决策表分析的统计依据

魏 玲[1,2] 张文修[2]

(西北大学数学系 西安710069)[1] (西安交通大学理学院 西安710049)[2]

摘 要 给出了决策表的条件属性约简的非参数统计检验方法。首先,给出与决策表相应的列联表,进行条件属性与决策属性间相关性的显著性检验,在一定的显著性水平上,依据相关性显著与否,来判别该属性相对于决策行为是否冗余,从而获得属性约简;进而,采用 Lambda 系数对与决策属性显著相关的属性进行相关性度量,说明用条件属性对决策属性进行预测将消减误差的比例。并在列联表的基础上,获得决策表的一级规则。病例决策表的实验表明,该方法简单,有效。

关键词 决策表,列联表,相关性度量,显著性检验,一级规则

Statistical Evidence for Decision System Analysis

WEI Ling[1,2] ZHANG Wen-Xiu[1]

(Faculty of Science,Xi'an Jiaotong University,Xi'an 710049)[1] (Department of Mathematics,Northwest University,Xi'an 710069)[2]

Abstract The nonparametric test method is introduced to the reduction of decision table,which gives the statistical evidence for decision system analysis. Firstly,we transfer the raw decision system into a series of corresponding contingency table between each condition attribute and decision attribute; then,test the correlation significance between them to make sure if a condition attribute is useless and should be deleted. Secondly,for those condition attributes that are correlated with decision attribute,we do correlated measure using Lambda coefficient. The set of condition attributes that are correlated with decision attribute is regarded to be the reduction of condition attributes,and the corresponding Lambda coefficient shows the proportionate reduction in error. At the same time,we get the first-order rules of the raw decision table based on the series of corresponding contingency table. Our test shows that our method is feasible and efficiently.

Keywords Decision table,Contingency table,Correlated measure,Significance test,First-order rule

知识获取是人工智能研究领域中最为活跃的一个分支,它对于专家系统、数据挖掘、机器学习等都起着极为重要的作用。

决策表(决策系统)是一类特殊而又重要的知识表达系统,它指当对象满足某些条件时,决策行为应当怎样进行。多数决策问题都可以用决策表形式来表达,因此决策表这一工具在决策应用中起着重要的作用。针对决策表,目前世界上的研究热点主要集中在属性约简,规则提取等方面。其目的是希望从由决策表所表达的知识系统中进行知识发现,以获取可以为人们所利用的规则和决策指导。

对决策表进行知识发现,如属性约简或规则提取的方法主要是粗糙集理论(Rough Sets)[1]。粗糙集理论是波兰数学家 Z. Pawlak 在1982年提出的一种分析数据的数学理论,主要用于知识的简化和知识依赖性的分析。无论是理论研究还是实际应用,有关粗糙集方面的文献颇多。尽管粗集理论现在发展迅速,理论也相对完善,但是,使用粗集理论对决策表进行分析却有一个明显的缺陷,就是缺乏统计依据[2]。我们知道,决策表中的对象是作为研究总体的样本来进行分析的,其分析结果如果没有统计检验作保证的话,就不能认为该分析结果反映了总体的信息。因此,我们提出了用统计检验进行决策表分析的方法。

本文以决策表的相应列联表为基础,从统计学角度给出了条件属性冗余与否的假设检验方法,基于列联表的一级规则获取方法,以及条件属性对决策属性的相关性度量;并与粗集约简结果进行了比较。

1 决策表的粗集约简

定义1 称(U,A,F)为一个信息系统或数据库系统。其中$U=\{x_1,\cdots,x_n\}$为对象集,每个$x_i(i\leqslant n)$称为一个对象;$A=\{a_1,\cdots,a_m\}$为属性集,每个$a_j(j\leqslant m)$称为一个属性;$F=\{f_j:j\leqslant m\}$为 U 和 A 的关系集,$f_j:U\to V_j(j\leqslant m)$,$V_j$ 为属性 a_j 的值域。

表1 决策表

U	a_1	…	a_j	…	a_m	d
x_1	f_{11}	…	f_{1j}	…	f_{1m}	g1
…			……		…	
x_i	f_{i1}	…	f_{ij}	…	f_{im}	g1
…			……	…		
x_n	f_{n1}	…	f_{nj}		… f_{nm}	g1

定义2 称(U,A,F,D,G)为决策表或决策系统,其中(U,A,F)是信息系统,此处,A 称为条件属性集,$D=\{d_1,d_2,\cdots,d_p\}$称为决策属性集;G 为 U 和 D 的关系集,即$G=\{g_j:j\leqslant p\}$,其中 $g_j:U\to V_j(j\leqslant p)$,$V_j$ 为决策属性 d_j 的值域。

表1给出了一个决策表(U,A,F,D,G)的样例,其中对象集$U=\{x_1,\cdots,x_n\}$,x_i 取值为 $F_i=(f_{i1},\cdots,f_{im})$,条件属性集$A=\{a_1,a_2,\cdots,a_m\}$,决策属性集 $D=\{d\}$,取值 g1。若将表中最后一列(决策属性)删除,则该表变成信息系统(U,A,F)。

在决策表 $DT=(U,A,F,D,G)$中,每一个对象是利用条

魏 玲 博士生,主要研究方向为人工智能;张文修 教授,博士生导师,主要研究方向为粗糙集,人工智能。

件属性 A 的属性值来描述的。然而，A 中的某些属性可能是冗余的，我们总可以找到 A 的一个极小子集 A_0，使得 $DT_{A0}(D)=DT_A(D)$，即用 A_0 来代替 A 不会丢失决策表的任何信息，从而得到一个简化的决策表。

在粗糙集理论中，决策表(U,A,F,D,G)的属性约简定义及方法如下。

定义3 决策属性 D 以 k 度依赖于条件属性 C，如果：

$$k=\gamma(C,D)=\sum_{X\in U/D}\frac{|\underline{R}(X)|}{|U|}$$

其中，$\underline{R}(X)=\{x\in U:C(x)\subseteq X\}$，称作 X 的 C-下近似。

定义4 如果对于 C 的最小子集 C'，成立 $\gamma(C,D)=\gamma(C',D)$则称 C'是 C 的 D-约简(即相对于决策属性集 D 的约简)。

依据粗集理论，我们还可以给出最简分类规则，即由一个属性就可以确定对象决策值的最简规则，后文称为一级规则。

定理1 如果一个条件属性 a 取值为 i 时关于决策属性 d 的某个取值 j 的下近似非空，且关于该决策属性的其他属性值的下近似皆为空，即$\underline{R}_a(D_j)=\{x\in U;[x]_{a=i}\subseteq D_j\}\neq\Phi$，而对于任意的 $h(h\neq j)$，$\underline{R}_a(D_h)=\{x\in U;[x]_{a=i}\subseteq D_h\}=\Phi$，则有一级规则：if $a=i$，then $d=j$。其中$[x]_{a=i}=\{y\in U;f_a(x)=f_a(y)=i\}$，$D_j=\{y\in U;g_d(y)=j\}$。

证明：由于条件属性 a 取值为 i 时关于决策属性 d 的某个取值 j 的下近似非空，且关于该决策属性的其他属性值的下近似皆为空，意即，条件属性 a 取值为 i 时，决策属性 d 只有唯一的取值 j，所以，显然有定理所描述的一级规则。

2 决策表分析的统计证据

本节，我们对决策表在其列联表的基础上，给出了条件属性冗余与否的非参数检验方法，以及一级规则获取方法，并给出用条件属性值去预测决策属性值时消减误差的比例。

2.1 假设检验基本理论

假设检验是近代统计学研究的中心课题之一。

根据对样本所属总体的性质假设，假设检验可分为参数检验和非参数检验。参数方法只能用于那些真正是数值的观察结果。许多非参数检验方法却只着眼于观察结果的顺序或等级，而不是它们的数值。还有一些非参数方法甚至可用于连编排顺序都做不到的观察结果，即分类资料。

事实上，我们要研究的决策表就是一个分类资料。这也是我们在决策表分析中引入非参数方法的理由。

表2 列联表

	A_1	A_2	…	A_n	Sum
B_1	x_{11}	x_{12}	…	x_{1n}	β_1
B_2	x_{21}	x_{22}	…	x_{2n}	β_2
…	…	…	…		
B_m	x_{m1}	x_{m2}	…	x_{mn}	β_m
Sum	α_1	α_2	…	α_n	$Sum=N$

2.2 列联表应用的基本理论

来自某一个总体的样本，同时按照两个或两个以上的标准进行分类，叫做交互分类。交互分类的资料可以排成一个行列交织的统计表，称为列联表。

列联表样例表2中的 x_{ij} 表示在样本容量为 N 的一个样本中，按照特征 $A_j(j=1,\cdots,n)$和 $B_i(i=1,\cdots,m)$分类的样本观察个数。列联表可以清楚地反映在 A 变量条件下，B 的次数分布情况。

随机样本中两个变量在总体中是否相关，需要通过显著性检验才能解决。而对于列联表给出的定类变量间是否独立的检验，卡方检验是最有效的。

定理2[3] 当列联表以表2的形式给出时，检验统计量

$$\chi^2=\sum_{i,j=1}^{m,n}\frac{(x_{ij}-\alpha_j\beta_i/N)^2}{\alpha_j\beta_i/N}\qquad(1)$$

服从自由度为$(m-1)(n-1)$的 χ^2分布。

任给一个 $m\times n$ 列联表，其自由度为 $df=(m-1)(n-1)$，通过确定 χ^2的观察值在 H_0(某两个特征在总体中无关)成立时出现的相伴概率 p，我们就可以确定相关度量的显著性。

两个特征在总体中若是相关的，就可以进一步应用 PRE (Proportionate Reduction in Error)测量方法考察以 A 对 B 进行预测将消减多大比例的误差。PRE 测量方法主要有 Lambda 相关测量法，Goodman-Kruskal Tau 相关测量法，Gamma 相关测量法，Somer's d 相关测量法等[4]。由于决策表是分类资料，我们选择用于定类变量间测量的 Lambda 相关测量法。Lambda 系数的计算公式如下。

$$\lambda_{BA}=\frac{\sum m_B-M_B}{N-M_B}\qquad(2)$$

其中，N 是样本容量，M_B 是 B 变量的众数，m_B 是各列中 B 变量的众数，其结果的意义是 B 受 A 影响的程度，即以 A 对 B 进行解释或预测时减少的误差。由于 Lambda 系数的取值属于[0,1]，所以其数值大小也体现了两个特征相关程度的大小。

2.3 决策表统计分析

表3 决策表的 a_l-d 列联表

	$a_l=1$	$a_l=2$	…	$a_l=\|V_l\|$	Sum
$d=1$	x_{11}	x_{12}	…	$x_{1,\|V_l\|}$	β_1
$d=2$	x_{21}	x_{22}	…	$x_{2,\|V_l\|}$	β_2
…		…	…	…	
$d=\|V_d\|$	$x_{\|V_d\|,1}$	$x_{\|V_d\|,2}$	…	$x_{\|V_d\|,\|V_l\|}$	β_m
Sum	α_1	α_2	…	α_n	N

对任意一个如表1所示的决策表(U,A,F,D,G)我们都可以写出每个条件属性和决策属性间的列联表，如表3所示。其中，$N=|U|$，而对某个 $a_l\in A$：

$\alpha_i=\|\{x_k:f_l(x_k)=i\}\|$，$(i\leqslant|V_l|)$，

$\beta_j=\|\{x_k:g(x_k)=j\}\|$，$(j\leqslant|V_d|)$，

$x_{ji}=\|\{x_k:f_l(x_k)=i,g(x_k)=j\}\|$，$(i\leqslant|V_l|,j\leqslant|V_d|)$。

对 a_l-d 列联表，用上述卡方检验方法来检验条件属性 a_l 和决策属性 d 间的相关程度。如果相关性不显著，则认为该属性是冗余属性，应该剔除；相关性显著的条件属性集合认为是条件属性相对于决策属性的约简。对于属性约简的结果，再用 Lambda 系数给出每一个条件属性与决策属性相关的程度，以及用该条件属性对决策属性进行预测时将消减多大比例的误差。

特别地，我们还可以由列联表的某一列中非零数字与零数字的出现情况来确定决策表的一级规则，由此实现了决策表部分知识的发现。通过对 a_l-d 列联表分析，不难得出以下结论。

定理3 对决策表(U,A,F,D,G)建立每个条件属性 a_l 与决策属性 d 间的列联表。在 a_l-d 列联表中，如果对于属性 a_l 的固定取值 i $(i=1,\cdots,|V_l|)$，只有一个非零的 x_{ij}，则有一级规则：if $a_l=i$，then $d=j$。

事实上，这一结论与第2小节中关于一级规则的描述结果相同，即它与粗集理论中决策类相对于各条件属性的下近似集的计算是等价的。由此，我们将决策系统的列联表分析方法

与粗糙集的分析方法联系起来。并且，由于列联表的计算与描述要比粗集理论中的关于下近似的计算与描述简单得多，因此，用列联表分析决策系统有简单，易掌握的优势。

3 实验

考察一个关于肺炎和肺结核两种疾病的病例数据库表4[5]。

该病例数据库共有20个病例，每个病例有4种症状：a—发烧、b—咳嗽、c—X光阴影、d—听诊，即条件属性集$A=\{a,b,c,d\}$；1个决策：e—诊断结果，即决策属性集$D=\{e\}$。

属性值$V_a=\{1,2,3,4\}$，其中1为不发烧，2为低烧，3为中度发烧，4为高烧。$V_b=\{1,2,3\}$，其中1为轻微咳嗽，2为中度咳嗽，3为剧烈咳嗽。$V_c=\{1,2,3,4\}$，其中1为片状，2为点状，3为索条状，4为空洞。$V_d=\{1,2,3\}$，其中1为正常，2为干鸣音，3为水泡音。$V_e=\{1,2\}$，其中1为肺炎，2为肺结核。

表4 病例信息表

	U																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
v_a	4	3	1	3	4	2	4	3	3	4	3	2	1	3	4	4	3	1	2	4
v_b	3	3	1	1	3	1	2	1	2	3	1	2	2	2	2	3	1	3	1	3
v_c	1	1	3	2	4	3	1	1	1	2	2	1	3	2	2	2	2	2	4	3
v_d	3	3	1	1	2	1	3	3	3	1	3	3	1	1	3	3	2	1	1	2
v_e	1	1	2	1	2	2	1	1	1	1	2	2	2	1	1	1	2	2	2	2

注：v_a、v_b、v_c、v_d、v_e分别表示V_a、V_b、V_c、V_d、V_e中的元素。

3.1 基于粗糙集理论的约简

利用第2节介绍的粗糙集约简方法，可以得到病例信息表的条件属性集A相对于决策属性e的约简为$\{a,c,d\}$，获取的全部规则如表5所示。其中，一级规则共5个，分别为第2，4，5，9，12号规则。

表5 病例数据库的规则

	1	2	3	4	5	6	7	8	9	10	11	12
v_a	3	2	4	*	*	3	4	4	1	3	4	*
v_c	*	*	*	*	3	1	2	*	*	2	1	4
v_d	1	*	3	2	*	*	*	1	*	3	*	*
v_e	1	2	1	2	2	1	1	1	2	2	1	2

3.2 基于属性相关显著性检验的约简

我们以条件属性a与决策属性e间的相关显著性检验为例，说明检验方法与步骤。其他类似。

条件属性a与决策属性e间的列联表如表6所示。

表6 a-e列联表

	$a=1$	$a=2$	$a=3$	$a=4$	Sum
$e=1$	0	0	5	5	10
$e=2$	3	3	2	2	10
Sum	3	3	7	7	20

建立假设组为：原假设H_0：a与e无关；

备择假设H_1：a与e相关。

利用式(1)，可以计算出$\chi^2=8.57$。此处，自由度$df=(2-1)(4-1)=3$。给定显著性水平$\alpha=0.05$，由卡方分布临界值表得，$\chi^2_\alpha(3)=7.82$。由于$\chi^2>\chi^2_\alpha(3)$，表明在5%的显著性水平上，拒绝H_0，即条件属性a与决策属性e存在相关。

在条件属性a与决策属性e相关的基础上，我们用式(2)计算出，$\lambda_a=0.6$。这一结果表明，属性a(发烧)与决策e(诊断结果)间的相关程度是较高的，用“发烧”去解释或预测“诊断结果”，可以减少60%的误差。

采用类似的方法，我们分别可以得到条件属性b,c,d与决策属性e间的列联表，在5%的显著性水平上，条件属性b与决策属性e不相关；条件属性c,d分别与决策属性e存在相关，相应的Lambda系数分别为$\lambda_c=0.6$，$\lambda_d=0.5$，即用属性c(X光阴影)和d(听诊)去解释或预测决策e(诊断结果)，分别可以减少60%和50%的误差。

3.3 基于列联表的一级规则获取

在获得的条件属性与决策属性的4个列联表上考察一级规则。利用定理2描述的一级规则获取方法知：

a-e列联表提供两个一级规则：if $a=1$，then $e=2$；if $a=2$，then $e=2$；

b-e列联表不提供一级规则。事实上，b相对于e来说是冗余属性，所以该列联表不提供任何规则；

c-e列联表提供两个一级规则：if $c=3$，then $e=2$；if $c=4$，then $e=2$；

d-e列联表提供一个一级规则：if $d=2$，then $e=2$。

而这5个规则恰恰就是粗糙集约简结果中的一级规则。

鉴于上述分析求解过程，我们得出结论：对肺炎和肺结核两种疾病的病例数据库表4，在其相应的列联表基础上应用非参数检验方法，可以认为该病例决策表的条件属性约简是$\{a,c,d\}$(在5%的显著性水平上)；同时获得了5个很简单的一级规则。这些结果与用粗糙集方法得到的结果完全相同。

结束语 决策表的数据本质上是定类数据，其资料中的数字只是用来表示事物所属的类别，而不能进行算术运算。我们引入用于定类数据分析的非参数统计检验方法，将其用于决策表的属性相关分析，获取属性间的相关关系以及相关程度，并且发现了决策表的最简规则。这一方法把统计理论引入到决策系统的分析当中，用统计检验的思想描述属性间的相关性，发现决策表的部分知识，更具有理论与实际意义。我们的实验表明，该想法是可行的，而且效果也较好，与传统的决策表分析方法——粗集理论得到的结论是一致的。不过，单纯分析每个条件属性与决策属性的关系其局限性太大。找到合适的方法分析条件属性的综合作用对决策属性的影响，将是非常有意义的工作。

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publishers, 1991

2 Tsumoto S. Statistical Evidence for Rough Set Analysis[A]. In: Proc. of the 2002 IEEE Intl. Conf. on Fuzzy Systems. FUZZ-IEEE'02, 1: 757～762

3 Rao C R. Linear Statistical Inference and Its Applications[M]. 2nd Edition, New York: John Wiley & Sons, 1973

4 易丹辉. 非参数统计——方法与应用[M]. 中国统计出版社, 1996

5 印勇，曹长修，张邦礼. 基于粗糙集理论的分类规则发现[J]. 重庆大学学报(自然科学版), 2000, 23(1): 63～65