Bayesian 网转化为神经元网

薛万欣[1], 董冠宇[2], 刘大有[3]
(1. 北京联合大学管理学院, 北京 100101; 2. 北京后勤指挥学院, 北京 100010; 3. 吉林大学计算机科学技术学院, 吉林长春 130021)

摘 要: Bayesian 网目前广泛应用于专家系统中, 用于处理大量以条件概率为形式的数据. 本文借用神经元网络结构, 根据专家给定的相关模型和部分观察集使用后向传播对条件概率进行估计, 并在训练中, 保持 Bayesian 网特性不变, 应用 Occam 修剪法则, 在化简过程中提炼其中的规律. 实践表明, 对于复杂的问题, 由化简的因果模型得出的神经元网络更有效.

关键词: Bayesian 网; SIGMA2PI 网络; 后向传播网
中图分类号: TP181 文献标识码: A 文章编号: 0372-2112 (2004) 02-0250-04

Compiling Bayesian Networks into Neural Networks

XUE Wan-xin[1], DONG Guan-yu[2], LIU Da-you[3]
(1. Beijing union university management engineering, Beijing 100101, China;
2. Beijing Logistics Conduction College, Beijing 100010, China;
3. Department of Computer Science, JI LIN University, Changchun, Jilin 130021, China)

Abstract: The criticism on the usage of Bayesian networks in expert systems is centered around the claim that the use of probability requires a massive amount of data in the form of conditional probabilities. This paper shows that with given information easily obtained from experts, the dependence model probabilities can be estimated using backpropagation, such that during training the Bayesian characteristic of the network is preserved. Applying the Occam's razor principle results in defining a partial order among neural network structures. Experiments show that for the Multiplexer problem, the network compiled from the more succinct causal model is better than the one compiled from the less succinct model.

Key words: bayesian networks; SIGMA2PI networks; propagation networks

1 引言

Bayesian 网采用图形方式表示不确定性知识, 有关它的研究在国外十分广泛, 理论和实践中的许多问题都可以通过 Bayesian 网建模实现. Bayesian 网可以通过常识或专家知识进行构造. 本文不是对条件概率进行专家估计, 而是在给定的相关模型以及部分观察集的基础上对条件概率进行估计和修正, 进而达到学习参数分布的目的.

2 相关工作介绍

随着修正逻辑理论的发展, 一些实际问题成功解决. 但实际生活中很多领域的问题运用逻辑必然性进行推理不合适, 需用概率推理才能实现.

Towel 和 Shavlik[1] 提出了 KBANN 系统, 它将满足 Horn 约定的命题知识库通过增加新的变量和连接转化成神经元网络, 并采用后向传播进行训练进而正确分类训练数据, 但得出的权重没有条件概率的含义. Mahoney 和 Mooney 提出了基于 MYCIN 模型法则进行修正的 Rapture 方法, 它运用后向传播进行确定性因子的传播和修正. 应用 Bayesian 网可以进行概率分布学习, 不过, 此时依赖关系的运用以及概念的归纳并没有引起人们的注目. Spiegelhalter 等人讨论了条件概率, 它用计数获得的说明频率的参数表示条件概率, 本文不同于简单的求频率的方法, 因为含有隐含命题, 即概率没有给出或无法测得. Cooper[2] 等人研究了隐含变量, 但没有讨论所得分布的归纳能力. Bayesian 网可以看作是通过独立关系组织在一起的概率分类或概率回归函数的集合. 产生概率输出的回归模型包括线性回归、广义线性回归、概率神经网络、概率决策树[3]、核密度方法以及字典方法[4] 等. 原则上, 这些方法都可以用于 Bayesian 网的概率学习.

本文对因果模型进行化简, 并在网络结构中进行偏序定义得出神经元网络, 然后对给定的神经元网络进行评估. 算法中输入的是相关模型, 相关模型是一些条件概率的估计值和

收稿日期: 2002-12-30; 修回日期: 2003-10-20
基金项目: 国家 863 项目(No. 863-306-ZD05-01-2); 国家自然科学基金(No. 69883003); 教育部高校博士点专项科研基金项目和教育部符号计算与知识工程重点实验室项目

以成对的输入/输出概率形式给出的观察集, 算法的任务是学习得出与给定观察集最佳拟合的条件概率分布. 算法中有输入、输出和隐含命题, 输入那些给定的先验概率, 输出基于输入先验概率所得到的概率.

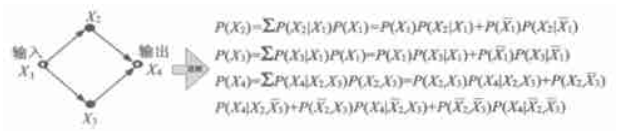
Bayesian 网以有向非循环图 DAG 表示相关模型, DAG 的根节点映射为输入节点, 结果节点映射为输出节点, 其他节点映射为隐藏节点. 给定的 Bayesian 网转换为标准的高位 SIG2MA2PI 网络[5] 从而可用标准的(或稍加修改) 后向传播方法来进行修正. 权重等同于条件概率, 学习权重就等同于学习条件概率, 从而直接修正 Bayesian 网. 后向传播方法可以使一些结构在全局中收敛到最简.

通常用网络容量来衡量简洁程度, 本文讨论的神经元网络是来自专家或其他方法的因果模型, Glymour 建议利用数据间的相互关系来发现因果结构; 最近的一项研究建议用条件独立关系来替代[6]; 本文建议通过衡量所得到的神经网络的归纳能力来衡量所建议的因果模型的好坏.

3　Bayesian 网和 SIGMA2PI 网

3l1　Bayesian 网

Bayesian 网是一个带有概率注释的有向非循环图, 图中节点与知识领域的随机变量一一对应; 每个节点有一个条件概率表, 定量描述其所有父亲节点对于该节点的作用效果. 图中的有向弧表示变量间的因果关系[6]. 假设 $x_1, , , x_n$ 表示命题集合, $p(x_i)$ 是 x_i = true 的概率, 在用 DAG 表示的 Bayesian 网中, 把每个节点看成一个命题, 如果 x_j 条件依赖 x_i, 则 i y j 存在弧, 并用 $p(x_j) | p(x_i)$ X $p(x_j)$ 表示. 如果一个节点有 m 个父亲, 则定义了 2^m 种传播. 以有 4 个命题的图 1 为例, 如果 $p(x_1)$ 给定, 那么 $p(x_2), p(x_3), p(x_4)$ 的条件概率为:



$$P(X_2)=\sum P(X_2|X_1)P(X_1)=P(X_1)P(X_2|X_1)+P(\bar{X}_1)P(X_2|\bar{X}_1)$$
$$P(X_3)=\sum P(X_3|X_1)P(X_1)=P(X_1)P(X_3|X_1)+P(\bar{X}_1)P(X_3|\bar{X}_1)$$
$$P(X_4)=\sum P(X_4|X_2,X_3)P(X_2,X_3)=P(X_2,X_3)P(X_4|X_2,X_3)+P(X_2\bar{X}_3)$$
$$P(X_4|\bar{X}_2X_3)+P(\bar{X}_2X_3)P(X_4|\bar{X}_2X_3)+P(\bar{X}_2\bar{X}_3)P(X_4|\bar{X}_2\bar{X}_3)$$

图 1　简单 DAG 及其分布

3l2　SIGMA2PI 网

多元模型中非线性判别函数结果表明使用单层高阶单元(HOUs) 比使用多层单元好处多, 而且在后向传播网络中表现也不错, (HOUs) 的一个特例就是 SIGMA2PI 单元. 其中权重之和表示为: $net_j = \sum w_{ij} \prod_k o_{ik}$, i 随着 j 的变化而变化. 而 k 代表变化的联合项. 以图 2 为例:

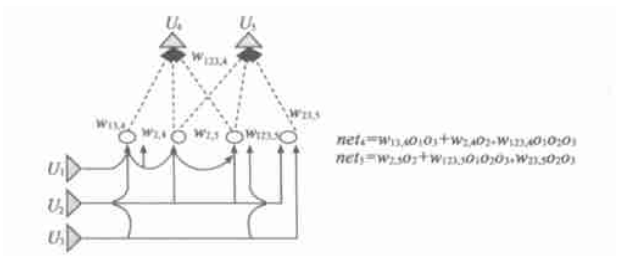


图 2　SIGMA2PI 网

4　Bayesian 后向传播网

对原有 Bayesian 网化简, 将 Bayesian 网映射成特殊的 SIG2MA2PI 网, 即后向传播网(BPN). 条件概率可以归纳为: $p(X_j) = \sum_{i=1}^{m} o_j^i w_{ij}$ 其中: o_j^i S $p(X_{i1}, , , X_{im})$, w_{ij} S $p(X_j | X_{i1}, , , X_{im})$.

这里定义两种单元: E 单元(图 2 中的三角形) 和 F 单元(图 2 中的圆形), E 单元 o_j 的输出对应概率 $p(X_j)$, F 单元 u_j^i 的输出对应联合概率 $p(X_{i1}, , , X_{im})$, 每个 E 单元可以是一个输出单元, 也可以代入 F 单元. 而且每个 E 单元既可以是输入单元也可以由 F 单元代入. 同样每个 F 单元可以代入和被代入 E 单元.

定义 1　一个给定 Bayesian 网的 SIGMA2PI BPN 是这样的:

Ó 对于 Bayesian 网 DAG 图中的每个节点 X_j, 在 BPN 中将创建两个 E 单元 u_j 和 u_j, 其中 E 单元的输出值 o_j, o_j 代表条件概率 $p(X_j) = o_j$, $p(X_j) = o_j$.

Ó 所有 BPN 的输入和输出单元都是 E 单元, DAG 的数据源映射为输入单元, DAG 的收点映射为输出单元.

Ó E 单元 u_j 和 u_j 由 2^m 个父亲 F 单元 $u_j^1, , , u_j^{2^m}$ 代入, 其中 m 是 Bayesian 网 DAG 图中 x_j 节点的父亲数目.

Ó F 单元的值 o_j^i 是 $p(X_{i1}, , , X_{im})$, 其中 $X_{i1}, , , X_{im}$ 是的父亲, 有 2^m 个不同的真/假组合.

Ó F 单元 u_j^i 来自 E 单元, E 单元是计算 $p(X_{i1}, , , X_{im})$ 的因子, 其中 X_{i1} 是 DAG 中 X_j 的父亲. 如果 $X_{i1}, , , X_{im}$ 是条件独立的, 那么 $p(X_{i1}, , , X_{im}) = o_j^i = \prod_{i=1}^{m} o_{i1} = \prod_{i=1}^{m} p(X_{i1})$, 否则(依赖), 需要沿着路径按链规则遍历, 定义更复杂.

Ó 权重 w_{ij} 就是条件概率 $p(X_j | X_{i1}, , , X_{im})$.

Ó 前向传播的值: E 单元: $o_j = \sum_{i=1}^{2^m} w_{ij} o_j^i$, E 单元: $o_j^i = \prod_{k1\ parents(u_j^i)} o_k$

Ó F 单元和权值必须满足: P j　1= $\sum_{i=1}^{2^m} o_j^i$, 1= $w_{ij} + w_{ij}$

注意: BPN 是把 X 和 X 定义成两种独立的单元来进行多值概念的一般表示的.

引理 1　在给定先验概率作为输入命题(根) 之后, BPN 的前向命题会计算出正确的概率值作为输出命题(收点).

根据修正后的 BPN 可以重建 Bayesian 网, 因为所需的条件概率值已经由权重得到了.

5　后向传播学习法则

这里描述的 BPN 学习规则与标准的后向传播算法不同, 加上了约束 $D_k = - D_k$, 从而使训练的每一步都遵从由 BPN 到 Bayesian 网的映射规则, 如果把 BPN 看作一个中间过程的转换形式, 事实上整个训练过程就是直接对 Bayesian 网进行修正.

定义 2　把 u_j 设成 E 单元, P_j 是代入 u_j 的 F 单元的集

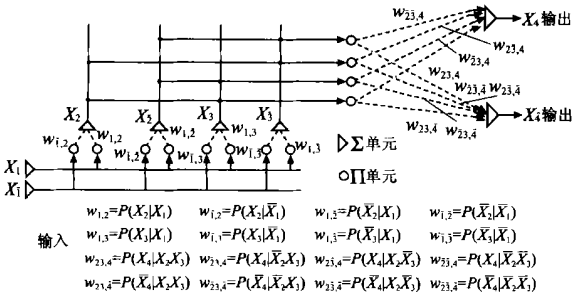


图 3 图 1 中 DAG 转换成的 BPN

合, 把 u_j^i 设成 F 单元, 其中 i∈ P_j, 把 P_j^i 设成代入 u_j^i 的 E2 单元的集合, 每个 E2 单元 u_j 和 E2 单元 u_k 的误差映射如下

$$P_i\in P_j, P_k\in P_{cj}, D_k z D_k+ D_j w_{ij}\frac{o_j^i}{o_k}$$

要保持 Bayesian 网的 $\sum p(X_{i1},\ ,\ ,X_{im})=1$ 特性, 就要保证 $P_j\sum_{i=1}^{2^m} o_j^i=1$, 这用到了 D_k, D_{ok} 的平均值 D_o. $D_{ok}=-D_{ok}=\frac{1}{2}(D_k-D_k)$, $\Delta w_{ij}=\eta D_{oj}o_j^i$, 因为保持了原始标记, 收敛性不受影响.

引理 2 给定 $P_j\ 1=\sum_{i=1}^{2^m} o_j^i$, $o_j=1-o_j$, 如果 P_k $D_{ok}=-D_{ok}$ 那么 $P_j\ 1=\sum_{i=1}^{2^m} o_j^{ci}$ 其中 $o_j^{ci}=\prod_{k\in R_i}(o_k+D_{ok})$

证明: 通过证明递归削减下面的联合项: $A=\sum_{i=1}^{2^m} o_j^i=\sum_{i=1}^{2^m}\prod_{k\in P_j^i}(o_k+D_{ok})$

由于上式因子中的 o_k 和 o_k 是不可能同时出现的(互余), 上式又可以写成: $A=(o_p+D_{op})\#Ac+(o_p+D_{op})\#Ac=(o_p+o_p^-+D_{op}+D_{op})\#Ac=Ac$, 其中 $Ac=\sum_{i=1}^{2^m}\prod_{k\in P_j^{i-(p,p)}}(o_k+D_{ok})$, 递归结束时 $Ac=1$.

同样, 如果 $p(X_j|X_{i1},\ X_{im})+p(X_j|X_{i1},\ X_{im})=1$, 那么 $w_{ij}+w_{ij}=1$, 故有:

引理 3 如果按 $P_j\ 1=\sum_{i=1}^{2^m} o_j^i$, $o_j=1-o_j$ 和 $w_{ij}+w_{ij}=1$ 初始化 BPN, 那么 $\Delta w_{ij}=\eta D_{oj}o_j^i$ 就可以保证所有训练过程中的初始化约束都是满足的.

推论 1 通过 $\Delta w_{ij}=\eta D_{oj}o_j^i$ 修改 BPN 中的权重, 就能实现直接对原有 Bayesian 网条件概率的修正.

6 分布的表示

BPN 可以表示条件概率的分布, 对 BPN 结构进行偏序定义, 同时定义出表征力最好的结构, 进而通过次表征力结构作后向传播, 能够进行最佳拟合给定数据的分布的学习.

完全双分支 DAG 图: 将 DAG 图中的节点划分成两个区域 X, Y, 所有的边都由 X 指向 Y, 所有 X 中点都有指向所有 Y 中点的边.

引理 4 由完全双分支 DAG 图转换而来的 BPN 能够表示所有的基于输入命题(源)的输出命题(收点)的条件概率的分布.

学习的任务是找到网络内部最佳拟合输入输出的分布.

定理 1 在由双分支 DAG(未必是完全的)转化来的 BPN 中进行后向传播, 可以实现最佳拟合给定观察的条件概率的学习.

Pearl 证明模型越简单, 约束性越强, 数据偏离的可能性越小, 更易纠错, 越可靠. 对结构有优先法则: 在两个 BPN 中, 取自更简洁因果模型的优先.

模型的简洁程度由表征能力衡量, 而不是描述语法, 但一般来说简洁结构的语法也是简明的. 以命题集合 $x_1,\ ,\ ,x_n$ 为例, 在 Markov 中标定一个有 m 个父亲的节点需要 2^m-1 个参数, 给定输入 n_i 和输出 n_0 命题的集合, 如果 DAG 是完全双分支, 最多需要 $(n_0(2^{n_i}-1)$ 个)参数, 表示也最完全, 其他任何包含相同信息的稀疏因果模型都是它的简化形式.

神经元网络需要尽量减小容量来提高可计算性, 多采用融合相关权重减少局部连通的方式, 缩减因果模型中的连通就会使其更简洁. BPN 的权重数和对应的因果模型的参数数密切相关, 可以用计算 BPN 中权重的简化方法来实现参数数量的缩减. 图 4 的环路, 相应的 DAG 图为图 4(a), 如果 x_1, x_2 未知, 则如图 4(b), 应用 Occam 修剪定理, 含有隐含命题的图 4(a)更简洁, 转化来的 BPN 的权重为 32 个, 图 4(b)权重则为 64 个.

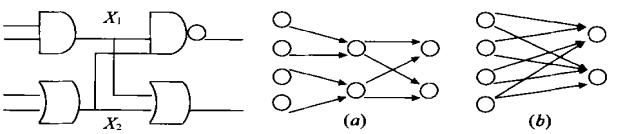


图 4 有隐含命题的实例, 对应两个可能的因果模型 (a)含隐含命题; (b)完全双分支 DAG

给定输入和输出的命题集合, 在网络结构中引入偏序定义, 使完全双分支 DAG 图成为极小, 相应的所有非完全双分支 DAG 都引入隐含命题使结构更简洁. 从给定的初始结构进行神经元网络的学习, 沿着尽量简洁的结构爬升, 可以得到好的效果. 初始的神经网络结构可以由诱导因果算法计算得到, 这里采用在多项式时间选择最大似然因果模型进行近似.

7 归纳的经验评估

在依赖关系完全已知和条件概率不完全已知的情况下, 都可以归纳出简洁的 BPN 网. 在具有归纳能力的经验评估中, 破坏的网络可以由实例归纳和修正, 破坏的结构用随机且分布一致的概率代替权重; 以低方差正态分布产生的权重替换破坏的所有权重, 使得分布的平均值还是以前的权重.

以标准的多态问题做实验: 给定一组输入输出的集合, 测定各逻辑部分在学习中的效能, 图 5DAG 图表示: 有 6 个输入命题 $x_1,\ ,\ ,x_6$, 一个输出命题 x_{11}和 4 个隐含命题 $x_7,\ ,\ ,x_{10}$, 表示的关系定义为$\{x_1,\ ,\ ,x_6, x_{11}|\ p(x_{11}|\ x_1,\ ,\ ,x_6)>\ p(x_{11}|\ x_1,\ ,\ ,x_6)\}$, 即如果 x_{11}的计算概率大于 0.5, 则输出是 1.

在这个实验中, 概念的所有知识点都是已知的, 这样, 评

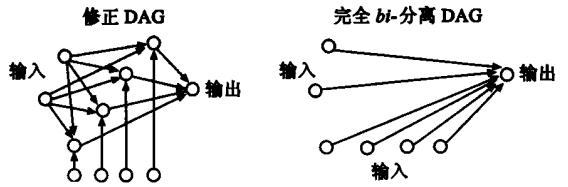


图 5　复杂的 DAG

估所学习得到的网络的最好的办法是对整个概念衡量准确度, 很明显, 训练集合的补图就是最大的可能测试集合.

实验过程: 首先进行正确 BPN 的计算, 里面有 64 个实例精确表示概念, 并受到一些干扰权重的破坏, 在子集实例中训练得出结果网络, 对 64 个知识点测试精确度, 图 6 中的每一点都是 18 段训练的平均值, 最短训练都通过 3000 次计算. 其中 G= 0.0005, 这样小的 G 可以保证在出现陡降时高阶多项式不会出错.

用不同的对称破坏噪声不断地进行训练就可以得到满意的极值点. 从网络独立集合的 6 个极值点选择 3 个最佳进行平均, 这个极值点的集合不能与训练集合相交, 并且 40% 保持了原样. 由完全双分支 DAG 转换来的网络常由三种不同算法得到同一个极值点: (1) 从正确 DAG 转换 BPN. (2) 从完全双分支 DAG 转换 BPN. (3) 学习决策树的 ID3 符号算法. 从正确 DAG 转换的 BPN 比 ID3 稍好一些, 但从完全双分支 DAG 转换的 BPN 在小训练集情况下归纳的更好一些.

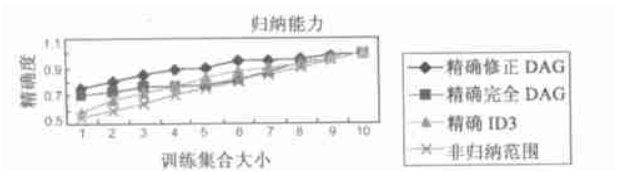


图 6　在 Q= 0.06 的破坏以及不同大小的训练集合上训练完成分布的归纳, 以 100% 概念测试精度, 每一点表示 18 个 BPN 训练以及 50 个 ID3 过程.

8　结论

Bayesian 网是概率分布的联合表示, 本文介绍一种模型, 它利用比较容易从专家获得的依赖模型和部分观察, 通过后向传播来学习条件概率, 从而在训练过程中保持 Bayesian 网特性不变. 这个模型要求神经元网络结构要始终反映因果模型的数据. 应用著名的 Occam 修剪理论得到了网络结构中的偏序特性, 这将应用到因果结构的发现和好的神经网络结构的学习中.

用 DAG 表示的 Bayesian 网, 能够转化为 SIGMA2PI 网的一种, 即 Bayesian 后向传播网, 为了确保在学习的每一个环节都能体现出由 BPN 到 Bayesian 网的映射, 需要在标准的后向传播过程中加入约束 $D_k = -D_l$, 这样做是因为可以将训练过程看成是直接针对 Bayesian 网的. 从一个完全双分支 DAG 转换来的 BPN 能够表示出输入端的所有概率分布的情况. 如果依赖模型是一个双分支 DAG, 就寻找最佳拟合给定观察的概率分布. 实验表明, 在多态问题中, 越简洁的神经网络, 作用越好.

Bayesian 网需要主观概率支持. 神经元网络能在最小均方差的情况下实现对样本输入输出对的逼近, 它的学习能力强、容易理解, 一般不需要主观知识的支持. 本文方法的缺陷是所得到的神经元网络的大小会随着依赖关系集合的数据增大而指数增长. 在解决类似 DNA 序列催化编码等问题时, 常会导致 10^6 以上的数量, 这显然不具备可操作性, 以后我们会利用本文中的方法解决一些实际问题, 来证明它的效用.

参考文献:

[1] Towel G G, Shavlik J W, Noordewier M o. Refinement of approximate domain theories by knowledge2based eural networks[A]. In Proc. of the Eighth National Conference on Artificial Intelligence (AAAI290)[C]. Menlo Park, CA: Portland or, AAAI Press, 1990. 861- 866.

[2] Cooper G F, Herskovits E. A bayesian method for the induction of prob2abilistic networks from data[J]. Machine Learning, 1992, 9: 309- 347.

[3] Roizen L, Pearl J. Learning link2probabilities in causal trees[A]. Proc of. the 2nd AAAI Workshop On Uncertainty In AI [C]. Menlo Park, CA: Portland or, AAAI Press, 1986.

[4] Friedman N, Goldszmidt M. Building classfiers using bayesian networks [A]. Proceedings AAAI296 Thirteenth National Conference On Artif2cial Intelligence [C]. Menlo Park, CA: Portland or, AAAI Press, 1996. 1277- 1284.

[5] Eddie schwalb. Compiling Bayesian networks into neural networks [J]. In AAAI, 2000, 8: 293- 297.

[6] Wang Shuang, Lin Shimin, Lu Yuchang. Learningstructures [J]. Sc2ence(in Chinese), 2000, 27(10): 77- 79.

作者简介:

薛万欣　女, 1966 年 5 月生于辽宁, 博士, 主要研究方向: 不确定性推理, 人工智能, 电子商务等.

董冠宇　男, 1967 年 3 月生于河北, 硕士, 主要研究方向: 计算机网络, 非线性编辑等.

刘大有　男, 1941 年生于河北, 教授, 博士导师, 主要研究方向: 人工智能, 不确定性推理, 神经网络与遗传算法等.