用于因果分析的混合贝叶斯网络结构学习

王双成[1] ,李小琳[2] ,侯彩虹[1]
(1. 上海立信会计学院 信息科学系 ,上海 201620 ;2. 南京大学 软件技术国家重点实验室 ,江苏 南京 210093)

摘 要:目前主要结合扩展的熵离散化方法和打分 —搜索方法进行混合贝叶斯网络结构学习 ,算法效率和可靠性低 ,而且易于陷入局部最优结构. 针对问题建立了一种新的混合贝叶斯网络结构迭代学习方法. 在迭代中 ,基于父结点结构和 Gibbs sampling 进行混合数据聚类 ,实现对连续变量的离散化 ,再结合贝叶斯网络结构优化调整 ,使贝叶斯网络结构序列逐渐趋于稳定 ,可避免使用扩展的熵离散化和打分 —搜索所带来的主要问题.

关键词:因果分析 ;混合贝叶斯网络 ;最大似然树 ; Gibbs 抽样
中图分类号: TP18 **文献标识码**: A **文章编号**:1673-4785 (2007) 06-0082-08

Learning in a hybrid Bayesian network structure for causal analysis

WANG Shuang-cheng[1] , LI Xiao-lin[2] , HOU Cai-hong[1]
(1. Department of Information Science , Shanghai Lixin University of Commerce , Shanghai 201620 , China ; 2. National Laboratory for Novel Software Technology , Nanjing University , Nanjing 210093 , China)

Abstract: At present , learning in a hybrid Bayesian network structure mainly depends on a combination of the searching & scoring method and the expanded entropy discretization algorithm. However , the algorithm is prone to fall into local optimal traps and its efficiency and reliability are not good. In this paper , a new iterative method for learning with hybrid Bayesian network structures is presented. In each iteration , mixed data clustering is carried out based on father mode structures and Gibbs sampling , so that continuous variables are discretized. Then , through optimization of the Bayesian network structure , the sequence of Bayesian network structures gradually tends to converge , avoiding the main problems encountered with the expanded entropy discretization algorithm.

Keywords: causal analysis ; hybrid Bayesian network ; maximum likelihood tree ; Gibbs sampling

人类对现实世界中现象的一种强烈渴望就是因果联系 ,通过因果关系能够揭示事物之间的内在联系 ,从而达到认识世界和改造世界的目的 ,因此 ,因果关系是哲学、自然科学和社会科学等的重要研究内容. 贝叶斯网络[1] 是描述随机变量之间依赖关系的图形模式 ,具有形象直观的知识表示形式 ,以及更接近人思维特征的推理方式 ,被广泛用于不确定性知识表示和推理. 在一些假设下 ,贝叶斯网络中边的方向具有因果语义 ,因此是变量之间因果分析的有力工具. 20 世纪 80 年代后期 ,加利福尼亚大学计算机科学系 Judea Dearl 给出了贝叶斯网络的严格定义 ,并建立了贝叶斯网络的基础理论体系[1] . 进入 90 年代以后 ,以 Pearl、Heckerman 、Peter Spirtes 、Chickering 和 Henson 等[2-6] 为代表相继进行了基于贝叶斯网络的因果分析的理论探索和应用研究 ,但这些研究都是基于离散变量的假设. 基于贝叶斯网络进行混合变量因果分析比较困难 ,其核心是混合贝叶斯网络结构学习. 以往对混合贝叶斯网络结构学习的研究主要从 2 个方面展开 :一方面是不离散化连续变量 ,通过打分 —搜索方法直接进行混合贝叶斯网络结构学习 , Thiesson 和 Murphy 等人曾经给出过一些近似打分函数[6-8] ,但由于运算过于复杂 ,不具有实用价值 ,至今还没有实质性的进展 ;另一方面是离散化连续变量 ,转化为离散变量贝叶斯网络结构学习问题 ,由于变量之间的因果结构具

收稿日期:2007-01-04.
基金项目:国家自然科学基金资助项目(60675036) ;上海市重点学科基金资助项目(P1601) ;上海市教委重点基金资助项目(05zz66) .

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

有相对稳定性,在发现因果结构时允许忽略一些细节信息,因此,基于离散化的方法是混合贝叶斯网络结构学习更为有效、实用的方法.

目前,在混合贝叶斯网络结构学习中连续变量的离散化主要采用扩展的熵离散化方法(Fayyad 和 Irani[9]提出的基于熵离散化方法的推广,早期用于分类器学习中连续变量的离散化).这种方法的实现是一个迭代过程,在迭代过程中,以父结点为条件的条件熵作为打分函数,对分点进行贪婪(greedy)或随机打分—搜索,使用 MDL[10](minimal description length)标准(或条件熵)确定分点数,并在新数据集的基础上使用打分—搜索方法重新进行结构学习.由于分点空间的大小随数据量的增加指数增长,无论采用哪种打分—搜索方法,当数据量大时都很难实现,而且得到的一般是局部最优划分.在打分—搜索结构学习中,由于打分函数的计算复杂性和结构搜索空间的大小也都随变量增加指数增长,因此一般要求结点有顺序,并根据打分函数的可分解性进行局部确定或随机搜索(完全搜索是 N-P 困难问题[11]),这样重新结构学习效率低,易于陷入局部最优结构.此外,基于扩展的熵离散化方法强调的是属性变量对类变量的分类贡献,不具有因果语义,这样易于导致因果信息的丢失和冗余.

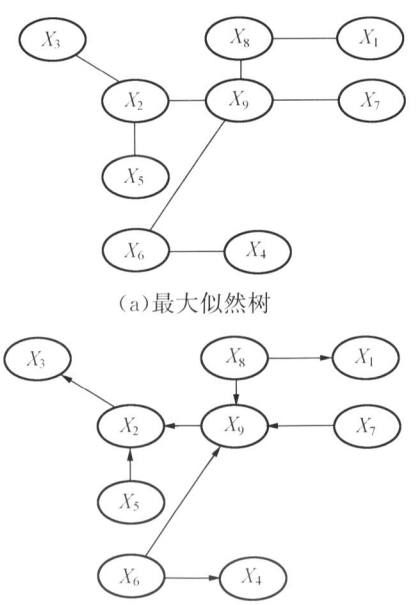
本文结合父结点结构(因果结构)和 Gibbs sampling[12-13]进行混合数据聚类,通过聚类实现连续变量的离散化,进行混合贝叶斯网络结构迭代学习,每一次离散化后进行贝叶斯网络因果关系优化调整,直到迭代收敛.按照父结点结构分解联合概率,解决了标准 Gibbs sampling 的指数复杂性问题[12],因此能够显著提高离散化效率,Gibbs sampling 过程收敛到全局平稳分布[12-13],这样可避免使用扩展的熵离散化方法的局部最优化分问题.贝叶斯网络因果关系优化调整将使变量之间的依赖关系逐渐趋于因果关系,实验结果显示,这种方法能够有效地进行混合贝叶斯网络结构学习.

用 $X_1, \cdots, X_n$ 表示连续和离散混合随机变量,$x_1, \cdots, x_n$ 表示变量的取值.D 表示具有 N 个例子的混合数据集,数据随机产生于联合分布 P.

1　初始化混合贝叶斯网络学习

初始混合贝叶斯网络学习包括连续变量的初始离散化和贝叶斯网络的初始化 2 部分.采用二分法对连续变量进行初始离散化(以均值作为离散化的分界点,进行二值离散化),把离散化后的数据集作为初始数据集,用 $D^{(0)}$ 表示.由于数据集 $D^{(0)}$ 中所蕴涵的变量之间的依赖关系可能比较混乱,这将导致直接进行贝叶斯网络学习不够可靠,影响随后的迭代收敛速度.最大似然树是与贝叶斯网络具有最好拟合的属性结构,在树中蕴含的依赖关系往往是贝叶斯网络中的重要依赖关系,而且最大似然树的结构相对稳定,学习效率高,因此,采用为最大似然树边因果定向的有向无环图作为初始贝叶斯网络.首先依据 Chow 和 Liu[14]的方法建立最大似然树,经过简单的碰撞识别定向后得到 Polytree[1](这时大部分边已经确定了方向,并且方向具有因果语义),然后基于链图和 MDL 标准为其他没有定向的边定向.

从数据集 $D^{(0)}$ 中建立最大似然树 T,并对 T 中的边进行碰撞识别定向得到 Polytree T_1.T_1 是一个链图[15](chain graph),设链图中没有确定方向的边为 $e_1, \cdots, e_s$,用 G_C^{i+} 和 C_C^{i-} 分别表示边 $e_1, \cdots, e_{i-1}$ 已定向、而边 $e_{i+1}, \cdots, e_s$ 还没定向、由边 e_i 方向所确定的不同链图.按照 Buntine[15]给出的基于链图分解联合概率的方法,便能计算一个链图的 MDL 打分.根据 MDL$(G_C^{i+} \mid D)$ 和 MDL$(G_C^{i-} \mid D)$ 的大小确定边 e_i 的方向,直到确定所有未定向边的方向.把最后得到有向无环图记为 $G^{(0)}$,并作为初始贝叶斯网络结构,由 $G^{(0)}$ 所确定的结点顺序记为 $X_1^{(0)}, \cdots, X_n^{(0)}$.将数据集 pima_indians_diabete 中的连续数据进行二分离散化,从中学习得到的最大似然树和初始贝叶斯网络结构如图 1 所示.



(a)最大似然树

(b)定向后的有向无环图

图 1　初始 pima_indians_diabetes 贝叶斯网络结构

Fig. 1　Initial Bayesian network structure of pima_indians_diabetes dataset

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2 混合贝叶斯网络迭代学习

混合贝叶斯网络迭代学习包括2个迭代过程,一个是在确定结构下的连续变量离散化内层迭代,另一个是结构外层迭代.迭代产生2个序列,它们是离散数据集序列$\{D^{(k)}\}$和贝叶斯网络结构序列$\{G^{(k)}\}$.每一次外层迭代又包括2个部分,一部分是基于上一次迭代得到的贝叶斯网络结构$G^{(k)}$重新离散化连续变量得到新的离散数据集$D^{(k)}$;另一部分是在$D^{(k)}$的基础上,优化调整贝叶斯网络结构得到$G^{(k+1)}$,直到满足结构迭代终止条件结束迭代.

2.1 基于混合数据聚类的连续变量离散化

对每一个连续变量的离散化是一个子迭代过程,按照贝叶斯网络结构$G^{(k)}$所确定的变量顺序依次对连续变量进行重新离散化,下一个连续变量的离散化在上一个离散化结果的基础上进行,直到离散化完所有的连续变量.为有效地继承因果信息,结合贝叶斯网络中的因果局部结构(父结点结构)和Gibbs sampling进行混合数据聚类,通过聚类实现连续变量的离散化.在对一个连续变量的离散化过程中,包括确定对应的离散变量取值(类值)和维数(类数)2部分内容.

在依次对连续变量重新离散化的过程中,贝叶斯网络结构保持不变,但数据集在不断的变化,这样产生一个离散数据集子序列$D_0^{(k)}=D^{(k)}, D_1^{(k)}, \cdots, D_t^{(k)}=D^{(k+1)}$,其中$D_j^{(k)}$是在$D_{j-1}^{(k)}$中用连续变量$X_j$的最新离散化数据代替原有离散化数据而得到的数据集.

2.1.1 确定连续变量对应的离散变量取值

设由$G^{(k)}$所确定的变量顺序为$X_1^{(k)}, \cdots, X_n^{(k)}$,为方便把离散化后的变量排在前部,在每一部分内部保持原来的顺序,得到新的变量序列为$X_1^{(k)}, \cdots, X_t^{(k)}, X_{t+1}, \cdots, X_n$.用$X_i$替换$X_i^{(k)}$,通过混合数据聚类得到$X_i$离散化后的离散变量(类变量),仍用$X_i^{(k)}$表示.

由贝叶斯和乘法公式可得

$$p(x_i^{(k)} \mid x_1^{(k)}, \cdots, x_{i-1}^{(k)}, x_i, G^{(k)}) = \alpha p(x_i^{(k)}, \pi_{x_i}^{(k)}, x_i, G^{(k)}) = \beta p(x_i \mid \pi_{x_i}^{(k)}, x_i^{(k)}, G^{(k)})\, p(x_i^{(k)} \mid \pi_{x_i}^{(k)}, G^{(k)}). \quad (1)$$

式中:$\pi_{x_i}^{(k)}$是$G^{(k)}$中$X_i^{(k)}$父结点集的配置,α和β是与$x_i^{(k)}$无关的量.

基于$G^{(k)}$和Gibbs sampling进行混合数据聚类,从而实现对变量X_i的重新离散化,用新的离散化变量的值代替$D_i^{(k)}$中$X_i^{(k)}$的值,并在$D_i^{(k)}$的基础上重新离散化X_{i+1}得到$D_{i+1}^{(k)}$,直到完成所有连续变量的重新离散化得到$D^{(k+1)}$.在式(1)中,对连续变量采用条件正态密度函数,也可采用其他密度函数(核密度或多项式密度函数等).在聚类迭代过程中,结合$G^{(k)}$和Gibbs sampling依次确定数据库中每一个记录中的$X_i^{(k)}$的值,确定了数据库中所有记录中的$X_i^{(k)}$的值便实现一次迭代,直到满足终止条件结束迭代.下面给出在一个记录中确定$X_i^{(k)}$值的方法:

$$p(x_i \mid \pi_{x_i}^{(k)}, x_i^{(k)}, G^{(k)}) = g(x_i; \mu_i(\pi_{x_i}^{(k)}, x_i^{(k)}); \sigma_i(\pi_{x_i}^{(k)}, x_i^{(k)}) \mid G^{(k)}) = \frac{1}{\sqrt{2\pi}\sigma_i(\pi_{x_i}^{(k)}, x_i^{(k)})} e^{-\frac{x_i-\mu_i(\pi_{x_i}^{(k)}, x_i^{(k)})}{2\sigma_i^2(\pi_{x_i}^{(k)}, x_i^{(k)})}}. \quad (2)$$

式中:$\mu_i(\pi_{x_i}^{(k)}, x_i^{(k)})$和$\sigma_i(\pi_{x_i}^{(k)}, x_i^{(k)})$分别表示变量$X_i$条件均值和条件标准差.

如果$p(x_i^{(k)} \mid \pi_{x_i}^{(k)}, G^{(k)})=0$,对$p_i(x^{(k)} \mid \pi_{x_i}^{(k)}, G^{(k)})$进行拉普拉斯修正(Laplace-corrected)[16]:

$$p(x_i^{(k)} \mid \pi_{x_i}^{(k)}, G^{(k)}) = (1/N)(N(\pi_{x_i}^{(k)}) + N(x_i^{(k)})(1/N)).$$

式中:$N(\pi_{x_i}^{(k)})$为$X_i^{(k)}$的父结点集$\prod_{X_i}^{(k)}$具有配置$\pi_{x_i}^{(k)}$的例子数量,$N(x_i^{(k)})$为$X_i^{(k)}=x_i^{(k)}$的例子数量.

对选择的维数l_i($2 \le l_i \le M_i^{(k)}$),首先随机初始化$X_i^{(k)}$的值,然后通过抽样对$D_i^{(k)}$中变量$X_i^{(k)}$的值进行修正,直到收敛.

设需要离散化的变量X_i离散化后对应的离散变量最大可能的维数为$M_i^{(k)}$(一般取5或6),按照数据库中记录的顺序依次对$X_i^{(k)}$的值进行修正.

设$X_i^{(k)}$在第m个记录的待修正值为$x_{im}^{(k)}$,$\hat{x}_{im}^{(k)}$表示修正后的值,变量$X_i^{(k)}$的可能取值为$x_i^1, \cdots, x_i^{l_i}$.对抽样式子进行归一化处理,记$w_i^{(k)}(h) = \frac{p(x_i \mid \pi_{x_i}, x_i^h, G^{(k)})\, p(x_i^h \mid \pi_{x_i}, G^{(k)})}{\sum_{j=1}^{l_i} p(x_i \mid \pi_{x_i}, x_i^j, G^{(k)})\, p(x_i^j \mid \pi_{x_i}, G^{(k)})}$, $h \in \{1, \cdots, l_i\}$,对生成的随机数λ,离散变量$X_i^{(k)}$的取值为

$$\hat{x}_{im}^{(k)} = \begin{cases} x_i^1, & 0 < \lambda \le w_i^{(k)}(j), \\ \cdots & \\ x_i^h, & \sum_{j}^{h-1} w_i^{(k)}(j) < \lambda \le \sum_{j}^{h} w_i^{(k)}(j), \\ \cdots & \\ x_i^{l_i}, & \lambda > \sum_{j}^{l_i-1} w_i^{(k)}(j). \end{cases} \quad (3)$$

2.1.2 确定最优的离散化方案

根据最优维数确定的离散化方案便是最优的离散化方案,因此确定最优的离散化方案的核心是确

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

定连续变量对应的离散变量的最优维数. 设 $D_{i2}^{(k)}$, …, $D_{iM_i^{(k)}}^{(k)}$ 是使用不同的离散化方案 Λ_h ($2 \le h \le M_i^{(k)}$),选择一个 h 的值便得到一个离散化方案) 离散化变量 X_i 而得到的离散数据集序列,对每一个离散化方案分别进行 MDL 标准打分,取 $h_0 = \min\limits_{M_i^{(k)} \ge h \ge 2} \{\mathrm{MDL}(\Lambda_h \mid D_{ih}^{(k)})\}$ 作为离散变量的维数.

$$\mathrm{MDL}(\Lambda_h \mid D_{ih}^{(k)}) = \frac{\lg N}{2} \mid \Lambda_h \mid - L(\Lambda_h \mid D_{ih}^{(k)}). \tag{4}$$

式中:$|\Lambda_h|$ 表示离散化方案 Λ_h 中变量 $X_i^{(k)}$ 在 $G^{(k)}$ 中马尔可夫毯[1]结构的参数数量:

$$L(\Lambda_h \mid D_{ih}^{(k)}) = \sum_{i}^{N} \lg(P(u_i \mid M_{X_i^{(k)}}, D_{ih}^{(k)}, \Lambda_h)) =$$

$$N \sum_{x_i^{(k)}, \pi_{x_i^{(k)}}} p(x_i^{(k)}, \pi_{x_i^{(k)}} \mid M_{X_i^{(k)}}, D_{ih}^{(k)}, \Lambda_h)$$

$$\prod_{x_i^{(k)} \in t_{x_j}} p(x_j, \pi_{x_j} \mid M_{X_i^{(k)}}, D_{ih}^{(k)}, \Lambda_h) \cdot$$

$$\lg(p(x_i^{(k)} \mid \pi_{x_i^{(k)}}, M_{X_i^{(k)}}, D_{ih}^{(k)}, \Lambda_h)$$

$$\prod_{x_i^{(k)} \in t_{x_j}} p(x_j \mid \pi_{x_j}, M_{X_i^{(k)}}, D_{ih}^{(k)}, \Lambda_h).$$

2.2　贝叶斯网络结构优化调整

贝叶斯网络结构优化调整包括结点顺序和结构调整 2 部分内容,通过优化将得到更好的贝叶斯网络结构,直到迭代中的贝叶斯网络结构趋于稳定.

2.2.1　依赖关系优化

利用贝叶斯网络的信息管道模型[17-18]描述变量之间存在的 3 种基本依赖关系(边):1) 传递依赖(transitive dependencies),结点之间存在直接的信息流动,而且信息流不能被其他结点所阻塞,结点所表示的变量之间边缘和条件依赖;2) 非传递依赖(non-transitive dependencies),结点之间不存在直接的信息流动,而是由连接两结点之间的开路(不含碰撞结点的路径)产生信息流,这种信息流被切割集中的结点所阻塞,2 个结点所表示的变量之间边缘依赖,但条件独立;3) 导出依赖(induced dependencies),这种依赖是由 V 结构所导致,结点之间不存在直接的信息流动,而是由 V 结构中的碰撞结点诱发的信息流,结点所表示的变量之间边缘独立,但条件依赖.

下面基于变量之间基本依赖关系和依赖分析思想进行依赖关系优化(包括补充丢失的依赖关系和删除冗余的依赖关系)和因果语义优化 2 部分内容.

使用互信息和条件互信息进行变量之间定量条件独立性检验,分别用 $I(X_i, X_j)$ 和 $I(X_i, X_j \mid X_{u_1}, \cdots, X_{u_h})$ 表示变量 X_i 和 X_j 之间的互信息和以 $X_{u_1}, \cdots, X_{u_h}$ ($u_k \ne i, j, k = 1, \cdots, h$) 为条件的条件互信息,对给定的小正数 ε(一般取 $\varepsilon = 0.01$),如果 $I(X_i, X_j \mid X_{u_1}, \cdots, X_{u_h}) < \varepsilon$,就认为 X_i 和 X_j 之间条件独立.

1) 补充丢失的依赖关系

对不存在边的结点对 X_i 和 X_j,依次进行互信息 $I(X_i, X_j)$ 计算. 如果 $I(X_i, X_j) > \varepsilon$,在 $G^{(k)}$ 中增加边 $X_i - X_j$. 用 $L_0^{(k)}$ 表示对应的边表,其中的元素是三元组(i, j, η), $j > i$, $i = 1, \cdots, n$, $j = 2, \cdots, n$, $\delta = 0, 1$, $\eta = 1$ 和 $\eta_0 = 0$ 分别表示存在和不存在边 $X_i - X_j$. 这一过程最多需要 $(n+1)(n+2)/2$ 次互信息计算.

2) 删除冗余的依赖关系

由于连续变量重新离散化,可能存在一些冗余的边,通过下面的方法去除冗余的边. 在 $G^{(k)}$ 中,用 $S_{X_i}(X_i, X_j)$ 和 $S_{X_j}(X_i, X_j)$ 分别表示 X_i 和 X_j 的邻域中在 X_i 和 X_j 链路[18]上的结点集,$S(X_i, X_j)$ 表示 $S_{X_i}(X_i, X_j)$ 和 $S_{X_j}(X_i, X_j)$ 中具有较少结点的结点集. 对存在边的结点对 X_i, X_j,设 $S(X_i, X_j) = \{X_1^{(i,j)}, \cdots, X_t^{(i,j)}\}$,如果存在 t_0 ($1 \le t_0 \le t$) 使 $I(X_i, X_j \mid X_{t_0}^{(i,j)}, D) < \varepsilon$,在 $G^{(k)}$ 中删除边 $X_i - X_j$ 和修改边表 $L_0^{(k)}$;否则,选取 $X_t^{i*} = \underset{X_k^{(i,j)}, t \ge k \ge 1}{\operatorname{argmin}} \{I(X_i, X_j) \mid X_k^{(i,j)}, D^{(k)}\}$,如果存在 h_0 ($1 \le h_0 \le t$, $h_0 \ne t^*$),使 $I(X_i, X_j \mid X_{h_0}^{(i,j)}, X_t^{(i,j)}, D^{(k)}) < \varepsilon$,在 $G^{(k)}$ 中删除边 $X_i - X_j$ 和修改边表 $L_0^{(k)}$;重新确定 $S(X_i, X_j)$,由于产生第 2 种依赖的信息流往往能被少数结点所阻塞(多数情况是 1 个或 2 个结点),因此大部分的第 2 种边已被清除,$S(X_i, X_j)$ 将具有很少的冗余结点,如果 $I(X_i, X_j \mid S(X_i, X_j), D^{(k)}) < \varepsilon$,在 $G^{(k)}$ 中删除边 $X_i - X_j$ 和修改边表 $L_0^{(k)}$. 这一过程最多需要 $2n^2$ 次条件互信息计算.

2.2.2　因果语义优化

因果语义优化就是边的方向优化,下面分别基于碰撞识别和条件预测能力优化边的方向,以便得到更适合于因果分析的贝叶斯网络.

基于碰撞识别优化边的方向:

在边表 $L_0^{(k)}$ 中查寻,选择结点对 X_i 和 X_j,设 X_i 和 X_j 之间不存在边,在 $G^{(k)}$ 中与 X_i 和 X_j 可能形成 V 结构的结点为 $X_{m_1}, \cdots, X_{m_t}$,对每一个可能的 V 结构进行碰撞识别(汇聚识别)[16-20]. 对给定的阈值 $\delta > 0$,如果 $\dfrac{I(X_i, X_j \mid X_{m_h}, D^{(k)})}{I(X_i, X_j \mid D^{(k)})} > (1 + \delta)$, $1 \le h \le t$,则 X_i, X_j 和 X_{m_h} 形成 V 结构,定向为 $X_i \to X_{m_h}$ 和 $X_j \to X_{m_h}$,这一过程最多需要 $n(n-1)$ 次条件互信息计算.

使用碰撞识别调整方向后,可能还有一部分边的方向没有得到调整,下面基于变量之间的预测能

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

力为其余的边调整定向.这种方法的基本思想是:对任意的2个变量 X_i 和 X_j,在排除其他变量影响的情况下,如果 X_i 对 X_j 比 X_j 对 X_i 更具条件预测能力,方向应该由 X_i 指向 X_j.

记 $F(X_{m_1},\cdots,X_{m_t}\rightarrow X_i)$ 为变量 $X_{m_1},\cdots,X_{m_t}$ 对 X_i 的预测能力

$$F(X_{m_1},\cdots,X_{m_t}\rightarrow X_i)=\sum_{x_{m_1},\cdots,x_{m_t}} p(x_{m_1},\cdots,x_{m_t})\max_{X_i(x_{m_1},\cdots,x_{m_t})}\{p(x_i\mid x_{m_1},\cdots,x_{m_t})\},$$

记 $R(X_{m_1},\cdots,X_{m_t},X_j\rightarrow X_i)$ 为以 $X_{m_1},\cdots,X_{m_t}$ 为条件,X_j 对 X_i 的相对预测能力

$$R(X_{m_1},\cdots,X_{m_t},X_j\rightarrow X_i)=\frac{F(X_{m_1},\cdots,X_{m_t},X_j\rightarrow X_i)}{F(X_{m_1},\cdots,X_{m_t}\rightarrow X_i)}-1,$$

式中:$j\neq i$,$j\neq m_h$,$h=1,\cdots,t$.

对选择的 X_i 和 X_j,分别用 $B(X_i)$ 和 $B(X_j)$ 表示 X_i 和 X_j 的马尔可夫毯,当 $B(X_i)$ 和 $B(X_j)$ 给定时,X_i 和 X_j 之间便没有其他因素的影响,按如下方法调整其他边的方向.

如果 $R(B(X_i)\cup B(X_j),X_i\rightarrow X_j)>R(B(X_i)\cup B(X_j),X_j\rightarrow X_i)$,而且 $X_i\rightarrow X_j$ 不产生环路,就定向为 $X_i\rightarrow X_j$,否则定向为 $X_i\leftarrow X_j$.

如果 $R(B(X_i)\cup B(X_j),X_i\rightarrow X_j)<R(B(X_i)\cup B(X_j),X_j\rightarrow X_i)$,而且 $X_i\leftarrow X_j$ 不产生环路,就定向为 $X_i\leftarrow X_j$,否则定向为 $X_i\rightarrow X_j$.

如果 $R(B(X_i)\cup B(X_j),X_i\rightarrow X_j)=R(B(X_i)\cup B(X_j),X_j\rightarrow X_i)$,选择不产生环路的情况定向,2种情况都不产生环路就随机定向.基于预测能力的定向最多需要进行 $\frac{1}{2}n(n-1)$ 次条件相对预测能力计算,对 $G^{(k)}$ 调整后得到 $G^{(k+1)}$.

综上所述,进行一次贝叶斯网络结构优化调整算法的时间复杂性是 $O(n^2)$.图2给出了经过迭代学习得到的混合贝叶斯网络结构.

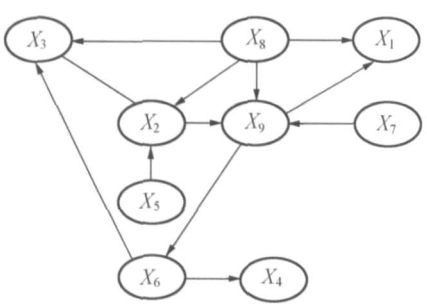


图2 迭代学习后的混合贝叶斯网络结构

Fig.2 Hybrid Bayesian network structure after iterative learning

关于混合贝叶斯网络的参数学习,如果使用离散化后的离散数据,可直接由数据统计得到;如果使用混合数据,由于父结点结构是分类或回归结构,对离散变量结点可采用分类技术,对连续变量结点可采用神经网络或Logistic回归确定参数,限于篇幅不予详述.

2.3 迭代终止检验

迭代终止检验包括离散化(聚类)迭代终止检验和结构迭代终止检验,离散化迭代是具有确定贝叶斯网络结构的内层迭代,结构迭代是外层迭代,结构迭代终止学习过程便结束.

2.3.1 离散化迭代终止检验

采用相邻2次迭代被离散化变量值序列的一致性检验进行终止迭代判断.设相邻2次迭代所得到的离散化值序列分别为 $x_{i1}^{(k)},x_{i2}^{(k)},\cdots,x_{iN}^{(k)}$ 和 $x_{i1}^{(k+1)},x_{i2}^{(k+1)},\cdots,x_{iN}^{(k+1)}$,那么 $\mathrm{sig}(x_{ij}^{(k)},x_{ij}^{(k+1)})=\begin{cases}0, x_{ij}^{(k)}=x_{ij}^{(k+1)}\\ 1, x_{ij}^{(k)}\neq x_{ij}^{(k+1)}\end{cases}$,$1\leqslant j\leqslant N$.对给定的阈值 $\eta_0>0$,如果 $\frac{1}{N}\sum_{j=1}^{N}\mathrm{sig}(x_{ij}^{(k)},x_{ij}^{(k+1)})<\eta_0$,则结束迭代.

2.3.2 结构迭代终止检验

按照变量的某一确定顺序(如在数据库中的顺序)建立贝叶斯网络结构数组,$G^{(k)}$ 所对应的结构数组为 $w^{(k)}=(a_{12}^{(k)},\cdots,a_{1n}^{(k)},\cdots,a_{i(i+1)}^{(k)},\cdots,a_{in}^{(k)},\cdots,a_{(n-1)n}^{(k)})$.当 X_i 和 X_j 之间存在边时,$a_{ij}^{(k)}=1(i<j)$,否则 $a_{ij}=0$.对给定的正整数阈值 n_0,如果 $\sum_{i<j}|a_{ij}^{(k+1)}-a_{ij}^{(k)}|<n_0$,结束迭代.

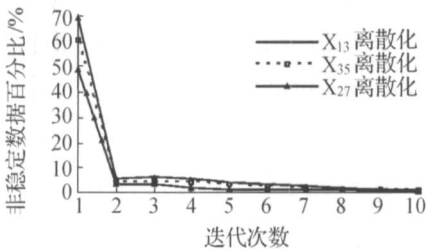
3 实 验

根据网站 http://www.norsys.com 提供的ALARM网概率分布表生成离散变量模拟数据集,使用文献[19]中的离散变量连续化的方法,把离散变量 $X_2,X_4,\cdots,X_{36}$ 连续化.分别取 $\eta_0=0.05$ 和 $n_0=4$ 进行实验.

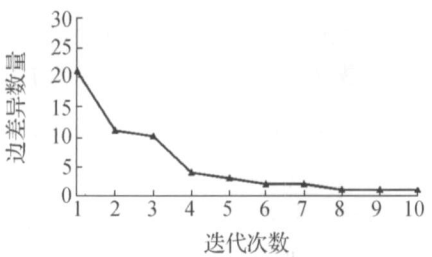
在第1次贝叶斯网络迭代中,对具有多父结点的连续变量 X_{13}、X_{35}、X_{27},分别取4、4、2个值的离散化迭代收敛情况如图3(a)所示.贝叶斯网络结构迭代收敛情况如图3(b)所示.

从图3(a)中可以看出,对 X_{13}、X_{35}、X_{27} 迭代5次后均收敛,显示了具有很高的离散化效率,而其他的离散化算法在离散化过程中都具有很大的系统开销.对其他连续变量(包括不同迭代次数)离散化迭代可得到类似的收敛情况.图3(b)中显示,对结构迭代6次后便收敛,表明学习算法具有很高的效率.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



(a)离散化迭代收敛情况

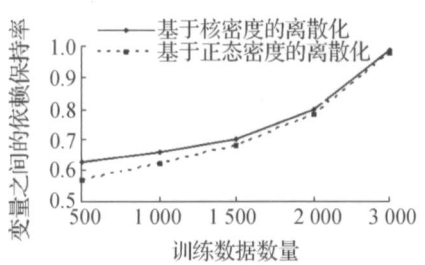


(b)结构迭代收敛情况

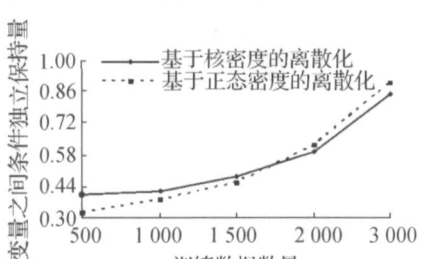
图 3　迭代收敛情况

Fig. 3　The iteration convergence situation

基于核密度离散化算法和基于正态密度的离散化算法的比较情况如图 4 所示.



(a)变量之间的依赖保持率



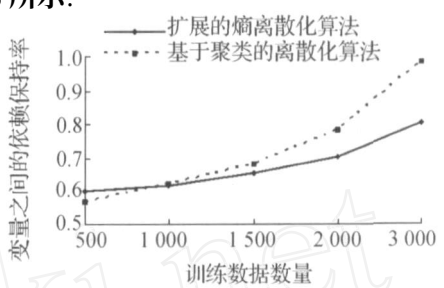
(b)变量之间的条件独立性保持率

图 4　基于正态密度和核密度的离散化情况比较

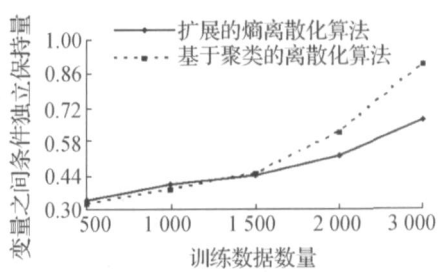
Fig. 4　The situation comparison of discretizing continuous variables between normal density and kernel density

从图 4 的实验结果可以看出,在依赖关系保持方面,当例子少时,基于核密度的离散化算法明显优于基于正态密度的离散化算法,当例子多时,在依赖保持率方面,基于核密度的离散化算法略占优势,在条件独立性保持率方面,基于正态密度的离散化方法略占优势.由于核密度估计随例子增加运算复杂程度的增长远大于正态密度,因此当例子多时适合于采用正态密度离散化方法,而当例子少时适合于采用核密度离散化方法.

使用扩展的熵离散化算法和本文建立的聚类离散化算法进行连续变量的离散化,其有效性的比较如图 5 所示.



(a)变量之间的依赖保持率比较



(b)变量之间的条件独立性保持率比较

图 5　基于聚类的离散化算法与扩展的熵离散化算法比较

Fig. 5　The situation comparison of discretizing continuous variables between clustering and extended entropy

图 5 中显示,当例子数据量小时,扩展的熵离散化算法略优于聚类算法,但随例子数据的增加,聚类算法明显具有优势.具有这种优势的主要原因是:随着例子数据的增加变量之间依赖和条件独立性信息能够得到充分的利用,且贝叶斯网络结构得到不断的优化调整,而在扩展的熵离散化算法中,随着例子数据的增加,离散化和结构学习的局部最优性影响逐渐增加,导致了最终的差距.

在 UCI 机器学习数据仓库[20]中选择 10 个分类数据集,脚标最大的一个是类变量,其他的是属性变量.从其中 3 个数据集 heart_disease、breast_cancer 和 cmc 学习得到的贝叶斯网络结构如图 6 所示.基于扩展的熵离散化和聚类离散化 2 种方法,对类变

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

量的预测能力(推理能力)比较情况如表1所示.

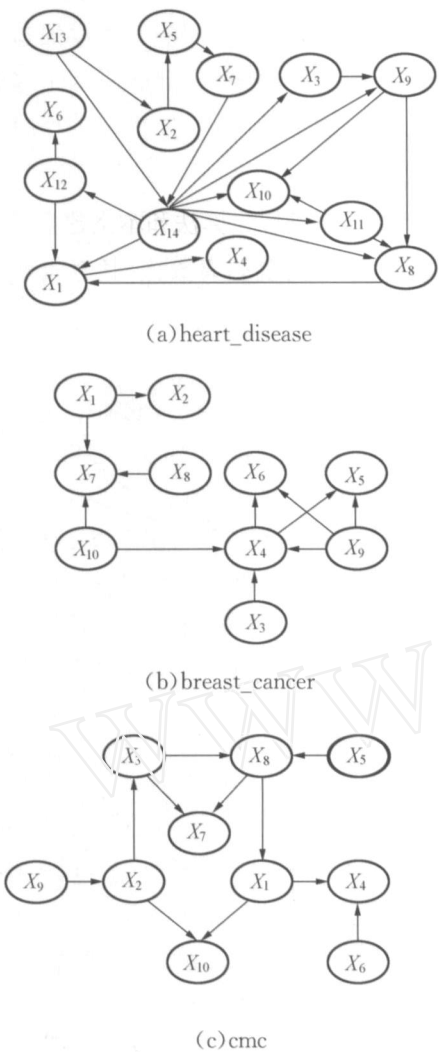


图6 学习得到的贝叶斯网络结构

Fig. 6 Learned Bayesian network structures

表1 2种离散化方法在推理能力方面的比较

Table 1 The comparison of two discretizing methods in inference respect

数据集	基于扩展熵离散化	基于聚类离散化	变量数量
iris	95.33 ±4.27	97.33 ±4.42	5
liver_disease	56.57 ±2.62	56.57 ±2.62	7
pima_indians_diabetes	70.78 ±3.83	74.55 ±2.63	9
heart_disease	77.40 ±5.80	86.66 ±5.28	14
new_thyroid	74.76 ±5.92	74.76 ±5.92	6
cmc	44.00 ±4.51	42.30 ±3.68	10
wdbc	93.16 ±1.66	97.02 ±1.93	32
breast_caneer	74.48 ±3.44	75.85 ±3.44	10
Thyroid0387	66.40 ±2.76	66.60 ±2.58	22
sick_euthyroid	90.76 ±1.17	91.55 ±1.11	25

从表1中可以看出,基于聚类离散化方法学习得到的贝叶斯网络对类变量的预测能力明显优于基于扩展熵离散化方法学习得到的贝叶斯网络,对属性变量的预测能力也可得到类似的结果,可知,基于聚类离散化方法学习得到的贝叶斯网络在推理方面更加可靠.

4 结束语

建立了具有连续变量的贝叶斯网络结构迭代学习方法,在迭代中过程中,一方面,通过基于父结点结构和 Gibbs sampling 进行混合数据聚类来实现连续变量的离散化,并根据离散变量的马尔可夫毯和 MDL 打分确定离散变量的最优维数,这样能够有效地动态继承变量之间的因果关系;另一方面,基于依赖分析方法的贝叶斯网络因果结构优化调整使变量之间因果关系不断得到改进,逐渐趋于稳定.该方法避免了使用基于扩展的熵离散化方法和打分-搜索结构学习方法所带来的主要问题,同时,也可用于其他具有连续变量的相关问题.

参考文献:

[1]PEARL J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Mateo, Morgan Kaufmann, 1988.

[2]HECKERMAN D, GEIGER D, CHICKERING D M. Learning Bayesian networks: the combination of knowledge and statistical data[J]. Machine Learning, 1995, 20(3): 197 - 243.

[3]SPIRTES P, MEEK C, RICHARDSON T. Causal inference in the presence of latent variables and selection bias[A]. Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence [C]. Pittsburgh, USA, 1995.

[4]CHICKERING D M. Learning equivalence classes of Bayesian network structures [J]. Machine Learning, 2002, 2(3): 445 - 498.

[5]HENSON J. Comparing causality principles[J]. Studies in History and Philosophy of Modern Physics, 2005, 36(3): 519 - 543.

[6]THIESSON B, MEEK C, CHICKERING D, HECKERMAN D. Learning mixtures of Bayesian networks[R]. MSR-TR-97-30, 1997.

[7]MURPHY K P. Inference and learning in hybrid Bayesian networks[R]. CSD-98-990, 1998.

[8]MONTI S, COOPER G F. learning hybrid Bayesian networks from data[R]. ISSP-97-01, 1997.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[9] FAYYAD U, IRANI K. Mult-interval discretization of continuous-valued attributes for calssification learning [A]. Proceedings International Joint Conference on Artificial Intelligence[C]. Chambery, France, 1993.

[10] LAM W, BACCHUS F. Learning Bayesian belief networks: an approach based on the MDL principle [J]. Computational Intelligence, 1994, 10(4): 269 - 293.

[11] CHICKERING D M. Learning Bayesian networks is NP-Hard[R]. MSR-TR-94-17, 1994.

[12] 茆诗松,王静龙,濮晓龙. 高等数理统计[M]. 北京: 高等教育出版社, 1998.

[13] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721 - 742.

[14] CHOW C K, LIU C N. Approximating discrete probability distributions with dependence trees [J]. IEEE Transactions on Information Theory, 1968, 14(3): 462 - 467.

[15] BUNTINE W L. Chain graphs for learning[A]. Proceedings of the 17th Conference Artificial Intelligence [C]. San Francisco, USA, 1995.

[16] DOMINGOS P, PAZZANI M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine Learning, 1997, 29(2 - 3): 103 - 130.

[17] 王双成,苑森淼. 具有丢失数据的贝叶斯网络结构学习研究[J]. 软件学报, 2004, 15(7): 1030 - 1041.
WANG Shuangcheng, YUAN Senmiao. Research on learning Bayesian networks structure with missing data [J]. Journal of Software, 2004, 15(7): 1030 - 1041.

[18] 王双成,苑森淼. 具有丢失数据的可分解马尔科夫网络结构学习[J]. 计算机学报, 2004, 27(9): 1221 - 1228.
WANG Shuangcheng, YUAN Senmiao. Learning decomposable Markov network structure with missing data [J]. Chinese Journal of Computers, 2004, 27(9): 1221 - 1228.

[19] 王　飞,刘大有,薛万欣. 基于遗传算法的 Bayesian 网中连续变量离散化的研究[J]. 计算机学报, 2002, 25(8): 794 - 800.
WANG Fei, LIU Dayou, XUE Wanxin. Discretizing continuous variables of Bayesian networks based on genetic algorithms [J]. Chinese Journal of Computers, 2002, 25(8): 794 - 800.

[20] MURPHY S L, AHA D W. UCI repository of machine learning databases [EB/OL]. http://www.ics.uci.edu/～mlearn/ML Repository, 2005 - 09 - 10.

作者简介:

王双成,男,1958 年生,教授,博士,主要研究方向为人工智能、机器学习和数据挖掘及其在风险管理中的应用,先后承担国家自然科学基金"面向智能信息处理的贝叶斯网络关键理论与方法研究"和"面向风险管理的贝叶斯网络与集成研究"等课题,发表学术论文 40 余篇.
E-mail: wangsc@lixin.edu.cn.

李小琳,女,1978 年生,讲师,博士,主要研究方向为机器学习和数据挖掘,先后参与国家自然科学基金 2 项,发表学术论文 16 篇.
E-mail: lixl_126@126.com.

侯彩虹,女,1978 年生,讲师,博士,主要研究方向是智能控制和数据挖掘,先后参与国家自然科学基金 2 项,发表学术论文 12 篇.
E-mail: dhuhch@mail.dhu.edu.cn.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net