学术论文

知识创新随机过程最大熵模型

陈 欣[1,2]，和金生[1]，董丽平[1]
（1. 天津大学管理学院，天津 300072；2. 天津医科大学，天津 300070）

[摘要] 根据创新知识的形成受已有相关知识影响的假设，综合运用概率统计、最优化、最大熵原理构建了知识创新随机过程最大熵模型，给出了以已有知识 $x\in X$ 为约束的新知识 $y\in Y$ 的条件概率 $p(y|x)$。模型具有随机性与因果性对立统一的特点。

[关键词] 最大熵；知识创新；条件熵；随机过程
[中图分类号] TP391 [文献标识码] A [文章编号] 1009－1742（2004）12－0043－04

1 引言

知识创新的核心问题之一是如何表达已有知识，以及如何应用已有知识进行分析、推理，从而得到新知识，其中尤以不确定性知识的推理和表达最为重要，也十分困难。国内外学者利用数学方法进行的理论与实践研究重点集中在知识创新规律探索和知识获取技术两个方面，其中知识获取的主要方法有：粗糙集、遗传算法、神经网络[1~8]，知识创新规律探索尚处起步阶段，方法相对分散且不成熟，有拟线性微分方程、SPA（set pair analysis）等[9,10]。这些模型共同的问题是对非结构化且大范围知识对象的处理有局限性，路径依赖性较强。本文的基本假设是：新知识的形成受已有相关知识的影响，即新知识不能孤立出现。以此为基础构建的知识创新随机过程最大熵模型试图反映、记录并利用知识创新中存在的规律，模型具有随机性与因果性对立统一的特点，是解决知识创新问题的一种尝试。

2 问题的数学表达[11]

设知识创新随机过程 P，输出的创新知识为有限集 Y，$\forall y\in Y$ 的生成受相关已有知识的影响或约束，定义与 Y 有关的已有相关知识集为 X，在给定 $\forall x\in X$ 条件下，随机过程建模的目的是求出形成新知识 $y\in Y$ 的条件概率 $P(y|x)$，即对 $P(y|x)$ 进行估计。用 P 表示所有条件概率分布集合即 $P(y|x)\in P$。

模型的输入是从经过处理的知识库中抽取的样本集 T，

$$T=\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_N,y_N)\}$$

$\forall x_i\in X$，$\forall y_i\in Y$，$1\leqslant i\leqslant N$，(x,y) 的经验分布为

$$\tilde{P}(x,y)=\frac{\mathrm{freq}(x,y)}{N} \tag{1}$$

其中 $\mathrm{freq}(x,y)$ 是 (x,y) 在样本中出现的次数。

3 特征与约束

考虑 Y 的知识环境 X，分析随机过程 P，若考虑所有与 Y 同现的已有相关知识信息 X，模型的建立相当复杂、烦琐，从知识创新规律角度分析，Y 的生成只与 X 中部分信息有关，因此，从 X 中找出对 Y 的取值有价值的知识才是模型所要求的。这些有价值的知识正是随机最大熵模型要寻找的特征。为此，先定义特征、特征函数、约束。

[收稿日期] 2004－02－03；修回日期 2004－02－28
[基金项目] 国家自然科学基金资助项目（70272044）
[作者简介] 陈 欣（1962－），女，浙江谨县人，天津医科大学副教授，天津大学博士生。研究方向：应用数学与管理

$\forall x \in X$ 且 $x = d$，$d \in D$，D 代表 x 的部分信息，若 d 对 $y \in Y$ 的出现有表征作用，则称 (d, y) 为模型的一个特征。同时，事件“d 出现时 $y \in Y$ 出现”在模型中以“0—1”函数的形式表示，即

$$f(x,y) = \begin{cases} 1 & d \text{ 出现且 } y \text{ 出现} \\ 0 & \text{否则} \end{cases}$$

称此函数为特征函数。

为了把有用的特征纳入模型，可通过增加约束使模型满足相应特征的期望值来实现。进一步的问题是求出在限制条件下具有最一致分布的模型，此时，若 f 对模型有价值，则期望概率值 $P(f)$ 等于经验概率值 $\tilde{P}(f)$，其中

$$\tilde{P}(f) = \sum_{x,y} \tilde{P}(x,y) \cdot f(x,y) \tag{2}$$

$$P(f) = \sum_{x,y} P(x)P(y \mid x)f(x,y) \tag{3}$$

设存在 n 个特征 f_i $(i=1, 2, \cdots, n)$，称

$$P(f_i) = \tilde{P}(f_i) i = 1,2,\cdots,n$$

即

$$\sum_{x,y} \tilde{P}(x,y) \cdot f(x,y) = \sum_{x,y} P(x)P(y \mid x)f(x,y)$$

为模型的约束，约束集合为

$$C = \{p \in P \mid P(f_i) = \tilde{P}(f_i), i = 1,2,\cdots,n\}。$$

4 随机过程最大熵模型[12]

由约束集合定义可知，满足约束条件的模型很多，目标是产生在约束集下具有最均匀分布的模型，即要求出最优的 $P(y|x)$ 值需要得到一个最为一致分布的模型，测量这种一致性的方法之一是条件熵，即

$$H(p) = -\sum_{x,y} \tilde{P}(x)P(y \mid x)\lg P(y \mid x) \tag{4}$$

或

$$H(p) = -\int \tilde{P}(x)P(y \mid x)\lg P(y \mid x)\mathrm{d}x \tag{5}$$

其中：$0 \leqslant H(p) \leqslant \lg \mid y \mid$。

由以上分析可知，概率预先未知待求，约束条件是使条件熵 $H(p)$ 等于最大值的概率。这正是最大熵原理所解决的问题。最大熵原理最初由 E. T. Jayness 提出[13]，1989 年 A. B. Tepleman 和李兴斯将最大熵法应用于不可微问题的优化[14]，Della Pietra 等于 1992 年首次将它应用于自然语言处理的模型建立中[15~17]，Jayness 的叙述为“当根据部分信息进行推理时，我们必须使用具有最大熵值且满足已知条件的概率分布。这是我们能够作出的唯一无偏选择，若采用任何其他分配就是对原来没有的信息做了任意假设。”

根据最大熵原理，应该使得 $P(x)$ 和 $f(x)$ 在已知的特征上表现出相同的统计特性，同时又要保证不作任何过多的假设，即要求 $p(x)$ 的熵尽可能大。

因此，在允许的概率分布 C 中选择模型，同时又保证不作任何人为的假设，则具有最大熵的模型 $p_x \in C$ 为所求，即

$$p^* = p_x = \arg\max_{p \in C} H(p) \tag{6}$$

作为被选定的值 p_x 是唯一的，在简单的情况下，p_x 可以通过分析得出，但大多数情况是复杂的，可通过寻找式（6）的等价函数的最优化来实现[14]。

对每个特征 f_i 引入 lagrange 乘子 λ_i，定义 lagrange 函数

$$\Lambda(p,\lambda) = H(p) + \sum_i \lambda_i (p(f_i) - \tilde{p}(f_i)) \tag{7}$$

保持 λ 固定

$$\frac{\partial \Lambda(p,\lambda)}{\partial p} = -\tilde{p}(x)(1 + \lg P(y \mid x)) - \sum_i \lambda_i \tilde{p}(x) f_i(x,y)$$

满足 $\frac{\partial \Lambda}{\partial p} = 0$ 的 p^* 为最优，并且使 $H(p)$ 得到限制条件下最大值，此时

$$p^*(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\Big(\sum_i \lambda_i f_i(x,y)\Big)$$

$Z_\lambda(x) = \sum_y \exp\Big(\sum_i \lambda_i f_i(x,y)\Big)$ 是保证对所有 x 使得 $\sum_y p_\lambda^*(y \mid x) = 1$ 的范化因子。令

$$\Psi(\Lambda) = \Lambda(p^*, \lambda)，则$$

$$\lambda^* = \arg\max_\lambda \Psi(\Lambda)$$

5 特征选择

以上建立的模型可以保证不含任何额外的假设，但不能保证所含特征是最有表征性的特征，因此，建立模型的重要环节是特征的选择。

进行特征选取时，是由特征的信息增益值做标准，一个特征对所处理的任务带来的信息越多，该特征越适合引入模型中[18]。有很多方法来测度这种差异，如 Fisher 准则、L^2 距离度量或 Kullback-leibler（KL）距离法，从准确性、便于计算的角度考虑，选择了 KL 法。首先，形式化一个特征空

间，所有可能的特征构成候补特征集合 F，根据信息增量标准选择 F 中的元素形成模型的特征集合。具体步骤：

Step1：设候补特征集合 F，模型选用的特征集合 S，对应模型 P_S 初始化选用集合，$S=\Phi$

Step2：$\forall f\in F$，求增益值 G_f，根据是 KL 距离，概率分布 p，q 的 KL 距离为

$$D(p\parallel q)=\sum_x p(x)\ln\frac{p(x)}{q(x)}。$$

因此，加入第 n 个特征前后，模型分布与样本分布间的 KL 距离分别为：

$$D(\tilde{p}\parallel p^{(n-1)})=\sum_x \tilde{p}(x)\ln\frac{\tilde{p}(x)}{p^{(n-1)}(x)},$$

$$D(\tilde{p}\parallel p^{(n)})=\sum_x \tilde{p}(x)\ln\frac{\tilde{p}(x)}{p^{(n)}(x)},$$

则引入第 n 个特征 f_n 后的增益值为

$$G(p,f_n)=D(\tilde{p}\parallel p^{(n-1)})-D(\tilde{p}\parallel p^{(n)}),$$

所以，选择的第 n 个特征为

$$f_{\max n}=\arg\max_{f\in F}G(p,f_n)。$$

Step3：选择具有最大增益值的特征 $f_{\max n}$。

Step4：把特征 $f_{\max n}$ 加入集合

$$S=(f_{\max 1},\cdots,f_{\max n})。$$

Step5：调整参数值，计算模型 P_S。

Step6：回到 Step2。

图 1 是模型系统流程。椭圆代表数据，矩形代表过程。

6　分析与结论

创新随机过程最大熵模型是概率统计、最优化、最大熵原理的完美组合，是随机性与因果性对立统一的数学模型，且简洁、易移植，应用范围广泛。模型的主要特点：

1）模型的结构决定对不同的任务只是选择不同的特征集合嵌入模型中，因此，模型可以被多次利用，模型的这种通用性和重用性允许使用者处理各种不同性质的任务。

2）不作未经验证的假设。利用最大熵原理承认已有的事实，对所选特征没有独立性假设。

3）采用 Kullback-Leibler 距离作为特征的约束条件，可以保证模型结论的准确性。同时，模型的结构决定其可比照语言处理模型编程，便于计算机实现。

4）原始知识得到有效利用。一个特征对所处理的任务带来的信息越多，该特征越优先引入模型中。从繁杂的原有知识中结构化选取有利于创新知识的集合，大大降低了知识创新的盲目性。

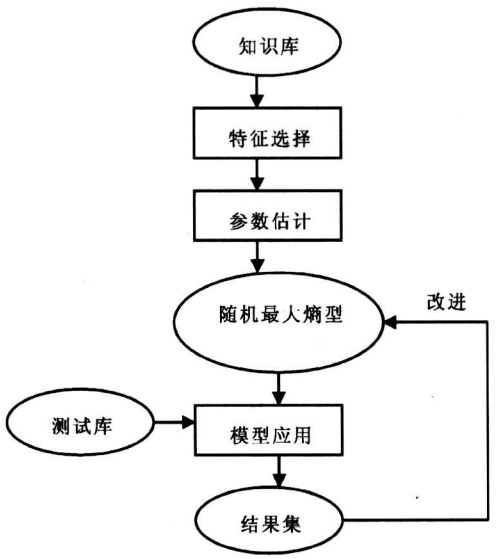

图 1　随机过程最大熵模型系统图

Fig.1　Systematic flow chart of model on maximum entropy of knowledge creating stochastic process

参考文献

[1] Pawlak. Z Rough sets Theoretical aspects of reasoning about data [M]. Dordrecht: Kluwer Academic Publishers, 1991. 58～162

[2] Pawlak Z. Rough set approach to knowledge-based decision support[J]. European Journal of Operational Rewearch 1999,(16):48～57

[3] Guan J W. Bell D A. Rough computational methods for information systems [J]. Artificial Intelligence, 1998, 105(1, 2): 77～103

[4] Goldberg D E. The design of innovation: Lessons from genetic algorithms, lessons for the real world [J]. Technological Forecasting and Social Change, 2000, (64):7～12

[5] Blas Payri. Knowledge base improvement through genetic algorithms [J]. Information Science. 1999, (114): 63～79

[6] Lin N, Cios K J. A Machine learning method for generation of neural network architecture: A Continuous ID_3 Algorithm[J]. IEEE Trans. on Neural Networks, 1992, 3(2): 280～290

[7] 马武瑜,等. 基于代数神经网络的不确定数据知识获取方法[J]. 计算机工程与设计, 2001, 22 (2): 74～76

[8] 彭志刚,等. 基于遗传算法的知识获取及其在故障诊断中的应用研究[J]. 信息与控制, 1999, 28(5): 391～395

[9] 蔡惠京, 吴晓红. 关于知识增长和知识创新的数学模型[J]. 数学的实践与认识, 2001, 31(3): 294～298

[10] 徐忆琳. 用SPA同异反系统理论研究知识创新规律[J]. 科学学研究, 2002, 20(3): 327～329

[11] 徐延勇. 基于最大熵方法的统计语言模型, [J]. 计算机工程与应用,2002, (5): 52～56

[12] Hyo Y W. Systematic bayes prior-assignment by coupling the mini-max entropy and moment-matching Methods [J] IEEE Transactions on Reliability, 1994, 43(2): 85～295

[13] Jayness E T. Information theory and statistical Mechanics [J]. Physical Review, 1957, 106(4): 620～630

[14] 李兴斯. 结构优化设计的最大熵方法 [J]. 计算结构力学及其应用, 1989, 6(1): 36～46

[15] Della Pietra S, Della Pietra V, Mercer R L, Roukos S. Adaptive language modeling using minimum discriminant estimation [J]. Acoustics, Speech, and Signal Processing, 1992, ICASSP-92, 1992 IEEE International Conference on 1992, 1.1 (23-26): 633～636

[16] Berger A L, Setal D P. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 40～724

[17] Rosenfeld R. A maximum entropy to adaptive statistical language learning [J]. Computer Speech and Language, 1996, 10(3): 187～228

[18] 李素建.汉语组块计算的若干研究 [D]. 北京: 中国科学院计算技术研究所, 2002. 23～29

The Model on Maximum Entropy of Knowledge Creating Stochastic Process

Chen Xin[1,2], He Jinsheng[1], Dong Liping[1]

(1. *Management School of Tianjin University, Tianjin* 300072, *China*;
2. *Tianjin Medical University, Tianjin* 300070, *China*)

[**Abstract**] On the basis of the hypothesis that the existing relevant knowledge influences the formation of the new knowledge, this thesis sets up the maximum entropy fundamental model of the stochastic prosess of knowledge innovation by applying the probability & statistics theory, the optimization theory and the maximum entropy principle comprehensively, and puts forward the conditional probability $P(y|x)$ of the new knowledge $y \in Y$ which is restrained by the existing knowledge $x \in X$. This model posesses such character that the randomness and the causality are unity of opposites

[**Key words**] maximum entropy; knowledge innovation; conditional entropy; stochastic process