

# 一种新的 Bayes 网络条件概率学习方法<sup>\*</sup>

汪荣贵, 高 隽, 张佑生, 彭青松

(合肥工业大学计算机与信息学院, 安徽合肥 230009)

**摘要:**针对大规模 Bayes 网络的条件概率赋值问题, 提出一种学习方法. 首先使用类层次结构定义一种新的基于层次的 Bayes 网络模型, 用于表示大规模 Bayes 网络. 然后将训练数据集由单个数据表的形式转化成多表数据库, 其中每个数据库表对应一个 Bayes 网络模块. 在此基础上导出条件概率计算公式, 从每个数据库表中算出相应的 Bayes 网络模块的条件概率表, 由此实现对整个层次 Bayes 网络的概率赋值. 通过适当增加数据库表的数目来控制每个表中属性的个数, 保证计算的可行性. 将层次 Bayes 网络及计算公式用于解决图像中文本的自动检测与定位问题, 实验结果表明了它们的有效性.

**关键词:** Bayes 网络; 类层次结构; 层次 Bayes 网络; 机器学习; 文本检测

**中图分类号:** TP181      **文献标识码:** A

## 0 引言

20 世纪 80 年代末, 加州大学的 Pearl 将概率论与图论相结合, 提出一种称为 Bayes 网络 (Bayesian Network) 的概率图模型<sup>[1, 2]</sup>, 作为不确定性信息和知识的处理工具. 该模型遵循经典的 Bayes 概率理论, 使用统一的联合概率分布作为知识表示形式, 保证了知识库的一致性; 通过巧妙使用条件独立性, 有效解决了确定联合概率分布的组合爆炸难题, 在条件独立性与可计算性之间取得了比较合理的折衷. 使用 Bayes 网络, 可进行预报推理、诊断推理和混合推理, 知识表示的一致性有效提高了推理结论的合理性. 该模型语义清晰、易于理解, 近年来在故障诊断、决策支持、数据挖掘、模式识别和图像理解<sup>[3]</sup>等领域得到了较为成功的应用.

Bayes 网络学习就是通过综合先验的专家知识和训练数据特征, 构造网络结构或条件概率表. 具体地说, 就是寻找网络结构或条件概率表, 在某种测度标准下与训练数据达到最佳拟合<sup>[4, 5]</sup>. 通常使用启发式搜索算法求取近似的全局最优解, 测度标准一般取最大的后验概率或最小的描述长度. 为使 Bayes 网络具有良好的机器学习能力, 人们提出了许多有效的

<sup>\*</sup> 收稿日期: 2004-03-01; 修回日期: 2004-09-01

基金项目: 国家自然科学基金 (60175011, 60375011); 安徽省自然科学基金 (03042207); 安徽省优秀青年科技基金 (04042044).

作者简介: 汪荣贵, 男, 1966 年生, 副教授/博士. 研究方向: 智能信息处理、知识工程.

E-mail: wangrgui@mail. hf. ah. cn

学习算法,它们可大致分为两类:①基于经典的统计学的学习算法,如适合处理完备数据的极大似然算法、适合处理不完备数据的 EM 算法和梯度上升算法等;②基于 Bayes 统计学的学习算法,如适合处理完备数据的 Bayes 预测(条件期望估计法)算法、适合处理不完备数据的 Gibbs 算法等<sup>[6~8]</sup>。

然而,大规模 Bayes 网络的学习问题一直没有得到有效地解决<sup>[6]</sup>。Bayes 网络学习的计算复杂度依赖于网络结构,最坏情况下是 NP-hard 问题<sup>[9]</sup>。变量较多时,难以对网络结构的进行合理地评判,现有的学习算法不可行。用层次化方法降低问题的复杂性是研究 Bayes 网络的重要方法。人们据此提出了许多改进的 Bayes 网络模型,代表性的有 Geiger 的 Bayesian Multinets<sup>[10]</sup>、Frydenberg 的 Chain Graph (CG)<sup>[11]</sup>、XIANG 的 Multiply Sectioned Bayesian Networks (MSBNs)<sup>[12]</sup>、Laskey 的 Network Fragments<sup>[13]</sup>、Koller 的 Object Oriented Bayesian Networks (OOBN)<sup>[14]</sup>等。这些模型的有效性主要在于不同程度地降低了大规模 Bayes 网络知识表示的复杂度。在推理方面,这些模型的做法可大致分为两类:(1)首先证明层次概率模型等价于某个 Bayes 网络,然后使用传统的 Bayes 网络推理算法实现模型的推理机制,这种做法没有对 Bayes 网络推理的复杂度问题做任何改进;(2)将层次概率模型中每个 Bayes 网络模块看成是一个相对独立的随机函数(stochastic function),通过对每个 Bayes 网络模块的推理来实现整个层次概率模型的推理,这在一定程度上降低了 Bayes 网络推理的复杂度,其计算复杂度是模型中 Bayes 网络模块的数目的多项式函数,因此只要每个 Bayes 网络模块的规模适度,数目不多,这些模型的推理是可行的。缺点是任何一次推理都需要模型中每个 Bayes 网络模块都参与计算。在变量取值状态较多的情况下, Bayes 网络模块的数目一般很大,此时推理的计算量比较大,难以满足实时计算的要求。

本文针对大规模 Bayes 网络的概率赋值问题,提出一种学习方法。首先用层次化方法扩展 Bayes 网络结构,定义一种新的基于层次的 Bayes 网络模型,来表示大规模 Bayes 网络。然后在此基础上提出条件概率表的学习算法,实现对这种层次 Bayes 网络的概率赋值。最后将层次 Bayes 网络及其学习算法用于解决图像中文本的自动检测与定位问题,验证它们的有效性。

与已有层次 Bayes 网络模型不同的是:本文的层次模型可以有效降低大规模 Bayes 网络推理和学习的复杂度。(1)将训练数据集由单个数据表的形式转化成多表数据库,其中每个数据库表对应一个 Bayes 网络模块。从每个数据库表中算出相应的 Bayes 网络模块的条件概率表,由此实现对整个层次 Bayes 网络的概率赋值。只要适当控制每个表的属性数目,就可保证算法的可行性。(2)通过推广消元,实现层次 Bayes 网络的一种推理算法。可以根据实际需要调节该算法中的控制参数,来有效地控制算法的复杂度。

## 1 Bayes 网络及类层次结构

概率论使用概率分布处理不确定性问题。对于  $n$  元随机向量  $U = \{x_1, x_2, \dots, x_n\}$ ,若每个分量最多有  $r$  个可能的取值,则要确定  $r^n - 1$  个概率。这是一个 NP-hard 问题。因此,通过直接对每个基本事件计算或指派概率的方法来确定联合概率分布  $P(U)$  是不现实的。需要建立适当的概率模型,借助模型的结构和数量特征来确定  $P(U)$ 。Bayes 网络就是一种这样的概率图模型。

**定义 1**  $n$  元离散随机向量  $U = \{x_1, x_2, \dots, x_n\}$  的 Bayes 网络是一个二元组  $B = (B_s, B_p)$ . 其中,  $B_s = (U, V)$  是有向无环图, 称为网络结构,  $U$  为节点集,  $V$  是有向边的集合, 每条边表示两节点间直接的依赖关系, 依赖程度决定于条件概率.  $B_p = \{P(x_i | \text{par}(x_i)); x_i \in U\}$  是一组条件概率分布的集合.  $\text{par}(x_i)$  表示在  $B_s$  中  $x_i$  所有父节点的集合 (若没有父节点则  $\text{par}(x_i) = \emptyset$ );  $P(x_i | \text{par}(x_i))$  表示节点  $x_i$  在其父节点某一取值状态下的条件概率分布.

Bayes 网络模型由网络结构和条件概率分布两部分组成. 网络结构用于定性描述变量间的依赖关系. 它是一个有向无环图, 图中每个节点表示问题领域中某一随机变量, 每条边表示节点间可能存在直接的概率依赖关系, 两节点间没有边连结则表示两节点间没有直接的概率依赖关系. 条件概率分布定量描述节点对其父节点的概率依赖程度. 对于网络结构中每个节点, 若它没有父节点, 则定义一个边际概率分布, 若它有父节点, 则定义一个条件概率表, 表中每一行为该节点在其父节点某一取值状态下的条件概率分布  $P(x_i | \text{par}(x_i))$ . Bayes 网络的基本思想是借助网络结构中所蕴含的变量之间独立性或条件独立性, 将联合概率分布分解为一系列边际概率和条件概率的乘积, 把问题转化为对边际概率和条件概率的确定, Bayes 网络的结构特征与变量间的条件独立性之间满足马尔可夫条件<sup>[2]</sup>, 即任一变量在已知其父节点取值状态的条件下, 独立于它的所有非子孙节点. 由链式法则, 有

$$P(U) = \prod_{i=1}^n P(X_i | X_1, X_2, \dots, X_{i-1}) = \prod_{i=1}^n P(X_i | \text{par}(X_i)) \quad (1)$$

Bayes 网络的构造一般分为三个步骤: (1) 根据问题域的特点及经验, 确定网络节点及其取值状态; (2) 根据经验或学习确定网络结构; (3) 根据经验或学习为每个节点确定条件概率表. 将 Bayes 网络作为知识库和推理引擎, 就可生成基于 Bayes 网络的智能系统. 机器学习是构造 Bayes 网络的重要途径, 然而对于大规模 Bayes 网络, 还没有有效的学习方法. 事实上, 大规模 Bayes 网络难以学习的根本原因在于它没有充分使用问题域中的结构信息. 它只使用了系统中变量之间的独立、依赖等结构信息, 忽视了系统中各个子系统或对象实体之间的结构信息. 为此, 本文用层次化方法推广 Bayes 网络模型, 利用层次关系中所蕴含的条件独立性, 降低 Bayes 网络学习的复杂度. 首先定义一种层次结构, 称之为类层次结构.

**定义 2** 称四元组  $\langle \mathcal{L}, \leq, \mathcal{A}, \mathcal{R} \rangle$  为类层次结构, 记为  $\mathcal{A}$ . 其中各要素定义为:

$\mathcal{L}$  是由有限个类组成的集合.  $\leq$  是  $\mathcal{L}$  上的一个偏序关系. 对于  $\mathcal{L}$  中的任意两个类  $C_1$  和  $C_2$ , 若  $C_1 \leq C_2$ , 则称  $C_1$  是  $C_2$  的子类, 或称  $C_2$  是  $C_1$  的超类. 对于满足  $C_1 \leq C_2$  的  $C_1$  和  $C_2$ , 若不存在  $C$  满足:  $C_1 \leq C \leq C_2$ , 则称  $C_1$  是  $C_2$  的相邻子类, 记为  $\text{sub}(C_2)$ , 或称  $C_2$  是  $C_1$  的相邻超类, 记为  $\text{sup}(C_1)$ . 规定若某个类  $C$  同时为某两个类  $C_1$  和  $C_2$  的子类, 则必有  $C_1 \leq C_2$  或  $C_2 \leq C_1$ .  $\mathcal{L}$  中存在惟一的一个类, 称之为顶层类,  $\mathcal{L}$  中所有的类都是它的子类, 记为  $C_T$ .

$\mathcal{A}$  是由一些变量组成的集合, 其中的每个变量  $A$  都有一个定义域  $C(A)$  和一个值域  $V(A)$ ,  $C(A)$  是  $\mathcal{L}$  中的某个类  $C$ ,  $V(A)$  是有限集. 称  $A$  为类  $C$  的简单属性.  $\mathcal{R}$  是由一些变量组成的集合, 其中的每个变量  $R$  都有一个定义域  $C(R)$  和一个值域  $V(R)$ ,  $C(R)$  是  $\mathcal{L}$  中的某个类  $C$ ,  $V(R)$  是  $\mathcal{L}$ . 称  $R$  为类  $C$  的复杂属性.

规定:  $C_T$  和所有没有子类的类只有一个复杂属性, 其余的类有两个复杂属性.

以上定义表明类库  $\mathcal{L}$  中每个类都有惟一确定的等级. 所有的类组成树结构: 每个类可以有多个相邻子类; 每个非顶层类只有一个相邻超类. 类的简单属性表示类自身的性质, 而

且  $\mathcal{A}$  中的每个属性不能同时属于两个类. 复杂属性表示类之间的联系, 用于类之间的信息传输. 下面在  $\mathcal{A}$  上定义的信息传输机制.

**定义 3** 设  $\mathcal{A} = \langle \mathcal{L}, \leq, \mathcal{A}, \mathcal{R} \rangle$  是一个类层次结构, 对于  $\mathcal{L}$  中任一类  $C$ , 若  $C = C_T$ , 则规定它的复杂属性的值域为  $C_T$  的所有相邻子类  $\text{sub}(C_T)$  组成的集合, 并称该属性为输出属性, 记为  $\text{output}(C_T)$ ; 若  $C$  没有子类, 则规定它的复杂属性的值域为  $\{C\}$ , 并称该属性为标识属性, 记为  $\text{id}(C)$ ; 在其余情况下, 将  $C$  的一个复杂属性的值域规定为所有  $\text{sub}(C)$  组成的集合, 并称该属性为输出属性, 记为  $\text{output}(C)$ , 将另一个复杂属性的值域规定为  $\{C\}$ , 并称该属性为标识属性, 也可称为输入属性. 将输入、输出属性统称为接口属性. 所有的接口属性组成的集合称为  $\mathcal{A}$  上的信息传输接口, 记为  $\mathcal{L}$ .

设  $\mathcal{A} = \langle \mathcal{L}, \leq, \mathcal{A}, \mathcal{R} \rangle$  是一个层次结构,  $\mathcal{L}$  是  $\mathcal{A}$  上的信息传输接口, 则可以使用如下方法构造一个具体的类层次模型:

(I) 将  $C_T$  实例化成一个对象, 称之为顶级对象  $O(C_T)$ ;

(II) 设  $O(C_T)$  的输出属性  $\text{output}(O(C_T))$  的取值为类  $C_1, C_2, \dots, C_n$ , 则分别将每个类例化成一个二级对象  $O(C_1), O(C_2), \dots, O(C_n)$ , 并且在  $O(C_i)$  的标识属性  $\text{id}(O(C_i))$  与  $\text{output}(O(C_T))$  的取值  $C_i$  之间建立一个数据绑定, 即在  $\text{output}(O(C_T))$ .  $C_i$  与  $\text{id}(O(C_i))$  之间用一有向边相连接, 箭头指向  $\text{id}(O(C_i))$ , 且规定  $\text{id}(O(C_i)) = \text{output}(O(C_T))$ .  $C_i$ ;

(III) 一般地, 对于每个存在输出属性的  $k$  级对象  $O$ , 设  $\text{output}(O)$  的取值为类  $C_1, C_2, \dots, C_n$ , 则分别将每个类实例化成一个  $k+1$  级对象:  $O(C_1), O(C_2), \dots, O(C_n)$ , 并且在  $O(C_i)$  的标识属性  $\text{id}(O(C_i))$  与  $\text{output}(O)$  的取值  $C_i$  之间建立一个数据绑定, 在  $\text{output}(O)$ .  $C_i$  与  $\text{id}(O(C_i))$  之间用一有向边相连接, 箭头指向  $\text{id}(O(C_i))$ , 且规定  $\text{id}(O(C_i)) = \text{output}(O)$ .  $C_i$ ,  $i=1, 2, \dots, n$ .

(IV) 所有对象组成的集合构成一个类层次模型.

类层次模型中对象的标识属性用于标识该对象所在的类, 并接收和保存来自相邻超对象的有关信息; 输出属性用于对该对象的类型作进一步的判别, 并向各个相邻子对象发送信息. 模型中的数据绑定有两个作用, 即用于表示对象之间的层次关系和传递 Bayes 网络块之间概率信息.

## 2 层次 Bayes 网络及其学习算法

在类层次结构的基础上定义一种基于层次的 Bayes 网络模型, 基本思想是为  $\mathcal{A}$  中每个类建立一个 Bayes 网络, 并使用  $\mathcal{L}$  将它们组织为层次结构. 具体做法如下:

(1) 对于顶级类  $C_T$ , 用它的输出属性, 以及所有与输出属性的取值有关的简单属性作为节点, 根据专家知识或通过学习构造一个 Bayes 网络模块, 使得输出属性是该模块的网络结构中惟一的根节点. 称该 Bayes 网络模块为由  $C_T$  生成的 Bayes 网络, 记为  $B(C_T)$ .

(2) 对于每个非顶级类  $C$ , 则将其标识属性的值域由  $\{C\}$  扩展为  $\{C, \text{其他类}\}$ .

(3) 对于每个具有输出属性的非顶级类  $C$ , 使用它的接口属性, 以及与输出属性取值有关的简单属性作为节点, 根据专家知识或通过学习构造一个满足如下三个条件的 Bayes 网络模块: (a) 标识属性是网络结构中惟一的根节点, 输出属性是标识属性惟一的子节点; (b) 简单属性节点是输出属性的后继节点; (c) 假设  $C$  的相邻超类  $\text{sup}(C)$  生成的 Bayes 网络模

块为  $B(\text{sup}(C))$ , 并且  $B(\text{sup}(C))$  的输出节点取值  $C$  的概率是  $p$ , 即有:  $P(\text{output}(\text{sup}(C))) = p$ , 则  $C$  的标识节点  $\text{id}(C)$  通过数据绑定机制接受这个概率. 此时  $\text{id}(C)$  的概率分布为:  $P(\text{id}(C)=C) = p; P(\text{id}(C)=\text{其他类}) = 1 - p$ . 称概率  $P(\text{id}(C)=C)$  为该网络的标识概率.

(4) 对于没有输出属性的非顶级类  $C$ , 仅由它的标识属性为节点, 生成一个只有一个节点的 Bayes 网络. 假设  $C$  的相邻超类  $\text{sup}(C)$  所生成的 Bayes 网络为  $B(\text{sup}(C))$ , 并且  $B(\text{sup}(C))$  的输出节点取值  $C$  的概率是  $p$ , 即有:  $P(\text{output}(\text{sup}(C))) = p$ , 则规定该节点的概率分布为:  $P(\text{id}(C)=C) = p; P(\text{id}(C)=\text{其他类}) = 1 - p$ .

**定义 4** 设  $\mathcal{A} = \langle \mathcal{L}, \leq, \mathcal{A}, \mathcal{R} \rangle$ ,  $\mathcal{L}$  是  $\mathcal{A}$  上的信息传输接口, 由上述方法构造的所有 Bayes 网络组成的集合称为 Bayes 网络库, 记为  $\mathcal{P}$ . 称三元组  $\mathcal{B} = \langle \mathcal{L}, \mathcal{A}, \mathcal{P} \rangle$  为层次 Bayes 网络,  $\mathcal{P}$  中的 Bayes 网络称为  $\mathcal{B}$  的 Bayes 网络模块.

在  $\mathcal{P}$  中, 可以将顶级类  $C_T$  生成的 Bayes 网络模块  $B(C_T)$  看成是一级 Bayes 网络模块. 由定义 1 可知,  $B(C_T)$  可以确定该网络所有节点的联合概率分布, 因而可以确定其输出节点的概率分布. 由  $C_T$  的每个相邻子类生成的 Bayes 网络模块可以看成是二级 Bayes 网络模块. 每个二级 Bayes 网络模块中标识节点取值的概率来自其相邻超类所生成 Bayes 网络模块, 即一级 Bayes 网络模块的输出节点取相应值的概率. 由此, 每个二级 Bayes 网络模块可以确定其所有节点的联合概率分布, 若它存在输出节点, 则可得到输出节点的概率分布. 此时, 该输出节点使用数据绑定, 将其取值的概率分布传给相应的相邻子类所生成三级 Bayes 网络模块的标识节点. 如此进行下去. 由此可知, 可以将  $\mathcal{P}$  中每个具有输出节点 Bayes 网络模块看成是一个函数. 函数的输出是输出节点的概率分布, 标识节点作为一个输入变量从上级 Bayes 网络模块的输出中接收有关的概率.

**命题 1** 设  $\langle \mathcal{L}, \mathcal{A}, \mathcal{P} \rangle$  是一个层次 Bayes 网络,  $B(C)$  是由类  $C$  生成的 Bayes 网络模块. 若知道  $\text{id}(C)$  的状态, 则  $B(C)$  中其余的任一节点变量独立于  $C$  的超类中的任一变量; 若知道  $\text{output}(C)$  的状态, 则  $B(C)$  中其余的任一节点变量独立于  $C$  的所有子类中的任一变量; 令  $R = (\text{id}(C), \text{output}(C))$ , 若知道  $R$  的状态, 则  $B(C)$  中其余的任一节点变量独立于  $\mathcal{L}$  中任一不属于  $C$  的变量.

**证明** (1) 由于  $B(C)$  不会向  $C$  的超类发送信息, 而且  $C$  的超类仅通过改变  $\text{id}(C)$  的概率分布影响  $C$  中变量的取值概率. 因此若已知  $\text{id}(C)$  的状态, 即  $\text{id}(C)$  的取值概率恒为 0 或 1, 则  $C$  的超类不能对  $C$  产生任何影响, 此时  $B(C)$  中任一其余节点独立于  $C$  的超类中的任一变量. (2) 由于  $B(C)$  仅通过  $\text{output}(C)$  影响它的所有生成的 Bayes 网络模块, 而且所有子类都不能影响它. 因此若已知  $\text{output}(C)$  的取值状态, 即  $\text{output}(C)$  的取值概率恒为 0 或 1, 则  $C$  就不能对其子类产生任何影响, 此时  $B(C)$  中任一其余变量独立于  $C$  的所有子类中的任一变量. (3) 结合(1)和(2)得证.

由命题 1 可知, 层次 Bayes 网络不仅通过使用 Bayes 网络模块保留了属于同一类变量之间的条件独立性, 而且充分使用了属于不同类的变量之间的条件独立性. 由此可以减少对复杂系统建模的难度. 更为重要的是这种性质为实现大规模 Bayes 网络的学习提供了一个有效的途径. 学习 Bayes 网络所使用的训练数据集通常是一个二维数据表, 表中的所有字段为学习 Bayes 网络的所有节点. 若节点数目比较多, 则表的规模庞大, 计算的复杂度较高,

而且存在数据冗余,影响计算的准确度.此时可以将所有的网络节点组成一种类层次结构,再将二维数据表根据这种类层次结构组织成一个数据库,每个类对应于一个数据库表,同时将类的输出属性作为数据库表的外键.这样就可以分别从每个数据库表中训练出相应的 Bayes 网络模块.将这些 Bayes 网络模块按类层次结构组织起来就构成了层次 Bayes 网络.此时只要适当增加数据库表的数目,就可以控制每个类中属性的个数,保证训练算法的可行性.

现在针对大规模 Bayes 网络的概率赋值问题,导出层次 Bayes 网络中条件概率表的学习算法.设  $\mathcal{B} = \langle \mathcal{L}, \mathcal{A}, \mathcal{P} \rangle$  是一个层次 Bayes 网络,网络结构为已知.  $B$  是  $\mathcal{B}$  的任一 Bayes 网络模块,需要学习的是  $B$  中各个节点的条件概率表.若  $B$  是由顶层类  $C_T$  生成的 Bayes 网络  $B(C_T)$ ,此时  $B$  与通常的 Bayes 网络没有区别,可以按通常的方法(如条件期望估计法)学习条件概率表.若  $B$  是由非顶层类生成的 Bayes 网络模块,则该模块中根节点的概率由  $B$  的上一层 Bayes 网络块决定,无需学习.只需考察该模块所有非根节点的学习算法.

设非根节点  $X$  的状态为  $x_1, x_2, \dots, x_r$ , 其父节点集  $\text{par}(X)$  的状态为  $\text{par}(X)$ . 对于  $\text{par}(X)$  的任一状态  $pr_1, pr_2, \dots, pr_s$ . 对于  $\text{par}(X)$  的任一状态  $pr_j$ , 令  $\theta_{ij} = P(X=x_i | \text{par}(X) = pr_j)$ ,  $\Theta_j = (\theta_1, \theta_2, \dots, \theta_r)$ . 当  $\Theta_j$  未知时,假设它服从参数为  $(\alpha_{1j}, \alpha_{2j}, \dots, \alpha_{rj})$  的狄利克莱(Dirichlet)分布<sup>[15]</sup>, 即有

$$P(\Theta_j | \alpha_{1j}, \alpha_{2j}, \dots, \alpha_{rj}) = \frac{\Gamma(\alpha_j)}{\prod_{k=1}^r \Gamma(\alpha_{kj})} \prod_{k=1}^r \theta_{kj}^{\alpha_{kj}-1} \quad (2)$$

设  $D$  是训练数据库中对应于 Bayes 网络模块  $B$  的数据库表. 假定  $D$  中的数据为简单随机样本, 在  $\Theta_j$  为已知时, 样本服从参数为  $\Theta_j$  的多项分布, 即有

$$P(D | \Theta_j) = \prod_{k=1}^r \theta_{kj}^{N_{kj}} \quad (3)$$

其中,  $N_{kj}$  表示  $D$  中同时满足  $X=x_k, \text{par}(X)=pr_j$  的记录数.

此时由狄利克莱分布的共轭性知, 后验概率分布  $P(\Theta_j | D)$  亦服从狄利克莱分布, 参数为  $(\alpha_{1j} + N_{1j} + \alpha_{2j} + N_{2j}, \dots, \alpha_{rj} + N_{rj})$ . 根据狄利克莱分布的基本性质, 该分布的数学期望  $E(\Theta_j | D)$  为  $\Theta_j$  的最优无偏估计量<sup>[9]</sup>, 即有

$$\hat{\theta}_{kj} = E(\theta_{kj} | D) = \frac{\alpha_{kj} + N_{kj}}{\sum_{k=1}^r \alpha_{kj} + \sum_{k=1}^r N_{kj}} \quad (4)$$

可以根据(4)式估计条件概率表中所有的条件概率. 其中  $\alpha_{ij}$  值为先验知识, 根据经验确定. 在没有经验的情况下, 令  $\alpha_{kj} = 1$ , 此时(4)式简化为

$$\hat{\theta}_{kj} = E(\theta_{kj} | D) = \frac{1 + N_{kj}}{r + \sum_{k=1}^r N_{kj}} \quad (5)$$

### 3 在文本检测中的应用

基于变异直方图的文本检测与定位算法, 通过使用图像中的文本特征与垂直变异直方

图  $VCH^C$  的中凸台区的规则性、以及水平变异直方图  $VCH^R$  中凹谷区规则性之间的映射关系,实现对文本的快速定位<sup>[16~18]</sup>. 由于图像内容的复杂性及噪音等因素的影响,图像特征与文本目标之间映射关系往往是不确定的,在检测算法中引入 Bayes 网络进行概率分析,可以提高检测效果. 我们曾经使用手工方法建立一个具有 20 多个节点的 Bayes 网络进行概率分析,效果不错<sup>[19]</sup>. 然而使用手工方法确定条件概率表非常麻烦,调试时间很长. 用现有的学习方法(如条件数学期望方法)训练条件概率,计算量大、训练时间长,而且检测效果不佳.

现在使用本文的层次 Bayes 网络及学习算法进行概率分析. 根据检测算法,构造四个类:即  $C = \{\text{焦点区域类, 规则凸台类, 规则凹谷类, 有文本类}\}$ . 它们的偏序关系为:有文本类  $\leq$  规则凹谷类  $\leq$  规则凸台类  $\leq$  焦点区域类. 图 1 是根据需要设计的一个简化的 Bayes 网络库  $P$ . 其中  $A_0$ 、 $B_0$ 、 $D_0$  分别为输出节点;  $B$ 、 $D$ 、 $E$  分别为标识节点; 有文本类没有输出属性,仅用于描述文本的位置、灰度等特征,是一个单节点 Bayes 网络. 表 1 列出了所有简单属性节点对应的图像特征. 为节约篇幅,略去了对每个节点的解释. 节点  $A_0$  的概率分布需要根据经验指定,若没有经验就取均匀分布,其余节点(不包括  $B$ 、 $D$ 、 $E$ )的条件概率表由公式(5)计算,结果如表 2 所示. 训练库为 ICDAR2003<sup>[20]</sup> 提供的两百多幅图像.

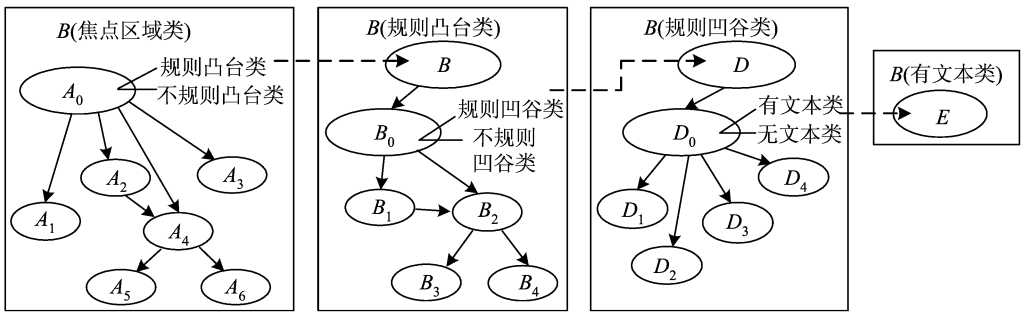


图 1 用于文本检测的 Bayes 网络库

Fig. 1 Database of Bayesian network structures for text detection

表 1 简单属性节点对应的图像特征

Tab. 1 Image characters corresponding to simple attribute nodes

$A_1$ 单元平均间距	$A_5$ 凸台区高度	$B_3$ 凹谷区高度	$D_3$ 凸台区中单元的长宽比
$A_2$ 单元密度	$A_6$ 凸台区宽度	$B_4$ 凹谷区宽度	$D_4$ 区域宽占图像行宽百分比
$A_3$ 单元的平均宽度	$B_1$ 凹谷区内边缘斜率	$D_1$ 凸台与凹谷高度比较	
$A_4$ 凸台区外形规则性	$B_2$ 凹谷区外形规则性	$D_2$ 区域的长宽比	

表 2 层次 Bayes 网络的条件概率表

Tab. 2 Some conditional probabilistic tables of hierarchical Bayesian network

	$A_0$	$A_2$	$A_4$		$B_0$	$B_1$	$B_2$	
			规则	不规则			规则	不规则
$D_0$	规则	正常	0.912 5	0.012 5	规则	正常	0.782 6	0.217 4
		不正常	0.432 7	0.567 3	规则	不正常	0.384 7	0.615 3
有文本	不规则	正常	0.394 2	0.405 8	不规则	正常	0.541 9	0.458 1
无文本	不规则	不正常	0.214 8	0.785 2	不规则	不正常	0.104 2	0.895 8

可以使用通过学习得到的层次 Bayes 网络对图像中文本进行检测与定位,图 2 是 1.6 GHz 微机上的运行的部分结果(取  $A_0$  为均匀分布),每幅图像检测时间一般仅为 10 s 左右.例如在第 1 幅图像中,取  $h=17$  个像素,检测到文本区的起止列为 62~241;起止行为 169~186.若只用一个 Bayes 网络进行检测,则所用的 Bayes 网络至少需要 18 个节点.实验表明,此时在与上述环境的相同条件下检测每幅图像至少需要 3 min.如表 3 所示,使用本文的层次 Bayes 网络进行本文检测,不仅便于学习,而且检测效果较好、速度快.其中  $P$  表示 Bayes 网络中诊断节点取值为“有文本”的后验概率.

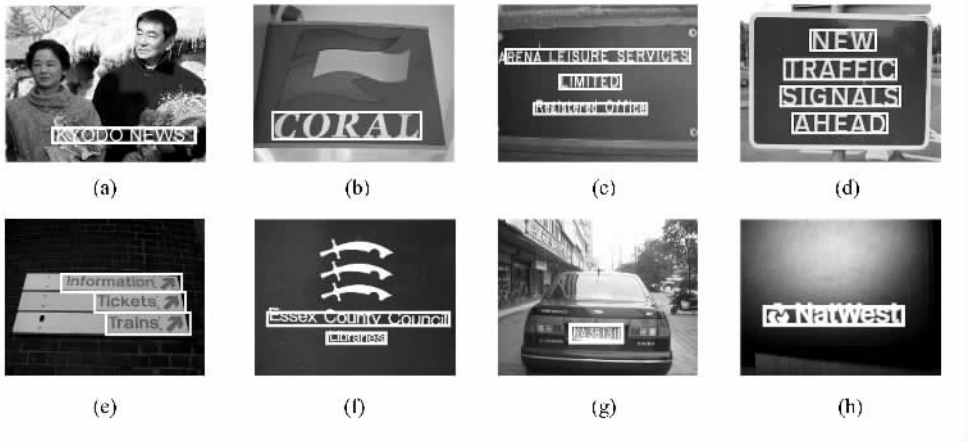


图 2 文本检测的实验结果

Fig. 2 The experimental results of text detection algorithm

表 3 两种方法的实验数据对比

Tab. 3 The comparison of experimental results for two solutions

图像编号		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	平均
概率 $P$	BN 方法	0.935	0.973	0.942	0.917	0.912	0.962	0.972	0.946	<b>0.945</b>
	本文方法	0.912	0.924	0.951	0.925	0.931	0.941	0.928	0.934	<b>0.931</b>
检测时间 (秒)	BN 方法	198.5	176.4	168.9	185.7	169.5	181.2	194.7	173.8	<b>181.1</b>
	本文方法	29.3	23.1	22.6	25.7	21.6	23.7	25.8	22.4	<b>24.3</b>
推理时间 (秒)	BN 方法	182	161.6	154.2	171.4	155.4	167.4	179.6	159.5	<b>166.4</b>
	本文方法	12.8	8.3	7.9	11.4	7.5	8.9	10.7	8.1	<b>9.5</b>

## 4 总结

为解决大规模 Bayes 网络的构造与学习等问题,本文提出了层次 Bayes 网络模型及相应的学习算法.文中首先定义一种层次结构,称之为类层次结构;然后使用这种结构来将大规模的 Bayes 网络分解成若干 Bayes 网络模块;最后用接口属性将这些模块组成一个层次结构,形成层次 Bayes 网络.这种模型通过利用层次关系中所蕴含的条件独立性,有效降低了学习大规模 Bayes 网络的复杂度.具体地说,对于这种层次 Bayes 网络,学习算法可以分



别从各个数据库表中学习相应的 Bayes 网络模块,只要适当控制每个表的属性数目,就可保证算法的可行性.

### 参 考 文 献

- [1] Pearl J. Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference [M]. San Mateo: Morgan Kaufmann, CA, 1988.
- [2] Jensen F V. Bayesian Networks and Decision Graphs [M]. New York: Springer-Verlag, 2001.
- [3] 汪荣贵,张佑生,高隽. 基于 Bayes 网络的航空图像理解模型[J]. 中国科学技术大学学报,2004,34(6):745-755.
- [4] Buntine W. A guide to the literature on learning probabilistic networks from data[J]. IEEE Trans. Knowledge Data Engrg. 1996, 8(2), 195-210.
- [5] Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data[J]. Machine Learning. 1995,20(3):197-243.
- [6] Neapolitan R E. Learning Bayesian Networks[M]. NJ. :Prentice Hall Upper Saddle River, 2003.
- [7] 汪荣贵,张佑生,彭青松. 分组样本下 Bayes 网络条件概率的学习算法[J]. 小型微型计算机系统, 2002,23:687-689.
- [8] 姚宏亮,王浩,胡学刚,汪荣贵. 基于遗传算法和 MDL 原则的 Bayes 网络结构优化算法[J]. 南京大学学报(自然科学版·计算机专刊),2002,38:23-27.
- [9] Chickering D M, Geiger D, Heckerman D E. Learning Bayesian networks is NP-Hard[R]. Microsoft Research Technical Report MSR-TR-94-17, 1994.
- [10] Geiger D, *et al.* Knowledge representation and inference in similarity networks and Bayesian multi-nets [J]. Artificial Intelligence, 1996,82:45-74.
- [11] Frydenberg M. The chain graph Markov property[J]. Scandinavian Journal of Statistics, 1990,17:333-353.
- [12] XIANG Y. Probabilistic Reasoning in Multi-agent Systems: A Graphical Models Approach[M]. London: Cambridge University Press, 2002.
- [13] Laskey K B, *et al.* Network fragments: representing knowledge for constructing probabilistic models[A]. Proc. of the 13th conference on uncertainty artificial in intelligence [C], Morgan Kaufmann, 1997,334-341.
- [14] Koller D, *et al.* Object-Oriented Bayesian networks[A], Proc. of the 13th conference on uncertainty artificial in intelligence [C], Morgan Kaufmann, 1997,302-313.
- [15] Geiger D, Heckerman D. A characterization of the dirichlet distribution applicable to learning Bayesian networks[R]. Technical Report, MSR-TR-94-16, Microsoft. Redmond, Wash. 1995.
- [16] PENG Qing-song, ZHANG You-sheng, WANG Rong-gui, *et al.* Detection caption using caption histograms[A]. Proceedings of Intel. Conf. on image and graphics [C], Hefei, China, 2002, SPIE. 4875:627-632.
- [17] 张佑生,彭青松,汪荣贵. 一种基于变异灰度直方图的视频字幕检测定位方法[J]. 电子学报,2004,32(2):314-317.
- [18] 张佑生,彭青松,汪荣贵. 一种新的数字图像灰度直方图及其应用研究[J]. 系统仿真学报,2002,14(12):1 655-1 658
- [19] 彭青松,张佑生,汪荣贵. 概率图模型在文本检测与定位中的应用研究[A]. 首届全国博士生学术论坛·计算机科学与技术分论坛论文集[C]. 北京:清华大学出版社,2003,10: 96-102.
- [20] <http://algoval.essex.ac.uk/icdar/Datasets.html>[DB/OL],2003.

# A New Approach to Learning Conditional Probabilities in Bayesian Networks

WANG Rong-gui, GAO Jun, ZHANG You-sheng, PENG Qing-song

*(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)*

**Abstract:** A learning approach is proposed to solve the problems of conditional probability assignment in large scale Bayesian networks. Firstly, a new hierarchical Bayesian Network model is defined based on class hierarchical structure, which is used to represent large scale Bayesian networks. Then, the train data set is changed from a single table to a database composed of some database tables. And each database table corresponds to a Bayesian network block. Based on that, a formula of conditional probability is developed. And each conditional probabilistic table of Bayesian network block can be calculated from the database tables respectively. Proper adjustment of the attribute number in each database table can assure the validity of this learning approach. Experiments in automatic detection and location of texts in images show the feasibility of this hierarchical Bayesian network and learning approach.

**Key words:** Bayesian networks; class hierarchical structure; hierarchical Bayesian network; machine learning; text detections