

概率逻辑模型与学习研究进展^{*})

徐光美 杨炳儒 张 伟 宁淑荣

(北京科技大学信息工程学院知识工程研究所 北京 100083)

摘 要 近年来,概率逻辑学习研究取得了很大进展,已经提出各种不同的形式化方法和学习方法,包括概率关系模型(PRM)、贝叶斯逻辑程序(BLPs)、逻辑贝叶斯网络(LBNs)和随机逻辑程序(SLPs)等。文章重点介绍了贝叶斯网络与一阶逻辑的结合,并以 PRMs、BLPs 和 LBNs 为例,描述了基于贝叶斯网络的概率逻辑模型(PLMs)的知识表示方法,给出了此类 PLMs 一般使用的参数估计方法和结构学习方法,并给出了建议的研究方向。

关键词 概率逻辑模型,概率关系模型,贝叶斯逻辑程序,逻辑贝叶斯网络,概率逻辑学习

Advances in Probabilistic Logic Model and Learning

XU Guang-Mei YANG Bing-Ru ZHANG Wei NING Shu-Rong

(School of Information Engineering, Beijing University of Science and Technology, Beijing 100083)

Abstract Probabilistic logic learning (PLL) research has made significant progress over the last years. A rich variety of different formalisms and learning techniques have been developed, including probabilistic relational models, bayesian logic programs, and logic bayesian networks and stochastic logic programs etc. This paper, focusing on the combination of bayesian networks and first-order logic, provides an introductory survey on probabilistic logic models based on bayesian networks through the investigation of knowledge representation, parameter estimation and structure learning algorithms. Although the PLL community has successfully demonstrated the feasibility of a number of probabilistic models for relational data, there is much work on efficiency and scalability to be done in order to begin generalizing the range and applicability of the various models.

Keywords Probabilistic logical models, Probabilistic relational models, Bayesian logic programs, Logic bayesian networks, Probabilistic logical learning

1 引言

概率逻辑学习(Probabilistic Logical Learning, PLL)也称统计关系学习(Statistical Relational Learning, SRL)或概率归纳逻辑程序设计(Probabilistic Inductive Logic Programming, PILP),它是近年来兴起的一个多学科融合的研究领域,旨在将概率推理、一阶逻辑表示和机器学习结合起来^[1,5,6]。PLL 中概率项是指以概率理论为基础使用概率表示和推理机制,如 BNs(Bayesian Networks)、HMMs(Hidden Markov Models)和概率文法(probabilistic grammars)以及相应的统计学习和推理技术等。逻辑或关系项则指一阶逻辑或关系表示,这种表示方法能够表示命题逻辑或特征向量无法表示的复杂的对象以及对象间的关系。

目前已经存在各种各样的概率逻辑模型(Probabilistic Logical Models, PLMs)。根据形成的方法不同,PLMs 大致可以分为两类:一类是根据基于知识的模型构建方法(Knowledge Based Model Construction, KBMC)将原有的概率模型(BNs, HMMs 等)和逻辑表示结合所形成的模型。Friedman, Getoor, Koller 和 Pfeffer 提出的 PRMs(Probabilistic Relational Models)^[1,2]是在关系知识中学习图模型的第

一次尝试。近来, Heckerman, Meek 和 Koller 扩展了 PRMs, 并将它们与其它的图模型进行了比较。其它的 PLMs 还有 BLPs(Bayesian Logic Programs)^[1,4], LBNs(Logic Bayesian Networks)^[1,3], RBNs(Relational Bayesian Networks), PLPs(Probabilistic Logic Programs)等。PRMs, BLPs 的学习算法也已经形成。第二类是将 ILP(Inductive Logic Programming)与概率技术整合,并尽可能与标准的 ILP 保持一致,已有的模型有 David Poole 的 probabilistic Horn abduction, Sato 的 PRISMs, Eisele, Muggleton 和 Cussen 的 SLPs(Stochastic Logic Programs)以及 Independent Choice Logic 和 Logic Programs with Annotated Disjunctions 等^[1,5,6]。PRISMs 和 SLPs 的学习算法也已经形成。本文重点介绍贝叶斯网络与一阶逻辑的结合。

2 知识表示

贝叶斯网络中的每一个节点都可以用一个命题子句来表示,因此贝叶斯网络本质上可以表示成命题子句集合,这使得可以通过提升相应的命题子句到一阶逻辑子句来扩展贝叶斯网络到一阶逻辑环境下。这些子句将指定所谓的内涵部分 I(指定一般规则)和外延部分 E(指定问题域中的特殊对象和

^{*})基金项目:《国家科技成果重点推广计划》项目(2003EC000001)。徐光美 博士生,主要研究方向为数据挖掘和概率逻辑学习;杨炳儒 教授,博士生导师,主要研究领域为知识发现与推理机制,柔性建模与系统集成;张 伟 博士生,主要研究方向为多关系数据挖掘;宁淑荣 博士生,主要研究方向为人工智能。

形式),这里内涵部分和外延部分的区分跟数据库理论和计算逻辑理论内的区分类似。给定了外延部分 E 和内涵部分 I,扩展 BNs 的关键思想就是:将 Herbrand 解释中的原子视为随机变量,这些随机变量将构成贝叶斯网络的节点;实例化子句 $H: -B_1, \dots, B_n$ 看作是概率依赖的定义,每一个子句都要赋予一个条件概率分布。还要注意的得到的贝叶斯网络必须是无环的,而且还要考虑如何限定随机变量的值域,最后还要解决多个基础子句都具有相同头部的问题,可用的方法有所谓的结合规则(combining rules)。在 PRMs 中则使用聚合函数解决这个问题,聚合函数在数据库领域是常用的术语,像最小、平均和求和等都是聚合函数。聚合函数以一种非常自然的方式来归纳信息,它降低了条件概率分布的规模,但是不可避免地会导致信息丢失问题。

以上是对贝叶斯网络进行逻辑拓展的原理,基于这些原理已经形成许多 PLMs。Peter Haddawy 第一次直接扩展贝叶斯网络到关系表示,在他的方法里,将基础原子视为随机变量,而且他要求以随机变量作为头部的子句只能有一个实例化,这就不需要联合规则。

2.1 PRMs

PRMs 是一种非常流行的形式化表示方法,它用实体关系模型作为表示框架。PRMs 的基本思想是:有关实体类型的信息是存在于关系中的。因此,每一个基础原子能存储多个属性的信息,而且也是在属性水平上建立依赖关系。PRMs 允许两类依赖关系存在:相同实体内属性间的依赖关系以及通过槽链(Slot Chain)形成的实体间的属性间的依赖关系。其中槽链是二元关系,它将两个不同实体间的属性相关联。具体而言,PRMs 由三个组件构成^[2]:关系模式、依赖结构和关系骨架。关系模式描述类集合及类的属性集合;依赖结构定义属性条件依赖的双亲集合,与依赖结构相关的是定量组件聚合函数(处理多实例化问题)集合和 CPDs(Conditional Probabilistic Distributions)集合;关系骨架指定所有类的所有对象以及所有对象的关键属性(主键和外键)的值,而对象的其它描述属性值则是未知的。PRMs 从关系骨架中归纳得到贝叶斯网络,归纳得到的贝叶斯网络负责为未知的描述属性指定概率分布,在 PRMs 中学习依赖结构和 CPDs 的算法已经形成。

PRMs 的优点是:表示方便,学习和测试无环性较为容易,但是它不能表示函数(函数可以用于表示结构的不确定性),也不能表示概念的否定(可以处理意外情况),这都限制了 PRMs 的表达能力。PRMs 现在已经成功地用于解决很多问题,包括关系聚类、超文本分类等,并且在生物信息学领域已经取得了较好的成果。

2.2 LBNs

Daan Fierens 和 Hendrik Blockeel 设计了 LBNs,它能被看作是 PRMs 的基于逻辑程序语言的一种变体。LBNs 从给定问题域的一般逻辑程序描述归纳得到的贝叶斯网络。设计 LBNs 的原则是使得知识表示尽可能简单,LBNs 不仅清楚地区分了逻辑知识和概率知识,而且它使用不同的语言组件分别确定 BNs 的不同部分(图的节点,有向边,以及条件概率分布)。具体而言 LBNs 有两类谓词:逻辑谓词和概率谓词。逻辑谓词用于描述问题域的逻辑背景知识(背景知识被假定是确定性信息)。概率谓词用来表示随机变量,而且每一个概率谓词都有与之相关的范围。在 LBNs 中,称概率谓词为概率原子(它与谓词有相同的范围),相似地也称逻辑谓词为逻辑

原子或逻辑文字。LBNs 有四个组件:定义随机变量的子句集合;定义条件依赖于子句的子句集合;表示条件依赖于子句间的依赖程度的逻辑条件概率分布集合(logicalCPDs);用来指定确定性背景知识的逻辑谓词^[3]。

LBNs 和 PRMs 组件间存在一一对应关系:LBNs 的逻辑和概率谓词相应于 PRMs 中关键属性和描述属性;LBNs 通过随机变量声明指定贝叶斯网络的节点,而 PRMs 则用关系模式来指定节点;LBNs 用条件依赖于子句指定贝叶斯网络中的有向边,而 PRMs 则用依赖结构指定;LBNs 用逻辑 CPDs 确定贝叶斯网络的 CPDs,而 PRMs 则用聚合函数将多个 CPDs 合并。这种对应关系的存在使得 PRMs 和 LBNs 间的转换相对较容易。但是 LBNs 可以由任意的逻辑谓词指定的知识确定随机变量和依赖关系,而 PRMs 仅仅可以由关于类成员和其关系的知识(包含在关系骨架中的知识)来确定,因此 LBNs 比 PRMs 更灵活、更具表达力。

2.3 BLPs

基于 Ngo 和 Haddawy 的工作,Kersting 和 De Raedt 引入了 BLPs 框架,它可以看作是 PLPs 的简化,而且确定子句逻辑(即纯 Prolog)和贝叶斯网络都可以看作是 BLPs 的实例。BLPs 也将原子看作是随机变量,更精确地,最小 Herbrand 模型的原子构成了贝叶斯网络的随机变量。

在 BLPs 的最初定义中,只有贝叶斯子句没有标准的逻辑谓词,逻辑谓词是具有真假二值的贝叶斯子句。在 BLPs 改进版本中,则清楚地区分了逻辑谓词和贝叶斯谓词(类似于 LBNs)。最新的定义中,只有基础贝叶斯原子是贝叶斯网络中的随机变量,基础逻辑原子不再是网络中的节点,而且每个子句都有相应的条件概率密度,每个谓词都有一个关联的结合规则^[4]。当存在具有相同头部的多个子句时,使用结合规则来定量这种共同的影响。结合规则将多个条件概率分布映射到同一个(联合的)条件概率分布,噪声或(noisy-or)、噪声与(noisy-and)等都可以作为这种结合规则。这里的结合规则与 LBNs 中的 logical CPDs 扮演相同的角色。这就意味着一个逻辑 CPD 不仅可以为单个的条件依赖于子句定量依赖关系,而且可以为多个具有相同头部的条件依赖于子句定量依赖关系。在独立同分布假定下,由 BLPs 形成的 BNs 可以由任何的贝叶斯网络推理机来计算覆盖关系,而且 BLPs 可以处理连续的随机变量,并允许使用函数。

BLPs 的最新定义和 LBNs 很接近,主要的不同是 LBNs 用子句集合来指定贝叶斯网络中的随机变量并用另一个子句集合来指定有向边,而 BLPs 则是使用同一个子句集合来指定。这点使得 LBNs 比 BLPs 更容易阅读(特别是对于已经熟悉了 PRMs 的人而言)。

3 学习

BLPs、PRMs、LBNs 等都是扩展(upgrading)贝叶斯网络形成的 PLMs,“upgrading”本身就暗示了新的 PLMs 的学习方法与原来的贝叶斯网络学习方法是密切相关的。在 PLL 中,一般也要像传统的命题表示的概率学习一样要区分两个任务:第一个任务就是要考虑概率部分,即数字从哪里来的问题,就是所谓的参数估计问题,这个任务中假定结构(一般是逻辑程序)是已知的并且是确定的,而参数是需要学习的。第二个任务即所谓的结构学习或者是模型选择问题,假定模型或者参数是确定的,但是需要学习才能得到。

3.1 参数估计

参数估计的任务就是估计与某确定的概率逻辑程序 L 相应的 λ^* 的值,以使得相应的 L 能够最好地解释例子 E 。模型拟和数据的好坏一般用数据的似然函数 $P(E|L, \lambda)$ 来评价,实际中一般使用最大似然估计 (MLE) 方法进行参数估计, MLE 方法旨在发现 $\lambda^* = \operatorname{argmax}_{\lambda} P(E|L, \lambda)$, 当然为了计算的简单,经常求对数似然 $\log P(E|L, \lambda)$, 而且一般常假定抽样是独立的, 即有 $\log P(E|L, \lambda) = \sum_{e \in E} \log P(e|L, \lambda)$ 。

当数据完全已知时, MLE 就简化为频率计数问题。当存在缺失值时, MLE 就成为数字优化问题, 有很多数字优化方法, 概率逻辑学习普遍采用的优化算法是期望最大化算法 (Expectation Maximization algorithm, EM) (Muggleton 在 SLPs 的结构学习中则使用基于梯度的方法来代替 EM 方法, 而 Kersting 和 De Raedt 在 BLPs 的参数估计中既形成了基于 EM 的算法也形成了基于梯度的算法)。EM 算法的原理是: 如果所有随机变量的值都是已知的, 那么学习将是比较容易的 (学习就是频率计数问题)。因此 EM 算法首先估计那些缺失的值, 接着用这些值去最大化似然函数, 依次迭代, 直至收敛^[5,6]。一般而言, EM 算法总是假定参数已经被初始化 (例如可以对参数进行随机初始化)。

PLPs, PRMs, BLPs 的参数估计方法都遵循相同的原则, 即给定数据和当前模型, 归纳解释数据实例的贝叶斯网络, 归纳得到的贝叶斯网络的参数则按照标准的贝叶斯网络参数估计方法来估计, 在 E 步依靠传统的贝叶斯网络推理方法填充缺失值, 并由此确定未观测值的分布, M 步中则更新参数值。不同的地方是网络中的不同节点的参数 (相应于相同子句的不同实例) 必须是同样的, 这与动态贝叶斯网络或循环的神经网络是类似的。

3.2 结构学习

为了从数据中既学习结构 L 又学习参数 λ , 往往会给出其它的一些信息, 如: 声明偏置 (declarative bias)、背景理论、初始假设、评分函数和覆盖标准等。

声明偏置: 在单关系情况下模式语言就包含大量模式, 多关系情况下模式规模更大, 所以必须通过明确的约束来限制模式空间的大小。在 PLL 中可以使用句法和语义偏置机制来限定搜索空间, 已经存在各种限制搜索空间大小的形式化方法, 但基本思想是一致的, 都是要限制子句数量。最直接的方法是限制变量的数目或子句的文字数目, 还可以给出子句的句法限制或用语法规则说明可接受子句的范围, 也可以通过模式声明指定子句如何求精。

背景理论: 实践证明, 合理的利用背景知识可以提高学习的效率和准确率, 因此需要与领域专家紧密合作, 以提供有效的背景知识, 降低学习的难度。

初始假设: 一般 PLL 学习算法, 都假定假设已经被初始化, 并由初始假设开始搜索假设空间, 因此初始假设的选择也将影响学习的结果, 为了提高学习准确率, 可以选择不同的初始假设多次学习。

评分函数: 常用的评分函数有考虑先验信息的贝叶斯方法和最小描述长度方法等, 一般评分函数都倾向于选择相对更简单的假设。

覆盖标准: 根据数据类型和表示方法的不同可以选择不同的覆盖标准, 如 Luc De Raedt 和 Kristian Kersting 将 PLL 分为三类: 从概率逻辑后承中学习 (Learning from Probabilis-

tic Entailment)、从概率解释中学习 (Learning from Probabilistic Interpretations) 和从概率证据中学习 (Learning from Probabilistic Proofs), 这三类学习方法分别选用不同的覆盖标准。

几乎所有的基于评分函数的结构学习方法都用启发式搜索方法, 一般使用爬山法或者柱状搜索方法搜索假设空间, 当找到的假设满足逻辑限制并且评分函数不再改进时则停止搜索。在搜索过程中, 一般要用求精操作, 即加入一个原子到子句中或者从子句中删除一个原子, 对于 PRMs、SLPs 和 BLPs, 当一个文字加入到逻辑假设时, 相当于向贝叶斯网络中加入多条边。从逻辑的观点看, 这些求精操作实现了基本的概化和特化操作 (这些操作通常是基于 θ 包含的)^[5,6]。此外, 贝叶斯网络结构的无环性可以通过检查求精算子和例子得到保证。

总结 概率逻辑学习是概率推理和一阶逻辑 (关系逻辑表示) 以及机器学习的结合, 它从关系数据集中构建概率模型, 扩展了 BNs 学习和其它的相关技术, 它可以处理多表领域, 可以表示相同表的不同行间的关系, 而且能够整合多个不同数据库中的数据。2000 年以来文献中已经介绍了各种概率逻辑模型, 其中很多模型 (如: PRMs 和 BLPs) 的参数估计方法和结构学习算法也都已经形成, 并在测试集上取得了不错的成果。大多有关 PLMs 的论文中都包括一些比较分析, 但是这些论文大都是同原有的命题学习系统和 ILP 系统作比较, 各种 PLMs 之间的性能比较很少提及。此外带噪声关系域本身的复杂性就决定了 PLL 的复杂性, 这就使得 PLL 难于被人理解而且学习效率较低, 这很大程度上限制了 PLMs 的发展和应用。为此, 深入研究目前提出的各种 PLMs, 分析各自的优劣, 比较性能, 建立更易于理解的概率逻辑模型, 提出更高效的概率逻辑学习方法都是值得研究的方向。

参考文献

- 1 Heckerman D, Meek C, Koller D. Probabilistic Models for Relational Data; [Technical Report MSR-TR-2004-30]. Redmond, WA: Microsoft Corporation, 2004
- 2 Koller D. Probabilistic relational models. In: S. Dzeroski, P. Flach, eds. Proceedings of 9th International Workshop on Inductive Logic Programming (ILP-99). Springer, 1999
- 3 Fierens D, Blockeel H, Bruynooghe M, Ramon J. Logical bayesian networks. In: Proceedings of the 3rd workshop on Multi-Relational Data Mining (MRDM-2004), Seattle, WA, USA, 2004. 19~30
- 4 Kersting K, De Raedt L. Bayesian logic programs. In: Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming, 2000
- 5 De Raedt L, Kersting K. Probabilistic Inductive Logic Programming. Invited paper. In: S. Ben-David, J. Case, A. Maruoka, eds. Proceedings of the 15th Int. Conf. on Algorithmic Learning Theory (ALT-2004), Padova, Italy, October 2004. 19~36
- 6 De Raedt L, Kersting K. Probabilistic Logic Learning. ACM-SIGKDD Explorations, special issue on Multi-Relational Data Mining, 2003, 5(1): 31~48