文章编号: 1000－8934(2018)12－0023－06
DOI:10.19484/j.cnki.1000-8934.2018.12.005

卡尔纳普式的归纳逻辑的局限与所罗门诺夫先验的优势

李 熙
(中南大学 哲学系,长沙 410083)

摘要: 20世纪50年代,卡尔纳普发展了归纳逻辑来表示证据相对于假设的"确证度"。随后,利斯塔、古德－图灵等人提出了各种平滑方法,这些平滑方法可以看作广义的卡尔纳普式的归纳逻辑。这些方法虽然都可以从某个层次的"无差别原则"导出,但这并不能构成其理论基础。"无差别原则"无论作用在这里的哪一层都不合适,根据机器学习领域的无免费午餐定理,都不具有通用性,只有作用在可能世界的产生方式这一层次上导出的所罗门诺夫先验才具有通用性,能够逼近任何可计算的模式。而且,不仅如此,在同时满足奥卡姆剃刀原则和最大熵原则的意义上,所罗门诺夫先验具有最优性。

关键词: 归纳逻辑; 古德－图灵估计; 无免费午餐定理; 所罗门诺夫先验
中图分类号: N031 **文献标识码:** A

19世纪末20世纪初,演绎逻辑日趋完善,在演绎逻辑成功形式化的刺激下,凯恩斯、卡尔纳普[1]等人发展了归纳逻辑。随后,很多人对卡尔纳普发展的归纳逻辑提出了批评和改进。辛提卡[2](J. Hintikka),扎贝尔[3](S. Zabell)等人从不能确证诸如"所有乌鸦都是黑的"这种全称命题的角度对卡尔纳普提出了批评并给出了自己的修正。这些修正都是基于卡尔纳普框架的微小变种,仍可以看作卡尔纳普式的归纳逻辑。利斯塔[4](E. Ristad)估计、古德－图灵[5]估计、胡特尔[6](M. Hutter)拟古德－图灵估计等平滑方法可以看作卡尔纳普归纳逻辑在序列预测、自然语言处理等领域的推广,只是"无差别原则"作用的层次不同。普特南[7]从通用学习机的角度对卡尔纳普进行了批评。受卡尔纳普的归纳逻辑的启发,所罗门诺夫[8-9](R. Solomonoff)给出了贝叶斯通用归纳模型。拉斯马纳、胡特尔[10](Rathmanner & Hutter)对所罗门诺夫的通用归纳模型做了较系统的哲学分析,但对于所罗门诺夫先验是否具有最优性没有给出解答。李熙[11]指出,通过把所罗门诺夫先验引入归纳逻辑,不仅可以进行通用归纳,还可以只关注某种具体的模式而忽略其他无关信息并证明收敛定理,甚至可以不用记录所有的相关信息而采用"随机采样"的方法确立合理的信念。

在人工智能诞生之前,归纳逻辑学家就试图在找寻某种证据对假设的合理的信念确证方式,这与人工智能领域关心的不确定性推理有直接关系,但卡尔纳普等人的工作在人工智能领域并没有太大影响。本文把卡尔纳普"确证度"函数与利斯塔估计、古德－图灵估计、拟古德－图灵估计、线性插值平滑方法等放在一个平台上讨论而不是孤立起来研究传统的归纳逻辑。

下面首先以卡尔纳普的归纳逻辑为框架比较这几种平滑方法,然后从机器学习的角度讨论这些方法的不足和局限,最后论证所罗门诺夫先验不仅可以克服这些不足,而且在相对于奥卡姆剃刀和最大熵原则的意义上具有最优性。

收稿日期: 2018－08－17
基金项目: 国家社科基金项目"通用人工智能的哲学基础研究"(17CZX020);国家社科基金重大招标项目"现代归纳逻辑的新发展、理论前沿与应用研究"(2015ZDB018)。
作者简介: 李熙(1985—),山东日照人,哲学博士,中南大学讲师,主要研究方向:通用人工智能、数理逻辑。

一、卡尔纳普式的归纳逻辑
——“无差别原则”应该作用在何处?

记号。字母表 x 上长度为 n 的序列 $x_1 \ldots x_n$ 简记为 $x_{1:n}$。函数 $f: x^n \to \mathbb{R}$ 相对于概率分布 P 的期望记为 $\mathbb{E}_P[f]$。【】为艾佛森括号。

假设一阶逻辑语言 $\mathcal{L}$ 包含一集 $\mathcal{C}$ 可数无穷多的常项符号和 m 个一元谓词符号 $\mathcal{R}=\{R_1, R_2, \cdots, R_m\}$,但没有任何函数符号,也不包含等词。$\mathcal{L}$ 上的句子集记为 $\mathcal{S}$。令 $Q_i(x) \equiv \bigwedge_{j=1}^{m} \pm R_j(x)$,其中 $1 \leq i \leq 2^m =: r$,$\pm R$ 指 $\{R, \neg R\}$,则 $\mathcal{Q}=\{Q_1, \cdots, Q_r\}$ 构成了某个以 $\mathcal{C}$ 为个体集的世界的 r 划分。世界中的任何一个个体都满足某个 Q 谓词,而哪个个体满足哪个谓词由状态描述函数 $h: \mathcal{C} \to \mathcal{Q}$ 决定。句子集上的概率赋值 $w: \mathcal{S} \to [0,1]$ 需要满足:

(P_1). $\vdash \psi \Rightarrow w(\psi)=1$

(P_2). $\psi_1 \vdash \neg \psi_2 \Rightarrow w(\psi_1 \vee \psi_2) = w(\psi_1) + w(\psi_2)$

(P_3). $w(\exists x \psi(x)) = \lim_{n\to\infty} w\left(\bigvee_{i=1}^{n} \psi(a_i)\right)$

对于无量词句子集上的 $w: \mathcal{S} \to [0,1]$,如果它满足 P_1, P_2,那么它有唯一扩张满足 P_1, P_2, P_3。除了要满足 P_1, P_2, P_3,如何只借助语法特征构造证据对假设的信念确证方式?这是卡尔纳普式的归纳逻辑追问的问题。

对个体序列 $a=(a_1, \cdots, a_n)$,状态描述 $\Theta(a) := \bigwedge_{i=1}^{n} Q_{h_i}(a_i)$。在 n 次试验 $\bigwedge_{i=1}^{n} Q_{h_i}(a_i)$ 中,事件 Q_i 发生的次数为 $n_i := \sum_{j=1}^{n} [\![h(j)=i]\!]$。卡尔纳普定义 $n := (n_1, n_2, \cdots, n_r)$ 为结构描述。类似的,定义“等级描述” $m := (m_0, m_1, \cdots, m_n)$,其中 $m_i := |\{j: n_j = i\}|$ 是 n 次试验中出现 i 次的事件种类数。

卡尔纳普相信合适的 w 必须满足某种对称性原则。比如他希望 w 尽量满足如下原则。

(1) w 对于个体常元的任意置换都保持不变(Ex),即,对于 $\mathbb{N}^+$ 上的置换 σ,$w(\psi(a_1, \cdots, a_n)) = w(\psi(a_{\sigma(1)}, \cdots, a_{\sigma(n)}))$。

(2) w 对于谓词 Q 的任意置换都保持不变(Ax),即,对于 $\{1,2,\cdots,r\}$ 上的置换 τ,$w\left(\bigwedge_{i=1}^{n} Q_{h_i}(a_i)\right) = w\left(\bigwedge_{i=1}^{n} Q_{\tau(h_i)}(a_i)\right)$。

(3) 充分性假设(SP),存在一系列函数 $\{f_i: 1 \leq i \leq r\}$ 使得 $w(Q_j(a_{n+1}) \mid \Theta(a)) = f_j(n_j, n)$。

对称性原则 Ex 说的是 $w\left(\bigwedge_{i=1}^{n} Q_{hi}(a_i)\right)$ 只依赖于 $\langle n_{hi}: 1 \leq i \leq n \rangle$,所以它与观察个体的次序无关。而有了原则 Ex,对称性原则 Ax 说的是 $w\left(\bigwedge_{i=1}^{n} Qh_i(a_i)\right)$ 只依赖于 $\{n_i: 1 \leq i \leq r\}$,所以对任意的 $1 \leq i \leq r$ 有 $w(Q_i(a_1)) = 1/r$。这些原则的背后揭示的是无差别原则的思想,但无差别原则应该作用在哪层?

(A) 所有的状态描述分享相同的权重。

(B) 所有的结构描述分享相同的权重。

(C) r 个可能事件的任何非空子集分享相同的权重。

(D) r 个可能事件的具有不同基数的子集簇分享相同的权重。

(E) 所有的“等级描述”分享相同的权重。

1. “无差别原则”应该作用在哪层?

给定 n 个个体常项,有 r^n 个状态描述;$|\{(n_1, \cdots, n_r): \sum_{i=1}^{r} n_i = n\}| = \binom{n+r-1}{r-1}$ 个结构描述;$p(n,r) := |\{(m_0, \cdots, m_n): \sum_{i=1}^{n} i \cdot m_i = n \,\&\, \sum_{i=0}^{n} m_i = r \,\&\, \forall i: m_i \geq 0\}|$ 个“等级描述”。根据(A),$m^\dagger(\Theta(a)) = 1/r^n$,$c^\dagger(Q_j(a_{n+1}) \mid \Theta(a)) = 1/r$。根据(B),$m^*(n_1, \cdots, n_r) = 1/\binom{n+r-1}{r-1}$,因为每个结构描述 $(n_1, \cdots, n_r)$ 对应 $\binom{n}{n_1, \cdots, n_r}$ 个可能的状态描述,根据 Ex,$m^*(\Theta(a)) = m^*(n_1, \cdots, n_r) / \binom{n}{n_1, \cdots, n_r} = 1/\binom{n+r-1}{r-1}\binom{n}{n_1, \cdots, n_r}$。卡尔纳普根据 m^* 定义了 $c^*(Q_j(a_{n+1}) \mid \Theta(a)) = (n_j+1)/(n+r)$。这是一种类似拉普拉斯的“+1”平滑方法,所有发生 k 次的事件都当作发生了 k+1 次,对未曾出现过的事件,也赋予一定的概率 $1/(n+r)$。

(C) 和(D) 是利斯塔提出的,主要用于处理当 $r \gg n$ 时,已发生事件赋予概率太小的问题 $\frac{n_i+1}{n+r} \ll \frac{n_i}{n}$。根据(C,D),类似 c^*,可以定义对应的确证度函数 $c^\$$,$c^\#$。当 $m_0 = 0$ 时,即,当所有的 r 种可能都发生了时,可以证明 $c^\# = c^\$ = c^*$。

在上述情形(A,B,C,D) 中,并没有涉及层级描述 m,能使层级描述 m 真正起作用的是古德 - 图灵估计和拟古德 - 图灵估计的方法。

2. 拟古德 - 图灵估计与古德 - 图灵估计

根据(E),$m^\tau(m_0, \cdots, m_n) = 1/p(n,r)$。每个“等级描述” $(m_0, \cdots, m_n)$ 对应 $\binom{r}{m_0, \cdots, m_n}$ 个结构

描述，每个结构描述$(n_1,\cdots,n_r)$对应$\binom{n}{n_1,\cdots,n_r}$个状态描述，所以，$m^{\tau}(\Theta(a))=1/\binom{n}{n_1,\cdots,n_r}\binom{r}{m_0,\cdots,m_n}p(n,r)$。然后可以定义“确证度”函数$c^{\tau}$，将其归一化$c^{\tau}_{norm}$。$c^{\tau}_{norm}$相当于用$n_j^*:=(n_j+1)(m_{n_j+1}+1)/m_{n_j}$而不是$n_j$来表示事件 j 发生的频次。这是胡特尔[6]提出的一种方法，可以看作一种拟古德－图灵估计。而古德－图灵估计[5]是源于图灵破译 Enigma 时发明的弥补很少发生甚至以前从未发生的事件的平滑方法，$c^{\tau 0}_{norm}(Q_j(a_{n+1})\mid\Theta(a))=\frac{(n_j+1)m_{n_j+1}}{m_{n_j}}/\sum_{j=1}^{r}\frac{(n_j+1)m_{n_j+1}}{m_{n_j}}$。差别在于古德－图灵估计$c^{\tau 0}_{norm}$用的是$\frac{(n_j+1)m_{n_j+1}}{m_{n_j}}$，而拟古德－图灵估计$c^{\tau}_{norm}$用的是$n_j^*$，这两种方法看上去很相似，实则不同。虽然$c^{\tau}_{norm}$，$c^{\tau 0}_{norm}$都是基于对频次$n_j$的修正，但当$m_{n_j+1}\ll m_{n_j}$时，可能使得$c^{\tau}_{norm}$，$c^{\tau 0}_{norm}$严重偏离频率。

在独立同分布的情形下，确证度趋近频率是合理的，但这需要引入额外的平滑方法，比如，在$c^{\tau 0}_{norm}$中用$\hat{m}_k:=a+b\log k$代替m_{n_j}可得到一种新的确证度函数$c^{\tau 1}_{norm}$，它可以强行将m_{n_j}与n_j建立联系，从而强行将$c^{\tau 1}_{norm}$收敛到频率。强制收敛到频率有很多种方法，比如，如果对$c^{\tau 0}_{norm}$用满足归纳定义$\frac{\hat{m}_{k+1}}{\hat{m}_k}=\frac{k}{k+1}\left(1-\frac{m_1[\![\prod_{i=1}^{r}n_i=0]\!]}{n}\right)$的函数$\hat{m}$做平滑，则可得到某种近似频率的确证度函数$c^{\tau 2}_{norm}$。借助诸如此类的强制手段，向频率的收敛虽然可以保住，但“等级描述”本来试图通过在更高的层次上划分等价类、分配权重、从而捕捉更深层次的模式，但这里m_{n_j}被消解掉了，“等级描述”本来的意义丧失了。m 与 n 既不能强制建立关联，也不能人为忽略 m 的意义，关键是要对 m 本身也给出一种合理的估计。下面我们讨论如何用最大熵原则来估计 m。

3. 从古德－图灵估计到卡尔纳普公式

已知$\sum_{i=1}^{n}i\cdot m_i=n$和$\sum_{i=0}^{n}m_i=r$，定义$p_i:=\frac{m_i}{r}$，$p:=(p_0,\cdots,p_n)$，那么有$\sum_{i=0}^{n}p_i\cdot i=\frac{n}{r}$和$\sum_{i=0}^{n}p_i=1$。然后在其约束下最大化香农熵 H(p)。拉格朗日乘子式为$L=-\sum_{i=0}^{n}p_i\log p_i-\beta(\sum_{i=0}^{n}p_i\cdot i-\frac{n}{r})-\lambda(\sum_{i=0}^{n}p_i-1)$。然后令$\nabla_p L=0\Rightarrow p_i=2^{-\lambda-\frac{1}{\ln 2}}\cdot 2^{-\beta i}$，当 n 足够大时，可估计，$\sum_{i=0}^{n}2^{-\beta i}\approx\int_0^n 2^{-\beta x}dx\approx 1/\beta\ln 2$，$\sum_{i=0}^{n}2^{-\beta i}\cdot i\approx\int_0^n x2^{-\beta x}dx\approx 1/(\beta\ln 2)^2$。将$p_i$代入约束条件可得，$2^{-\lambda-\frac{1}{\ln 2}}\approx\beta\ln 2\approx\frac{r}{n}$，所以有，$\hat{p}_i=\frac{r}{n}2^{-\frac{ir}{n\ln 2}}$和$\hat{m}_i=\frac{r^2}{n}2^{-\frac{ir}{n\ln 2}}$，而且，$\frac{\hat{m}_{i+1}}{\hat{m}_i}=2^{-\frac{r}{n\ln 2}}$，所以$\frac{\hat{m}_{i+1}}{\hat{m}_i}$可以看作样本数 n 和种类数 r 的函数。当样本数远远小于种类数$n\ll r$时，$\frac{\hat{m}_{i+1}}{\hat{m}_i}$很小。当 n 增大时，$\frac{\hat{m}_{i+1}}{\hat{m}_i}$也越来越大。但出现 i+1 次的种类仍然少于出现 i 次的种类。当样本数远远大于种类数$n\gg r$时，恰好出现某个特定的次数的种类不太可能，m 中大部分的值都是 0，所以这种平滑方式还是比较符合直观的。

将$\hat{m}_i$代入$c^{\tau 0}_{norm}$即得到一种新的平滑方法$c^{\tau 3}_{norm}$，易证$c^{\tau 3}_{norm}=c^*$。所以通过 m 对进行最大熵估计，从古德－图灵估计导出了卡尔纳普“确证度”函数c^*。将$\hat{m}_i$代入拟古德－图灵估计c^{τ}_{norm}得到$c^{\tau 4}_{norm}$。而$c^{\tau 4}_{norm}$却不太合理，这是因为，平滑就是要“损有余而补不足”，削减高频事件的权重分给低频事件，而$c^{\tau 4}_{norm}$却恰恰相反。当然，这里对拟古德－图灵估计相对于古德－图灵估计的不合理性的批评也只在一种比较弱的意义上进行的，因为，平滑方法本是为了处理样本不足的数据稀疏性问题而提出的，但这里用最大熵原则估计 m 的计算过程却要假设样本足够大，所以这里的批评是在弱的意义上的，但即使如此，对比古德－图灵估计$c^{\tau 0}_{norm}$，仍显示出拟古德－图灵估计c^{τ}_{norm}具有一定的不合理性。

不难看出，$c^{\#}$，$c^{\$}$，$c^{\tau}_{norm}$都违反了充分性假设 SP。卡尔纳普的 λ－连续统$c_\lambda$是$c^*$的简单扩充。卡尔纳普证明：如果语言$\mathcal{L}$至少包含两个谓词$m\geq 2$，那么$\mathcal{L}$上的概率赋值 w 满足 Ex 和 SP 当且仅当对于某个$0\leq\lambda\leq\infty$，$w=c_\lambda$。通过该定理，卡尔纳普给出了能满足原则 Ax 和 SP 的精确刻画。对于任何有限的 λ，c_λ最终都会收敛到频率，$c^{\#}$，$c^{\$}$也一样，后二者是针对 n 过小时对$c^*$的补充。当$m_0=0$时，m 就完全不起作用了，在古德－图灵估计$c^{\tau 0}_{norm}$中，m 本来起实质性作用，但它不能收敛到频率，需要引入额外的平滑办法，通过对 m 进行最大熵估计，古德－图灵估计退化为卡尔纳普“确证度”函数c^*。

4. 确证度与平滑方法

虽然这几种方案采用的无差别原则作用的层面不同，但从结果上看，当样本 n 足够大时，最终的“确证度”都收敛到了频率。所以当样本足够大时，这几种方案同样有效或同样无效。企图通过频率逼近概率的背后往往隐藏着“独立同分布”的假设。这几种方案有效或无效都取决于现实分布是否是独立同分布，对于高阶的马尔科夫过程都无能为力。事实上，对称性原则 Ex 和 Ax 说的是，事件发生的时间顺序无关紧要，这保证了事件背后概率分布的“独立性”，但这显然不合理。时间顺序是因果关系的重要体现，早先发生的事可能对后面遥远未来发生的事都具有不可忽略的影响，任意阶的马尔科夫过程都可能发生。对于高阶马尔科夫，因为训练样本规模的限制，大部分“高阶结构”很少出现在训练集中，所以无法通过统计频率给出合理的估计，所以也要削减高频弥补低频。但究竟怎么才算一种合理的通用的平滑方法？是否存在一个合理的“确证度”函数能够把时间因素考虑在内从而能把握住更复杂的结构？无差别原则作用在哪儿才能体现我们这种更深层次的“无知”？李熙[11]曾通过引入时间序列解决这个问题，并论证基于所罗门诺夫先验的归纳逻辑不仅具有“通用性”，还可以只针对某些具体模式在忽略无关信息的情况下进行归纳。下面对比卡尔纳普式的归纳逻辑的缺陷，对此方案给出进一步的辩护。

二、卡尔纳普式的归纳方法的局限与基于所罗门诺夫先验的归纳方法

一个具体的状态描述函数 h 唯一决定一个状态描述 Θ 或说可能世界，而它又可以被一个程序 p 输出，所以可以将 p 与 $h_{1:\infty}$ 等同看待。算法概率 $M(\Theta(a)) := \sum_{p:U(p)=h_{1:n^*}} 2^{-|p|}$，其中 U 是通用单调图灵机。算法概率可以有一种频率解释，就是与历史经验相一致的可能世界与所有可能世界的比。算法概率 $\approx \frac{|\text{一致的可能世界}|}{|\text{所有可能世界}|}$。$M(\Theta(a)) \approx \lim_{n\to\infty} \frac{|\{p: |p|=n \& \wedge_{i=1}^{n} Q_{<U(p)>_i}(a_i) \equiv \Theta(a)\}|}{2n}$。$M(\Theta(a))$ 也是在通用单调图灵机 U 的输入带上抛掷一枚质地均匀的硬币，正面写 1、反面写 0 所能输出历史 $\Theta(a)$ 的概率。这意味着，“无差别原则”作用在通用单调图灵机的前端。也就是说，“无差别原则”作用在可能世界的产生方式这一层，它不属于(A，B，C，D，E) 的任何一层，(A，B，C，D，E) 可以看作从“现象”的角度给出的区分，而算法概率直接作用于导致“现象”的“原因”这一层。

而在卡尔纳普式的归纳逻辑中，不管“无差别原则”作用在(A，B，C，D，E) 中的哪一层，最后给出的“先验”都具有一个共同点，那就是——具有相同“结构描述”的所有“状态描述”分享相同的权重。这一点决定了 $c^{\dagger}$，c^{*}，$c^{\$}$，$c^{\#}$，$c^{\tau}$ 等等都不可能具有通用性，这些“确证度”函数无法帮助我们学出所有可能的模式。这是因为，伊戈尔、图森特[12](Igel & Toussaint) 证明，无免费午餐成立的充要条件是——“假设空间”上的概率分布是“块均匀”的。“假设空间”$\mathcal{Y}^{\mathcal{X}}$ 上的概率分布 P 是“块均匀的”，是指 $\forall f, g \in \mathcal{Y}^{\mathcal{X}}: \forall y \in \mathcal{Y}(|f^{-1}(y)| = |g^{-1}(y)|) \Rightarrow P(f) = P(g)$，而具有相同“结构描述”的所有“状态描述”分享相同的权重恰恰是“块均匀”的。如果假设空间上的概率分布满足“块均匀性”，那么，对于任何算法来说，它们在整个假设空间上的期望表现是一样的。如果在一类函数上表现良好，那么在另一类函数上的表现必定很差。

定理 2.1(无免费午餐定理[12])。当且仅当概率分布 P 是块均匀的，那么对于任何算法 A，A'，任何 $k \in \mathbb{R}$，任何 $m \in \{1, \cdots, |\mathcal{X}|\}$，任何损失函数 L，都有

$$\sum_{f\in\mathcal{Y}^{\mathcal{X}}} P(f)\left[\!\left[k = L\left(T_m^{y}(A,f)\right)\right]\!\right] = \sum_{f\in\mathcal{Y}^{\mathcal{X}}} P(f)\left[\!\left[k = L\left(T_m^{y}(A',f)\right)\right]\!\right]$$

其中 $T_n := <(x_1, f(x_1)), \cdots, (x_n, f(x_n))>$，$T_n^{x} := <x_1, \cdots, x_n>$，$T_n^{y} := <f(x_1), \cdots, f(x_n)>$ 且 $A: T_n \mapsto x_{n+1} \in \mathcal{X} \setminus T_n^{x}$。

推论 1. 无免费午餐定理适用于基于“具有相同结构描述的所有状态描述分享相同的权重”这一假设的所有卡尔纳普式的归纳逻辑。

在上述无免费午餐定理 2.1 中，令 $\mathcal{X} := \mathcal{C}$，$\mathcal{Y} := \mathcal{Q}$，状态描述 h: $\mathcal{C} \to \mathcal{Q}$ 可以看作一个分类问题。具有相同“结构描述”的所有“状态描述”分享相同的权重这一事实使得 m^* 恰恰是“块均匀”的，这直接导致对任意的算法 A，A′

$$\sum_{h\in\mathcal{Q}^{\mathcal{C}}} m^*(h_{1:m})\left[\!\left[k = L\left(T_m^{y}(A,h)\right)\right]\!\right] = \sum_{h\in\mathcal{Q}^{\mathcal{C}}} m^*(h_{1:m})\left[\!\left[k = L\left(T_m^{y}(A',h)\right)\right]\!\right]$$

也就是说，所有的算法在状态描述的预测问题

上都同样好或同样差。而且，此结论对 $m^{\$}$，$m^{\#}$，$m^{\tau}$ 等都成立，只要接受了卡尔纳普“具有相同结构描述的所有状态描述分享相同的权重”这一预设，就不可能基于此预设构造出任何通用的归纳算法。比如，根据卡尔纳普“确证度”函数，最直接的归纳方法应该是 $A(h_{1:n}) := \operatorname{argmax}_j c^*(Q_j(a_{n+1}) \mid h_{1:n})$，不妨令损失函数 $L(A,n,h) := [\![A(h_{1:n}) \neq h_{n+1}]\!]$。令A′为随机预测算法，我们希望 A 会比A′表现好，但二者在所有可能的状态描述上的期望表现是一样的。

$$\sum_{h\in Q^{\mathcal{C}}} m^*(h_{1:n})L(A,n,h) = \sum_{h\in Q^{\mathcal{C}}} m^*(h_{1:n})L(A',n,h)$$

类似的，因为 (C,D,E) 都接受这一假设，所以也都不可能获得免费午餐。也就是说，如果考虑所有的可能，c^*，c_λ，$c^{\$}$，$c^{\#}$，$c^{\tau}$，$c^{\tau}_{\text{norm}}$，$c^{\tau_0}_{\text{norm}}$，$c^{\tau_1}_{\text{norm}}$，$c^{\tau_2}_{\text{norm}}$，$c^{\tau_3}_{\text{norm}}$，$c^{\tau_4}_{\text{norm}}$跟最简单的 $c^{\dagger}$ 一样，都无法帮助我们获得免费午餐。

要想享受“免费的午餐”，必须打破“块均匀性”。一个“假设空间”中的大部分函数都是“算法随机”的，而“块均匀性”也是一种弱的“均匀性”，它意味着需要分配大部分的权重给“算法随机”的“函数”。已知“算法概率”$M \approx \xi := \sum_{v\in\mathcal{M}} 2^{-K(v)} v$，其中$\mathcal{M}$是下半可计算的半测度的集合，如果状态描述不是由确定性的 h 决定，而是由某个概率分布 v 决定，则这种状态描述可以看作不确定的可能世界。$2^{-K(v)}$ 是所罗门诺夫先验概率，K 是柯尔莫哥洛夫复杂度函数。通过算法概率或所罗门诺夫先验，对“算法随机”的“函数”赋予 0 权重，只把宝贵的“权重”赋给那些有规律的可计算“函数”/“可能世界”，从而打破了“块均匀性”，这使得通用归纳成为可能(Everitt&Lattimore&Hutter[13])。所以，从无免费午餐定理的角度来说，基于所罗门诺夫先验的通用归纳要优于基于卡尔纳普假设的各种归纳方法。虽然所罗门诺夫先验优于卡尔纳普式的归纳模型，但它是否具有最优性？在什么意义上是最优的？

三、对所罗门诺夫先验的“最优性”的辩护

机器学习领域对奥卡姆剃刀的解释是：对于具有同样解释力的模型，越简单的越似真。这也是所罗门诺夫对它的用法：模型的权重应该与其复杂性呈反向增减的关系。即，对于任何 w，如果 w 与 v 的复杂性反向增减且 $0 \lneq w(v) \lneq 1$，$\sum_{v\in M} w(v) < \infty$，那么 w 就符合奥卡姆剃刀的哲学思想。能否给奥卡姆剃刀一个更强的解释、然后借助额外的约束条件给出一个合理的先验呢？奥卡姆剃刀原则的一种解释：极小化期望模型复杂度 $\mathbb{E}_w[K]$。这种解释下的奥卡姆剃刀与最大熵原则有什么关系？由香农编码定理知，最优码长近似等于香农熵。假设“上帝”对每个可能世界 v 的“最优”编码就是 $K(v)$，那么不管可能世界的分布如何，都会有，$H(w) \leq \mathbb{E}_w[K]$。而限定香农熵 $H(w)$，极小化 $\mathbb{E}_w[K]$，就是选择相信“上帝”对可能世界的编码是以一种最优的方式进行的——期望码长最短。所以这里奥卡姆剃刀与最大熵原则体现出一种对称性，如果限定期望码长 $\mathbb{E}_w[K]$，最大化香农熵 $H(w)$，那么仍然是在猜测“上帝”对可能世界的编码是以一种最优的方式进行的。一方面要极大化香农熵，一方面要借助奥卡姆剃刀极小化期望模型复杂度，所以结合奥卡姆剃刀和最大熵原则，最终的优化目标可以理解为使得 $\mathbb{E}_w[K]/H(w) \to 1$，所以不妨直接优化 $\mathbb{E}_w[K]/H(w)$。

定理 3.1 “归一化的”所罗门诺夫先验可以通过极小化 $\mathbb{E}_w[K]/H(w)$ 得到。

证明. 在 $\sum_{v\in M} w_v = 1$ 约束下极小化 $\mathbb{E}_w[K]/H(w)$。令 $L = \mathbb{E}_w[K]/H(w) - \lambda(\sum_{v\in M} w_v - 1)$，令偏导等于零 $\frac{\partial L}{\partial w_v} = 0$ 得 $w_v = \lambda' \cdot 2^{-\frac{H(w)}{\mathbb{E}_w[K]} K(v)}$，其中，$\lambda' = 2^{\lambda \cdot \frac{(H(w))^2}{\mathbb{E}_w[K]} - \frac{1}{\ln 2}}$。由 $\sum_{v\in M} w_v = 1 \Rightarrow \lambda' = (\sum_{v\in M} 2^{-\frac{H(w)}{\mathbb{E}_w[K]} K(v)})^{-1}$ 可知要求的 w 必是 $w_v = 2^{-\frac{K(v)}{T}} / \sum_{v\in M} 2^{-\frac{K(v)}{T}}$（公式 3.1）的不动点，其中 $T := \mathbb{E}_w[K]/H(w)$。因为 $H(w) \leq \mathbb{E}_w[K] \leq H(w) + K(w)$，所以，$T \geq 1 \Rightarrow \mathbb{E}_w[K] = H(w) = \infty$。如果 w 下半可计算，$K(w) < \infty \Rightarrow T = 1 \Rightarrow w^*_v = 2^{-K(v)} / \sum_{v\in M} 2^{-K(v)}$，这就得到了归一化的所罗门诺夫先验 w^*。如果要求 w 必须下半可计算，因为 $T = 1$，公式 3.1 右边不再包含 w，所以 w^* 是公式 3.1 的唯一解，即其唯一的不动点。

结　语

本文把卡尔纳普的工作放在人工智能领域普遍关心的数据稀疏问题这个大背景下来考虑，把卡

尔纳普公式与利斯塔估计、古德－图灵估计、拟古德－图灵估计、线性插值平滑方法等放在一个平台上讨论而不是孤立起来研究传统的归纳逻辑。为了处理数据稀疏性问题，人们提出了各种各样的平滑方法，不过大致方法都是通过削减高频事件的权重弥补低频事件的权重，通过低阶马尔科夫链估测高阶马尔科夫链。通过最大熵原则推测，拟古德－图灵估计可能不满足这个准则。更严重的是，基于“具有相同结构描述的所有状态描述分享相同的权重”的卡尔纳普式的归纳方法都面临着无免费午餐定理的质疑，也就是说，在考虑通用性方面，所有这些卡尔纳普式的方法都是同样好或同样差的，完全不具有通用性。而基于所罗门诺夫先验的归纳方法却可以帮助我们打破“块均匀”，赢得“免费的午餐”。而且，这种方法不仅具有通用性，在同时满足奥卡姆剃刀原则和最大熵原则的意义上还具有最优性。

参考文献

〔1〕Rudolf Carnap. Notes on probability and induction [J]. *Synthese*, 1973, 25(3): 269－298.

〔2〕Jaakko Hintikka. A two－dimensional continuum of inductive methods [J]. *Studies in Logic and the Foundations of Mathematics*, 1966, 43: 113－132.

〔3〕Sandy Zabell. Confirming universal generalizations [J]. *Erkenntnis*, 1996, 45(2－3): 267－283.

〔4〕Eric Sven Ristad. A natural law of succession [J]. *arXiv preprint cmp－lg/9508012*, 1995.

〔5〕Irving J Good. The population frequencies of species and the estimation of population parameters [J]. *Biometrika*, 1953, 40(3－4): 237－264.

〔6〕Marcus Hutter. Offline to online conversion [J]. *In proceeding of International Conference on Algorithmic Learning Theory*. Springer, Cham, 2014: 230－244.

〔7〕Hilary Putnam. ‘Degree of confirmation’ and inductive logic [J]. *The Philosophy of Rudolf Carnap*, *Schilpp*(*ed.*), 1963.

〔8〕Ray Solomonoff. A formal theory of inductive inference. Part Ⅰ and Part Ⅱ [J]. *Information and control*, 1964, 7(1－2): 1－22, 224－254.

〔9〕Ray Solomonoff. Complexity－based induction systems: comparisons and convergence theorems [J]. *IEEE transactions on Information Theory*, 1978, 24(4): 422－432.

〔10〕Samuel Rathmanner and Marcus Hutter. A philosophical treatise of universal induction [J]. *In Entropy*, 2011, 13(6): 1076－1136.

〔11〕LI Xi, Why Inductive Logic Needs Solomonoff’s Prior? [J]. *Studies in Logic*, 2014, 7(4): 48－68.

〔12〕Christian Igel and Marc Toussaint. A no－free－lunch theorem for non－uniform distributions of target functions [J]. *Journal of Mathematical Modelling and Algorithms*, 2004, 3(4): 313－322.

〔13〕Tom Everitt, Tor Lattimore, and Marcus Hutter. Free lunch for optimization under the universal distribution [J]. *In proceeding of IEEE Congress on Evolutionary Computation* (*CEC*14). IEEE, 2014: 167－174.

The Limits of Carnapian Inductive Logic and the Advantages of Solomonoff Prior

LI Xi

(Department of Philosophy, Central South University, Changsha 410083, China)

Abstract: In 1950s, Carnap develops inductive logic to express the degree of confirmation of some hypothesis relative to some evidence. After that, Ristad, Good develops serveral smoothing methods. Most of these smoothing methods can be taken as sort of general Carnapian inductive logic. Although they can be deduced from applying the so－called “indifference principle” to different levels, they are still lack of solid theoretical foundations. It seems that none of these “indifference principles” work, except the “indifference principle” on the level of the generation of possible worlds, which leads us to Solomonoff’s prior. According to the no－free－lunch theorem, the Carnapian logics are as good and as poor as any random algorithm, while Solomonoff’s induction model could achieve “universality”. Besides, Solomonoff’s prior is optimal with respect to the constraint of Occam’s razor and maximum entropy principle.

Key words: inductive logic; Good－Turing estimate; no－free－lunch theorem; Solomonoff’s prior

(本文责任编辑: 费多益)