维普资讯 http://www.cqvip.com

基于Skowron分明矩阵的快速约简算法*)

王元珍　裴小兵

(华中科技大学计算机学院数据库与多媒体研究所　武汉430074)

摘　要　属性约简是Rough集理论的核心内容之一,计算所有的属性约简已经被证明是NP完全问题。本文基于分而自治思想,在Skowron分明矩阵法的基础上,给出了最小析取范式的判定定理,从而提出了计算所有属性约简的算法。理论分析和实验结果表明,该约简算法在效率上较现有的算法有显著提高。

关键词　Rough集,属性约简,析取范式

A Fast Algorithm for Reduction Based on Skowron Discernibility Matrix

WANG Yuan-Zhen　PEI Xiao-Bing

(Institute of Database and Multimedia, Department of Computer Science, HuaZhong University of Science & Technology, Wuhan 430074)

Abstract　Attribute reduction is one of the basic contents, it is NP-complete problem to calculate all attribute reductions. Based on dividing and conquering thought, the judgement theorem with respect to minimal disjunctive normal form is obtained based on Skowron discernibility matrix, from which an algorithm for all attribute reductions is presented. Theoretical analysis and experimental results show that the algorithm is much more efficient in comparison with those existing algorithm.

Keywords　Rough set, Attribute reduction, Disjunctive normal form

1　引言

Rough集理论[1]是20世纪80年代由波兰的Pawlak教授提出,它是一种新型的处理模糊性和不确定性知识的数学工具,在数据挖掘、决策支持系统等领域得到了应用[2~5]。

属性约简是知识发现的基础,也是Rough集理论的核心内容之一,计算所有的属性约简已经被证明是NP完全问题[2]。目前,最常用的约简方法是基于分明矩阵的方法[3~5],但是基于分明矩阵的约简算法效率很低,为指数复杂度[3~5]$O(2^{|A|}|U|^2|A|)$。因此,当数据量很大时,这些约简算法的可行性将面临巨大的挑战。

现有Rough集算法计算的低效性在一定程度上限制了Rough集理论的广泛应用,因此寻求快速的Rough集算法具有重要的意义。本文基于分而自治思想,在Skowron分明矩阵法的基础上,给出了最小析取范式的判定定理,从而提出了计算所有属性约简的算法。理论分析和实验结果表明,该约简算法在效率上较现有的算法有显著提高。

2　Skowron分明矩阵法的基本概念

A.Skowron[3]于1991年提出了一种用分明矩阵表示知识的方法,这种表示有许多有利条件,成为人们最常用的计算所有属性约简的方法。

定义1　设信息系统 $S=\langle U,A,V,f\rangle$,其中 $U=\{x_0,x_1,\cdots,x_{n-1}\}$ 为论域,$A=\{a_1,\cdots,a_m\}$为属性集合,系统 S 的分明矩阵为 $M(S)=[c_{ij}]_{n\times n}$,其中矩阵项定义如下:$c_{ij}=\{a\in A, a(x_i)\neq a(x_j), x_i\in U, x_j\in U, i,j=1,2,\cdots,n-1\}$。

对于每一个分明矩阵 $M(S)$对应唯一的分明函数 $f_{M(S)}$,它的定义如下:信息系统 $S=\langle U,A,V,f\rangle$的分明函数 $f_{M(S)}$是一个有 m 一元变量 $a_1,\cdots,a_m$($a_i\in A, i=1,2,\cdots,m$)的布尔函数,它是 $\wedge c_{ij}$的合取,c_{ij}是矩阵$[c_{ij}]_{n\times n}$中各元素的析取($0<=j,i<=n-1$)[3~5]。

引理1　设 $S=\langle U,A,V,f\rangle$是一个信息系统,$f_{M(S)}$是 S 的分明函数,则该函数的最小简化的析取范式对应于 S 的全体属性约简[3~5]。

Skowron分明矩阵法的步骤如下[3~5]:

(1)计算信息系统 S 的分明矩阵 $M(S)$;(2)计算与分明矩阵 $M(S)$相关联的分明函数 $f_{M(S)}$;(3)计算分明函数 $f_{M(S)}$的最小简化的析取范式的合取项构成的集合 $DNF_{M(S)}$,它将给出所有的属性约简。

3　最小析取范式的判定定理

定义2　设信息系统 $S=\langle U,A,V,f\rangle$,$M(S)=[c_{ij}]_{n\times n}$为系统 S 的分明矩阵,$f_{M(S)}$为系统 S 的分明函数,$G\subseteq\{c_{ij}\mid 0<=j,i<=n-1\}$,定义在 G 上的分明子函数 f_G 为:

f_G是一个有 m 一元变量 $a_1,\cdots,a_m$($a_i\in A, i=1,2,\cdots,m$)的布尔函数,它是 $\wedge g_k$($g_k\in G$)的合取,g_k 是对应的 c_{ij}中各元素的析取。记 DNF_G 为分明子函数 f_G 的最小简化的析取范式的合取项构成的集合。

例　设 $M(S)=\{a\vee c, b\vee d\vee c, c\vee d, d\vee a\}$, $G=\{a\vee c, b\vee d\vee c\}$,则 $f_G=(a\vee c)\wedge(b\vee d\vee c)$, $DNF_G=\{(a\vee c),(b\vee d\vee c)\}$。

定义3　设信息系统 $S=\langle U,A,V,f\rangle$, $P\subseteq 2^A$,其中2^A 为 A 的幂集合,$R=\{\wedge a_k(a_k$ 的合取,其中 $a_k\in p)\mid p\in P\}$,则 R 的消简是指满足如下条件的 R':

1)$R'\subseteq R$;2)$\forall R'_i\in R', R'_j\in R'$都有 $R'_i\bigcap R'_j\neq R'_i$ 且 $R'_i\bigcap R'_j\neq R'_j$;3)对任意的 $R_k\in R$ 都存在 $R'_i\in R'$,使得 $R_k\bigcap R'_i=R'_i$。

*)本文得到科技部电子政务关键技术及应用系统研究项目的资助(项目编号:2001BA110B01)。王元珍　教授,博士生导师,主要研究方向为:现代数据库理论及实现技术;裴小兵　博士研究生,主要研究方向为:数据挖掘、数据库、网络及网络安全。

维普资讯 http://www.cqvip.com

例 取 $R=\{\{a\wedge c\},\{a\},\{b\wedge c\},\{d\},\{d\wedge e\}\}$，则 R 的消简为 $R'=\{\{a\},\{b\wedge c\},\{d\}\}$。

计算 R 的消简的时间复杂度为 $O(|A|m^2)$，其中 m 表示 R 中元素的个数。

定理1 设信息系统 $S=\langle U,A,V,f\rangle$，$M(S)=[c_{ij}]_{n\times n}$为系统 S 的分明矩阵，$G_1\subseteq\{c_{ij}|0<=j,i<=n-1\}$，$G_2\subseteq\{c_{ij}|0<=j,i<=n-1\}$，$G_1\cup G_2=\{c_{ij}|0<=j,i<=n-1\}$，$G_1\cap G_2=\Phi$，$\mathrm{DNF}_{M(S)}$为系统 S 的分明函数 $f_{M(S)}$的最小简化的析取范式的合取项构成的集合，DNF_{G_1}为集合 G_1的分明子函数 f_{G_1}的最小简化的析取范式的合取项构成的集合，DNF_{G_2}为集合 G_2的分明子函数 f_{G_2}的最小简化的析取范式的合取项构成的集合。$R=\{d_1\wedge d_2|d_1\in\mathrm{DNF}_{G_1},d_2\in\mathrm{DNF}_{G_2}\}$，则 $\mathrm{DNF}_{M(S)}$等于 R 的消简。

证明：记 R 的消简为 B。对任意的 $\beta\in\mathrm{DNF}_{M(S)}$，则存在 r_{ij} c_{ij} $(i,j\leqslant n)$，使得 $r_{ij}\subseteq\beta$，所以 $r_{00}\wedge\cdots\wedge r_{ij}\wedge\cdots r_{nn}\subseteq\beta$。又 $\mathrm{DNF}_{M(S)}$为分明函数 $f_{M(S)}$的最小析简化的取范式的合取项构成的集合，所以 $r_{00}\wedge\cdots\wedge r_{ij}\wedge\cdots\wedge r_{nn}=\beta$。假设 $r_{00}\wedge\cdots\wedge r_{ij}\wedge\cdots\wedge r_{nn}\notin B$，则存在 $d_1\in\mathrm{DNF}_{G_1},d_2\in\mathrm{DNF}_{G_2}$，使得 $d_1\wedge d_2\subset\beta$，又 $d_1\wedge d_2\in\{f_{M(S)}$的合取项构成的集合$\}$，这与 $\beta\in\mathrm{DNF}_{M(S)}$矛盾。所以：$r_{00}\wedge\cdots\wedge r_{ij}\wedge\cdots\wedge r_{nn}\in B$，即 $\beta\in B$。

另一方面，$\forall\gamma\in B$，则一定存在 $r_{ij}\in c_{ij}(i,j\leqslant n)$，使得 $\gamma=r_{00}\wedge\cdots\wedge r_{ij}\wedge\cdots\wedge r_{nn}$。假设 $\gamma\notin\mathrm{DNF}_{M(S)}$，则存在 $\gamma'\in\mathrm{DNF}_{M(S)}$且 $\gamma'\subset\gamma$，从而存在 $r'_{ij}\in c_{ij}(i,j\leqslant n)$，使得 $r'_{00}\wedge\cdots\wedge r'_{ij}\wedge\cdots\wedge r'_{nn}\subseteq\gamma'\subset\gamma$，即存在 $d'_1\in\mathrm{DNF}_{G_1},d'_2\in\mathrm{DNF}_{G_2}$使得 $d'_1\wedge d'_2\subset\gamma$，这与 $\gamma\in B$ 矛盾。所以 $\gamma\in\mathrm{DNF}_{M(S)}$。

4 快速计算属性约简的方法

计算所有的属性约简已经被证明是 NP 完全问题[2]。文[3～5]是基于分明矩阵的属性约简算法，效率很低，为指数复杂度 $O(2^{|A|}|U|^2|A|)$。

根据定理1，下面将给出快速计算属性约简的方法。

算法1 计算分明函数 $f_{M(S)}$的最小简化的析取范式的合取项构成的集合。

输入：信息系统 S 的分明矩阵 $M(S)=[c_{ij}]_{n\times n}$，集合 $M=\{c_{ij}|0<=j,i<=n-1\}$；

输出：分明函数 $f_{M(S)}$的最小析取范式的合取项构成的集合 $DNF_{M(S)}$。

FunctionMinDnf(M,DNF_M)

{(1)$\mathrm{DNF}_1=\Phi$；$\mathrm{DNF}_2=\Phi$；如果 $M=\Phi$；return Φ；如果 $|M|=1$，return $\mathrm{DNF}_M=M$；

(2)将集合 M 等分为两个集合 M_1、M_2。

(3) FunctionMinDnf (M_1, DNF_{M_1})；FunctionMinDnf (M_2,DNF_{M_2})；

(4)计算 $R=\{d_1\wedge d_2|d_1\in\mathrm{DNF}_{M_1},d_2\in\mathrm{DNF}_{M_2}\}$；

(5) $\mathrm{DNF}_M=R$ 的消简；Return DNF_M；

}

设算法时间复杂度为 $T(n)$，第2步的时间复杂度为 $T2=O(n)$；第3步的时间复杂度为 $T3=2T(\frac{n}{2})$；第4步的时间复杂度为 $T4=O(K^2)$；由于 $|R|\leqslant K^2$，因此，第5步的时间复杂度为 $T5=O(|A|K^4)$。所以，算法的时间复杂度为 $T(n)=2T(\frac{n}{2})+|A|K^4\leqslant O(n+(|A|K^4)\log^n)=(|A|K^4\log^{|U|}+|U|^2)$，其中 $K=\mathrm{Max}\{Card(\mathrm{DNF}_G)|G\subseteq M\}$，$Card(\mathrm{DNF}_G)$表示 DNF_G 中合取项的个数。

算法2 计算所有属性约简。

输入：信息系统 $S=(U,A,V,F)$；

输出：信息系统 $S=(U,A,V,F)$的所有属性约简集 R。

(1)计算信息系统 S 的分明矩阵 $M(S)=[c_{ij}]_{n\times n}$，构造集合 $M=\{c_{ij}|0<=j,i<=n-1\}$；

(2)FunctionMinDnf(M,DNF_M)；

(3)由 DNF_M 给出所有的属性约简。

根据算法2，易得算法3的时间复杂度 $T=O(|A|K^4\log^{|U|}+|A||U|^2)=O(|A|(K^4\log^{|U|}+|U|^2))$，其中 $K=\mathrm{Max}\{Card(\mathrm{DNF}_G)|G\subseteq M\}$，$Card(\mathrm{DNF}_G)$表示 DNF_G 中合取项的个数。

文[3～5]给出了计算所有属性约简的算法，但它们的时间复杂度都为指数复杂度，当 $|A|$ 比较大时文[3～5]中相应算法的时间复杂度将爆炸式地增长。因此，在 K 是关于 $|U|$ 与 $|A|$ 的多项式时，本文给出的计算所有属性约简算法的时间复杂度 $O(|A|(K^4\log^{|U|}+|U|^2))$明显优于文[3～5]$O(2^{|A|}|U|^2|A|)$中相应算法的时间复杂度。

5 实验结果及分析

我们选用 UCI 机器学习数据库中的9个数据库在 PC 机(Intel-Pentium，2GHz，256MB RAM，Win2000 Professional，Microsoft Access)上进行实验，分别采用文[3～5](简称算法 A，算法 A 是 Skowron 分明矩阵法)和本文中的约简算法(算法2)进行属性约简，实验结果如表1所示。

从表1可以看出，算法2在效率上明显优于算法 A。随着实例数增加，算法2在效率上的提高越显著，而算法 A 的效率急剧下降。实验结果和前面的分析相符，验证了算法2的高效性，高效性，表明该算法在效率上优于文[3～5]中相应的约简算法。

表1 约简算法比较

数据库名称	实例数	属性个数	算法2	算法 A
			执行时间(s)	执行时间(s)
Postoperative Patient	90	9	0.109	0.25
Hayes-Roth Database	132	6	0.172	0.344
Balance-scale database	625	5	5.609	183.437
Teaching Assistant Evaluation	152	6	0.232	0.297
Zoo Database	101	17	0.156	0.75
Tic-Tac-Toe Endgame Database	958	10	18.656	2628.25
Car Evaluation Database	1728	7	71.4672	5517.14
BUPA liver disorders	345	7	1.578	2.156
Monk's Problems (1)	432	7	2.688	60.891
Monk's Problems(2)	432	7	2.704	52.36
Monk's Problems(3)	432	7	2.500	66.454

维普资讯 http://www.cqvip.com

命名实体识别研究*)

张晓艳　王　挺　陈火旺
(国防科技大学计算机学院　长沙410073)

摘　要　命名实体识别是文本信息处理的重要基础,已经逐步成为自然语言处理的一项关键技术。其基于规则、统计、机器学习的研究方法及成果,都推动了自然语言处理研究的发展,促进了自然语言研究与应用的紧密结合。本文回顾了命名实体识别技术的发展过程,分析了主要的方法和技术,并展望了未来的发展趋势。

关键词　命名实体识别(NER),隐马尔可夫模型(HMM),最大熵模型(ME)

Research on Named Entity Recognition

ZHANG Xiao-Yan　WANG Ting　CHEN Huo-Wang
(Department of Computer, National University of Defense Technology, Changsha 410073)

Abstract　Named Entity Recognition (NER), as the important foundation of Text Processing, has become a key technology of Natural Language Processing (NLP). It has made a great impetus to the research of NLP with its method and result. Moreover, NER has built a more effective connection between NLP theories and applications. This paper will review the history of the research, analyze the methods and technologies and give a view of the future of NER.

Keywords　Named entity recognition (NER), Hidden markov model (HMM), Maximum entropy model (ME)

1　引言

随着计算机的普及以及各种电子文本的广泛应用,海量的信息给人们的信息获取带来了严峻的挑战,人们迫切需要一些自动化工具帮助进行海量信息处理。信息抽取、信息检索、机器翻译、文摘生成等技术正是在这种背景下产生的。在这些技术中,一个共同而基础的问题就是命名实体识别(Named Entity Recognition)。命名实体识别作为这些研究中非常重要并且是必不可少的关键技术,越来越得到人们的重视和关注,时至今日已经发展成一个独立的研究分支,COLING2002就有专门的命名实体识别专题。

在一篇文章中,实体名字是基本的信息元素,往往指示了文章的主要内容。命名实体识别是对文本进行理解的前提工作。命名实体识别的质量会直接影响到后续的一系列工作,例如在信息抽取中如果没有先识别实体,根本就不可能识别实体关系;在文摘生成中,很多时候是对固定模式的填充,填充内容大都是"谁","干什么","什么时候","在哪里"等等,这正是命名实体的内容,因此从文章中获取这些内容就离不开命名实体识别;又如,在机器翻译中命名实体的翻译往往需要特殊处理。由此可见,命名实体识别已经越来越成为自然语言处理中的关键技术。因此本文就命名实体识别技术研究的历史和现状进行分析和介绍,并对未来的发展趋势作一展望。

本文首先回顾一下命名实体识别的发展和存在的问题,然后介绍命名实体识别的方法和技术,并分析了两个具体的命名实体识别系统,最后对命名实体识别的研究趋势进行了展望。

2　命名实体识别问题概述

命名实体识别最初是在MUC-6(Message Understanding Conference)上作为一个子任务提出的[1]。命名实体识别任务主要是要识别出文本中出现的专有名称和有意义的数量短语并加以归类。所谓的名字实体(Named Entity)主要包括实体(组织名、人名、地名)、时间表达式(日期、时间)、数字表达式(货币值、百分数)等。就整个的命名实体识别的研究结果而言,时间表达式和数字表达式的识别相对简单,其规则的设计、数据的统计训练等也比较容易。而对于实体中的组织名、人名、地名,因为其具有开放性和发展性的特点,而且构成规律有很大的随意性,所以其识别就可能会有较多的错选或漏选。现在大多数的命名实体识别的研究都集中于对这三种实体的识别技术的研究。

命名实体识别研究至今已经有近二十年的发展历史,已经成为自然语言处理领域的一项重要技术,并取得了很多成

*)本课题得到国家"863"高技术研究发展计划资助(2001AA114110)。张晓艳　硕士生,主要研究方向为自然语言处理;王　挺　博士,副教授,主要研究方向为自然语言处理、计算机软件;陈火旺　院士,教授,博士生导师,主要研究领域为人工智能、计算机软件。

总结　知识约简是知识发现的基础,也是Rough集理论的核心内容之一,计算所有约简已经被证明是NP完全问题。本文在分明矩阵法的基础上,给出了最小析取范式的判定定理,从而提出了计算所有属性约简的方法。理论分析和实验结果表明,该约简算法在效率上较现有的算法有显著提高。

参考文献

1　Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982, 11(5): 341～356
2　Wong S K M, Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences, 1985(33): 693～696
3　Skowron A, Rauszer C. The discernibility matrics and function in information system. In: Slowinski R, ed. Intelligent Decision Support Handbook of Application and Advances of the Rough sets Theory. Dordreecht: Kluwer Academic Publishers, 1991. 331～362
4　刘清. Rough集及Rough推理. 北京:科学出版社,2001
5　张文修,吴伟志,梁吉业,李德玉. 粗糙集理论与方法. 北京:科学出版社,2001