基于TAN结构的启发式贝叶斯网络结构学习算法

程泽凯

(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:贝叶斯网络结构学习是个NP难题。一种有效且准确性较高的学习算法是K2算法。但K2算法要确定结点次序,在无先验信息时受到很大限制。提出了一种启发式结构学习G算法,该算法以学习树扩展朴素贝叶斯TAN结构作为启发式信息,由该启发式信息生成结点次序,再用K2算法生成贝叶斯网络结构。实验结果表明,G算法可以解决无先验信息时确定结点次序的问题。所添加的弧比较简洁,网络结构比TAN结构更加合理。

关键词:贝叶斯网络;树扩展朴素贝叶斯结构;结构学习;启发式

中图分类号:TP181;TP39 **文献标识符**:A **文章编号**:1673-629X(2007)08-0061-03

BN Structure Learning Heuristic Algorithm Based on TAN Structure

CHENG Ze-kai

(School of Computer Science, Anhui University of Technology, Maanshan 243002, China)

Abstract: The structure learning for Bayesian netwoks is NP-hard problem, K2 is one of efficacious and accurate algorithms. K2 confirms the order of nodes firstly. To a certain extent this limits in non-information. This paper purposes a new heuristic Bayesian networks structure learning G algorithm. This algorithm uses TAN structure which learns as heuristic information, using K2 algorithm learning Bayesian netwoks structure. The experimental result shows that G algorithm can solve nodes order in non-.information. Arcs is sententious, comparing TAN structure, it's more reasonable.

Key words: Bayesian networks; TAN structure; structure learning; heuristic

0 引 言

贝叶斯网络是一个带有概率注释的有向无环图,可表示为 $G=(S,P)$,其中 S 是网络的拓扑结构,P 是局部概率分布。贝叶斯网络的学习可以分成两步:(1) 网络结构 S 的学习;(2) 每个变量的局部条件概率分布 P 的参数学习。

常见的用于分类模型的贝叶斯网络结构有 Duda[1]提出的朴素贝叶斯结构 NB(Naive Bayes)和 Friedman[2]提出的树扩展朴素贝叶斯结构 TAN(Tree Augmented Naive Bayes)。它们均是贝叶斯网络结构的近似,在实际应用中取得了一定的成功。TAN结构无论结点间是否相关,$n-1$ 个属性结点必须要加入 $n-2$ 条扩展弧,仍然有不合理之处。贝叶斯网络结构考虑各个属性之间的相关性,结构更合理,但是完全的贝叶斯网络结构学习是NP难题。因此,构建简洁而且合理的结构及高效的学习算法是解决问题的关键。

收稿日期:2006-10-12

基金项目:安徽省教育厅自然科学项目(2006KJ061B)

作者简介:程泽凯(1975-),男,安徽马鞍山人,硕士,讲师,研究方向为人工智能、数据挖掘、机器学习。

Cooper等[3]提出了学习贝叶斯网络结构的K2算法,核心是使用评分函数进行模型选择。为了降低计算的复杂度,K2算法要求事先确定结点变量的次序。文献结论表明该算法是有效和准确的。但是K2算法没有解决如何确定结点变量的次序的问题。寻求最优的变量次序本身就是一个NP难题。文中提出了一种基于启发式的学习贝叶斯网络结构的G算法。G算法以学习TAN树形结构作为启发式信息,由该启发式信息生成结点次序,再用K2算法生成贝叶斯网络结构。实验结果表明,G算法所添加的弧比较简洁,网络结构比TAN结构更加合理,有更大的适应性。

1 K2贝叶斯网络结构学习算法

假设数据集 D 包含 m 个实例,基于如下4个假设:(1) 变量是离散的且可观察;(2) 各个事件的发生独立;(3) 数据完整,无缺失数据和空值;(4) 所有候选网络先验等价。文中约定:$A_1,A_2,\cdots,A_n$ 是属性变量,C 是类变量,a_i 是属性 A_i 的取值,实例 $x_i=(a_1,a_2,\cdots,a_n)$。A_i 有 r_i 个状态,其双亲集合 $pa(A_i)$ 有 q_i 个状态,$q_i=\prod_{Ai\in pai} r_i$,$N_{ijk}$ 是 A_i 的第 k 个状态,且 $pa(A_i)$ 的

第 j 个状态的记录数；N_{ij} 是 $pa(A_i)$ 的第 j 个状态的记录数，即 $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$；$\alpha_{ijk}$ 和 α_{ij} 是先验值，有 $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$。用贝叶斯方法求解贝叶斯网络结构的一般方法是：选择后验概率 $p(B_s \mid D)$ 最大的候选网络结构 B_s 作为学习结果。由贝叶斯公式可得，$p(B_s \mid D) \propto p(D, B_s) \propto p(B_s)\ p(D \mid B_s)$，$p(D)$ 是与结构无关的常量，所有候选网络结构 B_s 的先验概率 $p(B_s)$ 假定相等。故只需要计算 $p(D \mid B_s)$ 的数值。

Cooper 等[3] 证明：在无约束多项分布、参数独立、采用 Dirichlet 先验和数据完整的前提下，参数向量可以独立地更新，数据的边界似然正好等于每一个 $i-j$ 结点对的边界似然乘积。基于上述假设他们给出如下公式：$p(D \mid B_s) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} [\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}]$，Cooper 等建议均一先验 $\alpha_{ijk} = 1$。

网络结构模型空间的规模随网络结点个数 n 的增加而呈超指数增长（见表 1）。其近似递推公式是：$G(0) = 1, G(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} G(n-i)$，仅对具有 10 个结点的小型网络来说，则要确定约 420 亿亿个状态值。即使在确定网络结点次序（每个结点的父结点只能从排列在该结点之前的结点中选择），搜索空间大大减少的情况下，n 个结点所包含的网络数目是 $G'(n) = \prod_{k=1}^{n-1} \sum_{i=0}^{k} \binom{k}{i} = 2^{\frac{n(n-1)}{2}}$，10 个结点的小型网络的搜索空间也有 35 万亿之巨！用穷举法是无法进行模型选择的。

表 1 结点数目相应的网络状态数目

n	1	2	3	4	5	6	7	8	9	10
完全网络状态数	1	3	25	543	29281	3781503	$1.1*10^9$	$7.8*10^{11}$	$1.2*10^{15}$	$4.2*10^{18}$
确定结点次序后	1	2	8	64	1024	32768	$2.1*10^6$	$2.7*10^8$	$6.9*10^{10}$	$3.5*10^{13}$

K2 算法用贪婪搜索处理模型选择问题：先定义一种评价网络结构优劣的评分函数，再从一个空网络开始，根据事先确定的最大父结点数目和结点次序，选择分值最高的结点作为该结点的父结点。依法逐步为每一个变量添加最佳数目的父结点。由文献[3]，算法的时间复杂度为 $O(nmr2^n)$。K2 算法使用后验概率作为评分函数：

$p(D \mid B_s) = \prod_{i=1}^{n} \text{score}(i, pa_i)$，其中 $\text{score}(i, pa_i) = \prod_{j=1}^{q_i} [\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}]$

在文中限制最大父结点数目为 2，此时，算法的时间复杂度是 $O(n^2 mr)$。

K2 算法的伪码如下[4]：

输入参数：训练数据集 D，n 个结点及其状态数矩阵 ns，结点的次序集合 order( )，最大父结点数目 max_ fa，pred( ) 是次序排在结点 i 前的变量集合；

输出结果：dag 表，即贝叶斯分类器结构。

```
for i=1:n
  pai=[];
  S_old=score(i,pa_i); %计算当前状态的分值 S_old
  done=1;%结束标志
  %当结束标志或结点 i 的父结点数小于 max_fa 时结束循环
  while (done & num(pa_i)<max_fa)
    z=search(pred(i),pa_i);%在 pred(i)-pa_i 中搜索 z,使 score(i,pa_i∪{z})的值最大
    S_new=score(i,pa_i∪{z});%计算加入父结点后的分值 S_new
    if (S_new>S_old) %若 S_new大于 S_old
      S_old=S_new; %则保留新值和父结点 pa_i 的值
      pa_i=pa_i∪{z};
    else done=0;
    end if
  end while
  dag(pa_i,i)=1;%对于结点 i 生成的父结点 pa_i,记录到矩阵 dag 中
end for
  return(dag);
```

K2 算法每一迭循环中，添加一条弧的条件是最大化函数 score( )。用 K2 算法学习 ALARM 数据集，得到除 1 条缺失边和 1 条额外边的完整结构[3]，表明该算法的有效性和准确性。

2 基于 TAN 结构的启发式 G 算法及其性能分析

K2 算法要求先确定结点变量的次序，对先验知识的依赖性很大。在不了解相关的领域知识或没有专家指导的情况下，确定变量的次序相当困难。如果穷举所有变量的次序，其状态数是 $n!$，可见寻求最优的变量次序本身就是一个 NP 难问题。采用启发式方法确定结点次序，是避免盲目搜索、提高算法效率的有效方法。TAN 结构与数据集的拟合较好，可以作为贝叶斯网络分类器结构学习的启发式信息。

文中提出的基于启发式的贝叶斯网络结构学习 G 算法先从训练集数据学习得到 TAN 结构，以此为启发式信息确定变量次序，再用 K2 算法学习生成贝叶斯网络结构。其实质就是以 TAN 结构为基础，从次优结构导出更优结构。TAN 结构学习算法中包含：基于互信息测度的 MI 算法和 CMI 算法。相应地，基于 TAN

结构启发式信息的 G 算法这里分别称为 G－CMI 算法、G－MI 算法。

G 算法避免了 K2 算法由于缺乏先验知识无法确定,或只能盲目确定结点次序的不足。在 TAN 结构的基础上学习得到的贝叶斯网络结构更合理。

G 算法的时间开销包括两部分:次优结构的确立和 K2 算法的执行时间。由前文可知,K2 算法 的时间复杂度是 $O(n^2mr)$,学习 TAN 结构的时间复杂度是 $O(n^2m)$,m 是实例个数,n 是属性个数,r 是属性最大状态数[2]。由此可知 G 算法的时间复杂度是 $O(n^2mr)+O(n^2m)$,属于多项式时间复杂度算法。G 算法的时间开销仅比 K2 算法多出建构 TAN 结构的时间。在作者开发的贝叶斯分类器实验平台 MBNC (Bayesian Networks Classifier using Matlab)[3]实验中,用从 UCI(University of California in Irvine)[5]下载的标准数据集建构 TAN 结构,时间一般都不超过 5 秒。

3　实验设计与结果

基于 TAN 结构的启发式贝叶斯结构学习算法的测试在 MBNC 实验平台上完成。选择了从 UCI 下载的 Corral 数据集进行实验验证。先进行数据预处理,如识别各种格式的数据集,转换为统一的格式,以及打乱数据集记录的次序,缺失数据的例子被删除或补齐,连续属性的值进行离散化处理,明显对分类的作用微小的属性被忽略等等。

Corral 数据集是一个 7 个属性(A,B,C,D,E,F,Class)的人工数据集,其中,Class 是类结点,A 跟 B 相关,C 跟 D 相关,E 与 Class 不相关,F 与 Class 相关,干扰较大。表 2 列出了不同算法学习 Corral 数据集生成结构 BD 测度的分值。

表 2　G 算法和原算法生成结构的分值比较

分类算法	TANC－MI	TANC－CMI	G－MI	G－CMI
结构分值	－520.603728	－487.637297	－473.868599	－473.293974

用 G 算法得到结构的分值较大,表明所生成的结构跟数据拟合的更好。用 G－MI 算法和 G－CMI 算法都能较准确地学习得到“关键”特征,所添加的弧较少,网络结构比较合理,避免了大的存储空间和计算条件概率的复杂度。

G 算法取得了较好的实验效果,有如下特点:

(1)结合树形结构和网形结构,按“需”添加扩展弧,所添加的弧数目比较少,比 TAN 结构更加简洁,更加合理。

(2)G 算法不需要事先确定结点次序,无须先验知识。

与一般的分类相比,文本分类面临着文本数据的属性个数很多和属性之间存在某些依赖关系的情况。用 TAN 建构文本分类模型,添加的扩展弧较多,结构模型比较复杂,计算复杂度也较高。与此相比,文中提出的 G 启发式算法建构分类模型有很好的适应性。

总之,采用 G 算法所生成的结构较之 TAN 结构更加合理,在某些领域(如文本分类)内有更大的适应性。

4　结论与展望

提出了启发式 BNC 结构学习 G 算法,通过在 MBNC 实验平台上验证表明算法的有效性,进一步的研究如下:

(1)由 TAN 结构确定结点次序,还不是最优的结果,采用更好的启发式信息,减少节点次序的搜索空间是今后的研究方向。

(2)网络结构中最大父结点数限制为 2,如最大父结点数取为 3 或更高,虽然计算复杂度会大幅提高,但能学习得到更加合理的网络结构。

参考文献:

[1] Duda R, Hart P. Pattern Classification and Scene Analysis [M]. New York: John Wiley and Sons, 1973.

[2] Friedman N. Bayesian network classifiers[J]. Machine Learning, 1997(29): 131－163.

[3] Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9: 309－347.

[4] 程泽凯,林士敏,陆玉昌. 基于 Matlab 的贝叶斯分类器平台 MBNC[J]. 复旦学报,2004,43(5):729－732.

[5] Blake C, Keogh E, Merz C. UCI repository of machine learning database[EB/OL]. 1998. http://www.ics.uci.edu/mlearn/MLRepository.html.

(上接第 60 页)

[2] 罗　菲,何明一. 基于免疫进化算法的多层前向神经网络设计[J]. 计算机应用, 2005,25(7):1661－1663.

[3] 吴清佳. 遗传神经网络的智能天气预报系统[J]. 计算机工程,2005,31(14):176－189.

[4] 林　雄. 自适应模糊神经网络研究[J]. 微计算机信息, 2006,19(3):16－17.

[5] 史永胜,宋云雪. 基于进化算法和 BP 神经网络的故障诊断模型[J]. 计算机工程,2004,30(14):125－127.