基于主观贝叶斯的点击流数据分析应用研究

王军豪,彭 岩
(首都师范大学 信息工程学院,北京 100037)

摘 要:通过对不确定性推理和主观Bayes方法的分析研究,提出将主观贝叶斯方法应用到点击流数据分析系统中。在用主观贝叶斯方法进行推理计算时,针对Web日志文件中记录信息的不完备情况,应用了证据的不确定性推理,在系统中对用主观Bayes方法得出结论进行专家分析评估,来确定用户对网站内容的关注程度和上网喜好,从而掌握网站对用户的黏着度,进而为优化网站提供依据,为进一步建设更有吸引力的网站提供决策支持。

关键词:不确定性;贝叶斯;点击流;概率
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2008)07-0116-03

Application and Research of Clickstream Data Analysis Based on Bayes

WANG Jun-hao, PENG Yan
(College of Information Engineering, Capital Normal University, Beijing 100037, China)

Abstract: Through analysis and research of uncertainty reasoning and Bayes, presents a method that applied subjective Bayes approach in click stream data analysis system. In view of the incomplete information recorded in the Web log, while using subjective Bayes methods of reasoning, it has applied the evidence of the uncertainty to carry on expert evaluations with the conclusions drawn from the subjective Bayes method. It can help to master the degree of attention of the user for the site content and their preferences and interests; also, it can grasp the website's tenacity degree to users, then provide more personalized Web services support for optimizing the website, and provide decision support for further constructions with more attraction.

Key words: uncertainty; Bayes; clickstream; probability

0 引 言

点击流数据分析系统主要是对Web日志文件中记录的一些用户上网信息如用户浏览的每个站点、每个页面,在页面上滞留的时间以及点击的链接和图片等信息特征进行分析,根据分析结果可以推断出用户的行为习惯和个人喜好,以及对网站的内容的关注程度,为网站的进一步优化建设提供决策支持。主观贝叶斯方法由Duda和Hart等人在贝叶斯公式的基础上经过改进提出。它建立了相应的不确定性推理模型,并在地矿专家系统PROSPECTOR中得到了成功应用。针对点击流数据分析系统中存在的信息不完备和不确定性问题,文中提出了基于主观贝叶斯的分析方法。该方法具有直观和便于理解,易于发现数据间的因果关系,适于不确定性和不完备信息下进行有效的分析决策等优点。文中将主观贝叶斯方法引入点击流数据分析系统中,对这一方法进行了验证,得到了理想的效果。

收稿日期:2007-10-10
基金项目:北京市优秀人才培养资助项目(20061D0501600220);北京市强教计划资助项目
作者简介:王军豪(1981-),男,硕士研究生,主要研究领域为人工智能及其应用;彭 岩,副教授,博士,主要研究方向为人工智能及应用。

1 不确定性推理和主观Bayes方法

不确定性推理是建立在非经典逻辑基础上的一种推理,是对不确定性知识的应用和处理,严格地说,不确定性推理就是从不确定性的初始证据出发,通过运用不确定的知识最终推出具有一定程度的不确定性,但却是合理或者近乎合理的结论的思维过程[1]。对于许多比较复杂的人工智能系统,往往含有复杂性、不完全性、模糊性或不确定性。当采用产生式系统或专家系统的结构时,要求设计者建立某种不确定性问题的代数模型及其计算和推理过程。知识的不确定性与该领域问题的特征相关,只有根据该领域的问题特征来描述其知识的不确定性,同时还要适合于不确定推理过程中的不确定程度的推算。知识库是人工智能的核心,而知识库中的知识既有规律性的一般原理,又有大

量的不完全的专家知识,即知识带有模糊性、随机性、不可靠或不知道不确定因素。世界上几乎没有什么事情是完全确定的。不确定性推理即是通过某种推理得到问题的精确判断。

不确定性包括知识的不确定性和证据的不确定性。知识的不确定性通常为一个数值,表示相应知识的确定性程度。在实际应用中,知识的不确定性是由领域专家给出的。证据包括两种:求解问题时的初始证据和推理中得到的中间结果。一般来说,证据的不确定性表示应该与知识的不确定性表示保持一致,以便推理过程能对不确定性进行统一处理。初始证据的不确定性必然会造成结论的不确定性,在推理时,中间过程得到的结论往往作为当前证据存入数据库。由于初始证据具有不确定性,上一步推理所得的结论必然具有不确定性,而该结论又作为下一步的证据继续推理,如此往下直到推出结论,这样就将初始的证据的不确定性传递到了最终结论。不确定性推理的方法主要有经典概率方法、逆概率方法、主观 Bayes、证据理论方法和模糊推理方法等。但经典概率方法只适用简单的不确定性推理;逆概率方法则要求各个事件相互独立,不能处理证据之间有相互关联的事件;证据理论方法运算复杂[2]。主观 Bayes 方法是由杜达(R. O. Duda)等人于1976年提出的一种不精确推理模型,并成功地运用于地矿勘探专家系统 PROSPECTOR 中。其推理过程是:领域专家为每条规则提供两个规则强度 LS 和 LN (LS 表现规则 $A->B$ 成立的充分性,LN 表现规则 $A->B$ 成立的必要性。也就是说 LS 表现规则 $A->B$,A 为真时对 B 为真的支持程度,LN 表现了 A 不为真($\sim A$)对 B 为真的支持程度),同时还要给出每个命题的先验可能性,即命题单位元。原始证据的不确定性值由用户在系统运行时提供,其它所有命题的不确定性值均由不确定性的更新算法求出。主观 Bayes 方法通过使用专家的主观概率,避免了所需的大量统计计算工作[3]。

在主观 Bayes 方法中,知识是用产生式规则表示的,具体形式为:

IF E THEN(LS,LN) H ($P(H)$)

LS,LN 在上文已有定义,$P(H)$ 是专家给出的先验概率。推理就由 $P(H)$,$P(E)$,LS 和 LN 求出 $P(H\mid E)$ 或 $P(H\mid -E)$ 的过程。

主观贝叶斯方法是最早用于处理不精确推理的模型,它以概率论中的贝叶斯公式为基础。贝叶斯公式描述如下:设有事件 $B_1,B_2,\cdots,B_n$ 互不相容,$B_1\cup B_2\cup\cdots\cup B_n=\Omega$(全集),事件 A 能且只能与 $B_1,B_2,\cdots,B_n$ 中的一个同时发生,而且 $P(A)>0$,$P(B_i)>0$,$i=1,2,\cdots,n$,则有

$$P(B_i/A)=\frac{P(A/B_i)P(B_i)}{\sum_{j=1}^{n}P(A/B_j)P(B_j)}\quad i=1,2,\cdots,n$$

上式中 $P(B_i)$ 是事件 B_i 的先验概率,先验概率是在不考虑任何证据的情况下专家凭经验给出的,$P(A\mid B_i)$ 是在事件 B_i 发生条件下事件 A 的条件概率,$P(B_i\mid A)$ 是事件 A 发生条件下 B_i 的条件概率。贝叶斯公式的意义在于将 $P(B_i\mid A)$ 的概率计算转化为对 $P(A\mid B_i)$ 和 $P(B_i)$ 的计算[4]。

依据贝叶斯公式的计算方法直接、简单。但是,该公式的使用要求事件 $B_1,B_2,\cdots,B_n$ 互不相容,并且需要计算 $P(A\mid B_i)$ 和 $P(B_i)$,直接应用贝叶斯公式求解问题是困难的,因为必须知道 B_i 的先验概率和证据 A 出现的条件概率。主观贝叶斯方法在贝叶斯公式基础上确定了不确定性推理的模型,并具有实际意义。

在主观贝叶斯方法中,知识采用产生式表示法表示。其具体形式如下:IF A THEN (LS, LN)B。其中,A 是知识的前提,B 是结论。另外,为了度量知识的不确定性,引入了两个数值(LS , LN) 表示知识规则强度, LS 为规则成立的充分性,体现了前提 A 的成立对结论 B 的支持度; LN 为规则成立的必要性,体现了前提 A 的不成立对结论 B 的支持度。LS 和 LN 的具体定义如下:

$$LS=\frac{P(A/B)}{P(A/\sim B)}\quad LN=\frac{P(\sim A/B)}{P(\sim A/\sim B)}$$

为了方便后面的叙述,在这里建立几率函数 $O(x)$,它和概率 $P(x)$ 的关系为

$$O(x)=\frac{P(x)}{1-P(x)}\tag{1}$$

该函数体现的是 x 出现的概率与不出现的概率之比。显然 $O(x)$ 与 $P(x)$ 单调性一致,若 $P(x_1)>P(x_2)$,则 $O(x_1)>O(x_2)$。因为 $P(x)$ 的值域为[0,1],由此可知 $O(x)$ 的值域为[0,+∞)。根据 LS,LN 的定义,以及 $O(x)$ 和 $P(x)$ 的关系,可以推出

$$O(B/A)=LS\cdot O(B)\tag{2}$$

$$O(B/\sim A)=LN\cdot O(B)\tag{3}$$

由式(1) 代入式(2),可得

$$P(B/A)=\frac{LS\cdot P(B)}{(LS-1)\cdot P(B)+1}\tag{4}$$

同理,由式(1) 代入式(3) ,可得

$$P(B/\sim A)=\frac{LN\cdot P(B)}{(LN-1)\cdot P(B)+1}\tag{5}$$

式(4) 为证据 A 肯定为真时,将 B 的先验概率更新为其后验概率的公式;式(5) 为证据 A 肯定为假时,将 B 的先验概率更新为其后验概率的公式。

因为在实际应用中，LS 和 LN 的值均由领域专家根据经验给出，所以，进行不确定性推理时，只需知道 $P(B_i)$ 的值，就可以求得 $P(B_i \mid A)$，从而绕开对 $P(A \mid B_i)$ 的求解。

领域专家在为 LS 和 LN 赋值时，可依据 LS 和 LN 的性质。例如 LS 体现前提的成立对结论的支持度，由 LS 定义可知：当 $LS > 1$ 时，前提支持结论；当 $LS = 1$ 时，前提不影响结论；当 $LS < 1$ 时，前提不支持结论。LN 体现前提的不成立对结论的支持度，其性质可类推。由此，当前提越支持结论时，推理网络系统中 LS 的值就越大[4]。

2 点击流分析和 Bayes 应用

随着电子商务的发展，越来越多的交易在网上进行，网上购物已逐渐成为一种普遍的生活模式，因此各商务网站之间也进行着激烈的竞争。为了吸引公众的眼球，扩大网站的影响力，网站提供的信息价值和网站性能是网站竞争力的主要体现。与传统商业模式不同，在互联网上，Web 用户与网站信息提供者之间不存在直接的信息沟通和反馈渠道。所以，要了解和把握什么样的信息最受用户欢迎？不同用户对信息有什么不同的需求？增加或减少信息服务内容对用户点击量的影响如何？这些是网站经营者需要回答和了解的问题，也是经营者今后为网站发展进行决策的依据。对这些信息和数据的把握主要来自于用户登录网站所产生的一系列点击流数据，对这些数据进行分析研究，就可以把握用户的一些相关性信息，从而为网站建设提供决策支持。

点击流是指用户在网站访问的过程中所留下的行为踪迹。点击流数据(Clickstream Data)是 Web 服务器上一系列有序的日志记录，它不仅包括用户浏览的每个站点、每个页面，在页面上滞留的时间以及点击的链接和图片，还包括浏览页面的顺序以及用户参与的新闻组和收发邮件等信息，这些信息都被顺序地记录在网站的日志文件中[5]。

从日志文件中提取出信息如用户的来源、行为、兴趣等，应用 Bayes 方法进行不确定性推理，然后再对这些数据进行深层次的分析，可以把握网站用户的主要行为特点和行为路径，比如对某些网页可能疯狂点击而对其他可能不管不问，对某些网页可能长时间停留而对有些可能一点就过，进而掌握用户对网站的内容的关注程度和用户的兴趣所在，把握用户的上网习惯和喜好，为网站建设者优化网站的网页布局，优化网站建设，提高网页内容的针对性、交互性，吸引力提供有价值的数据。这样不仅能够增强网站的黏着度，而且也能够提高网站的点击率，更好地建设网站，吸引商业人士来关注网站，为网站投资献策。主观 Bayes 方法在点击流分析系统中的应用实例如下：

在网站的点击流数据分析系统中有下列规则：用户对当前页面感兴趣事件(设为 B) 的先验概率 $P(B) = 0.03$；如果用户在当前页面的停留时间大于 20 秒而且小于 600 秒(设为 A_1)，则认为用户对当前页面感兴趣，(LS_1, LN_1) 为 (12,1)，产生式表示为：IF A_1 THEN (12,1)B；如果用户点击当前页面的图片连接超过 3 次(设为 A_2)，则认为用户对当前页面感兴趣，(LS_2, LN_2) 为 (23,1)，产生式表示为：IF A_2 THEN (23,1)B。如果用户点击当前页面的文字链接超过 3 次(设为 A_3)，则认为用户对当前页面感兴趣，(LS_3, LN_3) 为 (76,1)，产生式表示为：IF A_3 THEN (76,1)B。当证据 A_1, A_2, A_3 必然发生，求用户对当前页面感兴趣 B 的概率。

解：逐步更新结论的后验概率

由题意可知 B 的先验概率，规则 A_1, A_2, A_3 必然发生，又因为 $LS > 1, LN = 1, \sim A_1$ 对结论 B 没有影响，所以直接引用式(4)：

$$P(B/A_1) = \frac{LS_1 \cdot P(B)}{(LS_1 - 1) \cdot P(B) + 1} = \frac{12 \times 0.03}{11 \times 0.03 + 1} = 0.2707$$

这一结果说明，证据 A_1 用户在当前页面的停留时间大于 20 秒小于 600 秒时，使得结论发生(认为用户对当前页面感兴趣) 的概率由 0.03 增加到 0.2707。

$$P(B/A_1A_2) = \frac{LS_2 \cdot P(B/A_1)}{(LS_2 - 1) \cdot P(B/A_1) + 1} = \frac{23 \times 0.2707}{22 \times 0.2707 + 1} = 0.8951$$

在证据 A_1 发生的基础上，证据 A_2 用户点击当前页面的图片连接超过 3 次也发生了，使用户对当前页面感兴趣的概率由 0.2707 增加到 0.8951。

$$P(B/A_1A_2A_3) = \frac{LS_3 \cdot P(B/A_1A_2)}{(LS_3 - 1) \cdot P(B/A_1A_2) + 1} = \frac{76 \times 0.8951}{75 \times 0.8951 + 1} = 0.9985$$

计算表明，证据 A_1, A_2, A_3 的发生，最终认为用户对当前页面感兴趣事件发生的概率增加到 0.9985。

通过应用主观贝叶斯方法和其他一些点击流数据的在线分析处理和数据挖掘系统，可以发掘有价值的行为模式规范，进而分析把握用户的行为喜好和上网习惯，确定他们对哪些页面的哪些信息和商品感兴趣，就可以提供有针对性和个性化的商业服务，提高商务网站的点击率，从而取得更大的商业利益。

(下转第 121 页)

施构成。

* 目标函数记为 $g(x,\tilde{t}) = \max_i \sum_{j\in x^i} \tilde{t}_j$。

* 代理 i 的值记为 $v^i(x,\tilde{t}) = -\sum_{j\in x^i} \tilde{t}_j$。

* 配置函数记为 $x(\hat{t}) = x^i(\hat{t}),\cdots,x^n(\hat{t})$,支付函数记为 $p(\hat{t},\tilde{t}) = p^1(\hat{t},\tilde{t}),\cdots,p^n(\hat{t},\tilde{t})$。

更进一步,给出一个具体的验证机制 —— 补偿 - 红利机制。它由最优配置算法和支付函数构成。支付函数是两项之和 —— 补偿和红利。给一代理的补偿表示为其值的相反数,即 $c^i(\hat{t},\tilde{t}) = \sum_{j\in x^i(i)} \tilde{t}_j$。这使得代理的效用等于其红利。代理的红利表示为 $b^i(\hat{t},\tilde{t}) = -g(x(\hat{t}),\mathrm{corr}^i(x(\hat{t}),\hat{t},\tilde{t}))$,其中 $\mathrm{corr}^i(x,\hat{t},\tilde{t})$ 表示代理 i 的相关时间向量,定义为

$$\mathrm{corr}^i(x,\hat{t},\tilde{t})_j = \begin{cases} \tilde{t}_j & j \in x^i \\ t_j^l & j \in x^l \text{ 且 } l \neq i \end{cases}$$

因此,对于代理 i,该向量包含 i 处理任务的实际处理次数和所有其它代理处理任务所宣布的次数。支付函数定义为 $p^i(\hat{t},\tilde{t}) = c^i(\hat{t},\tilde{t}) + b^i(\hat{t},\tilde{t})$。

定理4　补偿 - 红利机制是任务调度问题的严格真实实施。

事实上,由于代理的效用等于其红利,当他在最短时间内处理任务时,效用最大。又因为配置算法是最优的,所以找到了依赖于所宣布类型而时间花费最小化的配置。如果代理说谎,那么时间花费就增加。由此,宣布真实类型是唯一的最优策略[6]。如果所有的代理都执行最优策略,那么就获得了最佳可能的时间花费。

例如:

	j_1	j_2	j_3
A_1	10	30	45
A_2	100	60	100

考虑图中的类型矩阵[7]。首先假设两个代理是诚实的。这个例子中的最优分配是$\{\{j_1,j_2\},\{j_2\}\}$,时间花费是60,因此给予每个代理的红利是 - 60。考虑下面的情况:代理 1 试图"丢掉"j_3,宣称 t_3' 为 200。因此"最优"的时间花费减少到 100,结果每个代理的红利减少到 - 100。同样地,当代理 1 试图"赢得"j_2,宣称 t_2' 为 4,它的红利减少到 - 85。如果代理 1 是"懒惰的",在 100 个单位时间内处理任务,那么它的红利从 - 60 减少到 - 100。

4　结束语

任务调度问题的讨论强调了分布计算中的调度问题。在这些机制的讨论中,随机机制比确定机制好,模型延伸到验证机制也似乎是一个非常自然的延伸。未来的研究仍潜在着三个方向:一是对所提出的更严密方法进行复杂性研究;二是运用机制设计的其它概念,在分布计算中提出一些其它的问题;三是研究如何以分布方式实施机制。

参考文献:

[1] Nisan N, Ronen A. Algorithmic mechanism design[J]. Games and Economic Behavior, 2001, 35: 166 - 196.

[2] Mas - Colell, Whinston M D, Green J R. Microeconomic Theory[M]. Oxford: Oxford University Press, 1995.

[3] 迈尔林 R. 博弈论[M]. 北京:中国经济出版社,2001.

[4] 樊晓香,胡茂林. 基于 VGC 机制的最小支撑树问题研究[J]. 微机发展,2005,15(8):142 - 144.

[5] Nisan N. Algorithms for selfish agents[C]//In To appear in Proceedings of the 16th Symposium on Theoretical Aspects of Computer Science. Trier, Germany: [s. n.], 1999.

[6] 张维迎. 博弈论与信息经济学[M]. 上海:上海人民出版社,1996.

[7] 奥斯本 M J,鲁宾斯坦 R. 博弈论教程[M]. 北京:中国社会科学出版社,2000.

(上接第 118 页)

3　结束语

由于不确定性问题的研究具有现实意义,不确定性推理方法成为人工智能领域研究的重点课题。文中通过将主观贝叶斯应用到对网站 Web 日志的点击流数据分析系统中,说明了它在实际应用中的理论和研究价值,但由于其推理的复杂性,涉及推理方向、推理控制策略、不确定性知识的表示、不确定性的更新、结论的可信度等要素,在实际应用中还需根据领域问题的实际特点,不断进行深入研究,完善不确定性推理方法。

参考文献:

[1] 王万森. 人工智能原理及其应用[M]. 北京:电子工业出版社,2002:35 - 64.

[2] 余东峰,孙兆林. 基于贝叶斯网络不确定推理的研究[J]. 微型电脑应用,2004,20(8):6 - 8.

[3] 李　强,徐建政. 基于主观贝叶斯方法的电力系统故障诊断[J]. 电力系统自动化,2007,31(15):46 - 50.

[4] 饶　浩. 利用主观贝叶斯方法进行不确定性推理[J]. 韶关学院学报,2004,25(6):6 - 9.

[5] 蔡榆榕. 点击流分析技术在网上评教系统中的应用[J]. 实验室研究与探索,2006,25(12):1541 - 1542.