

关于贝叶斯粗糙集模型的知识约简^{*}王虹^{1,2} 张文修¹(西安交通大学理学院信息与科学研究所 西安 710049)¹(山西师范大学数学与计算机科学学院 临汾 041004)²

摘要 贝叶斯粗糙集模型是一种修正的变精度粗糙集模型,它既能克服粗糙集模型所描述的分类是完全精确的,同时又不需要象变精度粗糙集模型需要预先给定参数,因此在人工智能与认识科学领域有重要的作用。本文通过引入上、下分布约简给出该模型的一种约简方法。

关键词 贝叶斯粗糙集,信息系统,知识约简

Knowledge Reductions in the Bayesian Rough Set Model

WANG Hong^{1,2} ZHANG Wen-Xiu¹(Institute for Information and System Sciences, Faculty of Science, Xi'an Jiaotong University, Xi'an 710049)¹(College of Mathematics and Computer Science, Shanxi Teachers University, Linfen 041004)²

Abstract The Bayesian rough set model is a modified rough set model, it can avoid what Palawk rough set describes classification is precise completely, simultaneity it needn't give parameters previously like variable precision rough sets model. Hence, it acts an important role in fields of artificial intelligence and cognitive sciences. This paper gives a reduction method of this model by upper distribution reduction and lower distribution reduction.

Keywords Bayesian rough set, Information system, Knowledge reduction

粗糙集理论是由 Pawlak 于 1982 年提出的一种数据分析理论,这一理论为处理具有模糊不精确、不确定或不完全信息的分类提供了一种新的工具。基本的粗糙集模型的一个局限性是它处理的分类必须是完全正确的或肯定的,而没有某种程度上的“包含”或“属于”。这也限制了它在实际中的应用,因为大部分的决策表的属性之间并不一定存在严格的函数依赖关系,而只是表现出近似依赖关系。基于这种情况文[2]提出了一种可变精度的粗糙集模型,即允许一定程度的错误分类率存在,但在实际中我们只需根据所获得信息去处理问题,而不需要受预先给定某个参数的限制,因而文[3]进一步给出了贝叶斯粗糙集模型。本文利用文[4]给出的上下分布约简讨论了贝叶斯粗糙集模型的约简。

1 基本概念

定义 1 设 $(U, A \cup \{d\})$ 为信息系统, U 为对象的非空有限集合, $U = \{x_1, x_2, \dots, x_n\}$, A 为属性的非空集合, 即 $A = \{a_1, a_2, \dots, a_p\}$, d 为决策属性对于 $B \subseteq A$, 记

$$R_B = \{(x, y) : a_k(x) = a_k(y) \forall a_k \in B\} \cup R_B = \{[x]_B : x \in U\}$$

$$R_d = \{(x, y) : d(x) = d(y)\} \cup R_d = \{D_1, D_2, \dots, D_r\}$$

其中 $a_k(x)$, $d(x)$ 分别表示属性 a_k 和 d 在 U 中 x 处取值。

对于 $\forall X \subseteq U$, 记 $\beta \in (0, 0.5, 1]$ 记

$$\underline{R}_\beta(X) = \{x \in U | P(X|[x]_B) \geq \beta\} = \bigcup \{[x]_B | P(X|[x]_B) \geq \beta\}$$

$$\overline{R}_\beta(X) = \{x \in U | P(X|[x]_B) \geq 1 - \beta\} = \bigcup \{[x]_B | P(X|[x]_B) \geq 1 - \beta\}$$

则 $\underline{R}_\beta(X)$ 与 $\overline{R}_\beta(X)$ 分别为 X 关于 B 的 β 下近似和 β 上近似。

其中 $P(X/Y) = \frac{|X \cap Y|}{|Y|}$ 。基于 β 上、下近似而得到的粗糙集模型称为变精度粗糙集模型。

对于 $0 < P(x) < 1$, 记

$$\underline{R}_B^*(X) = \{x \in U : P(X|[x]_B) > P(X)\} = \bigcup \{[x]_B : P(X|[x]_B) > P(X)\}$$

$$\overline{R}_B^*(X) = \{x \in U : P(X|[x]_B) \geq P(X)\} = \bigcup \{[x]_B : P(X|[x]_B) \geq P(X)\}$$

则 $\underline{R}_B^*(X)$ 与 $\overline{R}_B^*(X)$ 分别为 X 关于 B 的贝叶斯下近似和上近似。

基于这种定义的上近似与下近似而得到的粗糙集模型称为贝叶斯粗糙集模型。

同样 $\underline{R}_B^*(x)$ 与 $\overline{R}_B^*(x)$ 满足对偶性质: $\underline{R}_B^*(x) = \sim \overline{R}_B^*(\sim x)$ 。

2 知识约简的定理及方法

定义 2 设 $(U, A \cup \{d\})$ 为目标信息系统, 对于 $B \subseteq A$ 记

$$\underline{B}(d) = (\underline{R}_B^*(D_1), \dots, \underline{R}_B^*(D_r))$$

$$\overline{B}(d) = (\overline{R}_B^*(D_1), \dots, \overline{R}_B^*(D_r))$$

(1) 若 $\underline{B}(d) = \underline{A}(d)$, 称 B 是 $(U, A \cup \{d\})$ 下分布协调集, 若 B 的任何真子集都不是下分布协调集, 则称 B 是 $(U, A \cup \{d\})$ 的下分布约简。

(2) 若 $\overline{B}(d) = \overline{A}(d)$, 称 B 是 $(U, A \cup \{d\})$ 上分布协调集, 若 B 的任何真子集都不是上分布协调集, 则称 B 是 $(U, A \cup \{d\})$ 的上分布约简。

定理 1 设 $(U, A \cup \{d\})$ 是目标信息系统, $B \subseteq A$, 记

$$M_B^{(1)}(x) = \{D_j | x \in \underline{R}_B^*(D_j)\} \quad x \in U$$

$$M_B^{(2)}(x) = \{D_j | x \in \overline{R}_B^*(D_j)\} \quad x \in U$$

^{*}资助项目: 973 计划资助 (编号: 2002CB312200)。王虹 副教授, 博士生, 研究方向: 人工智能的数学基础及粗糙集; 张文修 教授, 博士生导师, 研究方向: 人工智能的数学基础, 模糊集及粗糙集。

则

$$(1) \bar{B}(d) = \bar{A}(d) \Leftrightarrow M_B^{(1)}(x) = M_A^{(1)}(x) \quad \forall x \in U$$

$$(2) \underline{B}(d) = \underline{A}(d) \Leftrightarrow M_B^{(2)}(x) = M_A^{(2)}(x) \quad \forall x \in U$$

证: (1) 因 $x \in M_B^{(1)}(x) \Leftrightarrow x \in \bar{R}_B^*(D_j), x \in M_A^{(1)}(x) \Leftrightarrow x \in \bar{R}_A^*(D_j)$ 而 $\bar{B}(d) = \bar{A}(d)$, 即 $\bar{R}_B^*(D_j) = \bar{R}_A^*(D_j)$ 即证。

(2) 类似于(1)可证。

定理 2 设 $(U, A \cup \{d\})$ 是目标信息系统, $B \subseteq A$,

则

$$(1) \bar{B}(d) = \bar{A}(d) \Leftrightarrow \text{对 } \forall x, y \in U, \text{ 当 } M_A^{(1)}(x) \neq M_A^{(1)}(y)$$

时, $[x]_B \cap [y]_B = \emptyset$

$$(2) \underline{B}(d) = \underline{A}(d) \Leftrightarrow \text{对 } \forall x, y \in U, \text{ 当 } M_A^{(2)}(x) \neq M_A^{(2)}(y)$$

时, $[x]_B \cap [y]_B = \emptyset$

证: 必要性: 因 $\bar{B}(d) = \bar{A}(d)$, 由定理 1, 得 $M_B^{(1)}(x) = M_A^{(1)}(x), M_B^{(1)}(y) = M_A^{(1)}(y)$

若对 $\forall x, y \in U, [x]_B \cap [y]_B \neq \emptyset$, 一定有 $[x]_B = [y]_B$ 因此 $M_B^{(1)}(x) = M_B^{(1)}(y)$, 所以有 $M_A^{(1)}(x) = M_A^{(1)}(y)$ 。

充分性: 记 $J([x]_B) = \{[y]_A : [y]_A \subseteq [x]_B\}$

由于 $B \subseteq A$, 因此 $J([x]_B)$ 构成了 $[x]_B$ 的一个分划。

因对 $\forall x \in U$, 当 $[y]_A \subseteq [x]_B$ 时, 有 $[x]_B \cap [y]_B \neq \emptyset$

由已知有 $M_A(x) = M_A(y)$, 对 $\forall j \leq r$, 如果 $x \in \bar{R}_B^*(D_j)$, 则 $[x]_B \subseteq \bar{R}_B^*(D_j)$

由于 $[x]_B = \bigcup \{[y]_A : [y]_A \in J([x]_B)\}$ 因此对任意 $[y']_A \in J([x]_B)$, 有 $[y']_A \subseteq \bar{R}_B^*(D_j)$, 从而 $D_j \in M_A(y')$, $D_j \in M_A(x)$, 所以 $x \in \bar{R}_A^*(D_j)$, 即 $\bar{R}_B^*(D_j) \subseteq \bar{R}_A^*(D_j)$, 反过来, 若 $x \in \bar{R}_A^*(D_j)$, 则 $D_j \in M_A(x)$

当 $[y]_A \in J([x]_B)$ 时, $[y]_B \cap [x]_B \neq \emptyset$, 因此 $M_A(y) = M_A(x)$, 从而 $D_j \in M_A(y)$, 即 $P(D_j | [y]_A) \geq P(D_j)$, 因此, $P(D_j | [x]_B) = (\sum \{P(D_j | [y]_A) : [y]_A \in J([x]_B)\}) / |[x]_B| = \sum \{P(D_j | [y]_A) \cdot \frac{|[y]_A|}{|[x]_B|} : [y]_A \in J([x]_B)\} \geq P(D_j) \sum \{\frac{|[y]_A|}{|[x]_B|} : [y]_A \in J([x]_B)\} = P(D_j)$, 所以 $x \in \bar{R}_B^*(D_j)$, 即 $\bar{R}_A^*(D_j) \subseteq \bar{R}_B^*(D_j)$, 因此 $\bar{R}_B^*(D_j) = \bar{R}_A^*(D_j)$, 对 $\forall j \leq r$, 所以 $\bar{B}(d) = \bar{A}(d)$ 。

定义 3 设 $(U, A \cup \{d\})$ 是目标信息系统, 定义

$$D^{(i)}(x, y) = \begin{cases} \{a_i \in A : a_i(x) \neq a_i(y)\} & M_A^{(i)}(x) \neq M_A^{(i)}(y) \\ A & M_A^{(i)}(x) = M_A^{(i)}(y) \end{cases}$$

($i=1, 2$)

则 $D^{(i)}(x, y)$ ($i=1, 2$) 被分别称为对象 x, y 关于 $M_A^{(i)}(x)$ ($i=1, 2$) 的可辨识属性集, $D^{(i)} = \{D^{(i)}(x_i, y_j) : i, j \leq r\}$ ($i=1, 2$) 被分别称为目标信息系统关于 $M_A^{(i)}(x)$ ($i=1, 2$) 可辨识矩阵。

定理 3 设 $(U, A \cup \{d\})$ 是目标信息系统, $B \subseteq A$, 则

$$(1) \bar{B}(d) = \bar{A}(d) \Leftrightarrow B \cap D^{(1)}(x, y) \neq \emptyset \quad \forall x, y \in U$$

$$(2) \underline{B}(d) = \underline{A}(d) \Leftrightarrow B \cap D^{(2)}(x, y) \neq \emptyset \quad \forall x, y \in U$$

证: (1) 必要性

若 $\bar{B}(d) = \bar{A}(d)$, 对 $\forall x, y \in U$, 如果 $M_A^{(1)}(x) = M_A^{(1)}(y)$, 则 $D^{(1)}(x, y) = A$, 如果 $M_A^{(1)}(x) \neq M_A^{(1)}(y)$, 一定有 $B \cap D^{(1)}(x, y) \neq \emptyset$ 。

由定理 2, $[x]_B \cap [y]_B = \emptyset$, 于是存在 $a_k \in B$, 使 $a_k(x) \neq a_k(y)$, 所以 $a_k \in D^{(1)}(x, y)$, 因此 $B \cap D^{(1)}(x, y) \neq \emptyset$ 。

充分性:

若存在 $x, y \in U$, 使得 $B \cap D^{(1)}(x, y) \neq \emptyset$, 则 $M_A^{(1)}(x) = M_A^{(1)}(y)$, 对 $\forall a_k \in B$, 必有 $a_k \notin D^{(1)}(x, y)$, 所以 $a_k(x) = a_k(y)$, 因此 $[x]_B = [y]_B$, 由定理 2, $\bar{B}(d) = \bar{A}(d)$ 。(2) 类似于(1)可证得。

结论 属性约简是粗糙理论的主要内容之一, 即在保持信息不丢失的情况下约去冗余属性。本文利用上、下分布及可辨识属性矩阵, 讨论了贝叶斯粗糙集模型的知识约简, 这一方法比原来的约简方法在实际中具有可操作性, 在理论及应用上都是有意义的。

参考文献

- 1 Pawlak Z. Rough set [J]. International Journal of Computer and Information Science, 1982, 11, 341~356
- 2 Ziarko W. variable precision rough sets model [J]. Journal of computer and systems sciences, 1993, 46(1), 39~59
- 3 Slezak D, Ziarko W. Attribute reduction in the Bayesian Version of Variable precision rough set model [J]. Electronic Notes in Theoretical Computer Science, 2003(4), 1~11
- 4 张文修, 梁怡, 吴伟志, 等. 信息系统与知识发现[M]. 北京: 科学出版社, 2003
- 5 张文修, 米据生, 吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报, 2003, 26(1), 12~18

(上接第 149 页)

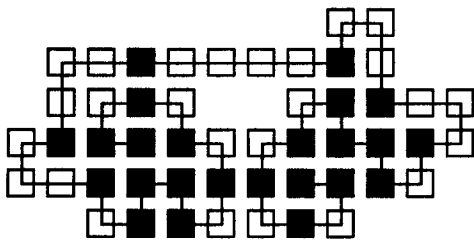


图 6 $N=50$, $E_{\min} = -21$ 的一个典型最低能量构形

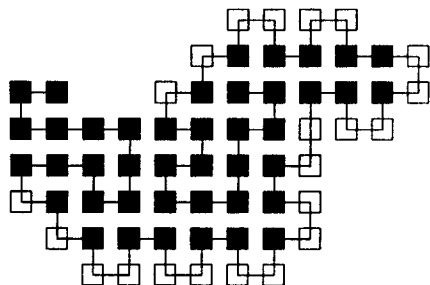


图 7 $N=64$, $E_{\min} = -42$ 的一个能量为 $E = -38$ 构形

通过上面的结果比较不难发现, 本文算法对大多数算例的计算都达到了最优解, 其计算效率要高于遗传算法和 Monte Carlo 方法。局部搜索算法具有局部寻优的特点, 而跳坑策略则增加了构形的多样性, 使算法在陷入僵局的情况下跳出“陷阱”, 同时保持原有构形的信息以引导算法走向有希望的区域。二者结合有效地提高了算法的计算效率。

参考文献

- 1 Dill K A, Bromberg S, Yue K, et al. Principles of protein folding: A perspective from simple exact models. Protein Sci., 1995, 4, 561~602
- 2 Unger R, Moult J. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull. Math. Biol., 1993, 55(6), 1183~1198
- 3 Unger R, Moult J. Genetic algorithms for protein folding simulations. J. Mol. Biol., 1993, 231, 75~81
- 4 王敏, 陈增强, 袁著社. 基于并行遗传算法的蛋白质空间结构预测. 计算机科学, 2003, 30(7): 147~150