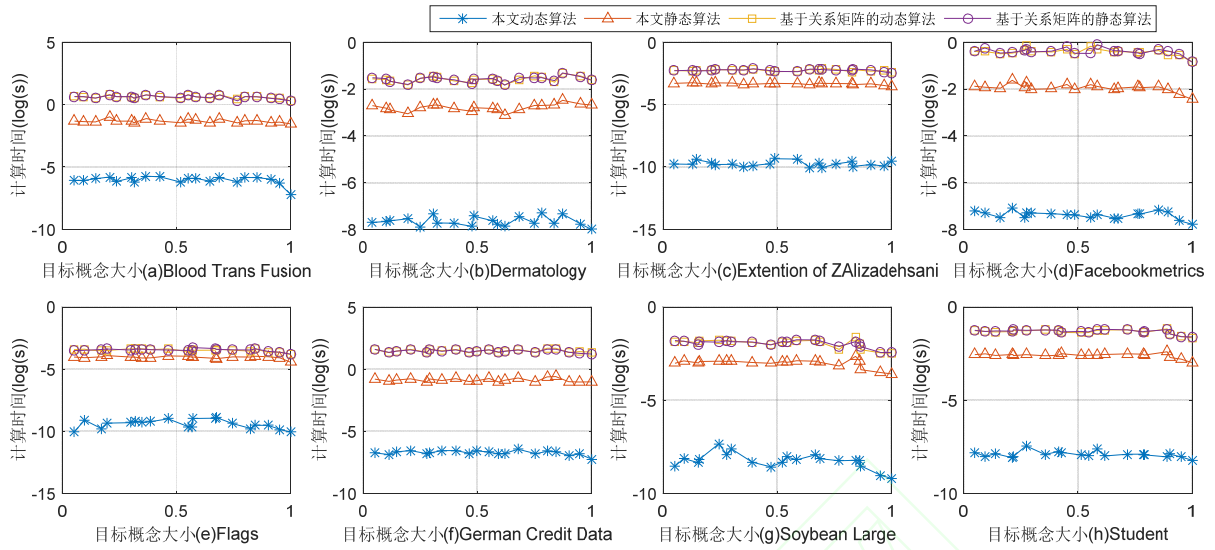
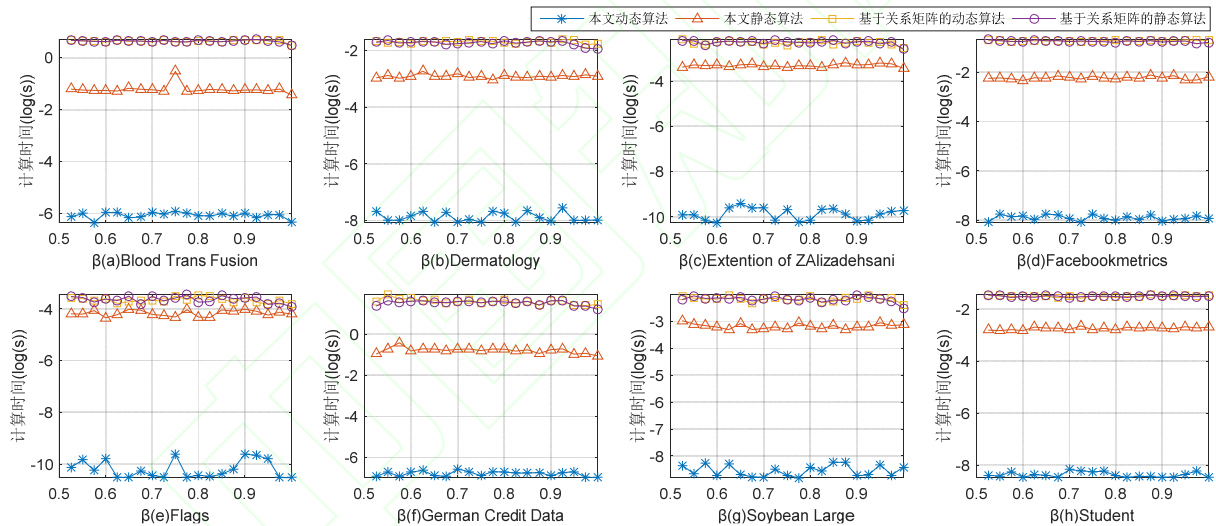


图2 目标概念增大时运行时间对比

Fig. 2 Comparison of computational time while target concept increasing

图3 β 增大时运行时间对比Fig. 3 Comparison of computational time while β increasing

参考文献

- [1] PAWLAK Z. Rough sets [J]. International Journal of Computer & Information Sciences, 1982, 11(5):341-356.
- [2] CHEN W. Evidence of electroconformational changes in membrane proteins: field-induced reductions in intra membrane nonlinear charge movement currents [J]. Bioelectrochemistry, 2004, 63(1):333-335.
- [3] MIN F, HU Q, ZHU W. Feature selection with test cost constraint [J]. International Journal of Approximate Reasoning, 2014, 55(1):167-179.
- [4] 韩建栋, 朱婷婷, 李月香. 结合粗糙集与分层思想的彩色图像分割算法[J]. 计算机应用, 2015, 35(7): 2020-2024.(HAN J D, ZHU T T, LI Y X. Color image segmentation algorithm based on rough-set and hierarchical idea [J]. Journal of Computer Applications, 2015, 35(7): 2020-2024.)
- [5] 陈世清, 唐志航, 肖建华. 基于粗糙集联系度的数据挖掘算法及应用研究[J]. 计算机应用, 2004, 24(6):74-77. (CHEN S Q, TANG Z H, XIAO J H. Algorithm of data mining based on rough sets pair analysis and its application [J]. Journal of Computer Applications, 2004, 24(6): 74-77.)
- [6] YAO Y, YAO B. Covering based rough set approximations [J]. Information Sciences, 2012, 200(1): 91-107.
- [7] KATZBERG J D, ZIARKO W. Variable precision rough sets with asymmetric bounds [M]. Rough Sets, Fuzzy Sets and Knowledge Discovery. Springer, London, 1994: 167-177.

- [8] 马建敏, 姚红娟, 潘笑晨. 串行概率粗糙集近似[J]. 计算机科学, 2018, 45(1): 79-83. (MA J M, YAO H J, PAN X C. Serial probabilistic rough set approximation [J]. Computer Science, 2018, 45 (1): 79-83.)
- [9] QIAN Y, LIANG J, YAO Y, et al. MGRS: A multi-granulation rough set [J]. Information Sciences, 2010, 180(6): 949-970.
- [10] 窦慧莉, 吴陈, 杨习贝, 等. 可变精度多粒度粗糙集模型[J]. 江苏科技大学学报(自然科学版), 2012, 26(1): 65-69. (DOU H L, WU C, YANG X B, et al. Variable precision multi-granularity rough set model [J]. Journal of Jiangsu University of Science and Technology (Natural Science Edition), 2012, 26 (1): 65 - 69.)
- [11] YANG X, et al. Updating multigranulation rough approximations with increasing of granular structures [J]. Knowledge Based Systems, 2014, 64(1): 59-69.
- [12] CHEN H, LI T, LUO C, et al. A rough set-based method for updating decision rules on attribute values' coarsening and refining [J]. IEEE transactions on knowledge and data engineering, 2014, 26(12): 2886-2899.
- [13] CHENG Y. Dynamic maintenance of approximations under fuzzy rough sets [J]. International Journal of Machine Learning and Cybernetics, 2018, 9(12): 2011-2026.
- [14] HU J, LI T, LUO C, et al. Incremental fuzzy probabilistic rough sets over two universes [J]. International Journal of Approximate Reasoning, 2017, 81: 28-48.
- [15] ZIARKO W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59.
- [16] 程燕. 基于矩阵的覆盖粗糙集算法研究[D]. 安徽大学, 2017. (CHENG Y. Study on covering rough set algorithm based on matrix [D]. Anhui University, 2017.)
- [17] HU C, LIU S, LIU G. Matrix-based approaches for dynamic updating approximations in multigranulation rough sets [J]. Knowledge-Based Systems, 2017, 122: 51-63.

This work is partially supported by the National Science Foundation of China (61379021, 11871259), the National Youth Natural Science Foundation of China (11701258).

ZHENG Wenbin, born in 1971. master, senior lecturer. **His research interests include** rough set, granular computing, data mining and artificial intelligence.

LI Jinjin, born in 1960. Ph. D, professor. **His research interests include** artificial Intelligence, granular computing, topology.

YU Peiqiu, born in 1991, M. S. candidate. **His research interests include** rough set, granular computing, data mining, artificial intelligence.

LIN Yidong, born in 1989, Ph.D. student, **His research interests include** uncertainty theory.