原

贝叶斯网络结构学习方法简介

2017年12月14日 16:56:46 jbb0523 阅读数：4814更多
个人分类：机器学习(Machine Learning)

https://blog.csdn.net/jbb0523/article/details/78804386
题目：贝叶斯网络结构学习方法简介

贝叶斯网络（Bayesiannetwork, BN）结构学习就是从给定的数据集中学出贝叶斯网络结构，即各节点之间的依赖关系；只有确定了结构才能继续学得网络参数，即表示各节点之间依赖强弱的条件概率。对于普通人来说（非贝叶斯网络的专业研究人员，仅一般使用者），希望的是能够有那么一个函数，函数的输入是数据集，输出即为贝叶斯网络结构。目前确实有很多贝叶斯网络工具箱，但新人上手还是有不小的门坎，原因有二：一是工具箱文件众多，很多时候就根本找不到你想要的函数；二是即使找到了还要输入一堆参数，而这些参数该如何设置又感觉无从下手，这是因为基础理论懂的太少，还有就是只要函数出错就会一脸茫然，为啥啊？

贝叶斯网络结构学习方法简介贝叶斯网络结构学习方法简介综上所述，尽管我们普通人并不需要从头编写贝叶斯网络的结构学习函数，但是了解一些有关贝叶斯网络结构学习的基础理论也还是需要的，哪怕仅仅是为了更好地使用工具箱现成的函数。以下主要综合了多篇学位论文的有关内容，若要了解更多可以拜读参考文献原文。

1、基于评分搜索的方法

那么如何根据已有的数据集学得贝叶斯网络结构呢？一种最简单的想法就是遍历所有可能的结构，然后用某个标准去衡量各个结构，进而找出最好的结构。

是的，这就是评分搜索的基本思想。你可以把所有可能的结构看为定义域，将衡量特定结构好坏的标准看为函数，寻找最好的结构的过程相当于在定义域上求函数的最优值，即这是一个最优化问题。但这里面有两个关键点：一是定义域一般几乎无穷大，不可能遍历，即确定合适的搜索策略；二是用什么样的衡量标准，即确定所谓的评分函数。

1.1、评分函数

【朱明敏. 贝叶斯网络结构学习与推理研究[D]. 西安电子科技大学, 2013.】

最常用的评分函数主要是基于贝叶斯统计的评分函数和基于信息理论的评分函数. 例如 K2 评分（又称 CH 评分）[37]，BD 评分[76]，MDL（Minimum Description Length）评分[77]，BIC（Bayesian Information Criterion）评分[78]，AIC（Akaike Information Criterion）评分[79]等等.

评分函数主要分为两类，一类是贝叶斯评分函数，另一类是基于信息论的评分函数。

1.1.1、贝叶斯评分函数

【胡春玲. 贝叶斯网络结构学习及其应用研究[D]. 合肥工业大学, 2011.】

贝叶斯评分的核心思想是：结合关于网络拓扑结构的先验知识，选择具有最大后验概率(Maximum A Posterior, 简称 MAP)的网络结构。假设网络拓扑结构 G 的先验概率为 $P(G)$，针对给定样本集 D，根据贝叶斯公式，可以得到网络结构 G 的后验概率为：

$$P(G|D)=\frac{P(G)P(D|G)}{P(D)} \qquad \text{（式 3.8）}$$

从公式（3.8）可以看出，因为 $P(D)$与网络拓扑结构无关，使 $P(G)P(D|G)$ 取得最大值的网络结构 G 就是具有最大后验概率的网络结构，因此，定义 $\log P(G,D)=\log(P(G)P(D|G))=\log P(G)+\log P(D|G)$ 为网络结构的贝叶斯评分，即为 MAP 测度。

上图中，所谓的拓扑结构G即为朝思暮想的贝叶斯网络结构，样本集D即为已有的数据集。贝叶斯评分函数主要包括K2评分、BD评分、BDe评分。几种评分函数的关系如下：

【刘峰. 贝叶斯网络结构学习算法研究[D].北京邮电大学, 2007.】

1992 年 Cooper 和 Herskovits 提出第一个贝叶斯评分函数，叫做 K2 评分函数[25,26]；1995 年 Heckerman 提出 BD 评分(Bayesian Dirichlet Score)函数[27,28]，该函数作为 K2 函数的泛化；同时，Heckerman 依据附加的似然等价假设，提出一个特例，叫做 BDe 评分函数[27,28]。

【胡春玲. 贝叶斯网络结构学习及其应用研究[D]. 合肥工业大学, 2011.】

假设网络参数 θ 的先验概率服从公式(3.5)所描述的 Dirichlet 分布，则有：

$$P(D|G)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \quad \text{(式 3.9)}$$

$$\log P(D|G)=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\right)+\sum_{k=1}^{r_i}\log\left(\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}\right)\right) \quad \text{(式 3.10)}$$

其中：$\alpha ij=\sum_k \alpha_{ijk}, N_{ij}=\sum_{k=1}^{r_i} N_{ijk}$。在公式(3.10)定义的 $\log P(G,D)$ 为 Cooper 和 Herskovits 给出的 CH 评分[95-96]。（又称为K2评分）

进一步假设网络结构的先验概率服从均匀分布，此时根据贝叶斯评分定义，按贝叶斯评分选择网络结构等同于按CH评分选择网络结构，这一假设条件下CH评分又被称为 BDe(Bayesian Dirichlet Equivalent)评分[97]。

【朱明敏. 贝叶斯网络结构学习与推理研究[D]. 西安电子科技大学, 2013.】

服从均匀分布，那么可以得到如下的K2 评分：

$$F_{K2}(G|D)=\log P(G)+\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log((r_i-1)!/(m_{ij\cdot}+r_i-1)!)+\sum_{k=1}^{r_i}\log(m_{ijk}!)\right] \quad (2\text{-}10)$$

若 D 和 G 满足上面的假设，并且 $p(\theta_G|G)$ 服从如下的 Dirichlet 分布：

$$p(\theta_G|G)\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1} \quad (2\text{-}11)$$

则得到相应的BD（Bayesian Dirichlet）评分：

$$F_{BD}(G|D)=\log(P(G))+\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log(\Gamma(\alpha_{ij\cdot})/\Gamma(\alpha_{ij\cdot}+m_{ij\cdot}))+\sum_{k=1}^{r_i}\log(\Gamma(\alpha_{ijk}+m_{ijk})/\Gamma(\alpha_{ijk}))\right] \quad (2\text{-}12)$$

α_{ijk} 表示狄利克雷分布中的超参数取值，$\alpha_{ij\cdot}=\sum_{k=1}^{r_i}\alpha_{ijk}$；$m_{ijk}$ 表示变量 X_i 取第 k 个值，同时 $pa(X_i)$ 取第 j 个值的样本数目，$m_{ij\cdot}=\sum_{k=1}^{r_i}m_{ijk}$. 容易看出，当所有的超参数值 $\alpha_{ijk}=1$ 时，BD 评分将退化为 K2 评分，因此，K2 评分是 BD 评分的特殊形式.

当公式（2-12）中的超参数 $\alpha_{ijk}=\alpha\times P(X_i=k,pa(X_i)=j|G)$ 时，其中 $P(\cdot|G)$ 表示网络 G 的先验分布，α 表示先验样本等价量. 取 $P(X_i=k,pa(X_i)=j|G)=1/(r_iq_i)$，即结构先验信息服从均匀分布，则称相应的评分为BDeu 评分，如下所示：

$$F_{BDeu}(G \mid D) = \log(P(G)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log(\Gamma(\frac{\alpha}{q_i}) \Big/ \Gamma(m_{ij\bullet} + \frac{\alpha}{q_i})) + \sum_{k=1}^{r_i}\log(\Gamma(m_{ijk} + \frac{\alpha}{r_i q_i}) \Big/ \Gamma(\frac{\alpha}{r_i q_i}))\right]. \quad (2\text{-}13)$$

以上仅简单列出了三篇学位论文中有关几种贝叶斯评分函数的描述，若需了解某具体评分函数，请根据原始文献仔细研究学习。

1.1.2、基于信息论的评分函数

【朱明敏. 贝叶斯网络结构学习与推理研究[D]. 西安电子科技大学, 2013.】

因此，根据公式（2-14）和（2-15）得到相应的MDL 评分函数（求函数最大值）：

$$F_{MDL}(G \mid D) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m_{ijk}\log(\frac{m_{ijk}}{m_{ij\bullet}}) - \frac{1}{2}\log m\sum_{i=1}^{n}(r_i - 1)q_i. \quad (2\text{-}16)$$

MDL 评分函数不依赖于先验概率；对给定的充分大的独立样本而言，具有最大 MDL 分值的网络可以任意接近于抽样分布；当实例数据 D 服从多项分布时，MDL 评分函数等于 BIC 评分函数 因此对公式（2-16）做进一步的简化，得到AIC 评分函数

$$F_{AIC}(G \mid D) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m_{ijk}\log(\frac{m_{ijk}}{m_{ij\bullet}}) - \sum_{i=1}^{n}(r_i - 1)q_i. \quad (2\text{-}17)$$

更多有关MDL准则、BIC评分函数、AIC评分函数的内容参见【周志华.机器学习[M]. 清华大学出版社，2016.】的7.5.2节（学习）。

1. 2、搜索策略

【胡春玲. 贝叶斯网络结构学习及其应用研究[D]. 合肥工业大学, 2011.】

在定义了评分函数的情况下，贝叶斯网络的学习问题就变成了一个搜索问题，通过搜索算法寻找具有最佳评分的网络结构，这是一个 NP-Hard 问题[4]，因此，通常采用启发式的搜索方法，常用的搜索算法有 K2 算法，爬山法（贪婪法）、模拟退火算法、演化算法以及抽样算法。

有关K2算法实际采样的就是爬山法搜索，后面会单独谈；爬山法(Hill-Climbing, HC)、模拟退火(Simulated Annealing, SA)、演化算法(Evolutionary Algorithm,EA)均为一些通用的方法，尤其是EA

是一大类算法，包括遗传算法(Genetic Algorithm, GA)、蚁群算法(Ant Colony Optimization,ACO)等等，抽样算法是一种特殊的搜索策略详见本文第4节，后面也会单独谈。

下表列出了一些基于评分搜索的贝叶斯网络结构学习算法：

【朱明敏. 贝叶斯网络结构学习与推理研究[D]. 西安电子科技大学, 2013.】

表 2.1 基于评分搜索的 BN 结构学习算法

年份	算法	搜索空间	评分函数	搜索策略
1968	Chow-Liu[81]	Tree	Entropy	-
1992	K2[37]	DAG	K2	Hill Climbing
1994	Lam-Bacchus[77]	DAG	MDL	Heuristic Search
1995	GHC[82]	DAG	BD	Hill Climbing
1996	GA-O[43]	Ordering	K2	Genetic Algorithm
1996	GA-B[42]	DAG	K2	Genetic Algorithm
1996	Suzuki[83]	DAG	MDL	Branch and Bound
2002	GES[30]	PDAG	BDeu	Greedy Search
2002	ACO-B[45]	DAG	K2	Ant Colony Optimization
2003	EDA-B[47]	DAG	Entropy	Estimation of Distribution Algorithm
2004	HEA[49]	DAG	MDL	Evolutionary Algorithm
2009	ACO-E[46]	PDAG	BIC	Ant Colony Optimization

2、基于约束（依赖分析或条件独立性测试）的方法

【李玮玮. 贝叶斯网络结构学习方法的研究[D]. 南京航空航天大学, 2008.】

基于约束的贝叶斯网络结构学习方法（也称为依赖分析的方法或基于条件独立性测试的方法），通常利用统计或信息论的方法定量地分析变量间的依赖关系以获取最优地表达这些关系的网络结构。这类方法的核心思想是：首先对训练数据集进行统计测试，尤其是条件独立性测试，确定出变量之间的条件独立性；然后，利用变量之间的条件独立性，构造一个有向无环图，以尽可能多地涵盖这些条件独立性。典型算法包括 SGS[22]算法，PC[23]算法，TPDA[10]算法等。

条件独立性测试是在给定条件集合下检测两个变量条件独立关系的一种典型度量。在贝叶斯网络结构学习中，条件独立性测试的基础是信息论中信息流的度量，常采用互信息和条件互信息进行变量之间条件独立性测试。

1993 年，Spines 等提出的 SGS 算法是典型的以条件独立性测试确定拓扑结构的算法。该算法从无向完全图出发，不需要节点次序，可以的利用条件独立性测试自动地确定边的方向，但是该算法需要指数次的条件独立性测试。

2000 年，Spirtes 等人对 SGS 算法进行增强，提出了 PC 算法。该算法当从隐含了稀疏模型的数据集中建立贝叶斯网络时很有效，用样本容量为 10000 的 ALARM 数据集，事先没有确定节点次序，能学习到比较完整的网络结构，只有 3 条边丢失和 2 条边多余。

2002 年，Cheng 将信息论与统计测试相结合，提出了 TPDA 算法。该算法通过计算互信息来确定节点间的条件独立性，从而构造多连接有向图模型。TPDA 算法也称之为三阶段算法，学习算法可分为三个阶段：第一阶段通过计算任意两个节点之间的互信息来测量节点间的相关程度，当互信息大于某个阈值时说明这两个节点间有边存在，并以此来构造一个初始网络；第二阶段通过计算条件互信息来决定任意两节点是否条件独立，如果不是则添加相应的边；第三阶段仍利用条件独立性测试检查当前网络中的每条边，如果删掉该边后，该边的两个节点满足条件独立性，则删掉该条边，否则保留该边。TPDA 算法可以处理事先知道节点次序和不知道节点次序两种情况，时间复杂度分别是 $o(n^2)$和 $o(n^4)$。

一般情况下如数据规模较大，该类算法的复杂度将是指数级的。当潜在的贝叶斯网络结构比较复杂时，整个算法的时间复杂度将是无法忍受的。由于时间复杂度的限制，基于条件独立性测试的方法更加适用于稀疏图，因为在稀疏图当中需要进行的条件独立测试的数量很少，但是在稠密图中则不然，在稠密图中需要进行的条件独立测试将达到指数级数量。然而在现实应用中，大多数数据集上潜在的贝叶斯网络是稀疏的有向无环图，所以这种方法在实际问题中可以被大量的应用。

3、基于评分搜索和基于约束相结合的混合方法

【李玮玮. 贝叶斯网络结构学习方法的研究[D]. 南京航空航天大学, 2008.】

基于评分搜索和基于约束相结合的混合方法，结合了两者长处，是目前的一个研究热点。

1993 年 Singh 等人提出的 CB 算法[32]是第一个混合学习算法，CB 算法先采用 PC[23]算法确定节点的次序，再用 K2 算法[14]进行结构学习。

1995 年，Spirtes 等人提出的 GBRS 算法[33]是很有效的混合学习算法之一，它先用 PC 算法得到初始网络，再采用贪婪搜索法从该初始网络出发在等价贝叶斯网络结构空间中搜索最优网络结构。

2006 年，Tsamardinos 等人在 Sparse Candidate[24] 算法的基础上提出了 MMHC[16]算法，算法第一阶段用 Max-Min Parents and Children（MMPC）[35]算法找到贝叶斯网络的骨架，第二阶段用评分搜索法对骨架进行定向。实验表明，无论在大数据集还是在小数据集上，无论是从时间复杂度还是从构建出的网络质量上看，MMHC 算法都比许多其他算法优越。但当变量很多时，MMHC 算法第二阶段中对边定向效率不高。MMHC 的提出者 Tsamardinos 等人提供了软件包 Causal_Explorer 下载，见 http://www.dsl-lab.org/supplements/mmhc_paper/mmhc_index.html，该软件包可以直接调用 MMPC、MMHC 等算法。

MMHC是Max-Min Hill-Climbing首字母的简写。

4、基于随机抽样的方法

随机抽样的代表是马尔可夫蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法，MCMC的重要代表是Metropolis-Hastings(简称MH)。更多有关MCMC和MH的内容可以参见【周志华. 机器学习[M]. 清华大学出版社，2016.】的14.5.1节（MCMC采样），以及【BishopC M. Pattern Recognition and Machine Learning (Information Science andStatistics)[M]. Springer-Verlag New York, Inc. 2006.】的11.2节（MarkovChain Monte Carlo）。

【胡春玲. 贝叶斯网络结构学习及其应用研究[D]. 合肥工业大学, 2011.】

MCMC 方法[98-99]是是源于统计物理学和生物学的一类重要的随机抽样方法，并被广泛应用于机器学习、统计和决策分析等领域的高维问题的推理求积运算。MHS 抽样算法[100]作为 MCMC 方法中常用的抽样方法之一，该方法通过构建一条马尔可夫链，模拟一个收敛于 Boltzmann 分布的系统。MHS 抽样算法抽样过程收敛之后的样本为来自于平稳 Boltzmann 分布的抽样，因而能够较好地保证了样本的多样性，所得样本可以直接用来对平稳分布进行矩估计，避免了高维积分的计算,该算法被评为 20 世纪对科学和工程领域产生重大影响的十大算法之一。

Madigan 等人将 MHS 抽样算法首次引入到贝叶斯网络的结构学习[101]，该算法采用局部的弧增加、删除和反向的均匀分布作为抽样过程的建议分布，并利用抽样过程收敛之后产生的来自目标平稳分布的网络结构样本来估计贝叶斯网络的结构特征，因此，MHS 抽样算法具有良好的学习精度，但 MHS 算法抽样过程的融合性差，收敛速度慢。MHS 抽样算法的不同的改进算法[45,102]都着力于改善 MHS 抽样算法的收敛速度，其中具有代表性的是由 Laskey 和 Myers 提出的 PopMCMC 算法[62]。总之， MHS 抽样算法能够较好地解决进化学习方法中由于个体趋同而产生的早熟问题，保证算法的学习精度，但该类算法目前仍存在收敛速度慢和收敛性判断困难等问题仍未能得到有效解决。

假设 MHS 抽样算法所面向问题系统温度为 T， 采用 $E(G)$表示问题领域状态空间 Θ 上任一状态 G 的能量，MHS 抽样算法通过构建一条收敛于平稳分布 $P(G)=\frac{1}{Z}\exp(\frac{-E(G)}{T}),(\forall G\in\Theta$,其中Z为正则化因子）的马尔可夫链，来模拟一个收敛于 Boltzmann 分布的系统。

MHS 抽样算法的抽样过程如下：

1. 假设当前状态为 $G^{(c)}$，则根据建议分布 $R(G^{(n)}\mid G^{(c)})$,生成下一个状态 $G^{(n)}$。
2. 按概率规则，决定所生成的新状态是否被接受。新状态的接受概率为：

$$A(G^{(n)}\mid G^{(c)})=\min\left[1,\frac{p(G^{(n)})R(G^{(c)}|G^{(n)})}{p(G^{(c)})R(G^{(n)}|G^{(c)})}\right] \qquad (式\ 3.18)$$

相应的转移概率为：

$$T(G^{(n)}\mid G^{(c)})=\begin{cases} R(G^{(n)}\mid G^{(c)})A(G^{(n)}\mid G^{(c)}) & G^{(n)}\neq G^{(c)} \\ 1-\sum_{G^{(n')}\neq G^{(c)}} R(G^{(n')}\mid G^{(c)})A(G^{(n')}\mid G^{(c)}) & G^{(n)}=G^{(c)} \end{cases} \qquad (式\ 3.19)$$

3. 若新状态被接受，则取代原有状态；若新状态被拒绝，则保留原有状态。

通过对随机抽样算法的分析发现：初始值和建议分布是影响MHS算法收敛速度的重要因素，若建议分布等于目标平稳分布，此时抽样成为来自于目标平稳分布的独立抽样[6]，因而具有较快的收敛速度，针对实际问题，因为其目标平稳分布未知，只能通过设计接近于目标平稳分布的初始值和建议分布来提高收敛速度。

5、总结

本篇主要是罗列一些贝叶斯网络结构学习策略，每种策略均包括了多种具体的方法（或算法），若需要深入了解某个算法时可以查找原文献，再根据原文献的参考文献找到原始资料，因此本篇相当于一个速查表或者是备忘录，不要期望能从本篇当中得到太具体的信息。接下来会结合一些贝叶斯网络的工具箱函数具体讨论几种算法的实现以供不同需求使用。.

后记

最近基于贝叶斯网络做一些事情，需要基于数据集学习贝叶斯网络的结构，然而搜索下载了一些贝叶斯网络工具箱后发现很难上手，而且网络上并没有多少入门的资料，因此就把自己学习的过程或者说是入门经验分享出来，希望能够降低贝叶斯网络结构学习的门坎。本系列文章仅为个人入门阶段所写，若有理解不当之处敬请谅解，也希望大家协手丰富这方面的入门技术资料^_^。