一种新的属性约简算法

高 亮[1], 王 伟[1], 吴 涛[1,2]
(1.安徽大学 数学与计算科学学院,安徽 合肥 230039;
2.安徽大学 智能计算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中的一个核心问题,为了有效获取属性最小相对约简,提出了一种新的基于相对差异比较表的属性约简算法。该算法给出了一种将信息表转化为相对差异比较表的方法,且该方法对于不相容决策表也是可行的,进而就将求解最小属性约简问题转化为求解一个0-1整数规划问题,并分别采用一般求解规划问题的方法和遗传算法两种方法来求解这个0-1整数规划问题。实验结果证明该算法结合遗传算法能够更加快速有效地进行属性约简。

关键词:粗糙集;属性约简;遗传算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)05-0019-03

A New Attribute Reduction Algorithm

GAO Liang[1], WANG Wei[1], WU Tao[1,2]
(1.School of Mathematics and Computational Science of Anhui University, Hefei 230039,China;
2.Ministry of Education Key Lab. of Intelligent Computing & Signal Processing,Anhui University,Hefei 230039,China)

Abstract: Attribute reduction is a key problem for rough set theory. In order to achieve effective attribute reductions, proposes a new rough set attribute reduction algorithm based on the relative difference comparison table . At first in this algorithm, a new method which information table is translated into the relative difference comparison table is discussed, and this method can calculate the incompatible decision table, then the problem about solving the best attribute reductions will be translated into a 0-1 integral programming problem , at the same time ,use a general method and genetic algorithm to calculate the 0-1 integral programming problem respectively. The experimentation results show the algorithm is more fast and effective.

Key words: rough set;attribute reduction;genetic algorithm

0 引 言

粗糙集理论是波兰数学家Z. Pawlak于1982年提出的一种数据分析理论,它是一种新的处理模糊和不确定性知识的数学工具,其主要思想就是在保持分类能力不变的前提下通过知识约简导出问题的决策或分类规则[1,2]。目前,粗糙集理论已被成功地应用于机器学习、决策分析、过程控制、模式识别与数据挖掘等领域。

属性约简是粗糙集理论中的一个核心问题,就是在保持知识库分类能力不变的条件下删除其中不相关或不重要的冗余知识,目的是想得到一个最简洁的决策,即最小(最优)约简。但是寻找决策表的最小约简已经被证明是一个NP-hard问题,解决这类问题的方法一般是启发式搜索。目前属性约简算法[3~6]很多,主要分为两类:一类是基于可区分矩阵和区分函数;一类是基于信息熵的启发式算法。

笔者在构造相对差异比较表的基础上,将属性约简问题转化为一个0-1整数规划问题,又进一步引入"贪婪思想"改进目标函数,同时利用"遗传算法"求解新的规划问题,既克服了一般求规划问题的缺点,又提高了算法的效率,而且对于不相容决策表也是有效可行的。

1 粗糙集相关基本概念

四元组 $S=(U,A,V,F)$ 是一个信息系统,其中 $U=\{u_1,u_2,\cdots,u_{|U|}\}$ 是有限非空集,称为论域,U 中元素称为对象;$A=\{a_1,a_2,\cdots,a_{|A|}\}$ 称为属性集,A

收稿日期:2007-08-31
基金项目:973计划资助项目(2004CB318108);国家自然科学基金(60475017,60675031);安徽省自然科学基金(050420208);安徽省高等学校省级自然科学研究项目(2006KJ244B);安徽大学学术创新团队和安徽大学人才队伍建设经费
作者简介:高 亮(1983-),男,硕士研究生,研究方向为计算智能与信息处理;吴 涛,博士,副教授,主要从事机器学习、智能计算及其应用的研究。

中元素称为属性；$V=\bigcup_{a\in A}V_a$ 称为属性值集，V_a 是属性 a 的值域；$F:U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即

$$\forall a\in A, x\in U, F(x,a)\in V_a$$

若 $A=C\cup D, C\cap D=\varnothing$，$C$ 称为条件属性集，D 称为决策属性集，则该信息系统称为决策表。

定义 1：设 $a\in A, P\subset A$。关于属性子集 P 的二元关系 $\mathrm{ind}(p)$ 称作不可分辨关系，定义为：

$$\mathrm{ind}(P)=\{(x,y)\in U\times U\mid \forall a\in P, a(x)=a(y)\}$$

则称 x 和 y 是不可分辨的，易知对 $\forall P\subset A$，不可分辨关系 $\mathrm{ind}(p)$ 是 U 上的一个等价关系构成 U 的一个划分，记作 $U/\mathrm{ind}(p)$。

定义 2：设 $X\subset U$ 是论域的一个子集，R 是 U 上的一个等价关系，定义

$$\underline{R}X=\{x\in U\mid [x]_R\subseteq X\}$$

$$\overline{R}X=\{x\in U\mid [x]_R\cap X\neq\varnothing\}$$

分别称为 X 的 R 下近似集和 R 上近似集，其中 $[x]_R$ 表示 U 中在等价关系 R 下的等价类元素构成的集。$\mathrm{POS}_P(X)=\underline{R}X$ 称为 X 的 R 正域。

定义 3：设 U 是一个论域，P、Q 为 U 中的等价关系，Q 的 P 正域记为 $\mathrm{POS}_P(Q)$，即：

$$\mathrm{POS}_P(Q)=\bigcup_{x\in U/Q}\underline{P}X$$

Q 的 P 正域是 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中去的对象集。

2 基于相对差异比较表的属性约简算法

首先给出相对差异比较表的定义：

给定决策表 $S=(U,A=C\cup D,V,F)$，C 为条件属性集，D 为决策属性集。由 S 构造一个新的信息表 $S^*=(U^*,A^*,V^*,F^*)$，如下

(1) $U^*=\{(u_i,u_j)\in U\times U\mid \omega(u_i,u_j)\}$[7]

对于 $\forall u_i,u_j\in U,\omega(u_i,u_j)$ 满足

$u_i\in\mathrm{POS}_C(D)$ 且 $u_j\notin\mathrm{POS}_C(D)$；或 $u_i\notin\mathrm{POS}_C(D)$ 且 $u_j\in\mathrm{POS}_C(D)$；或 $u_i,u_j\in\mathrm{POS}_C(D)$ 且 $(u_i,u_j)\notin\mathrm{ind}(D)$。

(2) $A^*=C$。

(3) 构造区分函数：$\forall u_i,u_j\in U, c\in A^*$，则令

$$c(u_i,u_j)=\begin{cases}1 & c(u_i)\neq c(u_j)\\ 0 & c(u_i)=c(u_j)\end{cases}\qquad F^*:U^*\times A^*\to V^*$$

于是就得到一个新的信息表 S^*，就称为相对差异比较表。文献[8]中也给出了一种构造相对差异比较表的方法，但是其方法只对相容决策表有效，对于不相容决策表其做法是不可行的。并且文献[8]在构造相对差异比较表时，对决策表中任意两个不同类的数据都要进行比较，如果所给数据类别很多则将大大增加计算量，而文中给出的相对差异比较表对于不相容决策表也是有效的，而且在构造方法上也能减少一定的计算量。

2.1 基于 0-1 整数规划属性约简算法

(1) 由此相对差异比较表可以将上述新的信息表换成矩阵形式[9]，即 $S^*\to \boldsymbol{M}=(m_{ij})_{p\times q}$，$m_{ij}$ 表示第 i 个序对能否被第 j 个属性分辨，p 表示序对数目，q 为条件属性的个数。若能分辨，则 $m_{ij}=1$，否则 $m_{ij}=0$。于是最优约简的选择问题就转化为在 $\boldsymbol{M}$ 中找到由 1 组成的路径，该路径覆盖最少的列，即寻找最少的列使得 $\boldsymbol{M}$ 中的每一行至少含有一个 1。

(2) 现在可以把条件属性用向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_q)$ 表示，其中 $x_i(i=1,2,\cdots,q)$ 表示第 i 个条件属性。利用矩阵 $\boldsymbol{M}$，最优约简就转化为下面的 0-1 整数规划问题：

$$\min f(\boldsymbol{X})=f(x_1,x_2,\cdots,x_q)=\sum_{i=1}^{q}x_i$$

$$\text{s.t. } \boldsymbol{M}\cdot\boldsymbol{X}^{\mathrm{T}}\geqslant I \text{ 即 } \sum_{j=1}^{q}m_{ij}x_j\geqslant 1$$

式中，$i=1,2,\cdots,q$；$I=(\underbrace{1,1,\cdots,1}_{p\uparrow 1})^{\mathrm{T}}$；$x_i\in\{0,1\}$

(3) 求解此 0-1 整数规划问题，得到一个最优解，若 $x_i=0$，说明第 i 个条件属性可删除；若 $x_i=1$，则说明保留第 i 个条件属性。

2.2 改进算法

上述的 0-1 整数规划的方法虽然是可行的，但其复杂度为 $(2^p\cdot q)$，对分类数或对象数较多的决策表，运算的效率则较低，这是求解最优约简必须考虑的因素。再来研究相对差异比较表 S^*，对 S^* 中的各行进行纵向相加，其结果设为 P，P 中的每一个元素 p_i 对应着它所在列的列和，列和越大表示该列所对应的条件属性能够区分对象序对的数量也就越多，从而它所占的重要程度也越大。就可以用 p_i 表示第 i 个属性在决策表中的重要度，而当 $p_i=0$ 时，说明该属性不能区分任意一组序对，因而它是最不重要的，把这一属性和它所对应的列删除。

在进行约简过程中，对于重要性越大的属性，总希望它能保留下来，这对于信息的保持是很重要的。因此利用“贪婪算法”的思想，优先考虑属性重要性大的且相应 $x_i=1$ 的属性。这时，目标函数 $\min f(\boldsymbol{X})=f(x_1,x_2,\cdots,x_q)=\sum_{i=1}^{q}x_i$ 就可以写成：

$$\min f(\boldsymbol{X})=f(x_1,x_2,\cdots,x_q)=\sum_{i=1}^{q}\frac{1}{p_i}x_i$$

s.t. $\boldsymbol{M}\cdot\boldsymbol{X}^{\mathrm{T}}\geqslant I$ 即 $\sum_{j=1}^{q}m_{ij}x_j\geqslant 1$

式中,$i=1,2,\cdots,p$;$I=(\underbrace{1,1,\cdots,1}_{p\times 1})^{\mathrm{T}}$;$x_i\in\{0,1\}$;$p_i\neq 0$

对于此规划问题,为了避免当分类数或对象数增多时运算效率变低的缺点,采用遗传算法来解决此整数规划问题[10]。遗传算法可以克服传统优化方法的缺点,是一种多线索而非单线索的全局优化方法,采用的是种群和随机搜索机制。因此不会因为分类数或对象数的增多而降低效率。

(1) 编码方案:采用二进制编码方法,编码长度为所给的条件属性的个数减去 $p_i=0$ 所对应的属性个数,染色体中每个基因对应一个条件属性,如果一位为0,则表示不选择对应属性;如果为1,则表示选择该属性。

(2) 适应度函数的选择:对于上述的编码方案,任意一个编码串或交叉,变异所产生的任意个体它们不一定是可行解即不满足约束条件,但是对于这样的解不是抛弃这些不可行的解,而是采用"罚函数"的方法,加入惩罚项,则适应度函数可写为:

$$f(\boldsymbol{X})=\sum_{i=1}^{q}\frac{1}{p_i}x_i+r\max\{p_i\}$$

其中 $r=\left|\left\{\sum_{j=1}^{q}m_{ij}x_j=0\right\}\right|$,$r$ 表示所给解不能区分的对象序队的个数,f 的值越小说明所给解的适应度越大。

(3) 选择操作:采用适应度比例选择方法,从当前群体中选出优良的个体,将其复制到下一代群体中,该方法也成为轮盘赌选择。

(4) 交叉操作:采用两点交叉算子,对于群体中的个体进行两两随机配对,设定交叉概率为 p_c。

(5) 变异操作:采用基本位变异方式,设变异概率为 p_m。

(6)最优保存策略:在得到新一代个体之后,用上一代最好的个体代替新一代最差的个体,确保算法收敛。

3　实证分析

为考察算法的有效性,对下述所给的一个关于汽车数据的信息系统(见表1),其中论域 $U=\{1,2,\cdots,21\}$,条件属性集 $C=\{$类型,汽缸,涡轮式,燃料,排气量,压缩率,功率,换档,重量$\}$,决策属性 $D=\{$里程$\}$ 分别采用2.1节的算法和2.2节的改进算法计算。

根据2.1节的算法用 Matlab 计算 0－1 整数规划,得最优解 $\boldsymbol{X}^*=(100110001)$,即属性约简结果为{类型,燃料,排气量,重量}。

表1　汽车数据表

类型	气缸	涡轮式	燃料	排气量	压缩率	功率	换档	重量	里程
小型	4	Y	1型	中	高	高	自动	中	中
小型	4	N	1型	中	中	高	手动	中	中
小型	4	N	1型	中	高	高	手动	中	中
小型	4	N	1型	中	中	中	手动	中	中
小型	4	N	2型	中	中	中	自动	重	低
小型	4	N	1型	中	中	高	手动	重	低
微型	4	N	2型	小	高	低	手动	轻	高
小型	4	N	2型	小	高	低	手动	中	中
小型	4	N	2型	小	高	中	自动	中	中
微型	4	N	1型	小	高	低	手动	轻	高
微型	4	N	1型	小	中	中	手动	中	高
小型	4	N	2型	中	中	中	手动	中	中
微型	4	Y	1型	小	高	高	手动	中	高
微型	4	N	2型	小	中	低	手动	中	高
小型	4	Y	1型	中	中	高	手动	中	中
小型	4	N	1型	中	中	高	自动	中	中
小型	4	N	1型	中	中	高	自动	中	中
微型	4	N	1型	小	高	中	手动	中	高
小型	4	N	1型	小	高	中	手动	中	高
小型	4	N	2型	小	高	中	手动	中	中

根据改进算法:令初始种群 $m=20$,$p_c=0.8$,$p_m=0.03$,终止条件:最优个体连续十代保持不变,求得最终结果为 $\boldsymbol{X}^*=(100110001)$。可以看出当分类数或对象数较少时,两种算法在运算效率上相差不是很大,且运算结果也是一致的。

为了进一步比较两种算法的约简效果,从 UCI 机器学习数据库中选择了 Zoo Database,论域 $U=\{1,2,\cdots,194\}$,条件属性集为 $C=\{$hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, cat size$\}$;决策属性集 $D=\{$type$\}$。

采用2.1节的算法计算 0－1 整数规划得最优解

$\boldsymbol{X}^*=(0010010101001000)$

即属性约简结果为{eggs,aquatic,toothed,breathes,legs}。

改进算法:

令初始种群 $m=50$,$p_c=0.8$,$p_m=0.03$。

终止条件:最优个体连续十代保持不变。

用改进算法求得最终结果为

$\boldsymbol{X}^*=(0011010100001000)$

即属性约简结果为{eggs, milk, aquatic, toothed, legs}。

这时就可以看出当分类数或对象数增多时,改进算法不仅在计算效率上明显优于2.1节的算法,而且所得约简结果也好于它。

最后,采用基于差别矩阵的属性约简策略,对上面

(下转第24页)

较软阈值函数好,但图像边缘有少许的振铃现象。改进阈值函数($\beta=0.76$ 在实验中比较挑选出来)处理的结果较前两种阈值处理函数视觉效果好些,对两种阈值函数的优点都有所保留而对缺点都有所抑制。

文中为对实验结果进行科学的评估,采用峰值信噪比(PSNR)的概念进行评价,对去噪后的图像来说,其 PSNR(见式(5))的计算结果越大,说明图像去噪的效果越好。

$$PSNR = 10\log\left[\frac{255^2}{\frac{\sum_{i=0}^{M-1}\sum_{j=1}^{N-1}(U(i,j)-U_1(i,j))^2}{M*N}}\right] \quad (5)$$

上式中 U 为原图,U_1 为去噪后的图像 ,M 和 N 为图像的行列值。

经过试验仿真表明改进后阈值函数比改进前的软硬阈值两种函数有更好的视觉效果,PSNR 值增大,表 1 为图 3 的结果比较值。

表 1 算法改进前后的 PSNR 对比

阈值函数	PSNR 值
硬阈值	27.84
软阈值	30.71
改进的阈值 $\beta=0.76$	31.65

在实验中对于不同的噪声干扰下,参数 β 的取值要根据实验结果去定,选取最优的 β 值,但经过大量的实验表明这个 β 值的最优取值一般在 0.7 至 0.9 之间。

4 结束语

文中主要研究小波阈值在图像去噪中的应用,试验发现小波阈值的选取和阈值函数的选择直接影响着最后的结果,文中的方法经过仿真结果表明去噪后的图像比原来的算法有所改善,在保留传统方法优点的同时有效地抑制传统方法的缺点,具有一定的使用价值。

参考文献:

[1] 谢杰成,张大力,徐立文. 小波图像去噪综述[J]. 中国图像图形学报,2002,7(3A):209-217.

[2] Mallat S. A theory for multiresolution signal decomposition: The wavelet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1989,11(7):674-693.

[3] Donoho D L, Johnstone I M. Ideal Spatial Adaptation Via Wavelet Shrinkage[J]. Biometrika,1994,81(12):425-455.

[4] Donoho D L. Denoising by Soft-thresholding[J]. IEEE Trans on IT,1995,41(3):613-627.

[5] 张维强,宋国乡. 基于一种新的阈值函数的小波阈值信号去噪[J]. 西安电子科技大学学报,2004,31(2):296-303.

(上接第 21 页)

两个例子进行属性约简,其结果是该策略的消耗时间与 2.1 节的算法的消耗时间基本相等。这是因为相对差异比较表实际上就是差别矩阵的展开,因此从这方面看,2.1 节的算法的时间复杂性与基于差别矩阵属性约简策略的时间复杂性相等。不过 2.1 节的算法将一些比较逻辑运算转化为矩阵运算,从而在一定程度上简化了计算,提高了运算速度;而改进算法是在 2.1 节的算法的基础上加入遗传算法,具有很高的并行性,因此在计算时间上大大地减少了。

4 结束语

在相对差异比较表的基础上将属性约简问题转化为求解整数规划问题,其过程简单易懂,效果也不错,同时为了克服求解规划问题存在的缺点,引入贪婪思想,并且结合遗传算法求解规划问题,使得算法具有更高的效率,实验结果也证明此方法是有效可行的。

参考文献:

[1] Pawlak Z. Rough sets and decision tables[J]. Lecture Notes in Computer Science,1985,208:187-196.

[2] Pawlak Z. Rough set theory and its applications to data analysis[J]. Cybernetics and Systems: An International Journal, 1998,29:661-688.

[3] 代建华,李元香. 一种基于粗糙集的决策系统属性约简算法[J]. 小型微型计算机系统,2003,3(3):523-526.

[4] 刘 清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.

[5] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001.

[6] 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.

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

[8] 潘 丹,郑启伦. 属性约简自寻优算法[J]. 计算机研究与发展,2001,38(8):904-910.

[9] 赵卫东,戴伟辉,蔡 斌. 遗传算法在决策表连续属性离散化中的应用研究[J]. 系统工程理论与实践,2003,1(1):62-67.

[10] 玄光男,程润伟. 遗传算法与工程优化[M]. 北京:清华大学出版社,1999.