云南大学学报 (自然科学版), 2002, 24 (3): 183~185 CN 53-1045/N ISSN 0258-7971
Journal of Yunnan University

Bayesian 网的信息熵*

张忠玉[1], 刘惟一[1], 张玉琢[2]
(1. 云南大学 计算机科学系, 云南 昆明 650091; 2. 云南师范大学 计算机科学系, 云南 昆明 650092)

摘要: 用信息熵的观点, 如果将 Bayesian 网看作 Agent 的背景知识, 采用与 Bayesian 网对应的概率分布作为信念函数的 Agent 的分布是最合理的, 说明了与最大熵相对应的概率分布正好是在条件独立性假设下由 Bayesian 网确定的特征概率分布.

关键词: Bayesian 网; Agent; 条件独立性; 信息熵
中图分类号: TP 18 文献标识码: A 文章编号: 0258-7971(2002)03-0183-03

概率方法已成功地处理了许多与不确定性有关的问题, 有丰富的理论和系统的方法[1]. 但对概率的本质长期以来存在着 2 种不同的解释, 一种认为概率是频率稳定性的基础, 它是客观的, 不随人的认识而改变; 另一种认为概率是人的信念程度, 是能随着信息的变化而改变的[1]. 与此相应, Bayesian 网作为一种不确定性问题的图形化表示的概率因果关系模型, 对它的语义的理解也存在着 2 种不同的解释[2], 一种认为 Bayesian 网中各节点的概率分布是客观的, 应由问题域中已有的数据统计得出; 另一种认为, Bayesian 网可以解释为一个 Agent 的背景知识, 这样它的各个节点的概率就是 Agent 对世界某一状态的主观信念, 它的概率可由专家知识予以主观确定. 与前一种观点相应的 Bayesian 网的构造方法是从大量样本数据中发掘 Bayesian 网, 大致有两类算法: 基于搜索和打分的算法或基于数据依赖分析的算法[3]. 这类算法的缺点是计算量大, 效率低. 近年来越来越多的学者倾向于后一种观点, 认为可以根据因果独立性分析, 由专家给出一个初始化的因果模型和联合概率分布, 然后, 通过机器学习算法获得新的后验概率信息, 并对 Bayesian 网的结构和概率分布进行更新, 修正所得到的 Bayesian 网. 如何评价一个算法的学习效率呢? 从信息论的角度看, 一个好的算法应使得系统的信息熵不断增加, 这样, 对 Bayesian 网的信息熵的研究就很有意义了, 事实上 Xiang[4] 的 Markov 网发现算法就是一种最大熵算法[3,4]. 核心问题是我们怎样才能决定合理的信度, 本文用信息论的观点对 Bayesian 网的信息熵作了语义说明, 并证明 Bayesian 网的概率分布对应于信息系统的最大熵, 还讨论了独立性假设在 Agent 信念传播中的不同语义解释.

1 信息熵

信息熵[1] 是对不确定性的一种度量, 香农对信息熵的定义为: $H = \log \frac{\text{后验概率}}{\text{先验概率}}$, 在无干扰的情况下, 传来的信息告诉某件事情发生, 则某件事情必是发生了, 所以上式分子为 1, 因而信息源输出的第 i 个消息的信息量为 $\log \frac{1}{P(i)} = -\log P(i)$, 第 i 个消息可能有 k 个状态, 那么输出这个消息的总信息量的期望值为: $-\sum_{i=1}^{k} P(i) \cdot \log P(i)$, 根据拉普拉斯无关性原理和 Jayne 的最大熵原理, 前者认为如果 X 不关心 J 中取值哪个为真, 则它将给每个信度为 $1/J$, 后者引申和概括前者如下: 在 $C_1, \cdots, C_N$ 上的概率函数可以通过给每一个参数 $x^{k_1, \cdots, k_N} = P(C_1 = v_1^{k_1} \wedge \cdots \wedge v_N^{k_N})$ 标值的方法而全部标识出来. 其中, $v_i^{k_i} \in v_i^{k_1} \cdots v_i^{k_N}$, $i =$

* 收稿日期: 2001-12-04
基金项目: 国家自然科学基金资助项目(69763003); 曲靖师院科技开发资助项目(0112904).
作者简介: 张忠玉(1967-), 男, 云南人, 曲靖师范学院讲师, 云南大学在读研究生, 主要从事数据库与知识工程研究.

$1, \cdots, N$ 和由背景知识蕴含的限制一起, $x^{k_1, \cdots, k_N} \in [0, 1]$, 并且, $\sum_{k_1, \cdots, k_N} x^{k_1, \cdots, k_N} = 1$, 最大熵原理认为如果缺乏进一步的信息, X 将通过选择一个最合理的 $x^{k_1, \cdots, k_N}$ 而选择一个最合理的信念, 服从熵最大的限制

$$H = -\sum_{k_1, \cdots, k_N} x^{k_1, \cdots, k_N} \log x^{k_1, \cdots, k_N}. \tag{1}$$

问题是, 由专家根据独立性假设给出的 Bayesian 网的概率分布函数与最大熵相应的概率函数是否一致?

2 Bayesian 网有最大的信息熵

由(1)知, Agent 的信念对应于系统的信息熵最大, 能否用一个 Bayesian 网表达 Agent 的信念知识库, 利用 Bayesian 网的良好性质和健壮的推理算法实现 Agent 的行为预测, 关键的问题是要证明 Bayesian 网在什么情况下有最大的信息熵. 结论是:

定理 1 给定 Bayesian 网的因果图和特征概率分布以及因果无关性, 那么, 最大化熵的分布就对应于由 Bayesian 网在条件独立性假设下决定的概率分布[2].

证明 方法是假设熵函数是凸函数, 具有极大值, 求出熵函数取极大值的条件, 再证明 Bayesian 网的概率分布满足这些条件. 由凸函数的定义: H 是严格的凸函数, 当且仅当: 参数 $x^{k_1, \cdots, k_N}$ 的 2 个不同向量 a, b 以及 $\lambda \in (0, 1)$, 有

$H(\lambda a + (1-\lambda)b) > \lambda H(a) + (1-\lambda)H(b)$ 将(1)式代入得

$$\lambda \sum a_i \log a_i + (1-\lambda) \sum b_i \log b_i - \sum (\lambda a_i + (1-\lambda) b_i) \log (\lambda a_i + (1-\lambda) b_i) > 0,$$

$$\lambda \sum a_i \log \frac{a_i}{\lambda a_i + (1-\lambda) b_i} + (1-\lambda) \sum b_i \log \frac{b_i}{\lambda a_i + (1-\lambda) b_i} > 0,$$

$$\lambda d(a, \lambda a + (1-\lambda)b)) + (1-\lambda) d(b, \lambda a + (1-\lambda) b) > 0. \tag{2}$$

其中 d 是交叉熵, a, b 及 λ 是非零的, $\sum a_i = 1 = \sum b_i$, $\lambda \in (0, 1)$; d 是严格正的, 因而, H 严格为正, 由拉格朗日乘法原理, 熵函数存在一个局部最大值.

假设一个 Bayesian 网的结点 $C_1, \cdots, C_n$ 按父子关系排序, 即 C_i 的父亲结点排在 C_i 的前面, 并设所有概率严格为正(实际上可以为 0). 用 $c_i^{k_i}$ 表示 $C = v_i^{k_i}$, $k_i = 1, \cdots, K_i$, $i = 1, \cdots, N$, 则参数可表示为: $y_{i, k_i}^{k_1, \cdots, k_{i-1}} = p(c_i^{k_i} \mid c_1^{k_1} \wedge \cdots \wedge c_{i-1}^{k_{i-1}})$, 概率链式公式为 $x^{k_1, \cdots, k_N} = \prod_{i=1}^{N} y_{i, k_i}^{k_1, \cdots, k_{i-1}}$, 代入(1)有

$$H = -\sum_{k_1, \cdots, k_N} x^{k_1, \cdots, k_N} \log x^{k_1, \cdots, k_N} = \sum_{k_1, \cdots, k_N} \left[\prod_{j=1}^{N} y_{j, k_j}^{k_1, \cdots, k_{j-1}} \right] \sum_{i=1}^{N} \log y_{i, k_i}^{k_1, \cdots, k_{i-1}} =$$

$$\sum_{i=1}^{N} \sum_{k_1, \cdots, k_N} \left[\prod_{j=1}^{N} y_{j, k_j}^{k_1, \cdots, k_{j-1}} \right] \log y_{i, k_i}^{k_1, \cdots, k_{i-1}} = \sum_{i=1}^{N} \sum_{k_1, \cdots, k_N} \left[\sum_{j=1}^{i} y_{j, k_j}^{k_1, \cdots, k_{j-1}} \right] \log y_{i, k_i}^{k_1, \cdots, k_{i-1}},$$

最后一步成立是因为对每一个 i, 可以分离出项: $\sum_{k_{i+1}, \cdots, k_N} \left[\prod_{j=i+1}^{N} y_{j, k_j}^{k_1, \cdots, k_{j-1}} \right]$.

对 Bayesian 网需要考虑 3 种限制: ① 概率特征分布限制; ② 由因果无关性确定的因果限制; ③ 由概率公理给出的额外附加限制. 通过选择合适的归纳方法, 可以得到一个只有较少限制的拉格朗日函数.

概率限制可表示为: $p(c_i^{k_i} \mid c_{r_1}^{k_1} \wedge \cdots \wedge c_{r_L}^{k_{r_L}}) = a_{i, k_i}^{r_1, \cdots, r_L, k_{r_L}}$, 其中 $c_{r_1}^{1}, \cdots, c_{r_L}, k_{r_L}$ 包含 c_i 的所有父结点, $r_1, \cdots, r_L < i$(按排序关系). $i = 1, \cdots, N$, 额外附加限制可表示为: $\sum_{k_i} y_{i, k_i}^{k_1, \cdots, k_{i-1}} = 1$, 对每个 $k_1, \cdots, k_{i-1}$, $i = 1, \cdots, N$; 将 H 分解为 $H = \sum_{i=1}^{N} H_i$, H_i 是只包含结点 C_i 及其所有祖先结点的 Bayesian 子网, 用归纳法可以证明: H 最大, 当且仅当各 H_i 最大. 现在考虑概率特征限制: 用 $b^{k_{r_1}, \cdots, k_{r_L}} = p(c_{r_1}^{k_{r_1}} \wedge \cdots \wedge c_{r_L}^{k_{r_L}})$ 及

$e^{k_1,\cdots,k_{N-1}} = \prod_{j<N} y_{j,k_j}^{k_1,\cdots,k_{j-1}}$ 分别表示由 $C_1, \cdots, C_{N-1}$ 相应的子网的熵最大时的约束条件, 则对 N 个结点的 Bayesian 网有

$$a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}} b^{k_{r_1},\cdots,k_{r_L}} = p(c_N^{k_N} \wedge c_{r_1}^{k_{r_1}} \wedge \cdots \wedge c_{r_L}^{k_{r_L}}) = \sum_{k_i \neq k_{r_1},\cdots,k_{r_L},k_N} p(c_1^{k_1} \wedge \cdots \wedge c_N^{k_N}) =$$

$$\sum_{k_i \neq k_{r_1},\cdots,k_{r_L},k_N} \prod_{j \leqslant N} y_{j,k_j}^{k_1,\cdots,k_{j-1}} = \sum_{k_i \neq k_{r_1},\cdots,k_{r_L},k_N} e^{k_1,\cdots,k_{N-1}} y_{N,k_N}^{k_1,\cdots,k_{N-1}}.$$

构造一拉格朗日函数如下, 使 $-H_N$ 最大.

$$\wedge_N = \sum_{k_1,\cdots,k_N} e^{k_1,\cdots,k_{N-1}} y_{N,K_N}^{k_1,\cdots,k_{N-1}} \log y_{N,k_N}^{k_1,\cdots,k_{N-1}} + \sum_{k_{r_1},\cdots,k_{r_L},k_N} \lambda^{k_{r_1},\cdots,k_{r_L}} \left[\sum_{k_i \neq k_{r_1},\cdots,k_{r_L},k_N} e^{k_1,\cdots,k_{N-1}} y_{N,k_N}^{k_1,\cdots,k_{N-1}} - a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}} b^{k_{r_1},\cdots,k_{r_L}} \right] + \sum_{k_1,\cdots,k_{N-1}} \mu^{k_1,\cdots,k_{N-1}} \left[\sum_{k_N} y_{N,k_N}^{k_1,\cdots,k_{N-1}} - 1 \right] =$$

$$\sum_{k_1,\cdots,k_N} (e^{k_1,\cdots,k_{N-1}} y_{N,k_N}^{k_1,\cdots,k_{N-1}} \log y_{N,k_N}^{k_1,\cdots,k_{N-1}} + \lambda^{k_{r_1},\cdots,k_{r_L}} [e^{k_1,\cdots,k_{N-1}} y_{N,K_N}^{k_1,\cdots,k_{N-1}} - a_{N,k_N}^{k_{r_1},\cdots k_{r_L}} b^{k_{r_1},\cdots,k_{r_L}}] + \mu^{k_1,\cdots,k_{N-1}} \left[\sum_{k_N} y_{N,k_N}^{k_1,\cdots,k_{N-1}} - 1/K_N \right]),$$

当偏导数为零时, 可得一极大值

$$\frac{\partial \wedge_N}{\partial y_{N,k_N}^{k_1,\cdots,k_{N-1}}} = e^{k_1,\cdots,k_{N-1}} [1 + \log y_{N,k_N}^{k_1,\cdots,k_{N-1}} + \lambda^{k_{r_1},\cdots,k_{r_L}}] + \mu^{k_1,\cdots,k_{N-1}} = 0.$$

考虑 $k_N \neq k'_N$, 上式与式子: $\frac{\partial \wedge_N}{\partial y_{N,k'_N}^{k_1,\cdots,k_{N-1}}} = 0$ 联立, 即可消去 $\mu^{k_1,\cdots,k_{N-1}}$, 从而

$$\lambda_{k_N}^{k_{r_1},\cdots,k_{r_L}} - \lambda_{k'_N}^{k_{r_1},\cdots,k_{r_L}} = \log y_{N,k'_N}^{k_1,\cdots,k_{N-1}} - \log y_{N,k_N}^{k_1,\cdots,k_{N-1}},$$

考虑存在另一组 k 值: $k', \cdots k'_{N-1}$ 使得: $k'_{r_1} = k_{r_1}, \cdots, k'_{r_L} = k_{r_L}$ 从而消去左边可得

$$\log y_{N,k'_N}^{k_1,\cdots,k_{N-1}} - \log y_{N,k_N}^{k_1,\cdots,k_{N-1}} = \log y_{N,k'_N}^{k'_1,\cdots,k'_{N-1}} - \log y_{N,k_N}^{k'_1,\cdots,k'_{N-1}}.$$

由 Bayesian 网的条件独立性, 蕴含下列条件

$$y_{N,k_N}^{k_1,\cdots,k_{N-1}} = y_{N,k_N}^{k_{r_1},\cdots,k_{r_L}} = a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}}, \; y_{N,k_N}^{k'_1,\cdots,k'_{N-1}} = y_{N,k_N}^{k'_{r_1},\cdots,k'_{r_L}} = y_{N,k_N}^{k_{r_1},\cdots,k_{r_L}} = a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}},$$

从而有

$$a_{N,k'_N}^{k_{r_1},\cdots,k_{r_L}} - a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}} = a_{N,k'_N}^{k_{r_1},\cdots,k_{r_L}} - a_{N,k_N}^{k_{r_1},\cdots,k_{r_L}},$$

即 Bayesian 网的分布满足最大熵的分布.

3 结　论

可以看出, 我们用信息熵的观点, 从假定 Bayesian 网作为 Agent 的背景知识和将 Bayesian 网作为因果概率网 2 个不同的角度, 将二者统一在一起, 从这个意义上说, 因果独立性既不是纯粹的因果事实, 也不是纯粹的 Agent 的背景知识断言, 而是一种概率机制, 可以用来从 Agent 的背景知识推出新的概率解释[3].

参考文献:

[1] 薛华成, 汪授泓. 管理信息系统[M]. 北京: 清华大学出版社, 1988.

[2] WILLIAMSON J. Foundation for Bayesian networks. kluwer applied logic series 2001[EB/OL]. http://q_squared.doc.ic.ac.uk/foundations.ps, 2001-11-20.

(下转第 191 页)

less signal is discussed and the broadband wireless access performance is analyzed by comparing the predicted data and the measured data. Solution of designing a broadband wireless access network in a building is presented.

Key words: broadband wireless access; 2.4 GHz band; in-building; path loss

* *

(上接第 185 页)

[3] LEMMER J F. The causal Markov condition, fact or artifact? [J]. Acm Sogart Bbulletin, 1996, 7(3): 3~ 16.

[4] WONG S K M, XIANG Y. Construction of a markov network from data for probabilistic inference[A]. In: Proc. Of the third International Workshop on Rough Sets and Soft Computing[C]. San Jose: Morgan kaufmann, 1994.

The entropy of Bayesian networks

ZHANG Zhong-yu[1], LIU Wei-yi[1], ZHANG Yun-zhuo[2]

(1. Department of Computer Science, Yunnan University, Kunming 650091, China;
2. Department of Computer Science, Yunnan Normal University, Kunming 650092, China)

Abstract: In the view of entropy, that if the graph and probability specification in a Bayesian network are thought of as an agent's background knowledge, the agent is most rational if she adopts the probability distribution determined by the Bayesian network as her belief function. It shows that the distribution determined by the Bayesian network maximises entropy given the causal and probability distribution of a Bayesian network under the conditional independence.

Key words: Bayesian networks; Agent; conditional independence; entropy

* *

“第九届全国高校固体物理科研与教学研讨会”将于 8 月在云南省大理市召开

全国高校固体物理研究会将于 2002 年 8 月中旬在云南省大理市召开“第九届全国高校固体物理科研与教学研讨会”. 会议由全国高校固体物理研究会主办, 南开大学、云南大学、四川师范大学承办. 会议旨在交流我国高等院校固体物理领域近年来所取得的创新性的科研和教学成果, 以促进我国高等院校固体物理领域的交流与协作, 加强固体物理学者之间的学术联系与沟通, 推动固体物理的科研与教学. 会议将以内容丰富、适应面宽、广聚同行、气氛活跃、讲求实效为指导思想, 热忱欢迎固体物理学、材料科学等方面的专家学者, 理工、师范、综合性大学的教师、研究生及对固体物理学感兴趣的各界人士参加.

会议内容: 固体的光、磁、电特性; 介观物理和超晶格; 强关联电子系统; 固体及其表面的电子结构; 固体光学性质研究的实验仪器; 新材料、新器件及应用; 固体物理教学研究; 专著及教材交流; 相关的交叉领域.

有关信息亦可参阅《云南大学学报》网页(http://yndz.chinajournal.net.cn)