概率图模型学习技术研究进展

刘建伟 [1] 黎海恩 [1] 罗雄麟 [1]

摘 要 概率图模型能有效处理不确定性推理, 从样本数据中准确高效地学习概率图模型是其在实际应用中的关键问题. 概率图模型的表示由参数和结构两部分组成, 其学习算法也相应分为参数学习与结构学习. 本文详细介绍了基于概率图模型网络的参数学习与结构学习算法, 并根据数据集是否完备而分别讨论各种情况下的参数学习算法, 还针对结构学习算法特点的不同把结构学习算法归纳为基于约束的学习、基于评分搜索的学习、混合学习、动态规划结构学习、模型平均结构学习和不完备数据集的结构学习. 并总结了马尔科夫网络的参数学习与结构学习算法. 最后指出了概率图模型学习的开放性问题以及进一步的研究方向.

关键词 概率图模型, 贝叶斯网络, 马尔科夫网络, 参数学习, 结构学习, 不完备数据集

引用格式 刘建伟, 黎海恩, 罗雄麟. 概率图模型学习技术研究进展. 自动化学报, 2014, **40**(6): 1025−1044

DOI 10.3724/SP.J.1004.2014.01025

Learning Technique of Probabilistic Graphical Models: a Review

LIU Jian-Wei[1] LI Hai-En[1] LUO Xiong-Lin[1]

Abstract Probabilistic graphical models are powerful techniques to deal with uncertainty inference efficiently, and learning probabilistic graphical models exactly and efficiently from data is the core problem to be solved for the application of graphical models. Since the representation of graphical models is composed of parameters and structure, their learning algorithms are divided into parameters learning and structure learning. In this paper, the parameters and structure learning algorithms of probabilistic graphical models are reviewed. In parameters learning, the dataset is complete or not is also considered. Structure learning algorithms are categorized into six principal classes according to their different characteristics. The parameters and structure learning of Markov networks are also presented. Finally, the open problems and a discussion of the future trend of probabilistic graphical models are given.

Key words Probabilistic graphical models, Bayesian network, Markov network, parameter learning, structure learning, incomplete dataset

Citation Liu Jian-Wei, Li Hai-En, Luo Xiong-Lin. Learning technique of probabilistic graphical models: a review. *Acta Automatica Sinica*, 2014, **40**(6): 1025−1044

概率图模型能简洁有效地表示变量间的相互关系, 为不确定性推理体系提供强有力的工具, 近年来已成为人工智能和机器学习的热门研究领域. 目前, 概率图模型已在图像分析、生物医学和计算机科学等多个领域成功应用. 在利用概率图模型进行不确定推理之前, 需要先根据领域知识构造出概率图模型. 过去, 概率图模型的构造通常由领域专家利用专业知识进行人工构造, 但是该过程耗时长而且实现过程复杂. 在当今的信息时代, 从样本数据中学习概率图模型已成为主要的构造手段.

目前, 常用的概率图模型主要有贝叶斯网络 (Bayesian network, BN) 和马尔科夫网络 (Markov network, MN). BN 是有向图模型, 而 MN 是无向图模型, 两者的学习算法存在一定的区别. 由于概率图模型的表示分参数表示和结构表示两个部分, 因此学习算法也分为参数学习与结构学习两大类.

BN 的参数学习假设结构已知, 然后从样本数据中学习每个变量的概率分布. 概率分布的形式一般已预先指定, 如多项式分布、高斯分布和泊松分布等, 只需利用一定的策略估计这些分布的参数. 学习过程中需要注意样本数据集的观测程度. 若每一个样本都是所有随机变量的充分观测样例, 则数据集为完备数据集; 若样本中缺失某些变量的观测样例或者存在不可能被观测的隐变量, 则数据集为不完备数据集. 而不完备数据集的数据缺失假设有三种情况: 随机缺失、完全随机缺失和非随机缺失. 根据

收稿日期 2013-06-05 录用日期 2013-08-01
Manuscript received June 5, 2013; accepted August 1, 2013
国家重点基础研究发展计划 (973 计划) (2012CB720500), 国家自然科学基金 (21006127), 中国石油大学 (北京) 基础学科研究基金 (JCXK-2011-07) 资助
Supported by National Basic Research Program of China (973 Program) (2012CB720500), National Natural Science Foundation of China (21006127), and Basic Subject Research Fund of China University of Petroleum (JCXK-2011-07)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 中国石油大学 (北京) 自动化研究所 北京 102249
1. Research Institute of Automation, China University of Petroleum, Beijing 102249

数据集的不同观测情况, 相应的学习算法也有所不同.

BN 的结构学习是贝叶斯网络学习的最主要部分, 此时 BN 的参数与结构均未知, 需要从样本数据中找到与数据匹配度最好的网络结构. 当确定网络结构之后, 参数学习只是相对简单的参数估计问题. 结构学习的精度取决于其学习目标, 而学习目标主要有两种: 知识发现和密度估计. 知识发现就是通过研究所学结构的节点依赖性, 发现相关变量的依赖关系. 这种依赖关系由结构中的边反映出来, 所以结构学习越精确越好, 最好的情况就是能恢复真实结构. 密度估计, 即估计真实分布的统计模型, 从而能预测新观测数据的后验概率, 典型的如分类任务. 密度估计需要结构对数据具有很好的泛化能力.

结构学习算法主要分为三类: 基于约束 (Constraint based, CB) 的学习算法、基于评分搜索 (Scoring and searching, SS) 的学习算法和混合学习算法. CB 算法认为结构学习问题是约束满足问题, 通过检验条件独立性 (Conditional independence, CI) 来构建结构. SS 算法则把结构学习问题表述为结构优化问题, 利用评分函数评价每个候选结构, 然后搜索出分数最高的结构. CB 算法的实现比较简单, 但是高阶 CI 检验很复杂并且结果不一定可靠. 虽然 SS 算法是目前使用最为广泛的结构学习算法, 可以灵活地把专家经验知识以结构先验概率分布的形式融入到学习过程中, 并可处理不完备数据情况, 但算法收敛速度慢, 计算复杂. 混合学习算法把 CB 算法和 SS 算法结合起来, 充分利用两者的优点, 可显著提高学习速度并可处理大型网络. 混合学习算法一般先使用 CI 检验构造变量序列或缩减搜索空间, 然后再通过 SS 算法进行结构学习.

除了这三类主要的结构学习算法外, 还有动态规划结构学习、模型平均结构学习和不完备数据集的结构学习等算法. 动态规划结构学习算法与 SS 算法很相似, 但不是使用搜索过程学习结构, 而是利用动态规划寻找最优模型. 动态规划结构学习是求新观测数据的后验概率预测值的精确方法, 而模型平均则是估计新观测数据后验概率的近似方法. 模型平均算法可返回多个模型, 而不是单个模型. 因为当数据样本不足时很难选择出分数最高的模型, 若非要选择一个相对较高分数的模型, 则这种选择带有任意性, 缺乏对结构的置信度.

MN 的学习任务比 BN 的学习更加复杂. MN 的参数学习中, 其划分函数耦合了结构中的所有参数, 使得参数学习问题不能分解, 不能分别独立估计每个局部参数. 而且 MN 的最优参数没有解析解, 所以一般使用迭代方法求解最优参数, 如梯度下降法. 庆幸的是, 迭代中的似然目标函数为凹函数, 迭代方法能保证收敛至全局最优解. 但是迭代中的每一步都需要在网络中进行推理, 使得计算成本相当高. MN 的结构学习也需要计算划分函数, 所以其参数学习存在的问题对结构学习有很大的影响. MN 的结构没有无环性约束, 能有效降低 BN 结构学习过程中遇到的困难, 但 MN 的结构学习要在网络上执行推理, 其学习效率不高. MN 的结构学习也可采用 CB 算法和 SS 算法.

本文对概率图模型的学习算法进行系统分类. 第 1 节先简单介绍 BN 和 MN 的表示理论. BN 的学习技术是目前最热门最广泛的研究内容, 所以第 2 节和第 3 节分别讨论 BN 的参数学习和结构学习算法. 第 4 节分析马尔科夫网络的学习. 第 5 节则给出概率图模型学习算法的新挑战, 而第 6 节指出概率图模型学习的研究趋势. 最后, 第 7 节对本文进行总结.

1 概率图模型

概率图模型结合图论与概率论来紧凑地描述多元统计关系. 概率图模型有多种表示类型, 如贝叶斯网络、马尔科夫网络、链图 (Chain graph, CG)、暂态模型 (Temporal model, TM) 和概率关系模型 (Probabilistic relational model, PRM) 等. 虽然这些表示类型各异, 但主要思想都是利用条件独立性假设来对联合概率分布进行因式分解, 简化表示形式和推理计算过程. 本文主要研究贝叶斯网络和马尔科夫网络的学习算法, 因此下面对这两种网络进行简单的介绍. 概率图模型更多的详细介绍可参考文献 [1−2].

1.1 贝叶斯网络

BN 能够表示随机变量集 $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ 的联合概率分布, 它由两个部分组成: 网络拓扑结构和参数. BN 的结构为有向无环图 (Directed acyclic graph, DAG), 节点表示变量, 有向边表示变量间的条件依赖关系. BN 的参数是节点随机变量的条件概率分布, 即已知父节点时该变量的条件概率分布. 根据结构中隐含的独立性假设: 已知父节点时, X_i 与其非子节点条件独立, 那么可以把联合概率分布分解为多个条件概率分布的乘积:

$$P(X_1, \cdots, X_n) = \prod_i P(X_i | Pa_{X_i}) \qquad (1)$$

其中, Pa_{X_i} 表示变量 X_i 的父节点.

1.2 贝叶斯网络的应用

贝叶斯网络最早的应用之一是奥尔堡大学开发的 MUNIN 专家系统, 用于辅助肌电图的诊断, 通过对人类神经肌肉系统建模, 能够处理多于 1 000 个变

量之间的关系学习. 同期开发的 Hugin 专家系统则通过比较直观和易于使用的界面利用 BN 进行医疗诊断, MUNIN 和 Hugin 系统极大推动了概率图模型的发展, 是 BN 在很多其他领域成功应用的基础.

由于 BN 的强大推理能力, BN 已在众多领域获得成功应用, 如医疗诊断、临床决策、生物信息学、法医学、语音识别、风险分析和可靠性分析等. 文献 [3] 详细综述了 BN 在多个领域的实际应用示例, 并针对每个领域给出具体的建模以及推理过程. 文献 [4] 给出了 BN 在法医学和基因学方面的应用综述, 归纳了 BN 在这两方面的应用发展. 而文献 [5] 综述了 BN 近几年在可靠性分析、风险分析和机器维护领域的应用.

随着 BN 的不断发展, 研究人员开发出很多 BN 的应用软件包. Korb 和 Nicholson 的书[6] 对 BN 的部分软件包进行了详细地比较. 目前, 关于 BN 的学术类软件包主要有 Elvira[7]、BN PowerConstructor[8]、BNT[9]、BUGS[10]、gr[11]、JavaBayes[12] 和 Tetrad[13] 等, 而商用型软件包主要有 Hugin[14]、BayesiaLab[15] 和 Netica[16] 等.

1.3 马尔科夫网络

MN 是一类表示随机变量间对称影响关系的概率图模型, 也是由拓扑结构和参数两部分组成. MN 的结构为无向图, 节点表示随机变量, 无向边表示变量间的依赖关系. MN 的参数为因子集合, 每个因子就是定义在无向图中某个团上的非负函数, 因子也称为团的势函数. MN 结构中的变量可划分为多个团, 那么联合概率分布可以分解为每个团的因子的乘积:

$$P(X_1,\cdots,X_n)=\frac{1}{Z}\prod_{k=1}^{m}\phi_k(\boldsymbol{C}_k) \tag{2}$$

其中, m 为团数, ϕ_k 为第 k 个团的因子, $\boldsymbol{C}_k$ 为第 k 个团的随机变量集, $Z=\sum_{X_1,\cdots,X_n}\prod_{k=1}^{m}\phi_k(\boldsymbol{C}_k)$ 为划分函数. 在学习任务中, MN 通常用对数线性模型表示, 即每个因子表示为变量集的特征函数的指数加权求和:

$$P(X_1,\cdots,X_n)=\frac{1}{Z}\exp\left(\sum_i\omega_i f_i(\boldsymbol{C}_i)\right) \tag{3}$$

特征函数 $f_i(\boldsymbol{C}_i)$ 为变量集 $\boldsymbol{C}_i$ 的任意实值函数, ω_i 为 $f_i(\boldsymbol{C}_i)$ 的权值.

1.4 马尔科夫网络的应用

由于 MN 为非因果模型, 能够灵活地表示变量间的相互作用, 因此 MN 最常用于计算机视觉和图像处理领域, 为像素点之间的关系建模. 目前, MN 的主要应用包括图像重构[17]、图像分割[18]、图像恢复[19]、3D 视觉[20]、目标识别和目标匹配[21] 等. 文献 [22] 总结了 MN 在图像处理领域的多种应用, 给出了 MN 在图像分析中的建模理论、方法和最新研究进展. 文献 [23] 介绍了 MN 在图像分割、图像超分辨率和图像恢复等方面的成功应用, 讨论了 MN 应用的最新的算法.

除了在计算机视觉方面的出色表现之外, MN 在基因网络建模中也十分热门. Wei 和 Li 利用离散马尔科夫随机场为基因网络建模[24], 而 Wei 和 Pan 把基因网络表达为高斯马尔科夫随机场[25]. 文献 [26] 对 MN 在基因表达中的不同应用进行了比较.

2 参数学习

贝叶斯网络的参数学习问题也是很多结构学习算法的一部分, 因为结构学习的前提是结构与参数均未知. 参数学习过程假设网络结构已知, 从数据集 $\boldsymbol{D}=\{\xi[1],\cdots,\xi[K]\}$ 中学习每个变量的条件概率分布. 条件概率分布的参数模型已预先指定, 只需估计其中的参数. 根据样本数据的观测程度, 数据集 $\boldsymbol{D}$ 可分为完备数据集和不完备数据集. 完备数据集就是每一个样本 $\xi[K]$ 都是所有随机变量的充分观测样例, 而不完备数据集则缺失某些变量的观测样例或者存在不可能被观测的隐变量.

当 $\boldsymbol{D}$ 为完备数据集时, BN 的参数学习算法主要有两种: 极大似然估计 (Maximum likelihood estimation, MLE) 算法和贝叶斯估计 (Bayesian estimation, BE) 算法[27]. MLE 和 BE 算法都是统计学中的典型算法, 但它们在 BN 的参数学习中所要估计的参数有所不同. MLE 的参数为条件概率表中每个节点的实际概率, 而 BE 的参数则为条件概率表中概率的条件密度函数.

当 $\boldsymbol{D}$ 为不完备数据集时, 参数学习问题需要借助近似方法来求解. 其算法分三种情况: 随机缺失、完全随机缺失和非随机缺失.

2.1 利用完备数据集学习参数

大多数参数学习研究工作都假设 BN 的变量为离散变量, 但是实际应用中却常常遇到连续变量. 下面根据离散变量和连续变量分别对参数学习算法进行总结.

2.1.1 离散变量

最常见的离散变量是多项式变量, 具有有限个可能取值. 含有多项式变量的 BN 参数学习通常可使用 MLE 或者 BE 方法求解. 但是, 当数据集为稀疏结构时, 即某些概率没有定义时, MLE 算法估计得到的条件概率表中很多项为 0, 使得后续的推理过程出现问题. 因此, 一般使用 BE 算法进行参数学习.

BE 算法先对变量定义一个先验分布, 然后根据数据来更新该先验分布. 假设参数互相独立并且参数的密度函数严格为正时, 先验分布可用 Dirichlet 分布表示, 而且更新后的分布依然为 Dirichlet 分布[28].

文献 [29] 提出一种收缩参数估计算法, 对 MLE 进行平滑, 能有效处理样本噪声带来的方差增大以及克服数据的过拟合现象, 并且能在结构学习中加速参数估计过程.

2.1.2 连续变量

连续变量的最简单的参数估计方法是对变量进行离散化后再估计, 但是会损失大量信息, 估计值准确性不高. 最常见的方法是对连续变量的概率密度函数进行模型假设. 当连续变量的概率密度函数假设为正态分布时, 父节点的概率密度函数服从条件高斯密度, 该网络也称为条件高斯网络, 此时可利用连续样本数据来学习参数[30−31]. 但是, 当网络较复杂时这种高斯近似并不是最优的近似.

为了消除高斯近似的过多约束条件, 可以使用核密度估计方法[32−33]. 核密度估计是一种非参数方法, 克服高斯近似过于简单的特点, 能够有效近似复杂网络分布.

另一种克服过多约束条件的方法是半参数方法, 不要求概率密度函数为某种特定类型, 近似后的模型大小只随求解问题的复杂度增加而增加. 最常用的半参数方法为混合模型 (Mixture models) 方法[34]. 混合模型的参数学习一般使用 MLE 算法或期望最大化 (Expectation-maximization, EM) 算法, 而文献 [35] 提出低阶矩法 (Method of moments) 对高维混合模型进行参数学习, 其估计值具有更低的方差, 并给出了严格的无监督学习结果, 精确参数估计的抽样复杂度也只是混合分量个数的多项式. Hsu 和 Kakade 则给出一种矩估计方法来进行球面高斯混合模型的参数学习[36], 其计算效率高并且能保证统计一致性.

2.2 利用不完备数据集学习参数

不完备数据集的数据缺失假设有三种: 随机缺失 (Missing-at-random, MAR)、完全随机缺失 (Missing-completely-at-random, MCAR) 和非随机缺失 (Missing-not-at-random, MNAR). MAR 假设下, 缺失值取决于观测数据, 可从观测数据中估计缺失数据. MCAR 假设下, 缺失值与观测数据和缺失数据都无关, 通常是由于隐变量的存在, 且不可能得到隐变量的观测值. MNAR 假设是最复杂的情况, 缺失值同时取决于观测数据和缺失数据, 这意味着需要已知缺失数据的数学模型, 而该模型一般未知. 文献 [37] 对不完备数据集下 BN 的参数学习方法进行了简单的综述, 并给出特定的应用例子.

2.2.1 随机缺失

MAR 情况下的参数学习算法主要为吉布斯抽样和 EM 算法.

吉布斯抽样[38] 是最基本的学习算法, 可应用于任意图模型, 如有向图和无向图模型, 其随机变量可以是离散的或者连续的. 吉布斯抽样从已知的数据中估计缺失的数据, 从而把数据集补充完整, 再利用该完整数据集来学习参数. 但是当连续样本之间具有相关性时, 吉布斯抽样不一定收敛.

EM 算法[39] 对模型参数的极大似然 (Maximum likelihood, ML) 函数和最大后验 (Maximum a posteriori, MAP) 函数进行搜索. 但是, 存在大量缺失数据时, EM 算法容易陷入局部最优值. 而且, EM 算法对初始点的选取很敏感, 若初始点离最优值较远, 则参数学习结果不可信.

研究人员提出很多方法来避免陷入局部极大值. 文献 [40] 提出 IB-EM (Information-bottleneck EM) 算法来学习带有隐变量的 BN 的参数. IB-EM 把学习问题看作是两个信息论目标函数之间的权衡, 其中一个目标函数使隐变量不为特定样例的识别提供任何信息, 而另一个目标函数则使隐变量为观测数据提供信息. 在简单贝叶斯网络中, IB-EM 算法比 EM 算法性能更优越. 文献 [41] 则通过随机置换训练数据来避免局部极大值. 虽然数据随机置换法能找到更优解, 但它依然是启发式方法, 不一定能避免陷入局部极大值, 且收敛速度较慢.

2.2.2 完全随机缺失

隐变量的存在使得参数学习过程中出现 MCAR 现象, 无法得到隐变量的观测值, 此时的参数学习问题为病态问题, 因此需要结合隐变量的先验知识来学习参数. 先验分布的参数约束方法包括 Dirichlet 先验分布法、参数共享法和定性约束法. 文献 [42] 指出使用 Dirichlet 先验分布表示先验分布时存在多个问题, 不能表示参数间的等式约束关系, 而且专家也无法说明参数的完备 Dirichlet 先验分布. 参数共享法允许多个先验分布模型共享相同的参数, 并且可使用参数间的等式约束关系. 但是, 参数共享法不能获得参数间的复杂关系, 如不等式约束关系. 而定性约束 (Qualitative constraints)[43−45] 能克服上述问题, 它根据先验知识使用一系列等式和不等式来表示参数的取值范围或参数间的关系, 清晰描述了参数的近似定性关系, 并把定性关系融合到参数估计过程中, 可有效简化学习空间并避免陷入局部最优值.

2.2.3 非随机缺失

针对数据的 MNAR 情况, Ramoni 和 Sebastiani 提出约束收缩法 (Bound and collapse,

BC)[46−47], 有效解决 MNAR 假设下的 BN 参数学习问题. 他们还把 BC 算法与 EM 和吉布斯抽样算法相比较, 指出 BC 算法的收敛速度比其他两种算法要快. 但是, BC 算法仍需要知道缺失数据的模型, 而该模型信息一般不可获取. 随后, 他们提出鲁棒贝叶斯估计法 (Robust Bayesian estimator, RBE)[48], 并不需要对缺失数据类型作出假设, 而是计算出含有估计值的概率区间, 此概率区间的长度是关于数据集信息的单调上升函数, 因此可度量数据所传递的信息. RBE 算法对任意类型的缺失数据都具有鲁棒性.

此外, 文献 [49] 还根据 EM 算法给出 AI & M 算法 (Adjusting imputation and maximization). 在 MNAR 假设下的参数估计过程中, AI & M 算法的估计精度比 EM 算法要高. 但是, 该算法依然存在具有多个极大值的问题, 估计的参数有可能无法辨识.

3 结构学习

BN 的结构学习问题就是在结构和参数都未知的前提下, 从样本数据中选择与数据匹配最好的模型. BN 的结构学习是 NP 难问题[50], 候选结构的数量会随着节点数的增加而指数倍增加, 而实际应用场景一般都具有大量变量, 因此评价所有候选结构的任务不现实. 结构学习算法主要分为基于约束的学习、基于评分搜索的学习和混合学习算法三类. CB 算法认为 BN 编码了一系列条件独立性关系, 通过 CI 检验来辨识变量间的依赖关系和独立关系, 然后建立能满足这些依赖和独立关系的网络结构. SS 算法是目前使用最广泛的结构学习算法, 它把结构学习问题处理为模型选择问题, 由评分函数和搜索算法组成. 混合学习算法把 CB 算法和 SS 算法结合起来, 利用两者的优点, 可显著提高学习速度并可处理大型网络.

除了这三类主要的结构学习算法外, 还有动态规划结构学习、模型平均结构学习以及不完备数据情况下的结构学习等. 下面分别讨论各种结构学习算法.

3.1 基于约束的学习算法

CB 学习算法, 一般使用 CI 检验或互信息来辨识变量间的依赖关系和独立关系, 然后建立满足这些关系的网络. 但是, CB 算法的性能取决于 CI 检验的次数和约束集的大小. 约束集越大, 高阶 CI 检验的次数越多, 则算法的精度越低, 所以 CB 算法一般适用于稀疏贝叶斯网络. 而且, 结构学习过程中对检验错误具有高度敏感性, 当某次 CI 检验出错时, 会直接误导后续的检验结果.

Spirtes 等归纳了 CB 算法的早期研究成果[51]. 最早利用 CI 检验来学习 DAG 结构的算法为 SGS 算法[51], 它首先构造一个无向图, 然后通过 CI 检验过程来删除图中的冗余边. 但是该算法效率很低, 每对变量间的 CI 检验都需要利用其他所有变量子集, 使得运算次数为变量个数的指数次幂. 在此基础上提出的 PC 算法[51] 能有效提高学习效率, 但是由于该算法只是利用邻节点子集来检验节点 X 和 Y 的 d-分离性, 所以在删除弧的过程中很容易出错. 文献 [52] 讨论了 PC 算法在高维数据中的适应性, 而文献 [53] 给出了如何控制 PC 算法执行过程中的错误率的方法.

SGS 算法的另一种变型为 Pearl 和 Verma 提出的 IC (Inductive causation) 算法[54]. 与 SGS 算法和 PC 算法不同, IC 算法考虑到隐变量的存在, 首先构造出一个能为变量关系建模的局部无向图, 图中的边可能为无向的、单方向的或双方向的. 后来, Spirtes 根据这种思想把 PC 算法改进为 IG 算法[51].

Cheng 等基于互信息和 CI 检验提出一种三阶段的依赖性分析算法[55], 称为 TPDA (Three-phase dependency analysis) 算法. TPDA 算法分 Drafting、Thickening 和 Thinning 三个阶段, 可以在 $O(n^4)$ 次 CI 检验内学习到具有 n 个变量的 DAG 结构. 然而, TPDA 算法需要变量依赖关系满足单调性假设, 文献 [56] 指出该假设只适用于少量特定情况, 而这些特定情况中已经有快速的学习算法.

为了有效减少计算时间和构建网络时出现的错误, Yehezkel 和 Lerner 提出 RAI (Recursive autonomy identification) 算法[57], 递归地进行 CI 检验、边定向和结构分解, 并对更小的结构使用更高阶的 CI 检验. Xie 和 Geng 还提出另一种递归算法[58], 他们递归地把变量分成两个子集, 直到子集不能再细分的时候就构造一个 DAG, 然后根据分离的方向来组合这些 DAG, 从而得到最终的网络结构.

文献 [59] 为了解决大型 BN 的结构学习问题, 提出两种 CB 算法用于学习 BN 的超结构, 即学习包含 BN 骨架的无向图模型. 其中一种算法称为 Opt01SS, 只利用零阶或一阶 CI 检验学习超结构; 另一种算法称为 OptHPC, 对 HPC (Hybrid parents and children) 算法[60] 进行计算优化并用于超结构学习.

3.2 基于评分搜索的学习算法

SS 学习算法把结构学习问题处理为模型选择问题, 由评分函数和搜索算法两部分组成. 评分函数用于评价候选结构与数据的拟合度, 拟合越好, 则评分越高. 搜索算法在候选结构组成的空间上搜索评

分最高的结构. 但是, 由于候选结构空间大小随着节点数的增加而指数增加, 所以搜索任务是 NP 难的. 搜索空间一般分三种: DAG 组成的空间、DAG 等价类组成的空间和变量序列组成的空间. 大多数搜索算法都是基于 DAG 空间开发的. 当在 DAG 等价类空间或变量序列空间上搜索时, 结构学习过程会更快.

3.2.1 评分函数

评分函数用于评价结构与数据样本的拟合程度, 拟合度越高, 则分数越高. 根据不同的假设条件和拟合度的不同定义, 评分函数可分为多种. 评分函数必须满足两个重要的假设: 评分等价性以及可分解性. 评分等价性即满足相同独立性关系的等价网络所获得的分数相同. 而可分解性则保证评分函数能分解为多个子函数的累加, 当结构发生局部变化时, 并不会影响其他结构部分的分数, 能有效减少搜索过程的计算复杂性. 表 1 给出了几种典型的评分函数. 本小节分别讨论各种评分函数.

1) BD 评分

Cooper 和 Herskovits 首先提出一种 BN 结构学习的贝叶斯评分函数, 称为 CH 评分或 K2 评分函数[61]. 已知结构 S 和数据集 $\boldsymbol{D}$ 时, K2 评分函数为

$$P(S,\boldsymbol{D}) = P(S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}! \tag{4}$$

其中, n 为变量数, q_i 为变量 i 的父节点数, r_i 为变量 i 的取值数, N_{ijk} 为当父节点是变量 j 时变量 i 取 k 值的次数, 并且有 $N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$. $P(S)$ 是结构 S 的先验概率分布, 而等式中的其他成分为已知数据时结构的似然函数.

后来, Heckerman 赋予该等式更加可靠的理论基础, 提出 BD 评分函数[62]:

$$P(S,\boldsymbol{D}) = P(S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}+\alpha_{ijk})}{\alpha_{ijk}} \tag{5}$$

这里, $\Gamma(x)=(x-1)!$ 为 Gamma 函数. 参数 α_{ijk} 表示配置 ijk 的先验知识, α_{ijk} 越高, 那么配置 ijk 的可能性越大, 并且有 $\alpha_{ij}=\sum_{k=1}^{r_i}\alpha_{ijk}$. BD 评分就是在已知结构时通过计算样本数据的边际似然函数来对 BN 结构进行评分. 而似然函数的计算需要对 BN 的参数作 Dirichlet 先验分布假设. BD 评分在实际中并不常用, 因为需要确定所有超参数 α_{ijk}, 其工作量过大.

当 $\alpha_{ijk}=1$ 时, BD 评分即为 K2 评分.

当给出似然等价性约束时, 即等价结构获得的分数相同时, 那么 BD 评分函数就改进为 BDe 评分函数.

当假设所有配置的可能性相同时, 即 $\alpha_{ijk}=\alpha/r_iq_i$ 时, α 为等价样本大小, 此时的 BD 评分函数

表 1 典型评分函数

Table 1 Classic scoring fuctions

Scoring function	Complete name	Author
CH/K2	Cooper and Herskouits	Cooper and Herskouits (1992)[61]
BD	Bayesian Dirichlet	Heckerman 等 (1995)[62]
BDe	Bayesian Dirichlet with likelihood equivalence	Heckerman 等 (1995)[62]
BDeu	Bayesian Dirichlet with likelihood equivalence and a uniform joint distribution	Heckerman 等 (1995)[62]
AIC	Akaike information criterion	Akaike (1974)[65]
BIC	Bayesian information criterion	Schwarz (1978)[66]
MDL	Minimum description length	Rissanen (1978)[67]
MML	Minimum message length	Wallace (1996)[69]
GU	Globally uniform	Kayaalp and Cooper (2002)[72]
MIT	Mutual information tests	de Campos (2006)[73]
MAP	Max a posterior	Riggelsen (2008)[74]
fNML	Factorized normalized maximum likelihood	Silander 等 (2010)[75]
fCLL	Factorized conditional log-likelihood	Carvalho 等 (2011)[76]

为 BDeu 评分函数. BDeu 评分函数只取决于参数 α. 结构学习对 α 的取值具有高度敏感性, 所以 α 的选择很重要. 文献 [63] 系统地分析了 α 在精确结构学习中的重要作用. Steck 还给出了 α 的优化选择方法[64].

2) AIC 评分和 BIC 评分

AIC 评分[65] 和 BIC 评分[66] 都是基于带惩罚函数的极大似然函数的评分方法. 惩罚极大似然函数的计算过程为

$$P(S,\boldsymbol{D})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}}-f(N)\dim(S) \tag{6}$$

式中第 1 项为似然函数. $\dim(S)$ 为贝叶斯网络的维数, 即表示模型所需的参数个数, $\dim(S)=\sum_{i=1}^{n}q_i(r_i-1)$. $f(N)$ 为非负罚函数, 取决于数据集的大小 N. 罚项是为了避免数据过拟合现象的出现. AIC 评分函数和 BIC 评分函数的区别正在于罚函数 $f(N)$ 的选取. AIC 评分的罚函数为 $f(N)=1$, 而 BIC 评分的罚函数为 $f(N)=\frac{1}{2}\log N$.

3) 最小描述长度 (MDL)

MDL 由 Rissanen 提出, 是一种基于信息论和数据压缩的评分方法. 利用 MDL 原理学习结构时[67], 已知结构 S, 编码数据集 $\boldsymbol{D}$ 的成本等价于描述结构的成本加上描述数据的成本: $Cost(S)+Cost(\boldsymbol{D}|S)$. 那么 MDL 评分过程就是选择总描述成本最小的模型, 这意味着学习过程要在网络结构的复杂性与网络结构表示样本的准确性之间选择平衡点.

文献 [68] 对 MDL 和 BIC 作出详细比较, 将 BIC 定义为

$$\mathrm{BIC}(S,\boldsymbol{D})=-\log P(\boldsymbol{D}|\Theta,S)+\frac{\dim(S)}{2}\log N \tag{7}$$

其中, Θ 为结构 S 的极大似然参数. 另外, 将 MDL 定义为

$$\mathrm{MDL}(S,\boldsymbol{D})=-\log P(\boldsymbol{D}|\Theta,S)+\frac{\dim(S)}{2}\log N+C_k \tag{8}$$

其中, k 为变量个数, $C_k=\sum_{i=1}^{k}(1+|Pa_{X_i}|)\log k$, $|Pa_{X_i}|$ 为变量 X_i 的父节点个数. 由此可以看出, MDL 评分函数实际上是带有模型复杂度惩罚项的 BIC 评分函数.

4) 最小信息长度 (MML)

Wallace 等受通信过程启发, 提出 MML 来对结构进行评分[69]. 类似于 MDL 中的惩罚项, MML 也对过于复杂的模型进行惩罚, 而对能较好拟合数据的模型进行奖励. 其基本思想是在发送器和接收器之间传递最小可能性的消息. Korb 和 Nicholson 以新的方式重新定义了 MML[70]. 文献 [71] 还研究了 MML 在带有节点局部相互作用的贝叶斯网络中的应用.

5) 其他评分函数

除了上述比较常见的评分函数外, 研究人员还给出了其他形式的评分函数. Kayaalp 和 Cooper 基于缺省参数先验分布的特定形式提出 GU 评分函数[72]. 而 de Campos 给出 MIT (Mutual information tests) 评分函数[73], 该评分函数基于信息论, 并且其性能比 K2 评分、BDeu 评分和 BIC 评分要好. Riggelsen 还在 BD 评分的基础上提出一种新的贝叶斯评分函数, 称为 MAP 评分函数[74], 不需要 Dirichlet 假设, 而且可以同时学习参数和结构. Silander 等提出 fNML 评分函数[75], 在无先验知识的前提下学习结构, 该评分函数满足可分解性, 可与目前多种搜索算法相结合. Carvalho 等提出 fCCL 评分函数[76], 为对数似然评分的近似. fCCL 评分与传统的对数似然评分具有相同的时间和空间复杂性, 但是能够获得更优的估计.

3.2.2 搜索算法

大多数 SS 算法都是在 DAG 空间中进行搜索. 即使约束每个节点的最大父节点数, 在如此庞大的空间中搜索最优结构仍然是 NP 难的. 由于搜索任务的困难性, 目前多采用启发式搜索算法, 如贪婪搜索、遗传算法 (Genetic algorithm, GA)、进化规划 (Evolutionary programming)、模拟退火算法、粒子群算法 (Particle swarm optimization, PSO) 和蚁群算法 (Ant colony optimization, ACO) 等. 表 2 给出了典型的 SS 算法, 并列出它们的评分函数、搜索算法和搜索空间. 下面介绍几种 SS 算法中常用的启发式搜索算法.

1) 贪婪搜索

贪婪搜索是结构搜索算法中最基本的启发式算法, 但贪婪搜索是局部搜索方法, 容易陷入局部极大值. 根据变量序列是否已知, 结构学习的贪婪搜索算法分两个研究方向. 若已知变量序列, 那么序列中靠前的变量有可能是后面变量的父节点, 此时的结构搜索比较简单.

基于贪婪搜索, Cooper 和 Herskovits 给出 K2 算法[61], 在已知数据样本和特定变量序列时构造 BN 结构, 所使用的评分函数为 K2 评分函数. Bouckaert 在 K2 算法的基础上提出 K3 算法[77], 也需要已知变量序列, 然后利用数据集来产生 DAG, 但是评分函数使用 MDL 评分函数. Liu 和 Zhu 也基

表 2 典型的基于评分搜索的学习算法

Table 2 Classic scoring and searching learning algorithms

Methods	Scoring functions	Searching procedure	Searching space
K2	K2	Greedy search	DAG
K3	MDL	Greedy search	DAG
K2GA	K2	Genetic algorithm	Ordering
ChainGA	K2	Genetic algorithm	Ordering
MDLEP	MDL	Evolutionary programming	DAG
HEP	CI test and MDL	Evolutionary programming	DAG
HEA	CI test and MDL	Evolutionary programming	DAG
ACO-B	K2	Ant colony optimization	DAG
ACO-E	K2	Ant colony optimization	DAG equivalence classes
K2ACO	K2	Ant colony optimization	Ordering
ChainACO	K2	Ant colony optimization	Ordering

于特定变量序列, 把结构搜索看做特征选择问题[78]. 文献 [79] 改进了贪婪爬山算法, 在搜索过程中对将要评估的候选结构进行动态约束, 证明了在特定条件下算法返回的解是训练数据联合概率分布的最小独立映射 (Independence map, I-map).

还有其他学者研究变量序列未知时的结构学习贪婪搜索算法. Heckerman 等提出 BD 评分[62], 并利用贪婪搜索来验证该评分函数, 搜索中的每一步都在当前 DAG 上增加边、删除边或对边取反方向.

2) 遗传算法和进化规划

GA 和进化规划的搜索过程能搜索出更加匹配的 BN 结构, 具有强大的计算能力. GA 在结构搜索中的使用逐渐受到人们的关注. 文献 [80] 首先使用 GA 来搜索序列空间, 并用 K2 算法对序列评分. 而 Faulkner 提出 K2GA 算法[81], 在遗传算法执行过程中使用修改后的 K2 算法进行评分. Kabli 等则把 GA 应用在链图上, 提出 ChainGA 算法[82], 首先基于模型评估过程使用 GA 求出最优序列, 然后在序列上连续运行 K2 算法, 返回最优的总体结构. ChainGA 算法能减少评估的数量, 但降低了网络结构质量. 文献 [83] 在孤岛模型 (Island model) 的实现过程中使用 K2GA 算法, 并通过拟合度和边的分析证明了其学习性能比原来的 K2GA 算法显著提升.

Wong 等把 MDL 评分与进化规划相结合, 提出在 DAG 空间上搜索的 MDLEP 结构学习算法[84]. 随后, Wong 等又把带有 CI 检验的评分搜索与进化规划相结合, 给出 HEP (Hybrid evolutionary algorithm) 算法[85]. HEP 算法中, DAG 搜索空间是有约束的, 当不同 DAG 之间有较强的依赖性时, 它们只能连接两个节点. HEP 算法寻找能使 MDL 评分最小的 DAG 结构. 基于前面的工作, Wong 和 Leung 提出 HEA (Hybrid evolutionary algorithm) 算法[86], HEA 算法同样是基于混合的方法. Wong 和 Guo 把 HEA 算法推广到不完备数据的情况, 称为 HEAm 算法[87].

文献 [88] 对进化算法在 BN 学习中的应用情况进行了详细的综述.

3) 模拟退火算法

模拟退火算法能够解决贪婪算法容易陷入局部极大值的问题, 并十分类似于遗传算法, 但是模拟退火算法在结构学习中的有效性还缺乏详细的理论分析. 遗传算法并不是选择最优的邻状态进行移动, 而模拟退火算法则根据状态的评分函数以及迭代的步数选择移动方向. Campos 和 Huete 比较了遗传算法和模拟退火算法在变量序列空间上搜索的不同[89], 他们的实验表明这两种搜索算法所得到的变量序列的质量并没有明显差异, 都能收敛到最优值, 但是模拟退火算法的收敛速度比遗传算法要快.

4) 粒子群算法

粒子群算法是目前比较受欢迎的结构学习的启发式搜索算法. 与 GA 和模拟退火相比, PSO 不仅实现简单, 需要调整的参数较少, 而且适用于各类搜索空间以及具有快速发现最优解的能力. 但是由于经典 PSO 算法只运行在连续实值空间中, 而 BN 的结构学习一般是基于离散空间, 所以目前的研究工作主要是把 PSO 改进为离散空间中的搜索算法. Heng 等通过字母顺序来表示候选结构, 然后定义了计算速率更新和位置更新的规则[90]. 他们还把该方法用于动态贝叶斯网络的结构学习[91]. Li 等提出一种二值 PSO 算法来搜索结构, 避免结构学习过程中的过早收敛[92]. 而文献 [93] 给出分布式粒子群算法.

Wang 和 Yang 对 PSO 中的速率和位置更新法则进行修改, 提出一种基于二值离散 PSO 的结构学习算法[94].

5) 蚁群算法

ACO 算法是一种仿生学算法, 其灵感来源于蚁群通过传播化学信息素而产生的协同合作行为. 它们能有效找出食物与巢穴之间的最短路径. 目前蚁群算法已经应用在 BN 的结构学习中.

de Campos 等首先使用 ACO 算法对 DAG 空间进行搜索, 称为 ACO-B 算法[95]. 后来 Daly 和 Shen 把该方法推广到 DAG 等价类空间的搜索中[96]. Pinto 等还提出 MMACO (Max-min ant colony optimization) 算法[97], 在可能边组成的空间中搜索, 把 ACO 与混合学习算法结合起来, 使用混合学习算法构建网络骨架, 再使用 ACO 算法为骨架中的边定向. 而 Wu 等基于 K2GA 算法和 ChainGA 算法, 把其中的 GA 过程修改为 ACO 过程, 提出两种基于 ACO 的结构学习算法, 分别称为 K2ACO 算法和 ChainACO 算法[98]. 这两种算法都是在变量序列上进行搜索的方法.

类似于 ACO 算法, 文献 [99] 提出一种元启发式算法用于 BN 的结构搜索, 称为人工蜂群 (Artificial bee colony, ABC) 算法. 他们把 ABC 算法与 ACO-B 算法进行比较, 实验结果表明 ABC 算法具有更好的有效性.

3.2.3 在等价类空间中搜索

当使用 DAG 表示 BN 的结构时, 由于满足相同独立关系的结构的 DAG 相同, 则导致结构的描述不是唯一的, 从而引起辨识问题. 为了简化结构学习的过程, 需要对每个特定结构使用独特的表示方式. DAG 等价类就是最经典的表示方式. DAG 等价类中, 满足独立等价关系的结构共享相同的骨架, 并同时含有无向边和有向边. 当所有结构的某条边都满足一个方向时, DAG 等价类中的该边也满足该方向, 而当某些结构出现某条边的方向不一致时, DAG 等价类中的该边为无向边. 前面在总结搜索算法的研究时, 已经提到了某些在等价类空间中搜索的结构学习算法. 文献 [100] 还提出一种基于交互信息的算法在等价类空间中学习 BN 结构. 他们首先利用交互信息和 CI 检验构造一个初始网络, 然后在等价类空间中通过贪婪搜索得到最优结构, 其迭代步数和运行时间显著减少.

Studený 提出了一种新的结构表示方法, 不再使用 DAG 等价类, 而是使用分量为整数的特定向量, 称为 Imsets[101]. 这是一种代数表示方法, 每个结构都统一使用标准 Imsets 表示. 其优点是使得评分函数成为标准 Imsets 的一个映射函数. Studený 等还给出了标准 Imsets 表示的几何观点描述, 认为固定变量集上的标准 Imsets 是某个多面体的顶点[102]. 该几何观点使得学习问题转化为线性规划问题, 在有约束多面集上优化线性函数. 而 Hemmecke 等提出另一种向量表示方式, 称为 Characteristic imsets[103], 该向量由标准 Imsets 通过一对一映射而获得, 向量中只包含 0、1 两种成分. 但是, 目前还没有专门针对向量表示法的结构学习算法.

3.2.4 在序列空间中搜索

序列空间比 DAG 空间要小得多, 因此在序列空间中搜索能更加高效. 而且, 序列空间搜索的每一步都对当前假设给出更加全局的修改, 有效避免陷入局部最优值. 另外, 序列空间无须考虑无环性, 而无环性检查是大型网络中成本很高的运算, 因此在序列空间中搜索能显著降低计算成本. 但是, 序列空间中的搜索需要事先对每个节点和每个父节点集计算充分统计量, 当样本数据集很大时该计算成本也很高.

Larranaga 等首先在序列空间上使用遗传算法进行搜索[80]. Friedman 和 Koller 提出用解析表达式对序列进行评分的有效方法[104], 其节点只具有有限数量的父节点. Teyssier 和 Koller 根据这种序列评分方法, 通过贪婪爬山算法能简单快速地求出最优序列[105].

3.3 混合学习算法

虽然 CB 学习算法相对快速, 并具有定义良好的停止规则, 但是其性能依赖于 CI 检验的次数, 而且前阶段的检验错误对后阶段的检验有连锁反应. 而 SS 学习算法可以灵活地把专家经验知识以结构先验概率分布的形式融入到学习过程中, 并且可以处理不完备数据情况. 但 SS 算法收敛速度慢, 计算复杂. 混合学习算法把这两种算法结合起来, 利用各自的优点, 不需要一开始就构建完整的网络, 而是构建目标节点附近的局部图. 混合学习算法一般先使用 CI 检验来构造变量序列或缩减搜索空间, 然后再通过 SS 算法学习结构.

Singh 和 Valtorta 首先提出混合学习算法, 先使用 CI 检验构建变量序列, 然后把变量序列作为 K2 算法的输入, 进行结构学习[106]. Dash 和 Druzdzel 结合两类学习算法的特征给出 EGS (Essential graph search) 算法, 使用 PC 算法获得局部 DAG 的初始推测, 再扩展为 DAG 并进行贪婪搜索[107]. 与此相反, de Campos 等则提出 IMAPR (I-map restart) 算法[108], 先以随机起始点进行贪婪搜索, 在获得的 DAG 上通过 CI 检验增加或删除边.

为了进一步提高搜索速度, 近年来出现很多对结构进一步约束的混合学习算法. Friedman 等提出

的 SC (Sparse candidate) 算法[109] 在初始阶段并没有构造一个完整结构, 而是使用 CI 检验找出较好的候选父节点, 因此约束了后面阶段的搜索空间大小. 该算法不仅能加速搜索, 还不会过度破坏评分过程. 在此基础上, Tsamardinos 等提出 MMHC (Max-min hill-climbing) 算法[110], 其候选父节点的产生使用 MMPC 算法[111] 实现, 然后使用贪婪搜索算法为边定向. MMHC 算法是目前能在合理时间内学习含有上千个节点的 BN 的最有效的算法. Perrier 等还提出一种算法能在结构约束的情况下学习到最优的 BN[112]. 他们定义这种约束的无向图为超结构, 算法执行过程中考虑到的每个图的骨架都强制约束为超结构的一个子图. 当超结构的平均入度为 2 时该算法能学习到含有最多 50 个节点的最优 BN. 后来, Kojima 根据超结构的概念并基于聚类结构约束提出一种优化搜索策略, 把超结构划分为多个聚类簇, 在每个聚类簇上优化搜索[113]. 该方法可学习具有几百个节点的最优 BN. de Campos 和 Ji 根据分支定界法约束搜索空间[114], 但保证最优解仍然位于约束后的空间中.

3.4 动态规划结构学习

动态规划结构学习算法类似于 SS 算法, 但不再使用搜索过程学习结构, 而是利用动态规划寻找较小子网络的最优模型, 从而利用这些较小的最优模型寻找较大的子网络的最优模型, 直到求出所有变量的最优模型. 动态规划是求新观测数据的后验概率预测值的精确方法.

Ott 等最早利用动态规划学习结构, 给出寻找最优模型的算法并证明了算法的正确性[115]. Ott 和 Miyano 指出当限制父节点个数时, 动态规划方法可应用于任意大小的 BN[116]. 与此同时, Koivisto 和 Sood 给出了类似的算法, 但更深入地分析了动态规划的结构学习问题[117]. 他们的出发点是要计算子网络的后验概率. 后来, Koivisto 给出了更加快速的动态规划算法, 可在 $\mathrm{O}(n2^n)$ 时间内计算所有边的后验概率[118]. 基于 Koivisto 和 Sood 的研究, Singh 和 Moore 提出一种类似的算法, 但优化等式的形式不同. 该算法结构更简单, 存储量更少, 但约束节点的输入边数时, 算法运算很慢[119]. 为了解决动态规划只能利用次模先验分布, 并只能计算次模特征上的后验概率, 而且很难计算密度函数的问题, Eaton 和 Murphy 把动态规划算法作为 MCMC (Markov chain Monte Carlo) 的建议分布, 从而得到学习精度更高的结构[120]. Malone 等对动态规划方法做出修改, 提高了其效率, 把存储量由 $\mathrm{O}(n2^n)$ 减少为 $\mathrm{O}(C(n, n/2))$, 并利用 MDL 评分函数的特点减少了算法的运行平均时间[121].

3.5 模型平均结构学习

BN 的结构学习通常是求出一个最优结构, 但是当样本不足时, 这种方式就会出现问题. 样本不足可能会使得没有某一个模型所取得的分数比其他模型都高得多, 那么就需要在相对较高分数的几个模型中进行选择. 这种选择带有任意性, 缺乏对最终得到的结构的置信度. 解决该问题的一种方法是, 令学习算法返回多个模型, 而不是单个模型, 然后在推理过程中根据模型的概率分别为其加权. 这就是模型平均的基本思想. 模型平均学习出来的结构通常用于密度估计. 虽然模型平均是一种近似方法, 但其预测值比由单个模型获得的预测值更具鲁棒性.

早期的模型平均研究都集中于 Madigan 等提出的随机方法, 他们使用 MCMC 模型分量来进行模型平均, 称为 MC^3[122]. MC^3 方法首先在模型空间上构造一条马尔科夫链, 并逐个模型前进, 每一步都计算出所需的模型并最后对结果求平均. 后来他们把 MC^3 方法推广到 DAG 等价类空间中[123]. Friedman 和 Koller 则在变量序列空间上使用 MCMC 过程进行模型平均, 其混合速度和估计值的鲁棒性都比 DAG 空间中的 MCMC 过程要好[104]. Giudici 和 Castelo 对 MC^3 方法作出改进, 在状态空间中使用新的前进方式, 并给出多个实际应用场景的概率分布的分析[124]. 在此基础上, Grzegorczyk 和 Husmeier 加速了 MC^3 方法的收敛性[125], 而 Liang 和 Zhang 提出了 SAMC (Stochastic approximation Monte Carlo) 方法[126]. 文献 [127] 使用广义动态规划结构算法求出 k 个最优结构, 并对这 k 个最优结构进行模型平均, 其预测性能比目前的模型选择方法和 MCMC 方法都要好.

其他相关的研究包括, Dash 和 Cooper 提出的加快模型平均过程的方法[128], 他们的算法能够找到与平均值等价的一个网络. 基于这种合并的思想, Kim 和 Cho 利用进化算法把多个 BN 合并为单个模型[129]. Gou 等提出的并行 TPDA 算法则在不同数据集上并行使用 TPDA 算法, 然后对所产生的 DAG 结构进行组合操作[130]. 而 Liu 等则使用 CI 检验学习结构, 再把生成的 DAG 组合起来[131].

3.6 利用不完备数据集学习结构

类似于参数学习的情况, 当数据缺失时结构学习也是比较复杂的问题. 此时, 数据缺失同样分三种情况: 随机缺失、完全随机缺失和非随机缺失. 由于多数结构学习算法能处理各种数据缺失情况, 所以下面的讨论不再根据数据缺失情况对结构学习算法分类.

早期, Cooper 和 Herskovits 首先开展在结构学习过程中处理数据缺失情况的研究工作, 他们提出

了 K2 算法和 K2 评分函数, 并给出了如何利用概率法则归纳出缺失数据的所有可能组合情况[61]. 然而这种可能组合事件为缺失项的指数倍. 虽然他们的方法可以应用于隐变量情况, 但是可能缺失情况也是数据个数的指数倍, 并且不知道存在哪些隐变量以及隐变量的取值是什么. CI 检验也可处理数据缺失, Kwoh 和 Gillies 提出一种 CI 方法, 为两个独立节点增加一个共享隐父节点[132], 而 Sanscartier 和 Neufeld 提出的方法可以从上下文独立中辨识出隐变量[133].

虽然上述方法的原理相当简单, 但是计算过程过于复杂. Geiger 等使用 BIC 评分函数为数据缺失情况下的结构打分[134], 其极大似然参数估计过程则使用数据缺失情况下的参数学习算法, 如 EM 算法. Ramoni 和 Sebastiani 使用 BC 算法学习数据非随机缺失下的结构[135]. 而文献 [136] 利用已观察数据的原始关系来辅助不完备数据情况下的结构学习, 获得良好的学习效果.

目前, 从不完备数据集中学习结构的最热门方法是 Friedman 提出的 SEM (Structural expectation maximization) 算法[137]. SEM 算法把模型选择与 EM 算法进行融合, 能估计参数, 并已证明能收敛到局部极大值. Beal 和 Ghahramani 利用变分贝叶斯 EM 算法对 SEM 算法进行改进[138], 而文献 [139] 证明了其有效性. Elidan 提出 BSEM (Bagged SEM) 算法[140], 能显著减少结构学习过程的运行时间.

4 马尔科夫网络的学习

MN 的学习任务比 BN 的学习任务更加复杂. MN 的划分函数包含网络中的所有参数, 使得参数学习问题不能分解, 也就不能分别独立地估计每个局部参数. 而且 MN 的最优参数没有解析解, 所以一般使用迭代方法求解最优参数, 如梯度法. 庆幸的是, 迭代中的似然目标函数为凹函数, 迭代方法能保证收敛至全局最优解. 但是迭代中的每一步都需要在网络中进行推理, 使得计算成本相当高. MN 的结构学习过程仍然需要计算划分函数. 所以其参数学习存在的问题对结构学习有很大的影响. MN 的结构没有无环性约束, 能有效降低 BN 结构学习过程中遇到的困难, 然而 MN 的结构学习要在网络上执行推理, 其学习效率仍然不高. MN 的结构学习也可采用 CB 算法和 SS 算法.

4.1 参数学习

MN 的参数就是其对数线性模型中的权值, 学习过程中的优化目标为权值的似然函数. 然而, 目标函数在 ω_i 点的梯度为数据特征函数期望与模型特征函数期望之差. 因此计算 ω_i 点的梯度时需要在模型上进行推理计算. 有两种方法可以降低参数学习的计算成本: 使用近似推理过程计算梯度, 或使用其他目标函数替代.

近似推理可有效简化参数学习过程, 主要有信任传播近似推理和抽样近似推理两种技术. 有很多学者研究了信任传播算法与 MN 学习之间的关系. Wainwright 等推导出信任传播学习中的伪矩匹配方法[141]. 矩匹配是一类等式约束问题, 即数据的特征期望要与模型的特征期望相等. 而伪矩匹配是指信任传播过程中的团信任要与数据中的经验边际分布相一致. 受此启发, Sutton 和 Minka 引入了分段训练目标函数, 直接在所有网络势函数上进行矩匹配[142]. 后来, Wainwright 还给出证据, 指出使用近似算法训练模型比使用精确推理算法训练模型的性能更好[143]. 他还指出在学习算法的内循环中使用非稳定推理算法 (如和积信任传播) 不利于学习. 而 Ganapathi 等定义了 CAMEL (Constrained approximate maximum entropy learning) 优化问题, 其优化函数为带有期望约束的极大熵近似, 可在单一联合目标函数中学习和推理[144].

除了信任传播近似推理技术之外, 还可使用抽样近似推理过程学习权值, 如 MCMC 方法. Murray 和 Ghahramani 提出几种抽样学习方法, 其平稳分布都只是参数后验分布的近似[145]. Murray 等还提出完美抽样方法, 不需要估计划分函数, 但实用性不高[146].

虽然极大似然函数是 MN 学习中最常用的目标函数, 但它带来的问题是参数学习时复杂的推理过程. 因此, 研究人员提出了其他替代目标函数, 如伪似然函数、对比离差 (Contrastive divergence) 函数和最大间隔函数等. Besag 最早提出了伪似然函数的概念[147]. 原似然函数中的条件概率分布为已知父节点时某节点的条件概率分布, 而伪似然函数中的条件概率分布则为已知其他所有变量时某节点的条件概率分布. 若只需求解少数变量的概率查询时, 伪似然函数能更好地学习模型. McCallum 等提出的多条件学习算法扩展了伪似然函数, 其目标函数中的条件概率分布不再是单个节点的条件概率分布, 而是多个节点上的条件概率分布[148]. 对比离差函数[149] 由 Hinton 提出, LeCun 等总结了对比离差函数在结构学习中的发展过程[150]. 最大间隔函数来源于支持向量机, 使用非线性核函数映射, 可在无穷维特征空间中学习判别函数. Taskar 等提出两种方法处理最大间隔目标函数中指数个约束条件的优化问题[151].

4.2 结构学习

MN 的结构学习也可使用 CB 算法和 SS 算法. 但是, 由于 MN 中的 CB 算法的约束假设比较苛刻,

在实际应用中通常不能实现, 所以其算法研究很少. 文献 [152] 给出了两种基于 CI 检验的 MN 结构学习算法.

最常用的 MN 结构学习算法为 SS 学习算法. 由于 MN 的结构学习中通常使用对数线性模型, 对于两两关系 MN (Pairwise Markov network) 来说, 两个节点间存在边时其对数线性模型相对应的权值不为零, 所以其结构学习问题可认为是特征归纳问题, 学习出相应的权值, 根据不为零的权值即可得到 MN 的结构. 因此, MN 的结构学习任务就是在样本空间中发现高概率区域, 用特征函数表示这些区域, 并且学习其相应的权值. 该过程有两种实现策略: 全局搜索策略, 但权值学习成本很高; 局部学习策略, 先学习出局部模型再组合为全局模型.

4.2.1 全局搜索

全局搜索策略中, 使用搜索算法求解结构学习问题. 该策略利用当前特征集构造出候选特征集, 再对每个特征进行评分, 得分最高的特征就作为所学模型的特征. 由于候选结构数量很多, 所以基于全局搜索的算法运行很慢. 而且, 为每个候选结构评分时需要学习每个特征的权值, 而权值学习需要进行迭代优化, 每次迭代都要在模型上执行推理过程, 计算成本很高.

Della 等给出了最经典的 MN 结构学习算法, 算法中进行自上而下的搜索, 也称为从一般到特殊的搜索[153]. 每一步搜索都把特征函数与使评分提高最多的特征相结合, 从而对特征函数特殊化. Kok 和 Domingos 也在马尔科夫逻辑网的学习中进行这种自上而下的搜索[154]. 但是, 自上而下的搜索会找到很多与数据不匹配的特征函数, 学习效率低, 并很容易陷入局部最优解.

Domingos 在规则归纳领域提出自下而上的学习算法[155], 能克服自上而下搜索的缺点. MN 的特征归纳问题与分类任务的规则归纳问题很相似. 自下而上的归纳从包含很多前提的规则体开始, 然后通过从规则体中去掉前提而把规则一般化, 这就扩大了与规则匹配的样例数. Mihalkova 和 Mooney 基于自下而上的思想提出 BUSL (Bottom-up structure learning) 算法[156], 用于学习马尔科夫逻辑网. 但是 BUSL 算法只是在预处理阶段使用自下而上思想减少候选特征数, 主要的特征归纳过程仍为自上而下. Davis 和 Domingos 提出 BLM (Bottom-up learning of Markov networks) 算法[157], 自下而上地学习 MN 结构. BLM 算法首先把完备样例作为初始特征集, 然后逐步令特征一般化, 即丢弃变量使特征与 k 个最近样本相匹配, 使特征覆盖邻近的高概率区域.

4.2.2 局部学习

局部学习在最近几年受到很多研究人员的关注. 局部学习就是先学习出一系列局部模型, 然后把局部模型组合为全局模型. 该策略依次考虑每个变量, 并构造一个模型来预测当已知其他变量时该变量的取值. 每个预测模型都会被转化为特征集, 这些特征集就构成了最后的全局模型. 但是当数据集包含较多变量时, 局部模型的学习成本相当高.

Ravikumar 等提出的算法把 L1 逻辑斯蒂回归作为局部模型[158]. 在此之前, 已有学者利用 L1 正则化来进行特征选择. L1 正则化同时实现参数学习与特征选择过程, 并强制约束大部分权值为 0. 给出带有初始特征集合的优化算法后, 模型选择就是在优化算法执行后选择带有非零权值的特征函数. 但是, 这些方法只能构造两两关系马尔科夫网络[159]. Pekins 等首次提出使用 L1 逻辑斯蒂回归算法选择特征, 并给出特征选择的梯度启发式算法和基于 L1 的停止准则[160]. 使用 L1 正则化目标函数学习 MN 结构的算法由 Lee 等提出[161]. 他们的算法理论上可学习任意长度的特征, 但实际中只能评价长度为 2 的特征. Yang 等把图模型表示为广义线性模型, 再利用 L1 正则化构造结构学习的目标函数, 实现能够表示为指数分布族的所有图模型的结构学习, 并给出统计保证[162].

Lowd 和 Davis 提出的 DTSL (Decision tree structure learner) 算法[163] 使用概率决策树作为局部模型. 决策树可表示变量间更丰富的结构. DTSL 算法使用概率决策树预测变量的值, 并把决策树转换为变量间的连接特征集. 他们还提出了多种不同的转换方法.

Haaren 和 Davis 提出的 GSSL (Generate select structure learning) 算法分两个阶段学习 MN 的结构, 结合了全局搜索和局部学习的特点[164]. 第一阶段为特征产生阶段, 通过自下而上的搜索产生候选特征空间; 第二阶段为特征选择阶段, 使用 L1 正则化目标函数学习权值. 该算法的精度显著提高, 时间复杂性也有效降低.

5 概率图模型学习算法的新挑战

随着概率图模型在实际领域中的应用日益增加, 不同的学习任务和学习环境对概率图模型的学习算法提出了不同的新要求. 由于概率图模型的结构学习是其学习算法的主要部分, 本节针对结构学习总结概率图模型学习算法所面临的新挑战, 并指出其待解决的问题.

5.1 概率图模型的并行学习

概率图模型的结构学习过程计算复杂性很高,

这势必成为实际应用的一大阻碍. 许多学者致力于提高结构学习的效率. 除了从算法上提高学习效率之外, 并行学习也是加速学习过程的解决策略. 并行学习使得学习问题可以同时由多个计算资源共同处理, 效率显著提高. 例如, SS 算法最大的问题就是求解所需的充分统计量. 当评价不同结构时, 若以并行方式进行, 能极大地解决 SS 算法的计算问题. 已有学者在其构建的算法中融入并行学习的思想. Sahin 和 Devasia 演示了如何利用粒子群算法进行并行学习[93]. Yu 等在 SEM 算法的 EM 部分使用并行学习的策略, 在每个样本上并行运行 E 步骤[165]. Gou 等的 TPDA 算法则并行执行 CI 检验, 然后把结果组合起来[130].

虽然并行学习能有效提高概率图模型的学习效率, 但其实际应用并不多, 因为并行学习在概率图模型中还存在关键的开放性问题仍未解决: 由于概率图模型的样本数据通常存在一定依赖性, 如何在并行学习中对数据进行调度是一个很关键的问题; MN 的参数和结构学习都需要迭代计算, 某些参数在迭代终止时被重新定义, 如何给出动态的迭代平行机制来直接编码这些迭代运算?

5.2 概率图模型的在线学习

概率图模型的学习一般是批处理过程, 即给定数据块, 从数据块中学习该部分数据的模型. 然而, 在实际应用中, 很多系统需要不断地处理新到达的数据, 以修改所学到的模型. 以往的批量学习把旧数据与新数据合并, 重新学习模型, 但该方法不能使用先前的学习结果, 而且所需处理的存储量很大. 研究人员提出了概率图模型的在线学习算法, 也称为增量学习算法. Lam 早就提出 BN 在新数据情况下的学习问题, 在旧网络与新数据间权衡考虑, 先使用新数据学习出一个局部网络, 然后通过该局部网络来改进旧网络[166]. Friedman 和 Goldszmidt 则对精确度和存储量进行取舍, 只保留一部分对结构再学习有用的旧数据[167]. Nielsen 等考虑了存在概念漂移现象的非平稳系统的增量学习[168]. 概念漂移即随着时间的变化, 样本可能来自于不同的概率分布. 他们的方法由两部分组成: 首先监控和检测模型何时应该进行更新, 然后使用局部搜索策略整合与观测数据矛盾的模型部分. Castillo 和 Gama 提出了变化环境中的自适应学习算法[169]. Yasin 和 Leray 把 MMHC 算法改进为增量形式的学习算法[170].

概率图模型的在线学习或增量学习还存在需要解决的问题: 目前多数在线学习算法都是从局部上对模型进行修改, 如何以增量形式来辨识整体的结构; 如何改进算法使其能实时检测系统的漂移情况, 并在漂移情况出现时进行参数的自适应修改; 如何通过计算和存储充分统计量来优化结构的连续更新过程; 如何在爬山搜索过程中使用更加精细的回溯方法等.

5.3 概率图模型的主动学习

从观测数据中学习模型的方法一般只能求出网络结构的等价类, 即表示相同独立集的 BN 网络集. 仅利用观测数据, 我们很难从等价类中精确辨识出网络结构. 然而, 很多实际应用需要知道网络的精确结构, 从而发现变量间的因果依赖关系. 概率图模型的主动学习能够解决该问题, 其学习器主动选择未标注数据, 并交由人类专家进行标注, 使某些特定变量被赋予特定的取值, 然后把人类专家标注的数据整合到原观测数据中, 从而在数据集较小的情况下获得较高的学习精度. 主动学习方法一般分为学习引擎和选择引擎两个部分. Cooper 和 Yoo 最早利用主动学习来发现图中的因果关系[171]. Tong 和 Koller 给出了最优的选择引擎算法[172]. He 和 Geng 使用主动学习的策略学习结构等价类, 并通过选择引擎来为无向边定向[173]. Hauser 和 Bühlmann 描述了等价类的图论特性[174], 随后他们还给出寻找有价值选择引擎的两种主动学习策略[175].

目前, 概率图模型的主动学习已吸引许多学者进一步研究, 然而, 其主动学习还存在有待解决的问题: 当前大多数研究都把 DAG 拆分成多个链图成分进行主动学习, 虽然这能简化学习过程, 但必然损失原图的部分信息, 如何就 DAG 整体进行主动学习还需研究; 如何在隐变量存在的情况下解决变量选择操作问题; MN 的主动学习研究算法还很少.

5.4 概率图模型的迁移学习、多任务学习和域自适应学习

迁移学习、多任务学习和域自适应学习是机器学习领域的新技术, 目前已逐渐受到研究学者的关注. 概率图模型的学习与这些新技术的结合显然十分必要, 但同时也带来了新挑战. Pan 和 Yang 给出了迁移学习的明确定义[176]: 已知源域 D_S 及其学习任务 T_S 以及目标域 D_T 及其学习任务 T_T, 其中 $D_S \neq D_T$ 和 $T_S \neq T_T$, 那么迁移学习就是利用 D_S 和 T_S 中的知识来辅助 D_T 中目标预测函数 f_T 的学习过程. 若学习过程还需同时改进 T_S 的性能, 那么这就是多任务学习. 若 T_S 与 T_T 相同, 那么学习过程为域自适应学习.

概率图模型的迁移学习就是当学习任务所需的样本个数不足时, 可以借助相关辅助任务的样本数据来实现原目标任务. Dai 等在朴素贝叶斯网络的参数学习中使用迁移学习, 其基本思想是首先利用有标签数据集估计朴素贝叶斯网络的初始概率分布, 然后利用另一个分布中的无标签数据通过 EM 算法

修改概率分布模型[177]. Roy 和 Kaelbling 则提出朴素贝叶斯分类器的另一种迁移学习方法, 把数据集划分为多个聚类簇, 每个聚类簇在所有学习任务中都具有相同的概率分布. 然后, 他们利用每个聚类簇分别训练得到一个分类器, 并通过 Dirichlet 过程组合得到最终的分类器[178]. Luis 等提出 BN 的归纳迁移学习方法, 该方法考虑了 BN 的结构学习与参数学习[179]. 概率图模型的迁移学习还需要考虑如何同时组合多个相关任务, 如何处理目标任务与相关任务的样本数据都不足的情况.

在 SS 算法中, 结构搜索空间通常根据评分近似等价的结构划分为多个大的凹空间. 而概率图模型的多任务结构学习则是选择搜索空间中相邻的网络作为输出网络集, 而不再输出评分相似的网络集. 这种选择有利于领域专家在数据集中发现独立关系, 最近已有学者对此展开研究[180−181]. 但是多任务结构学习中的任务关联性定义还没有系统的分析, 而且如何从数据中估计任务关联性, 如何有效学习大规模结构任务也是未来的工作.

在实际应用中, 训练数据和测试数据的概率分布一般不相同, 由训练数据学习得到的模型可能在测试数据中并不能获得很好的表现, 而域自适应学习可以解决该问题. 但是, 目前研究人员还没有对概率图模型的域自适应学习开展研究工作, 下一步可以考虑概率图模型的学习与域自适应学习的结合.

6 概率图模型学习的研究趋势

由于概率图模型能够简洁紧凑地表示多变量间的复杂依赖关系, 是不确定性推理体系的重要组成部分, 众多学者对概率图模型的研究热情一直高涨不退. 虽然概率图模型的学习算法已在多个方面取得重要的进展, 但是仍然存在不少问题需要解决. 本节总结概率图模型学习的未来研究热点.

1) 结构学习的 SS 算法中, 多数评分函数都对参数具有敏感性, 或者在有限样本大小的情况下得不到最优解, 这限制了结构学习的精度. 将来的研究应进一步开发性能更好的评分函数, 提高学习的精度.

2) 结构学习过程的计算成本很高, 而实际领域中却通常包含大量变量, 使得结构学习难以实现. 除了前一节介绍的并行学习策略之外, 未来应该从搜索算法上对结构学习过程提速, 优化中间计算过程.

3) 不完备数据是实际应用中经常出现的情况, 但目前能够处理不完备数据的学习算法仍不成熟, 特别是存在隐变量的情况. 如何在不完备数据情况中学习结构还有待进一步的研究.

4) MN 的学习由于需要执行推理任务, 其学习任务比 BN 要复杂很多. 而且 MN 的学习研究也没有 BN 的成熟. 由于马尔科夫逻辑网的提出, 借鉴其学习过程, MN 的学习才有了进一步的发展. 未来需要开发更加快速的权值学习算法; 在其局部结构学习中, 需要确定局部学习的渐进一致性的充分条件; 进一步提高算法速度, 如在算法中引入频繁项集挖掘算法.

5) 除了贝叶斯网络和马尔科夫网络之外, 动态贝叶斯网络、状态观测模型和概率关系模型等也是重要的概率图模型. 但是, 这些概率图模型的学习研究还比较缺乏, 下一步应该对 BN 和 MN 学习算法的新技术进行改良以适应这些概率图模型的学习过程, 并且根据这些概率图模型自身的特点开发出新的学习方法.

7 总结

本文系统综述了 BN 和 MN 的学习算法的研究进展. 其中, 主要以 BN 的学习为主, 详细介绍了参数学习和结构学习的算法. 参数学习是假设结构已知时的简单参数估计问题, 本文分别考虑其在数据完备和不完备时的不同学习情况. 结构学习是学习任务中的主要部分, 其算法可大致分为基于约束的学习算法、基于评分搜索的学习算法、混合学习算法、动态规划结构学习和模型平均结构学习, 本文还介绍了不完备数据的结构学习算法. 与 BN 的学习相比, MN 的研究则还不够成熟, 本文也分别对 MN 的参数和结构学习算法进行简单介绍. 最后, 还指出了概率图模型学习的新挑战以及其研究趋势.

概率图模型是不确定性推理的一种强有力工具, 在机器学习领域势必日益重要. 概率图模型应时刻结合机器学习领域出现的新理论新技术, 不断完善自身体系的理论与技术, 以便更好地在实际领域中应用.

References

1 Koller D, Firedman N. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: The MIT Press, 2009

2 Wainsright M J, Jordan M I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008, **1**(1−2): 1−305

3 Pourret O, Naim P, Marcot B. *Bayesian Networks: A Practical Guide to Applications*. Chichester: John Wiley, 2008

4 Larrañaga P, Moral S. Probabilistic graphical models in artificial intelligence. *Applied Soft Computing*, 2011, **11**(2): 1511−1528

5 Weber P, Medina-Oliva G, Simon C, Iung B. Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, 2012, **25**(4): 671−682

6 Korb K B, Nicholson A E. *Bayesian Artificial Intelligence* (*2nd edition*). Florida: CRC Press, 2010

7 Elvira Consortium. Elvira: an environment for probabilistic graphical models. In: Proceedings of the 1st European Workshop in Probabilistic Graphical Models. Cuenca, Spain, 2002. 222−230

8 Cheng J, Greiner R. Learning Bayesian belief network classifiers: algorithms and system. In: Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence. Ottawa, Canada: Springer, 2002. 141−151

9 Murphy K. The Bayes net toolbox for Matlab. *Computing Science and Statistics*, 2001, **33**(2): 1024−1034

10 Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0. 5: Bayesian Inference Using Gibbs Sampling Manual (version ii), Technical Report, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK. 1996

11 Lauritzen S L. gRaphical models in R. *R News*, 2002, **3**(2): 39

12 Cozman F G. The Javabayes system. *The International Society for Bayesian Analysis Bulletin*, 2001, **7**(4): 16−21

13 Scheines R, Spirtes P, Glymour C, Meek C. *TETRAD II: Tools for Discovery*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994

14 Andersen S K, Olesen K G, Jensen F V, Jensen F. HUGIN-a shell for building Bayesian belief universes for expert systems. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1989. 1080−1085

15 BayesiaLab, Bayesia home page [Online], available: http://www.bayesia.com, August 8, 2013

16 Netica, Norsys Software Corporation home page [Online], available: http://www.norsys.com/netica.html, August 8, 2013

17 Prelee M A, Neuhoff D L, Pappas T N. Image reconstruction from a Manhattan grid via piecewise plane fitting and Gaussian Markov random fields. In: Proceedings of the 19th IEEE International Conference on Image Processing. Orlando, Florida, USA: IEEE, 2012. 2061−2064

18 Dawoud A, Netchaev A. Preserving objects in Markov random fields region growing image segmentation. *Pattern Analysis and Applications*, 2012, **15**(2): 155−161

19 Yousefi S, Kehtarnavaz N, Cao Y, Razlighi Q R. Bilateral Markov mesh random field and its application to image restoration. *Visual Communication and Image Representation*, 2012, **23**(7): 1051−1059

20 Xiong R, Wang J N, Chu J. Face alignment based on 3D face shape model and Markov random field. In: Proceedings of the 12th International Conference on Intelligent Autonomous Systems. Berlin, Heidelberg: Springer, 2012. 249−261

21 Ghosh A, Subudhi B N, Ghosh S. Object detection from videos captured by moving camera by fuzzy edge incorporated Markov random field and local histogram matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, **22**(8): 1127−1135

22 Li S Z. *Markov Random Field Modeling in Image Analysis* (*3rd edition*). Tokyo, Japan: Springer, 2009

23 Blake A, Kohli P, Rother C. *Markov Random Fields for Vision and Image Processing*. Cambridge: The MIT Press, 2011

24 Wei Z, Li H Z. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 2007, **23**(12): 1537−1544

25 Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 2008, **24**(3): 404−411

26 Wei P, Pan W. Network-based genomic discovery: application and comparison of Markov random field models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2010, **59**(1): 105−125

27 Neapolitan R E. *Learning Bayesian Networks*. Upper Saddle River: Pearson Prentice Hall, 2004

28 Geiger D, Heckerman D. A characterization of the Dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 1997, **25**(3): 1344−1369

29 Burge J, Lane T. Shrinkage estimator for Bayesian network parameters. In: Proceedings of the 18th European Conference on Machine Learning. Berlin, Heidelberg: Springer, 2007. 67−78

30 Geiger D, Heckerman D. Learning Gaussian networks. In: Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1994. 235−243

31 Bøttcher S G. Learning Bayesian Networks with Mixed Variables [Ph. D. dissertation], Aalborg University, Denmark, 2004

32 John G H, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1995. 338−345

33 Pérez A, Larrañaga P, Inza I. Bayesian classifiers based on kernel density estimation: flexible classifiers. *International Journal of Approximate Reasoning*, 2009, **50**(2): 341−362

34 McLachlan G, Peel D. *Finite Mixture Models*. New York, USA: John Wiley and Sons, 2000

35 Anandkumar A, Hsu D, Kakade S M. A method of moments for mixture models and hidden Markov models. In: Proceedings of the 25th Annual Conference on Learning Theory. Edinburgh, Scotland, UK: The Journal of Machine Learning Research Workshop and Conference Proceedings, 2012, **23**: 33. 1−33. 34

36 Hsu D, Kakade S M. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In: Proceedings of the 4th Conference on Innovations in Theoretical Computer Science. New York, USA: Association for Computing Machinery, 2013. 11−20

37 Mahjoub M A, Bouzaiene A, Ghanmy N. Tutorial and selected approaches on parameter learning in Bayesian network with incomplete data. In: Proceedings of the 9th International Symposium on Neural Networks. Berlin, Heidelberg: Springer, 2012. 478−488

38 Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984, **6**(6): 721−741

39 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1977, **39**(1): 1−38

40 Elidan G, Friedman N. The information bottleneck EM algorithm. In: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2003. 200−208

41 Elidan G, Ninio M, Friedman N, Schuurmans D. Data perturbation for escaping local maxima in learning. In: Proceedings of the 18th National Conference on Artificial Intelligence. Menlo Park, USA: American Association for Artificial Intelligence, 2002. 132−139

42 Niculescu R S, Mitchell T M, Rao R B. A theoretical framework for learning Bayesian networks with parameter inequality constraints. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2007. 155−160

43 Druzdzel M J, Van Der Gaag L C. Elicitation of probabilities for belief networks: combining qualitative and quantitative information. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1995. 141−148

44 Feelders A, Van Der Gaag L C. Learning Bayesian network parameters with prior knowledge about context-specific qualitative influences. In: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2005. 193−200

45 Liao W H, Ji Q. Learning Bayesian network parameters under incomplete data with domain knowledge. *Pattern Recognition*, 2009, **42**(11): 3046−3056

46 Ramoni M, Sebastiani P. Learning Bayesian networks from incomplete databases. In: Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1997. 401−408

47 Ramoni M, Sebastiani P. The use of exogenous knowledge to learn Bayesian networks from incomplete databases. In: Proceedings of the 2nd International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data. London, UK: Springer-Verlag, 1997. 537−548

48 Ramoni M, Sebastiani P. Robust learning with missing data. *Machine Learning*, **45**(2): 147−170

49 Jaeger M. The AI&M procedure for learning from incomplete data. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2006. 225−232

50 Chickering D M. Learning Bayesian networks is NP-complete. *Learning from Data*, 1996, **112**: 121−130

51 Spirtes P, Glymour C N, Scheines R. *Causation, Prediction, and Search*. Cambridge: The MIT Press, 2000

52 Kalisch M, Bühlmann P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 2007, **8**: 613−636

53 Li J N, Wang Z J. Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *Journal of Machine Learning Research*, 2009, **10**: 475−514

54 Pearl J, Verma T S. A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics*, 1995, **134**: 789−811

55 Cheng J, Greiner R, Kelly J, Bell D, Liu W R. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 2002, **137**(1−2): 43−90

56 Chickering D M, Meek C. On the incompatibility of faithfulness and monotone DAG faithfulness. *Artificial Intelligence*, 2006, **170**(8−9): 653−666

57 Yehezkel R, Lerner B. Bayesian network structure learning by recursive autonomy identification. *Journal of Machine Learning Research*, 2009, **10**: 1527−1570

58 Xie X C, Geng Z. A recursive method for structural learning of directed acyclic graphs. *Journal of Machine Learning Research*, 2008, **9**: 459−483

59 Villanueva E, Maciel C D. Efficient methods for learning Bayesian network super-structures. *Neurocomputing*, 2014, **123**: 3−12

60 de Morais S R, Aussem A. An efficient and scalable algorithm for local Bayesian network structure discovery. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Barcelona, Spain: Springer-Verlag, 2010. 164−179

61 Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 1992, **9**(4): 309−347

62 Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 1995, **20**(3): 197−243

63 Silander T, Myllymäki P. A simple approach for finding the globally optimal Bayesian network structure. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2006. 445−452

64 Steck H. Learning the Bayesian network structure: Dirichlet prior versus data. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Corvallis, USA: AUAI Press, 2008. 511−518

65 Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, **19**(6): 716−723

66 Schwarz G. Estimating the dimension of a model. *Annals of Statistic*, 1978, **6**(2): 461−464

67 Rissanen J. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 1983, **11**(2): 416−431

68 Cruz-Ramírez N, Acosta-Mesa H G, Barrientos-Martínez R E, Nava-Fernández L A. How good are the Bayesian information criterion and the minimum description length principle for model selection? A Bayesian network analysis. In: Proceedings of the 5th Mexican International Conference on Artificial Intelligence. Berlin, Heidelberg: Springer-Verlag, 2006. 494−504

69 Wallace C S, Korb K B, Dai H H. Causal discovery via MML. In: Proceedings of the 13th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1996. 516−524

70 Korb K B, Nicholson A E. *Bayesian Artificial Intelligence (2nd edition)*. Boca Raton, USA: CRC Press, 2010

71 O′Donnell R T, Allison L, Korb K B. Learning hybrid Bayesian networks by MML. In: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence. Berlin: Springer-Verlag, 2006. 192−203

72 Kayaalp M, Cooper G F. A Bayesian network scoring metric that is based on globally uniform parameter priors. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2002. 251−258

73 de Campos L M. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 2006, **7**: 2149−2187

74 Riggelsen C. Learning Bayesian networks: a MAP criterion for joint selection of model structure and parameter. In: Proceedings of the 8th IEEE International Conference on Data Mining. Washington, USA: IEEE, 2008. 522−529

75 Silander T, Roos T, Myllymäki P. Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 2010, **51**(5): 544−557

76 Carvalho A M, Roos T T, Oliveira A L, Myllymäki P. Discriminative learning of Bayesian networks via factorized conditional log-likelihood. *Journal of Machine Learning Research*, 2011, **12**: 2181−2210

77 Bouckaert R R. Probabilistic network construction using the minimum description length principle. In: Proceedings of the 1993 European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. Berlin, Heidelberg: Springer-Verlag, 1993. 41−48

78 Liu F, Zhu Q L. Max-relevance and min-redundancy greedy Bayesian network learning on high dimensional data. In: Proceedings of the 3rd International Conference on Natural Computation. Haikou, China: IEEE, 2007. 217−221

79 Gámez J A, Mateo J L, Puerta J M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 2011, **22**(1−2): 106−148

80 Larranaga P, Kuijpers C, Murga R H, Yurramendi Y. Learning Bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans*, 1996, **26**(4): 487−493

81 Faulkner E. K2GA: heuristically guided evolution of Bayesian network structures from data. In: Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining. Honolulu, HI: IEEE, 2007. 18−25

82 Kabli R, Herrmann F, McCall J. A chain-model genetic algorithm for Bayesian network structure learning. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. New York, USA: ACM, 2007. 1264−1271

83 Regnier-Coudert O, McCall J. An island model genetic algorithm for Bayesian network structure learning. In: Proceedings of the 2012 IEEE World Congress on Computational Intelligence. Brisbane, Australia: IEEE, 2012. 1−8

84 Wong M L, Lam W, Leung K S. Using evolutionary programming and minimum description length principle for data mining of Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, **21**(2): 174−178

85 Wong M L, Lee S Y, Leung K S. A hybrid approach to discover Bayesian networks from databases using evolutionary programming. In: Proceedings of the 2002 IEEE International Conference on Data Mining. Washington, USA: IEEE, 2002. 498−505

86 Wong M L, Leung K S. An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation*, 2004, **8**(4): 378−404

87 Wong M L, Guo Y Y. Learning Bayesian networks from incomplete databases using a novel evolutionary algorithm. *Decision Support Systems*, 2008, **45**(2): 368−383

88 Larrañaga P, Karshenas H, Bielza C, Santana R. A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Information Sciences*, 2013, **233**: 109−125

89 de Campos L M, Huete J F. Approximating causal orderings for Bayesian networks using genetic algorithms and simulated annealing. In: Proceedings of the 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Madrid, Spain: Consejo Superior de Investigaciones Cientificas, 2000. 333−340

90 Heng X C, Qin Z, Tian L, Shao L P. Learning Bayesian network structures with discrete particle swarm optimization algorithm. In: Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence. Honolulu, HI: IEEE, 2007. 47−52

91 Heng X C, Qin Z, Tian L, Shao L P. Research on structure learning of dynamic Bayesian networks by particle swarm optimization. In: Proceedings of the 2007 IEEE Symposium on Artificial Life. Honolulu, HI: IEEE, 2007. 85−91

92 Li X L, Wang S C, He X D. Learning Bayesian networks structures based on memory binary particle swarm optimization. In: Proceedings of the 6th International Conference on Simulated Evolution and Learning. Berlin, Heidelberg: Springer-Verlag, 2006. 568−574

93 Sahin F, Devasia A. Distributed particle swarm optimization for structural Bayesian network learning. *Swarm Intelligence: Focus on Ant and Particle Swarm Optimization*. Vienna, Austria: I-Tech Education and Publishing, 2007, **27**: 505−532

94 Wang T, Yang J. A heuristic method for learning Bayesian networks using discrete particle swarm optimization. *Knowledge and Information Systems*, 2010, **24**(2): 269−281

95 de Campos L M, Fernández-Luna J M, Gámez J A, Puerta J M. Ant colony optimization for learning Bayesian networks. *International Journal of Approximate Reasoning*, 2002, **31**(3): 291−311

96 Daly R, Shen Q. Learning Bayesian network equivalence classes with ant colony optimization. *Journal of Artificial Intelligence Research*, 2009, **35**(1): 391−447

97 Pinto P C, Nagele A, Dejori M, Runkler T A, Sousa J M C. Using a local discovery ant algorithm for Bayesian network structure learning. *IEEE Transactions on Evolutionary Computation*, 2009, **13**(4): 767−779

98 Wu Y H, McCall J, Coles D. Two novel ant colony optimization approaches for Bayesian network structure learning. In: Proceedings of the 2010 IEEE Congress on Evolutionary Computation. Barcelona: IEEE, 2010. 1−7

99 Ji J Z, Wei H K, Liu C N. An artificial bee colony algorithm for learning Bayesian networks. *Soft Computing*, 2013, **17**(6): 983−994

100 Li B H, Liu S Y, Li Z G. Improved algorithm based on mutual information for learning Bayesian network structures in the space of equivalence classes. *Multimedia Tools and Applications*, 2012, **60**(1): 129−137

101 Studený M. *Probabilistic Conditional Independence Structures*. London: Springer-Verlag, 2005

102 Studený M, Vomlel J, Hemmecke R. A geometric view on learning Bayesian network structures. *International Journal of Approximate Reasoning*, 2010, **51**(5): 573−586

103 Hemmecke R, Lindner S, Studený M. Characteristic imsets for learning Bayesian network structure. *International Journal of Approximate Reasoning*, 2012, **53**(9): 1336−1349

104 Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 2003, **50**(1−2): 95−125

105 Teyssier M, Koller D. Ordering-based search: a simple and effective algorithm for learning Bayesian networks. In: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2005. 584−590

106 Singh M, Valtorta M. Construction of Bayesian network structures from data: a brief survey and an efficient algorithm. *International Journal of Approximate Reasoning*, 1995, **12**(2): 111−131

107 Dash D, Druzdzel M J. A hybrid anytime algorithm for the construction of causal models from sparse data. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1999. 142−149

108 de Campos L M, Fernández-Luna J M, Puerta J M. An iterated local search algorithm for learning Bayesian networks with restarts based on conditional independence tests. *International Journal of Intelligent Systems*, 2003, **18**(2): 221−235

109 Friedman N, Nachman I, Peér D. Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1999. 206−215

110 Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 2006, **65**(1): 31−78

111 Tsamardinos I, Aliferis C F, Statnikov A. Time and sample efficient discovery of Markov blankets and direct causal relations. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2003. 673−678

112 Perrier E, Imoto S, Miyano S, Chickering M. Finding optimal Bayesian network given a super-structure. *Journal of Machine Learning Research*, 2008, **9**: 2251−2286

113 Kojima K, Perrier E, Imoto S, Miyano S. Optimal search on clustered structural constraint for learning Bayesian network structure. *Journal of Machine Learning Research*, 2010, **11**: 285−310

114 de Campos C P, Ji Q. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 2011, **12**: 663−689

115 Ott S, Imoto S, Miyano S. Finding optimal models for small gene networks. *Pacific Symposium on Biocomputing*, 2004, **9**: 557−567

116 Ott S, Miyano S. Finding optimal gene networks using biological constraints. *Genome Informatics*, 2003, **14**: 124−133

117 Koivisto M, Sood K. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 2004, **5**: 549−573

118 Koivisto M. Advances in exact Bayesian structure discovery in Bayesian networks. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. Corvallis, USA: AUAI Press, 2006. 241−248

119 Singh A P, Moore A W. Finding Optimal Bayesian Networks by Dynamic Programming, Technical Report CMU-CALD-05-106, School of Computer Science, Carnegie Mellon University, USA, 2005

120 Eaton D, Murphy K. Bayesian structure learning using dynamic programming and MCMC. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2007. 101−108

121 Malone B, Yuan C H, Hansen E A. Memory-efficient dynamic programming for learning optimal Bayesian networks. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press, 2011. 1057−1062

122 Madigan D, York J, Allard D. Bayesian graphical models for discrete data. *International Statistical Review*, 1995, **63**(2): 215−232

123 Madigan D, Andersson S A, Perlman M D, Volinsky C T. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics-Theory and Methods*, 1996, **25**(11): 2493−2519

124 Giudici P, Castelo R. Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, 2003, **50**(1−2): 127−158

125 Grzegorczyk M, Husmeier D. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 2008, **71**(2−3): 265−305

126 Liang F M, Zhang J. Learning Bayesian networks for discrete data. *Computational Statistics and Data Analysis*, 2009, **53**(4): 865−876

127 Tian J, He R, Ram L. Bayesian model averaging using the k-best Bayesian network structures. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Corvallis, USA: AUAI Press, 2010. 589−597

128 Dash D, Cooper G F. Model averaging for prediction with discrete Bayesian networks. *Journal of Machine Learning Research*, 2004, **5**: 1177−1203

129 Kim K J, Cho S B. Evolutionary aggregation and refinement of Bayesian networks. In: Proceedings of the IEEE Congress on Evolutionary Computation. Vancouver, BC: IEEE, 2006. 1513−1520

130 Gou K X, Jun G X, Zhao Z. Learning Bayesian network structure from distributed homogeneous data. In: Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Washington, USA: IEEE Computer Society, 2007. 250−254

131 Liu F, Tian F Z, Zhu Q L. Bayesian network structure ensemble learning. In: Proceedings of the 3rd International Conference on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2007. 454−465

132 Kwoh C K, Gillies D F. Using hidden nodes in Bayesian networks. *Artificial Intelligence*, 1996, **88**(1−2): 1−38

133 Sanscartier M J, Neufeld E. Identifying hidden variables from context-specific independencies. In: Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference. Menlo Park, California, USA: The AAAI Press, 2007. 472−477

134 Geiger D, Heckerman D, Meek C. Asymptotic model selection for directed networks with hidden variables. *Learning in Graphical Models*. Netherlands: Springer, 1998, **89**: 461−477

135 Ramoni M, Sebastiani P. Learning Bayesian networks from incomplete databases. In: Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1997. 401−408

136 Parviainen P, Koivisto M. Ancestor relations in the presence of unobserved variables. In: Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer-Verlag, 2011. 581−596

137 Friedman N. The Bayesian structural EM algorithm. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1998. 129−138

138 Beal M J, Ghahramani Z. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In: Proceedings of the 7th Valencia International Meeting on Bayesian Statistics. Oxford: Oxford University Press, 2003. 453−464

139 Watanabe K, Shiga M, Watanabe S. Upper bound for variational free energy of Bayesian networks. *Machine Learning*, 2009, **75**(2): 199−215

140 Elidan G. Bagged structure learning of Bayesian networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL: JMLR Workshop and Conference, 2011. 251−259

141 Wainwright M J, Jaakkola T S, Willsky A S. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In: Proceedings of the 9th Workshop on Artificial Intelligence and Statistics. Key West, Florida: Society for Artificial Intelligence and Statistics, 2003. 97−105

142 Sutton C, Minka T. Local Training and Belief Propagation, Technical Report MSR-TR-2006-121, Microsoft Research, 2006

143 Wainwright M J. Estimating the "wrong" graphical model: benefits in the computation-limited setting. *Journal of Machine Learning Research*, 2006, **7**: 1829−1859

144 Ganapathi V, Vickrey D, Duchi J, Koller D. Constrained approximate maximum entropy learning. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Corvallis, USA: AUAI Press, 2008. 196−203

145 Murray I, Ghahramani Z. Bayesian learning in undirected graphical models: approximate MCMC algorithms. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2004. 392−399

146 Murray I, Ghahramani Z, Mackay D. MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. Arlington, USA: AUAI Press, 2006. 359−366

147 Besag J. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 1977, **64**(3): 616−618

148 McCallum A, Pal C, Druck G, Wang X. Multi-conditional learning: generative/discriminative training for clustering and classification. In: Proceedings of the 21st National Conference on Artificial Intelligence. Boston, USA: AAAI Press, 2006. 433−439

149 Hinton G E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002, **14**(8): 1771−1800

150 LeCun Y, Chopra S, Hadsell R, Marc′Aurelio R, Huang F J. A tutorial on energy-based learning. *Predicting Structured Data*. Cambridge, MA: MIT Press, 2006. 191−241

151 Taskar B, Guestrin C, Koller D. Max-margin Markov networks. In: Proceedings of the 17th Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2003. 24−32

152 Bromberg F, Margaritis D, Honavar V. Efficient Markov network structure discovery using independence tests. *Journal of Artificial Intelligence Research*, 2009, **35**(1): 449−484

153 Della P S, Della P V, Lafferty J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(4): 380−393

154 Kok S, Domingos P. Learning the structure of Markov logic networks. In: Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM, 2005. 441−448

155 Domingos P. Unifying instance-based and rule-based induction. *Machine Learning*, 1996, **24**(2): 141−168

156 Mihalkova L, Mooney R J. Bottom-up learning of Markov logic network structure. In: Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM, 2007. 625−632

157 Davis J, Domingos P. Bottom-up learning of Markov network structure. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress, 2010. 271−280

158 Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional ising model selection using L_1-regularized logistic regression. *Annals of Statistics*, 2010, **38**(3): 1287−1319

159 Höfling H, Tibshirani R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *Journal of Machine Learning Research*, 2009, **10**: 883−906

160 Pekins S, Lacker K, Theiler J. Grafting: fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 2003, **3**: 1333−1356

161 Lee S I, Ganapathi V, Koller D. Efficient structure learning of Markov networks using L_1-regularization. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2006. 817−824

162 Yang E, Ravikumar P, Allen G I, Liu Z D. Graphical models via generalized linear models. In: Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates, 2012. 1367−1375

163 Lowd D, Davis J. Learning Markov network structure with decision trees. In: Proceedings of the 10th IEEE International Conference on Data Mining. Washington, USA: IEEE, 2010. 334−343

164 Haaren J V, Davis J. Markov network structure learning: a randomized feature generation approach. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada: AAAI Press, 2012. 1148−1154

165 Yu K, Wang H, Wu X D. A parallel algorithm for learning Bayesian networks. In: Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2007. 1055−1063

166 Lam W, Bacchus F. Using new data to refine a Bayesian network. In: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1994. 383−390

167 Friedman N, Goldszmidt M. Sequential update of Bayesian network structure. In: Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1997. 165−174

168 Nielsen S H, Nielsen T D. Adapting Bayes network structures to non-stationary domains. *International Journal of Approximate Reasoning*, 2008, **49**(2): 379−397

169 Castillo G, Gama J. Adaptive Bayesian network classifiers. *Intelligent Data Analysis*, 2009, **13**(1): 39−59

170 Yasin A, Leray P. iMMPC: a local search approach for incremental Bayesian network structure learning. In: Proceedings of the 10th International Conference on Advances in Intelligent Data Analysis X. Berlin: Springer-Verlag, 2011. 401−412

171 Cooper G F, Yoo C. Causal discovery from a mixture of experimental and observational data. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1999. 116−125

172 Tong S, Koller D. Active learning for structure in Bayesian networks. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2001. 863−869

173 He Y B, Geng Z. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 2008, **9**: 2523−2547

174 Hauser A, Bühlmann P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 2012, **13**(1): 2409−2464

175 Hauser A, Bühlmann P. Two optimal strategies for active learning of causal models from interventions. In: Proceedings of the 6th European Workshop on Probabilistic Graphical Models. Granada, Spain, 2012. 123−130

176 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

177 Dai W Y, Xue G R, Yang Q, Yu Y. Transferring naive bayes classifiers for text classification. In: Proceedings of the 22nd AAAI conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2007. 540−545

178 Roy D M, Kaelbling L P. Efficient Bayesian task-level transfer learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2007. 2599−2604

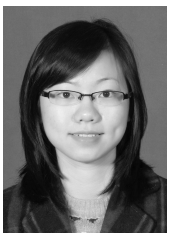
179 Luis R, Sucar L E, Morales E F. Inductive transfer for learning Bayesian networks. *Machine Learning*, 2010, **79**(1−2): 227−255

180 Honorio J, Samaras D. Multi-task learning of Gaussian graphical models. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress, 2010. 447−454

181 Oyen D, Lane T. Leveraging domain knowledge in multitask Bayesian network structure learning. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada: AAAI Press, 2012. 1091−1097

刘建伟 博士, 中国石油大学 (北京) 副研究员. 主要研究方向为智能信息处理, 机器学习, 复杂系统分析, 预测与控制, 算法分析与设计. 本文通信作者. E-mail: liujw@cup.edu.cn

(**LIU Jian-Wei** Ph. D., associate professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing. His research interest covers intelligent information processing, machine learning, analysis, prediction, controlling of complicated nonlinear system, and analysis of algorithm and designing. Corresponding author of this paper.)



黎海恩 中国石油大学 (北京) 地球物理与信息工程学院硕士研究生. 主要研究方向为机器学习, 概率图模型表示、学习和推理. E-mail: lihaien1988@163.com

(**LI Hai-En** Master student in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing. Her research interest covers representation, learning and reasoning of probabilistic graphical models.)

罗雄麟 博士, 中国石油大学 (北京) 教授. 主要研究方向为智能控制和复杂系统分析, 预测与控制. E-mail: luoxl@cup.edu.cn

(**LUO Xiong-Lin** Ph. D., professor in the Department of Automation, College of Geophysics and Information Engineering, China University of Petroleum, Beijing. His research interest covers intelligent control, and analysis, prediction, controlling of complicated nonlinear system.)