文章编号: 1001-0920(2013)12-1837-06

基于属性分辨度的最大相容块规则提取算法

纪 霞[a], 李龙澍[b]

(安徽大学 a. 计算智能与信号处理教育部重点实验室, b. 计算机科学与技术学院, 合肥 230601)

摘 要: 提出一种基于属性分辨度的不完备决策表规则提取算法, 它是一种例化方向的方法. 首先从空集开始, 逐步选择当前最重要的条件属性对对象集分类, 从广义决策值唯一的相容块提取确定规则, 从其他的相容块提取不确定规则; 然后设计属性必要性判断步骤去除每条规则的冗余属性; 最后通过规则约简过程来简化所获得的规则, 增强规则的泛化能力. 实验结果表明, 所提出的算法效率更高, 并且所获得的规则简洁有效.

关键词: 不完备决策表; 粗糙集; 属性分辨度; 最大相容块; 规则提取

中图分类号: TP18 **文献标志码:** A

Algorithm for rules acquisition from maximal consistent blocks based on attribute discernibility

JI Xia[a], *LI Long-shu*[b]

(a. Key Lab of Intelligent Computing and Signal Processing, Ministry of Education, b. Department of Computer Science and Technology, Anhui University, Hefei 230601, China. Correspondent: JI Xia, E-mail: ahuivy1983@sina.com.cn)

Abstract: An algorithm for rules acquisition from the incomplete decision table is proposed, which uses the attribute importance measure based on discernibility. This algorithm uses a method by specialization, in which condition attributes are considered to be added to selected attributes set in order of discernibility until the selected attributes set can make the classification. Certain rules are extracted from the consistent blocks with the single generalized decision, and the uncertain rules are extracted from other consistent blocks. An attribute necessity judgment step is constructed to remove redundant attributes of each rule. Besides, a rule reduction procedure is also constructed, which helps to enhance the rule generalization ability. The experiments and comparison show that the proposed algorithm can get the simple and effective rules.

Key words: incomplete decision table; rough sets; attribute discernibility; maximal consistent block; rule acquisition

0 引 言

在现实生活中, 由于受测量误差、对数据的理解或获取的限制等各种主客观因素的影响, 所得到的决策表往往是不完备的, 而信息不完备是从实例中归纳学习的最大障碍. 因此, 如何处理不完备决策表, 从中提取可靠的规则是人工智能和数据挖掘的研究热点之一. 目前, 对不完备信息的处理通常有两种方法: 1) 通过删除不完备对象或对不完备数据进行补齐[1]将不完备决策表转化为完备决策表, 再利用完备决策表的规则提取算法获取相应的规则. 该方法改变了不完备决策表的结构, 获取的规则不能真实反映属性间的关系, 可信度不高. 2) 不改变原信息表的结构, 基于粗糙集理论[2-3]直接从不完备决策表提取规则. 管延勇等[4]提出了不完备信息系统中可信决策规则和最优可信决策规则的概念, 并给出了基于最大协调相容块的区分矩阵规则获取算法; 瞿彬彬等[5]研究了非对称相似关系下基于粗糙集理论的确定规则推理算法; Leung 等[6]提出了相似类的概念, 并在此基础上提出了确定规则和关联规则的提取算法; 蒙祖强等[7]从粒度计算的角度, 在由完全覆盖构成的粒度世界中研究了知识的表示和获取, 提出了一种完整的知识获取算法; 骆公志等[8]研究了确定优势粗糙集模型下的规则提取问题.

作为 LERS 系统的一个核心算法[9-10], LEM2 算

收稿日期: 2012-08-31; **修回日期:** 2012-12-29.

基金项目: 安徽省自然科学基金项目(60273043); 安徽大学博士科研启动基金项目(33190081).

作者简介: 纪霞(1982–), 女, 讲师, 博士, 从事不精确信息处理、粗糙集理论的研究; 李龙澍(1956–), 男, 教授, 博士生导师, 从事智能软件、不精确信息处理等研究.

法利用所有的可能值取代缺失值来计算出每个决策类的上、下近似集, 再利用属性-值对对上、下近似集的覆盖分别从上近似集中提取不确定规则, 从下近似集中提取确定规则. 由于LEM2算法不改变原始信息系统的内容和结构, 提取的规则不受缺省值的影响, 规则的准确性较高, 近年来已成为最常用的规则提取算法之一.

在粗糙集模型中, 论域总是可以表达为总体决策类的正域和边界的并集, 而总体决策类的负域恒为空, 即单个决策类的负域必定可以划分到其补决策类的正域或者边界中[11-12], 因而LEM2算法依据决策类的上、下近似集来提取规则是不必要的. 此外, 当信息系统属性缺失较多时, LEM2算法计算复杂度过高, 效率低下, 在一定程度上限制了该方法的应用.

本文针对分类问题, 提出一种在不完备决策系统中基于属性分辨度的最大相容块规则提取算法(RMA). 规则提取算法通常要考虑3个问题, 即规则提取的效率、提取规则的简洁以及规则的覆盖度. 针对这3个问题, 本文算法将采取以下措施:

1) 不再计算各决策类的上、下近似集来分别提取规则, 而是直接在原始决策系统上提取规则, 以提高规则提取的效率;

2) 算法在分类过程中不断去除已经提取规则的相容块, 以逐步缩小样本空间, 进一步提高算法的收敛速度;

3) 定义属性分辨度来度量属性的重要程度, 以保证每次都能够选取当前最重要的属性, 既提高算法的效率, 又保证规则的简洁;

4) 因为对于整个系统重要的属性, 并不一定是每条规则的必要属性, 所以构建一个属性必要性判断步骤来去掉规则中的冗余属性, 以保证规则的简洁;

5) 对已获得的规则集进一步约简, 以增强规则的覆盖度.

为了获得更多的确定规则, 约简分别在确定规则集和不确定规则集中进行. 通过应用实例和仿真实验比较发现, 本文算法能获得简洁而有效的规则, 效率更高.

1 基本概念和理论

下面介绍文中将用到的基本概念和性质[13-17].

定义 1 信息表记为$S=(U,A,V,f)$. 其中: U是对象的非空有限集合; A是属性的非空有限集合; $V=\bigcup\limits_{a\in A}V_a$是属性值的集合, V_a是属性$a\in A$的值域; f是信息函数, $f:U\times A\to V$, 即$f(x,a)\in V_a$, 它指定了论域中每一个对象的属性值. 若信息系统中至少存在一个属性$a\in A$, 使其含有空值(用$*$表示), 则称S为不完备信息系统, 记为IS, 否则是完备信息系统.

定义 2 决策表记为$\mathrm{DT}=(U,A\bigcup d,V,f)$. 其中: d表示决策属性, $*\notin V_d$. 当决策表对应的信息系统S为不完备信息系统时, 称决策表为不完备决策表, 记为IDT.

定义 3 设不完备信息表$\mathrm{IS}=(U,A,V,f)$. 其中: $x,y\in U,B\subseteq A$. 相容关系定义如下:

$$S(B)=\{(x,y)|\forall a\in B,\ f(x,a)=f(y,a)\ \text{or}\ f(x,a)=*\ \text{or}\ f(y,a)=*\}.$$

相容关系是自反且对称的, 但不满足传递性.

$\forall x\in U$, 对象x的相容类定义为

$$S_B(x)=\{y\in U|(x,y)\in S(B)\}.$$

定义 4 设不完备信息表$\mathrm{IS}=(U,A,V,f)$. 其中: $B\subseteq A,X\subseteq U,S$为$U$上的相容关系. 如果$\forall x,y\in X$, 都有$(x,y)\in S(B)$, 则称$X$为属性集$B$上的相容块; 如果不存在相容块$Y\subseteq U$, 使得$X\subset Y$, 则称$X$为属性集$B$上的最大相容块.

将属性集B上最大相容块的集合记为$\zeta(B)$, 将包含对象x的最大相容块集合记为$\zeta_B(x)$. 显然

$$\zeta_B(x)\neq\varnothing,\ \bigcup_{x\in U}\zeta_B(x)=\bigcup\zeta(B)=U.$$

定理 1 设不完备信息表$\mathrm{IS}=(U,A,V,f),B\subseteq A,X\subseteq U$, 下面的等式成立:

1) $X\in\zeta(B)$当且仅当$X=\bigcap\limits_{x\in X}S_B(x)$;

2) $S_B(x)=\bigcup\{Y\in\zeta(B)|Y\subseteq S_B(x)\}=\bigcup\{Y\in\zeta_B(x)\}$.

定理1的证明请参阅文献[16]. 相容类表达了在属性集B上与对象x相容的最大对象集合, 但同一个相容类内部的对象之间并不满足两两相容, 而最大相容块描述了相容关系下不可分辨对象的最大集合. 它准确地表达了最大相容块覆盖下对象x的信息, 既没有丢失, 也不包含任何冗余或者无关紧要的信息. 虽然两者都是基于相容关系分类的, 但最大相容块改变了普通相容类结构上的缺陷, 它内部所有的对象都满足两两相容关系, 具有更高的分类精度. 因此, 本文依据最大相容块来提取规则, 既满足分类精度的要求, 又能使提取的规则具有更高的覆盖度.

定义 5 设不完备决策表$\mathrm{IDT}=(U,A\bigcup d,V,f),B\subseteq A,V_d$是决策属性的值域, $\zeta_B(x)$是$x\in U$的相容块, 则函数∂定义为

$$\partial(\zeta_B(x))=\{f(d,x)|x\in\zeta_B(x)\}.$$

显然, 函数∂是论域到值域V_d的幂集上的映射. $\partial(\zeta_B(x))$表示$\zeta_B(x)$中对象在决策d上取值的集合.

定义 6 设不完备信息表$\mathrm{IS}=(U,A,V,f)$, 设t

$=(a,v)$ 表示一个属性-值对. 其中: $a\in A, v\in V$, $[t]$ 表示满足该属性-值对的对象集合. 根据空值的语义, $[t]$ 的计算又分为两种情况: 对于遗漏型空值, $[t]=\{x\in U|a(x)=v \text{ or } a(x)=*\}$; 对于缺席型空值, $[t]=\{x\in U|a(x)=v\}$. 本文则基于遗漏型空值语义进行相关计算.

定义 7　设不完备决策表 $\mathrm{IDT}=(U, A\bigcup d, V, f), B\subseteq A$, 记 $U/B=\{X_1, X_2, \cdots, X_m\}$, 设 $X'=\{X_i||\partial(X_i)|=1\}$, 记 $U'=U-X', A'=A-B$, 称 $\mathrm{IDT}'=(U', A'\bigcup d, V, f)$ 为原决策表的 B 简化决策表.

2　属性重要性度量

通常, 属性的重要度都是通过比较去掉属性前后决策属性对条件属性集依赖度的变化表示的, 引起依赖度发生变化的直接原因是正域的变化. 进一步分析, 所谓正域变化就是去掉某属性之前某些对象是可分辨的, 而去掉该属性之后这些对象变成不可分辨的了, 因而正域和依赖度发生了变化. 本文将属性的这种能力称为分辨能力. 从外在表现看, 如果在同一个决策类内部, 属性的取值较为一致, 而在不同决策类之间该属性值的差异较大, 则该属性的分辨能力较强. 在不完备决策表中影响属性分辨度的还有一个重要因素, 就是属性值的缺失率. 缺失率越高, 属性的分辨能力越低.

对于属性约简或规则提取而言, 目的是找出整体分辨能力最大的最小属性集合, 而不是找出一个个单独具有高分辨度的属性集合, 所以属性彼此之间的重复关联程度也是影响属性分辨能力的一个重要因素. 同时, 本文算法在分类过程中不断去除已经提取规则的相容块, 以逐步缩小样本空间, 进一步提高算法的收敛速度, 因此在规则提取的计算过程中, 属性的分辨能力也是一个动态变化的过程. 根据以上分析, 给出属性分辨度的定义如下.

定义 8　不完备决策表的 B 简化决策表 $\mathrm{IDT}'=(U', A'\bigcup d, V, f)$, 假设 $U'/d=\{D_1, D_2, \cdots, D_k\}, \forall a\in A'$, 属性 a 相对决策 d 的 B 简化属性分辨度定义为

$$\mathrm{DIS}_B(a,d)=\sum_{V_{ai}\in V_a}\frac{\max(|X_i|/|D'_j|)}{|V_a|}-\frac{|N_a|}{|U|}.$$

其中: $X_i=\{x\in D'_j|f(x,a)=V_{ai}\}, N_a=\{x\in U'|f(x,a)=*\}$.

属性 a 相对决策 d 的属性分辨度由属性 a 的分辨能力和属性 a 的缺失率两部分组成. 其中分辨能力 $0\leqslant(\max(|X_i|/|D'_j|))/|V_a|\leqslant 1$, 缺失率 $0\leqslant|N_a|/|U|\leqslant 1$, 所以 $-1\leqslant\mathrm{DIS}_B(a,d)\leqslant 1$. 当属性 a 在同一个决策类中取值一致, 在不同决策类间取值完全不同, 且属性 a 没有缺失时, 属性 a 的分辨能力为 1, 缺失率为 0, 属性 a 相对决策 d 的分辨度最高, 为 1; 当属性 a 的缺失率为 1 时, a 的分辨能力为 0, 属性的分辨度最低, 为 -1.

3　规则提取算法

3.1　算法思想

规则提取通常有两种方法, 即泛化和例化. 本文基于属性分辨度的最大相容块规则提取算法采用例化的方法. 算法从空集开始, 依据属性的分辨度选择当前最重要的属性, 对当前的相容块依据该属性进一步细化. 当细化后相容块的广义决策函数值唯一时即可提取一条确定规则, 并删除该相容块; 对于细化后广义决策函数决策值不唯一的相容块, 判断该条件属性加入前后相容块广义决策函数值有无减少, 如果没有变化, 则说明该属性对于这条待提取规则是不必要的, 此时应从该条待提取规则的条件属性中删除该属性. 重复上述过程, 直至待选条件属性集为空或者规则已经覆盖全部论域. 最后依据规则覆盖的对象是否为最大相容块对获取的规则进行约简.

3.2　算法描述

算法 1　RMA.

输入: 不完备决策表 $\mathrm{IDT}=(U, A\bigcup d, V, f)$;

输出: 规则集 Rules.

begin

1) 初始化阶段.

$B_0=\varnothing$; / B_i 为已选择属性集/

Rules 1$=\varnothing$; / Rules 1 为已提取确定规则集/

Rules 2$=\varnothing$; / Rules 2 为已提取不确定规则集/

$U'=U$; / U' 是未被 Rules 规则覆盖的对象集/

$A'=A$; / A' 是未被选择的属性集/

$i=1$.

2) 规则获取阶段.

while $(U'\neq\varnothing)$ and $(A'\neq\varnothing)$ do

begin

计算 A' 中属性相对决策 d 的分辨度, 从中选择分辨度最高的属性, 若这样的属性不止一个, 则任取一个, 记为 a_i;

$B_i=B_{i-1}\bigcup a_i; A'=A'-a_i$;

$U'/B_i=(U'/B_{i-1})/a_i=X_{i1}, X_{i2}, \cdots, X_{ik}$;

/ 根据属性 a_i 对当前相容块进一步细化/

for $j=1$ to k do

begin

$B(X_{ik})=B(X_{ik})\bigcup a_i$; / 将当前分类属性加入到所有待提规则的相容块条件属性集/

if $|\partial(X_{ij})|=1$ then / 此时有新规则生成/

begin

$R=\{B(X_{ij})\rightarrow\partial(X_{ij})\}$;

$\text{Rules}\,1 = \text{Rules}\,1 \bigcup R$;

delete X_{ij};

end

if $|\partial(X_{ij})| \neq 1$ then { if $|\partial(X_{ij})| = |\partial(U')|$ then $B(X_{ij}) = B(X_{ij}) - (a_i = v_j)$;}; /说明该条件属性相对该相容块中对象集是冗余属性, 删除/

end

$i++$;

end

while $U'/B_i \neq \varnothing$ do /导出不确定规则/

begin

$R = B(X_{ij}) \to \partial(X_{ij})$;

$\text{Rules}\,2 = \text{Rules}\,2 \bigcup R$;

delete X_{ij};

end

3) 规则约简阶段.

for 规则集 Rules 1 中的每条规则 R do

判断规则 R 覆盖的相容块对象集是否为最大相容块, 若不是, 则删除对应的规则;

for 规则集 $\text{Rules}\,2$ 中的每条规则 R do

判断规则 R 覆盖的相容块对象集是否为最大相容块, 若不是, 则删除对应的规则;

end.

整个算法分为 3 个阶段: 第 1 阶段为初始化阶段, 为相关变量指定初始值; 第 2 阶段为原始规则提取阶段, 并通过条件属性的必要性判断步骤进行规则约简, 确保每条规则都没有冗余条件属性; 第 3 阶段为规则约简阶段, 只保留覆盖最大相容块的规则, 确保获取的规则具有最高的覆盖度, 泛化能力更强. 需要说明的是, 最大相容块在完备决策表中自然退化为等价类, 所以本文算法同样适用于完备决策表的规则提取.

3.3 算法复杂度分析

假设决策表的条件属性个数为 $|A|$, 对象个数为 $|U|$. LEM2 算法的复杂度为 $O(|A|^2|U|^3)$[4], 其相关改进算法大多通过在不完备信息系统中引入新的扩充粗糙集模型来实现, 缺少对算法本身的改进, 因而算法复杂度也是 $O(|A|^2|U|^3)$[17]. 在本文算法中, 第 2 个阶段计算属性分辨度的复杂度为 $O(|A||U|)$, 对相容块细化算法采用文献 [18] 的方法, 其复杂度为 $O(|U|\log|U|)$, 计算相容块的广义决策函数复杂度为 $O(|U|)$, 所以整个规则获取阶段总的算法复杂度为 $O(|A|^2|U|)$. 在规则约简阶段, 为方便计算, 假设共提取 m 条规则, 每条规则平均覆盖 k 个对象, 则规则约简阶段的算法复杂度为 $(m-1)k^2 + (m-2)k^2 + \cdots + k^2 = (m(m-1)/2)(k)2 = O(m^2k^2)$. 整个规则提取算法的复杂度为 $\max(O(|A|^2|U|), O(m^2k^2))$, 通常 $m \ll |U|, |U|/m \leqslant k \leqslant |U|$, 故与 LEM2 算法相比, RMA 算法有效地提高了规则提取的效率.

4 实验分析

首先以文献 [19] 中的不完备决策表为例, 如表 1 所示. 其中论域 $U = \{1,2,3,4,5,6\}$, 条件属性集 $A = \{\text{humidity, temperature, rainfall, air-quality}\}$, 分别简记为 H, P, R, Q, 条件属性值域为 {high, medium, low}, 决策属性集 D ={appraisal}, 简记为 D, 决策属性值域为 {verygood, good, bad}.

表 1 不完备决策表

day	H	P	R	Q	D
1	*	*	high	medium	bad
2	high	medium	low	medium	good
3	medium	*	low	medium	good
4	high	*	low	high	good
5	medium	high	low	*	good
6	*	*	low	high	verygood

表 1 有 3 个决策类, 分别为 $X_{\text{bad}} = \{1\}, X_{\text{good}} = \{2,3,4,5\}, X_{\text{verygood}} = \{6\}$.

容差关系下, 基于单个对象计算各决策类的上、下近似集如下:

$\overline{X}_{\text{bad}} = \{1\}$; $\overline{X}_{\text{good}} = \{2,3,4,5,6\}$;

$\overline{X}_{\text{verygood}} = \{4,5,6\}$; $\underline{X}_{\text{bad}} = \{1\}$;

$\underline{X}_{\text{good}} = \{2,3\}$; $\underline{X}_{\text{verygood}} = \varnothing$;

属性-值对如下:

$[t_1] = [(H, \text{high})] = \{1,2,4,6\}$;

$[t_2] = [(H, \text{medium})] = \{1,3,5,6\}$;

$[t_3] = [(P, \text{high})] = \{1,3,4,5,6\}$;

$[t_4] = [(P, \text{medium})] = \{1,2,3,4,6\}$;

$[t_5] = [(R, \text{high})] = \{1\}$;

$[t_6] = [(R, \text{low})] = \{2,3,4,5,6\}$;

$[t_7] = [(Q, \text{high})] = \{4,5,6\}$;

$[t_8] = [(Q, \text{medium})] = \{1,2,3,5\}$.

分别采用 LEM2 算法和文中的 RMA 算法对表 1 提取规则, 规则提取过程和结果如表 2 和表 3 所示.

从提取规则数、确定规则数、提取规则时间、提取规则的长度和覆盖度 5 个方面对两组实验结果进行比较, 结果如表 4 所示.

从提取规则的数目看, 两组算法获得的确定规则数目相同, 但本文 RMA 算法获取的不确定规则少一条. 原因是 RMA 算法是根据最大相容块来提取不确定规则, LEM 2 算法则是从不同的决策出发来提取规则, 因为信息的缺失, 最大相容块的广义决策个数通常大于 1, 所以提取的不确定规则会多一些. 从提取规则的时间、提取规则的长度和规则的覆盖度 3 个方面

表 2　基于 LEM2 算法对表 1 提取规则

近似集	算法过程简单描述	所得规则
$\underline{X}_{\text{bad}}=\{1\}$	$G=\{1\}$ 内循环 while, $T=t_5$ 获得一条确定规则, 结束循环	$(R,\text{high})\rightarrow(A,\text{bad})$
$\underline{X}_{\text{good}}=\{2,3\}$	$G=\{2,3\}$ 内循环 while, $T=t_8\bigcup t_4\bigcup t_6$ 获得一条确定规则, 结束循环	$(Q,\text{medium})\bigwedge(P,\text{medium})\bigwedge(R,\text{low})\rightarrow(A,\text{good})$
$\overline{X}_{\text{good}}=\{2,3,4,5,6\}$	$G=\{2,3,4,5,6\}$ 内循环 while, $T=t_6$ 获得一条不确定规则, 结束循环	$(R,\text{low})\rightarrow(A,\text{good})$
$\overline{X}_{\text{verygood}}=\{4,5,6\}$	$G=\{4,5,6\}$ 内循环 while, $T=t_7$ 获得一条不确定规则, 结束算法	$(Q,\text{high})\rightarrow(A,\text{verygood})$

表 3　基于 RMA 算法对表 1 提取规则

近似集	算法过程简单描述	所得规则
$H=2/4-2/6=1/6$, $P=1/4-4/6=-5/12$, $R=(1+1)/2=1$, $Q=(1+1)/2-1/6=5/6$	选择属性 R, $B_1=\{R\}$, $U'/B_1=\{\{1\},\{2,3,4,5,6\}\}$, 获得一条确定规则, $U'=2,3,4,5,6$, 结束循环	$(R,\text{high})\rightarrow(A,\text{bad})$
$H=2/4-1/5=3/10$, $P=1/4-3/5=-7/20$, $Q=(1+1)/2-1/5=4/5$	选择属性 Q, $B_2=\{R,Q\}$, $U'/B_2=\{\{2,3,5\},\{4,5,6\}\}$, 获得一条确定规则, $U'=\{4,5,6\}$, $\lvert\partial(\{4,5,6\})\rvert=\lvert\partial(\{2,3,4,5,6\})\rvert$, 从 $\{4,5,6\}$ 的条件属性里删除 Q, 结束循环	$(Q,\text{medium})\bigwedge(R,\text{low})\rightarrow(A,\text{good})$
$H=1-1/3=2/3$, $P=1-2/3=1/3$	选择属性 H, $B_3=\{R,H\}$, $U'/B_3=\{\{4,6\},\{5,6\}\}$, $\lvert\partial(\{4,6\})\rvert=\lvert\partial(\{4,5,6\})\rvert$, $\lvert\partial(\{5,6\})\rvert=\lvert\partial(\{4,5,6\})\rvert$, 删除 H, $B_3=\{R\}$, 结束循环	相容块的广义决策函数值大于 1, 并且待选属性集不为空, 此轮循环无法获取规则
$P=1-2/3=1/3$	选择属性 P, $B_4=\{R,P\}$, $U'/B4=\{\{4,6\},\{5,6\}\}$, 情况同上, 删除 P, $B_4=\{R\}$, 获取一条不确定规则, $A'=\varnothing$, 结束算法	$(R,\text{low})\rightarrow(A,\text{good})\bigvee(A,\text{verygood})$

表 4　两组实验结果数据

算法	规则数	确定规则数	提取时间/s	规则长度	平均覆盖度
LEM2	4	2	0.035	6	2.75
RMA	3	2	0.013	4	3

看, 本文的RMA算法优于LEM2算法, 尤其是规则提取时间, 6个对象的决策表规则提取时间缩短了62.9%.

下面通过两组大数据集的仿真实验进一步验证RMA算法提取规则的有效性.

第1组实验选用UCI数据库[20]中的完备数据集Iris、不完备数据集Breast cancer和Hepatitis进行实验. 其中: Iris数据集包含150个对象, 4个属性; Breast cancer数据集包含286个对象, 9个属性, 含0.2%的缺失属性; Hepatitis数据集包含155个对象, 19个属性, 含5.7%的缺失属性. 对上述3个数据集中的连续属性采用Entropy/MDL算法进行离散化处理[21]. 从3个数据集各随机选择50%的数据作为学习样本, 其余的50%作为分类测试样本. 分别采用LEM2算法和RMA算法对上述3个数据集提取规则, 从规则数目、规则长度、提取时间以及规则匹配度4个方面对实验结果进行比较(见表5). 其中规则匹配度定义为能够与规则集中规则正确匹配的测试样本占总测试样本的比例.

表 5　RMA 算法与 LEM2 算法性能比较

算法	数据集	规则数	平均规则长度	提取时间/s	匹配度/%
LEM2	Iris	10	1.40	0.87	96.00
	Breast cancer	37	4.36	2.46	86.01
	Hepatitis	22	3.60	3.38	79.49
RMA	Iris	6	1.33	0.23	94.67
	Breast cancer	32	4.05	1.33	84.61
	Hepatitis	17	3.17	1.59	79.23

第2组实验. 对完备数据集Iris通过随机去除1%、5%、10%和30%的属性值得到4个不完备数据集, 分别用RMA算法对4个数据集提取规则. 对每个数据集随机选择50%数据作为学习样本, 其余的50%作为分类测试样本. 从规则数目、规则长度、提取时间以及规则匹配度4个方面对结果进行比较. 为避免单次算法的随机性, 每组实验进行10次, 取10次结果的平均值, 如表6所示.

表 6　RMA 算法从不完备 Iris 数据集提取规则

数据集	规则数	平均规则长度	提取时间/s	匹配度
Iris (1%)	6.1	1.33	0.25	93.87
Iris (5%)	6.7	1.37	0.29	92.93
Iris (10%)	7.4	1.49	0.37	92.26
Iris (30%)	8.8	1.66	0.59	89.33

先分析表5中的数据. 在匹配度上, RMA算法与LEM2算法相差最多的是在Breast cancer数据集上, 差距为1.4%, 可见RMA算法具有与LEM2算法相当

的分类精度. 但在规则数目、平均规则长度和提取时间3项指标上, RMA算法明显优于LEM2算法, 尤其是提取时间, 在相同数据集的情况下, RMA算法几乎只用了LEM2算法时间的一半, 说明RMA算法用较少的规则获得了较高的分类精度, 是一种有效的不完备决策表规则提取算法, 兼顾了有用性与实用性.

下面分析表6中的数据. 随着Iris数据集属性值缺失比例的扩大, RMA算法在规则数目、规则长度、提取时间和规则匹配度4个方面性能都在下降, 但与数据缺失程度的增加幅度相比较, RMA算法在规则数目、规则长度和规则匹配度3个性能上都保持了稳定, 唯一随之变化的是规则提取时间. 这个实验结果是比较合理的, 因为规则数目、规则长度和规则匹配度虽然受属性缺失的影响, 但也与缺失属性的分布密切相关, 而提取时间则不然, 只要对象的属性值存在缺失, 就会影响对象的分类, 从而导致相容块细化过程中工作计算量的增加, 同时在最后的规则约简阶段, 最大相容块的查找也需要比较更多的相容块, 增加了时间的耗费. 综上分析, 在数据不完备程度较为严重的情况下, RMA算法也能保证规则提取的性能, 是一种比较实用的不完备决策表规则提取算法.

5 结 论

本文提出了一种基于属性分辨度的最大相容块规则提取算法, 它以条件属性相对决策的分辨能力来度量属性的重要性, 算法效率更高. 设计了属性必要性判定步骤, 以去除每条规则中的冗余属性, 从而保证规则的简洁. 规则约简过程只保留最大相容块上提取的规则, 使得规则具有更高的覆盖度和泛化能力. 实验结果表明, 算法获得的规则简洁、规模小, 具有较好的分类精度和较强的泛化能力. 本文算法仅适用于遗漏型的不完备决策表, 对于缺席型空值语义下不完备决策表的规则获取将是下一步的研究重点.

参考文献(References)

[1] Hong T P, Tseng L H, Wang S L. Learning rules from incomplete training examples by rough sets[J]. Expert Systems with Applications, 2002, 22(4): 258-293.

[2] Pawlak Z, Grzymala-Busse J W, Slowinski R, et al. Rough sets[J]. Communucations of the ACM, 1995, 38(11): 88-95.

[3] Pawlak Z, Skowron A. Rough sets: Some extensions[J]. Information Science, 2007, 177(1): 28-40.

[4] 管延勇, 薛佩军, 王洪凯. 不完备信息系统的可信决策规则提取与E-相对约简[J]. 系统工程理论与实践, 2005, 25(12): 76-82.
(Guan Y Y, Xue P J, Wang H K. Credible decision rules acquisition and E-relative ruduct in incomplete information systems[J]. Systems Engineering-Theory & Practice, 2005, 25(12): 76-82.)

[5] 瞿彬彬, 卢炎生. 基于粗糙集的不完备信息系统规则推理算法[J]. 小型微型计算机系统, 2006, 27(4): 798-700.
(Qu B B, Lu Y S. Rule induction algorithm based on rough sets for incomplete information system[J]. Mini-Micro Systems, 2006, 27(4): 798-700.)

[6] Yee Leung, Wu Wei-zhi, Zhang Wen-xiu. Knowledge acquisition in incomplete information systems: A rough set approach[J]. European J of Operational Research, 2006, 168(1): 164-180.

[7] 蒙祖强, 史忠植. 不完备信息系统中基于相容粒度计算的知识获取方法[J]. 计算机研究与发展, 2008, 45(增): 264-267.
(Meng Z Q, Shi Z Z. An approach to acquire knowledge in IIS based on tolerance granular computing[J]. J of Computer Research and Development, 2008, 45(S): 264-267.)

[8] 骆公志, 黄卫东. 不完备信息系统中的确定优势粗糙决策规则提取[J]. 南京邮电大学学报, 2011, 31(5): 114-120.
(Luo G Z, Huang W D. Extracting decision rules from incomplete information decision system by rough set model based on definitive dominance relation[J]. J of Nanjing University of Posts and Telecommunications, 2011, 31(5): 114-120.)

[9] Grzymala-Busse J W. A new version of the rule induction system LERS[J]. Fundamental Information, 1997, 31(1): 27-39.

[10] Grzymala-Besse J W. MLEM2: A new algorithm for rule induction from imperfect data[C]. Proc of the 9th Int Conf on Information Processing and Management of Uncertainty in Knowledge-based Systems. Annecy, 2002: 243-250.

[11] Yao Y Y, Wong S K M, Lingras P. A decision-theoretic rough set model[C]. Proc of the 5th Int Symposium on Methodologies for Intelligent Systems. Knoxville, 1990: 17-24.

[12] Yao Y Y, Wong S K M. A decision-theoretic framework for approximating concepts[J]. Int J of Man-machine Studies, 1992, 37(6): 793-809.

[13] Grzymala-Besse J W. Data with missing attributes values: Generalization of indiscernibility relation and rule induction[J]. Trans on Rough Setttsss, 2004, 31(1): 78-95.

[14] Kryszkiewicz M. Rough set approach to incomplete information system[J]. Information Sciences, 1998, 112(1): 39-49.

[15] Grzymala-Besse J W, Grzymala-Besse W J. Handling missing attribute values[M]. New York: Springer, 2006: 37-57.

[16] Leung Y, Li D Y. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Science, 2003, 153(1): 85-106.

[17] 徐怡, 李龙澍, 李学俊. 改进的LEM2规则提取算法[J]. 系统工程理论与实践, 2010, 30(10): 1841-1849.
(Xu Y, Li L S, Li X J. Improved LEM2 rule induction algorithm[J]. Systems Engineering-Theory & Practice, 2010, 30(10): 1841-1849.)

[18] Meng Z Q, Shi Z Z. A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets[J]. Information Sciences, 2009, 179(1/2): 2774-2793.

[19] 徐晓东, 沈惠璋, 王资凯. 基于非对称相似粗糙集的规则获取算法[J]. 计算机仿真, 2008, 25(10): 110-113.
(Xu X D, Shen H Z, Wang Z K. A rule extraction algorithm based on asymmetrical similarity rough set[J]. Computer Simulation, 2008, 25(10):110-113.)

[20] Blake C M. UCI machine learning repository[DB/EL]. [2009-03-08]. http://www. ics.uci.edu/mlearn/databases/.

[21] Komorowski O J. ROSETTA—A rough set toolkit for analysis of data[C]. Proc of the 3rd Int Joint Conf on Information Sciences. Berlin: Springer, 1997: 403-407.