基于粗糙集理论与灰色理论的属性约简算法

杜 晓[1],刘维亭[1],杜 茜[2],罗军生[1]
(1.江苏科技大学 电子信息学院,江苏 镇江 212003;
2.华中师范大学 电子信息工程系,湖北 武汉 430079)

摘 要:约简是粗集理论的重要概念,由定义计算约简是一个典型的NP问题且由于约简的不唯一,在面对大数据集或高维数据集问题时获得的属性集往往并非是最小的属性约简集。文中针对Rough sets理论的属性约简进行了研究。研究了通过可辨识矩阵求得属性约简集,利用Rough sets与灰色理论相结合,提出一种属性约简的启发式算法,拟合结果表明本约简算法合有效。

关键词:Rough sets理论;属性约简;分辨矩阵;灰色关联
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)01-0154-03

Algorithm for Attributes Reduction Based on Rough Set Theory and Gray Theory

DU Xiao[1], LIU Wei-ting[1], DU Qian[2], LUO Jun-sheng[1]
(1. Dept. of Electronic and Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China;
2. Dept. of Electronic and Information Engineering, Huazhong Normal University, Wuhan 430079, China)

Abstract: Reduction is an important concept in rough set theory, while computing reduction according to the definitions directly is a typical NP problem. The attribute set get from the problem with large and high-dimension database is not usually the minimum attribute set. Discusses the approaches for attribute reduction based on rough set theory. Following, studies the approaches to achieve attribute reduction set by applying recognized matrix. Researched how to get attribute reductions through discernibility matrix, combined rough sets and gray theory, and put forward a new heuristics algorithm for attribute reduction. The effectiveness of the result obtained is demonstrated by an example.

Key words: rough sets theory; attribute reductions; discernibility matrix; gray relation

0 引 言

粗糙集(Rough set)理论是一种处理模糊和不确定知识的数学工具,最早由波兰数学家Z. Pawlak于1982年提出[1]。它已经在数据挖掘、人工智能、模式识别与分类等领域获得了较广泛的应用。属性约简是Rought set理论研究的一个核心内容。人们希望找到最佳属性约简。然而已经证明它是NP-hard问题[2],因此对属性约简的获得,只能采用启发式规则。目前已提出了属性约简算法。例如:利用某一属性在决策矩阵中出现的概率作为启发式信息构造启发式算法,或利用属性依赖度的属性重要度等信息作为启发式知识进行属性约简等算法[3~10]。

文中考虑基于属性频率的约简规则。由于属性有多个不同的约简,导致一般能找到真正的约简而不是约简的超集。且差别矩阵中经常会出现属性频率函数值相同的情况,当频率函数值相同时属性的选取规则成为问题。所以如何选择启发式算法获得最有效、不影响分类精度,且最小的子集成为需要考虑的问题。结合灰色理论原理,提出的基于粗糙集理论与灰色理论的方法,可以有效去除贡献相同的属性保留分类的最佳属性,而且具有在时间上较快,并保证得到的是约简而不是约简的超集。

收稿日期:2007-03-27
基金项目:江苏省教育资助项目(2005DX006J)
作者简介:杜 晓(1979-),男,山东临沂人,硕士研究生,从事智能信息处理方面的研究;刘维亭,教授,博士,从事自动化、电气工程方面的研究。

1 基本概念

定义1 粗糙集理论中的一个信息系统可表示为:$S=(U,A,V,f)$,其中U是一个有限对象集;A是有限属性集,分为条件属性集C和决策属性集D,即

$A = C \cup \boldsymbol{D}, C \cap \boldsymbol{D} = \varnothing$;$V$ 是属性值组成的集合;f 是一个信息函数,它指定 U 中每一个对象的属性值。在信息系统 S 中,对于 $P \subseteq C$,则 P 在 S 的不分明关系 IND(P) 定义为:IND(P) = $\{(x,y) \in U^2 \mid \forall a \in P, a(x) = a(y)\}$。IND($P$) 把对象集 U 划分为 k 个等价类,记为:$U/P = \{x_1, x_2, \cdots, x_k\}$。

定义2　在信息系统 S 中,属性 $a \in B \subseteq C$ 是 B 中必要的,当且仅当 IND(B) ≠ IND($B - \{a\}$),否则,属性 a 在 B 中是冗余或可省略的。属性集 B 的约简 red(B) 是一个集合 $B' \subseteq B$,当且仅当满足:

(1)B' 是独立的;(2)IND(B') = IND(B)。属性集 $B \subseteq C$ 的所有约简族的交集称为属性集 B 的核,记为 core(B),有:core(B) = ∩ red(B)。

定义3　设信息系统 $S = (U, A, V, f)$,$R \in A$,R 是一组等价关系,$r \in R$,如果:IND(R) = IND($R - \{a\}$) 则称 r 是 R 中不必要的,否则称 r 为 R 中必要的;如果每一个 $r \in R$ 都为 R 中必要的,则称 R 为独立的,否则称 R 为信赖的。设 $Q \in R$,如果 Q 是独立的,且 IND(Q) = IND(P),则称 Q 为论域 U 在属性集 P 上的约简。

定义4[11]设有 n 个观测对象,每个对象观测 m 个特征数据,得到序列如下:$X_1, X_2, \cdots, X_n$。其中 $X_i = x_i(1), x_i(2), \cdots, x_i(n)$。对所有的 $i \leq j, i, j = 1, 2, \cdots, m$,计算出 x_i 与 x_j 的灰色绝对关联度 ε_{ij}。其在临界值 r 下的分类称为特征变量的灰色关联聚类。

定义5　分辨矩阵 $\boldsymbol{D}$ 的概念[12]:给定一个信息系统 $S = (U, A, V, f)$,$A = C \cup \boldsymbol{D}$ 是属性集合,$C, \boldsymbol{D}$ 是条件属性和决策属性。分辨矩阵 $\boldsymbol{D} = (d_{ij})$ 定义为:

$$d_{ij} = \begin{cases} \{a \in C : a(x_i) \neq a(x_j)\}, & \text{当 } \boldsymbol{D}(x_i) \neq \boldsymbol{D}(x_j) \\ \varnothing, & \text{当 } \boldsymbol{D}(x_i) = \boldsymbol{D}(x_j) \end{cases}$$

其中 $a(x)$ 是元组在属性 C 上的取值,$\boldsymbol{D}(x)$ 为在决策属性的取值。

2　启发式算法研究

目前研究粗糙集属性约简的启发式算法较多,利用差别矩阵的属性频率信息是一种重要的研究思想,胡可云算法虽然较其他一些算法在时间上较快,且一般能找到真正的约简而不是约简的超集,但算法并不一定能保证找到约简,马光志对 MIBARK 算法进行改进,改进算法从差别矩阵得到属性的核和频率。某些属性频率很大且同时与核属性的关联性很强,在差别矩阵中经常伴随核属性存在于矩阵项中,独立于核属性出现在差别矩阵中的频率很低。这些属性区别对象的能力可以由核属性代替。因此,即使它们的频率很大,在约简里也是冗余的。所以可从差别矩阵 $\boldsymbol{D}(S)$ 提取过滤差别矩阵,以参考属性集 R(一般为属性的核)为基础。凡是 $\boldsymbol{D}(S)$ 中包含 $a \in R$ 的矩阵项都置为空。过滤差别矩阵反映了非核属性区别对象的能力,可以尽可能地减少约简中的冗余属性数目。以上算法,在计算频率时,很难避免有频率相同的情况,相同频率的属性也存在贡献冗余,并且没有考虑分类属性中关联性大的属性。如果不对关联性大的属性进行适当限制,得到的结果可能会是约简的超集。灰色理论中的灰色关联聚类,可以找出可能是相关或是混同的指标,希望通过对少数对象的观测结果,删去不必要的指标简化考虑标准。综合考虑,可以采取以属性作为指标,样本作为观测对象,逐个求出属性之间的关联程度。

可以采取两种策略:

* 算法一:把临界值设为1,进一步对过滤后的差别矩阵进行过滤,确定属于一类的属性从中选择一个,并从差别矩阵中去除其余属性,然后计算属性频率 $f(a) = f(a) + |A| / |C|$,对于每个 $a \in C$,其中 $|A|$ 是信息系统总的条件属性个数。

* 算法二:计算两个值:① 计算过滤差别矩阵的属性频率 $f(a)$ ($f(a) = \lambda_{ij}/\boldsymbol{D}_{ij}$,其中 $\boldsymbol{D}_{ij}$ 表示矩阵元素中包含属性个数,对于每个 $a \in \boldsymbol{D}_{ij}$,$\lambda_{ij} = 1$,否则,$\lambda_{ij} = 0$);② 计算灰色关联度值,算法以 R 表示每次的约简。设属性重要度为 $M(a)$,

$$M(a) = f(a) - \sum \varepsilon_{ij} \tag{1}$$

其中 i 为 R 中的属性,j 为剩余属性,即 ε_{ij} 为当前约简属性与剩余属性的灰色绝对关联度值。ε_{ij} 值的范围为 $[0,1]$。

第一种方法去掉部分多余属性并对差别矩阵进行过滤,保证最终约简属性非约简超集。第二种方法对属性重要度函数进行改进,算法能对相关或混同属性进行去除。

3　约简算法描述

3.1　算法描述

粗糙集属性约简是在不损失信息的前提下删除冗余的属性,属性约简集的集合 R 可以表示为:$R = \{R : R \in C, \mathrm{POS}_R(\boldsymbol{D}) = \mathrm{POS}_C(\boldsymbol{D})\}$,所以算法的结束条件为 $\mathrm{POS}_R(\boldsymbol{D}) = \mathrm{POS}_C(\boldsymbol{D})$,即保证分类质量最小属性集。

输入:决策表 $T = (U, C \cup \boldsymbol{D}, V, f)$,其中,$C$ 为条件属性,$\boldsymbol{D}$ 为决策属性。

输出:决策表的一个最小属性约简集 R。

利用差别矩阵计算条件属性 C 相对于决策属性 $\boldsymbol{D}$ 的核 R = core(C)。$E = C - R$。求出过滤差别矩阵。

算法一:

(1) 计算差别矩阵以属性为指标,部分样为观测对象,设临界值为 1 求灰色关联聚类;

(2) 从聚为一类的属性中保留一个,并在差别矩阵中删除其余的属性。计算属性频率:

$f(a) = f(a) + |A| / |C|$

(3) 从 E 中选择最大 $f(a)$ 的属性 a 加入到 R 中;

(4) 令 $R = R + \{a\}, E = E - \{a\}$, 计算 $POS_R(D)$ 是否等于 $POS_C(D)$, 若是则结束, 否则, 转(3);

(5) 输出 R, R 即为属性约简。

算法二:

(1) 计算属性出现的频率: $f(a) = \lambda_{ij}/D_{ij}$, 当 $a \in D_{ij}, i, j = 1,2,\cdots,n$。$n$ 为样本个数;

(2) 对每个属性 $a \in E$, 根据式(1) 计算其属性重要度 $M(a)$;

(3) 选择 $M(a)$ 值最大的属性, 加入至 R 中;

(4) $R = R + \{a\}, E = E - \{a\}$, 计算 $POS_R(D)$ 是否等于 $POS_C(D)$, 若是则结束, 否则, 转(2);

(5) 输出 R, R 即为属性约简。

计算属性的灰色关联度, 及进行聚类仅对部分样本进行计算, 故算法时间复杂度和基于属性频率的算法复杂度相同, 故在时间上一样快。

3.2 算法实验分析

在职位的任职资格指标体系中, 如何精选那些代表性好、灵敏度高、特异性强、易于测评的指标是试图通过基于粗糙集数据挖掘要解决的问题。以某单位任职资格评审原有的指标为例进行数据集的属性约简。该测评共有 15 个指标(属性): 申请书印象、学术能力、讨人喜欢程度、自信程度、精明、诚实、推销能力、经验、积极性、报负、外貌、理解能力、潜力、交际能力、适应能力。而每个观测样本(具体的人)在指标(属性)中品质, 可由申请书印象、学术能力、诚实等描述, 所以指标(属性)具有一定的重复性。样本数为 1000。灰色关联算法显示学术能力与经验, 外貌与理解能力关联度很大。运用对数据进行属性约简。算法二实验结果较算法一及基于频率函数的方法得到属性集个数少。算法一仅在时间上较频率函数方法小。

4 结 语

粗糙集理论的主要思想是在保持分类能力不变的前提下, 对属性进行约简, 进而导出分类规则, 因此, 属性约简是粗糙集理论的核心内容。由于寻找决策表属性的最小约简是 NP - hard 问题, 因此应采用启发式算法进行求解。而启发式算法要着重考虑属性间的关联度才能保证在大数据集及高维属性下的有效性。文中的启发式信息, 由于考虑了属性间关联性, 算法在一定程度上还可以减少相对约简中的冗余属性, 且具有较好的时间性能。

参考文献:

[1] Pawlak Z. Roughset: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] Perkins C E, Royer E M. Ad Hoc on Demand Distance Vector (AODV) Routing[S]. IETF MANET WG Internet draft, 2000.

[3] 王 珏, 王 任, 苗夺谦, 等. 基于 Rough Set 理论的"数据浓缩"[J]. 计算机学报, 1998, 21(5): 393 - 399.

[4] 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.

[5] 石 峰, 娄臻亮, 张永清. 一种改进的粗糙集属性约简启发式算法[J]. 上海交通大学学报, 2002, 36(4): 478 - 481.

[6] 张冬玲. 基于粗糙集理论的属性约简算法的实现[J]. 计算机应用, 2006, 26: 78 - 82.

[7] 马光志, 吴黎明. 基于粗糙集理论的一种属性约简算法[J]. 计算机工程与应用, 2006(18): 171 - 175.

[8] 胡可云. 基于概念格和粗糙集的数据挖掘方法研究[D]. 北京: 清华大学, 2001.

[9] 卢佳华. 基于属性频率函数的粗糙集属性约简算法[J]. 武汉大学学报: 理学版, 2006(6): 331 - 334.

[10] 白秀玲, 王平普, 杰 信. 一种粗糙集值约简算法及其应用[J]. 微计算机信息, 2006(11)7: 207 - 209.

[11] 刘思峰, 党耀国, 方志耕. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2001.

[12] 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, 2005.

(上接第 153 页)

Practice[J]. Knowledge Engineering Review, 1995, 10(2): 115 - 152.

[4] 赵龙文, 侯义斌. 智能软件: 由面向对象到面向 Agent[J]. 计算机工程与应用, 2001(5): 41 - 43.

[5] 张云勇, 刘锦德. 移动 Agent 技术[M]. 北京: 清华大学出版社, 2003.

[6] 郑人杰, 殷人昆, 陶永雷. 实用软件工程[M]. 北京: 清华大学出版社, 1997.

[7] 侯惠芳, 彭成寒. 面向对象软件度量工具的设计实现[J]. 计算机工程与设计, 2005, 26(6): 1447 - 1449.

[8] Harrison R, Nithi R. An evaluation of the MOOD set of object - oriented software metrics[J]. IEEE Transactions on Software Engineering, 1998, 24(6): 491 - 496.

[9] 程显毅, 石纯一. Agent 社会理性的研究[J]. 软件学报, 2001, 12(12): 1825 - 1829.