一种基于变精度粗糙集理论的属性约简算法

周爱武,周闪闪,邹 武,李玉梅,王宝铜
(安徽大学 计算机科学与技术学院,安徽 合肥 230039;
安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:属性约简是粗糙集理论研究的核心问题之一,而且现已证明寻找一个决策表的最小约简是NP-hard问题。针对变精度粗糙集理论的属性约简问题,从相对可辨识矩阵,属性的重要度作为启发式的信息,给出变精度粗糙集的属性约简算法的改进,在一定程度上简化了属性约简的计算,提高了属性约简的效率。最后通过实例证明了改进的算法给出信息系统的属性约简结果的正确性。

关键词:属性约简;变精度粗糙集;相对差异矩阵;属性重要度
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2009)07-0035-03

An Algorithm of Attribution Reduction of Variable Precision Rough Sets Theory

ZHOU Ai-wu, ZHOU Shan-shan, ZOU Wu, LI Yu-mei, WANG Bao-tong
(School of Computer Science & Technology, Anhui University, Hefei 230039, China;
Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Attribute reduction is one of the key topics in the Rough Set theory field. It has been proved that computing the minimal reduction of decision table is an NP-hard problem. Relative discernibility matrix and attribute significance are considered to solve the attribute reduction of variable precision rough sets theory. Moreover, the improved algorithm of attribution reduction reduction is given. It can simplify the operation and enhance reduction the efficiency of seeking the reduction in some extent. At last, a practical example is given to show the validity of the algorithm.

Key words: attribute reduction; variable precision rough set relative discernibility Matrix; attribute significance

0 引 言

粗糙集(RS)[1]理论从提出到现在已经得到长足的发展,用于多个领域。不过粗糙集理存在一些局限。对此,研究者结合其他计算理论方法对经典粗糙集理论模型进行了不同的扩展。Ziarko[1]在1993年提出的变精度粗糙集模型(VPRS)[2]是对粗糙集理论的扩充,它是在基本粗糙集模型基础上引进$\beta(0.5<\beta\leqslant 1.0)$(文中采用此定义),即允许一定程度的错误分类率存在,这一方面完善了近似空间的概念,另一方面也有利于粗糙集理论从认为不相关的数据中发现相关的数据[3]。

属性约简是粗糙集理论中的一个重要的课题,通过属性约简可以去除数据库中的冗余、无用的成分,并且能够保证决策表的分类能力不会改变[4]。利用相对差异矩阵和属性的重要度的思想,结合变精度粗糙集理论给出一种启发式的属性约简的算法。

收稿日期:2008-10-15;**修回日期**:2008-12-20
基金项目:安徽省自然科学基金项目(050420204)
作者简介:周爱武(1965-),女,副教授,硕士研究生导师,研究方向为数据仓库与数据挖掘、数据库、粗糙集等。

1 基本概念

定义1 设X,Y为有限论域U的两个非空子集,令$P(X,Y)=|X\cap Y|/|X|$,称$P(X,Y)$为集合X相对于集合Y的正确分类率[5]。

定义2 设$S=(U,A=C\cup D)$为一决策信息系统,$P\subseteq C$为条件属性集,$Q\subseteq D$为决策属性集,分类$U/P=\{X_1,\cdots,X_n\}$,$U/Q=\{Y_1,\cdots,Y_n\}$,$0.5<\beta<1$,对任意的$Y\in U/Q$定义[5]:

$\mathrm{POS}_P^{\beta}(Y)=\bigcup\{X_i\in U/P\mid \mathrm{pr}(Y/X_i)\geqslant\beta\}$ 为β-正域

$\mathrm{NEG}_P^{\beta}(Y)=\bigcup\{X_i\in U/P\mid \mathrm{pr}(Y/X_i)\geqslant 1-\beta\}$ 为β-负域

$\mathrm{BND}_P^{\beta}(Y)=\bigcup\{X_i\in U/P\mid 1-\beta<\mathrm{pr}(Y/X_i)<\beta\}$

为 β - 边界域

其中 $\mathrm{pr}(Y/X_i)=\frac{|Y\cap X_i|}{|X_i|}$ 表示条件类相对于决策类的正确分类率，可以理解为可信度，β 可理解为可信度阈值。决策属性集 Q 与条件属性集 P 的 β 依赖性定义为

$$\gamma^{\beta}(P,Q)=\frac{|\mathrm{POS}(P,Q,\beta)|}{|U|}$$

其中 $|\mathrm{POS}(P,Q,\beta)|=\bigcup_{Y\in U/Q}\mathrm{POS}_P^{\beta}(Y)$。

定义 3 （相对差异矩阵）对于决策信息系统 $S=(U,A=C\cup D)$，论域 $U=\{x_1,x_2,\cdots,x_n\}$，D 为决策属性，条件属性 $C=\{c_1,c_2,\cdots,c_m\}$。设 $b((s,t),i)$ 是相对差异矩阵 $\boldsymbol{B}$ 中的一个元素，对应着属性 i 和决策表中的一对 (X_s,X_t) 记录，$s<t$。对于 $i\in\{1,2,\cdots,n\}$

$$b((s,t),n+1)=\begin{cases}1 & D(x_s)\neq D(x_t)\\ 0 & D(x_s)=D(x_t)\end{cases}$$

且 $$b((s,t),i)=\begin{cases}1 & c_i(x_s)\neq c_i(x_t)\\ 0 & c_i(x_s)=c_i(x_t)\end{cases}$$

设 $B1$ 是一个 $[(n-1)\times n/2]$ 行 $(n+1)$ 列的表。表中的每一列对应着一个属性，每一行对应着决策表中的一对不同的记录的比较，最后一列对应着决策属性。将 $B1$ 中所有的 $b((s,t),n+1)=0$ 以及 $B1$ 中的 $n+1$ 列去掉，剩余的部分就是相对差异矩阵 $\boldsymbol{B}$[6]。

定义 4 $\mathrm{red}^{\beta}(C,D)$ 是条件属性 C 相对于决策属性 D 的 β 近似约简[7]（简称为 β 约简）[7]。则有：$\mathrm{red}^{\beta}(C,D)\subseteq D$，且满足：

① $\gamma^{\beta}(C,D)=\gamma^{\beta}(\mathrm{red}^{\beta}(C,D),D)$。

②去掉 $\mathrm{red}^{\beta}(C,D)$ 中的任意一个属性都会使①不成立。

定义 5 （属性重要度）[5]设 $X\subseteq A$ 是一属性子集，$x\in A$ 是一属性，x 对于 X 的重要度记为 $\mathrm{SIG}_x(x)$。有

$$\mathrm{SIG}_x(x)=1-|X\cup\{x\}|/|X|$$

其中 $|X|$ 表示 $|\mathrm{IND}(X)|$。

设 $U/\mathrm{IND}(X)=U/X=\{X_1,X_2,\cdots,X_n\}$，则 $|X|=|\mathrm{IND}(X)|=\sum_{i=1}^{n}|X_i|^2$。

2 变精度粗糙集的属性约简

分别以上一节中给出的几个概念为基础，给出信息系统的 β 约简的算法。此种算法利用相对差异矩阵的矩阵运算来求得信息系统的属性约简，在一定程度上简化了计算，加快了速度，同时利用相对差异矩阵的运算还是收敛的[8]。

以下给出算法过程：首先要对给定数据集合的相对差异矩阵 $\boldsymbol{B}$ 找出含 1 最少的行，如果某一行只有一个 1，则该元素对应的列一定包含在属性约简中，并且通过属性的依赖度判断所加入的属性是不是所要求的结果。如果不是要删除该元素所在的行和列，再将剩余的矩阵赋给新的相对差异矩阵。如果所有行包含的 1 都不是 1 个的话，则找出这些元素所对应的列中所包含 1 最多的列，将此列加入到属性约简集合中，再次通过属性依赖度来判断所求的结果是不是约简的结果。如果不是则要将所选择的列中 1 对应的行和该列删除，再构成新的相对差异矩阵。依次这样做下去，一直到差异矩阵为空为止。

算法具体描述如下：

输入：决策系统 $S=(U,A=C\cup D)$，其中 U 为对象集合，C 为条件属性集合，D 为决策属性集合，正确分类率 β；

输出：条件属性集合 C 的约简 RED

步骤：

Step(1)：构造一个相对差异矩阵 $\boldsymbol{B}$，并且令 RED $=\varnothing$，计算 $\gamma(C,D,\beta)$；

Step(2)：分别对相对差异矩阵中的各行各列进行相加，结果分别为 R 和 C；

Step(3)：从各行相加的结果 C 中找出最小值 min 所在的行 $(r_1,r_2,\cdots,r_k)$

While（相对差异矩阵 $B\neq\varnothing$）

{

if 最小值为 min=1

找出最小值 1 所在的列。记属性 c_1，则 red=red$\cup c_1$，

If $\gamma(\mathrm{red},D,\beta)=\gamma(C,D,\beta)$

Break;

Else 删除 c_1 所在的行于列，$B\Leftarrow$新得到的矩阵

Else 最小值 min>1

在相对差异矩阵 B 中找出 min 所在行 $(r_1,r_2,\cdots,r_k)$ 中元素 1 所对应列中含有 1 最多的列 C_s，则 red=red$\cup c_s$，

If $\gamma(\mathrm{red},D,\beta)=\gamma(C,D,\beta)$

Break;

Else 删除 C_s 所在列中元素 1 所对应的行以及该列 C_s，

$B\Leftarrow$新得到的矩阵

}

Step(4)：输出属性约简 RED

按照上述算法的步骤，完全可以由一个数据集合来求出一个属性约简。每求出一个约简就要判断一下 $\gamma(\mathrm{red},D,\beta)=\gamma(C,D,\beta)$ 是否成立，如成立就可停止此算法，不必要再继续算下去。这样也提高了效率。

下面给出一个简单的算例来证明算法的正确性。利用给出的算法对表 1 进行属性约简，假定 $\beta=0.8$。根据算法的步骤，首先构造信息系统的相对差异矩阵，

并且计算 $\gamma(C,D,\beta) = 0.75$。然后根据给出的相对差异矩阵和 $\gamma(C,D,\beta)$ 的值逐步求出约简。最后求出属性约简的结果为 RED = $\{a_1,a_2,a_4\}$,而且 $\gamma(\text{red},D,\beta) = 0.75 = \gamma(C,D,\beta)$。

表 1　关于气象信息的决策系统

No	属性 C				分类 D
	Outlook(a1)	Temperature(a2)	Humidity(a3)	Windy(a4)	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Mild	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N
15	Sunny	Hot	High	True	P
16	Rain	Cool	Normal	False	N

根据给出的算法求出的属性约简是不是一个正确的约简,可以根据属性的重要度这个概念来进行检验。在参考文献[5]中给出定理:约简中的每一元对于约简中的其余的元都是重要的,约简集合外的每一元对于约简都是不重要的,重要度都为 0[5]。由这个定理,利用属性重要度公式就可以检验属性约简的结果。

3　结束语

属性约简是粗糙集理论中的一个重要的课题之一,但是信息系统的属性约简不唯一,要得到最简捷的决策规则必须得到最小约简[9]。但是要找出一个信息系统的最小约简已经证明是 NP - Hard 问题[10]。文中给出的算法通过实例证明可以有效地求出一个信息系统的 β 约简。而且给出利用属性重要度的概念来检验所求出的结果是不是正确的约简。

参考文献:

[1] Pawlak Z . Rough sets[J]. International journal of information and computer sciences,1982,11(5):341 - 356.

[2] Ziarko W. Variable precision rough set model[J]. Journal of computer and system science,1993,46(1):39 - 59.

[3] 张文修,吴伟志.粗糙集理论与方法[M].北京:科学出版社,2003.

[4] An A,Shan N,Chan C,et al. Discovering rules for water demand prediction: an enhanced rough set approach[J]. Engineering Application in Artificial Intelligence,1996,9(6):645 - 653.

[5] 史开泉,崔玉泉. S - 粗集与粗决策[M].北京:科学出版社,2006.

[6] 于兴刚.粗糙集属性约简算法在数据挖掘中的研究[D].重庆:重庆大学,2004.

[7] 于　冰.在数据挖掘中粗糙集信息约简算法的研究及应用[D].北京:中科院,2005.

[8] 夏春艳.基于粗集属性约简的数据挖掘技术的研究与应用[D].长春:长春大学,2006.

[9] 覃伟荣,秦亮曦.基于粗糙集理论的条件属性动态约简算法[J].计算机技术与发展,2008,18(8):23 - 25.

[10] 陶　志,许宝栋,汪定伟,等.基于可变精度粗糙集理论的粗糙规则的挖掘算法[J].信息与控制,2004,33(1):18 - 22.

(上接第 34 页)

4　结束语

在大中型应用程序里,无一例外地都会涉及到对象与关系数据库的处理。通过对 O/R Mapping 映射机制的深入研究,运用 NHibernate 技术将应用程序中的对象和数据库表之间建立适当的映射关系,有效地解决了面向对象设计与关系数据库设计之间的"阻抗不匹配"问题。同时,采用对象关系映射模式,使业务数据得以从业务逻辑中提取出来。随着面向对象和关系数据库的不断发展,O/R Mapping 技术将在软件开发过程中发挥日益重要的作用。

参考文献:

[1] Keller W. Persistence Options for Object - Oriented Programs Proceedings of OOP[M]. New York: Wiley Computer Publishing,2004.

[2] 孙卫琴.精通 Hiberuate:Java 对象持久化技术详解[M].北京:电子工业出版社,2005:5 - 39.

[3] 余俊新,孙　涌.J2EE 中对象关系映射的研究与实现[J].计算机技术与发展,2007,17(3):88 - 94.

[4] 林寒超,张南平. Hibernate 技术的研究[J].计算机技术与发展,2006,16(11):310 - 313 .

[5] 刘　伟.对象/关系映射在.NET 环境中的实现[D].长沙:中南大学,2007.

[6] 何　铮,陈志刚.对象/关系映射框架的研究与应用[J].计算机工程与应用,2003,39(26):188 - 189.

[7] 张淑全.基于 Hibernate 数据层设计模式的研究与实现[D].大连:大连海事大学,2007.

[8] Keller W. Object/Relational Access Layers[C]//Proceedings of the 3rd European Conference on Patterns Language of Programming and Computing. New York: Wiley Computer Publishing,1998.