DOI: 10.13656/j.cnki.gxkx.2009.04.029
广西科学 Guangxi Sciences 2009, 16(4): 389~391

基于贝叶斯粗糙集模型的属性约简算法*

Attribute Reduction Algorithm of Bayesian Rough Sets Model

陈 胜,曾雪兰,梅良才
CHEN Sheng, ZENG Xue-lan, MEI Liang-cai

(广西大学数学与信息科学学院,广西南宁 530004)
(College of Mathematical and Information Science, Guangxi University, Nanning, Guangxi, 530004, China)

摘要:在分析贝叶斯粗糙集模型已有的约简算法的基础上,从含有多个决策类情况下的全局相对增益函数的角度,利用二进制编码方法给出一种求贝叶斯粗糙集所有约简及核的算法,并基于实际应用,给出对求出的所有约简进行筛选的拓展算法。通过算例说明算法的实用性和有效性。

关键词:粗糙集 属性约简 R约简 贝叶斯粗糙集模型
中图法分类号:O159, TP301.6 文献标识码:A 文章编号:1005-9164(2009)04-0389-03

Abstract With an analysis based on previous method of redution algorithm of bayesian rough sets modle, from the view of the global relative gain function contained of multiple decision classes, based on the binary code method, an algorithm for solving all reductions and nuclear is given. Furthermore, the proposed algorithm is extended from the point of view of the practical appliction, a method to screening all of the reduction is propsed, and the output is the optimal reduction. Finally, an example is used to demonstrate that the proposed algorithm is practical and effective.

Key words rough sets, attribute reduction, R reduction, Bayesian rough sets mode

粗糙集理论是波兰数学家[1]1982提出的,现在已经成为一种新的处理模糊和不确定性知识的数学工具。粗糙集理论的中心问题是分类分析,而粗糙集模型在实际应用中存在一定的局限性,它所处理的分类必须是完全正确的或肯定的,即"包含"或"不包含",而没有某种程度上的"包含"或"属于"[2]。基于这些情况[3],于年对粗糙集模型进行推广提出变精度粗糙集模型,它是在粗糙集模型的基础上引入了U(0≤U<0.5),即允许一定程度的错误分类率存在。而在数据挖掘等实际应用中我们只需根据获得信息去处理问题,不需要受预先给定的某个参数的限制,并且参数也是有决策人主观设定的,因此文献[3]进一步给出了贝叶斯粗糙集模型,它是把变精度粗糙集中的参数U用先验概率来代替。

到目前为止关于贝叶斯粗糙集模型应用的文献很多,但是基于此模型的属性约简算法并不多[4,5]。文献[6]利用文献[5]的上、下分布约简讨论了贝叶斯粗糙集模型的属性约简,但是并没有给出具体的算法。文献[4]用全局相对增益作为属性重要度,并以此作为启发信息提出一种贝叶斯粗糙集属性约简的启发式算法。我们分析贝叶斯粗糙集模型后,在含有多个决策类[4]的情况下,从全局相对增益函数[7]的角度,利用二进制编码方法给出一种求贝叶斯粗糙集模型所有基于全局相对增益函数的约简(以下简记为 R 约简)及核的算法,并求出核。在做不确定决策实际应用时,通常只利用最重要的约简,因此本文又对算法进行拓展,给出对求出的约简进行筛选的拓展算法,输出重要性最好的约简。本算法实用,有效,在实际的数据处理中具有广阔的应用前景。

1 约简算法的理论基础及基本思想

首先给出全局相对增益函数的定义及相关定理。

收稿日期:2009-01-15
修回日期:2009-09-25
作者简介:陈 胜(1983-),男,硕士研究生,主要从事粗糙集研究。
* 国家自然科学基金项目(70861001),广西自然科学基金项目(桂科自0991027),广西研究生教育创新计划项目(105930903069)资助。

定义 **1**[7] 信息系统 S中,对于 $E\subseteq C, U/D=\{X_1, X_2, \cdots, X_r\}$, 则称 $R_E(D)=\sum_{[x]_E}\max\{P([x]_E|X_i), i=1,2,\cdots,r\}-1$为 E相对于决策属性 D的全局相对增益函数。

定理 **1**[7] 信息系统 S中,对于 $\forall B\subseteq A$和 $X\subseteq U$有: $R_B(X)\leqslant R_A(X)$。其中 $POS^*(X)=\cup\{E: P(X|E)>P(X)\}$

定理 **2**[7] 对于 $\forall B\subseteq A$,若 B为一个 R约简当且仅当 $R_B(X)=R_A(X)$,且不存在 B子集使得上式成立。

在含有多个决策类的情况下,由定理 2,利用二进制编码的方法求出不同编码组合的属性集的全局相对增益函数值,并以此得到所有 R约简及核的算法。算法的基本思想: 首先,将决策表条件属性的各种组合编成二进制编码,编码的方法在此不再赘述;其次,将所有编码组合对应的属性集相对于决策属性的分类质量与条件属性相对于决策属性的分类质量相比较,如果相等则将该属性集作为矩阵的一行;再次,根据准则删除那些子集已经是约简的行,以及全为零的行;最后,求出信息系统的所有 R约简组成的矩阵,即得到所有 R约简。

2 算法描述

算法的步骤是输入 $S=(U, A, V, f)$, $A=C\cup D$, $C=\{c_1, c_2, \cdots, c_m\}$,输出贝叶斯粗糙集的所有的 R约简。

步骤 1 令 $w=2^m-1$,令 M为 $w\times m$的零矩阵,并令 $j=2^m-1$;

步骤 2 对 j进行二进制编码,根据 $P(X)$用轮盘赌的形式从大到小逐个计算每个二进制编码中 1所对应的属性组合 P的全局相对增益函数值 $R(X)$;

步骤 3 如果 $R_P(X)=P_C(X)$成立,则用属性组合 P所对应的二进制编码替换 M中的第 j行,否则继续运算;

步骤 4 $j=j-1$,如果 $j\geqslant 1$,返回步骤 2,否则继续;

步骤 5 根据准则删除 M中冗余的行,并删除行全为零的行。剩余的行对应的属性集即为贝叶斯粗糙集的所有属性约简;

步骤 6 求出核 $CoreC=\cap redP$。

算法的时间复杂度为 $O(2^{|C|-1}|U|^2)$。

3 算例分析

以表 1为例对贝叶斯粗糙集 R约简算法进行说明。表 1中 U包含 7个对象集 $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$,决策表的前 4列为条件属性集 $\{a, b, c, d\}$,第 5列为决策属性集 $\{e\}$,属性值集合为 $V_a=\{1, 2, 3, 4\}$, $U/D=\{Y_1, Y_2, Y_3\}$, $X_1=\{X_1\}$, $X_2=\{X_2, X_3\}$, $X_3=\{X_4, X_5, X_6, X_7\}$。步骤 1由二进制编码的 $w=15$;步骤 2计算得 $j=15$时 $R_C(X)=2$,然后以轮盘赌的形式从大到小逐个计算每个 j对应的全局相对增益函数值 $R(X)$;步骤 3逐个验证 $R_P(X)=P_C(X)$成立,其中当 $j=5, 7, 12, 13, 14, 15$时, $R_P(X)=P_C(X)=2$,并得到矩阵 M;跳过步骤 4;步骤 5根据准则除去 M中冗余的行,剩余的行所对应的属性集为 $\{a, c\}$和 $\{c, d\}$,即为贝叶斯粗糙集表 1的全部属性约简,并且核为 $CoreC=\{a, c\}\cap\{c, d\}=\{c\}$。此结果与文献[7]相比,得到了决策表的所有属性约简及核。

表 **1** 决策表

Table 1 Dicision table

U	a	b	c	d	e
X_1	1	3	1	3	1
X_2	2	1	2	1	3
X_3	4	3	2	2	3
X_4	1	3	3	3	2
X_5	3	1	3	3	2
X_6	2	1	3	1	2
X_7	1	3	2	3	2

4 算法拓展

在利用粗糙集做不确定决策时,有必要求出决策表的所有约简及核,但是在实际应用中通常只用最重要的属性约简,因此有必要对上述贝叶斯粗糙集 R约简算法得到的属性约简进行筛选。对求出的约简进行筛选的拓展算法步骤是输入所有 R约简,输出最优约简。

步骤 1 求出包含所有约简属性的集合 $N=\cup redP$;

步骤 2 对集合 N中的属性进行二进制编码并算出每个属性的全局相对增益函数值 $R(X)$;

步骤 3 求出所有约简所包含属性的全局相对增益函数值 $R(X)$的和: $R_{redP}=\sum_{r\in redP}R_r(X)$;

步骤 4 将所有 R_{redP}从大到小排列,那么输出最大值所对应的 R约简,即为属性重要性最大的属性约简。

应用拓展算法计算表 1算例,计算各属性的重要性分别为 $R_a(X)=1.25, R_c(X)=1.75, R_d(X)=1$。所以 2个属性约简的属性重要性总和分别为 $\sum_{a,c}R(X)=3$, $\sum_{a,c}R(X)=2.75$;最后经过筛选输出约简 $Red=\{a, c\}$。

5 结束语

我们在分析贝叶斯粗糙集模型的基础上,从含有多个决策类情况下的全局相对增益函数的角度,利用二进制编码方法给出一种求贝叶斯粗糙集模型所有 R 约简及核的算法。而在实际做不确定决策应用时,通常只利用最重要的约简,因此我们又对该算法进行拓展,给出对求出的约简进行筛选的拓展算法,输出重要性最好的约简。

本算法思想简洁易懂,在实际处理大数据量时可以采用 MATLAB编程实现,在实际的数据处理中具有广阔的应用前景。但是我们同时也发现该算法的时间复杂度以属性个数呈几何增长,所以算法应用要约定属性的个数不超过 20个。另外,在贝叶斯粗糙集模型的定义中我们还发现一些不足:(1) 在实际的数据处理时,当 $P(X)$很小时,满足 $P(X|E) > P(X)$的所有等价类 E都成为了正域的集合,这就意味着正域中有一些等价类含有很少的 X 中的元素。这很不符合定义正域的思想。当 $P(X)$ 很大时,使用负域的定义会损失大量的信息。这也不符合改进传统粗糙集模型以获得更多信息的思想。(2) 在该模型中满足 $P(X|E) = P(X)$ 条件的等价类在实际计算中很少,因此在多数情况下,边界区域是空集。这不符合改进传统粗糙集模型更柔性化的思想。

基于以上不足,对贝叶斯粗糙集模型定义的改进,能够更好地应用于智能信息处理,这将是我们下一步要解决的问题。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences, 1982, 11(5): 341–356.

[2] 张文修,吴德伟,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.

[3] Ziarko W. Variable precision rough set model[J]. Journal of Computer and System Science, 1993, 46(1): 39–59.

[4] 蔡娜,张雪峰.基于贝叶斯粗糙集模型的属性约简[J].计算机工程,2007, 12(33): 93–96.

[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.

[6] 王虹,张文修.关于贝叶斯粗糙集约简模型的知识约简[J].计算机科学,2005, 32(11): 150–151.

[7] Dominik S, Ziarko W. The investigation of bayesian rough set model international[J]. Journal of Approximate Reasoning, 2005, 40: 81–91.

(责任编辑:邓大玉)

模拟光合作用产生石油替代能源

光合作用广泛存在于自然界,叶绿体收集太阳光能,将水和二氧化碳转化为有机物(首先是葡萄糖),并释放出氧气。但这只是最终结果,整个过程一开始是将水和二氧化碳气转化为氧,自由的质子和电子。在光合作用中产生了两个化学反应,叶绿素分子失去两个电子,水分子发生分解。尽管光合作用在各种教科书中都得到了详尽的阐述,但是想人工实现这一过程却绝非易事,主要的问题在于缺少有效地电解水的媒介,在植物中充当这一媒介的是叶绿体。最近美国科学家把用铟氧化物和锡氧化物做成的电极放置在钴离子和磷酸钾水溶液中,然后在溶液中通入太阳能电池的电流,产生了相当于叶绿体的触媒,把水分解成氧气和自由的氢离子,这些氢离子聚集在电极上,并在那里形成氢气。白天,用通常方法获得的太阳能一部分可用于日常所需要,一部分用来将水分解成氢气和氧气并将氢气储存起来。晚上,氢气和氧气转化为燃料用来发电。美国科学家在实验实内模拟光合作用的过程,将水分解成氢和氧,并产生了可供燃烧的氢气和氧气。这是利用太阳能历史上的一个大创举,将使氢气生产成为可能,使太阳能使用步入新的时代。

(据科学网)