基于信息量的完备覆盖约简算法

覃丽珍[1] 姚炳学[1] 李金海[2]
(聊城大学数学科学学院 聊城 252059)[1] (西安交通大学理学院 西安 710049)[2]

摘 要 覆盖粗糙集是Pawlak粗糙集的一种重要推广。类似于Pawlak粗糙集,约简也是覆盖粗糙集中的核心问题之一。通过引入覆盖族的信息量的概念,讨论了覆盖协调集、约简以及核的等价判定定理,同时对覆盖的重要性进行了度量;在此基础上,提出一种完备的启发式覆盖约简算法,它能够从搜索空间中逐步删除不重要覆盖,避免对其重要性的重复计算;最后,通过一个购房综合评价的实例说明了该算法的可行性与有效性。

关键词 粗糙集,覆盖粗糙集,约简,信息量
中图法分类号 TP18 **文献标识码** A

Complete Algorithm for Covering Reduction Based on Information Quantity

QIN Li-zhen[1] YAO Bing-xue[1] LI Jin-hai[2]
(School of Mathematics Science, Liaocheng University, Liaocheng 252059, China)[1]
(School of Science, Xi'an Jiaotong University, Xi'an 710049, China)[2]

Abstract Covering generalized rough sets are an important extension of Pawlak's rough sets. Similar to the Pawlak's rough sets, reduction is also one of the key issues in covering generalized rough sets. Information quantity of a family of coverings was introduced to characterize the consistent set, reduct and core of a family of coverings. A novel approach for measuring the significance of coverings was presented. Then, a complete heuristic algorithm was proposed to reduce a family of coverings. In the algorithm, unimportant coverings can gradually be removed from the search space to effectively avoid the repeated computation of the significance of the unimportant coverings. Finally, a real example of evaluating houses was used to illustrate the feasibility and effectiveness of the proposed algorithm.

Keywords Rough sets, Covering generalized rough sets, Reduction, Information quantity

1 引言

粗糙集理论[1]是波兰数学家Pawlak于1982年提出的一种处理含糊、不精确、不确定性知识的数学方法,已被广泛应用于机器学习、模式识别、决策分析、过程控制、数据挖掘、知识发现等领域[2-6]。

随着Pawlak粗糙集在实际应用中的不断深入,人们发现由于它是基于划分(或等价关系)建立的,极大地限制了其在一些复杂数据集中的应用。为了弥补Pawlak粗糙集在这方面存在的不足,很多学者从不同角度给出了不同的拓广粗糙集模型。例如,Zakowski在文献[7]中使用覆盖替换Pawlak粗糙集中的划分,提出了覆盖粗糙集。对于该覆盖粗糙集,Bonikowski等[8]研究了有关性质。Pomykala于1987年[9]提出了另外一种覆盖粗糙集,它保留了Zakowski定义的下近似算子,但上近似算子从外延上来讲要更加宽松。Tsang等在文献[10]中指出,从规则提取的角度来看,Zakowski[7]定义的上近似算子会丢失不精确规则信息,而Pomykala[9]定义的上近似算子往往又导致不精确规则信息的冗余。因此,Tsang等[10]介绍了一种新的上近似算子,它在外延上介于Zakowski与Pomykala定义的上近似算子之间。换句话说,就规则提取而言,Tsang等定义的上近似算子要更加合理[11]。为了方便叙述,本文将Zakowski[7]、Pomykala[9]、Tsang等[10]给出的覆盖粗糙集分别称为覆盖粗糙集Ⅰ、Ⅱ、Ⅲ。有关这3种覆盖粗糙集的性质见文献[11-13]。

类似于Pawlak粗糙集,覆盖粗糙集中的约简也吸引了众多学者的关注和研究兴趣。Zhu和Wang[14]在覆盖粗糙集Ⅰ中,讨论了保持上、下近似算子不变的覆盖约简。为了满足特定情况下对知识约简的需要,胡军和张闽[15]在覆盖粗糙集Ⅰ中又提出了保持上、下近似算子对给定对象集近似不变的相对覆盖约简,并称Zhu和Wang[14]定义的覆盖约简为绝对覆盖约简。事实上,不管是相对覆盖约简,还是绝对覆盖约简,它们都是在保持上、下近似算子不变或部分不变的前提下,删除单一覆盖中的可约知识。文献[16]在覆盖粗糙集Ⅱ中,基于辨识矩阵提出了覆盖族的约简方法,并从信息论的角度给出了计算最小约简的启发式算法。Tsang等[17]在覆盖粗糙集Ⅲ中,提出了与Pawlak粗糙集中的属性约简相一致的约简方法,并指出当覆盖均退化为划分时,所给出的覆盖约简就退化为Pawlak粗糙集下定义的属性约简。鉴于覆盖粗糙集Ⅲ

到稿日期:2011-12-25 返修日期:2012-03-18 本文受国家自然科学基金(71140008)资助。
覃丽珍 女,硕士生,主要研究方向为模糊代数与粗糙集理论,E-mail:qinlizhen2010@126.com;姚炳学(1963-),男,教授,硕士生导师,主要研究方向为模糊代数与粗糙集理论;李金海(1984-),男,博士生,主要研究方向为粗糙集与概念格。

在规则提取中的实用性较强，其约简方法的执行效率备受关注。由于文献[17]中提出的覆盖约简方法是基于辨识矩阵建立的，因此它的时间复杂度为指数级，这意味着它在实际中并不容易实现，特别是遇到大型覆盖数据集。为了有效降低覆盖粗糙集Ⅲ中计算约简的时间复杂度，本文在文献[17]的基础上，继续提出完备的具有多项式时间复杂度的启发式覆盖约简算法。该算法具有较大的灵活性，它能从搜索空间中逐步删除不重要覆盖，避免对其重要性的重复计算，以提高算法的搜索效率。

2 基础知识

为了后面几节讨论的方便，这一节简单回顾一下覆盖粗糙集中的一些基本概念。

定义 1[8] 设 U 为有限论域，$C=\{K_1,K_2,\cdots,K_s\}$ 为 U 的幂集的一个子集。若 C 中的每个集合均非空，且 $\bigcup_{i=1}^{s}K_i=U$，则称 C 为 U 的一个覆盖。

显然，覆盖是划分的一种推广。

定义 2[8] 称二元组 (U,C) 为一个覆盖近似空间，其中 U 是有限论域，C 是 U 上的一个覆盖。

定义 3[8] 设 (U,C) 为覆盖近似空间。对于任意 $x\in U$，称 $Md(x)=\{K\in C\mid x\in K\wedge\forall K'\in C(x\in K'\wedge K'\subseteq K)\Rightarrow K'=K\}$ 为对象 x 的最小描述。

从定义 3 可以看出，最小描述 $Md(x)$ 是由包含对象 x 的所有最小元（相对于 C 而言）组成的集合。

定义 4[8] 设 (U,C) 为覆盖近似空间。对于任意 $X\subseteq U$，称 $X_*=\bigcup\{K\in C\mid K\subseteq X\}$ 为集合 X 的下近似。

定义 5[17] 设 (U,C) 为覆盖近似空间。对于任意 $X\subseteq U$，称 $X^*=\bigcup\{K\in Md(x)\mid x\in X\}$ 为集合 X 的上近似。

有序对 (X_*,X^*) 称为 X 的覆盖粗糙集。依据引言中的约定，称其为覆盖粗糙集Ⅲ。本文将在覆盖粗糙集Ⅲ下讨论覆盖族的约简问题。

定义 6[17] 设 $C=\{K_1,K_2,\cdots,K_s\}$ 为 U 的一个覆盖。对于任意 $x\in U$，记 $K_x=\bigcap_{i=1}^{s}\{K_i\mid x\in K_i\in C\}$，称 $Cov(C)=\{K_x\mid x\in U\}$ 为 C 的诱导覆盖。

定义 7[17] 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，即 Δ 是 U 上的一族覆盖组成的集合。对于任意 $x\in U$，记 $\Delta_x=\bigcap_{j=1}^{t}\{K_x^j\mid K_x^j\in Cov(C_j)\}$，则称 $Cov(\Delta)=\{\Delta_x\mid x\in U\}$ 为 Δ 的诱导覆盖。

设 $G=\{C_1,C_2,\cdots,C_t\}$ 和 $\Delta'=\{C_1',C_2',\cdots,C_t'\}$ 为 U 的两个覆盖族。若对于任意 $x\in U$，都有 $\Delta_x=\Delta_x'$ 成立，则称这两个覆盖族的诱导覆盖相同，记为 $Cov(\Delta)=Cov(\Delta')$。

定义 8[17] 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族。若 $Cov(\Delta-\{C_i\})=Cov(\Delta)$，则称覆盖 C_i 在 Δ 中是不必要的；否则，称 C_i 在 Δ 中是必要的。

定义 9[17] 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族。对于 $R\subseteq\Delta$，若 $Cov(R)=Cov(\Delta)$，则称 R 为 Δ 的一个协调集；如果 R 是 Δ 的一个协调集，且 R 的任一真子集都不是 Δ 的协调集，那么称 R 为 Δ 的一个约简。Δ 的所有约简的交集称为 Δ 的核，记为 $CORE(\Delta)$。

由定义 9 可知，Δ 的一个约简 R 是指 Δ 的一个最小子集使其诱导覆盖保持不变。这说明，对 Δ 进行约简，可以消除覆盖的冗余，且不会影响诱导覆盖的结果。

3 覆盖族的信息量

文献[16,18-21]从信息论的角度讨论了属性（或覆盖）的重要性的度量方法。受此启发，本文将信息量引入到覆盖粗糙集Ⅲ中，以讨论覆盖族的协调集、约简以及核的等价判定定理。

定义 10 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$P,Q\subseteq\Delta$。如果对任意 $x\in U$，有 $P_x\subseteq Q_x$ 成立，那么称诱导覆盖 $Cov(P)$ 比 $Cov(Q)$ 细（或 $Cov(Q)$ 比 $Cov(P)$ 粗），记为 $Cov(P)\leqslant Cov(Q)$。

由定义 6 和定义 7 可知，下述性质成立。

性质 1 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族。则对任意 $R\subseteq\Delta$，有 $Cov(\Delta)\leqslant Cov(R)$。

性质 1 说明，删除一个覆盖族的部分覆盖，会导致其诱导覆盖变粗。

定义 11 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$。定义覆盖族 R 的信息量为

$$I(R)=\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|R_x|}{|U|}\right)$$

式中，$\{R_x\mid x\in U\}$ 为 R 的诱导覆盖，$|R_x|$ 表示 R_x 的基数。

容易证明，当覆盖族 R 全部退化为划分时，$I(R)$ 与文献[18]中定义的信息量是完全一致的，即覆盖粗糙集Ⅲ中的覆盖族的信息量是 Pawlak 粗糙集中的信息量的自然推广。

性质 2 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族。则对任意 $R\subseteq\Delta$，有 $I(R)\leqslant I(\Delta)$。

证明：由性质 1 可知，对任意 $x\in U$，有 $\Delta_x\subseteq R_x$ 成立，因此

$$\frac{1}{|U|^2}\sum_{x\in U}|\Delta_x|\leqslant\frac{1}{|U|^2}\sum_{x\in U}|R_x|$$

故，

$$I(R)=\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|R_x|}{|U|}\right)=1-\frac{1}{|U|^2}\sum_{x\in U}|R_x|\leqslant 1-\frac{1}{|U|^2}\sum_{x\in U}|\Delta_x|=\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|\Delta_x|}{|U|}\right)=I(\Delta)$$

性质 2 说明，删除一个覆盖族的部分覆盖，会导致其信息量减少。

定理 1 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$。R 是 Δ 的协调集当且仅当 $I(R)=I(\Delta)$。

证明：若 R 是 Δ 的协调集，由定义 9 知，对任意 $x\in U$，有 $R_x=\Delta_x$。根据定义 11，可得 $I(R)=I(\Delta)$。反之，若 $I(R)=I(\Delta)$，由性质 1、性质 2 可知，对任意 $x\in U$，有 $R_x=\Delta_x$ 成立，即 $Cov(R)=Cov(\Delta)$，因此 R 是 Δ 的一个协调集。

定理 2 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$C_i\in\Delta$。则 C_i 在 Δ 中是必要的当且仅当 $I(\Delta-\{C_i\})<I(\Delta)$。

证明：记 $R=\Delta-\{C_i\}$。若 C_i 在 Δ 中是必要的，由定义 8 知，$Cov(\Delta)\neq Cov(R)$。根据性质 1 有，对任意 $x\in U$，有 $\Delta_x\subseteq R_x$，且存在 $x_0\in U$，使得 $\Delta_x\subset R_x$。根据定义 10，有 $I(R)<I(\Delta)$ 成立，即 $I(\Delta-\{C_i\})<I(\Delta)$。由于上述各步都是可逆的，因此充分性也是成立的。

定理 3 设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族。则 Δ 的

核满足：$CORE(\Delta)=\{C_i\in\Delta\mid I(\Delta)>I(\Delta-\{C_i\})\}$。

证明：要证 $CORE(\Delta)=\{C_i\in\Delta\mid I(\Delta)>I(\Delta-\{C_i\})\}$，由定理 2 可知，只需证 $CORE(\Delta)$ 是 Δ 的所有必要覆盖组成的集合。为了方便，不妨设 Δ 的所有约简为 $\{R_j\mid j=1,2,\cdots,t\}$。

(1)证明 Δ 的每一个必要覆盖都属于 $CORE(\Delta)$

对于 Δ 的任一必要覆盖 C_i，根据定理 2，有 $I(\Delta-\{C_i\})<I(\Delta)$，由此可得：对于任意 $j\in\{1,2,\cdots,t\}$，有 $C_i\in R_j$ 成立。(反证法)若存在 $j_0\in\{1,2,\cdots,t\}$ 使得 $C_i\notin R_{j_0}$，那么 $I(R_{j_0})\leqslant I(\Delta-\{C_i\})<I(\Delta)$，这与 R_{j_0} 为 Δ 的一个约简相矛盾。因此，$C_i\in\bigcap_{j\in\{1,2,\cdots,t\}}R_j=CORE(\Delta)$。

(2)证明 $CORE(\Delta)$ 的每个元素必是 Δ 的必要覆盖

对于任意 $C_i\in CORE(\Delta)$，根据定义 9 可得：对任意 $j\in\{1,2,\cdots,t\}$，有 $C_i\in R_j$ 成立。下面用反证法证明 C_i 为 Δ 的一个必要覆盖，即证明 $I(\Delta-\{C_i\})<I(\Delta)$ 成立。如果 $I(\Delta-\{C_i\})=I(\Delta)$，那么 $\Delta-\{C_i\}$ 为 Δ 的一个协调集。由于每个协调集都至少包含 Δ 的一个约简，因此存在子集 $R_{j_0}\subseteq\Delta-\{C_i\}$ 使得 R_{j_0} 为 Δ 的一个约简，这与 $C_i\in R_{j_0}$ 是相矛盾的。所以假设不成立，即 $I(\Delta-\{C_i\})<I(\Delta)$ 成立。

定理 4　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$。R 是 Δ 的一个约简当且仅当 $I(R)=I(\Delta)$，且 $\forall C\in R$，$I(R-\{C\})<I(R)$。

证明：由定理 1 和定义 9 即可得证。

4　覆盖重要性的度量方法

在粗糙集理论中，属性重要性的合理度量是一个非常重要的课题，它可以用来作为约简算法的启发式信息。下面从信息量的角度给出覆盖重要性的一种度量方法。

定义 12　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$，$C_i\in R$。定义覆盖 C_i 在 R 中的重要性为

$$SIG(R\mid C_i)=I(R)-I(R-C_i)$$

定义 12 表明，C_i 在 R 中的重要性是通过从 R 中删除覆盖 C_i 后，所减少的信息量来度量的。如果删除覆盖 C_i 后 R 减少的信息量越大，说明 C_i 在 R 中就越重要。

由定义 8 和定理 2 可知，下述性质成立。

性质 3　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$C_i\in\Delta$。则 C_i 在 Δ 中是必要的当且仅当 $SIG(\Delta\mid C_i)>0$。

性质 3 说明，删除 Δ 的任一必要覆盖，都会引起其信息量的减少。因此，所有的必要覆盖在约简的过程中，都应该得到保留。

定义 13　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$，$C_i\in\Delta-R$。定义 C_i 相对于 R 的重要性为

$$SIG(C_i\mid R)=I(R\cup\{C_i\})-I(R)$$

定义 13 表明，C_i 相对于 R 的重要性，是通过将覆盖 C_i 添加到 R 中所增加的信息量来度量的。如果添加覆盖 C_i 引起 R 的信息增量越大，说明 C_i 相对于 R 越重要。根据 Pawlak 粗糙集中约简算法设计的经验，这种重要性可以用来作为增量式覆盖约简算法的启发式信息。另外，为了使得所设计的覆盖约简算法能够更加有效，还需继续引入不重要覆盖的定义。

定义 14　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$R\subseteq\Delta$，$C_i\in\Delta-R$。若 $SIG(C_i\mid R)=0$，则称 C_i 相对于 R 是不重要的。

定理 5　设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，$P\subseteq Q\subseteq\Delta$，$C_i\in\Delta-Q$。若 $SIG(C_i\mid P)=0$，则 $SIG(C_i\mid Q)=0$。

证明：由于

$$\begin{aligned}
&I(P\cup\{C_i\})-I(P)\\
&=\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|(P\cup\{C_i\})_x|}{|U|}\right)-\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|P_x|}{|U|}\right)\\
&=\sum_{x\in U}\frac{1}{|U|}\left(\frac{|P_x|}{|U|}-\frac{|(P\cup\{C_i\})_x|}{|U|}\right)\\
&=\sum_{x\in U}\frac{1}{|U|}\left(\frac{|P_x|-|P_x\cap K_x|}{|U|}\right),K_x\in Cov(C_i)\\
&=\sum_{x\in U}\frac{1}{|U|}\frac{|P_x-K_x|}{|U|}
\end{aligned}$$

因此，若 $SIG(C_i\mid P)=0$，那么对任意 $x\in U$，都有 $P_x\subseteq K_x$ 成立，其中 $\{P_x\mid x\in U\}$ 和 $\{K_x\mid x\in U\}$ 分别为 P、C_i 的诱导覆盖。又由于 $P\subseteq Q$，因此

$$\begin{aligned}
&I(Q\cup\{C_i\})-I(Q)\\
&=\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|(Q\cup\{C_i\})_x|}{|U|}\right)-\sum_{x\in U}\frac{1}{|U|}\left(1-\frac{|Q_x|}{|U|}\right)\\
&=\sum_{x\in U}\frac{1}{|U|}\frac{|Q_x-K_x|}{|U|}\\
&\leqslant\sum_{x\in U}\frac{1}{|U|}\frac{|P_x-K_x|}{|U|}=0
\end{aligned}$$

定理 5 表明，如果 C_i 相对于 P 是不重要的，那么 C_i 相对于 $Q\supseteq P$ 也是不重要的。这一点会被即将提出的覆盖约简算法充分利用起来，以加快算法的收敛速度。

5　覆盖约简的启发式算法

通过前面几节的准备，这一节正式给出覆盖约简的启发式算法。为了便于理解，在提出启发式覆盖约简算法之前，先简单描述一下算法的基本原理。

设 $\Delta=\{C_1,C_2,\cdots,C_t\}$ 为 U 的覆盖族，由定义 9 可知，Δ 的核 $CORE(\Delta)$ 中的每个覆盖包含在任一约简之中。因此，为了提高约简的效率，先计算 $CORE(\Delta)$（$CORE(\Delta)$ 可以通过定理 3 快速求得），然后以它作为起点去搜索 Δ 的一个约简。首先，判断 $I(CORE(\Delta))=I(\Delta)$ 是否成立。如果成立，则由定理 4 可知 $CORE(\Delta)$ 已是 Δ 的一个约简；如果不成立，那么对 $\Delta-CORE(\Delta)$ 中的每个覆盖 C_i，计算其相对于 $CORE(\Delta)$ 的重要性，并选择一个最重要的覆盖 C_j 添加到 $CORE(\Delta)$ 中进行更新，同时从 Δ 中删除不重要的覆盖。上述步骤将被反复执行。由于 Δ 是一个有限集，因此 $I(CORE(\Delta))=I(\Delta)$ 必能在有限步之内得到满足。最后，再从 $CORE(\Delta)$ 中依次去掉满足 $I(R-\{C_i\})<I(R)$ 的覆盖 C_i，那么最终得到的覆盖族 $CORE(\Delta)$ 便是 Δ 的一个约简。上述约简思想可以归纳成如下算法。

算法 1　覆盖约简的启发式算法

输入：U 的覆盖族 $\Delta=\{C_1,C_2,\cdots,C_t\}$

输出：Δ 的一个约简

1)初始化 $CORE(\Delta)=\emptyset$。

2)对于 Δ 中的每个覆盖 C_i，根据定义 12，计算 C_i 在 Δ 中的重要性 $SIG(\Delta\mid C_i)$，并将所有满足 $SIG(\Delta\mid C_i)>0$ 的覆盖 C_i 添加到 $CORE(\Delta)$ 中。

3)判断 $I(CORE(\Delta))=I(\Delta)$ 是否成立。若成立，则转步骤 5)；否则转步骤 4)。

4)对 $\Delta-\mathrm{CORE}(\Delta)$ 中的每个覆盖 C_i,根据定义 13,计算 C_i 相对于 $\mathrm{CORE}(\Delta)$ 的重要性 $\mathrm{SIG}(C_i|\mathrm{CORE}(\Delta))$;在此基础上,从 $\Delta-\mathrm{CORE}(\Delta)$ 中选择满足

$$\mathrm{SIG}(C_j|\mathrm{CORE}(\Delta))=\max_{C_i\in\Delta-\mathrm{CORE}(\Delta)}\{\mathrm{SIG}(C_i|\mathrm{CORE}(\Delta))\}$$

的覆盖 C_j 添加到 $\mathrm{CORE}(\Delta)$ 中对其进行更新,并从 Δ 中去掉那些既属于 $\Delta-\mathrm{CORE}(\Delta)$ 但相对于 $\mathrm{CORE}(\Delta)$ 又不重要的覆盖,然后返回步骤 3)。

5)对覆盖族 $\mathrm{CORE}(\Delta)$ 进行一次扫描,找出它的所有不必要覆盖,即从 $\mathrm{CORE}(\Delta)$ 中找出使等式 $I(\mathrm{CORE}(\Delta)-\{C_i\})=I(\mathrm{CORE}(\Delta))$ 成立的所有覆盖 C_i,并令 Ω 为 $\mathrm{CORE}(\Delta)$ 的所有不必要覆盖组成的集合。

6)若 Ω 为空,则转步骤 8);否则从 Ω 中选出一个满足 $I(\mathrm{CORE}(\Delta)-\{C_k\})=I(\mathrm{CORE}(\Delta))$ 的覆盖 C_k,将其从 $\mathrm{CORE}(\Delta)$ 中删除。

7)从 Ω 中删除步骤 6)中已选出的覆盖 C_k 和已判断过不满足等式 $I(\mathrm{CORE}(\Delta)-\{C_i\})=I(\mathrm{CORE}(\Delta))$ 的覆盖,然后返回步骤 6)。

8)算法结束,输出 $\mathrm{CORE}(\Delta)$。

定理 6 算法 1 中的步骤 4),每次执行时从 Δ 中去掉那些既属于 $\Delta-CORE(\Delta)$ 但相对于 $CORE(\Delta)$ 又不重要的覆盖,不会影响寻找满足 $I(CORE(\Delta))=I(\Delta)$ 的覆盖族 $CORE(\Delta)$。

证明:不妨假设步骤 3)与 4)为了得到满足 $I(CORE(\Delta))=I(\Delta)$ 的覆盖族 $CORE(\Delta)$ 而反复执行时依次去掉的覆盖为 $\{C_{i_1},C_{i_2},\cdots,C_{i_k}\}$。为了方便描述,称这种做法为删除策略。

首先证明在不采取上述删除策略时,步骤 4)每次得到执行时,添加到 $CORE(\Delta)$ 中进行更新的覆盖 C_j 都满足 $SIG(C_j|CORE(\Delta))>0$。(反证法)若步骤 4)得到执行,但添加到 $CORE(\Delta)$ 中进行更新的覆盖 C_j 满足 $SIG(C_j|CORE(\Delta))=0$。那么依据步骤 4)的设计特点,对所有 $C_i\in\Delta-CORE(\Delta)$,都应该有 $SIG(C_i|CORE(\Delta))=0$ 成立。由此可得 $I(CORE(\Delta))=I(\Delta)$,这与步骤 4)得到执行是矛盾的,因为要执行步骤 4)的前提条件是 $I(CORE(\Delta))=I(\Delta)$ 不成立。

对于任一覆盖 $C_{i_j}\in\{C_{i_1},C_{i_2},\cdots,C_{i_k}\}$,存在某时刻的 $CORE(\Delta)$ 使得 $SIG(C_{i_j}|CORE(\Delta))=0$。根据定理 5 可知,一旦 C_{i_j} 相对于 $CORE(\Delta)$ 是不重要的,那么不管如何通过添加覆盖对 $CORE(\Delta)$ 进行更新,C_{i_j} 相对于 $CORE(\Delta)$ 都是不重要的。又由于添加到 $CORE(\Delta)$ 中进行更新的覆盖 C_j 都满足 $SIG(C_j|CORE(\Delta))>0$,因此,C_{i_j} 一旦相对于 $CORE(\Delta)$ 是不重要的,就可以从搜索空间中删除,不会影响寻找满足 $I(CORE(\Delta))=I(\Delta)$ 的覆盖族 $CORE(\Delta)$。

定理 6 表明,采取删除不重要覆盖的策略不但不会影响覆盖约简的计算,反而可以减少求解约简所需的计算量。这是因为一旦发现不重要覆盖就进行删除,在下次循环时,就不需要再计算这些被删除的不重要覆盖的重要性了。

定理 7 算法 1 是完备的,即它输出的覆盖族 $CORE(\Delta)$ 一定是 Δ 的一个约简。

证明:由算法 1 中的步骤 3)、4)以及定理 6 可知,等式 $I(CORE(\Delta))=I(\Delta)$ 是成立的。再根据步骤 5)—7)可知,$CORE(\Delta)$ 还满足:$I(CORE(\Delta)-\{C_i\})<I(CORE(\Delta))$,$\forall C_i\in CORE(\Delta)$。因此,算法 1 输出的 $CORE(\Delta)$ 是 Δ 的一个约简。

文献[16,18-20]中给出的基于信息量(或条件信息量)的启发式算法都不能保证是完备的,即这些算法输出的均是相对约简,不能保证一定是约简,原因是它们都没有对最后得到的协调集进行一次冗余过滤。

下面继续分析算法 1 的时间复杂度:由定义 6、定义 7 可知,计算 Δ 的诱导覆盖 $Cov(\Delta)$ 的时间复杂度为 $O(|\Delta||U|^2)$,再由定义 11、定义 12 可得,执行步骤 1)、2)的时间复杂度为 $O(|\Delta|^2|U|^2)$。步骤 3)、4)在最坏的情况下,需要计算 $|\Delta|(|\Delta|+1)/2$ 次 $SIG(C_i|CORE(\Delta))$。此处,为了减少计算量,每次运算完之后,用中间变量替换并保存已得到的诱导覆盖 $Cov((CORE(\Delta))$,在此情况下求解 $SIG(C_i|CORE(\Delta))$ 的时间复杂度为 $O(|U|^2)$,所以执行步骤 3)、4)的时间复杂度为 $O(|\Delta|^2|U|^2)$。步骤 5)—7)在最坏的情况下,需要判断 $2|CORE(\Delta)|$ 次 $I(CORE(\Delta)-\{C_i\})=I(CORE(\Delta))$ 是否成立,而每次判断它是否成立的时间复杂度为 $O(|\Delta\|U|^2)$,故执行步骤 5)—7)的时间复杂性为 $O(|\Delta|^2|U|^2)$。综上可知,算法 1 的时间复杂度为 $O(|\Delta|^2|U|^2)$。显然,该时间复杂度是多项式的。

下面通过一个购房综合评价的实例(文献[17]中例 4.9 节的进一步扩充得到)来进一步说明算法 1 的具体实施过程。

例 1 假设论域 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$ 为 9 套房组成的集合,现在通过价格、颜色、内部布局、周边环境、交通、教育设施、停车位、楼距这 8 个方面综合评价这 9 套房。价格的评价值:高、中、低,颜色的评价值:好、不好,内部布局的评价值:合理、一般、不合理,周边环境的评价值:安静、有点吵、比较吵、相当吵,交通的评价值:便利、不便利,教育设施的评价值:齐全、一般、欠缺,停车位的评价值:满意、不满意,楼距的评价值:合理、不合理。目前,已邀请 4 位专家甲、乙、丙、丁分别就上述 8 个方面对这 9 套房进行评估,详细评估信息如下:

专家对于价格的评价:

甲:高$=\{x_1,x_4,x_5,x_7\}$,中$=\{x_2,x_8\}$,低$=\{x_3,x_6,x_9\}$;

乙:高$=\{x_1,x_2,x_4,x_7,x_8\}$,中$=\{x_5\}$,低$=\{x_3,x_6,x_9\}$;

丙:高$=\{x_1,x_4,x_7\}$,中$=\{x_8\}$,低$=\{x_2,x_3,x_5,x_6,x_9\}$;

丁:高$=\{x_1,x_4,x_7\}$,中$=\{x_5\}$,低$=\{x_2,x_3,x_6,x_8,x_9\}$。

专家对于颜色的评价:

甲:好$=\{x_1,x_2,x_3,x_6\}$,不好$=\{x_4,x_5,x_7,x_8,x_9\}$;

乙:好$=\{x_1,x_2,x_3,x_5\}$,不好$=\{x_4,x_6,x_7,x_8,x_9\}$;

丙:好$=\{x_1,x_2,x_3,x_4\}$,不好$=\{x_5,x_6,x_7,x_8,x_9\}$;

丁:好$=\{x_1,x_2,x_3\}$,不好$=\{x_4,x_5,x_6,x_7,x_8,x_9\}$。

专家对于内部布局的评价:

甲:合理$=\{x_1,x_2,x_3\}$,一般$=\{x_4,x_5,x_6,x_7,x_8\}$,不合理$=\{x_9\}$;

乙:合理$=\{x_1,x_2,x_3\}$,一般$=\{x_4,x_5,x_6,x_7,x_9\}$,不合理$=\{x_8\}$;

丙:合理$=\{x_1,x_2,x_3\}$,一般$=\{x_4,x_5,x_6,x_8,x_9\}$,不合理$=\{x_7\}$;

丁:合理$=\{x_1,x_2,x_3\}$,一般$=\{x_4,x_5,x_6\}$,不合理$=\{x_7,x_8,x_9\}$。

专家对于周边环境的评价:

甲:安静$=\{x_1,x_2\}$,一点吵$=\{x_3,x_6\}$,比较吵$=\{x_4,x_5,x_7\}$,相当吵$=\{x_8,x_9\}$;

乙:安静$=\{x_1,x_5\}$,一点吵$=\{x_2,x_3\}$,比较吵$=\{x_4,x_7,$

x_8},相当吵={x_6,x_9};

丙:安静={x_1},一点吵={x_2,x_3},比较吵={x_4,x_7,x_8},相当吵={x_5,x_6,x_9};

丁:安静={x_1,x_2,x_4},一点吵={x_3,x_5},比较吵={x_7,x_8},相当吵={x_6,x_9}。

专家对于交通的评价:

甲:便利={x_1,x_2,x_3,x_5},不便利={x_4,x_6,x_7,x_8,x_9};

乙:便利={x_1,x_2,x_3,x_4,x_5},不便利={x_6,x_7,x_8,x_9};

丙:便利={x_1,x_2,x_3,x_5},不便利={x_4,x_6,x_7,x_8,x_9};

丁:便利={x_1,x_2,x_3,x_5,x_6},不便利={x_4,x_7,x_8,x_9}。

专家对于教育设施的评价:

甲:丰富={x_1,x_2,x_3},一般={x_5,x_6},欠缺={x_4,x_7,x_8,x_9};

乙:丰富={x_1,x_2,x_3,x_4},一般={x_5,x_6},欠缺={x_7,x_8,x_9};

丙:丰富={x_1,x_2,x_3},一般={x_4,x_5},欠缺={x_6,x_7,x_8,x_9};

丁:丰富={x_1,x_2,x_3},一般={x_4,x_5,x_6},欠缺={x_7,x_8,x_9}。

专家对于停车位的评价:

甲:满意={x_1,x_2,x_3,x_6,x_7,x_8},不满意={x_4,x_5,x_9};

乙:满意={x_1,x_2,x_3,x_8},不满意={x_4,x_5,x_6,x_7,x_9};

丙:满意={x_1,x_2,x_3,x_6,x_8},不满意={x_4,x_5,x_7,x_9};

丁:满意={x_1,x_2,x_3,x_7,x_8},不满意={x_4,x_5,x_6,x_9}。

专家对于楼距的评价:

甲:合理={x_1,x_2,x_3,x_4,x_6,x_8},不合理={x_5,x_7,x_9};

乙:合理={x_1,x_2,x_3,x_4,x_7,x_8},不合理={x_5,x_6,x_9};

丙:合理={x_1,x_2,x_3,x_4,x_8},不合理={x_5,x_6,x_7,x_9};

丁:合理={x_1,x_2,x_3,x_5,x_8},不合理={x_4,x_6,x_7,x_9}。

为了不丢失专家给出的评估信息,需要对这些信息进行融合。以价格为例,融合甲、乙、丙、丁4位专家的评价信息后,得到价格的总评价信息为

$C_1=\{\{x_1,x_2,x_4,x_5,x_7,x_8\},\{x_2,x_5,x_8\},\{x_2,x_3,x_5,x_6,x_8,x_9\}\}$

其中,高={x_1,x_2,x_4,x_5,x_7,x_8},中={x_2,x_5,x_8},低={x_3,x_6,x_9}。类似地,对于另外7个方面,融合甲、乙、丙、丁4位专家的评价信息后,可以得到:

颜色:

$C_2=\{\{x_1,x_2,x_3,x_4,x_5,x_6\},\{x_4,x_5,x_6,x_7,x_8,x_9\}\}$

内部布局:

$C_3=\{\{x_1,x_2,x_3\},\{x_4,x_5,x_6,x_7,x_8,x_9\},\{x_7,x_8,x_9\}\}$

周边环境:

$C_4=\{\{x_1,x_2,x_4,x_5\},\{x_2,x_3,x_5,x_6\},\{x_4,x_5,x_7,x_8\},\{x_5,x_6,x_8,x_9\}\}$

交通:

$C_5=\{\{x_1,x_2,x_3,x_4,x_5,x_6\},\{x_4,x_6,x_7,x_8,x_9\}\}$

教育设施:

$C_6=\{\{x_1,x_2,x_3,x_4\},\{x_4,x_5,x_6\},\{x_4,x_6,x_7,x_8,x_9\}\}$

停车位:

$C_7=\{\{x_1,x_2,x_3,x_6,x_7,x_8\},\{x_4,x_5,x_6,x_7,x_9\}\}$

楼距:

$C_8=\{\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\},\{x_4,x_5,x_6,x_7,x_9\}\}$

因此,综合4位专家给出的全部评价信息后,得到U的8个覆盖$\Delta=\{C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8\}$。为了对上述评价信息进行精简,下面利用算法1求$\Delta$进行约简。

(1)由定义12可知,Δ中的每个覆盖的重要性为:$SIG(\Delta|C_i)=0(i=1,2,4,5,6,7,8)$,$SIG(\Delta|C_3)=1/81$,因此,$\Delta$的核为$CORE(\Delta)=\{C_3\}$。由于$I(CORE(\Delta))=45/81<70/81=I(\Delta)$,下面对$CORE(\Delta)$进行不断更新以得到$\Delta$的一个约简。令$R_1=CORE(\Delta)=\{C_3\}$。

(2)由定义13可知,$\Delta-R_1$中的覆盖相对于R_1的重要性依次为:$SIG(C_1|R_1)=16/81$,$SIG(C_2|R_1)=9/81$,$SIG(C_4|R_1)=21/81$,$SIG(C_5|R_1)=11/81$,$SIG(C_6|R_1)=12/81$,$SIG(C_7|R_1)=SIG(C_8|R_1)=10/81$,故将$C_4$添加到$R_1$中。记$R_2=R_1\cup\{C_4\}=\{C_3,C_4\}$。由于$I(R_2)=66/81<70/81=I(\Delta)$,因此为了得到$\Delta$的一个约简,还需继续对$R_2$进行更新。

(3)$\Delta-R_2=\{C_1,C_2,C_5,C_6,C_7,C_8\}$中的覆盖相对于$R_2$的重要性依次为:$SIG(C_1|R_2)=SIG(C_2|R_2)=0$,$SIG(C_5|R_2)=SIG(C_6|R_2)=SIG(C_8|R_2)=2/81$,$SIG(C_7|R_2)=3/81$。故将覆盖$C_7$添加到$R_2$中,并从$\Delta$中删除不重要覆盖$C_1$和$C_2$。记$R_3=R_2\cup\{C_7\}=\{C_3,C_4,C_7\}$,$\Delta'=\Delta-\{C_1,C_2\}$。由于$I(R_3)=69/81<70/81=I(\Delta)$,因此$R_3$仍然需要添加新覆盖以得到$\Delta$的一个约简。

(4)$\Delta'-R_3=\{C_5,C_6,C_8\}$中的覆盖相对于$R_3$的重要性为:$SIG(C_5|R_3)=SIG(C_6|R_3)=1/81$,$SIG(C_8|R_3)=0$。故将覆盖$C_5$添加到$R_3$中,并记$R=R_3\cup\{C_5\}=\{C_3,C_4,C_5,C_7\}$。

容易验证,$I(R)=I(\Delta)$,且$I(R)>I(R-C_i)(i=3,4,5,7)$。由定理4可知,R是Δ的一个约简。

在步骤(3)中,由于从Δ中删除了不重要覆盖C_1和C_2,在步骤(4)中就避免了计算覆盖C_1和C_2相对于R_3的重要性,从而减少了计算量。当覆盖族Δ较大时,这种删除不重要覆盖的策略可以大大降低算法的计算量。因为对于大型覆盖数据集而言,在计算覆盖约简的整个过程中,能够删除的不重要覆盖在数量上一般是相当可观的。

根据上述约简得到的结果$R=\{C_3,C_4,C_5,C_7\}$,可以得到如下结论:对于这9套房,并不需要参考4位专家就上述8个方面给出的全部评价信息,而仅仅只需参考内部布局、周边环境、交通、停车位这4个方面的信息即可。

结束语 覆盖粗糙集Ⅲ是Pawlak粗糙集的一种推广。类似于Pawlak粗糙集,约简也是覆盖粗糙集Ⅲ中的核心问题之一。由于现有的基于辨识矩阵的覆盖约简方法在计算上存在较大的困难,因此继续研究高效的覆盖约简算法是有意义的。

该文在覆盖粗糙集Ⅲ中,引入了覆盖族的信息量的概念,讨论了覆盖协调集、约简以及核的等价判定定理,同时给出了覆盖重要性的一种度量方法;在此基础上,以覆盖重要性作为启发式信息,提出了一种完备的具有多项式时间复杂度的覆盖约简算法。该算法具有较大的灵活性,它能从搜索空间中逐步删除不重要覆盖,避免了对其重要性的重复计算,有效提

(下转第263页)

MIT Press,2007
[2] Zadeh L A. Fuzzy sets[J]. Information and Control,1965,8:338-353
[3] 薛艳,雷红轩,李永明.基于可能性测度的计算树逻辑[J].计算机工程与科学,2011,33(9):70-75
[4] Drakopoulos J A. Probabilities, possibilities, and fuzzy sets[J]. Fuzzy Sets and Systems,1995,75:1-15
[5] 李永明.模糊系统分析[M].北京:科学出版社,2005
[6] 欧阳丹彤,欧阳继红.基于模型的诊断方法[J].南京大学学报,2000,36:187-192
[7] 雷丽晖,段振华.使用扩展区间时序逻辑为并发工作流建模[J].西安电子科技大学学报,2007,34(4):673-680
[8] 赵林,吴尽昭.基于吴方法的多值模型检验[J].系统科学与数学,2008,28(8):1020-1029
[9] 杨军,葛海通,郑飞君,等.一种形式化验证方法:模型检验[J].浙江大学学报,2006,33(4):403-407
[10] 哈明虎,吴从炘.模糊测度与模糊积分[M].北京:科学出版社,1998
[11] Clarke E, Grumberg O, Peled D. Model Checking [M]. MIT Press,1999
[12] Jenhani I, Benferhat S, Elouedi Z. Learning and Evaluating Possibilistic DecisionTrees using Information Affinity[J]. World Academy of Science, Engineering and Technology, 2010, 63: 599-605
[13] Droste M, Kuich W, Vogler H, et al. Handbook of Weighted Automata [M]. An EATCS Series, Birlin-Heidelberg: Springer-Verlag,2009
[14] Kwiatkowska M. Quantitative Verification: Models, Techniques and Tools[C]//ESEC-FSE '07. 2007:449-458
[15] Hart S, Sharir M. Probabilistic propositional temporal logics[J]. Information and Control,1986,70(2/3):97-155
[16] Boutouhami S, Mokhtari A. Possibilistic Explanation[J]. International Journal of Computer Sicence and Applicatongs, 2006, 3(2):57-73

(上接第 239 页)
高了算法的计算效率。此外,通过购房综合评价的实例表明,该算法是可行且有效的。另外,该文提出的‘删除不重要覆盖的策略’也可以用来加速其它基于信息量的约简算法,例如文献[16,18]中的约简算法。

相对于 Pawlak 粗糙集中的高效约简算法,该文提出的完备覆盖约简算法的时间复杂度可能仍有下降的空间,因此继续提出一些有效的措施以提高本文算法的运行效率是很有必要的。对于这方面的研究,可以借鉴 Pawlak 粗糙集中已取得的成熟经验,例如先排序,再求论域的划分。与此同时,困难也是存在的,比如“先排序,再求论域的划分”这种策略,它在一定程度上依赖于对象在各属性下的取值情况,而以覆盖族进行表示的数据集中对象在各属性下的取值有些并不唯一,故直接利用属性值进行排序存在一定的困难。因此,这个问题还有待于进一步研究,也是本文后继工作的一部分。此外,继续讨论覆盖决策系统[21,22]下的启发式加速约简算法是本文的另一个后继工作,将另文发表。

参考文献

[1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences,1982,11(5):341-356
[2] Pawlak Z, Skowron A. Rudiments of rough sets [J]. Information Sciences,2006,177(1):3-27
[3] 张文修,梁怡,吴伟志.信息系统与知识发现[M].北京:科学出版社,2003
[4] 张文修,吴伟志.粗糙集理论介绍和研究综述[J].模糊系统与数学,2000,14(4):1-12
[5] 王国胤,姚一豫,于洪.粗糙集理论与应用研究综述[J].计算机学报,2009,32(7):1229-1246
[6] 李金海,吕跃进.决策系统的快速属性约简算法[J].电子科技大学学报,2007,36(6):1237-1240
[7] Zakowski W. Approximations in the Space(u,π)[J]. Demonstratio Mathematica,1983,16:761-769
[8] Bonikowski Z, Bryniarski E, Wybraniec U. Extensions and intensions in the rough set theory [J]. Information Sciences, 1998,107:149-167
[9] Pomykala J A. Approximation operations in approximation space [J]. Bulletin of the Polish Academy of Sciences, 1987, 35: 653-662
[10] Tsang E, Chen D G, Lee J, et al. On the upper approximations of covering generalized rough sets [C]//IEEE Proceedings of the Third International Conference on Machine Learning and Cybernetics. Shanghai,2004:4200-4203
[11] Zhu W, Wang F Y. On three types of covering-based rough sets [J]. IEEE Transactions on Knowledge and Data Engineering, 2007,19:1131-1144
[12] 魏荣,刘保仓,史开泉.基于覆盖广义粗集的模糊性[J].计算机科学,2007,34(1):153-155
[13] 杨勇,朱晓钟,李廉.覆盖粗糙集的公理化[J].计算机科学,2009,36(5):181-182
[14] Zhu W, Wang F Y. Reduction and axiomization of covering generalized rough sets [J]. Information Sciences, 2003, 152: 217-230
[15] 胡军,张闽.覆盖近似空间的约简理论[J].计算机工程与应用,2007,43(28):86-88
[16] 张燕兰,李进金.广义覆盖粗集的约简[J].模糊系统与数学,2010,24(3):138-143
[17] Tsang E, Chen D G, Yeung D. Approximations and reducts with covering generalized rough sets [J]. Computers and Mathematics with Applications,2008,56:279-289
[18] 梁吉业,曲开社,徐宗本.信息系统的属性约简[J].系统工程理论与实践,2001,21(12):76-80
[19] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简[J].计算机学报,2002,25(7):759-766
[20] 苗夺谦,胡桂荣.知识约简的一种启发式算法[J].计算机研究与发展,1999,36(6):681-684
[21] Li F, Yin Y Q. Approaches to knowledge reduction of covering decision systems based on information theory [J]. Information Sciences,2009,179:1694-1704
[22] Chen D G, Wang C Z, Hu Q H. A new approach to attribute reduction of consistent and inconsistent covering decision systems [J]. Information Sciences,2007,177:3500-3518