软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2476−2497 [doi: 10.3724/SP.J.1001.2013.04486]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

概率图模型研究进展综述*

张宏毅[1,2], 王立威[1,2], 陈瑜希[1,2]

[1](机器感知与智能教育部重点实验室(北京大学),北京 100871)

[2](北京大学 信息科学技术学院 智能科学系,北京 100871)

通讯作者: 张宏毅, E-mail: hongyi.zhang.pku@gmail.com

摘 要: 概率图模型作为一类有力的工具,能够简洁地表示复杂的概率分布,有效地(近似)计算边缘分布和条件分布,方便地学习概率模型中的参数和超参数.因此,它作为一种处理不确定性的形式化方法,被广泛应用于需要进行自动的概率推理的场合,例如计算机视觉、自然语言处理.回顾了有关概率图模型的表示、推理和学习的基本概念和主要结果,并详细介绍了这些方法在两种重要的概率模型中的应用.还回顾了在加速经典近似推理算法方面的新进展.最后讨论了相关方向的研究前景.

关键词: 概率图模型;概率推理;机器学习

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 张宏毅,王立威,陈瑜希.概率图模型研究进展综述.软件学报,2013,24(11):2476−2497. http://www.jos.org.cn/1000-9825/4486.htm

英文引用格式: Zhang HY, Wang LW, Chen YX. Research progress of probabilistic graphical models: A survey. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2476−2497 (in Chinese). http://www.jos.org.cn/1000-9825/4486.htm

Research Progress of Probabilistic Graphical Models: A Survey

ZHANG Hong-Yi[1,2], WANG Li-Wei[1,2], CHEN Yu-Xi[1,2]

[1](Key Laboratory of Machine Perception (Peking University), Ministry of Education, Beijing 100871, China)

[2](Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

Corresponding author: ZHANG Hong-Yi, E-mail: hongyi.zhang.pku@gmail.com

Abstract: Probabilistic graphical models are powerful tools for compactly representing complex probability distributions, efficiently computing (approximate) marginal and conditional distributions, and conveniently learning parameters and hyperparameters in probabilistic models. As a result, they have been widely used in applications that require some sort of automated probabilistic reasoning, such as computer vision and natural language processing, as a formal approach to deal with uncertainty. This paper surveys the basic concepts and key results of representation, inference and learning in probabilistic graphical models, and demonstrates their uses in two important probabilistic models. It also reviews some recent advances in speeding up classic approximate inference algorithms, followed by a discussion of promising research directions.

Key words: probabilistic graphical model; probabilistic reasoning; machine learning

我们工作和生活中的许多问题都需要通过推理来解决.通过推理,我们综合已有的信息,对我们感兴趣的未知量做出估计,或者决定采取某种行动.例如,程序员通过观察程序在测试中的输出判断程序是否有错误以及需要进一步调试的代码位置,医生通过患者的自我报告、患者体征、医学检测结果和流行病爆发的状态判断患者可能罹患的疾病.一直以来,计算机科学都在努力将推理自动化,例如,编写能够自动对程序进行测试并且诊断

* 基金项目: 国家自然科学基金(61222307, 61075003)
收稿时间: 2013-07-17; 修改时间: 2013-08-02; 定稿时间: 2013-08-27

错误位置的调试工具、能够辅助医生诊断患者病情的医疗诊断专家系统、理解英语文本的含义并将其转换为汉语的自动翻译系统、从机场监控视频中发现可疑人员的安全监控系统等等.人们设计出多种多样的方法来实现这些应用,其中,将知识陈述式地表示为概率模型,通过计算我们所关心变量的概率分布实现推理的途径具有独特优势:

首先,它提供了一个描述框架,使我们能够将不同领域的知识抽象为概率模型,将各种应用中的问题都归结为计算概率模型里某些变量的概率分布,从而将知识表示和推理分离开来[1].模型的设计主要关心如何根据领域知识设计出反映问题本质的概率模型,同时兼顾有效推理的可行性,而推理算法的设计只需关心如何有效地在一般的或者特定的概率模型中进行推理.这种一定程度上的正交性极大地扩展了概率模型的应用,也加快了它的发展速度.

其次,它能够评估未知量取值的可能性,对不同取值的概率给出量化的估计.这在涉及风险的决策系统中非常重要.

另外,它常常具有良好的复用性.例如,我们不需要为预测父亲和儿子患上某种家族遗传病的概率分别设计算法,只需一个关于基因和表现型的家族遗传路径的概率模型,就能处理关于遗传病风险的各种推理问题.

概率图模型就是一类描述这种陈述式表示的概率模型的建模和推理框架,它为简洁地表示、有效地推理和学习各种类型的概率模型提供了可能.在历史上,曾经有来自不同学科的尝试使用图的形式表示高维分布的变量间的相关关系的例子[2,3].在人工智能领域,概率方法始于构造专家系统的早期尝试[4,5].到 20 世纪 80 年代末,由于在贝叶斯网络和一般的概率图模型中进行推理的一系列重要进展[6,7],以及大规模专家系统的成功应用[8],以概率图模型为代表的概率方法重新受到了重视.如今,经过 20 余年的发展,概率图模型的推断和学习已广泛应用于机器学习、计算机视觉、自然语言处理、语音识别、专家系统、用户推荐、社交网络挖掘、生物信息学等研究领域的最新成果中,成为人工智能相关研究中不可或缺的一门技术.概率图模型的研究方兴未艾,而且应用范围和研究热度仍在继续增长.

本文首先介绍概率模型中的推断和学习问题的相关背景,并引入条件独立性这一重要概念;然后,根据研究主题依次综述概率图模型的表示、推理和学习问题核心内容的研究进展;我们还将介绍两种近年来有较大影响的概率图模型——条件随机场和主题模型,以说明概率图模型的表示、推理和学习这 3 个环节的联系;最后,讨论关于大规模图模型的一些延伸主题,包括效率更高的推理算法、并行和分布式推理以及针对查询的推理问题.

在本文中,我们将统一使用大写字母(例如 A,X)表示随机变量,如未指明变量类型,则默认为离散变量;使用小写字母(例如 x,y)表示随机变量的赋值;使用大写字母表示集合(例如 A,X)或者某种数据结构(例如 F,H).

- 推断问题

多数与人工智能相关的应用所解决的问题都可以形式化地表述为概率模型中的推断问题.在推断问题中,目标是推断我们感兴趣的随机变量集合 S 中变量的取值分布,而我们采用生成式模型或判别式模型为问题建模,并运用一般的或针对具体模型的推断算法来计算这一分布.在生成式模型中,我们已知包含感兴趣变量集合在内的一些相互联系的变量的联合分布,以及其中可观测变量的观测值(或真实值),目标是以可观测变量为条件计算目标变量的条件概率.在判别式模型中,我们已知包含感兴趣变量集合 S 在内的一些相互联系的变量与另一些可观测变量之间的联系,即以可观测变量为条件的条件分布,以及可观测变量的观测值,目标同样是计算感兴趣变量集合 S 中的变量的条件概率.

例如,在计算机视觉应用中,人们可能感兴趣一个图像区域所表示的物体类别;在自然语言处理应用中,人们可能感兴趣一句汉语文本的语法分析结果;而在用户推荐应用中,人们可能感兴趣某用户对某产品的喜好程度.这些来自不同领域的问题都可以表示成概率模型中的推断问题,并得到统一的处理.

在以上描述中,要计算感兴趣变量的条件概率,需要知道感兴趣变量及其相关变量包含可观测变量的联合分布或以可观测变量为条件的条件分布.一般情况下,设全体随机变量的集合为 S,感兴趣的变量集合为 $\{X_1,X_2,\ldots,X_n\}=X\subset S$,可观测变量集合为 $\{O_1,O_2,\ldots,O_m\}=O\subset S$,其他变量的集合记为 $Y=S\backslash(X\cup O)$,则生成式模型确定了

联合分布 $P(X,Y,O)$,而判别式模型确定了条件分布 $P(X,Y|O)$.给定观测值,即 O 的一个赋值 $\{o_1,o_2,\ldots,o_m\}$,在生成式模型中,我们需要使用概率求和规则消去 Y 中的变量,并重新归一化,得到条件概率分布 $P(X|O)$,在判别式模型中,我们只需求和消去 Y 中的变量即可.

然而在实际的推断问题中,我们还要考虑到数据结构的表示开销和运算开销(时间和空间复杂度).假设在某模型中,每个变量可以有两种取值,如果我们简单地定义以上概率分布,并使用求和规则推断目标分布,容易验证时间开销和空间开销都至少是 $\Omega(2^{|Y|})$.因此,我们必须寻找更紧凑的表示概率分布的数据结构以及能够在其中有效运行的推断算法.

- 学习问题

推断问题研究如何在已有的模型基础上,根据观测计算目标变量的分布,并没有考虑如何构建模型的问题.一方面,模型可以由领域专家构建,模型的结构以及参数可以由专家根据经验来指定;另一方面,实际应用中可能需要对人类尚不了解的问题建立模型,或建立参数众多的大规模模型,或历史经验以数据的形式而不是人类知识存储等等,在这些情况下,模型的结构和参数并不适合人工指定,因此,我们希望设计算法从已往的数据中学习得到模型的参数和结构.

从更一般的角度来讲,学习问题可以看作是推断问题的一类特例:我们感兴趣的随机变量是模型的参数或结构.因此,对学习问题的简单处理将会遇到与推断问题相同的困难,即表示和计算的时间和空间复杂度关于模型的变量数目都是指数级的,而我们需要能够处理实际应用数据规模的有效的学习算法.

学习问题特有的困难在于:用于学习的训练样本通常是有限的,并且算法允许的训练时间也是有限的.当我们允许复杂的模型尝试从数据中估计联合分布的每一项概率时,我们将面对所谓的维数灾难.相对于呈指数增长的参数,样本量往往太少,以至于我们对真实分布的估计几乎必定有很大的误差.

- 条件独立

考虑变量集 A,B,C,如果 $P(A,B|C=c)=P(A|C=c)=P(B|C=c),\forall c$ 成立,我们就称以 C 为条件,变量集 A,B 相互独立.此时容易验证,$P(A|B,C)=P(A|C),P(B|A,C)=P(B|C)$.模型中的条件独立性是对推断问题和学习问题进行有效计算的基础.例如,考虑对式 $P(A,B,C)$ 求和以消去 B,C,利用上述条件独立性,我们可以写出:

$$\sum_{B,C}P(A,B,C)=\sum_B\sum_C P(A|C)P(B|C)P(C)=\sum_C P(A|C)P(C)\Big(\sum_B P(B|C)\Big).$$

经过变换,在模型的表示上,需要指定的项从 $O(|A||B||C|)$ 个减少到 $O((|A|+|B|)|C|)$ 个,运算次数从 $O(|B||C|)$ 减少到 $O(|B|+|C|)$.注意到,我们可利用集合 A(或 B,C)内的条件独立性进一步简化问题的表示和计算.事实上,如果一个概率分布能够分解为一些包含不超过 d 个变量的项的乘积,且每个变量的可取值不超过 m,则表示和推断的复杂度有上界 $O(m^d)$.其中,d 表示一个复杂的概率分布分解为若干较简单分布的乘积性质的强弱,或者说表示变量之间的条件独立性的强弱.

条件独立性是概率图模型里非常基本的核心概念,贯穿模型的表示、推理和学习等各方面.概率图模型将概率论与图论结合,提供了直观地表示随机变量间条件独立性质的工具,便于人们分析模型的性质,同时使得有关图论的结论和算法可以用于处理概率模型的推断和学习问题[1].

1 概率图模型的表示

概率图模型的表示刻画了模型的随机变量在变量层面的依赖关系,反映出问题的概率结构以及推理的难易程度,也为推理算法提供了可以操作的数据结构.概率图模型的表示方法有多种,我们主要介绍最常见的贝叶斯网络、马尔可夫网络、因子图等表示,以及一些简化表示的记法.

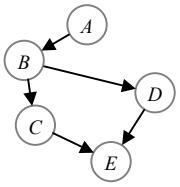
1.1 贝叶斯网络

对应于有向无环图的概率模型称为贝叶斯网络(如图 1 所示).图的每个顶点代表随机变量或随机向量,边代表变量间的条件相关关系,常常也被用于表示因果关系.对于任意一条边和它所连接的两个顶点 $A\to B$,A 称为 B 的父节点,B 称为 A 的子节点.贝叶斯网络中每个顶点 X 和它的父节点 $U(X)$ 表示一个条件分布 $P(X|U(X))$,称为一

个因子.

整个概率分布由所有因子的乘积表示:

$$P(X)=\prod P(X\,|\,U(X)).$$



$P(A,B,C,D,E)=P(A)P(B|A)P(C|B)P(D|B)P(E|C,D)$

Fig.1 Bayesian network of five variables and its union distribution

图 1 含有 5 个变量的贝叶斯网络及其表示的联合分布

1.1.1 贝叶斯网络中的独立性

在贝叶斯网络中,条件独立性可以直接根据图的结构判定.我们首先考虑 3 个变量之间的相互关系.由 X,Y,Z 这 3 个变量构成的 X,Y 不直接相关的贝叶斯网络,X,Y 之间的关系有以下几种形式:

1) X 到 Y 是成因路径($X\rightarrow Z\rightarrow Y$).当且仅当 Z 未被观测时,X,Y 不相互影响(即 X 的取值不影响 Y 的条件分布,反之亦然).因此,X,Y 不相互独立,但给定 Z 时,X,Y 条件独立.
2) X 到 Y 是证据路径($X\leftarrow Z\leftarrow Y$).当且仅当 Z 未被观测时,X,Y 不相互影响.因此,X,Y 不相互独立,但给定 Z 时,X,Y 条件独立.
3) X,Y 有共同原因($X\leftarrow Z\rightarrow Y$).当且仅当 Z 未被观测时,X,Y 不相互影响.因此,X,Y 不相互独立,但给定 Z 时,X,Y 条件独立.
4) X,Y 有共同效果($X\rightarrow Z\leftarrow Y$).当且仅当 Z 或 Z 的一个子代节点被观测时,X,Y 相互影响.因此,X,Y 相互独立,但给定 Z 或其子代节点时,X,Y 不相互独立.

对于贝叶斯网络中的一条路径 $X_1\rightleftarrows\ldots\rightleftarrows X_n$ 和观测变量的子集 $\boldsymbol{Z}$,当 X_1 和 X_n 的取值能够相互影响时,我们称路径 $X_1\rightleftarrows\ldots\rightleftarrows X_n$ 是有效的.在一般情况下,我们有以下结论:给定 $\boldsymbol{Z}$ 时,$X_1\rightleftarrows\ldots\rightleftarrows X_n$ 是有效的当且仅当:

1) 对该路径上的每个 V 字结构 $X_{i-1}\rightarrow X_i\leftarrow X_{i+1}$,存在 X_i 或其子代在 $\boldsymbol{Z}$ 中;
2) 路径上的其他任何节点都不在 $\boldsymbol{Z}$ 中.

1.2 马尔可夫网络

注意到除了用若干条件概率分布的乘积构造联合分布以外,还有更一般的构造概率分布的方法.考虑定义在随机变量集合 X 的子集 Ψ 变量值域上的非负实值函数 ϕ,若对任意 $i,\phi_i(\Psi_i)\geqslant 0,\bigcup_i\Psi_i=X$ 且 $Z=\sum_X\prod_i\phi_i(\Psi_i)>0$,则 $\frac{1}{Z}\prod_i\phi_i(\Psi_i)$ 定义了一个有效的分布.其中,Z 称为归一化常数,由模型参数确定而与观测值无关.

对应于无向图的概率模型称为马尔可夫网络,图的每个顶点代表随机变量或随机向量,边代表变量间的相关关系.对任意一条边,其所在的最大的团(全连通子图)称为一个因子.每一条边都唯一地属于一个因子,由于有了上述构造概率分布的方法,只需为每个因子 Ψ 指定一个非负函数 $\phi(\Psi)$,并对所有因子的乘积归一化,我们就可以构造出由马尔可夫网络表示的概率模型.

归一化常数是马尔可夫网络与贝叶斯网络的重要区别之一.在许多模型中,直接计算归一化常数的复杂度是指数级的,因此必须寻找其他替代方法解决推断或学习问题.

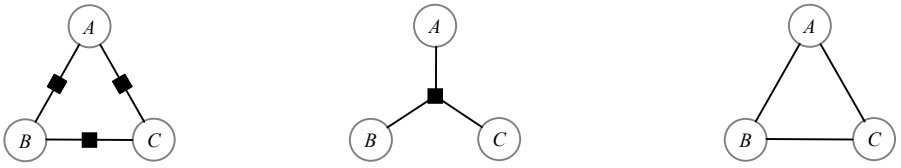
1.2.1 马尔可夫网络中的独立性

马尔可夫网络中的独立性情况比贝叶斯网络要简单.与之前定义类似,对马尔可夫网络中的一条路径 X_1-…-X_n 和观测变量的子集 $\boldsymbol{Z}$,当 X_1 和 X_n 的取值能够相互影响时,我们称路径 X_1-…-X_n 是有效的.在一般情况下,我们有以下结论:给定 $\boldsymbol{Z}$ 时,X_1-…-X_n 是有效的,当且仅当路径上的其他任何节点都不在 $\boldsymbol{Z}$ 中.

1.3 因子图

图模型的另一种常见表示是因子图[9].在因子图 $G=(X,F)$中,我们规定存在两种顶点:随机变量和因子.边是无向的,连接因子和它所包含的随机变量.因此,每个随机变量的邻居顶点是包含它的各个因子,而每个因子的邻居顶点是它的辖域内的各个随机变量.

因子图是一种比马尔可夫网络更精细的模型表示,例如在图 2 所示的因子图中,我们可以区分不同的分布究竟是定义为因子 $\Psi_{AB},\Psi_{BC},\Psi_{AC}$ 的乘积还是一个辖域为 A,B,C 的大因子 Ψ_{ABC};而在马尔可夫网络中,它们有相同的表示而无法从图结构上进行区分.



(a) 含有 3 个二元因子的因子图 (b) 仅含有 1 个三元因子的因子图 (c) 对应的马尔可夫网络仅有 1 种表示

Fig.2

图 2

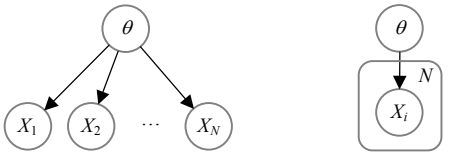
与马尔可夫网络相似,因子图中也有马尔可夫网络的概念.

从马尔可夫网络的特征和因子图的定义中容易推导,在因子图中,设边集为 E,变量 X_i 的马尔可夫网络是由 $B(X_i)=\{X_j:(X_j,\alpha)\in E$ 且 $(X_i,\alpha)\in E$ 且 $j\neq i\}$ 构成的变量集合.

因子图数据结构由于显式地表示出构造概率分布的因子,因此特别适合一类通过在变量与因子之间传递消息的推理算法(参见第 2.2.1 节)的执行.包括微软的 Infer.NET 项目[10]在内,许多采用这类算法的推理系统都采用了因子图的表示.

1.4 盘式记法、模板

盘式记法(plate notation)是一种常用的图模型的简化记法,考虑如下简单的概率模型:在一个装有白球和黑球的盒子里进行有放回抽样,n 次抽样的结果记为 $X_1,\ldots,X_n$,盒子中白球的比例记为 θ,则该概率模型可以采用标准的记法表示,也可以用盘式模型表示(如图 3 所示).在盘式模型中,我们用一个框(称为盘)圈住图模型中重复的部分,并在框内标注重复的次数.



(a) 展开的摸球模型 (b) 盘式记法

Fig.3

图 3

盘式记法能够为我们表示和分析许多概率模型提供很大的方便,但它也有一定的局限性.例如,它无法表示盘内变量不同拷贝间的相关性,而这种相关性广泛出现于动态贝叶斯网络中.在动态贝叶斯网络中,概率模型符合 k 阶马尔可夫假设,因此,t 时刻的系统状态只与[max(0,$t-k$)]时刻的状态有关.为了简明地表示这类含有重复的基本结构的动态贝叶斯网络,我们可以使用称为 kBN 的模板.例如,图 4 展示了隐马尔可夫模型及其 2BN 模板记法.

(a) 展开的隐马尔可夫模型 (b) 2BN-记法

(该记法要求使用盘分别圈住 t 时刻的变量和它们的父节点变量)

Fig.4

图 4

2 概率图模型的推理

在本节中,我们介绍图模型推理所使用的一些重要算法.这些算法可大致分为 3 类:

- 精确推理算法.能够计算查询变量的边缘分布或条件分布的精确值,但其计算复杂度随着图模型的树宽呈指数增长,因此适用范围有限.
- 基于优化的推理算法.将图模型的推断形式化为优化问题,通过松弛优化问题的约束条件或者调整优化的目标函数,算法能够在较低的时间复杂度下获得原问题的近似解;对于多数难以进行精确推理的图模型,对其变量进行抽样并不困难.
- 基于抽样的推理算法.试图产生所求边缘分布或条件分布的抽样,使用样本的经验分布来近似真实分布.

一般的图模型中的推理是困难的.关于最坏情况下推理的复杂度,我们已经知道一系列负面的结果:图模型精确推理是#*P*-完全的;相对误差ε的图模型近似推理是 NP-难的;绝对误差$\varepsilon\in(0,1/2)$的图模型近似推理有多项式的时间复杂度,但以证据为条件的近似推理依旧是 NP-难的[11].然而,实际应用中的问题往往具有良好的条件独立结构,因而图模型的精确推理和近似推理算法带来了许多成功的应用.

从推理的目标来区分,有时我们希望计算目标变量的(条件)边缘分布,称为后验推理;有时我们希望计算目标变量(集合)最可能的(联合)赋值,称为最大后验推理.我们将首先介绍 3 类后验推理方法,最后统一介绍最大后验推理方法.

由于贝叶斯网络中的条件分布 $P(X|Y)$也可看作因子$\psi(X|Y)$,本节中,我们将统一地采用因子ψ来构造分布.

2.1 精确推理

图模型的精确推理算法实质是一类动态规划算法,它利用条件独立性导致的联合分布的因子化特征,压缩图模型变量之间传递的信息.

2.1.1 变量消去算法

变量消去算法[12]是最直观且容易推导的精确推理算法,然而它却是其他更高级的精确推理算法的基础.考虑贝叶斯网络(如图 1 所示)所定义的联合分布 $P(A,B,C,D,E)$,我们如果观测到变量 $E=e$,并且想要计算变量 $C=c$ 的条件概率 $P(c|e)$,则可以简单地写出:

$$P(c\mid e)=\frac{1}{Z}\sum_{a,b,d}P(a,b,c,d,e),$$

其中, $Z=\sum_{a,b,c,d}P(a,b,c,d,e)$.

利用贝叶斯网络所蕴含的独立性,以上求和可分别化简为

$$\sum_{a,b,d}P(a,b,c,d,e)=\sum_{a,b,d}P(a)P(b\mid a)P(c\mid b)P(d\mid b)P(e\mid c,d)=\sum_{d}\left(P(e\mid c,d)\sum_{b}\left(P(c\mid b)P(d\mid b)\sum_{a}P(a)P(b\mid a)\right)\right) \quad (1)$$

简而言之,变量消去算法实际是利用了乘法对加法的分配律,将对多个变量的积的求和分解为对部分变量

交替进行的求积与求和.变量消去算法的缺点在于:一次变量消去只能求出本次查询变量的条件分布,不同的查询将带来大量的重复计算.为实现一次计算、多次使用,就需要更加仔细地进行算法设计.为了保证算法的正确性,需要保证所有与求和变量有关的项都在求和符号内,这也是以下两种常用精确推理算法的理论基础.

2.1.2 团树算法

为了实现精确推理的更高效算法,需要设计一种支持推理运算的数据结构:团树.团树是一种具有如下性质的树结构:

1) 团树的每个顶点都是对应图模型中因子的集合.
2) 族保持性:原图模型中的每个因子都属于至少 1 个团树顶点.
3) 流动相交性:如果变量 X 在团树的两个顶点 $\boldsymbol{C}_i,\boldsymbol{C}_j$ 中出现,则 X 出现在 $\boldsymbol{C}_i,\boldsymbol{C}_j$ 之间的每一个顶点中.

团树中,相邻两顶点的变量的交 $\boldsymbol{S}_{ij}=\boldsymbol{C}_i\cap\boldsymbol{C}_j$ 称为两顶点的割集.

从同一图模型可以生成多种团树,它们的推理代价并不相同,但寻找推理代价最小的团树是困难的.一个比较好的方法是在所谓"最大消去团"的全连通图上寻找好的团树.指定一个变量消去的顺序,考虑依次消去变量时所生成的团,以其中的最大团为顶点生成全连通图,用割集的基数为边赋权,则该全连通图的最大权生成树就是一个较好的团树.这是因为在变量消去的过程中,我们只需要消去未出现在割集中的变量,因此对于给定的顶点集合,割集中包含的变量较多,计算代价相对就较小.

团树生成之后,需要计算团树顶点对应的因子.为此,只需对每个顶点遍历原图模型因子的集合,若该因子所含变量是该团树顶点所含变量的子集,则将其乘入该顶点的因子中.团树对应的因子集合称为可分解密度,这是因为它们是原概率分布在团树上分解的结果.

有了团树和可分解密度,就可以执行以下精确推理算法了:

- 消息传递(也称 Shafer-Shenoy 算法[13]、和-积算法).

在消息传递算法中,当且仅当节点 A 收到了除邻居节点 B 外所有邻居节点的消息时,A 才向 B 传递消息.消息传递算法按如下步骤执行:

1) 在团树中选择一个根节点,由根节点出发构造一个单根树.
2) 从根节点起递归地执行:如果该节点的子节点收到了其他所有邻居的消息,则子节点向该节点传递消息.
3) 从根节点起递归地执行:如果该节点收到了除某个子节点外其他所有邻居的消息,则该节点向子节点传递消息.

- 信念传播(也称 HUGIN 算法[14]).

在信念传播算法中,我们为每个节点附加一个称为信念的变量.在算法运行的每一步,每个节点都同时向其相邻节点传递消息,消息的内容是该节点当前关于其割集中变量的边缘分布的信念.每个节点在接收到来自其他节点的消息后立即更新其信念:将自身的信念乘以新接收的消息,并除去上一次从相同节点接收到的消息.如果任一节点在收到其邻居节点的消息后自身的信念都不再发生变化,算法就收敛了.可以证明:对于团树上的信念传播,每对节点间传递两次消息后,算法就会收敛,且收敛于正确的分布(可能需要归一化).

2.2 基于优化的近似推理

在上一节中,我们介绍了用于精确推理的信念传播算法,其中使用到了团树这一数据结构.团树可以在对原始的图模型进行变量消去的过程中生成,团树上信念传播的复杂度是团树中最大团所含变量数目的指数量级.而如果我们对于在计算机视觉等领域常用的伊辛(Ising)模型生成团树(如图 5 所示,其中,灰色加粗的变量集合为变量 $X_{2,2}$ 的马尔可夫毯),设图模型的变量数量为 $\Theta(N^2)$,则一个团中将含有 $\Omega(N)$ 个变量,精确推理的复杂度是 $2^{\Omega(N)}$.因此,对于具有这类变量耦合性质的模型,团树上的信念传播算法就不适用了.

通过对团树信念传播算法的进一步分析人们发现:它可以表示成一个关于团节点信念的优化问题的求解算法,其约束就是节点信念关于割集中变量的边缘分布一致.可以证明,这也保证了联合分布整体的一致性.类似地,我们也可以把其他图模型(例如伊辛模型)中更困难的推理表示成优化问题,考虑其近似求解方案,对应地

设计出推理算法.

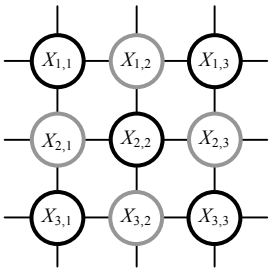


Fig.5 Representation of graph models of Ising model

图 5 伊辛(Ising)模型的图模型表示

2.2.1 环路信念传播

环路信念传播(loopy belief propagation,简称 LBP[6])是团树信念传播方法的自然推广.从算法上看,LBP 与团树算法的区别仅在于它放松了团树算法所要求的流动相交性这个性质,因此允许在有环图上执行算法.保留环路的结果,是从节点 $\boldsymbol{C}_i$ 到节点 $\boldsymbol{C}_j$ 关于 X 变量的信息可能影响了 $\boldsymbol{C}_j$ 对 Y 的信念,既而又传播回 $\boldsymbol{C}_i$ 并对其关于 X 的信念造成反馈.这可能导致 X 的信念无法收敛,或者收敛到错误的分布.

尽管环路信念传播的运行结果取决于具体模型及其参数,缺乏理论上好的保证,但在实际中其表现常常令人满意.在编解码问题研究中,著名的 Turbocode 解码算法就被证明与环路信念传播算法等价[15].

可以证明,在伊辛模型上直接运行环路信念传播算法,等价于求解原优化问题的一个近似[16]:目标函数等价于优化 Bethe 近似的能量函数,而约束条件仅要求割集变量的边缘分布一致.在环路图中,这并不保证联合分布的整体一致.

2.2.2 平均场近似

通过对可行域施加更多约束条件,平均场(mean field)算法简化优化目标函数,从另一个角度解决在类似伊辛模型的高树宽模型中近似推理的问题.例如对于伊辛模型,原优化问题为

$$A(\theta)=\sup_{\mu\in M}(\langle\theta,\mu\rangle+H(p_{\theta(\mu)})) \tag{2}$$

其中,θ是模型参数;μ是对应的因子在该参数下的充分统计量的期望;M 称为边缘(概率)多面体剖分,是所有合理的边缘分布的集合;$\langle\cdot,\cdot\rangle$为内积,$H(\cdot)$是分布的熵函数,而 $p_{\theta(\mu)}$是充分统计量期望μ对应的模型参数θ所定义的分布.平均场算法采用 $\hat{M}=\{\mu:\mu_{u,v}=\mu_u\mu_v\}$ 替代实际可行域 M,这相当于在去掉原伊辛模型所有边连接的一类模型中,寻找伊辛模型的最佳近似.从而,其对应的熵函数化简为易于优化的形式:

$$H(p_{\theta(\mu)})=-\sum_{v\in V}(\mu_v\log\mu_v+(1-\mu_v)\log(1-\mu_v)) \tag{3}$$

新的优化问题对每个μ_v 分别是凹的,因此可以采用迭代地关于每个μ_v 进行优化的方法求解,得到μ的更新方程为

$$\mu_v=\frac{1}{\exp\left(-\theta_v-\sum_{u\in N(v)}\theta_u\mu_u\right)} \tag{4}$$

当算法收敛时,得到的μ_v和 $A(\theta)$就是原问题解的平均场近似.

2.3 基于抽样方法的推理

当精确推理计算困难时,抽样方法提供了计算随机变量边缘(条件)分布和其他函数的另一种途径[17].抽样方法的种类很多,包括拒绝性采样、重要性采样[18−20]、粒子滤波[21]、马尔可夫链蒙特卡罗(MCMC)[22,23]算法等.与变分推理方法不同,基于马尔可夫链的抽样推理方法在一定条件下,理论上能够渐近地收敛于真实分布;不仅如此,它还具有普遍的适用性.但对复杂的推理问题,马尔可夫链的混合时间通常无法估计,其实际效率也往往

低于针对具体模型设计的变分优化推理算法.

2.3.1 拒绝性采样

拒绝式采样意指与观测不一致的采样将被拒绝,它是贝叶斯网络中采样的最简单方法.考虑依条件概率分布从图模型 G 中抽样,则得到样本 x 的概率为 $P(x)$.若观测变量 $X_e=x_e$,查询变量 X_q,则有:

$$P(x_q \mid x_e)=\frac{\sum_{x\setminus\{X_q\}} P(x)\mathrm{I}[X_e=x_e]}{\sum_{x} P(x)\mathrm{I}[X_e=x_e]} \tag{5}$$

即在所有与观测一致的抽样中,查询变量 $X_q=x_q$ 的比例即其取值 x_q 的条件概率.注意到接受概率 $1[X_e=x_e]$ 关于观测变量的数目呈指数趋于 0,因此,拒绝性采样不适用于连续变量以及有大量观测的模型.

2.3.2 重要性采样

重要性采样意指首先获得服从另一分布的样本,再根据重要性对样本加权,以使样本等效于来自目标分布.考虑欲估计关于 X 的函数 $f(X)$,抽样分布 $q(X)$,目标分布 $p(X)$,则由

$$E_p(f(X))=\int f(X)\frac{p(X)}{q(X)}q(X)\mathrm{d}X\approx\frac{1}{M}\sum_{i=1}^{M}f(X_i)w_i,\ w_i=\frac{p(X_i)}{q(X_i)} \tag{6}$$

可依分布 $q(x)$抽样得到$\{X_1,\ldots,X_M\}$,计算 w_i 进行重要性加权,即可得到 $E_p(f(X))$的估计.对一般的图模型,往往只知道 $\tilde{p}(X)=p(X)Z_p,\tilde{q}(X)=q(X)Z_q$, 而不知道其归一化常数.此时,可利用:

$$E_p(f(X))=\frac{Z_q}{Z_p}E_q\left[f(X)\frac{\tilde{p}(X)}{\tilde{q}(X)}\right],\ \frac{Z_p}{Z_q}=E_q\left[\frac{\tilde{p}(X)}{\tilde{q}(X)}\right],$$

得到:

$$E_p(f(X))=\frac{E_q\left[f(X)\frac{\tilde{p}(X)}{\tilde{q}(X)}\right]}{E_q\left[\frac{\tilde{p}(X)}{\tilde{q}(X)}\right]}\approx\frac{\sum_{i=1}^{M}f(X_i)w_i}{\sum_{i=1}^{M}w_i} \tag{7}$$

其中, $w_i=\frac{p(X_i)}{q(X_i)}$.

因此,同样可以从样本中获得关于目标分布的函数值估计.

在应用中,虽然只要求 $q(x)>0,x\in\{y:p(y)>0\}$,但若 $q(x)$与 $p(x)$差别较大,估计的误差也往往较大.结合退火等技术发展出的退火重要性采样(annealed importance sampling,简称 AIS)可以减少估计误差.

似然度加权也是一种常用的图模型抽样算法.它可以看作重要性采样的一个特例.在似然度加权中,样本及其权重按如下算法产生:

算法 1. 似然度加权.

输入:观测集合(E,x_E).

输出:样本 x 以及重要性权值 w.

(1) $w\leftarrow 1$

(2) **foreach** 拓扑排序中的第 i 个变量 **do**

(3) **if** $i\in E$ **then**

(4) $X_i\leftarrow x_i$

(5) $w\leftarrow wp(x_i|X_{\pi(i)})$

(6) **else**

(7) 依 $X_i\sim p(X_i|X_{N(i)})$抽样

(8) **return** (x,w)

2.3.3 MCMC 算法

考虑随机向量 X 状态空间上的随机过程$\{X^{(1)},\ldots,X^{(t)}\}$,如果适当选择状态转移矩阵并作用于 X,则可构造出马尔可夫链使 $X^{(t)}$的分布收敛于后验分布.因此,这一类算法称为 MCMC 算法.由不同的状态转移方法确定的 MCMC 算法具有不同的性质.下面主要介绍最简单、最常见的两种算法:Gibbs 抽样和 Metropolis-Hastings 算法.

- Gibbs 抽样

Gibbs 抽样算法假定对任意变量 X_i,我们可以从 $p(x_i|x_{-i})$中抽样.算法的每一步依次选择一个变量,以其他变量的当前值为条件,从该变量的条件分布中抽样更新其值.

Gibbs 抽样的优势在于算法简单,缺点在于一次只能更新一个变量,效率较低.改进的联锁(blocked)Gibbs 抽样在每一步能够对相互条件独立的变量同时抽样.但当从 $p(x_i|x_{-i})$中抽样比较困难时,Gibbs 抽样就不适用了.Gibbs 抽样可以看作下面介绍的 Metropolis-Hastings 算法的一个特例.

- Metropolis-Hastings 算法

Metropolis-Hastings 算法具有更大的灵活性.它只假定我们可以产生遍历状态空间的马尔可夫链,并且能够计算样本的生成概率和比较不同样本的似然比.我们可以为算法设计不同的建议分布(proposal distribution)Q,通过控制接受概率(acceptance probability),使状态转移分布对目标分布满足细致平衡(detailed balance)性质.具体地,设 $Q(x\to x')$表示由状态转移函数从 x 生成状态 x'的概率密度,$A(x\to x')$表示状态转移 $x\to x'$的接受概率,则由细致平衡原理有:

$$\frac{A(x\to x')}{A(x'\to x)}=\frac{p(x')Q(x'\to x)}{p(x)Q(x\to x')} \tag{8}$$

取接受概率为 $A(x\to x')=\min\left(1,\frac{p(x')Q(x'\to x)}{p(x)Q(x\to x')}\right)$,即可使上式成立.

通常,我们希望建议分布 Q 的选择能够使算法有较高的接受概率,同时又能较快而全面地探索状态空间的不同区域.但实际中,算法的表现往往依赖于概率分布的复杂程度以及算法设计者对问题的理解.一些高级的 MCMC 算法,例如推广的 Swendsen-Wang 算法[24,25]以及 Hamiltonian Monte Carlo[26]方法,针对不同的方面对基本算法进行了改进.

3 概率图模型的学习

对于复杂的、缺乏专家经验的概率模型,如果我们有一定数量的观测数据,通常希望能够从观测数据中获得模型的参数甚至结构.这又分为几种不同的情况:贝叶斯网络或马尔可夫网络;所有变量都可观测的模型或只可观测部分变量的模型;结构已知的模型或结构未知的模型.不同的情况产生出一系列不同难度的问题.

在参数估计的不同准则中:一种常用的估计方法是最大似然估计(maximum likelihood estimation,简称 MLE),即在参数空间中寻找使观测数据的似然最大的参数;另一种称为最大后验(maximum a posteriori,简称 MAP)的估计方法寻找参数空间中后验概率最大的参数.以下为叙述简单,我们只介绍最大似然估计的方法,对应的最大后验估计可通过相同方法导出.在有缺失数据的参数估计中,我们还将介绍视参数为潜在的随机变量,估计其后验分布的贝叶斯推断方法.

3.1 完全可观测模型的参数估计

由于已知的信息最多,完全可观测模型的学习问题是各类学习问题中最简单的.然而即使在这种情况下,贝叶斯网络和马尔可夫网络的学习算法代价也有着本质的不同.我们将会看到:贝叶斯网络的参数估计问题是可分解的,因而容易求解;但是马尔可夫网络由于其因子缺少局部的归一化性质,由全局归一化常数带来的模型参数之间的耦合增大了学习的难度.

3.1.1 估计贝叶斯网络的参数

对于贝叶斯网络 G,记其变量集合为 X,参数为θ,观测样本为 $D=\{X^{(1)},\ldots,X^{(N)}\}$,则似然函数为

$$L(\theta;D)=\prod_{i=1}^{N}\prod_{X_j\in x}p(x_j^{(i)}\mid u(x_j)^{(i)};\theta_j) \tag{9}$$

最大化似然函数等价于最大化对数似然,即

$$\max_{\theta} l(\theta;D)=\sum_{X_j\in x}\max_{\theta_j}\sum_{i=1}^{N}\log p(x_j^{(i)}\mid u(x_j)^{(i)};\theta_j) \tag{10}$$

由此可见,在完全可观测的贝叶斯网络中,最大似然参数估计问题可分解为对每个条件概率密度的参数估计问题.

3.1.2 估计马尔可夫网络的参数

对于马尔可夫网络 H,一般研究一类称为对数线性模型(log-linear model)的模型比较方便.记其因子集合为 $\boldsymbol{C}$,参数为 θ,对数线性模型的似然定义为

$$p(\boldsymbol{y}\mid\theta)=\frac{1}{Z(\theta)}\exp\left(\sum_{c}\theta_c^T\varphi_c(\boldsymbol{y})\right) \tag{11}$$

可以证明,完全可观测的对数线性模型的对数似然函数关于参数 θ 是凸的,因此可以采用梯度方法求解最大似然估计问题.对数似然函数关于参数的偏导为

$$\frac{\partial l(\theta)}{\partial\theta_c}=\left[\frac{1}{N}\sum_{i}\phi_c(\boldsymbol{y}_i)\right]-E_{y|\theta}[\phi_c(\boldsymbol{y})] \tag{12}$$

其中,第 1 项可以从训练集的经验分布估计;第 2 项需要在当前模型中进行推理,对于低树宽的模型,可以采用精确推理的方法求得此项;而对于高树宽的模型,则需要采用上文介绍的近似推理方法.一种常用的方法是对模型中的目标变量进行抽样,以在抽样经验分布中的均值近似上式第 2 项,从而求得近似 t 的梯度.对于大训练集,每次迭代求解上式第 1 项也有较大的时间开销,因此可以用小批训练样本(minibatch)诱导的分布近似整个训练集的经验分布求第 1 项.由于梯度下降第 t 步与 t+1 步的参数一般差异不大,可以用第 t 步的抽样初始化第 t+1 步的 MCMC 算法,从而加快马尔可夫链的混合速度.随机最大似然方法综合了这些近似技巧.

另一种称为最大化伪似然的典型方法试图回避由归一化常数导致的参数估计的困难.该方法定义对数伪似然为

$$l_{PL}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}\log p(y_{id}\mid\boldsymbol{y}_{i,-d},\theta) \tag{13}$$

修改参数估计的目标为最大化所有条件分布之积(又称组合似然).在新的目标函数下,可以与在贝叶斯网络中类似地将参数估计问题分解为每个条件分布的参数估计问题.在高斯马尔可夫随机场中,最大伪似然与最大似然目标等价,然而这种等价性对一般模型并不成立.

算法 2. 拟合 MRF 的随机最大似然算法.

(1) 随机初始化权值 θ

(2) k=0, η=1

(3) **foreach** 轮次 **do**

(4) **foreach** 大小为 B 的小批训练样本 **do**

(5) **foreach** s=1:S **do**

(6) 抽样 $\boldsymbol{y}^{s,k}\sim p(\boldsymbol{y}\mid\theta_k)$

(7) $\hat{E}(\phi(\boldsymbol{y}))=\frac{1}{S}\sum_{s=1}^{S}\phi(\boldsymbol{y}^{s,k})$

(8) **foreach** 小批训练样本中的第 i 个样本 **do**

(9) $\boldsymbol{g}_{ik}=\phi(\boldsymbol{y}_i)-\hat{E}(\phi(\boldsymbol{y}))$

(10) $\boldsymbol{g}_k=\frac{1}{B}\sum_{i\in B}\boldsymbol{g}_{ik}$

(11) $\theta_{k+1}=\theta_k-\eta \boldsymbol{g}_k$

(12) $k=k+1$

(13) 减小步长η

3.2 有缺失数据的参数估计

对于某一变量,数据缺失分为两种情况:部分训练样本的数据缺失以及数据完全缺失(隐变量).我们主要讨论模型含有隐变量的情况.在有缺失数据的情况下,即使对于贝叶斯网络,最大化数据集的似然函数的问题也不能分解成更小的问题.一般有如下两种参数估计的方法:

- 梯度上升算法是优化非线性目标函数的标准方法.算法迭代的每一步计算数据的似然关于模型参数的偏导,并在梯度方向上升.由于似然函数有界,算法将收敛于一个局部最优解.
- 期望-最大化算法(又称 EM 算法)是一种处理含有缺失数据的最大似然估计问题的专门方法.该方法的思路来源于两个较为简单的问题:给定模型真实参数和可观测变量,推理隐变量的分布,是标准的推理问题;给定可观测变量和隐变量的真实值,求最大似然参数估计,是标准的学习问题.基于以上想法,算法将求解最优参数的过程分解为交替进行的两个步骤:在“期望”步骤,我们根据模型参数的当前估计推理,求得隐变量的分布,以此获得(对数)似然函数关于隐变量分布的期望;在“最大化”步骤,我们求解最大化(对数)似然函数期望的模型参数,得到参数的新估计值.可以证明,每个步骤均能增大似然函数,因此,算法将收敛于一个局部最优解.

然而,与完全可观测模型的似然函数不同,在有隐变量的情况下,目标函数一般有多个局部极大值.为解决算法陷于局部极大值的问题,有时采用退火的方法,或者对训练数据加入扰动(pertubation).但是在实际中,依然无法保证算法能够找到最大似然估计的全局最优解.

对于部分训练样本缺失某变量数据的情形,还需要考虑数据缺失事件是随机发生的,还是与当前模型的其他变量有关:若为前者,则可以用对待隐变量的类似方法处理;若为后者,则需要做进一步的修正.

3.3 结构学习

从数据中推断变量之间的依赖关系(即图模型结构)的问题称为结构学习.结构学习可以用于发现变量之间的依赖关系,例如发现生物体不同基因表达的依赖关系.此外,通过结构学习获得数据的合理模型,还可以用来对数据做密度估计.具有稀疏性的结构有助于防止密度估计的过拟合问题.我们主要介绍贝叶斯网络的结构学习问题.马尔可夫网络的结构学习方法与之相近,可以参考文献[1]中第 20 章第 7 节的相关内容.

贝叶斯网络结构学习的方法大体可分为 3 类,下面分别进行介绍.

3.3.1 基于约束的方法

若数据产生于某真实模型 G^*,则对于 G^*中条件独立的变量(集合),在训练数据中也应当近似条件独立;相反,若 G^*完整地反映了数据的条件独立关系,则 G^*中条件依赖的变量在训练数据中也很可能是条件依赖的.这构成了基于约束的方法的依据.具体说来,在关于变量间条件独立性的统计测试基础上,该方法对训练数据集进行一系列的条件独立性查询.在一定的假设下[1],该方法能够使用多项式个关于有限多变量的条件独立性查询找到最优的图模型结构.

3.3.2 基于得分的方法

基于得分的方法将结构学习问题转化为最优化问题处理.该方法定义某种得分函数,根据训练样本为每一种候选结构打分,搜索得分较高的结构.选择得分函数时,我们首先可能会想到采用似然函数——例如对于图模型 G,首先对参数做最大似然估计,用数据集的对数似然作为 G 的得分.然而,这种称为似然得分的方法会导致过拟合——一般来说,复杂的模型将获得更高的得分.

贝叶斯得分能够有效地避免过拟合训练数据的问题.该得分函数采用模型 G,以证据 x 为条件的后验概率作为 G 的得分,为此,需要指定关于模型结构 G 和参数θ_G的先验分布,并对参数的所有取值求积分.因此,这一得分只在具有特定分布的简单模型上适用.

贝叶斯信息准则(Bayesian information criterions,简称 BIC)得分是贝叶斯得分的一种简化和近似,可以避免计算对模型参数空间的多重积分.可以证明,若假设贝叶斯网络的变量都是离散的,且采用 Dirichlet 分布作为参数的先验分布,则随着样本数量趋向无穷,BIC 得分:

$$score_{BIC}(G:D)=l(\hat{\theta}_g:D)-\frac{\log M}{2}Dim[G] \tag{14}$$

与贝叶斯得分之差将趋向于一个常数.BIC 得分还具有统计一致性,即随着样本数量趋向无穷,真实的模型结构(以及与其有等价条件独立性质的结构)将获得最大的 BIC 得分.

选定了打分规则,还需要解决搜索得分较高的图模型结构的问题.对于树结构的图模型,有 $O(n^2)$时间的算法找到最优树结构;对于给定了序关系的变量集合,找到最优结构的计算复杂度大致是图模型节点最大入度的指数;对于一般的图模型,则问题常常是 NP-难的,一般采用启发式的方法近似求解.一种常用的称为爬山(hill-climbing)法的方法,通过反复试探增边、删边、逆转边的方向等基本操作,寻找具有更高得分的结构.

3.3.3 贝叶斯模型平均

如上所述,当训练样本趋于无穷时,模型的后验概率将集中于真实模型和它的等价模型;然而,当训练样本较少时,模型结构的后验分布往往并不集中.这时,使用最大后验的模型来代表我们对模型后验的估计就未必合适了.我们可以采用 MCMC 方法,构造在不同结构的图模型之间转换的马尔可夫链,设计 Metropolis-Hastings 算法的接受概率,使其平稳分布恰好是模型结构的后验分布,利用马尔可夫链抽样的经验分布估计模型结构.

4 代表模型

本节中,我们将介绍两类在近 10 年中有较大学术影响且获得广泛应用的概率图模型:

- 条件随机场[27]是马尔可夫随机场的一种变体,其相对于马尔可夫随机场的优势在于不需要对输入进行建模,可以综合各种特征、打分函数,因此能够方便地与特征提取、检测器、分类器等输出的中间结果结合起来.
- 主题模型[28]是一类文档聚类和文本分析的概率模型.它提出了更合理的模型假设,弥补了传统聚类方法的种种不足,提供了更精细的分析结果,且有有效的近似推理算法.

4.1 条件随机场

条件随机场(conditional random field,简称 CRF)是一类判别式概率图模型,在句法分析和计算机视觉等问题中有广泛的应用.具体来说,条件随机场是一种所有的势函数都以输入特征为条件的马尔可夫随机场:

$$p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{w})=\frac{1}{Z(\boldsymbol{x},\boldsymbol{w})}\prod_c\psi_c(\boldsymbol{y}_c\mid\boldsymbol{x},\boldsymbol{w}) \tag{15}$$

通常,我们假设势函数关于输入特征有对数线性的表示:

$$\psi_c(\boldsymbol{y}_c\mid\boldsymbol{x},\boldsymbol{w})=\exp(\boldsymbol{w}_c^T\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}_c)) \tag{16}$$

其中,$\varphi(\boldsymbol{x},\boldsymbol{y}_c)$是关于输入 $\boldsymbol{x}$ 和标签 $\boldsymbol{y}_c$ 的特征向量.由此可以写出条件随机场的对数似然函数:

$$l(\boldsymbol{w})\triangleq\frac{1}{N}\sum_i\left[\sum_c\boldsymbol{w}_c^T\boldsymbol{\phi}_c(\boldsymbol{y}_i,\boldsymbol{x}_i)-\log Z(\boldsymbol{w},\boldsymbol{x}_i)\right] \tag{17}$$

以及梯度:

$$\frac{\partial l}{\partial\boldsymbol{w}_c}=\frac{1}{N}\sum_i[\boldsymbol{\phi}_c(\boldsymbol{y}_i,\boldsymbol{x}_i)-E[\boldsymbol{\phi}_c(\boldsymbol{y},\boldsymbol{x}_i)]] \tag{18}$$

注意到 $E[\varphi_c(\boldsymbol{y},\boldsymbol{x}_i)]$依赖于第 i 个训练样本,因此在采用梯度方法训练的每一步,针对每一个训练样本都需要在模型中进行推断.这一特点使得对于大小为 N 的训练集,条件随机场的训练时间比马尔可夫随机场要大一个 $O(N)$的因子.然而,这也为条件随机场带来不可替代的优势:因子以输入为条件(而不是包含输入),意味着我们可以向模型中加入全局特征,而不必担心变量相互耦合导致的推理困难.

4.2 主题模型

主题模型是一类生成式的贝叶斯模型,在文本分析、社交网络分析、计算机视觉等问题中有广泛的应用.如图 6 所示,变分推断的目标是选择合适的变分参数 $\tilde{\pi}_i,\tilde{q}_{il},\tilde{B}$, 使得图 6(b)中 π_i,q_{il},B 的分布尽可能接近图 6(a)中对应变量的后验分布.这里,我们从文本分析的角度介绍主题模型.假设语料库由 K 个主题组成,每篇文档的内容是不同主题的混合,用 K 项分布 π_i 表示.每个主题对应一个该主题的词库,$\boldsymbol{b}_k$ 表示第 k 个主题词库产生不同词汇的概率分布.文档的生成过程由算法 3 来描述.

算法 3. 主题模型的文本生成方法.

输入:超参数 α,γ,K,V,N;

输出:文本语料库.

(1) **for** 第 $k\in[1,\ldots,K]$个主题 **do**

(2) 抽样词汇的概率分布 $\boldsymbol{b}_k|\gamma\sim Dir(\gamma 1_V)$

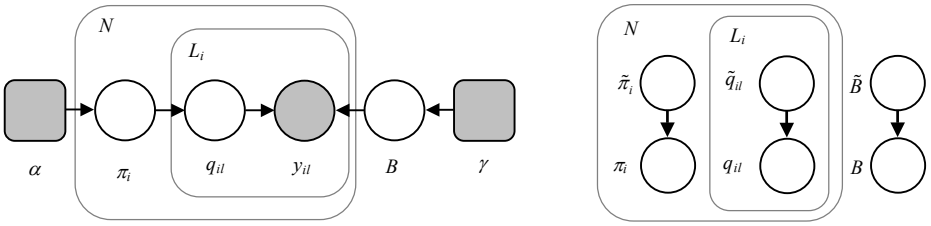
(3) **for** 第 $i\in[1,\ldots,D]$篇文档 **do**

(4) 抽样第 i 篇文档的主题分布 $\pi_i|\alpha\sim Dir(\alpha 1_K)$

(5) **for** 第 $l\in[1,\ldots,N_i]$个单词 **do**

(6) 抽样该单词的主题 $q_{il}|\pi_i\sim Cat(\pi_i)$

(7) 抽样单词 $y_{il}|q_{il}=k,\boldsymbol{B}\sim Cat(\boldsymbol{b}_k)$



(a) 主题模型 (b) 主题模型的变分近似

Fig.6

图 6

4.2.1 主题模型的推理

- (坍塌的)Gibbs 抽样[29]

我们从朴素的 LDA 的 Gibbs 抽样算法开始.首先,我们可以写出所有变量的完全条件分布:

$$p(q_{il}=k\,|\,\cdot)\propto \exp[\log\pi_{ik}+\log b_{k,x_{il}}] \tag{19}$$

$$p(\pi_i\,|\,\cdot)=Dir\left(\left\{\alpha_k+\sum_l I(z_{il}=k)\right\}\right) \tag{20}$$

$$p(b_k\,|\,\cdot)=Dir\left(\left\{\gamma_v+\sum_i\sum_l I(x_{il}=v,z_{il}=k)\right\}\right) \tag{21}$$

然而,由于 π_i 和 b_k 都服从狄利克雷(Dirichlet)分布,我们可以积分消去这些变量,导出坍塌的 Gibbs 抽样(collapsed Gibbs sampling)[29,30]:

$$p(q_{i,l}=k\,|\,\boldsymbol{q}_{-i,l},\boldsymbol{y},\alpha,\gamma)\propto\frac{c^{-}_{v,k}+\gamma}{c^{-}_{k}+V_\gamma}\frac{c^{-}_{i,k}+\alpha}{L_i+K\alpha} \tag{22}$$

其中,记 $v=y_{il}$ 表示第 i 篇文档的第 l 个单词, $c^{-}_{v,k}$ 表示除文档 i 的第 l 个单词外所有被归入主题 k 的语料库单词 v 的数目, $c^{-}_{i,k}$ 表示除文档 i 的第 l 个单词外所有被归入主题 k 的文档 i 的单词的数目, c^{-}_{k} 表示除文档 i 的第 l 个单

词外所有被归入主题 k 的单词的数目,而 L_i 表示第 i 个文档的单词总数.该公式有直观的含义:文档 i 的第 l 个单词被归入主题 k 的概率正比于该单词在主题 k 中产生的概率(第 1 项)以及该主题在文档中使用的概率(第 2 项).

基于公式(22),我们可以实现坍塌的 Gibbs 抽样算法(算法 4).

算法 4. 坍塌的 Gibbs 抽样.

输入:文本语料库.

输出:文本语料库 q_{il}.

(1) 初始化:随机为每个单词指定一个主题 $q_{il}\in\{1,\ldots,K\}$

(2) **for** 第 $i\in[1,\ldots,D]$篇文档 **do**

(3) **for** 第 $l\in[1,\ldots,N_i]$个单词 **do**

(4) 统计 $c^{-}_{v,k}, c^{-}_{i,k}$和c^{-}_{k}

(5) 按照公式(22)抽样该单词的主题 q_{il}

- 批量变分推断[28]

由于后验分布不易直接求解,可以采用变分推断方法,从一类独立性更强的图模型中寻找原模型的最优近似.最简单的变分推断使用平均场近似,即假设待推断变量的联合后验分布可以完全因子化:

$$q(\pi_i, c_i, B) = B(\pi_i \mid \tilde{\pi}) \prod_v Mul(c_{iv\cdot} \mid n_{iv}, \tilde{c}_{iv\cdot}) \prod_k Dir(b_k \mid \tilde{b}_k) \tag{23}$$

在这种假设下,我们最小化泛函 $q(\pi_i,c_i,B)$与目标分布的 KL-散度,将得到如下更新规则:

$$\tilde{\pi}_{ik} = \alpha_k + \sum_v n_{iv} \tilde{c}_{ivk} \tag{24}$$

$$\tilde{c}_{ivk} \propto \exp(E[\log b_{vk}] + E[\log \pi_{ik}]) \tag{25}$$

$$\tilde{b}_{vk} = \gamma_v + \sum_i \tilde{c}_{ivk} \tag{26}$$

由公式(24),我们得到批量变分推断的更新算法(算法 5、算法 6).

算法 5. LDA 的批量变分近似推理.

输入:超参数α_k,γ_v,K,词汇统计 n_{iv}.

输出:π_{ik},c_{ivk},b_{vk}.

(1) 使用多项分布混合模型的 EM 方法估计 $\tilde{b}_{vk}$ 的值

(2) 初始化计数 n_{iv}

(3) **while** 未收敛 **do**

(4) s_{vk}=0

(5) **for** 第 $i\in[1,\ldots,D]$篇文档 **do**

(6) $(\tilde{\pi}_i, \tilde{c}_i) = Estep(n_i, \tilde{B}, \alpha)$

(7) $s_{vk} = s_{vk} + n_{iv}\tilde{c}_{ivk}$

(8) **for** 第 $k\in[1,\ldots,K]$个主题 **do**

(9) $\tilde{b}_{vk} = \gamma_v + s_{vk}$

算法 6. $Estep(n_i,\tilde{B},\alpha)$.

输入:超参数α_k,γ_v,K,词汇统计 n_{iv}.

输出:π_{ik},c_{ivk},b_{vk}.

(1) 初始化 $\tilde{\pi}_{ik} = \alpha_k$

(2) **repeat**

(3) $\tilde{\pi}^{old}_{i\cdot} = \tilde{\pi}_{i\cdot}, \tilde{\pi}_{ik} = \alpha_k$

(4) **foreach** 单词 $v\in\{1,\ldots,V\}$ **do**

(5) **foreach** 主题 $k\in\{1,\ldots,K\}$ **do**

(6) $\tilde{c}_{ivk} = \exp(\psi_k(\tilde{b}_{v\cdot}) + \psi_k(\tilde{\pi}_{i\cdot}^{old}))$

(7) 归一化 $\tilde{c}_{iv\cdot}$

(8) $\tilde{\pi}_{ik} = \tilde{\pi}_{ik} + n_{iv}\tilde{c}_{ivk}$

(9) **until** $\frac{1}{K}\sum_k |\tilde{\pi}_{ik} - \tilde{\pi}_{ik}^{old}| < thresh$

- 在线变分推断[31]

注意到在批量变分推断中,求期望的步骤需要遍历所有 N 个文档,在文档集很大的情形下,这将花费太多的时间.利用随机梯度下降的方法,在每次迭代中我们可以使用 1 个或数个数据的期望充分统计量代替整个数据集上的期望充分统计量,然后在最大化步骤中对 B 的变分参数做部分更新.这样,就得到了 LDA 的在线变分推断算法.

算法 7. LDA 的在线变分近似推理.

输入:超参数 α_k, γ_v, K,词汇统计 n_{iv}, τ_0, κ.

(1) 随机初始化 $\tilde{b}_{vk}$

(2) **for** t=1:∞ **do**

(3) 步长 $\rho_t=(\tau_0+t)^{-\kappa}$

(4) 选择小批文档 $i=i(t)$

(5) $(\tilde{\pi}_i, \tilde{c}_i) = Estep(n_i, \tilde{B}, \alpha)$

(6) $\tilde{b}_{vk}^{new} = \rho_v + Nn_{iv}\tilde{c}_{ivk}$

(7) $\tilde{b}_{vk} = (1-\rho_t)\tilde{b}_{vk} + \rho_t\tilde{b}_{vk}^{new}$

5 近似推理的加速

随着互联网、传感器网络等大规模应用的兴起,图模型的推理问题规模也迅速扩大.在这一类大规模图模型应用中,如何迅速而有效地对图模型执行近似推理,是我们关心的核心问题.针对这一问题,目前的方法主要结合以下几种解决途径:设计好的消息传递调度方法,利用多核硬件实现并行和分布式推理,针对查询变量有选择地执行推理.

5.1 消息传递调度

设计消息更新调度是一种有效且广泛使用的提高信念传播计算效率的方法.在标准的环路信念传播中,每次迭代将重新计算和更新所有的消息.然而多数情况下,只有少数消息在相继的两次更新中有较大变化,而大多数的消息更新几乎不改变计算结果,因此浪费了计算资源.

- 残余信念传播

Elidan[32]提出了一种称为残余信念传播(residual belief propagation)的消息更新调度算法.该算法的更新规则是:每次只更新在标准环路信念传播算法的消息更新中改变最大的那个消息.

具体来说,残余信念传播算法为因子图的每个边记录两个消息:一个是当前消息 $\nu_{\alpha-j}^{(t)}$; 一个是新消息,即给定所有其他当前消息时信念传播的更新结果:

$$\hat{\nu}_{\alpha-i}^{(t)}(x_i) = \log \sum_{X_{\alpha\setminus i}} \left(\psi_\alpha(X_{\alpha\setminus i}, x_i) \exp \sum_{j\in\Gamma_\alpha\setminus i} (\tilde{Y}^{(t)}(x_j) - \nu_{\alpha-j}^{(t)}(x_j)) \right) \tag{27}$$

新消息与当前消息之间的差异称为残余 $r_{\alpha-i}$:

$$r_{\alpha-i} \equiv \| \nu_{\alpha-i}^{(t)} - \nu_{\alpha-i}^{(t)} \| \tag{28}$$

在第 t+1 轮迭代,算法只更新有最大残余的消息.换言之,

$$\nu_{\alpha-i}^{(t+1)}=\begin{cases}\hat{\nu}_{\alpha-i}^{(t)}, & (\alpha-i)=\arg\max\limits_{(\alpha-i)\in E} r_{\alpha-i}\\ \nu_{\alpha-i}^{(t)}, & \text{其他情况}\end{cases} \tag{29}$$

在更新 $\nu_{\alpha-i}$ 后,只有与 i 直接相关的因子 $\beta\in\Gamma_i$ 发送的新消息 $\hat{\nu}_{\beta-j}^{(t+1)}$ 需要重新计算.算法伪代码如下:

算法 8. 残余信念传播.

输入:G=($\{X,F\}$,T).

(1) 定义优先队列 M

(2) **foreach** $(\alpha-i)\in T$ **do**

(3) 初始化消息 $\nu_{\alpha-i}$

(4) **foreach** $(\alpha-i)\in T$ **do**

(5) 根据公式(27)、公式(28)计算 $\hat{\nu}_{\alpha-i}$,$r_{\alpha-i}$

(6) 令边$(\alpha-i)$在 M 中的优先权为 $r_{\alpha-i}$,将$(\alpha-i)$加入 M

(7) **while** 消息未收敛 **do**

(8) 记 M 的堆顶$(\alpha-i)$

(9) 令 $\nu_{\alpha-i}$ 为 $\hat{\nu}_{\alpha-i}$

(10) 令$(\alpha-i)$在 M 中的优先权为 0

(11) **foreach** $\beta\in\Gamma_i\backslash\alpha$ **do**

(12) **foreach** $j\in\Gamma_\beta\backslash i$ **do**

(13) 根据公式(27)、公式(28)重新计算 $\hat{\nu}_{\beta-j}$,$r_{\beta-j}$

(14) 令边$(\beta-i)$在 M 中的优先权为 $r_{\beta-i}$

(15) 由公式 $\tilde{P}(x_i)\propto\exp\left(\sum_{\alpha\in\Gamma_i}\nu_{\alpha-i}(x_i)\right)$ 计算变量 $x\in X$ 的边缘分步 $\tilde{P}(x)$

(16) 返回 $\tilde{P}(x)$

直观上,残余信念传播总是更新最大残余的消息,因此,每单位计算量所获得的消息改变更多,从而在信念空间中能够更快地趋向不动点.

5.2 并行和分布式推理

信念传播算法由于消息传递的局部性,自然适合于在多核机器上并行实现.然而,同步的并行信念传播算法并不高效.Gonzalez 等人[33]提出的残余飞溅信念传播(residual splash belief propagation)算法是一种异步的信念传播方法.它通过设计消息更新顺序、合理分布计算任务,能够减少无用的计算,在多核机器上获得明显的性能提升.算法名称中的“飞溅”一词是对算法运行过程的直观描述:在算法运行过程中,消息更新总是在以某个顶点(称为根)为中心的一定直径的子图上进行,且消息首先由叶节点向根传播,再由根节点传向叶子,如同水面上溅起水花的过程.

该算法的设计受到了以下两点的启发:首先,在同一路径上变量之间的影响随着距离增大而减小,因而,为了获得较为精确的近似,常常只需要考虑顶点周围的子图;其次,对于一个树状图模型,尽管容易设计同步的并行信念传播算法,但是对每个顶点而言,使得算法收敛的消息却要在其接受到其余邻居顶点的消息之后才能形成,因此在此之前的消息传递实际上都浪费了.于是,算法设计使得每个机器独自处理一个由宽度优先搜索生成的树状子图,对该子图采用串行的和-积算法执行推理(称为飞溅操作).在飞溅操作结束后,树状子图的内部将被标定,没有残余的消息;消息残余只存在于子图的叶节点.之后再分配机器对残余的消息执行飞溅操作,直到所有顶点有关的最大残余消息都小于某一阈值.若信念传播算法能够收敛于不动点,则某时刻消息与不动点的距离和残余消息同阶,因此,算法终止时能够获得较好的近似结果.

算法的主要伪代码参见算法 9.

算法 9. 残余飞溅信念传播算法.

输入:飞溅操作的树深 h,残余消息阈值β.

(1) 初始化顶点优先队列 Q

(2) 令所有消息残余为∞

(3) **forall** 处理器 **in parallel**

(4) **while** Q 中最大残余消息大于β **do**

(5) 记堆顶的顶点为 v,弹出堆顶

(6) 飞溅(v,h)

(7) **foreach** 飞溅操作影响的顶点 u **do**

(8) 在 Q 中更新 u 的消息残余

(9) 将(v,v 的消息残余)推入堆 Q

算法 10. 飞溅(v,h).

输入:飞溅操作的树深 h,顶点 v.

(1) 构造以 v 为起点,深度为 h 的广度优先搜索序(bfs-order)

(2) **foreach** i∈逆 bfs-order **do**

(3) 向 v 的方向传递消息

(4) **foreach** i∈bfs-order **do**

(5) 向背离 v 的方向传递消息

5.3 针对查询的推理

在实际的应用中,我们常常并不关心所有变量的取值,而只求能够获得感兴趣的变量,即查询变量的条件分布.因此,如果信念传播中的消息更新与我们的查询变量没有关系或者只存在微弱联系,那么所做的计算在我们看来就是浪费了.针对查询的推理,就是研究如何利用图模型的结构和变量依赖关系对查询变量进行有针对性的近似推理,以加快推理速度,适应大规模图模型推理的需求.

- 针对查询的信念传播

针对查询的信念传播目标就在于估计信念传播消息更新对于查询变量的分布的影响,只更新对查询变量分布影响较大的消息,从而改善查询变量分布的计算效率[34].

具体来说,我们记因子图模型为 $G=(\{X,F\},T)$,查询变量为 $q\equiv i_k$,一条影响查询变量的消息传播路径为 $\pi=(\beta_1\to i_1\to\ldots\to\beta_k\to q)$.若固定所有不在路径$\pi$中的消息$\nu_{-\pi}$,应用环路信念传播的更新规则,则 ν_{β_k-q} 可以表示为 $\nu_{\beta_1-i_1}$ 的函数 $F_\pi(\nu_{\beta_1-i_1},\nu_{-\pi})$. 因此,我们可以写出 ν_{β_k-q} 关于 $\nu_{\beta_1-i_1}$ 的变化的敏感性:

$$\left.\frac{\partial \nu_{\beta_1-i_1}}{\partial \nu_{\beta_k-q}}\right|_{\pi}\equiv\frac{\partial F_\pi}{\partial \nu_{\beta_k-q}}=\prod_{d=1}^{k-1}\frac{\partial \nu_{\beta_{d+1}-i_{d+1}}}{\partial \nu_{\beta_d-i_d}} \tag{30}$$

利用不等式放缩,我们可以得到如下关于 $\nu_{\beta_1-i_1}$ 的变化对 ν_{β_k-q} 的影响的界:

$$\|\Delta\nu_{\beta_k-q}|_{\pi}\|\leqslant\|\Delta\nu_{\beta_1-i_1}\|\cdot\prod_{d=1}^{k-1}\sup_{\nu_{-\pi}}\left\|\frac{\partial \nu_{\beta_{d+1}-i_{d+1}}}{\partial \nu_{\beta_d-i_d}}\right\| \tag{31}$$

为此,定义一个有向路径$\pi=(\beta_1\to i_1\to\ldots\to\beta_k\to q)$的敏感强度为

$$sensitivity(\pi)=\prod_{d=1}^{k-1}\sup_{\nu_{-\pi}}\left\|\frac{\partial \nu_{\beta_{d+1}-i_{d+1}}}{\partial \nu_{\beta_d-i_d}}\right\| \tag{32}$$

其中,最大值的计算复杂度取决于范数的选择.利用 Mooij 和 Kappen[35]的结果,若取范数为对数动态范围,则公式(32)中的最大值可以解析地求出.为衡量消息α–i 对于查询变量 x_q 的影响,理论上需要考虑所有从α–i 指向 q 的有向路径.然而,由于环路的存在,这样的路径有无穷多条,计算它们的影响之和是困难的.为此,我们可以采用对查询变量影响最大的路径来近似代替所有路径的影响.定义α–i 的最大敏感重要性值为

$$\max-sensitivity(\alpha-i,q)\equiv\max_{\pi\in\prod(\alpha-i,q)}sensitivity(\pi) \tag{33}$$

其中,$\prod(\alpha-i,q)$是由所有从α–i指向q的有向路径组成的集合.

注意到$\sup_v\left\|\frac{\partial v_{\alpha-i}}{\partial v_{\beta-j}}\right\|\in[0,1)$,因此,计算最大敏感重要性值的问题与计算最短路径问题一样具有贪心最优性质.因此,仿照求最短路径的 Dijkstra 算法,我们可以得到计算边的重要性值的算法.

算法 11. 计算边重要性.

输入:因子图 $G=(\{X,F\},T)$,查询 $Q\in X$.

(1) 定义优先队列 $L=\varnothing$,边的优先权$\rho_{\alpha-i}$

(2) **foreach** $(\alpha$–$i)\in T$ **do**

(3) 令$(\alpha$–$i)$的优先权为 $\rho_{\alpha-i}=\begin{cases}1, & \text{如果} i\in Q\\ 0, & \text{其他}\end{cases}$

(4) 将$(\alpha$–$i)$加入优先队列 L

(5) **while** $L\neq\varnothing$ **do**

(6) 记 L 的堆顶为$(\alpha$–$i)$

(7) 令 $w_{\alpha-i}$ 为$\rho_{\alpha-i}$,删除 L 的堆顶元素$(\alpha$–$i)$

(8) **foreach** $\beta\in\Gamma_i\backslash\alpha$ **do**

(9) **foreach** $j\in\Gamma_\beta\backslash i$ **do**

(10) **if** $(\beta$–$j)\in L$ **then**

(11) 令$\rho_{\beta-j}$为 $\max\left(\rho_{\beta-j},\rho_{\alpha-i}\cdot\sup_v\left\|\frac{\partial v_{\alpha-i}}{\partial v_{\beta-j}}\right\|\right)$

(12) 根据新的优先权维护优先队列 L

(13) 返回 $W=\{w_{\alpha-i}|(\alpha$–$i)\in T\}$——所有边的重要性值

有了算法 11,对残余信念传播稍做修改,即可得到针对查询的信念传播算法(算法 12).

算法 12. 针对查询的信念传播.

输入:因子图 $G=(\{X,F\},T)$,查询 $Q\in X$;

(1) 定义优先队列 M

(2) 令 W 为算法 11 的返回值

(3) **foreach** $(\alpha$–$i)\in T$ **do**

(4) 初始化消息 $v_{\alpha-i}$

(5) **foreach** $(\alpha$–$i)\in T$ **do**

(6) 根据公式(27)、公式(28)计算 $\hat{v}_{\alpha-i},r_{\alpha-i}$

(7) 令边$(\alpha$–$i)$在 M 中的优先权为 $w_{\alpha-i}\times r_{\alpha-i}$,将$(\alpha$–$i)$加入 M

(8) **while** 消息未收敛 **do**

(9) 记 M 的堆顶$(\alpha$–$i)$

(10) 令 $v_{\alpha-i}$ 为 $\hat{v}_{\alpha-i}$

(11) 令$(\alpha$–$i)$在 M 中的优先权为 0

(12) **foreach** $\beta\in\Gamma_i\backslash\alpha$ **do**

(13) **foreach** $j\in\Gamma_\beta\backslash i$ **do**

(14) 根据公式(27)、公式(28)重新计算 $\hat{v}_{\beta-j},r_{\beta-j}$

(15) 令边$(\beta$–$i)$在 M 中的优先权为 $r_{\beta-i}$

(16) 计算查询变量 $x\in Q$ 的边缘分步 $\tilde{P}(x)$

(17) 返回 $\tilde{P}(x)$

- 查询敏感的 MCMC

考虑由以下两个步骤诱导的提议分布:

(1) 从当前状态开始,依定义于变量下标(1,2,...,n)上的分布 p 抽样,选择一个变量 $x_i \in \boldsymbol{x}$;

(2) 从某个定义在 x_i 值域上的分布 $q(X_i)$抽样 x_i 的新值,固定其余变量的值不变,返回新状态 s'.

简而言之,这一提议分布每次由当前状态 s 更新一个变量的值得到新状态 s'.在传统的 MCMC 推理中,我们感兴趣的是所有变量的边缘分布,变量更新抽样通常采用固定的顺序,或者由 $p(i)=\frac{1}{n}$ 诱导的顺序.针对这类提议分布,Wick 和 McCallum[36]提出了一种查询敏感的 MCMC 算法.该算法的关键在于,注意到图模型的变量并非对查询变量有相同的影响:查询变量的取值与某些变量的值有强依赖关系;而与另一些变量仅有微弱的依赖关系,甚至不相关.因此,一种更好的提议分布应当更频繁地抽样与查询变量依赖性强的变量.为此,可以定义变量间的影响的概念.设 x 和 y 是两个随机变量,其联合分布和边缘分布分别为 $\pi(x,y),\pi(x),\pi(y)$.令 $f(\pi_1(\cdot),\pi_2(\cdot))\mapsto r$, $r\in R_+$为概率分布的散度,因而是实值、非负的.定义 x 和 y 的影响 $\iota(x,y)$为

$$\iota(x,y)\equiv f(\pi(x,y),\pi(x)\pi(y)) \tag{34}$$

取 f 为总变差范数,得到 $\iota_{tv}(x,y)\equiv P\|\pi(x,y)-\pi(x)\pi(y)P_{tv}\|$.

然而,对于图模型中的任意两个变量,精确计算 ι_{tv} 需要在图模型上推理.为此,文献[36]进一步提出了影响路径得分的概念.对于一条从查询变量 $x_0=x_q$ 到任意变量 $x_s=x_s$ 的有效路径 $\rho=(x_0,x_1,\ldots,x_r)$,设 $\phi(x_i,x_j)$是由包含 x_i,x_j 的因子所定义的关于 x_i,x_j 的联合分布,作为真实的联合分布的近似,而 $\phi(x_i),\phi(x_j)$为其边缘分布,则有效路径 ρ的影响路径得分为

$$\tau_\rho(x_q,x_s)\equiv\prod_{i=1}^{r-1} f(\phi_i(x_i,x_{i+1}),\phi_i(x_i)\phi_i(x_{i+1})) \tag{35}$$

令 $p(i)\propto\tau_\rho(x_q,x_i)$,即得到查询敏感的提议分布.

对应的 Metropolis-Hastings 算法称为查询敏感的 MCMC(QAM).

6 未来工作研究与展望

概率图模型作为描述和处理具有条件独立结构的概率模型的一般框架,其研究在近年来已经取得了长足的进展,获得了来自领域内外的持续关注,催生了丰富的应用.本文介绍了概率图模型的表示、推理和学习问题及主要进展,比较了不同的推理和学习算法的适用范围和执行的复杂度,简要介绍了近年来较为重要的两类概率图模型以及推理问题的最新研究进展.

从本文的介绍中可以看出:尽管在一般意义下概率图模型的推理是困难的,但是仍然有相当一部分模型和推理算法能够在我们所关心的应用中取得较好的效果.然而,概率图模型的研究中还有许多问题值得在今后进一步探索:

(1) 概率图模型的表示

通常情况下,我们仍然需要针对具体问题设计特定的概率图模型.在设计中既要对所处理的问题中变量的依赖关系做出合理抽象,又要权衡模型的推理代价或近似推理的难度,以在实际应用中获得推理精度和推理效率的折中.许多经典的图模型即反映出这种折中的理念.然而,一方面由于新的应用问题不断出现,另一方面由于从精确推理到近似推理、批量学习到在线学习有一系列可选的折中方案,可以期待新的概率图模型设计将会不断涌现.例如,为解决 LDA 模型需要事先指定主题数量的问题,诞生出一系列基于贝叶斯非参数统计的模型.

(2) 概率图模型的学习和推理

大规模概率图模型的有效推理方法.目前,概率图模型应用规模逐渐扩大,例如 Google Scholar 等学术搜索应用需要识别区分同名作者,可考虑建立关于学术研究作者及其发表论文的概率模型,而这样的模型中随机变量的规模可达到百万量级.在大规模问题中,利用单机 CPU 和传统推理方法进行的推理将因效率过低而失去实

用性.因此,进一步研究加速推理的方法,包括利用 GPU 加速、分布式推理、针对查询的推理、效率和准确性的折中、新的变分推理和 MCMC 方法等等都是有意义的研究方向.

(3) 描述和处理概率图模型的计算机语言

从概率图模型的图形表示到其数据结构表示之间还有一定的距离,而模型的高效近似推理算法常常也需要研究者针对模型专门设计.据统计,概率图模型的应用研究中,模型的设计只占到整个研究周期的一小部分时间,而大部分时间(通常需要 1 个月~数月)都花费在推导推理和学习算法以及模型和算法的代码实现上.这表明,在概率图模型研究效率上还有很大的改善空间.如果存在一种工具,使得关心模型应用的研究者只需提供关于模型的形式化描述而由计算机完成模型和推理算法的实现,将给图模型的应用带来很大的方便.一些只需要使用者提供模型的描述而由程序自动实现推理和学习算法的工具(例如 WinBUGS[37],Infer.NET[10],Church[38])已经在部分应用领域获得了成功.在这一称为概率式编程(probabilistic programming)的研究主题下,研究者们正在尝试扩展这些工具的通用性,并且提高其推理效率,使其能够在速度和准确性上媲美人工设计的算法.

References:

[1] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[2] Gibbs JW. Elementary principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics. Yale University Press, 1902.

[3] Wright S. Systems of mating. I. the biometric relations between parent and offspring. Genetics, 1921,6(2):111−123.

[4] Croft DJ, Machol RE. Mathematical methods in medical diagnosis. Annals of Biomedical Engineering, 1974,(2):69−89. [doi: 10.1007/BF02368087]

[5] Gorry GA, Barnett G. Experience with a model of sequential diagnosis. Computers and Biomedical Research, 1968,1(5):490−507. [doi: 10.1016/0010-4809(68)90016-5]

[6] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference. Morgan Kaufmann Publishers, 1988.

[7] Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society. Series B (Methodological), 1988,50(2):157−224.

[8] Heckerman DE, Horvitz EJ, Nathwani BN. Toward Normative Expert Systems: The Pathfinder Project. Stanford: Knowledge Systems Laboratory, Stanford University, 1990.

[9] Kschischang FR, Frey BJ, Loeliger HA. Factor graphs and the sum-product algorithm. IEEE Trans. on Information Theory, 2001, 47(2):498−519. [doi: 10.1109/18.910572]

[10] Winn J, Bishop CM. Variational message passing. Journal of Machine Learning Research, 2006,6(1):661−694.

[11] Dagum P, Luby M. Approximating probabilistic inference in bayesian belief networks is NP-hard. Artificial Intelligence, 1993, 60(1):141−153. [doi: 10.1016/0004-3702(93)90036-B]

[12] Zhang NL, Poole D. A simple approach to bayesian network computations. In: Proc. of the 10th Canadian Conference on Artificial Intelligence. 1994. 171−178.

[13] Shafer GR, Shenoy PP. Probability propagation. Annals of Mathematics and Artificial Intelligence, 1990,2(1-4):327−351. [doi: 10.1007/BF01531015]

[14] Andersen SK, Olesen KG, Jensen FV, Jensen F. Hugin—A shell for building bayesian belief universes for expert systems. In: Proc. of the 11th Int'l Joint Conf. on Artificial Intelligence, Vol.2. 1989. 1080−1085.

[15] McEliece RJ, MacKay DJC, Cheng JF. Turbo decoding as an instance of pearl's "belief propagation" algorithm. IEEE Journal on Selected Areas in Communications, 1998,16(2):140−152. [doi: 10.1109/49.661103]

[16] Yedidia JS, Freeman WT Weiss Y. Generalized belief propagation. In: Proc. of the NIPS. 2000. 689−695.

[17] Robert CP, Casella G. Monte Carlo Statistical Methods. New York: Springer-Verlag, 2004.

[18] Fung R, Chang KC. Weighting and integrating evidence for stochastic simulation in Bayesian networks. In: Proc. of the 5th Conf. on Uncertainty in Artificial Intelligence (UAI 1989), Vol.5. 1989. 209−219.

[19] Shachter RD, Peot MA. Simulation approaches to general probabilistic inference on belief networks. In: Proc. of the 5th Conf. on Uncertainty in Artificial Intelligence (UAI 1989), Vol.5. 1989. 221−231.

[20] Dagum P, LubyM. An optimal approximation algorithm for bayesian inference. Artificial Intelligence, 1997,93(1):1−27.

[21] Doucet A, De Freitas N, Gordon N. An Introduction to Sequential Monte Carlo Methods. Springer-Verlag, 2001.

[22] Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984,(6):721−741.

[23] Neal RM. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Reort, CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[24] Swendsen RH, Wang JS. Nonuniversal critical dynamics in Monte Carlo simulations. Physical review letters, 1987,58(2):86−88. [doi: 10.1103/PhysRevLett.58.86]

[25] Barbu A, Zhu SC. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(8):1239−1253. [doi: 10.1109/TPAMI.2005.161]

[26] Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Physics Letters B, 1987,195(2):216−222. [doi: 10.1016/0370-2693(87)91197-X]

[27] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the ICML. 2001. 282−289.

[28] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[29] Griffiths TL, Steyvers M. Finding scientific topics. Proc. of the National Academy of Sciences of the United States of America, 2004,101(Suppl.1):5228−5235. [doi: 10.1073/pnas.0307752101]

[30] Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press, 2012.

[31] Hoffman M, Blei DM, Bach F. Online learning for latent dirichlet allocation. In: Proc. of the NIPS. 2010.

[32] Elidan G, McGraw I, Koller D. Residual belief propagation: Informed scheduling for asynchronous message passing. In: Proc. of the UAI. 2006.

[33] Gonzalez J, Low Y, Guestrin C. Residual splash for optimally parallelizing belief propagation. Journal of Machine Learning Research—Proc. Track, 2009,5:177−184.

[34] Chechetka A, Guestrin C. Focused belief propagation for query-specific inference. Journal of Machine Learning Research—Proc. Track, 2010,9:89−96.

[35] Mooij JM, Kappen HJ. Sufficient conditions for convergence of the sum-product algorithm. IEEE Trans. on Information Theory, 2007,53(12):4422−4437. [doi: 10.1109/TIT.2007.909166]

[36] Wick ML, McCallum A. Query-Aware MCMC. In: Proc. of the NIPS. 2011. 2564−2572.

[37] Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing, 2000,10(4):325−337. [doi: 10.1023/A:1008929526011]

[38] Goodman ND, Mansinghka VK, Roy DM, Bonawitz K, Tenenbaum JB. Church: A language for generative models. In: Proc. of the UAI. 2008. 220−229.

张宏毅(1990－),男,湖北襄阳人,博士生,主要研究领域为机器学习,概率图模型.
E-mail: hongyi.zhang.pku@gmail.com

陈瑜希(1990－),女,硕士生,主要研究领域为机器学习,计算机视觉.
E-mail: chenyuxi90@pku.edu.cn

王立威(1975－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习.
E-mail: wanglw@cis.pku.edu.cn