一种对应约束的决策表属性约简算法

成红红[1] 张晓琴[2] 李飞江[1] 钱宇华[1]
(山西大学计算机与信息技术学院 太原 030006)[1] (山西大学数学科学学院 太原 030006)[2]

摘 要 决策表属性约简是粗糙集理论中的重要问题,经典决策表属性约简方法从保持论域划分能力的角度出发,选择最优条件属性约简集。从决策属性与条件属性的相关性角度出发,将决策表属性约简思想与传统统计学中的对应分析方法相结合,提出了一种量化决策属性与条件属性之间依赖关系的度量,称为投影区分度,并基于此发展了一种决策表属性约简算法。最后用简单实例说明了该方法的正确性。

关键词 决策表,属性约简,对应约束,投影区分度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.011

Decision Table Attribute Reduction Algorithm Based on Correspondence Constraints

CHENG Hong-hong[1] ZHANG Xiao-qin[2] LI Fei-jiang[1] QIAN Yu-hua[1]
(School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China)[1]
(School of Mathematics,Shanxi University,Taiyuan 030006,China)[2]

Abstract Decision table attributes reduction is an important problem in rough set theory,and classical decision table attributes reduction methods choose the optimal condition attribute reduction set from the perspective of maintaining the classification ability of universe. Taking the correlation of decision attributes and condition attributes into account, by combining attributes reduction idea with correspondence analysis method in traditional statistical methods, this paper proposed a quantitative measurment to measure the dependent relationship between decision attributes and condition attributes, called projection differentiation. Based on the measurement, we developed a decision table attributes reduction algorithm. Finally, a simple example was given to illustrate the correctness of the proposed method.

Keywords Decision table, Attribute reduction, Correspondence constraints, Projection differentiation

1 引言

属性约简是粗糙集理论中最重要的问题之一,目的是在产生决策规则之前,明确知识系统中潜在的信息,在保持分类能力不变的前提下删减知识库中冗余的属性,从大量的属性集中求取最小的属性约简集[1]。自粗糙集理论提出至今的30年里,已经发展了很多有代表性的属性约简算法,比如Hu等人在区分矩阵基础上提出的基于属性重要性的启发式约简算法[2]、Jelonek等提出的以属性增益为重要性的启发式算法[3]、苗夺谦提出的基于互信息的属性约简算法[4]、王国胤提出的基于信息熵的属性约简算法[5]。

粗糙集理论作为一种处理不确定性、不精确、不完全数据的数学方法,已在数据挖掘、人工智能、金融决策分析、图像处理等很多领域广泛应用[6-11]。随着大数据时代的到来,实际数据量剧增,如何从海量数据中找到有用的数据,尽可能提高现有算法的效率并降低时间空间复杂度,也是粗糙集理论的研究方向。粗糙集方法与别的数据分析方法的融合也是一种趋势。统计学分析方法是典型的数据分析学科,已有研究者将粗糙集理论方法与传统的统计学方法相结合,比如刘宏杰提出粗糙集属性约简判别分析方法来提高预测速度[12]。

对应分析是近年来发展起来的一种多元统计分析方法,由法国统计学家Benzenci提出[13,14]。该方法通过分析定性变量构成的交互汇总表来揭示变量间的联系,不仅能揭示同一变量的各类别之间的差异,还能反映不同变量各个类别之间的对应关系。对应分析的基本思想是将一个列联表的行和列中各元素的比例结构以点的形式在低维的空间中表示出来,其最大的特点就是将多个行点和列点同时表示在一张图上,将行信息和列信息直观明了地在图上标示出来。其因直观、简单、方便,在市场细分、地质研究、社会调查及计算机工程等各个领域得到了广泛应用。

对应分析实质是一种图表示方法,可表示在低维空间中的变量间的关系。粗糙集决策表的条件属性的取值类型多是定性变量,且在之前的决策表属性约简研究中,大多数方法是从约简属性对论域的划分与原来属性集的划分相同的角度出发,在寻找最优子集的过程中决策属性与条件属性的依赖关系比较抽象,而且只注重最后的约简结果,决策属性与条件属

到稿日期:2014-04-23 返修日期:2014-05-15 本文受国家自然科学基金重点项目(71031006),国家青年基金项目(41101440),山西省专项科研项目(20102003)资助。

成红红(1986－),女,博士生,CCF会员,主要研究方向为统计机器学习,E-mail:chhsxdx@163.com;**张晓琴**(1975－)女,博士,副教授,主要研究方向为统计机器学习、应用统计;**李飞江**(1990－),男,博士生,主要研究方向为大数据机器学习;**钱宇华**(1976－),男,博士,教授,主要研究方向为模式识别、特征选择、粗糙集、粒计算、人工智能。

性之间的相依关系没有引起足够关注。

本文针对粗糙集决策表的特点及对应分析的优越性，结合粗糙集属性约简的思想，从条件属性对决策的影响大小的角度出发，通过量化条件属性与决策属性之间的关系，提出一种量化决策属性与条件属性之间依赖关系的度量，称为投影区分度。基于此发展了一种决策表属性约简算法，旨在定量地刻画决策属性与条件属性之间的相关关系。最后通过实例表明该方法是合理并且行之有效的。

2 基于对应分析的粗糙集决策表属性约简

2.1 粗糙集决策表属性约简思想

给定决策表 $S=(U,C\cup D,V,f)$，条件属性集 C 中的有些属性可能并不重要。传统的约简思路是从属性重要性的角度出发，删除决策表中的某一条件属性，观察决策表的分类能力变化程度，如果变化大则认为该属性比较重要，反之，认为其不重要[15,16]。本文沿用传统约简思路，但是采用前向贪婪约简思想[17]，逐次增加约简集中属性的个数，使得决策表的划分能力不变。

2.2 对应分析

对应分析是在因子分析的基础上发展起来的一种多元相似变量统计技术[18]，其本质是一种数据降维技术，把R型因子和Q型因子结合起来研究行之间、列之间、行列之间的相关性，并将数据信息投影到"最佳"二维坐标系中，也即将列联表的各种性质以图像的形式清晰地显示。

对应分析的步骤：

(1)计算样本概率矩阵：假设原始数据集 $X_{n\times m}$ 表示 n 个 m 维空间的观测值，其中 x_{ij} 表示频数或计数，设 N 为 X 中的总频数，即 $N=\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}$，构造样本概率矩阵 $P=\{p_{ij}\}_{n\times m}$，其中 $p_{ij}=\frac{x_{ij}}{N}$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。定义样本行条件概率为 $r_i=\sum_{j=1}^{m}p_{ij}$，$i=1,2,\cdots,n$；列条件概率为 $c_j=\sum_{i=1}^{n}p_{ij}$，$j=1,2,\cdots,m$。

(2)对应变换：计算对应矩阵 $Z=\{z_{ij}\}_{n\times m}$，且 $z_{ij}=\frac{p_{ij}-r_ic_j}{\sqrt{r_ic_j}}$，$i=1,2,\cdots,n,j=1,2,\cdots,m$。$A$ 表示列属性间的协方差阵用于R型因子分析，则 $A=Z^TZ$；B 表示行样本间的协方差阵用于Q型因子分析，则 $B=ZZ^T$。

(3)因子分析：根据奇异值分解，$Z_{n\times m}$ 可以表示为 $Z=U_1\Lambda V_1$，其中 Λ 表示矩阵 $Z_{n\times m}$ 所有特征根组成的对角矩阵，$\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_{l=\min(n,m)},0,0,\cdots,0)$，$U_1$ 为矩阵 B 的特征根的特征向量矩阵，V_1 为矩阵 A 的特征根的特征向量矩阵。因此R型因子的载荷矩阵为 $R=V_1\sqrt{\Lambda}$，Q型因子的载荷矩阵为 $Q=U_1\sqrt{\Lambda}$。

(4)绘制对应分析图：Eckart-Young 定理已经证明了数据集 $Z_{n\times p}$ 的第 i 个观测值 z_i 的最佳二维近似就是用它的前两个主成分的样本值逼近[19]，也即对于R或Q型因子在 R_1-R_2 或 Q_1-Q_2 直角坐标中系绘制投影图，即得到对应分析因子聚点图。由步骤(3)可知，R和Q型因子的载荷矩阵具有线性变换关系，因此以特征值为尺度，可将列信息和行信息表示在一张图上。本文的主要目的是分析决策表中属性间的相关关系，因此更多关注R型因子分析图。

2.3 决策表的转化

对应分析特别延用于处理定性数据，如处理项目类目反映表[20]，因此本文将决策表按一定的规则转化成项目类目表进行对应分析。

给定决策表 $S=(U,C\cup D,V,f)$，U 为对象的论域，C 为由 m 个属性组成的条件属性集，且第 j 个属性有 r_j 个取值，$V=\bigcup_{a\in C}V_a$，V_a 是属性 a 的值域，D 为决策属性，$f:U\times C\rightarrow V$ 是一个信息函数，为每个对象的每个属性赋予一个信息值，即 $\forall a\in A,x\in U,f(x,a)\in V_a$。本文按如下定义将决策表转化成项目类目表。

定义 1 将决策表中 m 个属性看作 m 个项目，r_j 个取值看作是第 j 个项目含有 $r_j(j=1,2,\cdots,m)$ 个类目，则共有 $q=\sum_{j=1}^{m}r_j$ 个项目，此时决策表 S 转换成列联表(项目类目表)K。K 中的元素为：

$$x_{ij}^{k}=\begin{cases}1, & \text{第 } i \text{ 个元组在第 } j \text{ 个项目的第 } k \text{ 类目上有反映}\\ 0, & \text{第 } i \text{ 个元组在第 } j \text{ 个项目的第 } k \text{ 类目上没有反映}\end{cases}$$

其中，$i=1,2,\cdots,n;j=1,2,\cdots,m;k=1,2,\cdots,r_j$。

2.4 基于对应分析的决策表属性约简算法

对应分析中R因子载荷矩阵的前两维能很好地近似列属性信息，本文也采用对应图的性质给出一种决策属性与条件属性的依赖性度量。

定义 2 设 $v_C=(v_{C1},v_{C2})$ 为条件属性 C 的因子载荷得分向量，v_{C1},v_{C2} 分别为属性 C 在对应图中的坐标，$v_D=(v_{D1},v_{D2})$ 为决策属性 D 的因子载荷得分向量，v_{D1},v_{D2} 分别为属性 D 在对应图中的坐标。$\|v_C\|=\sqrt{v_{C1}^2+v_{C2}^2}$，$\|v_D\|=\sqrt{v_{D1}^2+v_{D2}^2}$ 分别表示向量 v_C,v_D 的模长，$\cos\langle v_C,v_D\rangle=\frac{\langle v_C,v_D\rangle}{\|v_C\|\,\|v_D\|}$ 表示向量 v_C 与 v_D 之间的夹角。

注：$\|v_C\|$，$\|v_D\|$ 的模长表示向量所代表的变差，模长越长，向量代表的变差越大；$\cos\langle v_C,v_D\rangle$ 表示两个向量之间的相似程度，夹角越大，向量间的相似程度越强。

在此定义基础上，给出决策属性与条件属性的依赖程度度量，称为投影区分度：

$$\gamma(C,D)=\|v_D\|\cos\langle v_C,v_D\rangle \tag{1}$$

表示决策属性 v_D 所代表的变差在条件属性 v_C 方向上的投影，$\gamma(C,D)$ 值越大，表示决策属性 D 受条件属性 C 的影响较大，也即两者的依赖程度较高，也可认为条件属性 C 相对于决策属性 D 更重要，或者决策属性 D 越偏好于条件属性 C。

图1给出 γ 的几何解释。

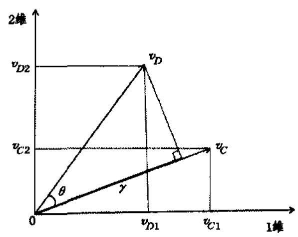

图1 γ 的几何解释

图1中 $\theta=\cos\langle v_C,v_D\rangle$ 表示 v_C 与 v_D 之间的相关关系，也

即量化后的条件属性与决策属性之间的相关关系,γ 表示 v_D 在 v_C 上的投影长度,描述决策属性受条件属性的影响程度。由图 1 可知,γ 越长,表明依赖程度越高,也可理解为决策属性越偏好于条件属性。

本文结合粗糙集理论前向贪婪属性约简思想,根据判断准则,即寻找到的最优条件属性集使其与决策属性的依赖度和原属性集的相同,提出一种对应约束的决策表属性约简算法。

算法 基于对应分析的决策表属性约简算法

输入:决策表 $S=(U,C\cup D,V,f)$

输出:条件属性 C 相对于决策属性 D 的一个相对约简 B

第一步:根据定义 1,将决策表 S 转化成列联表 K,并将 K 分割成 m 个子列联表 K_j,$j=1,2,\cdots,m$,K_j 表示第 j 个属性与决策属性的列联表;

第二步:令 $B=\emptyset$;

第三步:计算每个 K_j 中的 $\gamma(a_j,D)$,$j\leqslant m$,并比较大小,若 $\gamma(a_0,D)\geqslant\gamma(a_i,D)$,$i=1,2,\cdots,m$,令 $B=B\cap\{a_0\}$;

第四步:若 $\gamma(B,D)\neq\gamma(C,D)$,将 $BDa_k\in(C-B)$ 组成的决策表转化为 $k=|C-B|$ 个相应的列联表,并计算 $\gamma(B,D)=\max\{\gamma(a_k+B,D),a_k\in(C-B)\}$,若 $\gamma(B,D)=\gamma(C,D)$,算法终止;

第五步:输出属性约简 B。

3 实验与分析

如表 1 所列的决策表,条件属性 $C=\{$头疼,肌肉疼,体温$\}$,决策属性 $D=\{$流感$\}$。根据上述算法对决策表进行属性约简。

表 1 流感病人决策表

病人	头疼	肌肉疼	体温	流感
1	是	是	正常	否
2	是	是	高	是
3	是	是	很高	是
4	否	是	正常	否
5	否	否	高	否
6	否	是	很高	是
7	否	否	高	是
8	否	是	很高	否

(1)根据定义 1 将表 1 转化成列联表。

文中只展示所有条件属性与决策属性的列联表,其余类似。表 2 为转化后的列联表。

表 2 综合列联表

	综合属性(头疼/肌肉疼/体温)						流感		$\sum r_i$
	是/是/正常	是/是/高	是/是/很高	否/是/正常	否/否/高	否/是/很高	是	否	
1	1	0	0	0	0	0	0	1	2
2	0	1	0	0	0	0	1	0	2
3	0	0	1	0	0	0	1	0	2
4	0	0	0	1	0	0	0	1	2
5	0	0	0	0	1	0	0	1	2
6	0	0	0	0	0	1	1	0	2
7	0	0	0	0	1	0	1	0	2
8	0	0	0	0	0	1	0	1	2
$\sum c_j$	1	1	1	1	2	2	4	4	16

(2)分别计算决策属性流感与条件属性头疼、肌肉疼、体温的投影区分度以及与综合属性的区分度,得到单个识别度量表如表 3 所列。

表 3 单个识别度量表

	维数		区分度
	1	2	γ
流感	0.629	0.371	0.73026
头疼	0.629	0.371	
流感	0.259	0.741	0.48899
肌肉疼	0.741	0.259	
流感	0.789	0.000	0.48872
体温	0.789	1.000	
流感	0.854	0.000	0.55460
综合变量	0.854	1.00	

根据表 3 中各属性与决策属性区分度度量 γ 的大小,可知流感与头疼的投影区分度为 0.73026 与流感与综合属性的投影区分度 0.5546 最接近,因此将条件属性头疼加入到约简属性集中。将属性头疼分别与属性肌肉疼、体温进行组合,并按定义 1 转化成列联表进行对应分析,得到组合识别度量表如表 4 所列。

表 4 组合识别度量表

	维数		区分度
	1	2	γ
流感	0.644	0.000	0.34869
头疼+肌肉疼	0.644	1.000	
流感	0.854	0.000	0.55460
头疼+体温	0.854	1.000	

由表 4 可知,决策属性流感与条件属性头疼+体温组合的投影区分度最大,且和综合变量与流感的投影区分度相同。因此按照本文的算法,{头疼、体温}是该决策表的最优约简集,说明了该算法是行之有效的。

最后给出算法实现的简单示意图,如图 2 所示。

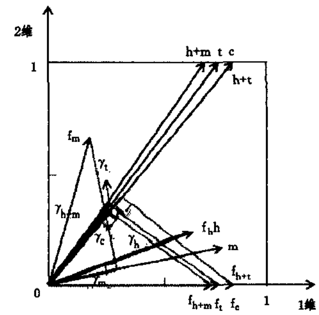

图 2 算法的实现过程

图 2 中,h、m、t 分别表示头疼、肌肉疼、体温 3 个条件属性的得分向量,c 表示 3 个属性的综合属性的得分向量,$h+m$、$h+t$表示头疼和肌肉疼的组合属性得分向量、头疼和体温的组合属性得分向量。f_h、f_m、f_t 分别表示决策属性流感在头疼、肌肉疼、体温 3 个条件属性下的得分向量,f_{h+m}、f_{h+t}、f_c 分别表示决策属性流感在头疼+肌肉疼组合属性、头疼+体温组合属性、综合属性下的得分向量。γ_h、γ_m、γ_t 分别表示 f_h、f_m、f_t 在 h、m、t 方向上的投影。γ_{h+m}、γ_{h+t}、γ_c 分别表示 f_{h+m}、f_{h+t}、f_c 在 $h+m$、$h+t$、c 方向上的投影。

单个条件属性与决策属性比较时各投影长度不同,决策属性在头疼属性上的投影 γ_h 最长,因此优先选择头疼属性作为约简集中的元素,在头疼+肌肉疼与头疼+体温两个组合中,γ_{h+t} 与 γ_c 的投影长度相同,按照本文的算法收敛条件,已经找到最优子集{头疼+体温}。

结束语 本文将统计学中的对应分析与粗糙集理论中的

决策表属性约简结合起来，提出一种定量度量决策属性与条件属性依赖关系的方法。该度量不仅能反映各条件属性与决策属性的相关关系，还能很好地识别最优约简子集，也可用来判断决策属性在条件属性上的偏好，在实际问题中有很好的应用前景。由于对应分析有很强的图形解释功能，下一步将研究如何结合实际问题解释决策属性与条件属性之间的依赖关系。

参考文献

[1] 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，2001：12-32

Zhang Wen-xiu, Wu Wei-zhi, Liang Ji-ye, et al. Rough set theory and method [M]. Beijing: Science Press, 2001: 12-32

[2] Hu X H, Cercone. Learning in relational database: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338

[3] Jelonek J. Rough set reduction of attributes and their domains for neural network[J]. Computational Intelligence, 1995, 11(2): 339-347

[4] 苗夺谦，李道国. 粗糙集理论、算法和应用[M]. 北京：清华大学出版社，2008：34-41

Miao Duo-qian, Li Guo-dao. Rough set Theory, Method and Applications[M]. Beijing: Tsinghua University Press, 2008: 34-41

[5] 王国胤，姚一豫，于洪. 粗糙集理论与应用研究综述[J]. 计算机学报，2009，32(7)：1229-1245

Wang Guo-yin, Yao Yi-yu, Yu Hong. A Survey on Rough set Theory and Applications [J]. Chinese Journal of Computers, 2009, 32(7): 1229-1245

[6] Mitra S, Pal S K, Mitra P. Data mining in soft computing framework: a survey [J]. IEEE Transactions on Neural Networks, 2002, 13(1): 3-14

[7] Estaji A A, Hooshmandscl M R, Davvaz B. Rough set theory applied to lattice theory[J]. Information Science, 2012, 200: 108-122

[8] Hegland M. Data mining techniques[M]. Cambridge University Press, 2001, 10: 313-355

[9] 韩丽丽. 决策粗糙集的属性约简算法研究[D]. 安徽：安徽大学，2013

Han Li-li. The Research of Attribute Reduction Algorithm in Decision Theoretic Rough Set [D]. Anhui: Anhui University, 2013

[10] Lingras P J, Yao Y Y. Data mining using extensions of the rough set model[J]. Journal of the American Society for Information Sciences, 1998, 49(5): 415-422

[11] Golan R, Ziarko W O. A methodelogy for stock market analysis utilizing Rough set theory[C]// Proc. of IEEE/IAFE Conference on Computation Intelligence for Financial Engineering. New Jersey, 1995: 32-40

[12] 刘宏杰，冯博琴，李文捷，等. 粗糙集属性约简判别分析方法以及应用[J]. 西安交通大学学报，2007，41(8)：939-943

Liu Hong-jie, Feng Bo-qin, Li Wen-jie, et al. Discrimination Method of Rough Set Attribute Reduction and Its Applications [J]. Journal of Xi'an Jiaotong University, 2007, 41(8): 939-943

[13] Greenacre M J. Theory and Applications of Correspondence Analysis[M]. London: Academic Press, 1984

[14] Greenacre M J. Correspondence Analysis of Square Asymmetric Matrices[J]. Applied Statistics, 2000, 49: 297-310

[15] Miao D Q. Analysis on Attribute Reduction Strategies of Rough Set[J]. Computer Science and Technology, 1998, 13(2): 189-194

[16] Yao Y Y, Zhao Yan. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178: 3356-3373

[17] Qian Yu-hua, Liang Ji-ye, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in rough set theory [J]. Artificial Intelligence, 2010, 174: 597-618

[18] 高慧旋. 应用多元统计分析[M]. 北京：北京大学出版社，2005：110-120

Gao Hui-xuan. Applications of Multivariate Statistical Analysis [M]. Beijing: Peking University Press, 2005: 110-120

[19] Johnson R A, Wichern D W. 实用多元统计分析(第六版)[M]. 陆璇，叶俊，译. 北京：清华大学出版社，2008：145-150

Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis(Sixth Edition)[M]. Lu Xuan, Ye Jun. Beijing: Tsinghua University Press, 2008: 145-150

[20] 陶凤梅. 对应分析的数学模型[D]. 吉林：吉林大学，2005

Tao Feng-mei. Mathematical Model of Correspondence[D]. Jilin: Jilin University, 2005

(上接第 17 页)

[11] 陈涛，应振根，申世飞，等. 相对速度影响下社会力模型的疏散模拟与分析[J]. 自然科学进展，2006，16(12)：1606-1612

Chen Tao, Ying Zhen-gen, Shen Shi-fei, et al. Evacuation simulation and analysis of the relative velocity under social force model [J]. Progress in Natural Science, 2006, 16(12): 1606-1612

[12] 叶青，夏时洪，毛天露，等. Agent-Based 群体模拟中的朝向计算方法[J]. 计算机辅助设计与图形学学报，2011，23(8)：1349-1356

Ye Qing, Xia Shi-hong, Mao Tian-lu, et al. Orientation Computing in Agent-Based Crowd Simulation [J]. Journal of Computer-Aided Design & Computer Graphics, 2011, 23(8): 1349-1356

[13] Mehdi M, Niriaska P, Simon G, et al. The Walking Behaviour of Pedestrian Social Groups and its Impact on Crowd Dynamics [J]. PLoS One(S1932-6203), 2010, 5(4): e10047

[14] Daamen W. Modelling passenger flows in public transport facilities [M]. Delft, Netherlands: DUP Science, 2004

[15] Sakuma T, Mukai T, Kuriyama S. Psychological model for animating crowded pedestrians [J]. Computer Animation and Virtual Worlds, 2005, 16(3/4): 343-351

[16] van Toll W G, Cook A F, Geraerts R. Real-time density-based crowd simulation [J]. Computer Animation and Virtual Worlds, 2012, 23(1): 59-69

[17] 赵欣欣，张勇，孔德慧，等. 基于场的人群运动仿真[J]. 中国图像图形学报，2013，18(3)：344-350

Zhao Xin-xin, Zhang Yong, Kong De-hui, et al. Field-based crowd simulation [J]. Journal of Image and Graphics, 2013, 18(3): 344-350

[18] Song Wei-guo, Yu Yan-fei, Wang Bing-hong, et al. Evacuation behaviors at exit in CA model with force essentials: A comparison with social force model[J]. Physica A: Statistical Mechanics and its Applications, 2006, 371(2): 658-666