维普资讯 http://www.cqvip.com

基于边界的Markov网的发现*)

Learning Markov Network Based on the Boundary

何盈捷　刘惟一

(云南大学计算机科学系　昆明650091)

Abstract Markov network is an another powerful tool besides Bayesian network which can be used to do uncertain inference. A method of learning Markov network automaticly from mass data based on boundary has been discussed in this paper. Taking advantage of an important conclusion in information theory, we present an efficient boundary based Markov network learning algorithm. This algorithm only demands $O(N^2)$ times CI (conditional independence) test. We prove if the joint probability is strictly positive, then the found Markov network must be the minimal I_map of the sample.

Keywords Markov network, Boundary, CI test, I_map

一、引言

Bayesian网是一种进行不确定性推理的有力工具[1],被广泛应用于人工智能、专家系统以及数据挖掘等领域。近年来,越来越多的研究者开始研究如何从大量的样本数据中发现Bayesian网络,提出了许多Bayesian网的学习算法。这些算法大致可以分为两类:基于搜索和打分的算法与基于依赖分析的算法。

基于搜索和打分的算法的基本思想是根据评分函数搜索得到对样本数据拟合得最好的Bayesian网络。评分函数主要对待选的网络结构进行打分,选择与数据拟合得最好的网络结构。由于不可能对所有的网络结构进行测试,所以需要运用局部搜索算法进行网络结构的搜索,通常是从初始网络结构(可以是空结构、随机指定的结构或先验网络结构等)开始,通过增加、删除或转向操作使得局部最优化(根据评分函数),再逐渐扩展到整个网络最优化。常用的局部搜索算法是爬山法和模拟退火法。

依据评分函数,基于搜索和打分的算法又可以分为:基于统计学Bayesian的算法、最小描述长度(MDL)算法和最大熵的算法。基于统计学Bayesian的算法采用对数似然评分标准;基于MDL的算法采用最小描述长度评分函数(最小描述长度包括网络结构本身的描述长度和使用网络结构描述数据的长度)。基于最大熵的算法的评分函数标准是使得信息增长得最快。Herckerman等[4]的Bayesia网发现算法是一种基于传统统计学的Bayesian方法。Xiang和Wong(1994)[2]的Markov网发现算法是一种最大熵的算法,Suzuki[3] Bayesian网发现算法是一种MDL的算法。

基于搜索和打分的算法存在如下问题:1)局部搜索算法基于评分函数的可分解性,通过局部测度保证了局部最优化,但并不能保证全局最优化,搜索的结果往往是局部最优的;2)基于搜索和打分的算法根据评分函数搜索得到对样本数据拟合得最好的网络,往往所得的网络对样本数据过度拟合,并不适应于所有的情况。MDL算法相对较好地解决了这个问题;3)算法往往缺乏效率,影响效率的主要因素在于评分函数和搜索算法。

与基于搜索和打分的算法相比,基于依赖分析的算法并不是通过打分来选择网络,而是通过分析样本中蕴涵的依赖关系(条件独立关系)来构造网络。Cheng等[5]的Bayesian网发现算法是一种基于依赖分析的算法。

基于依赖分析的算法存在如下问题:1)由于Bayesian网是一个有向无环图,多数算法需要一个结点序,不同的结点序所得到的Bayesian网不同;2)多数算法缺乏效率,并具有指数级的计算复杂度。

Markov网是类似于Bayesian网的另一种进行不确定性推理的有力工具[1]。Morkov网是一个无向图,发现Markov网不需要发现边的方向,因此要比发现Bayesian网容易得多。本文我们提出了一个基于边界

*)国家自然科学基金资助项目(项目编号69763003)。何盈捷　硕士生,主要研究方向为数据与知识工程。刘惟一　教授,主要研究方向为数据与知识工程。

维普资讯 http://www.cqvip.com

的方法发现 Markov 网，所谓边界 S 就是指与属性 a 有直接关系的属性集，给定 S，属性 a 不受其它属性 ($U-S-a$，设 U 为属性全集)所影响，即给定 S，属性 a 条件独立于其它属性 $U-a-S$，S 是满足这一条件的最小集。在无向图中 a 的边界 S 表现为 a 的直接近邻。对每个属性 a，我们均求其边界，然后将 a 与它的边界相连，即可得到反映样本数据的 Markov 网。为了衡量发现的 Markov 网的好坏，我们引入了 I-图的概念，我们希望得到的 Markov 网是尽可能多地反映样本中蕴涵的条件独立关系的无向图，即最小 I-图。测试样本中的条件独立，我们利用信息论中验证信息独立的一个重要结论，从而大大提高了效率。由此，我们提出了一个有效的基于边界的 Markov 网发现算法，该算法是一种基于依赖分析的算法，仅需 $O(n^2)$次 CI(条件独立)测试。经证明，假如全局概率函数严格为正，该算法发现的 Markov 网一定是最能反映样本数据的最小 I-图。和已有算法相比，该算法具有以下优点：

1)整个算法基于依赖分析，不需要评分函数对网络进行选择，也不需要运用局部搜索算法反复搜索，大大提高了效率；

2)不需要给出一个结点序，实现起来更为容易；

3)此算法是一个多项式界的算法，只需 $O(n^2)$次 CI(条件独立)测试，与已有的算法相比，大大提高了效率，是一种可行的算法。

最后，为了进一步验证算法的有效性，我们对一个含有1000条医疗数据的样本数据库进行 Markov 网的发现，实验证明，该算法是有效的。

二、基本概念

2.1 Bayesian 网

定义1 Bayesian 网是一个有向无环图，它可以表示为一个三元组(N,E,P)。N 是一组结点的集合，$N=\{x_1,x_2,\cdots,x_n\}$，每一个结点代表一个变量(属性)。E 是一组有向边的集合，$E=\{(x_i,x_j)\mid x_i\neq x_j$ 并且 $x_i,x_j\in N\}$，每一条边(x_i,x_j)表示变量 x_i,x_j 具有因果关系(x_i 是 x_j 的因，x_j 是 x_i 的果)，P 是一组条件概率的集合，$P=\{p(x_i|\pi_i)\mid p(x_i|\pi_i)$表示 x_i 的父亲们 π_i 对 x_i 的影响，$p(x_i|\pi_i)$存于结点 x_i 上$\}$。

引理1 一个 Bayesian 网可以定义一个联合概率分布 $p(x_1,x_2,\cdots,x_n)$为：(假设 Bayesian 网有 n 个结点) $p(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p(x_i|\pi_i)$，其中 π_i 为 x_i 的父亲。

例1 图一的 Bayesian 网所表示的全局概率函数为：

$p(x_1,x_2,\cdots,x_6)=p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_1,x_2)p(x_5|x_2,x_3)p(x_6|x_5)$

2.2 Markov 网

定义2 Markov 网是一个无向图，它也可以表示为一个三元组(N,E,P)。N 的定义与 Bayesian 网络相同。E 是一组无向边的集合，$E=\{(x_i,x_j)\mid x_i\neq x_j$ 并且 $x_i,x_j\in N\}$，每一条边(x_i,x_j)表示变量 x_i,x_j 具有依赖关系。P 是一组条件概率的集合，$P=\{p(x_i|B_i)\mid p(x_i|B_i)$表示 x_i 的直接近邻们 B_i 对 x_i 的影响。$p(x_i|B_i)$存于结点 x_i 上$\}$。

定义3 无向图 G=(V,E)是弦化的，当图中每一个长度大于等于4的圈(回路)都有一条弦(一条连接不相邻顶点的边)。

引理2 一个弦化的 Markov 网所表示的全局概率函数可以写成图中各个圈(最大完全子图)对应的边缘分布的乘积除以各个圈的交集对应的边缘分布的乘积。

例2 图2的 Markov 网所表示的全局概率函数为：

$$p(x_1,x_2,\cdots,x_6)=\frac{p(x_1,x_2,x_3)p(x_1,x_2,x_4)p(x_2,x_3,x_5)p(x_5,x_6)}{p(x_1,x_2)p(x_2,x_3)p(x_5)}$$

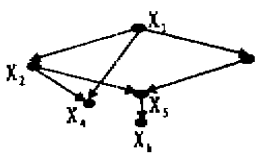

图1

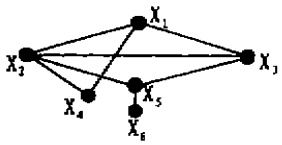

图2

三、基于边界的 Markov 网发现算法

3.1 依赖模型与 Markov 网

现实世界中，知识往往可以用一组条件独立和条件概率表示，Markov 网(无向图)可以用于表示这些条件独立。我们希望发现的 Markov 网尽可能多地表示样本中蕴涵的条件独立。下面我们主要讨论如何用 Markov 网表示一个依赖模型(一组条件概率的集合)以及如何衡量 Markov 网的好坏(引入 I-图和最小 I-图)。

定义4 依赖模型 M 定义为一组条件独立的集合，设 X,Y,Z 是全集 U 的三个子集，$M=\{I(X,Z,Y)\mid I(X,Z,Y)$表示给定 Z 的条件下，X 条件独立于 Y，即 $p(X,Y|Z)=p(X|Z)p(Y|Z)\}$。

定理1[1] 依赖模型 M 中的 $I(X,Z,Y)$满足以下四个性质，设 X,Y,Z 是三个不相交的子集。

·对称性：$I(X,Z,Y)\Leftrightarrow I(Y,Z,X)$

·分解律：$I(X,Z,Y\cup W)\Rightarrow I(X,Z,Y)\&I(X,Z,$

维普资讯 http://www.cqvip.com

W)

·弱归并律：$I(X,Z,Y\cup W)\Rightarrow I(X,Z\cup W,Y)$

·减缩律：$I(X,Z,Y)\&I(X,Z\cup Y,W)\Rightarrow I(X,Z,Y\cup W)$

假如概率函数 p 严格为正，$\forall x,p(x)>0$，则相交律成立。

·相交律：$I(X,Z\cup W,Y)\&I(X,Z\cup Y,W)\Rightarrow I(X,Z,Y\cup W)$。

用图形表示依赖模型，较符合人们的思维方式，具有许多好处。给定一个依赖模型 M，如何用一个无向图来表示 M 呢？我们可以利用无向图中结点分割的概念来表示依赖模型中的条件独立。

定义5 在无向图 G=(V,E)中，设 X,Y,Z 为结点集 V 的三个子集，删去结点集 Z 及其相应的边，使结点集 X,Y 之间再无边相连，称 Z 将 X、Y 分割开，记作$(X|Z|Y)_G$。

我们可以用$(X|Z|Y)_G$来表示依赖模型 M 中的依赖关系(条件独立)$I(X,Z,Y)$，从而用无向图来表示依赖模型。

一个无向图是否能正确完全地反映依赖模型 M 的条件独立信息呢？我们采用 I_图、D_图和 P_图进行衡量。

定义6 设 M 为依赖模型，$I(X,Z,Y)_M$ 表示依赖模型 M 所蕴涵的依赖关系(条件独立)$I(X,Z,Y)$。无向图 G=(V,E)的 I_图、D_图、P_图定义如下：

·G 是 M 的 I_图(独立图)，当$(X|Z|Y)_G\Rightarrow I(X,Z,Y)_M$；

·G 是 M 的 D_图(依赖图)，当$I(X,Z,Y)_M\Rightarrow(X|Z|Y)_G$；

·G 是 M 的 P_图(理想图)，当 G 既是 M 的 D_图又是 M 的 I_图。

由上述定义，I_图不一定包含依赖模型 M 所蕴涵的所有依赖关系，但 I_图中蕴涵的依赖关系 M 中一定蕴涵；D_图恰好相反，D_图包含依赖模型 M 所蕴涵的所有依赖关系，但 D_图中蕴涵的依赖关系 M 中不一定蕴涵；P_图是最理想的情况，P_图与 M 形成一一对应关系。

空图(不含任何边的无向图)是一个平凡的 D_图，而完全图(包含所有的边)是一个不平凡的 I_图。

例3 表示依赖模型 $M=\{I(x_1,x_2x_3,x_4),I(x_2,x_3x_4,x_5),I(x_1x_2x_3,x_4,x_5)\}$可用图3表示。

图 G 是依赖模型 M 的 I_图，因为$(X|Z|Y)_G\Rightarrow I(X,Z,Y)_M$，例如，$(x_2|x_3|x_5)_G$ 在图 G 中成立，$I(X_2,X_4,X_5)$在依赖模型 M 中也成立(因为 $I(x_2,x_3x_4,x_5)$成立，由定理1的分解律可知 $I(x_2,x_4,x_5)$成立。)图 G 同时也是依赖模型 M 的 D_图，因为$I(X,Z,Y)_M\Rightarrow(X|Z|Y)_G$，例如，$I(x_1,x_2x_3,x_4)$在 M 中成立，通过观察，$(x_1|x_2x_3|x_4)_G$ 在 G 中也成立。因此，图 G 是 M 的 P_图。

求依赖模型 M 的 P_图十分困难，甚至于不存在。例如，$M=\{I(x_1,x_2,x_3),I(x_2,x_1,x_3)\}$就不存在 P_图(因为不可能同时在图中表示这两个依赖关系)。退后一步，我们的精力主要放在 M 的 I_图上。但是给定 M 的两个 I_图(G_1,G_2)，我们怎样判定 G_1 与 G_2 的好坏呢？我们标准是能尽可能多地表示依赖模型中的依赖关系(条件独立)。

例4 下图中 G_1 和 G_2 都可以表示依赖模型 $M=\{I(x_1,x_2x_3,x_4),I(x_2,x_1x_4,x_3)\}$，但 G_1 其实什么也没有表示，但 G_2 却表示了 $I(x_1,x_2x_3,x_4)$和 $I(x_2,x_1x_4,x_3)$，可见 G_2 要比 G_1 好。

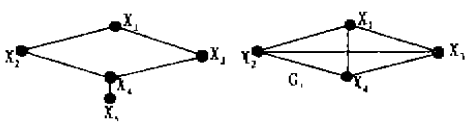

图3

图4

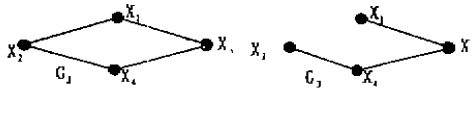

图5

定义7 一个无向图 G=(V,E)是 M 的一个 I_图，若删去 G 的任何一条边后，使得 G 不再是 M 的 I_图，则称 G 为 M 的一个最小 I_图。

显然，由定义，最小 I_图能够最多地表示依赖模型 M 中的依赖关系(条件独立)。我们希望发现的 Markov 网是样本的一个最小 I_图。

例5 图4中 G_2 是依赖模型 $M=\{I(x_1,x_2x_3,x_4),I(x_2,x_1x_4,x_3)\}$的最小 I_图，因为删去 G_2 中的任意一条边后，G_2 将不再是 M 的 I_图，例如，删去边(x_1,x_2)，G_2 变为 G_3，如图5所示。

$(x_1|x_3|x_4)_{G_3}$在 G_3 中成立，但 $I(x_1,x_3,x_4)$在 M 中并不成立，由定义6可知 G_3 不是 M 的 I_图。

3.2 基于边界的 Markov 网发现算法

给定一样本集(n 个属性的一张二维表)，我们希望发现的 Markov 网尽可能多地反映样本所蕴涵的依赖关系(条件独立)，即求样本的最小 I_图。例如：在医疗专家系统中，我们往往希望根据某个病人的病情来判断该病人是否患了某种疾病。我们可以根据过去

维普资讯 http://www.cqvip.com

1000个病人患此病时的症状，通过 Markov 网发现算法发现这些症状与该疾病的关系，即 Markov 网，并利用得到的 Markov 网进行不确定性推理[1]，求出病人患此病的概率，从而对病人的病情作出诊断。

发现 Markov 网包含两个部分：1）结构估计：发现 Markov 网的结构；2）参数估计：Markov 网各结点上概率的估计。

参数估计十分简单，只要用在样本中出现的频率来代表相应概率即可，我们的重点放在结构估计上。

我们采用基于边界的方法来发现 Markov 网。

定义8 设依赖模型 M，U 为全集，$a\in U$，$S\subseteq U$，a 的边界是使 $I(a,S,U-S-a)$ 成立的最小集 S，其中 $I(a,S,U-S-a)$ 表示给定 S，a 条件独立于 $U-S-a$。

a 的边界 S 可以解释为与 a 有直接关系的属性集，给定 S，a 不受其它属性 $(U-S-a)$ 所影响。由于无向图中用结点分割的概念 $(\langle X|Z|Y\rangle_G)$ 来表示依赖模型中的条件独立 $(I(X,Z,Y))$，所以 a 的边界在无向图中表现为 a 的直接近邻。

例6 图 $4G_2$ 中 x_1 的边界是 $\{x_2,x_3\}$，因为虽然 $I(x_1,x_2x_3x_4,\phi)$，$I(x_1,x_2x_3,x_4)$ 均成立，但 $\{x_2,x_3\}$ 是最小集。$\{x_2,x_3\}$ 是 x_1 的直接近邻。

假设被观察的属性总共有 n 个，我们对每一属性都求它的边界，最后将该每个属性与它的边界连接起来，即可求出样本所蕴涵的 Markov 网。

怎样测试 $I(X,Z,Y)$ 呢？我们采用信息论中测试信息独立的一个结论来进行测试。

定义9 设 X，Y，Z 为全集 U 的三个子集，给定条件 Z，X，Y 的条件共同信息定义为：

$$I(X,Y|Z)=\sum_{X,Y,Z\text{的所有取值}}p(X,Y,Z)\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$

假如 $I(X,Y|Z)\leqslant\varepsilon$，则给定 Z，X，Y 条件独立，即 $I(X,Z,Y)$ 成立。否则，给定 Z，X，Y 条件不独立，$I(X,Z,Y)$ 不成立。其中 ε 为门槛值，为大于零的一个较小的正实数。

式中的边缘概率 $p(X,Y,Z)$ 和条件概率 $p(X,Y|Z)$、$p(X|Z)$ 和 $p(Y|Z)$ 可由样本数据（假设是一张二维表）的相应频率计算而得：

$$p(X=x,Y=y,Z=z)=\frac{\text{样本数据中 }X=x,Y=y,Z=z\text{ 的元组个数}}{\text{样本数据包含的元组总数}}\quad(1)$$

$$p(X=x,Y=y|Z=z)=\frac{p(X=x,Y=y,Z=z)}{p(Z=z)}$$

其中 $p(X,Y,Z)$ 和 $p(Z)$ 见式(1)，$p(X|Z)$ 和 $p(Y|Z)$ 类似计算。

我们的算法中需要多次进行 CI 测试，而每次计算边缘概率或条件概率都要统计元组个数，为了避免这种重复统计，我们首先由样本数据得到一个全局概率函数表，假设我们关心的只是样本数据的其中 n 个属性 $\{x_1,x_2,\cdots,x_n\}$，我们将样本数据（关系表）进行投影，并统计元组 $t_i=(a_1,a_2,\cdots,a_n)$ 在样本中出现的频率，作为 t_i 的概率 $p(t_i)$。如下面的例子所示。

例7 给定一样本数据（一张二维表，$R\{K,A,B,C,D\}$），如表1所示。

表1 样本数据 $R\{K,A,B,C,D\}$

$R=$

K	A	B	C	D	K	A	B	C	D
k_1	a_2	b_1	c_1	d_1	k_6	a_1	b_1	c_2	d_1
k_2	a_2	b_1	c_2	d_2	k_7	a_1	b_1	c_2	d_1
k_3	a_1	b_1	c_2	d_1	k_8	a_1	b_2	c_2	d_2
k_4	a_1	b_1	c_2	d_2	k_9	a_1	b_1	c_2	d_1
k_5	a_1	b_2	c_1	d_2	k_{10}	a_1	b_1	c_2	d_2

经过观察，R{K，A，B，C，D}的码为{K}，我们要由样本数据得到一个全局概率函数表 p(A，B，C，D)。我们并不关心码{K}，所以对 R{K，A，B，C，D}在{A，B，C，D}上进行投影，记下元组的重复次数得表 R′{A，B，C，D}，如表2所示。

表2 $R'\{A,B,C,D\}$

$R'=$

A	B	C	D	COUNT
a_1	b_1	c_1	d_1	1
a_1	b_1	c_2	d_2	2
a_1	b_2	c_2	d_1	4
a_1	b_1	c_2	d_2	3

最后，由 $R'\{A,B,C,D\}$ 得到全局概率函数表 $p(A,B,C,D)$，如表3所示，每一元组 t_i 对应的概率函数值 $p(t_i)$ 为 t_i 出现的频率，如：$p(a_1,b_1,c_2,d_2)=3/(1+2+4+3)=0.3$。

表3 全局概率表 $p(A,B,C,D)$

$\Phi_R=$

A	B	C	D	$p(t)$
a_1	b_1	c_1	d_1	$p(a_1,b_1,c_1,d_1)=0.1$
a_1	b_1	c_1	d_2	$p(a_1,b_1,c_1,d_2)=0.2$
a_1	b_1	c_2	d_1	$p(a_1,b_1,c_2,d_1)=0.4$
a_1	b_1	c_2	d_2	$p(a_1,b_1,c_2,d_2)=0.3$

通过全局概率函数表，计算边缘概率 $p(X,Y,Z)$ 和条件概率 $p(X,Y|Z)$、$P(X|Z)$ 和 $p(Y|Z)$ 如下：

$$p(X=x,Y=y,Z=z)=\sum_{t[XYZ]=xyz}p(t)\quad(2)$$

$$p(x=x,Y=y|Z=z)=\frac{p(X=x,Y=y,Z=z)}{p(Z=z)}\quad(3)$$

其中 $p(X,Y,Z)$ 和 $p(Z)$ 见式(2)，$p(X|Z)$ 和 $p(Y|Z)$ 类似计算。

设 U 为属性全集，求属性 $x(x\in U)$ 的边界 S 我们

维普资讯 http://www.cqvip.com

采用剥皮的办法。首先从 $S=U-x$ 开始(因为 $I(x,U-x,\phi)$显然成立),除去S中的一个变量y,测试条件共同信息 $I(x,U-(S-y)-x|S-y)\leqslant\varepsilon$(即判断条件独立 $I(x,S-y,U-(S-y)-x)$是否成立),如果成立,从S中除去y,否则,S不变。这样循环做下去,测试完S中的所有变量,S为使 $I(x,S,U-S-x)$成立的最小集,即 x 的边界。在求属性 x_i 边界时,我们可以参照属性 $x_j(j<i)$的边界,如 x_i 在 x_j 的边界中,则 x_j 也在 x_i 的边界中。这样做可以避免一些不必要的计算,加速计算过程。

假设我们处理的样本数据是完整的,即不存在数据残缺的情况,每一变量(属性)取值为离散型。若样本数据中某一变量取值为连续型,可以首先对其进行离散化(数据归纳)。

基于边界的Morkov网发现算法(伪码表示):

输入:样本数据集W(一张二维表,假设我们感兴趣的是其中n个属性 $U=\{x_1,x_2,\cdots,x_n\}$ 之间的关系),判定两属性是否条件独立的门槛值 ε。

输出:Morkov网,并用图形表示发现的Morkov网。

```
procedure Boundary_learning;
Begin
  由样本数据W得到全局概率函数表 Φ_R{x1,x2,…,xn,f_φR}
  (如表3所示)//f_φR=p(x1,x2,…,xn)可通过每一元组在
  样本中出现的频率求得,见例7。
  For i←1 to n do
    Begin
      s←U-{xi};
      For j←1 to n do
        Begin
          If j<>i then
            Begin
              If xj 在已求出的 xi(j<i)的边界中 then Continue;
              T←S-{xj};
              I ← Σ_{U的所有取值} p(xi,U-{xi}-T,T) log
                  p(xi,U-{xi}-T|T) / (p(xi|T)p(U-{xi}-T|T));
  //计算条件共同信息,式中边缘概率和条件概率的计算见式(2)、(3)
              if I≤ε then S←T;
            End
        End
      Bi←S;//Bi 为 xi 的边界
    End
  图形化表示发现的Markov网。
End.
```

3.3 算法分析及证明

Boundary_learning算法的时间复杂度分析,可以看出基于边界的Markov网发现算法最多只需 $O(n^2)$ 次CI(条件独立)测试,因为对n个属性求边界,求 x_i 的边界,最多只需 $n-1$ 次循环。

下面的定理保证了算法Boundary_learning所得到的Markov网是样本的最小I_图。

定理2[Pearl and Paz 1985] 设U为全集,$\forall a\in U$,满足对称性、分解律、弱归并律和相交律的依赖模型M有唯一的一个边界 $B_I(a)$。并且,$B_I(a)$与M的最小I_图 G_0 的边界 $B_{G_0}(a)$相一致。

定理2要求样本数据满足对称性、分解律、弱归并律和相交律,由定理1,显然样本数据满足对称性、分解律、弱归并律。相交律要求样本数据严格为正,即对x的任意取值,p(x)均大于0,这就需要样本数据尽可能地包含所有可能出现的情况,样本越大,Boundary_learning算法发现的Markov网越接近样本的最小I_图。

四、实验结果

Chest_Clinic网是一个根据专家知识构造的医疗Bayesian网,在给定病人的Smoking(抽烟),Xray(X_光),Dyspnea(呼吸困难)和VisitAsia(去过亚洲)状态的情况下,判断病人是否患有Tuberculosis(肺结核),lung Cancer(肺癌)或Bronchitis(支气管炎)。Chest_Clinic网络的结构及概率描述见http://www.norsys.com/netlib/Asia.dnet。它包含8个结点和8条有向边,每个结点的取值只有两个。

根据Chest_Clinic的概率描述,我们抽样产生1000个数据的样本(Access数据库),该样本包含8个属性{VisitAsia,Tuberculosis,Smoking,Cancer,TborCa,XRay,Bronchitis,Dyspnea}。

注意,样本中的一个元组可能会重复出现多次,我们用元组出现的频率来代替元组的概率,求得全局概率函数表。

应用算法Boundary_learning对样本进行Markov网发现,当门槛值为0.03时,求出了每个属性的边界。

注意到TborCa在XRay的边界中,但Xray却不在TborCa的边界中,这是因为样本数据并不严格为正,即有一些情况没有被考虑到的缘故。假如样本严格为正,由定理2可知,x_i 在 x_j 的边界中,则 x_j 一定在 x_i 的边界中。前面我们提到,样本越大,Boundary_learning算法发现的Markov网越接近样本的最小I_图。

Boundary_learning算法发现的Markov网与Chest_Clinic网相差三条边,如果进一步增大样本,Boundary_learning算法发现的Markov网还将进一步贴近Chest_Clinic网。

Boundary_learning算法发现所有属性的边界共需4分钟。由此可知,Boundary_learning算法是一个有效的Markov网发现算法。

总结 Markov网络是类似于Bayesian网络的一种进行不确定性推理和知识表示的有力工具。本文提出了一个基于边界的Markov网发现算法,该算法只需 $O(n^2)$ 次CI测试,经证明,假如全局概率函数严格为正,则该算法发现的Markov网一定是最能反映样

(下转第65页)

维普资讯 http://www.cqvip.com

	面向对象数据库	RO 模式
使用对象概念的范围	对象主要用来组织数据	具有层次化的对象模型。数据、关系和事务的组织均是面向对象。
数据库设计和处理的重心	重心数据存储本身。面向对象数据库是对象的一种持久化。	重心是 DG 间关系的组织和转化。关系是 RODBMS 的核心。

4.3 RO 模式与数据仓库

RO 模式支持 OLAP 和 OLTP，实现了数据仓库的复杂数据分析能力，其主要区别如下：

	数据仓库	RO 模式
灵活的数据组织	数据仓库往往按固定的算法确定要提取的数据和结构。缺乏灵活性。适合算法固定的工程数据处理。	RO 模式可以按多种算法动态重组数据关系。方法改变灵活。尤其适用于数据挖掘中的试探性算法。
元数据的来源	元数据来自基本的传统数据库。数据仓库缺乏对元数据的控制能力。	元数据来自 RODBMS 的 DG，关系组织可以自由控制元数据。
分析的反馈能力	数据处理的结果并不能直接反馈到元数据库。	处理结果可以通过反馈方法直接对元数据进行反馈。

结束语 RO 模式最大的特点在于它独特的动态数据组织能力。RODBMS 具有出色的可伸缩性，既可组织成传统数据库，又可扩展为适应实际工作的复杂数据库，甚至可作为可动态操作的数据仓库。RO 模式未来的研究重点在以下三个方面：①RO 模式是一个复杂而灵活的数据模式，效率是它最主要的问题；②RDL 的实现。怎样实现动态关系的建立和转化？③DGDL 的实现。DGDL 具体怎样组织数据？

参 考 文 献

1 Ramakrishna R, Gehrke J Database Management System (Second Edition). 清华大学出版社，2000

2 王能斌. 数据库系统原理. 机械工业出版社，2000

3 Ambler S W. Mapping Objects To Relational Databases SIGS Books/Cambridge University 发行，1998，版本号. 1999 年 2 月 26 日。Available at: http. // www. AmbySoft. com/mappingObjects. pdf

（上接第 82 页）

本数据的最小 I-图。从实验结果可以看出，该算法是有效的。由于 Markov 网是一个无向图，不能表示属性之间的因果关系，如何由发现的 Markov 网得到等价的 Bayesian 网是我们下一步研究的课题。

参 考 文 献

1 Pearl J Probabilistic reasoning in intelligent systems. Morgan Kaufmann，1988

2 Xiang Y，Wong S K M. Learning conditional independence relations from a probabilistic model. [Technical report]. University of Regina，1994

3 Suzuki J. Learning bayesian belief network based on the MDL principle. An efficient algorithm using the branch and bound technique. In: Proc. of the intl conf on machine learning，Bari，Italy，1996

4 Herckerman D，Geiger D，Chickering D M Learning bayesian networks: the combination of knowledge and statistical data: [Technical Report MSR-TR-94-09]. Microsoft Research，1994 To apperar，Michine Learning Journal

5 Cheng J，Bell D A，Liu W. An algorithm for bayesian belief network construction from data. In: Proc. of sixth AI & STAT'97 (pp. 83-90)，Ft. Lauderdale，Florida，1997