基于属性核的遗传约简算法*)

郭 平 刘潭仁 刘 然 贺 琼
(重庆大学计算机学院 重庆400044)

摘 要 属性最小约简是NP完全问题,该问题的研究一直被关注。如,以不可分辨矩阵为基础的传统约简方法[1],基于属性重要性的约简方法[1]等等,这些方法对于大数据集都是不实用的。文[8]提出了以遗传算法全局搜寻能力为基础的属性约简方法,文[3]通过引进属性依赖启发信息改进了文[8]中的方法。本文中,先给出了一个时间复杂度为O($k\times n\times\log n$),空间复杂度为O(n)的核属性判别方法。然后,以此为基础给出了较文[3]和文[8]中更有效的遗传粗糙约简算法。

关键词 粗糙集,属性约简,遗传算法,核属性

An Genetic-Enhanced Core Attributes Based Reduct Method

GUO Ping LIU Tan-Ren LIU Ran HE Qiang
(College of Computer Science, Chongqing University, Chongqing 400044)

Abstract The problem of finding minimal reduct belongs to the class of NP-complete Problems. Some articles solved the problem in different ways. The traditional reduct methods such as the method based on discernability matrix[1], the induction method based on the importance of the attributes[1] and so on, are all impractical for large database. Paper [8] brings out a minimal reduct methods which take advantage of the global search ability of Genetic Algorithm. Paper [3] improves the methos in Paper [8] mentioned by inducing the heuristic information: the dependency of Attributes. In this paper, firstly we improve the method of finding core attributes, and only O($k\times n\times\log n$) time complexity and O(n) memory space required in our method. Then we combine the method with methods advised in [3,8], and bring out a genetic-enhanced Attribute reduct Method which shows good qualities in some aspects.

Keywords Rough set, Attribute reduct, Genetic algorithm, Core atrributes

1 引言

粗糙集理论是一种研究不完整与不确定知识和数据的表达、学习和归纳的理论方法。其研究对象是多值属性集合描述的对象集合。这种对象集合可以抽象为二维表格,其中行表示不同的对象,列表示对象的属性。在这样的表格中,当将列划分为条件属性和决策属性时称为决策表。由于粗糙集在处理不完整与不确定性方面具有很好性能,它已在人工智能、机器学习和模式识别等领域广泛地使用。在数据挖掘领域,粗糙集已成功地应用到包括数据处理[5]、属性约简[5]和规则提取[5]等各个阶段以及分类规则、聚类规则和关联规则提取等方面。随着空间数据库的发展,粗集理论在空间数据挖掘方面也得到了广泛的应用[9,10]。

属性的约简一直以来是粗糙集理论在数据挖掘中的研究重点。已经证明,寻找决策属性的最小约简是一个NP完全问题。文[3]提出了基于属性依赖度的遗传约简算法,本文进一步研究了属性核的性质,提出了一种较为方便的属性核判断方法,并以此为基础提出了一个利用核属性为启发信息的属性约简算法。

2 决策表与属性约简

这里先给出与本文相关的一些概念。

定义1 一个决策表是一个信息表知识表达系统 $S=\langle U,C,D,V,F\rangle$,$R=C\cup D$ 称为属性集,C 和 D 分别称为条件属性集和决策属性集,$D\neq\varnothing$。特别,如果 $POS_C(D)=U$,则称此知识表达系统是一致的。

定义2 设一致的知识表达系统 S 对决策属性 d_x,有 $[x]_C\subseteq[x]_D$。若$\forall\ r\in C$,有$[x]_{C\setminus\{r\}}\subseteq[x]_D$ 不成立,称 r 为 d_x 的核值属性,r 为 d_x 中不可省略的;若$[x]_{C\setminus\{r\}}\subseteq[x]_D$ 成立,称 r 不是 d_x 的核值属性,r 为 d_x 可省略的。

定义3 设 U 是论域,P 是定义在 U 上的等价关系簇,$R\in P$。如果 $IND(P\setminus R)=IND(P)$,称 R 在 P 中是绝对不必要的;否则称 R 在 P 中是绝对必要的;P 中所有绝对必要的关系组成的集合称为 P 的核。

本文要引用的主要结论有:

定理1[2] 设 U 是一个论域,P 和 Q 是定义在 U 上的两个等价关系簇,$r\subset P$。若 $POS_P(Q)=POS_{P\setminus\{r\}}(Q)$,则 r 为 P 中相对于 Q 是可省略的(不必要的);否则 r 为 P 中相对于 Q 是不可省略的(必要的)。

定理2[2] 对一致的知识表达系统 $S=\langle U,C,D,V,F\rangle$,则以属性 a 为核值属性的决策规则的集合为

$$core(a)=\{d_x\mid x\in(U\setminus POS_{C\setminus\{a\}}(D))\}$$

定理3[7] 在决策表 $S=\langle U,C,D,V,F\rangle$ 中,$C\cup D$ 的任何子集 B 的不可分辨关系 $IND(B)$ 可以在时间复杂度 O($n\log n$)和空间复杂度 O(n)下求得。n 是决策表中记录的个数。

定理4[7] 在决策表 $S=\langle U,C,D,V,F\rangle$ 中,$POS_C(D)$ 可

*)本文的研究得到国家十五攻关项目(编号:2002BA107B)的资助。郭 平 副教授,主研方向:AI、GIS;刘潭仁 硕士研究生,主研方向:粗集理论与应用,GIS。

以在时间复杂度 $O(k\times n\log n)$ 和空间复杂度 $O(n)$ 下求得。n 是决策表中记录的个数，k 为 C 中条件属性的个数。

3 核属性判别方法

在属性约简中，核属性是不能被约简的属性。因此，判别核属性是属性约简中至关重要的。

定理5 在一致的知识表达系统 $S=\langle U,C,D,V,F\rangle$ 中，核属性集合为

$$core(C)=\{a\mid a\in C\wedge core(a)\neq\varnothing\}$$

证明：如果 $core(a)=\varnothing$，则 $POS_{C-\{a\}}(D)=U$，即 $C\backslash\{a\}$ 为 C 的一个约简，所以 a 必然不是核属性。

如果 $core(a)\neq\varnothing$，则有 $POS_{C\backslash\{a\}}(D)\neq U$，又由于表达系统 $S=\langle U,C,D,V,F\rangle$ 是一致的，因此有 $POS_C(D)=U$ 成立。所以 a 为核属性。

综上，定理得证。

定理6 在一致的知识表达系统 $S=\langle U,C,D,V,F\rangle$ 中，$a\in C$，条件子类 $X_1,X_2,\cdots,X_k\in IND(C/a)$，决策类 $Y_1,Y_2,\cdots,Y_m\in IND(D)$，其中 k 为 $IND(C/a)$ 中条件子类个数，m 为决策类个数。如果存在 i,j,n 使得

$X_i\cap Y_j\neq\varnothing\wedge X_i\cap Y_n\neq\varnothing$　$1\leqslant i\leqslant k,1\leqslant j,n\leqslant m$，且 $j\neq n$

则 a 必为核属性。

证明：因为 $X_i\cap Y_j\neq\varnothing\wedge X_i\cap Y_n\neq\varnothing,1\leqslant i\leqslant k,1\leqslant j,n\leqslant m$，且 $j\neq n$，则必有 $POS_{C\backslash\{a\}}(D)\neq U$，所以有 $core(a)=\{d_x\mid x\in(U\backslash POS_{C\backslash\{a\}}(D))\}$ 为非空。根据定理5的证明，a 必为核属性。

这个定理给出了在一致的知识表达系统中判定核属性的方法，对于不一致的知识表达系统采用以下的判别方式：

定理7 在知识表达系统 $S=\langle U,C,D,V,F\rangle$ 中，$U_C=POS_C(D)$。则对于新的知识表达系统 $S'=(U_C,C,D,V,F)$，任意的 $a\in C$，条件子类 $X_1,X_2,\cdots,X_k\in IND(C/a)$，决策类 $Y_1,Y_2,\cdots,Y_m\in IND(D)$，其中 k 为 $IND(C/a)$ 中条件子类个数，m 为决策类个数，如果存在 i,j,n 使得

$X_i\cap Y_j\neq\varnothing\wedge X_i\cap Y_n\neq\varnothing$　$1\leqslant i\leqslant k,1\leqslant j,n\leqslant m$，且 $j\neq n$

则 a 必为 S 的核属性。

证明：因为 $U_C=POS_C(D)$ 是下近似集合，所以知识表达系统 $S'=\langle U_C,C,D,V,F\rangle$ 一定是一致的。由定理6知 a 为 S' 的核属性。

又由于 $U_C\subset U$，故 a 为 S 的核属性，定理得证。

定理6和定理7给出了快速判断一个属性是否为核属性的启发式方法，根据这两个定理和文[7]介绍的求 $POS_C(D)$ 方法，我们可以在时间复杂度 $O(k\times n\times\log n)$ 和空间复杂度 $O(n)$ 下求得核集合。其主要原因在于只要存在任何的一条记录满足定理6或者定理7中的条件，我们就可以停止有关该属性的是否为核属性的搜索工作，从而提高了效率。

4 遗传约简算法

这里给出的约简算法是根据定理6和定理7得到的。

4.1 算法框架

基于属性核的遗传约简算法

```
输入：知识表达系统 S=⟨U,C,D,V,F⟩
输出：约简后的知识表达系统
    第1步：求核属性
        对每一个属性，进行如下处理
        (1) 获取相关的样本数据；
        (2) 对样本数据进行快速排序；
        (3) 判断是否是核属性；
    第2步：随机生成遗传初始种群；
    第3步：遗传计算
        对于每一代种群进行如下处理：
        (1) 适应度评估｛
              取样本数据；
              对样本数据进行排序；
              计算个体对应的下近似和适应度；
            ｝
        (2) 选择运算；
        (3) 交叉运算；
        (4) 变异运算；
        (5) 优化运算——以至今最好的个体取代当代最差的个体；
    第4步：结果输出；
```

如果要求各个属性集的不可分辨关系，同样可以在快速排序的基础上，对样本数据进行一次扫描获得，其时间复杂度同样为 $O(n*\log n)+O(n)$。

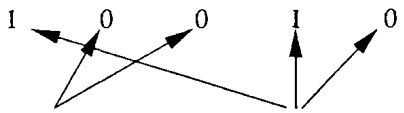
4.2 遗传编码与操作的限制

由于核属性是不能被约简的，因此在编码与操作中须满足以下限制：

(1)核属性对应的编码位必须始终为0，同时不参与变异操作；

(2)对于初始种群，由变异操作引进核属性集合，须进行特殊处理。

例如，如果属性个数为5，其中2，3为核属性，则编码规则解释如图1所示。



核属性（必须为0）　非核属性（可为0或1）

图1 遗传编码示意图

4.3 遗传算法适应性函数的选择

适应性函数在文[3]的基础上做如下调整

$$((m-l_r)/m)*w1+\text{Dependency}*w2$$

其中 m 是所有属性的个数，l_r 是该染色体实际所取的属性个数。$w1,w2$ 为调正整系数，且 $w1\geqslant 0,w2\geqslant 0,w1+w2=1$。当 $w1=0,w2=1$ 时为文[3]中的适应性函数。

4.4 遗传终止条件

遗传终止条件为达到设定的遗传代数或者适应性函数的值达到所要求的水平。

4.5 算法适应范围分析

设知识表达系统总共的条件属性个数为 m，则在不考虑任何启发性信息的条件下总的搜索空间为 2^m，如果核属性个数为 k，则采用基于核属性的算法的搜索空间最大为 2^{m-k}，可以看到引入核属性集后算法的搜索空间迅速下降。由于判定一个属性是否非核属性的时间复杂度为 $O(n\times\log n)$，因此搜索出所有的核属性的时间复杂度为 $O(m\times n\times\log n)$。而如果对搜索空间的每一种可能进行依赖性计算，其时间复杂度为 $O(2^m\times n\times\log n)$，增加核属性启发后为 $O(m\times n\times\log n)+O(2^{m-k}\times n\times\log n)$。当然这是在没有其他任何启发性信息的条件下的理论分析，在算法中引进依赖性(重要性)等启发性信息后，搜索空间将降为更小。可以预见，在 m 小，同时 n 也小的时候，算法效率的提高不是很多，该算法最适合的情况是 m 大，n 也大的情况。

4.6 时间复杂度和空间复杂度分析

设知识表达系统的记录数为 n，属性个数为 m，则寻找属性核的时间复杂度为 $O(m\times n\times\log n)$，空间复杂度为 $O(n)$。由文[1]，可以知道计算属性的依赖度的时间和空间复杂度均

为$O(n\times\log n)$。

5 算法的实验结果及分析

在测试平台 Windows 2000 Professional，CPU 为赛扬 III，内存256M 下，我们实现了文[3]中的算法和本文提出的算法。在算法中遗传算法的种群数为30，遗传代数为50，交叉概率为0.90，变异概率为0.05，终止条件为适应度达到0.980，调整系数 $w1=0.05$，$w2=0.95$。仿真结果列于表1中。

表1 仿真实验结果

序号	样本数目	属性个数	核属性个数	文[3]中算法		本文中算法	
				代数	时间(秒)	代数	时间(秒)
1	100	10	4	15	1.192	7	0.2
2	100	20	6	25	3.356	12	1.342
3	1000	20	8	32	36.465	18	24.802
4	4000	22	8	36	164.607	24	108.548
5	10000	26	10	44	422.095	29	311.206

由于本算法是基于核属性的，因此在数据样本的选择上，我们选择了核属性个数较多的数据进行测试，实验表明，本算法在属性较多且具有核属性时较文[3]中的算法有较明显的优势。

结束语 本文所提出的方法利用了决策系统本身的启发信息，通过采用系统核属性的改进算法，使得时间复杂度降为$O(m\times n\times\log n)$。本文给出的算法的特点是充分利用属性核信息，提高算法的效率，采用简洁的编码方式满足算法思想，利用遗传提高全局搜索的能力，避免局部最优。合适的适应函数的选择也是此方法的一大特点。

实验表明，在样本具有核属性时是一个较文[3]中的算法更有效的算法。

参考文献

1 Khoo L-P, Zhai L-Y. A Prototype Genetic Algorithm-enhanced Rought Set-based Rule Induction System. Computers in Industry, 2000,46:95～106

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

3 熊晖，肖人岳. 用遗传算法求解粗糙集约简的改进算法. 计算机科学，2002,29(9)

4 Http://www.idi.ntnu.no/～aleks/rosetta/help/manual.pdf

5 Skowron A. Rough Sets in KDD, Presented during PAKDD 2000, Kyoto, Japan Pacific-Asia Conference on Knowledge Discovery and Data mining

6 Hu X. Knowledge discovery in databases: An attributes-oriented Rough Sets approach: [Ph. D. Thesis]. University of Regina, 1995

7 Nguyen S H, Nguyen H S. Some Efficient Algorithms for Rough Set Methods. In: Proc. of the Conf. of Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'96, Granada, Spain, 1996. 1451～1456

8 尹旭日，陈世福. 一种基于 Rough 集缺省规则挖掘算法. 计算机研究发展，2000,37(12):1441～1445

9 石云，孙玉芳，左春. 基于 Rough Set 的空间数据分类方法. 软件学报，2000,11(5):673～678

10 Ester M, Kriegel H-P, Sander J. Knowledge Discovery in Spatial Databases. Lecture notes in computer science, 1999,1701:61～74

(上接第165页)

层次(可能不是共享的抽象层次)。Agent2将自身的理解发送给 Agent1，Agent1根据 Agent2的理解情况以及自身的本体对 Agent2的理解予以确认或者是加以更多的解释。如果加以了更多的解释，Agent2将重复以上的理解过程直到收到 Agent1的确认消息为止。收到/发送确认消息以后，Agent1和 Agent2将分别地根据最终理解结果以及自身的本体知识进行相关性分析，以完成各自的本体进化过程。

总结 多 Agent 系统之间的本体异构对知识的共享与集成以及 Agent 之间的协作都是一个非常大的障碍。为了解决这个问题，我们提出了一种结合元本体理论以及本体协商的方法。该方法以本体协商为框架，基于元本体进行协商。该方法通过引入元本体使得 Agent 之间就本体知识进行协商的效率大大提高。由于本体协商过程中不再需要进行请求说明，使得协商的原语大大减少。同时将协商的扩展原语在本体中加以定义也提高了其灵活性。

但是，该方法必须基于一定的元本体，如何建立普遍的通用的元本体是该方法所面临的一个重大问题。只有建立了相应的标准，才能建立一个大家都认可的元本体知识。此外，该方法中 Agent 进行理解和抽象时如何判定其最终的抽象层次也是尚待研究的问题。由于 Agent 理解时需要不断地进行抽象，因此 Agent 理解可能需要更多的时间，但是与通过通信来进行说明相比，Agent 自身进行抽象可能是一个更好的方法。

参考文献

1 Uschold M. Ontologies: Principles Methods and Applications. The Knowledge Engineering Review, 1996,11(2):93～136

2 Cranefield S, Willmott S. Introduction to the Special Issue on Ontologies in Agents Systems. The Knowledge Engineering Review, 2002,17(1):1～5

3 金芝. 知识工程中的本体论研究. 世纪之交的知识工程，陆汝钤编著，2001,6:447～465

4 Pease A, Niles I. IEEE Standard Upper Ontology: a Progress Report. The Knowledge Engineering Review, 2002,17(1):65～70

5 Bailin S C, Truszkowski W. Ontology Negotiation Between Intelligent Information Agents. The Knowledge Engineering Review, 2002,17(1):7～19

6 Tamma V, Bench-Capon T. An Ontology model to facilitate knowledge-sharing in multi-agent systems. The Knowledge Engineering Review, 2002,17(1):41～60

7 Tate A. Towards a Plan Ontology. AI-IA Notizie, Special Issue on Aspects of Planning Research, 9(1):19～26

8 刘志忠，姚莉. 基于本体的计划表示. 见：第七届中国人工智能年会，中国，广西

9 Dimitrov M, et al. OntoMap: Upper-Ontology Service Agent. In: Proc. of Workshop on Ontologies in Agent Systems Autonomous Agent 2001, Montreal Canada on May, 2001

10 Omelayenko B. RDFT: A Mapping Meta-Ontology for Business Integration. In: Proc. of Workshop on Knowledge Transaction for Semantic for Semantic Web at the 15th European Conf. on AI (KTSW-2002), Lyon France 2002. 77～84