以随机事件为证据的一种新推理模型*)

李冠英　游荣彦（华南师范大学计算机科学系）

摘　要

本文在假定推理规则中，仅证据为概率空间的随机事件，而结论可相当自由地表现为定性或定量；文中合理地解决了证据间的相似程度的刻画问题并建立由新证据推出结论的可信度的计算方法；最后举例说明本推理模型的意义。

一、问题的提出

专家系统中常用的主观 Bayes 方法实际上要求所建立的一切推理规则

IF　E　THEN　C　　(1)

中的条件E及结论C都同属某一给定的概率空间，均为随机事件。在不少实际问题中，我们都是面对仅证据为某一概率空间的随机事件，而结论却不然的情形，此时若再套用主观 Bayes 方法，则显得牵强且根据不足；若用证据理论方法[1]或发生率计算方法[2]，则许多关于概率空间的知识未能得到有效的应用而被浪费。

针对这个左右为难的问题，本文提出一种尽量运用现有知识的一种新的推理模型。在本模型中，一切证据均为随机事件，而结论则不然，可以是某一数量，也可以是某个定性描述的性状。因为结论不是随机事件，故此（1）不再是似然规则，我们只称之为规则。

在 PROSPECTOR 系统中，规则强度用似然比$LS=P(E|C)/P(E|\overline{C})$作描述[3]，它可取值于区间$[0, +\infty)$，然而在本模型中，结论$C$及其逆$\overline{C}$已非随机事件，无法沿用$LS$。解决规则（1）的规则强度，有两条途径，一是由专家赋值，另一是试验定值。设有足够次数的试验作为依据算出：平均每n次出现证据E运用规则（1），有n_0次成功，而有$n-n_0\neq 0$次不成功。循似然比定义的思路，确定规则（1）的规则强度为

$$t=\frac{n_0}{n-n_0} \qquad (2)$$

若证据E与结论C有某种规律支配下的相依性，则当n增大时，(2)式的值将会有某种意义的收敛；否则，（2）式将会呈现无规律的变化。

若记

$$S=\frac{n_0}{n} \qquad (3)$$

则可得到

$$t=\frac{S}{1-S} \quad 或 \quad S=\frac{t}{1+t}$$

由此可知，S与t单值依存，所以由（3）式确定的S值，亦可以作为规则（1）的规则强度，用t或S描述规则强度，在本质上并无二致，表面上的差异来源于度量的尺度，使用(2)，相当于用"无穷大的尺"；使用（3），相当于使用"最大刻度为1"的尺，由于S的规范性能带来与其他学科联系的方便，在此推荐使用（3）式确定的S来描述规则强度。

二、证据的相似度

设所讨论的证据属于概率空间Ω，任一随机事件A的概率值为$P(A)$，换言之，设我

*) 广东省高等教育局资助项目

们已经掌握了整个证据所在的概率空间的三要素：Ω，一切证据的集合以及每一证据作为随机事件的概率。

若$P(A)=0$，则A为不可能事件，以之作为证据是没有意义的，因此约定，凡讨论的证据，均有大于零的概率值。对于证据A和证据B，规定它们的相似度为

$$r(A,\ B)=\frac{2P(A\cap B)}{P(A)+P(B)} \qquad (4)$$

不难证明以下事实：①$0\leqslant r(A,\ B)\leqslant 1$；②$r(A,\ B)=r(B,\ A)$；③当且仅当$A\cap B=\phi$时，$r(A,B)=0$；④当且仅当$A=B$时，$r(A,B)=1$。

证据相似度的前两点性质可分别称为规范性和对称性，第③点性质可称为明辨性，第④点性质表示只有与自身相比，证据才达到最大相似度1。这说明相似度的定义是合理的。

三、新证据的推理问题

设有一规则集，由n条异证据同结论的规则组成，第i条规则$R_i(i=1,\ \cdots,\ n)$是：

IF E_i THEN C WITH S_i (5)

其中S_i为规则强度，如前所述，它具有规范性，即$0\leqslant S_i\leqslant 1$。我们假定，当$i\neq j$，$E_i\cap E_j=\phi$，即证据$E_i$与$E_j$互斥，由上节可知，它们之间的相似度等于零。以下称规则(5)中的证据$E_1,\ E_2,\cdots,E_n$为基本证据。

设E为一新证据。若至少有一个基本证据E_i，使$E_i\cap E\neq\phi$，此时称E为有效新证据。若$E\cap(\bigcup_{i=1}^{n}E_i)=\phi$，称新证据对规则集(5)是无效的，出于慎重，此时不应作任何推理。

当新证据E是有效的，则至少有一个E_i，使$r(E,\ E_i)>0$。于是对有效证据E，可定义权

$$W_i=\frac{r(E,\ E_i)}{\sum_{i=1}^{n}r(E,\ E_i)},\quad i=1,\ \cdots,n \qquad (6)$$

它反映出新证据E在n个基本证据中，与E_i的相对接近程度。显然，诸W_i满足权的基本性质：$0\leqslant W_i\leqslant 1$；$\sum_{i=1}^{n}W_i=1$。

基于规则集(5)，由有效新证据推出结论C的可信度，借用Shortliffe的符号而记为$CF(E)$，它的计算，应满足下面的设计原则：

第一，$0\leqslant CF(E)\leqslant \mathrm{Max}\{S_1,\ \cdots,\ S_n\}$，这样，体现了对规则集(5)即专家知识的根本确认，任何新证据的出现，都不能强化(5)的规则强度的最大值，任何新证据都不能更胜于基本证据。

第二，由新证据E推出结论C的可信度值是新证据E与n个基本证据的相似度$r(E,E_i)$，以及n条规则的规则强度S_i的函数：

$$CF(E)=f(r(E,\ E_1),\ \cdots,\ r(E,\ E_n);\ S_1,\ \cdots,\ S_n)$$

并且此函数必须对每一个变元都是单调增加的。

现规定

$$CF(E)=\sum_{i=1}^{n}W_iS_i \qquad (7)$$

由权W_i的性质，不难验证，$CF(E)$满足上述的两条原则。

例 设某地年降雨量$X\sim N(2,000,\ 100^2)$，(单位：毫米)，已由知识工程师和专家建立了五条规则：

R_i：IF E_i THEN C WITH S_i，$i=1,\ \cdots,\ 5$.

其中C为"作物Y丰收"——一个非数量的定性结论，五个基本证据依次为如下的随机事件：

$E_1=(1,400\leqslant X<1,600)$,

$E_2=(1,600\leqslant X<1,800)$,

$E_3=(1,800\leqslant X<2,000)$,

$E_4=(2,000\leqslant X<2,200)$,

$E_5=(2,200\leqslant X<2,400)$.

五条规则的强度依次为：$S_1=0.3$，$S_2=0.5$，$S_3=0.8$，$S_4=0.7$，$S_5=0.4$.

今预测明年的降雨量在1,900～2,200毫米之间，以此作为新证据E，求结论C成立的可信度值$CF(E)$

因为$E_1\cap E=E_2\cap E=E_5\cap E=\phi$，故知$r(E,$

$E_1)=r(E,\ E_2)=r(E,\ E_5)=0$，再由(6)式即得$W_1=W_2=W_5=0$，以下求$W_3$和$W_4$。

降雨量X服从于以2,000为期望，100^2为方差的正态分布，于是它的分布密度函数是

$$g(x)=\frac{1}{100\sqrt{2\pi}}\exp\{-\frac{1}{2}\cdot\frac{(x-2000)^2}{100^2}\}$$

以$\Phi(x)$记标准正态随机变量的分布密度函数在区间$(-\infty,\ x)$上的积分值，则有[4]

$$\int_a^b g(t)dt=\Phi\left(\frac{b-2000}{100}\right)-\Phi\left(\frac{a-2000}{100}\right)$$

并且对任意$\delta>0$，$\Phi(-\delta)=1-\Phi(\delta)$，于是可算出

$$P(E_3)=\int_{1900}^{2000}g(x)dx=\Phi(0)-\Phi(-2)$$
$$=\Phi(0)+\Phi(2)-1=0.477$$

类似地可算出

$$P(E_4)=0.477,\ P(E)=0.819$$

再只须注意$E\cap E_3=(1,900\leqslant X\leqslant 2,200)\cap(1,800\leqslant X\leqslant 2,000)=(1,900\leqslant X\leqslant 2,000)$以及$E\cap E_4=E_4$，即可算出$P(E\cap E_3)=0.341$，$P(E\cap E_4)=0.477$。根据(4)算出相似度

$$r(E,\ E_3)=\frac{2\times 0.341}{0.819+0.477}$$

$$r(E,\ E_4)=\frac{2\times 0.477}{0.819+0.477}$$

根据(6)，可算出以下两个非零权数：

$$W_3=\frac{r(E,\ E_3)}{r(E,\ E_3)+r(E,\ E_4)}=0.42$$

$$W_4=\frac{r(E,\ E_4)}{r(E,\ E_3)+r(E,\ E_4)}=0.58$$

按(7)式，以新证据E推出结论C的可信度值

$$CF(E)=\sum_{i=1}^{5}W_iS_i=W_3S_3+W_4S_4$$
$$=0.42\times 0.8+0.59\times 0.7=0.742$$

四、多维证据的推理

若要同时考虑多种不同性质的证据，就涉及到多维证据的推理问题。一个多维证据可视作若干个一维证据的直积，或以各个一维证据为分量的多维向量。多维证据的推理比上一节的一维证据的推理更常见，不过所遇到的困难却不少，其中最主要的是不同性质的证据之间有些独立，有些存在各种形式的联系[5]，即使把证据看成多维随机变量的表现，也难以弄清其联合分布。以下讨论解决这一困难的方法。

设不同性质的证据共有m个，第i种性质的证据$E_i=(a_i\leqslant X_i<b_i)$表现为随机变量$X_i$取值于某区间$[a_i,\ b_i)$的随机事件，$i=1,\ \cdots,\ m$。在许多实际问题中，诸$X_i$的联合分布不可知，在此仅假设知道边缘分布。又设已根据专业知识或专家经验，对m种性质的不同证据对应的随机变量作了如下的分组：共分成k个组，在同一组里，随机变量存在相关关系，而不同组的随机变量独立。我们认为，在同一组里各不同性质的证据，在某种规律的支配之下，互相关联，各从不同的侧面同时支持同一结论的成立，例如白血球数高和体温升高，都支持"有炎症"的结论，不过这两方面的证据，同受一个规则——身体对入侵细菌的反应——所支配；不同组的证据互相独立，互相独立的证据合在一起，会更有力地支持结论，例如体温升高与皮肤有轻度破损是独立的（两者无因果关系），但若两者作为证据同时出现，则会更强烈地支持"有炎症"的结论。

设有关于m方面的证据导出同一结论C的规则，第i条规则是$R_i(i=1,\ \cdots,\ m)$：

IF E_i THEN C WITH S_i

又设我们获得m个新证据$E_1',E_2',\cdots,E_m'$，它们构成m维新证据$E'=(E_1',E_2',\ \cdots,E_m')$，根据上面的分析，由$m$维新证据$E'$推出结论$C$的可信度值$CF(E')$按下面的步骤计算：

第一，把m个新证据分成k个组，组的数目$k<m$只能依实际情况而定，各组所含的证据个数也只能由实际情况而定，基本的原则是：同一组内的证据不独立，不同组的证据互相独立；

第二，在每个组内，运用上节的方法，使用对应的规则，分别求出由各组证据推出结论C的可信度值CF_j，$(j=1,\ \cdots,\ k)$；

第三，基于有k方面独立新证据支持结论C的事实，推广文献[6]关于证据积累的公式，

通用产生式系统语言模糊化扩充研究

康建初 （北京航空航天大学计算机系）

摘 要

Forward chaining production system, as a kind of computational model of knowledge representation,has been used to implement some of the most significant expert systems.However unfortunately,most of forward chaining production system languages make no provision for dealing with lexical imprecision. This paper briefly presents a fuzzy production system language model, which supports fuzzy matching between condition patterns of productions and facts in working memory, and its discussion is also focussed on the part played by the truth-values of fuzzy production instantiations in the conflict resolution of the language.

作为一种有用的知识表示形式，正向产生式系统结构已被人们广泛地用以建造各种专家系统。不过，大多数通用的正向产生式系统语言，如OPS5、YAPS等均未提供处理不精确语词（即模糊语词）的机制。另一方面，虽然一些有关模糊推理的实用模糊系统亦称为模糊产生式系统，那不过是因为它是数据驱动的。单从推理形式这一点看，它们被看成了正向产生式系统。然而，就其控制机制而言，模糊产生式系统与正向产生式系统有着很大的区别。

正向产生式系统的控制机制有这样的特征：在合一匹配阶段，一般有若干条规则的条件部份同时为事实库描述的外部世界的当前状态所满足，由此形成一个触发规则实例冲突集。然而，在系统的执

把由m维新证据E'推出结论C的可信度值确定为

$$CF(E')=\sum_{j=1}^{k}CF_j-\sum_{1\le j<l\le k}CF_j\cdot CF_l+\cdots+(-1)^{k-1}CF_1\cdot\cdots\cdot CF_k \qquad (8)$$

例 讨论位于某河流下游的城市八月份供水问题。证据有三方面：上游七月份的流量、下游八月份的气温和下游流域的降雨量。若气温高，则降雨量低，于是后两方面的证据不独立，应属于同一组。第一方面的证据独立于后两方面的证据，单独成一组。讨论的结论C是该市八月份缺水。

设由第一方面的证据推出C成立的可信度值$CF_1=0.6$，又设按上节的方法，第二，第三方面的证据推出结论C成立的可信度值$CF_2=0.5$

根据(8)，由三维新证据E'推出结论C的可信度值

$$CF(E')=CF_1+CF_2-CF_1\cdot CF_2$$
$$=0.6+0.5-0.6\times0.5=0.8$$

参考文献

〔1〕 Shafer, G., A Mathematical Theory of Evidence, Princeton University Press, 1976.

〔2〕 L. 约翰逊，专家系统技术指南，11.3.4，世界图书出版公司，1989.7.

〔3〕 Duda, R.O., Hart, P. E., Subjective Bayesian Methods for Rule-Based Inference System, AFIPS, 1976, 1075-1082.

〔4〕 魏宗舒等，概率论与数理统计教程，高等教育出版社，1986，117—118.

〔5〕 姚玉川等，知识系统，大连理工大学出版社，1988，194—201.

〔6〕 R. 福西斯等，专家系统原理和实例研究，中国铁道出版社，1989，64—66.