

概率图模型表示理论

刘建伟 黎海恩 罗雄麟

(中国石油大学(北京)自动化研究所 北京 102249)

摘要 概率图模型结合概率论与图论的知识,利用图来表示与模型有关的变量的联合概率分布。近年它已成为不确定性推理的研究热点,在人工智能、机器学习和计算机视觉等领域有广阔的应用前景。主要研究概率图模型的表示方法,讨论如何利用概率网络中的独立性来简化联合概率分布的方法表示。首先介绍了单个节点上的条件概率分布的表示模型及其引起的独立性,包括表格 CPD、确定性 CPD、特定上下文 CPD、因果影响 CPD、高斯模型和混合模型,并把单个分布模型推广到指数分布族中。然后详细介绍贝叶斯网络中的独立性以及图与概率分布的关系,讨论了高斯分布和指数分布族的贝叶斯网络表示理论。再详细描述马尔可夫网络的参数化问题及其独立性,也讨论高斯分布和指数分布族的马尔可夫网络表示理论。还给出两种局部有向图模型:条件随机场和链图。并且描述基于模板的概率模型表示,包括动态贝叶斯网络和状态观测模型这两种暂态模型,以及盘模型和概率关系模型这两种对象关系领域的有向概率模型,而且给出对象关系领域的无向表示。最后对概率图模型表示理论和方法所面临的问题及前景进行展望。

关键词 概率图模型,贝叶斯网络,马尔可夫网络,动态贝叶斯网络,概率关系模型,条件随机场,链图,指数分布族,局部概率模型

中图法分类号 TP181 文献标识码 A DOI 10.11896/j.issn.1002-137X.2014.09.001

Representation Theory of Probabilistic Graphical Models

LIU Jian-wei LI Hai-en LUO Xiong-lin

(Research Institute of Automation, China University of Petroleum, Beijing 102249, China)

Abstract Probabilistic graphical models bring together graph theory and probability theory in a formalism, so the joint probabilistic distribution of variables in the model can be represented using graph. In recent years, probabilistic graphical models have become the focus of the research in uncertainty inference, because of its bright prospect for the application in artificial intelligence, machine learning, computer vision and so on. This work introduced the representations of probabilistic graphical models and discussed how to represent the joint probabilistic distribution compactly using the independences in the network. First, models of the conditional probabilistic distribution of single node and their independences were introduced, including tabular CPD, deterministic CPD, context-specific CPD, CPD of causal influence, Gaussian models and hybrid models, and a general framework called the exponential family that encompasses a broad range of distributions was defined. Then Bayesian network representation and its independence properties were described in detail, as well as the Bayesian network of Gaussian distribution and exponential family. This work also introduced Markov network representation, independence properties in Markov network and Markov network of Gaussian distribution and exponential family. We also gave two partially directed models, conditional random fields and chain graph models. In addition, this work discussed template-based representations, including temporal models that contain dynamic Bayesian network and state-observation models, directed probabilistic models for object-relational domains which contain plate models and probabilistic relational models, and undirected representation for object-relational domains. Finally, this work raised some questions that probabilistic graphical models face with and discussed its development in the future.

Keywords Probabilistic graphical model, Bayesian network, Markov network, Dynamic bayesian network, Probabilistic relational model, Conditional random field, Chain graph, Exponential family, Local probabilistic model

1 引言

概率图模型把概率论与图论结合在一起,为多变量统计

建模提供有力的框架,它利用图来表达随机变量间的依赖关系。概率图模型由两部分组成:网络结构图和参数模型。

网络结构图 $G=(N,E)$ 由节点集 N 和边集 E 构成。 N 中

到稿日期:2013-10-15 返修日期:2013-12-13 本文受国家重点基础研究发展计划项目(973 计划)(2012CB720500),国家自然科学基金项目(21006127),中国石油大学(北京)基础学科研究基金项目(JCXX-2011-07)资助。

刘建伟(1966—),男,博士,副研究员,主要研究方向为智能信息处理,复杂系统分析、预测与控制,算法分析与设计,E-mail:liujw@cup.edu.cn;黎海恩(1988—),女,硕士生,主要研究方向为机器学习等;罗雄麟(1963—),男,博士,教授,主要研究方向为智能控制。

的节点表示随机变量集 $\mathcal{X} = \{X_1, \dots, X_n\}$ 中的变量, 变量集的取值为 $\mathbf{x} = \{x_1, \dots, x_n\}$ 。边集 E 表示两个变量间的概率关系, 可以是有向的或者无向的。图 1 给出随机变量集 $\mathbf{X} = \{X_1, X_2, X_3, X_4, X_5\}$ 上的一个网络结构图。

参数模型是网络结构图的概率分布表示。整个网络结构图的概率分布为联合概率分布, 即所有变量的联合概率分布。联合概率分布的显式表示难以得到且不利于推理。因此, 通常需要将概率图模型参数化, 即利用网络结构中的独立性, 把高维联合概率分布分解为低维概率分布的乘积。例如, 利用贝叶斯链式法则, 图 1 的联合概率分布可简单分解为

$$\begin{aligned} P(X_1, X_2, X_3, X_4, X_5) \\ = P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2)P(X_4 | X_1, X_2, X_3) \\ P(X_5 | X_1, X_2, X_3, X_4) \end{aligned} \quad (1)$$

贝叶斯网络中隐含着局部条件独立假设 $I_e(G)$, 即已知父节点时节点条件独立于其非子节点, 那么利用该独立性, 联合概率分布可进一步简化为

$$\begin{aligned} P(X_1, X_2, X_3, X_4, X_5) \\ = P(X_1)P(X_2)P(X_3 | X_1, X_2)P(X_4 | X_2)P(X_5 | X_3) \end{aligned} \quad (2)$$

因此, 独立性在概率图模型的表示中十分重要, 可以简化联合概率分布的表示。讨论概率图模型的表示, 主要是讨论其网络结构中的独立性, 以及网络结构与概率分布之间的关系。

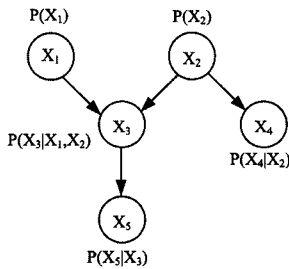


图 1 简单贝叶斯网络

根据边有无方向性, 概率图模型可以分为 3 类:

(1) 有向图模型

有向图模型, 也称贝叶斯网络 (Bayesian Network, BN)^[1], 其网络结构图使用有向无环图, 有向边表示节点间的有向条件依赖关系。其参数模型中, 需要为每个节点定义条件概率分布 (Conditional Probabilistic Distribution, CPD), 联合概率分布就是所有 CPD 的乘积。

(2) 无向图模型

无向图模型, 也称马尔可夫网络 (Markov Network, MN)^[2,3], 其网络结构图为无向图, 无向边表示相邻变量间的直接概率相互作用。其参数模型中, 需要为图中每个团定义势函数, 或者叫做因子 (Factor), 联合概率分布就是这些因子的乘积。

(3) 局部有向模型

局部有向模型, 即同时存在有向边和无向边的模型, 包括条件随机场 (Conditional Random Field, CRF)^[4] 和链图 (Chain Graph)^[5]。

根据表示的侧重点不同, 概率图模型又可分两类:

(1) 基于变量的概率图模型

基于变量的概率图模型确定某个固定随机变量集 $\mathbf{X}$ 的

联合概率分布, 如贝叶斯网络、马尔可夫网络、条件随机场和链图等。

(2) 基于模板的概率图模型^[6]

基于模板的概率图模型能够在结构丰富的、由多个不同类型且相互关联的对象组成的空间上定义联合概率分布, 它构建的单个紧凑模型为同一类型的整个分布类提供模板。这类模型根据应用场景不同又可分为两种: 暂态模型, 包括动态贝叶斯网络 (Dynamic Bayesian Network, DBN)^[7] 和状态观测模型; 对象关系领域的概率模型^[8], 包括盘模型 (Plate Model, PM) 和概率关系模型 (Probabilistic Relational Model, PRM)^[9,10]。

概率图模型的发展给不确定性推理的研究带来空前的变化。早在 1979 年 Duda 等人开发的 Prospector 专家系统, 以及 1984 年 Buchanan 等人开发的 Mycin 专家诊断系统就开始了进行不确定性推理的研究工作。但是这些体系使用产生式规则框架, 缺乏正式的语义, 在推理模式中也存在重大问题。另外, 人工智能领域研究者提出的模糊逻辑、可能性理论和 D-S 可信函数^[11,12], 也能较好地处理不确定性情况。然而, 在 Pearl 等人开发的贝叶斯网络框架^[13]下以及在进行有效推理的概率模型推动下^[14], 概率图模型在不确定性推理上显示出很大的优势。

贝叶斯网络是由路径分析 (Path Analysis)、因果模式 (Cause Pattern) 和影响图 (Influence Diagram)^[15] 等发展演变过来的。1988 年, Pearl 建立起贝叶斯网络体系, 给出许多基本概念, 如独立映射 (Independence-map, I-映射)、完美映射 (Perfect map, P-映射)、d-分离 (d-separation)、独立等价 (Independence-equivalence, I-等价) 等^[16,17]。后来这些基本理论得到不断发展和补充^[18-21]。早期的贝叶斯网络主要应用在医疗专家系统中, 如 David Heckerman 及其同事研究开发出的 Pathfinder 系统^[22,23]。作为一个诊断工具, 该系统的诊断效果超过了专业医师, 这极大推动了概率图方法的发展, 使得概率图模型在 20 世纪 90 年代成为研究的热点。目前, BN 逐渐成为人工智能^[24]、模式识别^[25] 和数据挖掘^[26] 等领域的研究热点。Pearl 指出贝叶斯网络是表示独立关系的最基本网络。还存在许多贝叶斯网络的变种, 它们使用非概率“参数化”的局部依赖模型, 即节点上的参数模型, 而不使用概率分布。这样的例子有逻辑微积分学^[27]、D-S 可信函数^[28]、可能性理论^[29]、定性数量级概率^[30], 以及区间约束概率^[31]。

贝叶斯网络的无环性约束, 限制了它对变量相互作用的表示能力。而基于无向图的马尔可夫网络能解决相互作用的表示问题。Hammersley 和 Clifford 的研究工作解决了无向图表示和吉布斯分布之间的联系^[32], 使得马尔可夫网络的参数化得以实现。Pearl 和 Paz 深入研究 MN 的独立性, 推动了 MN 独立性研究的发展^[33,34]。MN 在计算机视觉中的应用包括图像处理^[35]、三维重构^[36] 和目标识别^[37] 等, 这也极大推动了无向图模型的研究。近年来, 还提出了可分解马尔可夫网络 (Decomposable Markov Network)^[38-40] 和马尔可夫逻辑网络 (Markov Logic Network)^[41,42] 等。

基于变量的表示不能处理暂态过程和多个对象复杂关系的问题。80 年代末出现的隐马尔可夫模型 (Hidden Markov Model, HMM)^[43] 和动态贝叶斯网络^[44] 解决了暂态过程的概率表示问题。后来的研究工作根据这两种模型开发出更丰富

的表示语言,如混合记忆马尔可夫模型(Mixed-memory Markov Model)^[45]、因式化HMM(Factorial HMM)^[46,47]、层次HMM(Hierarchical HMM)^[48,49]等。HMM与DBN之间有着密切的关系,在一定情况下HMM可用来表示DBN^[50-52]。对象关系领域的概率模型则能处理多个不同类型对象的复杂关系,它起源于90年代提出的基于知识的模型构建理论。对象关系领域的概率模型包括盘模型^[53]和概率关系模型^[54,55]。Heckerman等人还基于有依赖的PRM表示,定义了一种使PM和PRM相一致的表示语言^[56]。近几年的研究焦点主要是为结构不确定性建模,包括为关系不确定性和对象不确定性建模。其中,对象不确定性的对象身份推理已经在多个领域中应用,包括数据关联^[57,58]、自然语言文本中的共指消解^[59]以及引文匹配^[60,61]等。Milch等人定义了BLOG(贝叶斯逻辑)^[62,63]来为对象数量不确定性建模。

本文首先从单节点的局部概率模型入手,第2节先介绍节点变量的概率分布表示;第3节—第6节再从整体上讨论概率网络的性质,其中第3节说明贝叶斯网络的相关基础概念和在多种情况下的表示,第4节讨论马尔可夫网络的参数化、独立性以及其在多种情况下的表示,第5节分析条件随机场和链图这两种局部有向模型,第6节描述基于模板表示的概率图模型;最后给出总结与展望。

为了表示的统一性,先约定文中所使用的符号。由于节点表示随机变量,因此文中提到的节点也即随机变量。随机变量集为 $\mathcal{X}=\{X_1, \dots, X_n\}$,变量集的取值为 $x=\{x_1, \dots, x_n\}$ 。一般而言,节点为大写体,其取值为小写体。某个节点 X 的所有取值为其值域 $R(X)$ 。节点 X 的父节点为 F_X ,非子节点为 $\bar{C}_X$,邻结点为 N_X ,边界节点为 B_X 。离散概率分布用 P 表示,连续概率分布用 p 表示。满足某一独立性的分布,如满足 $(X \perp Y | Z)$ 的分布 P 表示为 $P |=(X \perp Y | Z)$ 。

2 概率图中节点变量的概率分布表示

概率图模型参数化是利用全局结构中的独立性把高维联合分布分解为低维CPD或因子的乘积。节点变量的CPD或因子的概率分布表示,称为局部概率分布表示。由于CPD比因子带有更多的约束条件,下面选择根据CPD来表示局部概率分布。

离散节点变量的概率分布表示,包括表格CPD、确定性CPD、上下文说明CPD和因果影响CPD。连续节点变量的概率分布表示,主要是高斯模型。同时带有离散节点变量和连续节点变量的概率分布表示,主要使用混合模型。前述局部概率分布都是单个分布的概率表示,而指数分布族可表示一类分布族的概率分布。下面详细介绍这些局部概率分布模型。

2.1 表格CPD

表格CPD,也称为条件概率表(Conditional Probability Table, CPT),它把条件概率分布 $P(X|F_X)$ 编码为一个表格。变量 X 和 F_X 的每个联合取值的条件概率分布都对应表格中的一项,表格中的每一项都不为负,并且 $\sum_{x \in R(X)} P(x|f_X)=1$ 。

几乎所有离散CPD都可以用表格CPD来表示,在推理算法中也可以很自然地使用表格CPD,因此它是贝叶斯网络的一个重要部分。但是表格CPD有两个显著的缺点:(1)当节点随机变量的值域为无穷时(节点为连续随机变量),不可

能在一个表格中存储所有CPD。即使节点为离散随机变量,描述表格CPD所需的参数个数是 X 和 F_X 联合取值的个数,即 $|R(F_X)| \cdot |R(X)|$,该参数个数随父节点个数的增加呈指数增加。(2)表格CPD忽略了CPD内部的结构,没有利用CPD内部结构的规律性来减少参数个数。

2.2 确定性CPD

2.2.1 表示

确定性CPD中变量 X 是其父节点 F_X 的确定性函数,即存在函数 $f:R(F_X) \mapsto R(X)$,使得

$$P(x|f_X)=\begin{cases} 1, & x=f(f_X) \\ 0, & \text{其他} \end{cases} \quad (3)$$

确定性CPD比表格CPD更加简洁,在计算机中占用更少空间。在很多领域的建模中都会用到确定性关系,而且确定性变量也有助于简化复杂模型的变量上下文依赖性。

2.2.2 独立性

定义1 令 G 为网络结构图, $X_1 \Leftrightarrow \dots \Leftrightarrow X_n$ 为 G 中的一条迹。令 Z 为观测变量的一个子集。已知 Z 时,如果满足下面的条件,迹 $X_1 \Leftrightarrow \dots \Leftrightarrow X_n$ 是活动的:

(a)当有一个v-结构 $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$ 时,即两个父节点指向同一子节点时, X_i 或它的一个子节点属于 Z 。

(b)此迹上没有其他节点属于 Z 。

定义2 令 X, Y, Z 为 G 中3个节点集。如果已知 Z 并任意节点 $X \in X$ 和 $Y \in Y$ 之间不存在活动迹,则称已知 Z 时 X 和 Y 是d-分离的,表示为 $d_G(X, Y | Z)$ 。用 $I(G)$ 表示与d-分离对应的全局独立集:

$$I(G)=\{(X \perp Y | Z); d_G(X, Y | Z)\} \quad (4)$$

定义3 令 X, Y, Z 为 G 中3个节点集,其中 Z 不仅包含所有观测变量,而且包含可由观测变量确定的所有变量。如果已知 Z 时任意节点 $X \in X$ 和 $Y \in Y$ 之间不存在活动迹,那么说已知 Z 时 X 和 Y 是确定性分离的。

定义4 令 X, Y, Z 为两两不相交变量集合, C 为某个变量集合(可能与 $X \cup Y \cup Z$ 重叠的集合),并令 $c \in R(C)$ 。如果

$$P(X|Y, Z, c)=P(X|Z, c) \quad (5)$$

其中, $P(Y, Z, c) > 0$,则称已知 Z 和上下文 c 时 X 和 Y 是特定上下文独立的(Context-specific Independency, CSI)^[64],表示为 $(X \perp_c Y | Z, c)$ 。

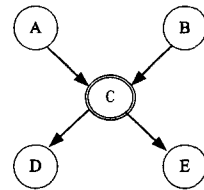


图2 带有确定性CPD的网络

考虑图2的例子, C 是 A 和 B 的确定性函数。已知 A 和 B 时, C 的值也确定了,因此 D 和 E 是独立的,即 $(D \perp E | A, B)$ 。但是,Pearl提出的d-分离概念并不能发现该独立性。Geiger等人把d-分离的概念扩展到带有确定性CPD的贝叶斯网络中,这称为确定性分离^[65]。令 G 为网络结构,并令 D, X, Y, Z 为变量集合。若已知 Z 时 X 与 Y 是确定性分离的,那么对所有 $P | = I_c(G)$ 的分布 P 来说,其中每个 $X \in D$ 时 $P(X|F_X)$ 都是确定性CPD,则有 $P |=(X \perp Y | Z)$ 。

只已知图结构和确定性CPD的集合时,只能发现当变量

为其父节点的确定性函数时的独立性。但是,某些特殊确定性函数可能隐含着其他的独立性,这类独立性只有当某些变量取特殊值时才存在,如 CSI 中的情况。

2.3 特定上下文 CPD

2.3.1 表示

在特定上下文 CPD 中,已知子集 $U \subseteq F_X$ 的特定取值为 u 时,其余父节点的取值与 CPD 是不相关的,即 f_X 取不同值时分布 $P(X|f_X)$ 是等价的。发现这种特定上下文独立的表示有两种:树 CPD 和规则 CPD。

(1) 树 CPD

Boutilier 等人提出使用完全划分的树 CPD 来编码上下文和参数之间的映射^[66]。为了避免混淆,与 BN 中使用的节点和边相对比,对树 CPD 使用 t-节点和箭头来描述。树 CPD 是一棵有根树,它表示变量 X 的 CPD。树中的 t-节点有两种:树叶 t-节点和内部 t-节点。树叶 t-节点表示分布 $P(X)$,内部 t-节点表示某个变量 $Z \in F_X$ 。每个内部 t-节点有一组向外的箭头指向其子节点,对 $z_i \in R(Z)$ 而言,每个箭头都与唯一赋值 $Z=z_i$ 相关联。

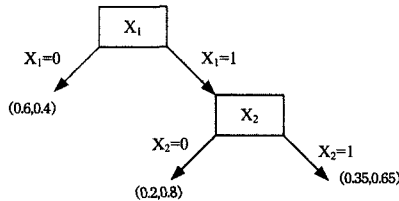


图3 $P(X_3 | X_1 X_2)$ 的树 CPD

例如,考虑前面图 1 所给出的 CPD $P(X_3 | X_1 X_2)$,所有变量都是二值变量。假设已知 $x_1=0$ 时 x_2 的取值与该 CPD 不相关,即 $P(X_3 | x_1=0, x_2=0)$ 与 $P(X_3 | x_1=0, x_2=1)$ 等价。那么 $P(X_3 | X_1 X_2)$ 的树 CPD 如图 3 所示,其内部 t-节点是 X_1 和 X_2 ,树叶 t-节点表示概率分布 $P(X_3)$ 。该树 CPD 能发现已知 $x_1=0$ 时 X_3 与 X_2 是特定上下文独立的。

(2) 规则 CPD

定义 5 规则 ρ 是一对 $\langle c; p \rangle$,其中 c 是变量集 C 的某个子集的赋值, $p \in [0, 1]$ 。定义 C 为 ρ 的定义域,表示为 $D[\rho]$ 。

树 CPD 是全局表示模型,能发现单个变量的完整 CPD。而规则 CPD^[67] 能实现特定上下文依赖的细粒度表示。每个规则对应 CPD 中的一项,说明该项所应用的上下文及其取值。如基于规则的 CPD $P(X|F_X)$ 是一组规则 R ,使得:

- 对每个规则 $\rho \in R$ 有 $D[\rho] \subseteq \{X\} \cup F_X$ 。
- 对 $\{X\} \cup F_X$ 的每个赋值 (x, f_X) ,有一个规则 $\langle c; p \rangle \in R$ 使得 c 与 (x, f_X) 相容,即 c 等价于 (x, f_X) 或者 c 是 (x, f_X) 的子集,表示为 $c \sim (x, f_X)$ 。此时有 $P(X=x | F_X=f_X) = p$ 。
- 生成的 CPD $P(X|F_X)$ 是一个合法 CPD,其中 $\sum_x P(x | f_X) = 1$ 。

前面所述的 CPD $P(X_3 | X_1 X_2)$ 也可以用一组规则来表示:

- $\rho_1: \langle x_1=0, x_3=0; 0.6 \rangle$
- $\rho_2: \langle x_1=0, x_3=1; 0.4 \rangle$
- $\rho_3: \langle x_1=1, x_2=0, x_3=0; 0.2 \rangle$
- $\rho_4: \langle x_1=1, x_2=0, x_3=1; 0.8 \rangle$
- $\rho_5: \langle x_1=1, x_2=1, x_3=0; 0.35 \rangle$
- $\rho_6: \langle x_1=1, x_2=1, x_3=1; 0.65 \rangle$

由规则 ρ_1 和 ρ_2 可以看出,已知 $x_1=0$ 时 X_3 与 X_2 是特定上下文独立的。

前面把规则看作 CPD 的分解,我们也可以把规则看作整个联合分布的细粒度分解。令 B 为贝叶斯网络,假设 B 中的每个 CPD $P(X|F_X)$ 都表示为一组规则 R_X 。令 R 定义为并集 $\bigcup_{X \in \mathcal{X}} R_X$,变量集 X 的任意取值 x 的概率计算如下:

$$P(x) = \prod_{\langle c; p \rangle \in R, x \sim c} p \quad (7)$$

2.3.2 独立性

定义 6 令 $\rho = \langle c; p \rangle$ 为规则, $C=c$ 为上下文。如果 c' 与 c 相容,则称 $\rho \sim c$ 。此时,令 $c'' = c' \setminus D[c'] - D[c]$,即 c'' 的定义域为 $D[c'] - D[c]$,而 $D[c'] - D[c]$ 中的变量与 c' 中该变量的取值一致。定义简化规则为 $\rho[c] = \langle c''; p \rangle$ 。对一组规则 R 来说,简化规则集合为 $R[c] = \{\rho[c]; \rho \in R, \rho \sim c\}$ 。

定义 7 令 $P(X|F_X)$ 为 CPD, $Y \in F_X$,并令 c 为上下文。如果 $P(X|F_X)$ 满足 $(X \perp_c Y | F_X - \{Y\}, c')$,其中 $c' = c \setminus F_X$,即 F_X 中变量的取值要与 c 中该变量的取值一致,则称已知上下文 c 时边 $Y \rightarrow X$ 是虚假的。

定义 8 令 X, Y, Z 为 G 中 3 个节点集, c 为上下文。如果消去已知上下文 c 时的所有虚假边之后,已知 Z 时 X 和 Y 还是 d-分离的,则称已知 Z 时 X 和 Y 是 CSI-分离的。

一般地,如果定义 c 为与 X 的树型 CPD 有关的上下文,那么已知上下文 c 时 X 独立于其他的父节点 $F_X - D[c]$ 。但是,还可能会存在其他 CSI 的表述。为了描述的一般化,下面采用规则 CPD 进行分析,以寻找所有 CSI 的表述。

先举例说明简化规则和虚假边的概念。如果把 CPD $P(X_3 | X_1 X_2)$ 的规则约束在上下文 $x_1=0$ 上,那么其简化规则集为:

- $\rho_1: \langle x_3=0; 0.6 \rangle$
- $\rho_2: \langle x_3=1; 0.4 \rangle$

只留下与 $x_1=0$ 相容的规则,并且把 $x_1=1$ 从所出现的规则中消去。由前文可知,已知 $x_1=0$ 时 X_3 与 X_2 是特定上下文独立的,因此,在 $x_1=0$ 时边 $X_2 \rightarrow X_3$ 是虚假的。

令 R 为规则 CPD 的规则, $R[c]$ 为 R 中与 c 相容的简化规则。令 $Y \subseteq F_X$ 为 X 的某个父节点子集,使得 $Y \cap D[c] = \emptyset$ 。如果对每个简化规则 $\rho \in R[c]$ 有 $Y \cap D[\rho] = \emptyset$,那么 $(X \perp_c Y | F_X - Y, c)$ 。但是,该检查过程并不完整,没有考察实际参数值,因此会漏掉对特定参数取值而言为真的其他独立;而且该过程只检验已知上下文和其他父节点时 X 和某个父节点间的独立性。为此, Boutilier 等人定义了 CSI-分离的概念,把 d-分离的概念扩展到包含特定上下文独立的网络中。令 G 为一个网络结构, P 为使得 $P \models I_t(G)$ 的一个分布, c 为上下文, X, Y, Z 为变量集。如果在上下文 c 中已知 Z 时 X 与 Y 是 CSI-分离的,那么 $P \models (X \perp_c Y | Z, c)$ 。

2.4 因果影响 CPD

因果影响 CPD 中变量 Y 的分布由 k 个原因 $X_1, \dots, X_k$ 的某个子集决定。 Y 以任意方式依赖于其父节点,父节点 X_i 以复杂方式相互作用,但其每个取值组合都与其他取值组合无关,即相互保持独立分布。每个原因 X_i 独立地影响 Y ,并且几个原因的影响能以某种方式简单组合。下面介绍因果影响的两种模型:噪声或模型 (Noisy Or Model) 和广义线性模型,然后再给出因果影响相互作用的一般定义。

2.4.1 噪声或模型

噪声或模型中^[68],原因之间的相互作用为逻辑或关系,并且每个原因带有噪声。令 Y 为二值随机变量,它有 k 个二值父节点 $X_1, \dots, X_k$ 。如果存在 $k+1$ 个噪声参数 $\lambda_0, \lambda_1, \dots, \lambda_k$ 使得

$$P(y=0|X_1, \dots, X_k) = (1-\lambda_0) \prod_{i: X_i=x_i^1} (1-\lambda_i) \quad (9)$$

$$P(y=1|X_1, \dots, X_k) = 1 - [(1-\lambda_0) \prod_{i: X_i=x_i^1} (1-\lambda_i)] \quad (10)$$

那么CPD $P(Y|X_1, \dots, X_k)$ 是一个噪声或模型,其中 λ_0 为泄露概率,表示无论任何情况 $y=1$ 总是发生的概率,即 $P(y=1)$; λ_i 为 $x_i=1$ 引起 $y=1$ 发生的概率。

2.4.2 广义线性模型

广义线性模型中,子节点 Y 的值由父节点 X_i 所决定,且两者保持线性函数关系。

令 Y 为可取 m 个值的变量,它有 k 个父节点 $X_1, \dots, X_k$ 。如果对每个 $j=1, \dots, m$ 都存在 $k+1$ 个权值 $\omega_{j,0}, \omega_{j,1}, \dots, \omega_{j,k}$,使得

$$\ell_j(X_1, \dots, X_k) = \omega_{j,0} + \sum_{i=1}^k \omega_{j,i} X_i \quad (11)$$

$$P(y=j|X_1, \dots, X_k) = \frac{\exp(\ell_j(X_1, \dots, X_k))}{\sum_{j=1}^m \exp(\ell_j(X_1, \dots, X_k))} \quad (12)$$

那么,CPD $P(Y|X_1, \dots, X_k)$ 是一个多类逻辑斯蒂模型。

2.4.3 因果影响相互作用的一般定义

定义 9 如果 $x \diamond y = y \diamond x$,则称确定性二值函数 $x \diamond y$ 是可交换的;如果 $(x \diamond y) \diamond z = x \diamond (y \diamond z)$,则称 $x \diamond y$ 是可结合的。如果存在一个可交换可结合的函数 $x \diamond y$ 使得 $f(x_1, \dots, x_k) = x_1 \diamond x_2 \diamond \dots \diamond x_k$,则称函数 $f(x_1, \dots, x_k)$ 是一个对称可分解的函数。

定义 10 如果条件分布 $P(Y|X_1, \dots, X_k)$ 的局部结构具有如图4所示的结构,则称 $P(Y|X_1, \dots, X_k)$ 具有对称的因果影响独立(Independency of Causal Influence, ICI)^[69],其中 Z 的CPD是一个确定性对称可分解函数 f 。若不同变量 Z_i 的CPD是相同的,则该CPD显示出完全对称的ICI属性。

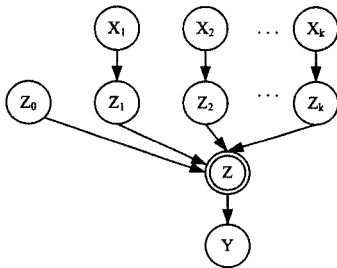


图4 因果影响独立

含有不同噪声模型 $P(Z_i|X_i)$ 和不同对称可分解函数的对称ICI模型的实例有很多。噪声或模型使用对称可分解的逻辑或函数和简单的带有二值变量的噪声模型。广义线性模型使用 Z_i 来生成 $\omega_i X_i$,然后相加后作为对称可分解函数 f ,最后的阈值效果在已知 Z 时 Y 的分布中体现。

2.5 高斯模型

前面的局部概率模型中变量是基于离散空间的。为连续变量建模的通用方法是把连续变量离散化,但是离散值会相当多,并且会损失两个离散值之间的结构特征。虽然高斯模型带有强假设,如远离均值时分布指数衰减,以及变量间存在

线性相互作用,但是其在数学上易于处理。而且根据中心极限定理,高斯模型能作为实际连续分布的一个较好近似。高斯模型能表示节点的连续分布。

定义 11 令 Y 为有连续父节点 $X_1, \dots, X_k$ 的连续变量。我们说 Y 服从线性高斯模型,如果存在参数 $\beta_0, \dots, \beta_k$ 和 σ^2 使得

$$p(Y|x_1, \dots, x_k) = N(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k; \sigma^2) \quad (13)$$

其向量表示为

$$p(Y|x) = N(\beta_0 + \beta^T x; \sigma^2) \quad (14)$$

2.6 混合模型

混合模型是同时含有离散和连续变量的模型。在此模型中,需要处理两种情况:带有连续和离散父节点的连续变量、带有连续和离散父节点的离散变量^[70]。

定义 12 令 X 为连续变量,并令 $U = \{U_1, \dots, U_m\}$ 为其离散父节点, $Y = \{Y_1, \dots, Y_k\}$ 为其连续父节点。如果对每个取值 $u \in R(U)$ 有 $k+1$ 个系数 $a_{u,0}, \dots, a_{u,k+1}$ 和一个方差 σ_u^2 使得

$$p(X|u, y) = N(a_{u,0} + \sum_{i=1}^k a_{u,i} y_i; \sigma_u^2) \quad (15)$$

则称 X 服从条件线性高斯模型(Conditional Linear Gaussian, CLG)CPD^[71]。

定义 13 如果每个离散变量只有离散父节点,并且每个连续变量有一个CLG的CPD,那么贝叶斯网络称为CLG网络^[72]。

首先考虑连续子节点 X 的情况,采用CLG模型。CLG模型不允许连续变量有离散子节点,它引入一个混合权平均高斯模型的概率分布。该混合形式对离散变量的每个取值包含一个高斯分量;分量的权值是取值的概率。因此,混合分量的个数在最坏情况下是离散变量个数的指数倍。

在带有连续父节点的离散子节点的情况下,一般使用阈值模型。例如,假设二值离散变量 U 有连续父节点 Y ,其阈值模型为

$$P\{u=1\} = \begin{cases} 0.8, & y \leq 80 \\ 0.05, & \text{其它} \end{cases} \quad (16)$$

但是在阈值模型中,概率的变化是 Y 的一个非连续函数,从数学角度来看是很不方便的,则可以利用逻辑斯蒂模型或者它的多项式展开来解决这个问题。

2.7 指数分布族

前面只关注于单个分布的表示,现在考虑指数分布族^[73]。直观上,分布族就是具有相同参数化形式、只在特定参数的选择上有所不同的分布集合。一旦选择了网络的全局结构和局部结构,就定义了一个分布族,它可由CPD的特殊参数选择来获得。

2.7.1 基本定义

定义 14 令 χ 为变量集。 χ 上的指数分布族的定义包括4部分:

- (a) 从 χ 的取值到 R^K 的充分统计量函数 $\tau: \chi \rightarrow R^K$;
- (b) 由凸集 $\theta \subseteq R^M$ 定义的参数空间;
- (c) 从 R^M 到 R^K 的自然参数函数 $t: R^M \rightarrow R^K$;
- (d) χ 上的测度 A 。

每个参数向量 $\theta \in \theta$ 说明了分布族中的一个分布 P_θ :

$$P_\theta(\xi) = \frac{1}{Z(\theta)} A(\xi) \exp\{\langle t(\theta), \tau(\xi) \rangle\} \quad (17)$$

其中, $\langle t(\theta), \tau(\xi) \rangle$ 是向量 $t(\theta)$ 和 $\tau(\xi)$ 的内积, $Z(\theta) = \sum_{\xi} A(\xi) \exp\{\langle t(\theta), \tau(\xi) \rangle\}$ 是 P 的有限划分函数。指数分布族 P 定义为: $P = \{P_{\theta}, \theta \in \Theta\}$ 。

指数分布族是有相似函数形式的概率分布类的简洁表示。分布族中的成员分布是由参数向量 θ 来确定的。充分统计量函数 τ 是对概率所有可能取值的总结。函数 t 把参数映射到充分统计量空间中。测度 A 给不依赖于参数的取值赋予权值。

实际中所遇到的几乎所有参数分布都可以表示成指数分布族, 包括泊松分布、指数分布、几何分布、伽马分布等。

通常可以构建编码多个相同分布类的指数分布族, 但是分布类的指数分布族表示需要满足以下要求: (1) 参数空间 θ 有较好的特性, 特别是要作为 R^M 的开凸集; (2) 指数分布族是非冗余的, 即参数的每个选择表示唯一分布, 也就是当 $\theta \neq \theta'$ 时有 $P_{\theta} \neq P_{\theta'}$ 。可以验证, 当且仅当函数 t 是可逆时, 分布族才是非冗余的。这样的指数分布族叫做可逆指数分布族。

2.7.2 线性指数分布族

线性指数分布族是指数分布族的一个特殊类型, 由函数 t 为恒等式函数的分布族组成, 这意味着参数的维数与数据表示的维数一样。对已知的充分统计量函数来说, 这样的参数也叫做自然参数。当使用自然参数时, 分布简化为

$$P_{\theta}(\xi) = \frac{1}{Z(\theta)} \exp\{\langle \theta, \tau(\xi) \rangle\} \quad (18)$$

参数空间 θ 需满足容许性条件: 容许自然参数集合 (自然参数空间) 为可被归一化的自然参数集合。

$$\theta = \{\theta \in R^K : \int \exp\{\langle \theta, \tau(\xi) \rangle\} d\xi < \infty\} \quad (19)$$

在有限离散空间的分布情况下, 所有参数选择都会产生归一化分布, 并且 $\theta = R^K$ 。自然参数空间是开凸集的指数分布族叫做线性指数分布族。

3 贝叶斯网络

BN 由两个部分构成: 网络结构图 G 和局部 CPD 集 P 。

(1) BN 的结构图 G 是一个有向无环图, 其节点表示随机变量 $X_1, \dots, X_n$, 边表示随机变量间的有向依赖。贝叶斯网络的结构图 G 隐含着局部条件独立假设: 对每个变量 X_i 有 $(X_i \perp \bar{C}_X | F_{X_i})$ 。这表明, 已知父节点时, X_i 与其非子节点条件独立。

(2) 贝叶斯网络的 CPD 集 P , 是网络结构图 G 中每个节点所对应的 CPD 的集合。 P 中的概率值表示子节点与其父节点间的关联程度, 没有父节点的节点概率则为其先验概率。

设 $X_1, \dots, X_n$ 为随机变量, 那么贝叶斯网络的联合概率分布可参数化地表示为

$$P(X_1, \dots, X_n) = \prod_i P(X_i | F_{X_i}) \quad (20)$$

结构图 G 与分布 P 之间的关系是通过独立映射和因式分解的概念来描述的。图 G 本身就隐含着局部条件独立性, 而 Pearl 又给出 d-分离的概念来定义图 G 中的全局独立性。两个不同的结构图只要满足相同的独立性, 那么这两个图就具有独立等价性。已知分布 P , 还可以构造反映其独立性的图 G 。讨论这些独立性及相关概念, 目的主要是使联合概率分布可简化表示。本节最后将介绍 BN 中的一种特殊网络, 即条件贝叶斯网络; 以及简单讨论高斯模型的 BN 和指数分

布族的 BN。

3.1 图 G 与分布 P 的关系

3.1.1 独立映射 (Independence-map, I-映射)

定义 15 令 P 为 χ 的分布, 定义 $I(P)$ 为 P 的独立断言集, 即在 P 中使 $(X \perp Y | Z)$ 成立的集合。

定义 16 令 K 为任意图对象, 其独立集为 $I(K)$, 如果 $I(K) \subseteq I$, 那么说 K 是独立集 I 的一个 I-映射。

如果 G 是 $I(P)$ 的 I-映射, 那么说 G 也是 P 的 I-映射。由定义 16 可看出, 当 G 是 P 的一个 I-映射时, G 中的任意独立性必须包含在 P 中, 而 P 可能含有没有反映在 G 中的额外独立性。

3.1.2 因式分解

定义 17 令 G 为变量 $X_1, \dots, X_n$ 的 BN 图, 那么相同空间上的分布 P 可以根据 G 来因式分解, 如果 P 能够表示为:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | F_{X_i}) \quad (21)$$

如果 G 是 P 的一个 I-映射, 那么 P 可以根据 G 来因式分解, 即 G 隐含的条件独立假设可以将分布 P 因式分解成小的 CPD。相反, 如果能根据结构图 G 来分解 P , 那么 G 是 P 的一个 I-映射。

3.2 图 G 中的独立

BN 利用独立性把高维联合概率分布分解为局部 CPD 的乘积以简洁表示。所以, 独立性在贝叶斯网络中尤其重要。前面提到, 图 G 隐含局部条件独立假设: 对每个变量 X_i 有 $(X_i \perp \bar{C}_X | F_{X_i})$, 该独立集表示为 $I_i(G)$ 。先介绍由 d-分离概念定义的全局独立性, 再引出两个图的独立等价概念。

3.2.1 d-分离

由定义 16 可知, P 可能含有没有反映在 G 中的额外独立性, 那根据 G 因式分解的每一个分布 P 是否存在其他独立集? Pearl 提出的 d-分离概念能从结构 G 的细节中说明 P 包含的独立性, 很好地解决该问题。

由定义 2 知道, d-分离概念给出网络的全局独立集:

$$I(G) = \{(X \perp Y | Z); d_G(X; Y | Z)\} \quad (22)$$

(1) d-分离的健全性

d-分离的健全性是指, 如果分布 P 根据 G 来因式分解, 那么 $I(G) \subseteq I(P)$ 。也就是说, P 满足 d-分离提出的所有全局独立性。

(2) d-分离的完备性

定义 18 如果当 $(X \perp Y | Z) \in I(P)$ 时有 $d_G(X; Y | Z)$, 那么分布 P 关于图 G 是可信的, 也就是 P 中的任意独立性都反映在图的 d-分离属性中。

d-分离的完备性^[74]是指, 如果 G 中已知 Z 时 X 和 Y 不是 d-分离的, 那么根据 G 因式分解的分布 P 中已知 Z 时 X 和 Y 是依赖的。

Meek 还给出 d-分离完备性的测度论概念^[75]: 对在图 G 上因式分解的几乎全部分布 P 来说, 也就是对除了 CPD 参数化空间中的零测度的集合之外的所有分布来说, 有 $I(P) = I(G)$ 。这个结论在两个方面强化了 d-分离的完备性: (1) 该观点揭示了存在一个分布关于图是可信的, 也就是图中的所有依赖关系同时成立于分布中; (2) 这个属性不仅适用于单个分布, 而且也适用于 G 上分解的几乎所有分布。

Shachter 提出了贝叶斯球算法^[76], 用于求 d-分离及其相

问题。

3.2.2 独立等价(Independence equivalence, I-等价)

定义 19 如果 $I(K_1) = I(K_2)$, 那么 χ 上的两个结构图 K_1 和 K_2 是 I-等价的^[77]。 χ 上所有图的集合被划分为一系列相互排斥并且穷举的 I-等价类, 这是由 I-等价关系诱导产生的等价类集合。

定义 20 χ 上 BN 图 G 的骨架是 χ 上的一个无向图, 若 G 包含一条有向边 (X, Y) , 那骨架包含一条无向边 $\{X, Y\}$ 。

定义 21 如果 X 和 Y 之间无有向边, 那么 v -结构 $X \rightarrow Z \leftarrow Y$ 是一个不道德集。若存在这样一条边, 它就叫做 v -结构的覆盖边。

d-分离的概念用结构 G 的细节来说明分布 P 的独立性, 该观点有一个重要的含义: 不同的 BN 结构可以完全等价, 只要它们编码了相同的独立假设集。

两个图的 I-等价表明, 根据其中一个图因式分解的任意分布 P 也可根据另一个图因式分解得到。判断两个图是否 I-等价的方法是: G_1 和 G_2 为 χ 上的两个图, 如果 G_1 和 G_2 有相同的骨架和相同的 v -结构集合, 那么它们是 I-等价的。但反向不成立。引入不道德集的概念, 可给出其反方向: G_1 和 G_2 为 χ 上的两个图, 只有在它们是 I-等价时, G_1 和 G_2 有相同的骨架和相同的不道德集。

根据覆盖边的方向, 可给出 I-等价的另一种描述方法。当且仅当存在网络图序列 $G = G_1, \dots, G_k = G'$, 并且序列中每一个图都有 I-等价于 G , 使得 G_i 和 G_{i+1} 的唯一差别是覆盖边的反向, 则称两个图 G 和 G' 是 I-等价的。

3.3 由分布集 P 来构造结构图 G

希望构造能反映概率分布 P 中独立性的图 G , 并通过检测 G 中的 d-分离性来检测 P 的独立性。即在某一程度上构建一个网络结构图 G , 使 G 的独立集是 P 的独立集的一个合理替代。但是, 以下的讨论并不会给出分布 P 的充分说明以及构造图 G 的实际过程, 只是从概念上解释其构造过程。构造过程依赖于两个重要概念: 最小 I-映射和完美映射 (Perfect Map, P-映射)。

3.3.1 最小 I-映射

定义 22 如果图 K 是独立集 I 的一个 I-映射, 并且移动 K 中的任一条边都会使它不再是 I-映射, 那么 K 是 I 的最小 I-映射。

构造表示分布 P 的图 G 的一个方法是, 构造分布 P 的任意一个最小 I-映射的图 G 。怎样获得分布 P 独立集的最小 I-映射? 假设已知某一变量序列 $\{X_1, \dots, X_n\}$, 依次检测每个变量 $X_i, i=1, \dots, n$ 。对每个 X_i , 选择 $\{X_1, \dots, X_n\}$ 中的某个最小子集 U 作为 X_i 在 G 中的父节点集, 要求 U 满足 $(X_i \perp \{X_1, \dots, X_{i-1}\} - U | U)$, 并且从 U 中任意去掉一个节点后将不再满足该性质。那么令 U 为 X_i 的父节点集。 U 的选择并不唯一。但是, 如果分布为正, 即对所有变量 χ 的任意取值 x 有 $P(x) > 0$, 那么给定变量序列时父节点集的选择是唯一的。

然而, G 是 P 的最小 I-映射时, 并不能保证 G 确实包含了 P 的独立性结构, 因而引出 P-映射概念。

3.3.2 P-映射

定义 23 若 $I(K) = I$, 则称图 K 是独立集 I 的一个 P-映射。若 $I(K) = I(P)$, 那么 K 是 P 的一个 P-映射。

当 G 是 P 的最小 I-映射时, 并不能保证 G 可以发现 P 的所有独立性, 而我们的目的是找出能精确发现分布 P 独立性的结构 G , 所以引入完美映射的概念。如果图 G 是分布 P 的 P-映射, 那么可以从 G 中直接读出 P 的所有独立性。

Pearl 和 Verma 以 I-等价类的形式提出构建分布 P 映射的第一个算法^[78]。该算法后来得到发展和补充^[79]。由于分布 P 可以有多个 I-等价的 P-映射, 因此算法需要返回整个等价类。前面说到, 两个图有相同骨架和相同不道德集, 那这两个图 I-等价。所以通过从已知分布 P 的独立性中辨识 P-映射 G^* 的骨架和不道德集, 就可以为 P-映射 G^* 构建 I-等价类。

3.4 条件贝叶斯网络

定义 24 已知 X 时 Y 上的条件贝叶斯网络 B 定义为一个有向无环图, 其节点为 $X \cup Y \cup Z$, 其中 X, Y 和 Z 互不相交。 X 中的变量称为输入变量, Y 中的变量称为输出变量, 而 Z 中的变量则称为压缩变量。图 G 中, X 中的变量没有父节点。 $Y \cup Z$ 中的变量具有 CPD。利用链式法则, 该网络定义一个条件概率分布:

$$P_B(Y, Z | X) = \prod_{X \in Y \cup Z} P(X | F_X) \quad (23)$$

分布 $P_B(Y | X)$ 定义为 $P_B(Y, Z | X)$ 的边缘分布:

$$P_B(Y | X) = \sum_Z P_B(Y, Z | X) \quad (24)$$

定义 25 令 Y 为有 k 个父节点 $X_1, \dots, X_k$ 的随机变量。如果 CPD $P(Y | X_1, \dots, X_k)$ 表示为已知 $X_1, \dots, X_k$ 时 Y 上的条件贝叶斯网络, 那么它就是一个压缩 CPD。

条件 BN 只是一个 BN 片段, 并不是完整的 BN。该片段没有说明片段中所有变量的联合概率分布, 只说明了某个变量的 CPD。因此, 条件 BN 可以用来说明节点上的 CPD, 可以作为第 2 节所描述的局部概率模型的表示^[80]。

条件 BN 的一个重要应用是定义压缩 CPD。压缩表示不仅可以使模型简化、更容易理解, 而且它在描述一个由低级子系统组成的复杂系统时, 只需为每个子系统单独地建模, 而不用考虑其低级成分的内部模型。

3.5 高斯模型的贝叶斯网络

定义 26 定义高斯模型的贝叶斯网络为, 全部变量是连续的、全部 CPD 是线性高斯模型的贝叶斯网络。

在高斯模型情况下的 BN 中^[81], G 为线性高斯模型网络, P 为联合多元高斯分布。先介绍多元高斯模型及其特性, 然后说明线性高斯模型网络和联合多元高斯分布之间的等价性。

3.5.1 基本参数化

$X_1, \dots, X_n$ 组成的多元高斯分布是一个具有 n 维均值向量 μ 和对称的 $n \times n$ 维协方差阵 Σ 的概率分布, 其密度函数定义为

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] \quad (25)$$

其中, $|\Sigma|$ 是 Σ 的行列式。矩阵 Σ 是正定的: 对任意 $x \in R^n$ 且 $x \neq 0$, 有 $x^T \Sigma x > 0$ 。正定矩阵 Σ 能保证是非奇异的, 因此它有非零行列式值。

因为正定矩阵是可逆的, 可以根据协方差阵的逆矩阵 $J = \Sigma^{-1}$ 来定义高斯模型, 该逆阵叫做信息阵。该表示方法引出高斯密度的另一种形式。由于

$$\begin{aligned}
-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) &= -\frac{1}{2}(x-\mu)^T J(x-\mu) \\
&= -\frac{1}{2} [x^T J x - 2x^T J \mu + \mu^T J \mu] \quad (26)
\end{aligned}$$

最后一项为常数,因此得到

$$p(x) \propto \exp[-\frac{1}{2}x^T J x + (J\mu)^T x] \quad (27)$$

该高斯密度的公式一般称为信息形式,向量 $h = J\mu$ 叫做势能向量。由于当且仅当 Σ^{-1} 是正定时 Σ 是正定的,因此当且仅当信息阵是对称正定时,信息形式才定义一个有效的高斯密度。

3.5.2 高斯模型的运算

对分布的操作主要有两种运算:计算变量集 Y 某个子集的边缘分布,以及计算已知 $Z=z$ 的某个取值时的条件分布。

(1)边缘分布的计算采用协方差形式,只需要计算两两协方差。若 $\{X, Y\}$ 的联合分布的均值和协方差作如下分解:

$$p(X, Y) = N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}; \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right) \quad (28)$$

其中, $\mu_X \in R^n$, $\mu_Y \in R^m$, Σ_{XX} 是 $n \times n$ 维矩阵, Σ_{XY} 是 $n \times m$ 维矩阵, $\Sigma_{YX} = \Sigma_{XY}^T$ 是 $m \times n$ 维矩阵, Σ_{YY} 是 $m \times m$ 维矩阵,那么 Y 的边缘分布是正态分布 $N(\mu_Y; \Sigma_{YY})$ 。

(2)条件分布计算采用信息形式,只需要对协方差阵求逆。把 $Z=z$ 的值代入信息形式中,该过程把某些二次项转换成线性项或者常数项,并把某些线性项转换成常数项。

3.5.3 高斯模型的独立性

对多元高斯模型来说,从分布的参数中可直接得到其独立性。令 $X = X_1, \dots, X_n$ 服从联合正态分布 $N(\mu; \Sigma)$, 那么当且仅当 $\Sigma_{i,j} = 0$ 时, X_i 与 X_j 独立。

考虑高斯分布 $p(X_1, \dots, X_n) = N(\mu; \Sigma)$, 并令 $J = \Sigma^{-1}$ 为信息阵。那么,当且仅当 $p \mid = (X_i \perp X_j \mid \chi - \{X_i, X_j\})$ 时, $J_{i,j} = 0$ 。信息阵发现了已知其它所有变量时两个变量间的独立性。

3.5.4 线性高斯模型网络和联合多元高斯分布之间的等价性

(1)线性高斯模型网络总是能定义联合多元高斯分布

令父节点为 $X_1, \dots, X_k$ 的变量 Y 的线性高斯模型为: $p(Y \mid x) = N(\beta_0 + \beta^T x; \sigma^2)$ 。假设 $X_1, \dots, X_k$ 具有联合高斯分布,其分布为 $N(\mu; \Sigma)$ 。那么:

(a) Y 的分布是正态分布 $p(Y) = N(\mu_Y; \sigma_Y^2)$, 其中 $\mu_Y = \beta_0 + \beta^T \mu$, $\sigma_Y^2 = \sigma^2 + \beta^T \Sigma \beta$;

(b) $\{X, Y\}$ 上的联合分布是正态分布,其中 $Cov[X_i; Y] = \sum_{j=1}^k \beta_j \Sigma_{i,j}$ 。

因此,如果 B 是线性高斯模型的贝叶斯网络,那么它具有联合高斯模型分布。

(2)由多元高斯分布能构造线性高斯模型网络

令 $\chi = \{X_1, \dots, X_n\}$, 并令 p 为 χ 的联合高斯分布。给定 χ 上 $X_1, \dots, X_n$ 的任意序列,可以构建一个贝叶斯网络图 G 和一个 G 上的贝叶斯网络 B ,使得

(a) $F_{X_i} \subseteq \{X_1, \dots, X_{i-1}\}$;

(b) B 中 X_i 的 CPD 是其父节点的线性高斯模型;

(c) G 是 p 的最小 I-映射。

3.6 贝叶斯网络的指数分布族

贝叶斯网络中,如果所有 CPD 都来自一个指数分布族,那么这些 CPD 的乘积也是在该指数分布族中。因此,带有指数 CPD 的贝叶斯网络定义了一个指数分布族。但是,要想得到贝叶斯网络的指数分布族表示,需要保证每个 CPD 都是局部归一化的。对指数 CPD 来说,把 τ 的维数扩大常数维就可以实现,那么 $t(\theta)$ 中相匹配的元素就能作为划分函数的对数。

一般来说,虽然带有适当 CPD 的贝叶斯网络定义了一个指数分布族,但是该分布族一般不是线性分布族。特别地,任意包含不道德集的网络都不能产生一个线性指数分布族。

4 马尔可夫网络

在为实际现象建模时,有许多现象变量之间的相互影响关系是没有方向性的,因此,无向图模型能更好地表示这类现象的概率模型。

与 BN 一样, MN 也由两部分构成:网络结构图 H 和因子集 Φ 。

(1) MN 的结构图 H 是无向图,其节点表示随机变量,其无向边表示相邻变量间的直接概率相互作用。若 H 的子图中任意两个节点都有边相连,那么称该子图为完全子图(complete subgraph)。完全子图的变量集叫做团(clique)。

(2) MN 的因子集 $\Phi = \{\phi_1(D_1), \dots, \phi_K(D_K)\}$, 也叫做势函数集(potential function)。因子并没有像 CPD 那样对应于条件概率,而是表示变量间的相互作用。下面只讨论非负实因子。

由于因子并不对应于条件概率,因此 MN 的参数化表示并不像 BN 的那样直观。需要把 MN 与吉布斯分布联系起来,建立吉布斯分布与因子乘积的等价关系,那么 MN 的联合概率分布可分解为因子的乘积。利用因子图和对数线性模型, MN 可进行更高细粒度的参数化。MN 中包含 3 种独立性^[33]:全局独立、两两独立和局部独立。通过分析其独立性,可以使 MN 的联合概率分布表示更加简洁。本节最后还会介绍高斯模型的马尔科夫随机场和 MN 的指数分布族表示形式。

4.1 参数化

4.1.1 吉布斯分布和马尔可夫网络

Hammersley 和 Clifford 首次发现马尔可夫随机场和吉布斯分布之间的联系:一个随机场是关于邻域系统的马尔可夫随机场,当且仅当该随机场是关于邻域系统的吉布斯分布。这个等价条件解决了 MN 的参数化问题,建立了吉布斯分布与因子乘积的等价关系。下面先介绍因子的概念以及因子的运算,再利用因子来定义吉布斯分布的参数化过程,并与图 H 联系起来。

(1)因子

定义 27 令 D 为随机变量集合,定义因子 ϕ 为从 $R(D)$ 到 R 的函数。如果因子所有的输入是非负的,那么这个因子也是非负的。变量集合 D 叫做因子的定义域,表示为 $D[\phi]$ 。

定义 28 令 X, Y, Z 为互不相交的变量集,并令 $\phi_1(X, Y)$ 和 $\phi_2(Y, Z)$ 为两个因子。定义因子乘积 $\phi_1 \times \phi_2$ 也为一个因子 $\phi_1 R(X, Y, Z) \mapsto R$, 表示如下:

$$\phi(X, Y, Z) = \phi_1(X, Y) \times \phi_2(Y, Z) \quad (29)$$

由定义 27 可知,联合概率分布和 CPD 也是因子。然而,

CPD 和联合概率分布必须满足归一化约束,即分布中数值之和必须为 1。

根据定义 28 给出的因子乘积的概念,下面可以定义吉布斯分布的无向参数化,即把吉布斯分布分解为因子的乘积。

(2) 吉布斯分布的参数化以及与图 H 的关系

定义 29 分布 P_Φ 是由因子集 $\Phi = \{\phi_1(\mathbf{D}_1), \dots, \phi_K(\mathbf{D}_K)\}$ 来参数化的吉布斯分布,其定义如下:

$$P_\Phi(X_1, \dots, X_n) = \frac{1}{Z} \tilde{P}_\Phi(X_1, \dots, X_n) \quad (30)$$

其中, $\tilde{P}_\Phi(X_1, \dots, X_n) = \phi_1(\mathbf{D}_1) \times \phi_2(\mathbf{D}_2) \times \dots \times \phi_m(\mathbf{D}_m)$ 是非归一化测度值,而 $Z = \sum_{X_1, \dots, X_n} \tilde{P}_\Phi(X_1, \dots, X_n)$ 是一个归一化常量,叫做划分函数。

定义 30 如果每个 $\mathbf{D}_k (k=1, \dots, K)$ 都是 H 的完全子图,则称分布 P_Φ 根据马尔可夫网络 H 分解,其中 $\Phi = \{\phi_1(\mathbf{D}_1), \dots, \phi_K(\mathbf{D}_K)\}$ 。

由于参数化过程把因子与完全子图的团联系起来,马尔可夫网络参数化的因子通常也叫做团势。因为每个完全子图都是某个团的一个子集,所以只考虑最大团的因子可以减少参数化中的因子个数。

因此, MN 的联合概率分布是一个吉布斯分布,其参数化可表示为:

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{k=1}^m \phi_k(\mathbf{D}_k) \quad (31)$$

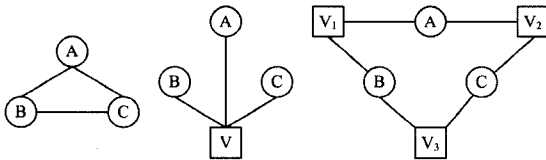
其中, m 为 H 中团的个数, ϕ_k 为第 k 个团的因子, $\mathbf{D}_k$ 为第 k 个团的所有随机变量, $Z = \sum_{X_1, \dots, X_n} \prod_{k=1}^m \phi_k(\mathbf{D}_k)$ 为划分函数。

4.1.2 高细粒度参数化

(1) 因子图

定义 31 因子图^[82] F 是包含两类节点的无向图:变量节点(用椭圆形表示)和因子节点(用正方形表示)。该图只包含变量节点和因子节点之间的边。因子图 F 通过因子集来参数化,每个因子节点 V 与一个因子 ϕ 有关, ϕ 的定义域是图中 V 的邻节点变量集。如果分布 P 能够表示为这种形式的因子集,则称 P 在 F 上分解。

马尔可夫网络结构一般没有发现吉布斯参数化的所有结构,不能从图结构中看出参数化中的因子是否包含最大团或其子集,而因子图能使网络中的因子结构更加明确^[83]。例如,如图 5 所示,从因子图中可直接看到因子与团的关系。



(a) 表示团 A, B, C 上的马尔可夫网络完全子图; (b) 表示 3 个变量上只有一个因子的因子图; (c) 表示两两变量上都有因子的因子图。

图 5 相同马尔可夫网络的不同因子图

(2) 对数线性模型

定义 32 令 $\mathbf{D}$ 为变量子集。定义特征 $f(\mathbf{D})$ 为从 $\mathbf{D}$ 到 R 的函数。

定义 33 如果分布 P 与下面两个集合:

(a) 特征集 $\mathbf{F} = \{f_1(\mathbf{D}_1), \dots, f_k(\mathbf{D}_k)\}$, 其中每个 $\mathbf{D}_i$ 是 H 的完全子图;

(b) 权值集 $\omega_1, \dots, \omega_k$;

有关,使得

$$P(X_1, \dots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^k \omega_i f_i(\mathbf{D}_i)\right] \quad (32)$$

则称 P 是马尔可夫网络图 H 上的对数线性模型。

把因子变换到对数空间中,则有 $\phi(\mathbf{D}) = \exp(-\epsilon(\mathbf{D}))$, 其中 $\epsilon(\mathbf{D}) = -\ln \phi(\mathbf{D})$ 称为能量函数。特征与能量函数有关,是一个没有非负要求的因子。利用特征集和权值集可以给出分布 P 的对数线性模型。当变量值域较大时,对数线性模型可以为分布提供更加简洁的表示。

现在,马尔可夫网络的参数化有 3 种表示:因子乘积、因子图、对数线性模型。对数线性模型比因子图具有更高细粒度的表示,而因子图比因子乘积具有更高细粒度的表示能力。因子图适用于学习推理,而对数线性模型适用于模型参数化。

4.1.3 过参数化

在使用细粒度因子图或对数线性模型表示时,马尔可夫网络的参数化一般都是过参数化的。也就是说,对任意给定的分布,模型中存在多种参数选择方式来描述概率分布。例如,图中有两个团 $\{A, B\}$ 和 $\{B, C\}$ 。能量函数 $\epsilon_1(A, B)$ 不仅包含 A 和 B 之间相互作用的信息,而且还包含单个变量 A 和 B 的分布信息。相同地, $\epsilon_2(B, C)$ 也给出了单个变量 B 和 C 的分布信息。与 B 有关的信息可以放在任意一个团中,或者 B 的作用可以用某种方式分离在两个团中,从而导致在说明同一个分布时有多种不同的方法。解决过参数化的方法有两种:正规参数化和消除冗余。

(1) 正规参数化

正规参数化能很自然地避免吉布斯分布 P 的参数化中的含糊不清,它要求分布 P 为正。 H 上吉布斯分布的正规参数化是通过所有非空团上的能量函数集来定义的。

先给出一些符号定义。 $\xi^* = (x_1^*, \dots, x_n^*)$ 是 χ 的特定取值,这种取值可任意选取。对于任意变量集 Z 和包含 Z 的 χ 的某个子集的任意赋值 x 来说,定义赋值 x_Z 为 $x \langle Z \rangle$, 也就是 Z 中变量的取值与 x 中该变量的取值一致。相反地,定义 ξ^*_{-Z} 为 $\xi^* \langle \chi - Z \rangle$, 也就是 ξ^* 中的值为 Z 外变量的取值。现在构造一个赋值 (x_Z, ξ^*_{-Z}) 。那么团 $\mathbf{D}$ 的正规能量函数为 $\epsilon_D^*(\mathbf{d}) = \sum_{Z \supset \mathbf{D}} (-1)^{|\mathbf{D}-Z|} \ln P(\mathbf{d}_Z, \xi^*_{-Z})$ 。

所以正吉布斯分布 P 的正规参数化表示为:

$$P(\xi) = \exp\left[-\sum_i \epsilon_{D_i}^*(\xi \langle \mathbf{D}_i \rangle)\right] \quad (33)$$

(2) 消除冗余

由于对数线性模型参数化表示有足够的细粒度,因此可以消除冗余而又不会损失表达能力,下面使用对数线性模型参数化表示。检测和消除冗余的方法来自于线性代数。令 $f_1, \dots, f_k$ 为具有权值 $\omega = \{\omega_1, \dots, \omega_k\}$ 的特征集,这样形成分布 P 的对数线性表示。如果存在系数 $\alpha_0, \alpha_1, \dots, \alpha_k$ 使得对所有 ξ 有 $\alpha_0 + \sum_i \alpha_i f_i(\xi) = 0$, 那么具有权值 $\omega' = \{\omega_1 + \alpha_1, \dots, \omega_k + \alpha_k\}$ 的对数线性模型同样能表示 P 。因此,线性相关的特征集是冗余的,非冗余的特征集中的特征是非线性相关的。只有当特征集是非冗余时,每个权值集才描述一个唯一的分布。也就是说,当 $f_1, \dots, f_k$ 为非冗余特征集时,令 $\omega, \omega' \in R^k$, 若 $\omega \neq \omega'$, 有 $P_\omega \neq P_{\omega'}$ 。

4.2 独立性

马尔可夫网络中的独立性有全局独立、两两独立和局部

独立 3 种。根据马尔科夫网络的独立性,可以由分布 P 构造网络图 H 。

4.2.1 全局独立假设

定义 34 令 H 为一个马尔科夫网络结构图,并令 $X_1 - X_2 - \dots - X_k$ 为 H 中的一条路径。令 $Z \subseteq \mathcal{X}$ 为观测变量集。如果任意 X_i 都不包含于 Z 中, $i=1, \dots, k$, 那么路径 $X_1 - X_2 - \dots - X_k$ 在已知 Z 时是活动的。

定义 35 如果已知 Z 时,在任意节点 $X \in \mathcal{X}$ 和 $Y \in \mathcal{Y}$ 之间没有活动路径,那么说已知 Z 时 X 和 Y 是分离的,表示为 $s_H(X; Y | Z)$ 。定义与 H 有关的全局独立集为:

$$I(H) = \{ (X \perp Y | Z); s_H(X; Y | Z) \} \quad (34)$$

(1) 分离的健全性

任意根据 H 分解的分布都满足分离隐含的全局独立性。令 P 为 $\mathcal{X}$ 上的一个分布, H 为 $\mathcal{X}$ 上的一个 MN 结构。如果 P 是根据 H 分解的一个吉布斯分布,那么 H 是 P 的一个 I-映射。但其反方向只适用于 P 为正分布的情况,即对变量集 $\mathcal{X}$ 的任意取值 x 都有 $P(x) > 0$ 的情况。也就是说,令 P 为 $\mathcal{X}$ 上的正分布, H 为 $\mathcal{X}$ 上的马尔科夫网络图,如果 H 是 P 的 I-映射,那么 P 是根据 H 分解的吉布斯分布。

(2) 分离的完备性

如果 H 中 X 和 Y 在给定 Z 时不是分离的,那么给定 Z 时在某个根据 H 分解的分布 P 中 X 和 Y 是依赖的。利用测度论可以得到,对于几乎所有在 H 上分解的分布 P 来说,也就是对于所有除因子参数化空间中的测度为零点集之外的所有分布来说,有 $I(P) = I(H)$ 。

4.2.2 局部马尔科夫假设

定义 36 令 H 为 MN, 定义与 H 有关的两两独立集 (pairwise independencies) 为:

$$I_p(H) = \{ (X \perp Y | \mathcal{X} - \{X, Y\}); X - Y \notin H \} \quad (35)$$

定义 37 对给定图 H , H 中 X 的马尔科夫毯 (Markov blanket) 为 H 中 X 的邻点,表示为 $MB_H(X)$ 。定义与 H 有关的局部独立集 (local independencies) 为:

$$I_l(H) = \{ (X \perp \mathcal{X} - \{X\} - MB_H(X) | MB_H(X)); X \in \mathcal{X} \} \quad (36)$$

在很多情况下,局部马尔科夫独立假设也是很有价值的。它关注于分布的较小独立性集合,并能有效简化分布 P 寻找 I-映射的过程。

两两独立集的定义较为直观,当给定图中所有其它结点时, X 与 Y 是独立的。

局部独立集的定义是间接定义,在给定直接邻点时 X 与图中其它的结点独立。

此时,已经给出 3 种独立假设集合:全局独立集 $I(H)$ 、两两独立集 $I_p(H)$ 和局部独立集 $I_l(H)$ 。对于一般分布来说, $I_p(H)$ 严格弱于 $I_l(H)$, 而 $I_l(H)$ 也严格弱于 $I(H)$ 。但是,对于正分布来说,3 种独立假设是等价的。

4.2.3 由分布 P 构建网络结构图 H

定义 38 若 $X \notin U$ 并且 U 是节点的最小集合,使得 $(X \perp \mathcal{X} - \{X\} - U | U) \in I(P)$, 那么集合 U 是分布 P 中 X 的马尔科夫毯。

有了结构图 H 和分布 P 的独立性关系后,如何构造分布 P 的最小 I-映射? 前面提到的局部马尔科夫假设为构建最小

I-映射提供了 2 种方法:一种是基于两两马尔科夫独立,另一种是基于局部独立。

第一种方法,如果 X 和 Y 是两两独立的,即给定所有其他结点时 X 和 Y 是独立的,那么在 X 和 Y 之间引入一条边 $\{X, Y\}$, 所得到的马尔科夫网络图 H 是 P 的唯一最小 I-映射。

第二种方法,令 P 为正分布。对每一个节点 X , 令 $MB_P(X)$ 为满足定义 38 的最小节点集 U 。通过对所有 X 和所有 $Y \in MB_P(X)$ 引入一条边 $\{X, Y\}$ 来定义图 H , 那么马尔科夫网络 H 是 P 的唯一最小 I-映射。

4.3 高斯模型马尔科夫随机场 (Gaussian Markov Random Field, GMRF)

定义 39 如果对所有 $i \neq j$, 有 $-\frac{J_{i,j}}{\sqrt{J_{i,i}J_{j,j}}} \geq 0$, 则由信息阵 $J = \Sigma^{-1}$ 参数化的二次 MRF 是吸引的 (attractive)。如果对所有 i , 有 $\sum_{j \neq i} |J_{i,j}| < J_{i,i}$, 则 MRF 是对角占优的。

定义 40 二次 MRF 的参数化如下:

$$\epsilon_i(x_i) = d_0^i + d_1^i x_i + d_2^i x_i^2 \quad (37)$$

$$\epsilon_{i,j}(x_i, x_j) = a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{i,j} x_j + a_{11}^{i,j} x_i x_j + a_{02}^{i,j} x_i^2 + a_{20}^{i,j} x_j^2 \quad (38)$$

若对所有 i , 有 $d_2^i > 0$; 并对所有 i, j , 2×2 矩阵 $\begin{pmatrix} a_{02}^{i,j} & a_{11}^{i,j}/2 \\ a_{11}^{i,j}/2 & a_{20}^{i,j} \end{pmatrix}$ 是半正定的, 那么称二次 MRF 为两两归一化的。

高斯模型马尔科夫随机场^[84,85]是通过无向图模型来表示多元高斯分布的。首先说明高斯分布如何表示 MRF, 然后说明 MRF 如何定义高斯分布, 并给出两个检验条件来判断 MRF 是否能有效定义高斯分布。

(1) 高斯分布如何表示为 MRF

高斯分布想要表示 MRF, 可以直接从高斯模型的信息形式中将其推导出来。由 3.5 节知道, 高斯模型的信息形式的参数化使用协方差的逆阵, 即信息阵 $J = \Sigma^{-1}$, 概率密度的信息形式为:

$$p(x) \propto \exp \left[-\frac{1}{2} x^T J x + (J \mu)^T x \right] \quad (39)$$

把信息形式分解为 2 类:一类包含单变量 X_i , 另一类包含变量 X_i, X_j 的组合。只包含单变量 X_i 的项为:

$$-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i \quad (40)$$

包含 X_i, X_j 组合对的项为:

$$-\frac{1}{2} [J_{i,j} x_i x_j + J_{j,i} x_j x_i] = -J_{i,j} x_i x_j \quad (41)$$

信息形式直接引出一个两两关系马尔科夫网络, 其节点势能来自于势能向量和信息阵的对角元素, 其边势能来自于信息阵的非对角元素。因此, 任意高斯分布可表示为带有二次节点和边势能的两个关系马尔科夫网络。该马尔科夫网络一般称为高斯模型马尔科夫随机场 (GMRF)。

(2) MRF 如何定义高斯分布

考虑任意带有二次节点和边势能的两个关系马尔科夫网络。忽略常数因子, 常数因子可被合并到划分函数中去, 这样就可以写出节点和边的能量函数:

$$\epsilon_i(x_i) = d_0^i + d_1^i x_i + d_2^i x_i^2 \quad (42)$$

$$\epsilon_{i,j}(x_i, x_j) = a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{i,j} x_j + a_{11}^{i,j} x_i x_j + a_{02}^{i,j} x_i^2 + a_{20}^{i,j} x_j^2 \quad (43)$$

通过合并,可以用对数二次形式重新表示这样的势能集合:

$$p'(x) = \exp(-\frac{1}{2}x^T J x + h^T x) \quad (44)$$

其中不失一般性地假设 J 是对称的。当且仅当 J 是正定阵时,该马尔可夫网络定义一个有效的高斯分布。

然而,与高斯模型的贝叶斯网络的情况不一样,不是每个二次节点和边势能的集合都引出一个合法的高斯分布。要检验 MRF 是否有效,需要产生完整的信息阵,并检验它是否正定。首先介绍两个简单检验。

第一个检验是,令 p' 为二次两两关系 MRF,若 J 是吸引的,那么 p' 定义一个有效高斯模型 MRF;若 J 是对角占优的,那么 p' 定义一个有效高斯模型 MRF。

第二个检验是,令 p' 为二次两两关系 MRF,如果 p' 是两两归一化的,那么它定义一个有效高斯分布。

4.4 马尔可夫网络的指数分布族

定义 41 非归一化指数因子分布族 Φ 由 τ, t, A 和 θ 定义,该分布族中的一个因子为 $\phi_\theta(\xi) = A(\xi) \exp\{\langle t(\theta), \tau(\xi) \rangle\}$ 。

定义 42 令 $\Phi_1, \dots, \Phi_k$ 为指数因子分布族,其中每个 Φ_i 由 τ_i, t_i, A_i 和 θ_i 定义。 $\Phi_1, \dots, \Phi_k$ 的组合为分布族 $\Phi_1 \times \Phi_2 \times \dots \times \Phi_k$,它是由 $\theta = \theta_1 \circ \theta_2 \circ \dots \circ \theta_k \in \theta_1 \times \theta_2 \times \dots \times \theta_k$ 来参数化的,定义如下:

$$P_\theta(\xi) \propto \prod_i \phi_{\theta_i}(\xi) = (\prod_i A_i(\xi)) \exp\{\sum_i \langle t_i(\theta_i), \tau_i(\xi) \rangle\} \quad (45)$$

其中, ϕ_{θ_i} 是第 i 个因子分布族中的一个因子。

在 MN 的对数线性模型中,有如下形式的分布:

$$P(X_1, \dots, X_n) = \frac{1}{Z} \exp[-\sum_{i=1}^k \omega_i f_i(D_i)] \quad (46)$$

这样的分布显然是一个线性指数分布族,其充分统计量函数是特征函数的向量 $\tau(\xi) = \langle f_1(d_1), \dots, f_k(d_k) \rangle$ 。通过选择合适的特征函数,可以设计一个对数线性模型来表示已知的离散马尔可夫网络结构。这足以说明离散马尔可夫网络是线性指数分布族。

在乘积形式的分布中,各项的乘积形式对应于指数分布族的一个简单组合。由定义 42 看出,指数因子的乘积是带有 $\tau(\xi) = \tau_1(\xi) \circ \tau_2(\xi) \circ \dots \circ \tau_k(\xi)$ 和自然参数 $t(\theta) = t_1(\theta_1) \circ t_2(\theta_2) \circ \dots \circ t_k(\theta_k)$ 的指数分布族。这说明,如果马尔可夫网络中的势能有指数表示形式,那么势能的乘积也是一个指数分布族。而且,线性指数因子分布族的乘积还是一个线性指数分布族。

5 局部有向模型

本节介绍两种局部有向模型:条件随机场和链图。

5.1 条件随机场

定义 43 条件随机场是节点对应于 $X \cup Y$ 的无向图 H ;参数模型为因子集 $\phi_1(D_1), \dots, \phi_m(D_m)$,变量集 $D_i \subset X$ 。网络编码的条件分布如下:

$$P(Y|X) = \frac{1}{Z(X)} \tilde{P}(Y, X) \quad (47)$$

$$\tilde{P}(Y, X) = \prod_{i=1}^m \phi_i(D_i) \quad (48)$$

$$Z(X) = \sum_Y \tilde{P}(Y, X) \quad (49)$$

只要 H 中的两个变量共同出现在某个因子的值域中,它们就由一条无向边连接。

CRF 是一个无向图,其参数化方式与马尔可夫网络的参数化一样,但是 CRF 可以编码条件概率分布 $P(Y|X)$,这里 Y 是目标变量集, X 是观测变量的不相交集。CRF 的定义与马尔可夫网络的定义的唯一区别是划分函数 $Z(X)$ 中使用了不同的归一化,当 X 取不同值 x 时,其划分函数的值也不相同。

CRF 表示的一个主要优点是避免编码 X 中变量的分布,这样就可以将大量相当复杂的或难以理解的观测变量独立性集融入到模型中。CRF 同样可以包含没有简单参数形式的连续变量分布。所以可以使用领域知识来定义一个丰富特征集合,而不用考虑联合概率分布的建模。

Lafferty 等人首次提出 CRF 的表示框架^[86],随后 CRF 在许多领域中得到广泛应用,如自然语言处理、计算机视觉等。后来 Sutton 和 McCallum 又引入跳跃链 CRF^[87],而 Sutton 等人引入因子 CRF^[88]。

5.2 链图

链图^[89]建立于 CRF 表达之上,并可以处理局部有向模型的独立假设。在局部有向无环图(Partially Directed Acyclic Graph, PDAG)中,节点可以不相交地划分为多个链单元。在同一个链单元中两个节点间的边一定是无向的,而不同链单元中两个节点间的边一定是有向的。因此,PDAG 也称为链图。下面讨论链图的分解和独立假设。

5.2.1 链图的因式分解

定义 44 令 K 为一个 PDAG, $K_1, \dots, K_\ell$ 为其链单元。定义 F_{K_i} 为 K_i 中节点的父节点。 K 的道德图是一个无向图 $M[K]$,首先把任意节点对 $X, Y \in F_{K_i}, i=1, \dots, \ell$ 用无向边连接起来,然后把所有有向边转换为无向边,就可以产生 $M[K]$ 。

定义 45 令 K 为一个 PDAG, $K_1, \dots, K_\ell$ 为其链单元。链图分布是通过因子集 $\phi_i(D_i) (i=1, \dots, m)$ 来定义的,每个 D_i 是道德图 $M[K]$ 中的一个完全子图。将每个因子 $\phi_i(D_i)$ 与一个链单元 K_j 联系起来,使得 $D_i \subseteq K_j \cup F_{K_j}$,并且定义 $P(K_i | K_{K_i})$ 为含有这些因子的 CRF, $Y_i = K_i, X_i = F_{K_i}$ 。现在定义

$$P(X) = \prod_{i=1}^{\ell} P(K_i | F_{K_i}) \quad (50)$$

PDAG K 的结构可用来因式分解联合概率分布,把联合概率分布作为已知父节点时每个链单元的乘积。链图模型中的每个链单元 K_i 都与一个定义 $P(K_i | F_{K_i})$ 的 CRF 有关。如果分布 P 能表示为 K 上的链图分布,那么说分布 P 根据 K 来因式分解。

5.2.2 链图中的独立

定义 46 令 K 为 PDAG,定义与 K 有关的两两独立集为:

$$I_P(K) = \{(X \perp Y | (ND_X - \{X, Y\})); X, Y \text{ 不相交}, Y \in ND_X\} \quad (51)$$

定义 47 令 K 为 PDAG,定义与 K 有关的局部独立集为:

$$I_\ell(K) = \{(X \perp_i ND_X - B_X | B_X); X \in \mathcal{X}\} \quad (52)$$

定义 48 令 $X, Y, Z \subset \mathcal{X}$ 为 3 个不相交集,并令 $U = X \cup Y \cup Z$ 。如果在无向图 $M[K^+][X \cup Y \cup Z]$ 中已知 Z 时 X 与 Y 是分离的,那么说已知 Z 时 X 与 Y 是 c -分离的。

定义 49 令 K 为 PDAG,定义与 K 有关的全局独立为:

$I(K) = \{(X \perp Y | Z) : X, Y, Z \subset \chi, \text{已知 } Z \text{ 时 } X \text{ 与 } Y \text{ 是 } c\text{-分离的}\}$ (53)

在 PDAG 中, X 的父节点集和 X 的邻节点集的并集是 X 的边界节点集, 表示为 B_X 。定义 46 泛化了无向图的两两独立: 在无向图中, 节点没有子节点, 所以 $\bar{C}_X = \chi$ 。类似地, 不难说明这些独立同样存在于有向图中。定义 47 泛化了有向和无向图的局部独立。对于有向图而言, $\bar{C}_X$ 是非子节点集, 而 B_X 是父节点集。对于无向图而言, $\bar{C}_X$ 是 χ , 而 $B_X = N_X$ 。 c -分离^[90]的概念泛化了无向图的分离概念以及有向图的 d -分离概念。

对于非正分布来说, 这 3 个独立假设并不等价。全局独立隐含局部独立, 而局部独立隐含两两独立。但是对于正分布来说, 这 3 种定义等价。

6 基于模板的表示

无论是贝叶斯网络还是马尔可夫网络, 都是用来表示某个固定随机变量集 χ 的联合概率分布。由于这类表示方法的侧重点是随机变量集, 因此称此类模型为基于变量的模型。然而, 在许多领域中, 概率模型要应用于更加复杂的空间, 而不能把模型编码为某个固定的变量集。基于模板的表示主要应用于两个场景。第一个是暂态建模, 用动态贝叶斯网络语言构造一个描述系统动态属性的单个紧凑模型, 在不同时间轨迹上产生不同分布。第二个应用场景是为具有相互关联关系的多个对象建模。

首先描述基于模板表示的基本框架, 主要有 3 个重要概念: 模板属性、实例化和模板因子。

(1) 模板属性 (Template Attributes)

定义 50 属性 A 为函数 $A(U_1, \dots, U_k)$, 值域为集合 $R(A)$, 参数 U_i 为与特定类 $Q[U_i]$ 相关联的类型逻辑变量。元组 $U_1, \dots, U_k$ 称为属性 A 的参数特征, 表示为 $\alpha(A)$ 。

模板在模型中被多次实例化, 它可以对具有相同值域和同一语义的多个随机变量模型进行编码。模板属性可以是单个对象的性质, 也可能是整个对象元组的性质。一般把对象分为满足 $Q \cap Q_i = \emptyset$ 和 $\bigcup Q_i = Q$ 的类。

(2) 实例化 (Instantiations)

定义 51 令 $\mathcal{Q}$ 为类集合, N 为 $\mathcal{Q}$ 上的模板属性集合。对每个 $Q \in \mathcal{Q}$, 对象骨架 K 为一个固定的有限对象集合 $O^K[Q]$ 。定义

$$O^K[U_1, \dots, U_k] = O^K[Q[U_1]] \times \dots \times O^K[Q[U_k]] \quad (54)$$

缺省时 $\Gamma_K[A] = O^K[\alpha(A)]$ 为 A 的参数特征中逻辑变量可能取值的集合。对象骨架也确定了合法赋值 $\Gamma_K[A] \subset O^K[\alpha(A)]$ 的子集。

定义 52 令 K 为 $\mathcal{Q}, N$ 上的对象骨架。定义底层随机变量集为:

$$\chi_K[A] = \{A(\gamma) : \gamma \in \Gamma_K[A]\} \quad (55)$$

$$\chi_K[N] = \bigcup_{A \in N} \chi_K[A] \quad (56)$$

给定模板属性集合, 可以把其实例化来产生同一类型多个随机变量的概率空间。

(3) 模板因子

定义 53 模板因子是定义在模板属性元组 $A_1, \dots, A_\ell$ 上的函数 ξ, A_j 的值域为 $R(A)$ 。它定义了从 $R(A_1) \times \dots \times$

$R(A_\ell)$ 到 R 的一个映射。给定随机变量元组 $X_1, \dots, X_\ell$, 使得对所有 $j=1, \dots, \ell$, 有 $R(X_j) = R(A_j)$ 。定义 $\xi(X_1, \dots, X_\ell)$ 为从 X 到 R 的实例化因子。

本节介绍的暂态模型分为动态贝叶斯网络和状态观测模型, 其中动态观测模型又包括隐马尔可夫模型和线性动态系统。HMM 与 DBN 有着密切关系, 也可以用 HMM 来表示 DBN。对象关系领域的概率模型可以表示多个对象间的复杂相互作用, 它的有向图表示包括盘模型和概率关系模型, 它也可以使用无向图来表示。对象关系领域的概率模型还可以编码结构不确定性。结构不确定性主要有两种类型: 关系不确定性, 为对象间相关关系不确定的分布建模; 对象不确定性, 为领域中实际对象的存在不确定或数量不确定的分布建模^[91]。但本文不对结构不确定性的建模进行展开讨论。

6.1 暂态模型

暂态模型用于动态场景建模, 系统状态随着时间的变化而改变。假定系统状态表示为某个随机变量集 χ 的值。用 $X_i^{(t)}$ 表示变量 X_i 在时刻 t 的实例化, X_i 不再是取某个值的变量, 而是模板变量, 模板变量在时间 t 内的不同时刻上实例化。 $X_i^{(t)}$ 是值域为 $R(X_i)$ 的变量。对于变量集 $X \subset \chi$, 用 $X^{(t_1, t_2)} (t_1 < t_2)$ 来表示变量集 $\{X^{(t)} : t \in [t_1, t_2]\}$ 。用符号 $x^{(t, t')}$ 表示该变量集的一个取值。

以下给出 3 个简化假设, 使模型表示问题更易于处理。

假设 1 离散假设。把时间轴离散为时间段: 系统状态的测量值由预先设定的等间隔时间粒度 Δ 采集。这样就把随机变量集限制为 $\chi^{(0)}, \chi^{(1)}, \dots$, 其中 $\chi^{(t)}$ 是表示时刻 $t \cdot \Delta$ 的系统状态的底层随机变量。

假设 2 马尔可夫假设。如果对所有 $t \geq 0$, 有 $(\chi^{(t+1)} \perp \chi^{(0:t-1)} | \chi^{(t)})$, 那么模板变量 χ 的动态系统满足马尔可夫假设, 这样的系统称为马尔可夫系统。概率分布可以简化为:

$$P(\chi^{(0:T)}) = \prod_{t=0}^{T-1} P(\chi^{(t+1)} | \chi^{(0:t)}) = \prod_{t=0}^{T-1} P(\chi^{(t+1)} | \chi^{(t)}) \quad (57)$$

假设 3 平稳假设。如果对所有 $t, P(\chi^{(t+1)} | \chi^{(t)})$ 都是相同的, 那么说马尔可夫动态系统是平稳的。这样可以使用转换模型 $P(\chi' | \chi)$ 来表示过程, 所以对任意 $t \geq 0$, 有

$$P(\chi^{(t+1)} = \xi' | \chi^{(t)} = \xi) = P(\chi' = \xi' | \chi = \xi) \quad (58)$$

6.1.1 动态贝叶斯网络 (Dynamic Bayesian Network, DBN)

定义 54 χ 的二时间段贝叶斯网络 (2-time-slice Bayesian Network, 2-TBN) 是已知 χ_t 时 χ' 上的条件贝叶斯网络, 其中 $\chi_t \in \chi$ 是接口变量集, 接口变量 χ_t 在时刻 t 的值对时刻 $t+1$ 的变量有直接影响。

定义 55 动态贝叶斯网络^[7,44] 是 $[B_0, B_-]$ 的组合, 其中 B_0 是 $\chi^{(0)}$ 的贝叶斯网络, 表示状态的初始分布, 而 B_- 是随机过程的 2-TBN。对任意时间跨度 $T \geq 0$, $\chi^{(0:T)}$ 的分布定义为展开的贝叶斯网络, 其中, 对任意 $i=1, \dots, n$, 有:

(a) $X_i^{(0)}$ 的结构和 CPD 与 B_0 中 X_i 的结构和 CPD 一样;

(b) 对 $t > 0$, $X_i^{(t)}$ 的结构和 CPD 与 B_- 中 X_i' 的结构和 CPD 一样。

在条件贝叶斯网络中, 只有变量 χ' 有父节点或 CPD。接口变量 χ_t 在时刻 t 的值对时刻 $t+1$ 的变量有直接影响。因此, 只有 χ_t 中的变量才能成为 χ' 中变量的父节点。如图 6 所

示, X_1, X_2, X_3 和 X_4 是接口变量。2-TBN 表示条件概率分布

$$P(X'|\chi) = P(X'|\chi_t) = \prod_{i=1}^n P(X_i' | Pa_{X_i'}) \quad (59)$$

对每个模板变量 X_i 来说, CPD $P(X_i' | F_{X_i'})$ 是一个模板因子; 由于存在多个变量 $X_i^{(t)}$, 该模板因子将在模型中多次被实例化。

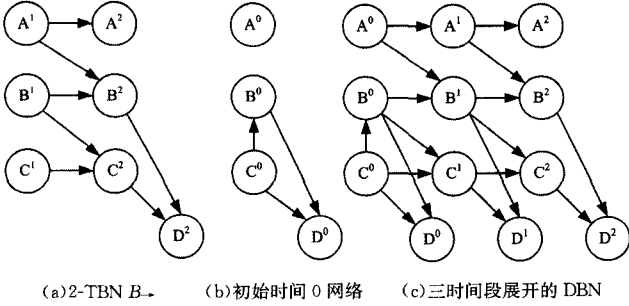


图 6

根据前面的假设, 只需要表示初始状态分布和转换模型 $P(X'|\chi)$, 就可以在无限轨线上表示概率分布。转换模型是用 2-TBN 表示的 CPD。

由 HMM 构建的 DBN 有两类^[50]: 因式化 HMM^[92] 和耦合 HMM^[93], 如图 7 所示。因式化 HMM 的 2-TBN 结构为链 $X_i \rightarrow X_i' (i=1, \dots, n)$ 的集合, 并带有一个所有变量 X_i' 子节点的观测变量 Y 。耦合 HMM 也是由链 $X_i \rightarrow X_i'$ 的集合构成, 但是每条链都是一个带有观测变量 Y_i 的 HMM。这些链通过状态变量直接相互作用, 每条链都能作用于其相邻链。

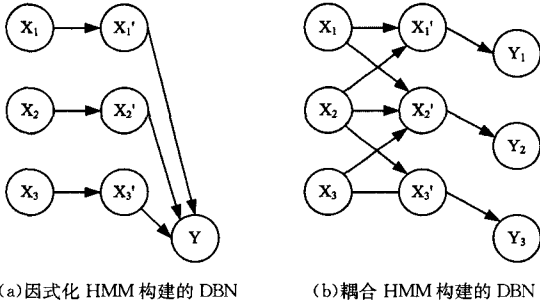


图 7 两类 HMM 构建的 DBN

6.1.2 状态观测模型(State-observation Model)

状态观测模型是另一种暂态模型。状态观测模型假定两个独立性假设成立: 状态变量由马尔可夫方式产生, 使得 $(X^{(t+1)} \perp X^{(0, (t-1))} | X^{(t)})$; 以及已知时刻 t 的状态变量时, 时刻 t 的观测变量条件独立于整个状态序列: $(O^{(t)} \perp X^{(0, (t-1))}, X^{(t+1, \infty)} | X^{(t)})$ 。因此, 状态观测模型由两部分组成: 转换模型 $P(X' | X)$ 和观测模型 $P(O | X)$ 。

状态观测模型包含两类: 隐马尔可夫模型和线性动态系统。

(1) 隐马尔可夫模型 (Hidden Markov Model, HMM)

HMM^[43, 94] 是比较简单的状态观测模型, 如前面所说, 它也可作为 DBN 的特殊情况。HMM 的转换模型 $P(S' | S)$ 可用有向带环图来编码, 节点表示系统的不同状态, 即 $R(S)$ 中的值。如果有可能实现从 s 到 s' 的转换, 那么从 s 到 s' 之间有一条有向弧, 即 $P(s' | s) > 0$ 。从 s 到 s' 的边标注了转换概率 $P(s' | s)$ 。

HMM 的转换图与编码图模型的图, 在本质上是不同的。

转换图的节点是状态变量, 或者是状态变量的可能值; 有向边表示状态之间的可能转换, 或者表示 CPD 中概率为零的项。因此, 离开节点的边权之和必须为 1。注意到, 这种基于图的表示方法没有编码 HMM 的观测模型。在某些情况中, 观测模型是确定性的, 对每个 s 都只存在一个观测变量 o , $P(o | s) = 1$ 。此时, 观测变量标注在与状态有关的节点上。

(2) 线性动态系统 (Linear Dynamical System, LDS)

LDS 表示带有高斯噪声的一个或多个随时间演化的实值变量的系统。这样的系统也叫卡尔曼滤波器。LDS 可视为所有变量为连续变量且变量依赖关系为线性高斯模型的 DBN。

LDS 表示为状态观测模型, 其状态变量和观测变量都是随机向量, 转换模型和观测模型都使用矩阵来编码, 定义如下:

$$P(X^{(t)} | X^{(t-1)}) = N(AX^{(t-1)}; Q) \quad (60)$$

$$P(O^{(t)} | X^{(t)}) = N(HX^{(t)}; R) \quad (61)$$

其中, A 是定义线性转换模型的 $n \times n$ 矩阵, Q 是定义与系统动态性有关的高斯噪声的 $n \times m$ 矩阵, H 是定义线性观测模型的 $n \times m$ 矩阵, R 是定义与观测变量有关的高斯噪声的 $m \times m$ 矩阵。

开关线性动态系统 (Switching Linear Dynamical System, SLDS)^[95] 是既包含线性连续动态性也包含离散变量的系统, 系统在离散模式之间进行开关切换。

6.2 对象关系域的有向概率模型

对象关系域上有向概率模型, 是基于模板表示的另一个应用, 包括盘模型和概率关系模型。

6.2.1 盘模型

定义 56 对于每个带有参数特征 $U_1, \dots, U_k$ 的模板属性 $A \in N$, 盘模型^[53] M_{Plate} 定义如下:

(a) 模板父节点集合 $F_A = \{B_1(U_1), \dots, B_k(U_k)\}$, 使得对每个 $B_i(U_i)$ 有 $U_i \in \{U_1, \dots, U_k\}$ 。变量 U_i 是父节点 B_i 的参数特征。

(b) 模板 CPD $P(A | F_A)$ 。

定义 57 盘模型 M_{Plate} 和对象骨架 K 定义如下的底层贝叶斯网络 B_{Plate}^M 。令 $A(U_1, \dots, U_k)$ 为 N 中的任意模板属性, 对任意赋值 $\gamma = \langle U_1 \mapsto u_1, \dots, U_k \mapsto u_k \rangle \in \Gamma_K[A]$, 底层网络中存在变量 $A(\gamma)$, $A(\gamma)$ 的父节点为 $B(\gamma)$ (所有 $B \in F_A$), $A(\gamma)$ 的实例化 CPD 为 $P(A(\gamma) | Pa_A(\gamma))$ 。

盘模型由 Gilks 等人提出, 后来它作为在模型内或模型间共享参数的语言。盘模型中, 对象类型称为盘。多个同类对象共享相同的属性集和相同的概率模型, “盘”表示一摞同类对象。模板属性嵌入在盘中, 如果属性 A 嵌入在盘 $Q_1, \dots, Q_k$ 的集合中, 可把 A 视为与参数特征 $U_1, \dots, U_k$ 相关, 其中每个逻辑变量 U_i 在盘(类) Q_i 的对象上取值。盘模型允许多个盘以各种方式叠加在一起, 如盘嵌套和盘相交。

当已知对象集时, 实例化这些对象能诱导出一个底层贝叶斯网络 B_{Plate}^M , 该网络说明了 $\chi_K[N]$ 上的联合分布。

但是盘模型有一个显著缺点, 它约束 $A(U_1, \dots, U_k)$ 只依赖于形式为 $B(U_{i_1}, \dots, U_{i_k})$ 的属性, 其中 $U_{i_1}, \dots, U_{i_k}$ 是 $U_1, \dots, U_k$ 的一个子元组。这个约束条件使我们不能编码对象之间的概率依赖关系网络。

6.2.2 概率关系模型(Probabilistic Relational Models, PRM)

盘模型的最大限制是对属性父节点的参数特征的约束。下面对盘模型放松这种限制, Friedman 等人提出了概率关系模型^[9,10,96,97]。首先引入或有依赖(Contingent Dependencies)模型^[56]的概念,再描述在底层网络的 CPD 中如何编码或有依赖,以及如何处理带环的问题。

(1) 或有依赖模型

定义 58 对模板属性 A 来说,定义或有依赖模型为一个元组,由以下成分组成:

(a) 父节点参数特征 $\alpha(F_A)$,它是类型逻辑变量 U_i 的一个元组,使得 $\alpha(F_A) \supseteq \alpha(A)$ 。

(b) 警卫谓词表达式 Γ ,它是根据参数特征 $\alpha(F_A)$ 上的模板属性集 F_A^Γ 来定义的二值逻辑表达式。

(c) 模板父节点集 $F_A = \{B_1(U_1), \dots, B_\ell(U_\ell)\}$,使得对每个 $B_i(U_i)$ 有 $U_i \subseteq \alpha(F_A)$ 。

定义 59 概率关系模型 M_{PRM} 和对对象骨架 K 定义一个如下描述的底层贝叶斯网络 B_{PRM}^K 。令 $A(U_1, \dots, U_k)$ 为 N 中的任意模板属性。对任意 $\gamma \in \Gamma_K[A]$,底层网络存在变量 $A(\gamma)$ 。对任意 $B \in F_A^\Gamma \cup F_A$ 和 $\alpha(F_A) - \alpha(A)$ 的任意取值 γ' ,变量 $A(\gamma)$ 的父节点为实例化变量 $B(\gamma, \gamma')$ 。

或有依赖模型说明了特定依赖存在的上下文。PRM M_{PRM} 对每个 $A \in N$ 定义一个或有依赖模型和一个模板 CPD。PRM 中的模板父节点定义为底层网络中父节点取值的模板,这个模板对应于适当类型对象逻辑变量的特定赋值。然而,父节点逻辑变量集不一定是子节点逻辑变量的子集。由定义 59 可以看出,出现在警卫谓词表达式中的属性也是底层变量的父节点。

(2) 底层网络中的 CPD

首先利用警卫谓词表达式结构,假设警卫谓词表达式是基本关系结构的特征,并且事先固定好。通过定义关系骨架 K_Γ ,可以编码这个假设。 K_Γ 定义了事先给定的某个事件集(通常是对象间的关系集),并且 K_Γ 不是概率模型的一部分。当警卫谓词表达式中的属性值被说明为关系骨架的一部分时,可以利用该信息去确定模型中的有限父节点集。使用或有依赖可以利用由骨架确定的关系来产生更简化的模型,并且明确使模型为无环的。

然而,警卫谓词表达式的使用不能完全处理 PRM 中 CPD 参数化的定义问题。因为父节点的实例数量事先未知,并且可能是无限的。因此需要定义一种机制来说明模板水平上的父节点数量为变量的局部依赖模型。这可以利用聚合 CPD(Aggregator CPD)来实现,聚合 CPD 使用聚合统计数或者变量父节点集合的和。它把同类父节点组合到一个 CPD 中,把 CPD 集合 $\{P(X|Y_{i,1}, \dots, Y_{i,j_i}) : i=1, \dots, l\}$ 组合到一个联合 CPD $P(X|Y_{1,1}, \dots, Y_{1,j_1}, \dots, Y_{l,1}, \dots, Y_{l,j_l})$ 中。

(3) 检查无环性

定义 60 模板依赖模型 M_{PRM} 的模板依赖图,对每个模板,水平属性 A 都包含一个节点,无论 $F_A^\Gamma \cup F_A$ 中是否存在类型 B 的属性,图中都包含一条从 B 到 A 的有向边。

通过检查生成的任意底层网络是否无环,可以获得一个一致概率分布。其中一种方法是构建一个编码底层级潜在依赖关系的模板依赖图。如果模型 M_{PRM} 的模板依赖图是无环的,那么由 M_{PRM} 产生的任意底层网络也是无环的。然而,

M_{PRM} 的模板依赖图中的一个环并不意味着由 M_{PRM} 引起的每一个底层网络是有环的。

6.3 无向表示

定义 61 关系马尔可夫网络(Relational Markov Network, RMN)是根据模板特征集 Γ 来定义的,其中每个 $\gamma \in \Gamma$ 包含:

(a) 实值模板特征 f_γ ,其参数为 $N(\gamma) = \{A_1(U_1), \dots, A_\ell(U_\ell)\}$;

(b) 权值 $\omega_\gamma \in R$ 。

定义 $\alpha(\gamma)$ 使得对所有 i 有 $U_i \in \alpha(\gamma)$ 。

定义 62 给定 RMN M_{RMN} 和对象框架 K ,可以定义如下的底层吉布斯分布 P_{RMN}^K :

(a) 网络中的变量为 $\chi_K[N]$;

(b) 对每个特征模板 $\gamma \in \Gamma$ 和每个赋值 $\gamma \in \Gamma_K[\alpha(\gamma)]$, P_{RMN}^K 包含一项 $\exp(\omega_\gamma \cdot f_\gamma(\gamma))$ 。

前面描述的基于模板的表示使用有向图模型,其实也可以用无向图模型来表示。关系域的无向概率模型起源于 Taskar 等人提出的关系马尔可夫网络框架^[8,98]。Richardson 和 Domingos 给出一种基于逻辑公式的特征表示方法,即马尔可夫逻辑网^[41,42,99]。在马尔可夫逻辑网络中,团的说明与势能中特征的说明没有分离。当然,模型是根据逻辑公式集来定义的,每个公式都与一个权值关联。在避免局部归一化要求和无环性约束中,无向表示具有更大的灵活性。

但是,无向模型也存在着问题:不利于对特定大小的值域通过学习或手工构建基于模板的模型,而且也不利于把无向模型应用于不同大小的模型中。值域大小对概率分布的影响是变化的,因此在每次应用时都需要评估模型在此场景中的学习能力。

结束语 本文综述了概率图模型的表示方法及其独立性,梳理了节点上的多种局部概率模型,并总结了贝叶斯网络及其扩展、马尔科夫网络及其扩展、局部有向模型和基于模板表示模型的表示方法和独立性。

虽然概率图模型已成功应用于多个领域,但是仍有许多关于概率图模型的表示方面的问题有待进一步解决。

(1) 在 BN 中,参数的选择会对推理的结论产生影响。我们希望网络参数即使发生相当大的变化,也只会引起网络性能上的较小退化。如何分析和尽量消除这种影响有待进一步研究。

(2) 目前,可以通过与不同参数有关的网络来测量网络对参数方差的灵敏度,但是如何评价灵敏度的性质以及有效估计灵敏度的问题还有待回答。

(3) DBN 结构假设了所应用的系统为平稳动态系统,即系统为时不变齐次系统,但是在实际中所遇到的多数系统都不符合平稳假设。因此,非平稳 DBN 是一个研究方向。另外,DBN 需要把时间轴离散化,然而这样会损失信息。虽然已提出连续 DBN,但是其理论还需深化。

(4) 对象关系领域的概率模型可以为对象间相互关系网络的不确定性建模,甚至可以为实际对象集的不确定性建模。但是在结构丰富的概率空间中如何选择合适的分布类型来表示联合概率分布,特别是引入当前比较热门的关系语言,如逻辑符号、函数符号和量词等,是一个更加复杂的问题。

(5) 对象关系领域的概率模型表示方法是基于知识的模

型构建方法,其语义依赖于模型实例化所产生的底层网络。然而,该方法是有问题的,该过程产生的模型为推理算法带来了很大的挑战。首先,所产生的底层网络相当巨大,比通过手工构建的模型大得多;其次,由于变量间的多相互作用,该模型的连接相当稠密;而且,关系的结构不确定性和对象存在的结构不确定性也会导致模型的连接更加稠密。这些问题有待进一步解决。

(6)基于模板的概率模型,除了本文所描述的基于逻辑的表示之外,近年来还出现一种程序语言框架,称为归纳逻辑程序(Inductive Logic Programming, ILP)。它把随机变量视为从其输入到其输出的随机函数。定义随机函数后,就可以在多个空间重复使用该函数。更重要的是,可以定义函数来调用其它函数,或者定义函数来递归调用函数本身。基于该框架的语言包括概率上下文自由语法和面向对象贝叶斯网络。但是,该表示方法与概率结合的理论还需深入研究。

参 考 文 献

- [1] Langseth H, Portinale L. Bayesian networks in reliability [J]. Reliability Engineering and System Safety, 2007, 92 (1): 92-108
- [2] Santana R, Shakya S. Probabilistic Graphical Models and Markov Networks[J]. Markov Networks in Evolutionary Computation, 2012, 14(1): 3-19
- [3] Shakya S, Santana R. A Review of Estimation of Distribution Algorithms and Markov Networks[J]. Markov Networks in Evolutionary Computation, 2012, 14(1): 21-37
- [4] Pal C J, Weinman J J, Tran L C, et al. On Learning Conditional Random Fields for Stereo: Exploring Model Structures and Approximate Inference[J]. International Journal of Computer Vision, 2012, 99(3): 319-337
- [5] Golumbic M C, Maffray F, Morel G. A characterization of chain probe graphs[J]. Annals of Operation Research, 2009, 188(1): 175-183
- [6] Sun X, Huang D. Nested Template-based Model for Chinese-Japanese Machine Translation[C]// Proceedings of 14th International Conference on Computational Science and Engineering, 2011: 161-166
- [7] Murphy K. Dynamic Bayesian networks: representation, inference and learning[D]. Berkeley: University of California, 2002
- [8] Getoor L, Taskar B. Introduction to Statistical Relational Learning[M]. MIT Press, 2007
- [9] Friedman N, Getoor L, Koller D, et al. Learning probabilistic relational models[C]// Proceedings of the 16th International Joint Conference on Artificial Intelligence, 1999: 1300-1307
- [10] Medina-Oliva G, Weber P, Levrat E, et al. Use of probabilistic relational model (PRM) for dependability analysis of complex systems[C]// Proceedings of 12th IFAC Symposium on Large Scale Systems: Theory and Applications, 2010
- [11] Shafer G, Pearl J. Readings in uncertain reasoning [M]. San Francisco, CA: Morgan Kaufmann Publishers, 1990
- [12] Halpern J Y. A logical approach to reasoning about uncertainty: a tutorial[M]. Netherlands: Springer, 1998: 141-155
- [13] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Mateo, California, Morgan Kaufmann Publishers, 1988: 117-133
- [14] Lauritzen S L, Spiegelhalter D J. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems[J]. Journal of the Royal Statistical Society, Series B (Methodological), 1988, 50(2): 157-224
- [15] Dawid A P. Influence Diagrams for Causal Modelling and Inference[J]. International Statistical Review, 2002, 70(2): 161-189
- [16] Geiger D, Verma T, Pearl J. Identifying independence in Bayesian networks[J]. Networks, 1990, 20(5): 507-534
- [17] Verma T, Pearl J. Causal networks: semantics and expressiveness[C]// Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence. The Netherlands: North-Holland Publishing, 1990: 69-78
- [18] Marchetti G M, Wermuth N. Matrix representations and independencies in directed acyclic graphs[J]. Annals of Statistics, 2009, 37(2): 961-978
- [19] VanderWeele T J, Robins J M. Minimal sufficient causation and directed acyclic graphs[J]. Annals of Statistics, 2009, 37(3): 1437-1465
- [20] Heckerman D. A Tutorial on Learning with Bayesian Networks [J]. Studies in Computational Intelligence, 2008, 15: 33-82
- [21] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques[M]. MIT Press, 2009
- [22] Heckerman D E, Nathwani B N. An evaluation of the diagnostic accuracy of Pathfinder[J]. Computers and Biomedical Research, 1992, 25(1): 56-74
- [23] Sajda P. Machine learning for detection and diagnosis of disease [J]. Annual Review of Biomedical Engineering, 2006, 8: 537-565
- [24] Larrañaga P, Moral S. Probabilistic graphical models in artificial intelligence[J]. Applied Soft Computing, 2011, 11 (2): 1511-1528
- [25] Chang L, Duarte M M, Sucar L E. A Bayesian approach for object classification based on clusters of SIFT local features[J]. Expert Systems with Applications, 2012, 39(2): 1679-1686
- [26] Raghuram S, Xia Y, Ge J, et al. AutoBayesian: Developing Bayesian Networks Based on Text Mining[C]// Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2011: 450-453
- [27] Rautenberg W. Complete logical Calculi [M]. A Concise Introduction to Mathematical Logic. New York: Springer 2010: 91-134
- [28] Martin R, Zhang J, Liu C. Dempster-Shafer Theory and Statistical Inference with Weak Beliefs[J]. Statistical Science, 2010, 25 (1): 72-87
- [29] Dubois D, Prade H. Possibility theory and formal concept analysis: Characterizing independent sub-contexts[J]. Fuzzy Sets and Systems, 2012, 196: 4-6
- [30] Ognjanović Z, Perović A, Rašković M. Logics with the Qualitative Probability Operator[J]. Logic Journal of the IGPL, 2008, 16(2): 105-120
- [31] Cozman F G. Graphical models for imprecise probabilities[J]. International Journal of Approximate Reasoning, 2005, 39 (2/3): 167-184
- [32] Cressie N, Lele S. New Models for Markov Random Fields[J]. Journal of Applied Probability, 1992, 29(4): 877-884
- [33] Geiger D, Pearl J. Logical and Algorithmic Properties of Conditional Independence and Graphical Models[J]. Annals of Statistics, 1993, 21(4): 2001-2021

- [34] Wermuth N. Probability distributions with summary graph structure[J]. *Bernoulli*, 2011, 17(3):845-879
- [35] Murthy V S, Gupta S, Mohanta D K. Digital image processing approach using combined wavelet hidden Markov model for well-being analysis of insulators[J]. *Image Processing*, 2011, 5(2): 171-183
- [36] Ye J. Image Filtering with Associative Markov Networks for ECT With Distinctive Phase Origins[J]. *IEEE Sensors Journal*, 2012, 12(7):2435-2443
- [37] Morariu V I, Davis L S. Multi-agent event recognition in structured scenarios[C]// *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011:3289-3296
- [38] Gregor J, Thomason M G. Hybrid pattern recognition using Markov networks[J]. *IEEE. Trans. on Pattern Analysis and Machine and Intelligence*, 1993, 15(6):651-656
- [39] Xiang Y, Lesser V. A constructive graphical model approach for knowledge-based systems: a vehicle monitoring case study[J]. *Computational Intelligence*, 2003, 19(3):284-309
- [40] Xiang Y. Comparing alternative methods for inference in multiply sectioned Bayesian networks[J]. *International Journal of Approximate Reasoning*, 2003, 33(3):235-254
- [41] Richardson M, Domingos P. Markov logic networks[J]. *Machine Learning*, 2006, 62(1/2):107-136
- [42] Andrzejewski D, Zhu X, Craven M. A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-Order Logic[C]// *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. AAAI Press, 2011:1171-1177
- [43] Rabiner L, Juang B. An introduction to hidden Markov models [J]. *ASSP Magazine, IEEE*, 1986, 3(1):4-16
- [44] Dean T, Kanazawa K. A model for reasoning about persistence and causation[J]. *Computational Intelligence*, 1989, 5(2):142-150
- [45] Saul L K, Jordan M I. Mixed Memory Markov Models: Decomposing Complex Stochastic Processes as Mixtures of Simpler Ones[J]. *Machine Learning*, 1999, 37(1):75-87
- [46] Ghahramani Z, Jordan M I. Factorial Hidden Markov Models [J]. *Machine Learning*, 1997, 29(2/3):245-273
- [47] Wainwright M J, Jordan M I. Graphical Models, Exponential Families, and Variational Inference[J]. *Machine Learning*, 2008, 1(1/2):1-305
- [48] Fine S, Singer Y, Tishby N. The Hierarchical Hidden Markov Model: Analysis and Applications[J]. *Machine Learning*, 1998, 32(1):41-62
- [49] Duong T, Phung D, Bui H. Efficient duration and hierarchical modeling for human activity recognition[J]. *Artificial Intelligence*, 2009, 173(7/8):830-856
- [50] Smyth P, Heckerman D, Jordan M I. Probabilistic Independence Networks for Hidden Markov Probability Models[J]. *Neural Computation*, 1997, 9(2):227-269
- [51] Murphy K P. *Dynamic bayesian networks: Representation, inference and learning*[D]. Berkeley: University of California, Computer Science Division, 2002
- [52] Baah G K, Podgurski A, Harrold M J. The Probabilistic Program Dependence Graph and Its Application to Fault Diagnosis[J]. *Software Engineering*, 2010, 36(4):528-545
- [53] Gilks W R, Thomas A, Spiegelhalter D J. A Language and Program for Complex Bayesian Modelling [J]. *The Statistician*, 1994, 43(1):169-177
- [54] Getoor L, Friedman N, Koller D. Learning probabilistic models of link structure[J]. *Machine Learning*, 2003, 3:679-707
- [55] Fersini E, Messina E, Archetti F. Web page classification: a probabilistic model with relational uncertainty[C]// *Proceedings of the Computational intelligence for knowledge-based systems design, and 13th international conference on Information processing and management of uncertainty*. Dortmund, Germany: Springer Berlin Heidelberg, 2010:109-118
- [56] Heckerman D, Meek C, Koller D. Probabilistic entity-relationship models, PRMs, and plate models[C]// *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*. Banff, Canada: IMIS, 2007:55-60
- [57] Oh S, Sastry S. A polynomial-time approximation algorithm for joint probabilistic data association[C]// *Proceedings of American Control Conference*. IEEE, 2005:1283-1288
- [58] Oh S, Russell S, Sastry S. Markov Chain Monte Carlo Data Association for Multi-Target Tracking[J]. *Automatic Control*, 2009, 54(3):481-497
- [59] Culotta A, Wick M, McCallum A. First-order probabilistic models for coreference resolution[C]// *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. New York, USA: the Association of Computational Linguistics, 2007:81-88
- [60] Wellner B, McCallum A, Peng F. An integrated, conditional model of information extraction and coreference with application to citation matching[C]// *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2004:593-601
- [61] Poon H, Domingos P. Joint inference in information extraction [C]// *Proceedings of the 22nd National Conference on Artificial Intelligence*. AAAI Press, 2007:913-918
- [62] Milch B, Marthi B, Russell S, et al. BLOG: Probabilistic models with unknown objects[J]. *Statistical relational learning*, 2007:373
- [63] Kersting K, De Raedt L. Basic principles of learning Bayesian logic programs[C]// *Probabilistic inductive logic programming*. Springer Berlin Heidelberg, 2008:189-221
- [64] Glass D H. Likelihoods and Explanations in Bayesian Networks [J]. *Studies in Computational Intelligence*, 2008, 109:325-341
- [65] VanderWeele T J, Robins J M. Signed directed acyclic graphs for causal inference[J]. *Journal of the Royal Statistical Society; Series B (Statistical Methodology)*, 2010, 72:111-127
- [66] Boutilier C, Friedman N, Goldszmidt M, et al. Context-specific independence in Bayesian networks[C]// *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*. 1996:115-123
- [67] Poole D. First-order probabilistic inference[C]// *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Mexico: Morgan Kaufmann, 2003:985-991
- [68] Li G, Gao J, Chen F. Construction of causality diagram model for diagnostics[C]// *Proceedings of Reliability and Maintainability Symposium*. IEEE, 2008:97-102
- [69] van Gerven M A J, Lucas P J F, van der Weide T P. A generic

- qualitative characterization of independence of causal influence [J]. *International Journal of Approximate Reasoning*, 2008, 48 (1):214-236
- [70] Lerner U, Segal E, Koller D. Exact inference in networks with discrete children of continuous parents[C]// *Proceedings of the 17th conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. ,2001:319-328
- [71] Giang P H, Shenoy P P. Decision making on the sole basis of statistical likelihood[J]. *Artificial Intelligence*, 2005, 165 (2): 137-163
- [72] Daly R, Shen Q, Aitken S. Learning Bayesian networks; approaches and issues[J]. *The Knowledge Engineering Review*, 2011, 26(2):99-157
- [73] Liang P, Jordan M I, Klein D. Learning from measurements in exponential families[C]// *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009:641-648
- [74] Niepert M, Van Gucht D, Gyssens M. Logical and algorithmic properties of stable conditional independence[J]. *International Journal of Approximate Reasoning*, 2010, 51(5):531-543
- [75] Zhang J, Spirtes P. Detection of Unfaithfulness and Robust Causal Inference[J]. *Minds and Machines*, 2008, 18(2):239-271
- [76] Shachter R D. Bayes-ball; The rational pastime [C]// *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. ,1998:480-487
- [77] Daly R, Shen Q. Learning Bayesian network equivalence classes with ant colony optimization[J]. *Journal of Artificial Intelligence Research*, 2009, 35(1):391-447
- [78] Pearl J, Verma T. An algorithm for deciding if a set of observed independencies has a causal explanation [C] // *Proceedings of Eighth Conference in Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. ,1992:323-330
- [79] Biaoletti M, Busanello G, Vantaggi B. Finding P-Maps and I-Maps to Represent Conditional Independencies [J]. *Lecture Notes in Computer Science*, 2011, 6717:239-250
- [80] Chickering D M, Heckerman D, Meek C. A Bayesian approach to learning Bayesian networks with local structure[C]// *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. ,1997:80-89
- [81] Sullivan S. Algebraic geometry of Gaussian Bayesian networks [J]. *Advances in Applied Mathematics*, 2008, 44(4):482-513
- [82] Loeliger H A. An introduction to factor graphs[J]. *Signal Processing Magazine*, 2004, 21(1):28-41
- [83] Frey B J. Extending factor graphs so as to unify directed and undirected graphical models[C]// *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc. ,2003:257-264
- [84] Song H-R, Fuentes M, Ghosh S. A comparative study of Gaussian geostatistical models and Gaussian Markov random field models[J]. *Journal of Multivariate Analysis*, 2008, 99(8):1681-1697
- [85] Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields; the stochastic partial differential equation approach[J]. *Journal of the Royal Statistical Society; Series B (Statistical Methodology)*, 2011, 73: 423-498
- [86] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields; Probabilistic Models for Segmenting and Labeling Sequence Data[C]// *Proceedings of the Eighteenth International Conference on Machine Learning*. ACM, 2001:282-289
- [87] Galley M. A skip-chain conditional random field for ranking meeting utterances by importance[C]// *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006:364-372
- [88] Sutton C, Rohanimanesh K, McCallum A. Dynamic conditional random fields; factorized probabilistic models for labeling and segmenting sequence data [J]. *Machine Learning Research*, 2007, 8:693-723
- [89] Drton M. Discrete chain graph models[J]. *Bernoulli*, 2009, 15 (3):736-753
- [90] Studeny M, Bouckaert R R. On Chain Graph Models for Description of Conditional Independence Structures[J]. *The Annals of Statistics*, 1998, 26(4):1434-1495
- [91] Fersini E, Messina E, Archetti F. Probabilistic Relational Models with Relational Uncertainty; An Early Study in Web Page Classification[C]// *Proceedings of International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*. ACM, 2009:139-142
- [92] Wohlmayr M, Stark M, Pernkopf F. A Probabilistic Interaction Model for Multipitch Tracking With Factorial Hidden Markov Models[J]. *Audio, Speech, and Language Processing*, 2011, 19 (4):799-810
- [93] Malcolm W P, Quadrianto N, Aggoun L. State Estimation Schemes for Independent Component Coupled Hidden Markov Models[J]. *Stochastic Analysis and Applications*, 2010, 18(3): 430-446
- [94] Gales M, Young S. The application of hidden Markov models in speech recognition[J]. *Foundations and Trends in Signal Processing*, 2008, 1(3):195-304
- [95] Quinn J A, Williams C K I, McIntosh N. Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring [J]. *Pattern Analysis and Machine Intelligence*, 2009, 31(9):1537-1551
- [96] Howard C, Stumptner M. Automated compilation of Object-Oriented Probabilistic Relational Models[J]. *International Journal of Approximate Reasoning*, 2009, 50(9):1369-1398
- [97] König J, Nordstrom L, Ekstedt M. Probabilistic Relational Models for assessment of reliability of active distribution management systems[C]// *Proceedings of IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems*. IEEE, 2010:454-459
- [98] Wan H, Lin Y, Jia C, et al. Community-Based Relational Markov Networks in Complex Networks[J]. *Lecture Notes in Computer Science*, 2011, 6954:301-310
- [99] Yoshikawa K, Riedel S, Asahara M, et al. Jointly identifying temporal relations with Markov Logic[C]// *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 2009:405-413