一种新的使用辨识集的属性约简算法*

史岳鹏[1],朱颢东[2,3]

(1.郑州牧业工程高等专科学校 信息工程系,河南 郑州 450011;
2.中国科学院成都计算机应用研究所,四川 成都 610041;
3.中国科学院研究生院,北京 100039)

摘 要:为基于差别矩阵的属性约简算法求解时,先要求出差别矩阵,问题规模增大,将导致存放差别矩阵的空间过大和算法执行时间过长。针对这一问题,本文提出了辨识集的定义,并利用辨识集设计了新的属性约简算法,减少了存储量和计算量,提高了算法的效率。

关键词:粗糙集;差别矩阵;辨识集;属性约简

中图分类号:TP301 **文献标识码**:A

New attribute reduction algorithm employed discernible sets

SHI Yue Peng[1], ZHU Hao Dong[2,3]

(1.Department of Information Engineering, Zhengzhou College of Animal Husbandry Engineering, Zhengzhou 450011, China;
2.Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China;
3.The Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: In attribute reduction algorithms based on discernible matrix, discernible matrix must be acquired firstly. But the space of storing the discernible matrix in computer is very difficulty when the scale of the problem is very large. Moreover, the computing cost of the algorithm is higher. In order to solve above-mentioned problems, the definition of discernible sets is firstly provided, and then a new attribute reduction algorithm based on the discernible sets is designed. so it can cut down the computing and storing capacity greatly.

Key words: rough set; discernible matrix; discernible set; attribute reduction

粗糙集RS(Rough sets)理论[1]是由Pawlak在上世纪80年代初提出来的一种处理不精确、不相容、不完全和不确定知识的软计算工具,它以从新的角度对知识进行了定义,把知识看作是关于论域的划分,从而认为知识是具有粒度(granularity)的,知识的不精确性是知识的粒度大小引起的。粗糙集理论已经引起了许多数学家、逻辑学家、计算机研究人员,特别是人工智能研究人员的兴趣。目前已被广泛应用于机器学习、决策分析、数据挖掘、过程控制、数据分析等领域[2]。

知识约简是粗糙集理论的核心内容之一。所谓知识约简,就是在保持知识库的分类或决策能力不变的条件下,删除其中不相关或不重要的知识,导出问题的决策或分类规则。知识库中的知识(属性)并不是同等重要的,甚至其中某些知识是冗余的,特别当知识库数据随机采集时,其冗余性更为普遍。冗余知识的存在,一方面是对资源的浪费(需要存储空间);另一方面,干扰人们做出正确而简洁的决策。基于粗糙集理论的知识获取,主要是在保持决策表的决策属性和条件属性之间的依赖关系不发生变化的前提下对原始决策表进行约简。

属性约简是知识约简中非常重要的概念,它反映了决策表的本质信息。基于差别矩阵的属性约简算法是粗糙集理论中一类经典的属性约简算法[3]。但是随着问题规模的增大,这类算法存放差别矩阵的空间和算法执行时间的代价都很大。参考文献[3-5]中给出了相关的改进算法,但这些算法仍要存放差别矩阵,参考文献[6]中虽将差别矩阵转换成特征矩阵,但特征矩阵的存放和计

* 基金项目:四川省科技计划项目(2008GZ0003)

算与差别矩阵并无区别。为解决上述问题，本文提出了辨识集的定义，进而给出了基于辨识集的属性约简的定义。同时证明了该定义与基于差别矩阵的属性约简定义是等价的。在此基础上，设计了一个新的属性约简算法，由于这一算法在求属性约简的过程中不用生成差别矩阵和大量的无用元素，从而减少了存储量和计算量，提高了算法的效率。

1 粗糙集相关基础知识

定义 1：信息系统[7] $S=<U,R,V,f>$，其中 U 为对象集合，$R=C\cup D$ 是属性集合，C 为条件属性集，D 为决策属性集，$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示属性 r 的值域，$f:UR\rightarrow V$ 是一个映射函数，它指定 U 中每一个对象 X 的属性值。信息系统也可用二维表来表示，称之为决策表，其中行代表对象 x_i，列代表属性 r，$r(x_i)$ 表示第 i 个对象在属性 r 上的取值。

定义 2：知识约简[7]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知识。

设 P 为属性集，$R\in P$，$\mathrm{Ind}(P)J$ 表示在 P 上的等价关系。如果 $\mathrm{Ind}(P)=\mathrm{Ind}(P-[R])$。则称属性 R 为 P 中可省的，否则称 R 为 P 中不可省。

设 $Q\subseteq P$，若 Q 中所有属性都是不可省的，且 $\mathrm{Ind}(Q)=\mathrm{Ind}(P)$，则称 Q 为 P 的约简。记为 $\mathrm{red}(P)$。P 中所有不可省关系组成的集合称为 P 的核，记为 $\mathrm{core}(P)$。核与约简有如下关系：$\mathrm{core}(P)=\cap\mathrm{red}(P)$。

定义 3：决策表[7] $S=<U,C,D,V,f,d>$ 的差别矩阵是 $n\times n$ 的对称矩阵 $M_{n\times n}=(m_{ij})$，其元素定义为：

$$m_{ij}=\{a\mid a\in C,f(x_i,a)\neq f(x_j,a)\}\cap(\exists s\in D,f(x_i,s)\neq f(x_j,s))\}$$

其中 $i,j=1,2,\cdots,n$。

2 基于差别矩阵的属性约简算法

基于差别矩阵的属性约简算法[3-5]是粗糙集理论中一类经典的属性约简算法，在这类算法中，通常是先求出差别矩阵，然后再根据所设置的启发信息选取 1 个属性放入属性约简中，在差别矩阵的元素中删除所有包含该属性 a 的元素，直至差别矩阵为空。基于差别矩阵的属性约简算法描述如下[3-5]：

Step1：生成差别矩阵 $\boldsymbol{M}_{n\times n}=(m_{ij})$，$B=\phi$，其中 B 用来存放属性约简。

Step2：若 $\boldsymbol{M}_{n\times n}=\phi$，则输出 B，否则转下一步。

Step3：根据启发信息从 $C-B$ 中选择属性 a，并加入到 $B=B\cup\{a\}$。

Step4：在 $\boldsymbol{M}_{n\times n}$ 的所有非空元素中去掉包含属性 a 的元素，转 Step2。

经过分析可知，该算法存在如下缺点：

(1)存放差别矩阵的空间可能很大。例如，当对象个数为 1 000 000 单元，条件属性的个数为 100 单元时，则存放差别矩阵的最大空间为 $100\times1\ 000\ 000\times(1\ 000\ 000-1)/2=5\times1\ 013$ 单元，这对算法的实现是很不利的；

(2)在所存放的元素中有很多是重复的，造成了存储空间的极大浪费。因为在属性约简算法中，显然要删除包含某一元素的所有元素，由于这些元素存放时占用了大量的空间，删除时就要花费大量的比较时间，显然这对算法的运行也是很不利的。

3 改进的属性约简算法

基于差别矩阵的属性约简算法求解，要先求出差别矩阵，当问题规模增大时，将导致存放差别矩阵的空间过大和算法执行时间过长。针对此问题，本文提出了辨识集定义，并利用辨识集设计了一个新的属性约简算法。

定义 4：决策表 $S=<U,C,D,V,f>$ 中 $\forall x_i\in U$，记 $D(U,C)=m\mid m=\{p\in C:f(x_i,p)\neq f(x_j,p)\}$ 且 $\exists d\in D,f(x_i,d)\neq f(x_j,d),i<j=1,2,\cdots,n\}$，称 $D(U,C)$ 为属性集 C 的辨识集。

记 $D(U,c)=m\mid c\in m,m=\{p\in C:f(x_i,p)\neq f(x_j,p)\}$ 且 $\exists d\in D,f(x_i,d)\neq f(x_j,d),i,j=1,2,\cdots,n\}$ 称 $D(U,c)$ 为属性 $c\in C$ 的辨识。

推理：在决策表 $S=<U,C,D,V,f>$ 中，$\boldsymbol{M}_{n\times n}=(m_{ij})$ 为决策表的辨识矩阵，$D(U,C)$ 为决策表的辨识集，则有：$\forall m_{ij}\in M_{n\times n}$ 且 $m_{ij}\neq\phi\Leftrightarrow m_{ij}\in D(U,C)$。

证明：对比辨识矩阵的定义 3 和辨识集的定义 4，很明显此推理成立。

根据推理，本文基于辨识集的属性约简算法为：

(1)给定信息系统 $S=<U,C,D,V,f>$；

(2)初始约简属性集 $B=\mathrm{NULL}$；

(3)求 $D(U,C)$；

(4) $\forall m\in D(U,C)$，若 $|m|=1$，则把 m 加入到 B，即 $B=\{m\}+B$，$D(U,C)=D(U,C)-\{d\mid m\in d,d\in D(U,C)\}$；

//* 相当于求核。

(5) $\forall b\in C-B$，选择 $\mathrm{MAX}\{|\{d\mid b\in d,d\in D(D,C)\}|\}$（若不止 1 个，可根据具体情况选择其一），$D(U,C)=D(U,C)-\{b\mid m\in d,d\in D(U,C)\}$，$B=\{b\}+B$；

(6)若 $D(U,C)$ 为 null，输出 B，算法结束；否则转向(5)。

4 改进的算法效率分析

新算法的最大优点是没有生成无用的元素，因为在(2)中获得了核属性，这样做可使得(3)开始就生成一个较小的搜索空间，显然这可以提高算法的效率，在(3)中生成 $D(U,C)-\{b\}$，其意义是能由属性 b 区分的元素就不用生成，相当于在基于差别矩阵的属性约简算法中删除差别矩阵中所有包含属性 b 的元素，由于在新算法中没有生成这样的元素，所以也就不用删除，极大地压缩了占用的存储空间，提高了算法的效率。

5 改进的算法例证

使用参考文献[6]中的决策表，分别采用本文算法与基于差别矩阵的属性约简算法(老算法)进行比较。表1为所用决策表，表2为其对应的特征矩阵，表3为本文算法所对应的辨识集。

表1 决策表[6]

U/C	O	T	H	W	D
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	cool	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

表2 对应的特征矩阵

O	OT	OTH	OTHW	TH	OTH	THW	OTW	OH
OW	OTW	OTHW	OTW	THW	OTHW	TH	OT	OHW
OTHW	THW	W	O	OW	TW	OT	OTH	OTW
OT	OT	THW	OTHW	TH	OH	HW	OW	OTH
OTW	W	THW	OTH	OTHW	HW	TH	OW	OTHW

表3 本文算法的辨识集

O,OT,OTH,OTHW,TH,THW,OTW,OH,OW,OHW,W,TH,HW

采用本文算法生成 $D(U,C)=\{O,OT,OTH,OTHW,TH,THW,OTW,OH,OW,OHW,W,TH,HW\}$，需要比较的次数为：$(9+9+3\times3+6+2+5+5)\times5=45\times5=225$。选择核时 D(U,C)内部比较的次数为13次，各个核生成$\{d|m\in d,d\in D(V,C)\}$时需比较$13\times2=26$次(有2个核)，核选择后，$B=\{W,O\}$ $D(U,C)=\{TH\}$；下一步选择 T 或 H 只需要比较1次即可，$B=\{T,W,O\}$ 或 $B=\{H,W,O\}$，$D(U,C)=\{\}$，算法结束。因此本文算法总的比较次数为：225+13+26+1=265次。每个属性元素存储时占1个存储单元，则本文算法只需要30个存储单元。而在参考文献[7]中分析的老算法的比较次数为447次，存储单元需116个。性能对比如表4所示。可见本文改进的属性约简算法大大提高了算法的效率。

表4 新老算法性能比较

	本文改进算法	老算法
比较次数	265	447
所需存储单元数	30	116

本文在分析传统基于辨识矩阵属性约简的基础上，提出了一种基于辨识集的属性约简算法，该算法无论在空间使用上还是在时间性能上都优于传统的基于辨识矩阵的属性约简算法，例证证明了本算法的优点，在属性约简中有一定的实用价值。

参考文献

[1] PAWLAK Z. Rough sets[J]. International Journal of Information and Computer Sciences.1982,11(5):341-383.

[2] LIANG J Y, CHIN K S, DANG C Y, et al. A new method for measuring uncertainty and fuzziness in rough set theory[J]. International Journal of General Systems, 2002, 31(4):331-342.

[3] 徐章艳，杨炳儒，宋威，等.一种快速计算HU差别矩阵的属性约简算法[J].小型微型计算机系统，2008，29(10)：1820-1827.

[4] 杨明，杨萍.基于广义差别矩阵的核和属性约简算法[J].控制与决策，2008(29)：1049-1054.

[5] 王柯，朱启兵.一种基于差别矩阵的启发式属性约简算法[J].计算机工程与科学，2008，30(6)：73-75.

[6] 赵卫东，戴伟辉.基于特征矩阵的决策表约简研究[J].系统工程理论与实践，2003，23(3)：65-69.

[7] 曾黄麟.智能计算[M].重庆：重庆大学出版社，2004.

(收稿日期：2009-06-04)

(上接第51页)

TR-069。虽然通过部署基于TR-069的网管系统，可以在很大程度上减少用户的配置和管理工作，提高设备的易用性和可管理性，便于家庭网络中设备的快速部署和业务的迅速开展。但从协议目前的发展情况来看，TR-069仍然处于一个不断完善的过程中，在业务参数模型上还需要加入对更多的终端业务和特性的支持。

参考文献

[1] 宋臣.终端自动配置管理研究[D].成都：西南交通大学，2006.

[2] 彭淑静.SNMP网络管理系统技术.甘肃科技纵横，2008，38(2).

[3] LIU Yun Xin, ZHANG Yao Xue. HNMP: a digital home network management model. ACTA Electronica Sinica,2001.

[4] 王勇，张尧学，方存好.一种改进的家庭网络管理协议—ExHNMP.小型计算机微型系统，2004,25(7).

[5] DSL Forum. CPE WAN management protocol [S]. Tech. rep. 069, May 2004.

[6] 王远波.家庭网关远程管理功能模块的设计与实现[D].武汉：华中科技大学，2009.

(收稿日期：2009-06-04)