DOI:10.3969/j.issn.1673-4785.201112005
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120416.0843.001.html

逻辑回归分析的马尔可夫毯学习算法

郭坤,王浩,姚宏亮,李俊照
(合肥工业大学 计算机与信息学院,安徽 合肥230009)

摘 要:针对当前的马尔可夫毯学习算法会引入不正确的父子节点和配偶节点的问题,提出了一种基于逻辑回归分析的马尔可夫毯学习算法RA-MMMB.利用MMMB算法得到候选的马尔可夫毯,建立目标变量与候选马尔可夫毯的逻辑回归方程,通过回归分析在保留与目标变量相关性很强的变量的同时,去掉MMMB等算法所引入的弱相关性的错误变量以及其他的弱相关性变量;然后利用G^2测试去掉回归分析后候选马尔可夫毯中的兄弟节点,得到目标变量的马尔可夫毯.RA-MMMB算法通过回归分析,减少了条件独立测试的次数,提高了学习的精度.实验比较和分析表明,RA-MMMB算法能有效地发现变量的马尔可夫毯.

关键词:贝叶斯网络;马尔可夫毯;逻辑回归分析;条件独立测试
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2012)02-0153-08

An algorithm for a Markov blanket based on logistic regression analysis

GUO Kun, WANG Hao, YAO Hongliang, LI Junzhao
(College of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: To solve the problem of incorrect parent, child, and spouse nodes being brought into the current algorithms, an improved algorithm called a regression analysis-max min Markov blanket (RA-MMMB) was presented using the Markov Blanket based on logistic regression analysis. First, a logistic regression equation was established between the target variable and a set of its candidate Markov blankets obtained from the max-min Markov blanket (MMMB) algorithm. Regression analysis can retain the variables strongly correlated with the target variable, and can remove the error variables and other variables weakly correlated with it as well. The incorrect nodes in the MMMB algorithm were also removed from the candidate Markov blanket; then, after the G^2 conditiond independence test, which removed the brother node of the target variable in the candidate Markov blanket, returned after the regression analysis, the Markov blanket of the target variable was obtained. By the method of regression analysis, the RA-MMMB algorithm reduces the number of condition tests of independence and improves the accuracy of discovering the Markov blanket for the target variable. The result shows that the method can discover the Markov blanket of the target variable efficiently.

Keywords: Bayesian networks; Markov blanket; logistic regression analysis; conditional independence test

在给定贝叶斯网络(Bayesian networks)中一个变量的马尔可夫毯(Markov blanket)时,贝叶斯网络中其他变量与该变量条件独立,一个变量的马尔可夫毯能够屏蔽贝叶斯网络中其他变量对该变量的影响,可用来预测、分类和因果发现等.

确定目标变量的马尔可夫毯有2类方法:利用打分—搜索方法等建立贝叶斯网络结构,然后基于贝叶斯网络结构确定目标变量的马尔可夫毯,但该类方法得到的马尔可夫毯不准确,且学习方法效率低;另一类是利用局部学习的方法直接学习目标变量的马尔可夫毯.当前研究者主要采用基于局部学习的方法学习马尔可夫毯,相关工作如Margaritis和Thrun提出了GS(Grow-Shrink)算法[1],首先启发式地搜索所有与目标变量依赖的变量,然后去除冗余的变量.由于配偶节点较晚进入候选的马尔可夫毯,

收稿日期:2011-12-08. 网络出版日期:2012-04-160.
基金项目:国家自然科学基金资助项目(61070131,61175051).
通信作者:郭坤. E-mail:guokun19871005@163.com.

导致候选的马尔可夫毯中引入了较多的错误节点，降低了后面的条件独立测试的有效性和可靠性. Tsamardinos 等对 GS 进行了改进，提出了 IAMB(incremental association Markov blanket)算法[2]，每入选一个变量，就对该变量进行条件独立测试，减少了错误变量的引入；但该算法的条件独立测试是在给定整个马尔可夫毯下进行的，条件独立测试要求的数据量较大[3]. Tsamardinos 等提出的 MMMB(max-min Markov blanket)算法[4]首先利用 MMPC(max-min parents and children)算法[4]寻找目标节点的父节点和子节点，然后找到它的配偶节点，但该方法会引入错误的父子节点和配偶节点[5]. 与此相似的算法还有 Hiton-MB(Hiton-Markov blanket)算法[6]. Tsamardinos 等在贝叶斯网络结构学习算法 MMHC(max-min Hill-climbing)[7]中调用 MMPC 算法时，利用父节点与子节点对称的性质，去除 MMPC 算法引入的错误父子节点. 而 PCMB(parents-children Markov blanket)算法[5]利用完整的条件独立测试去除错误节点，但存在时间复杂度较大的问题.

针对上述算法存在引入错误节点和时间复杂度较大的问题，为了提高学习马尔可夫毯的精度和效率，在马尔可夫毯学习算法中引入回归分析[8]. 回归分析可以在发现与目标变量相关性很强的变量的同时，去掉与目标变量相关性弱或无关的变量. 回归分析广泛应用于机器学习中的特征选择，从变量集合中选取最优的特征子集[9]. 根据变量数据取值是否连续，将回归分析分为线性回归分析和逻辑回归分析[10](logistic regression analysis)2 类. 逻辑回归分析可以有效处理贝叶斯网络中的离散数据. 因此，如何让学习到的马尔可夫毯更加精确，学习过程效率更高，是马尔可夫毯学习算法的核心问题. 提出一种基于逻辑回归分析对 MMMB 算法改进的 RA-MMMB(regression analysis-max min Markov blanket)算法. 该算法对 MMMB 算法过程中的候选马尔可夫毯与目标变量进行逻辑回归分析，去掉相关性弱的变量，然后进行条件独立测试，去掉候选马尔可夫毯存在的兄弟节点，得到最终的马尔可夫毯. 本文采用文献[11]中的 G^2 测试来判断 2 个变量在给定变量集时是否条件独立，实验证明，该方法能有效地去掉 MMMB 算法包含的错误变量，并减少了条件独立测试的次数.

1 贝叶斯网络及相关定义

设 V 代表一组离散随机变量，用〈G,θ〉来表示贝叶斯网络，其中有向无环图 G 中的节点对应 V 中的变量，是指 G 中每个节点 X 在给定它的父节点 $P_a(X)$ 下的条件概率分布 $p(X|P_a(X))$. 贝叶斯网络的联合概率分布可表示如下：

$$p(V) = \prod_{X\in V}(p(X|P_a(X))).$$

定义 1 碰撞节点. 如果路径 P 中的节点 W 含有 2 条指向它的边，那么节点 W 在 P 中是碰撞节点. 在给定 W 时，它的 2 个父节点条件依赖.

定义 2 d 分离. 如果下列任意一条成立：1)路径 P 上存在一个包含于集合 Z 的非碰撞节点；2)路径 P 上的碰撞节点和它的子孙节点均不包含在 Z 中，那么称节点 X 到节点 Y 的一条路径 P 被节点集合 Z 阻塞. 当且仅当从 X 到 Y 的每条路径均被 Z 阻塞，称节点 X 和 Y 被 Z 集合 d 分离. 当有向无环图 G 和联合概率分布满足忠实性条件时，d 分离与条件独立等价. 本文中 Ind(X;$T|Z$)表示变量 T 和 X 在给定变量集 Z 时条件独立；Dep(X;$T|Z$)表示变量 T 和 X 在给定 Z 时条件依赖.

定义 3 马尔可夫毯. 一个变量 T 的马尔可夫毯 MB(T)是在给定该集合时，变量集 V 中所有其他的节点与 T 条件独立的最小集合. 即对 $\forall X \in V \setminus (\mathrm{MB}(T)\cup T)$，Ind($X$;$T|MB\cup T$)). 有向无环图 G 中的每个节点 T，它的马尔可夫毯 MB(T)包括 T 的父节点、子节点和配偶节点(与 T 共同有一个子节点).

图 1 中节点 T 的马尔可夫毯包括父节点{B, E}、子节点{C,D}和配偶节点 F.

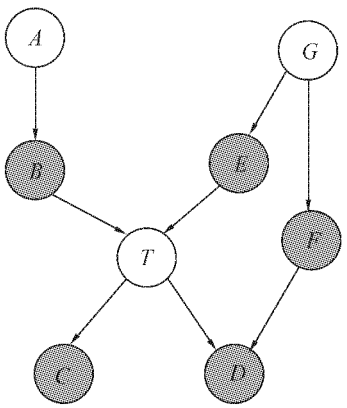


图 1 T 的马尔可夫毯(阴影节点)

Fig. 1 The Markov blanket of node T (shadow nodes)

2 MMMB 算法

2.1 MMMB 算法描述

MMMB 算法采用分治法发现目标变量 T 的马尔可夫毯 MB(T)，首先调用 MMPC 算法找到 T 的父子节点集 PC(T)，然后找到 T 的配偶节点. T 的父子节点集 PC(T)和配偶节点组成了 T 的马尔可夫毯 MB(T). MMPC 算法首先利用启发式搜索策略使与 T

相关的变量依次进入 T 的候选父子节点集 CPC,然后移去 CPC 中被错误引入的变量[11];MMMB 算法对 PC(T)中的每一个元素调用 MMPC 算法,得到 T 的候选马尔可夫毯 CMB,经过条件独立性测试,找到 T 的配偶节点. 然而,MMPC 算法会包含未去掉的 T 的错误父子节点,MMMB 算法也会引入 T 的错误的配偶节点. MMPC 算法和 MMMB 算法描述如下.

1)MMPC 算法.

输入:目标变量 T.
输出:T 的父子节点集合 PC(T).
/* 添加节点到 CPC */
CPC = ∅;
repeat;
　foreach $X \in V \backslash \text{CPC} \backslash \{T\}$;
　　assoc(X) = arg: $\min_{s \subseteq \text{CPC}} \text{Dep}(X; T \mid s)$;
　　//寻找 $s \subseteq$ CPC,使得 Dep(X;T|s)的值最小
　Y = arg: $\max_{X \in V \backslash \text{CPC} \backslash T} \text{Dep}(X; T \mid \text{assoc}(X))$;
　　//寻找 $Y \in V \backslash (\{T\} \cup \text{CPC})$,使得 Dep($Y$;$T$|assoc($Y$))的值最大
　if Dep (Y;T|assoc(Y));
　　CPC = CPC ∪ {Y};
until CPC 不再改变;
/* 从 CPC 中去掉错误的节点 */
　　foreach $X \in$ CPC;
　　if $\exists s \subseteq \text{CPC} \backslash \{X\}$,使得 Ind($X$;$T$|$s$);
　　CPC = CPC \ {X};//把 X 从 CPC 中移除
return CPC.

2)MMMB 算法描述:

输入:目标变量 T.
输出:T 的马尔可夫毯 MB(T).
/* 得到 MB(T)的候选马尔可夫毯 */
PC(T) = MMPC(T);
MB = PC(T);
CMB = PC(T) $\cup_{C \in \text{PC}(T)}$ MMPC(C) \ {T};
/* 找到 T 的配偶节点 */
foreach $X \in$ CMB \ {PC(T)};
　　寻找集合 s 使得 Ind(X;T|s);
　　foreach $Y \in$ PC(T);
　　if Dep(X;T|{Y} ∪ s);
　　将 X 标记;
if(X 有标记);
　MB = MB ∪ {X};
return MB.

2.2　MMMB 算法存在的问题

MMPC 算法去掉错误节点的依据为:如果 $X \notin$ PC(T),在给定 $Z \subseteq$ PC(T)下,X 与 T 条件独立,通过条件独立测试可以将添加到 CPC 中的错误节点去掉. 但存在有些错误节点不能被去掉,以图 2(a)为例,节点 T 的父子节点集合 PC(T) = {A},对 T 调用 MMPC 算法:

1)CPC 添加节点.

①CPC = ∅,A 与 T 邻接,Dep(A;T|∅)的值最大,节点 A 首先进入到 CPC;

②CPC = {A},路径 $T \rightarrow A \leftarrow B \rightarrow C$ 中的碰撞节点 A 包含在{A}中,该路径未被{A}阻塞,Dep(C;T|A);而 Ind(B;T|∅),节点 C 被添加到 CPC;

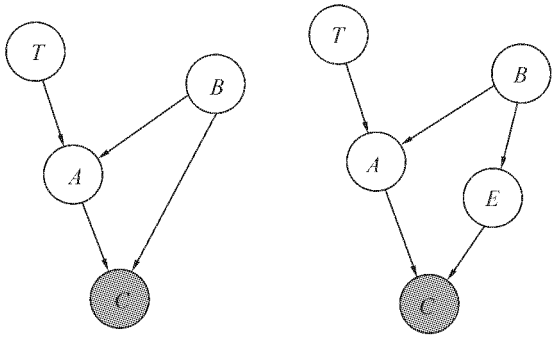
③CPC = {A,C},由于 Ind(B;T|∅),节点 B 不能被添加到 CPC.

2)CPC = {A,C}去掉错误节点.

①给定任意的集合 Z,Dep(A;T|Z),所以 A 不会从 CPC 中移除.

②由于路径 $T \rightarrow A \rightarrow C$ 中的非碰撞节点 A 并不包含在∅中,该路径未被∅阻塞,Dep(C;T|∅),且 Dep (C;T|A). 所以不存在 CPC\{C}的子集 s 使得 Ind(C;T|s),节点 C 并不能被移除,CPC = {A,C}. 但节点 C 并不在真实的 PC(T)中.

同理,图 2(b)中的节点 C 也会包含在 MMPC 算法输出的父子节点集合中.



(a)情况1　　(b)情况2

图 2　MMPC 算法引入错误节点 C

Fig. 2　Incorrect node C included in CPC returned

MMMB 算法寻找配偶节点的依据为:对 $X \in$ CMB\PC(T),$Y \in$ PC(T),如果存在集合 Z(X,$T \notin Z$),且 Ind(X;T|Z),使得 Dep(X;T|{Y} ∪ Z),那么 X 为 Y 的配偶节点. 即使 MMPC 算法输出的 CPC 为正确的 PC(T),MMMB 算法返回的 MB(T)也会包含错误

的配偶节点. 例如图 3 中节点 T 的父子节点 PC(T)为 $\{B,D\}$,候选马尔可夫毯 CMB = $\{A,B,C,D\}$. 由图 3 易知 Ind($A;T|\{B\}$),路径 $A\rightarrow C\rightarrow D\leftarrow T$ 中的碰撞节点 D 包含在集合$\{D\}\cup\{B\}$中,所以 Dep($A;T|\{D\}\cup\{B\}$) . A 被添加到马尔可夫毯中,但是实际上 A 并不在节点 T 的马尔可夫毯 MB(T)中.

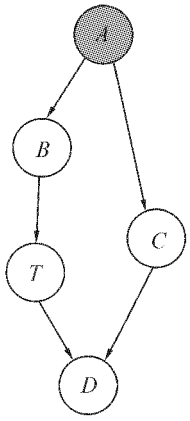


图 3 MMMB 算法引入 T 的错误的配偶节点 A

Fig. 3 Incorrect spouse node A included in MMMB

MMPC 算法添加的错误节点会包含在最终的马尔可夫毯 MB(T)中,而且这些节点会引入 T 的错误的配偶节点. 即使 MMPC 算法返回的是正确的 PC(T),MMMB 本身也会引入错误的配偶节点. 所以,为了提高学习马尔可夫毯算法的精度和效率,需要去掉这些错误的变量,而这些变量与目标变量的相关性不强.

3 RA-MMMB 算法

3.1 根据相关性将节点分类

根据贝叶斯网络中各节点与目标节点 T 的相关性关系,把 MMMB 算法中候选马尔可夫毯CMD = PC(T) $\cup_{C\in \mathrm{PC}(T)}$ MMPC(C) $\setminus\{T\}$ 中的节点分为 4 类:

1)T 的父节点和 T 的子节点:这类节点与 T 有很强的相关性;

2)T 的父节点的父节点和 T 的子节点的子节点:由于贝叶斯网络已存在 T 的父节点和 T 的子节点,故这类节点与 T 的相关性较弱;

3)T 的兄弟节点(与 T 共同有一个父节点)和 T 的配偶节点:这类节点和 T 有共同的原因或结果,当给定 T 的父节点或 T 的子节点时,与 T 的相关性较强;

4)MMPC 算法引入的错误节点的父子节点中上述 3 类以外的节点,这类节点与 T 的相关性最弱.

3.2 候选马尔可夫毯的逻辑回归分析

回归分析[12]可以从自变量集合中选入与因变量相关性强的自变量,并去掉那些与因变量无关的变量和与因变量相关性弱的次要变量. 以目标变量 T 为因变量,MMMB 算法中的候选马尔可夫毯 CMB 为自变量集合,进行回归分析,可以从 CMB 中去掉与目标变量相关性不强的变量.

3.2.1 候选马尔可夫毯的逻辑回归分析模型

一般贝叶斯网络中的数据为离散值,所以对目标变量和候选马尔可夫毯采用逻辑回归分析[11]. 当目标变量 T 为 0-1 型(取值为 2 个)因变量,CMB 为自变量集合时,二元逻辑回归模型为

$$\ln\frac{p}{1-p}=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k.$$

式中:$p=P(T=1)$,$X_1,X_2,\cdots,X_k\in$ CMB,$\beta_0,\beta_1,\cdots,\beta_k$ 为未知参数,称为回归系数. 采用极大似然估计方法得到回归系数的估计值$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$. 当因变量取值为多个(大于 2 个)时,采用多元逻辑回归. 当目标变量 T 的取值有 a、b、c 3 种,CMB 为自变量集合时,多元逻辑回归的模型为:

$$\ln\frac{p(T=b)}{P(T=a)}=\beta_1+\beta_{11}X_1+\beta_{12}X_2+\cdots+\beta_{1k}X_k,$$

$$\ln\frac{p(T=c)}{P(T=a)}=\beta_2+\beta_{21}X_1+\beta_{22}X_2+\cdots+\beta_{2k}X_k.$$

回归分析通过假设检验判断回归系数是否为零来决定是否去掉候选马尔可夫毯中的变量. 假设 H_0:$\beta_i=0,i=1,2,\cdots,k$, 逻辑回归中回归系数的检验统计量采用 Wald 统计量,即

$$\mathrm{Wald}_i=\left(\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right)^2.$$

式中:$S_{\hat{\beta}_i}$为回归系数的标准误差. Wald 统计量服从自由度为 1 的 χ^2 分布. 原假设是正确的却拒绝了该假设,犯这类错误的概率记为 p. 当概率值 p 小于给定的显著性水平 α(一般取 $\alpha=0.05$)时,则拒绝原假设,认定该回归系数不为零;反之,认定该回归系数为零,则将该变量从方程中去掉. 概率 p 值越小,表明对应的变量对目标变量 T 的影响就越显著.

3.2.2 候选马尔可夫毯的逻辑回归分析过程

采用逐步后向回归依次去掉候选马尔可夫毯 CMB 中与目标变量 T 相关性弱的变量. 如果逻辑回归方程中自变量集合存在回归系数为零的概率值 p 大于显著性水平的变量,则将回归方程中 p 值最大的变量从 CMB 中去掉,然后建立 CMB 中剩余的变量与目标变量的逻辑回归方程,继续进行回归分析,再将方程中概率值 p 最大的变量从 CMB 去掉,继续回归分析直至回归方程中不再含有 p 值大于显著性水平的变量.

由于 CMB 中的第 4)类节点与 T 的相关性最弱,所以会在逐步后向回归中最先被去掉;因为回归

方程中含有 T 的父节点和子节点,接下来与 T 的相关性较弱的第 2)类节点会作为回归方程中的次要变量从 CMB 中被去掉;第 3)类节点由于 T 的父节点和 T 的子节点的存在,与 T 的相关性较强,所以会保留在 CMB 中;而第 1)类节点与 T 的相关性最强,也包含在 CMB 中.

3.3　去除兄弟节点

经过逐步后向回归分析,最终的 CMB 中包含 T 的父节点和子节点、配偶节点和兄弟节点. 在给定 T 的父节点时,T 的兄弟节点与 T 条件独立. 对于 $X \in \mathrm{CMB}\backslash\{\mathrm{PC}(T)\}$,$Y \in \mathrm{PC}(T)$,如果 $\mathrm{Ind}(X;T|Y)$,则将 X 从 CMB 中去掉. 而 $\mathrm{CMB}\backslash\{\mathrm{PC}(T)\}$ 中不存在给定子节点时与 T 条件独立的变量,所以不会去掉马尔可夫毯中的变量. 通过条件独立测试,去掉 T 的兄弟节点,得到最终的马尔可夫毯. 如果 $\mathrm{PC}(T)$ 中的元素在逐步后向回归分析中被去掉,说明该元素为 T 的错误的父子节点,把它从候选马尔可夫毯 CMB 中去掉的同时,从 $\mathrm{PC}(T)$ 中也把它去掉,减少不必要的条件独立测试,并且避免在马尔可夫毯中引入其他错误的变量.

3.4　RA-MMMB 算法描述

基于逻辑回归分析的马尔可夫毯学习算法 RA-MMMB 描述如下:

RA-MMMB 算法:

输入:数据集 Data 和目标变量 T.

输出:T 的马尔可夫毯 $\mathrm{MB}(T)$.

/* 对候选马尔可夫毯进行逐步回归分析 */

$\mathrm{PC}(T) = \mathrm{MMPC}(\mathrm{Data}, T)$;

$\mathrm{CMB} = \mathrm{PC}(T) \cup_{C \in \mathrm{PC}(T)} \mathrm{MMPC}(\mathrm{Data}, C) \backslash \{T\}$;

repeat;

建立以 T 为因变量,CMB 为自变量集合的逻辑回归方程,进行回归分析;

$Y = \arg\max_{X \in \mathrm{CMB}} p(X)$;//寻找回归方程中 Wald 统计量 p 值最大的变量

if $P(Y) > \alpha$;　　//α 为显著性水平

$\mathrm{CMB} = \mathrm{CMB}\backslash\{Y\}$;

if $Y \in \mathrm{PC}(T)$;

$\mathrm{PC}(T) = \mathrm{PC}(T)\backslash\{Y\}$;

until $P(Y) \leqslant \alpha$;

/* 去除兄弟节点 */

foreach $X \in \mathrm{CMB}\backslash\{\mathrm{PC}(T)\}$;

foreach $Y \in \mathrm{PC}(T)$;

if $\mathrm{Ind}(X;T|Y)$;

$\mathrm{CMB} = \mathrm{CMB}\backslash\{X\}$;

return CMB.

RA-MMMB 算法运用逐步后向回归依次把 MMMB 算法中的候选马尔可夫毯 CMB 中与目标变量相关性弱的变量去掉,再经过条件独立测试,去掉兄弟节点,返回最终的马尔可夫毯. 由于 MMPC 算法引入的错误节点是目标变量的子节点的子节点,MMMB 算法引入的错误的配偶节点是目标变量的父节点的父节点,都属于上述第 2)类节点,它们会在回归分析中被去掉. RA-MMMB 算法的回归分析过程去掉与目标变量相关性弱的变量后,只需去掉回归分析后的 CMB 中的兄弟节点就能得到马尔可夫毯,与 MMMB 算法相比,减少了大量条件独立性测试,并且由于条件变量集很小,保证了条件独立测试的可靠性.

4　实验分析与算法比较

在 Matlab 7.0 和 SPSS17 的软件环境下,利用 Insurance 网(含有 27 个节点)和 Alarm 网(含有 37 个节点)的 500、1 000、5 000 组数据,对这 2 个网络中的每个节点分别使用 MMMB 算法、PCMB 算法和 RA-MMMB 算法输出它的马尔可夫毯,并进行对比. 由于实验数据为离散数据,对取值为 2 的目标变量,RA-MMMB 算法在 SPSS 软件里采用二元逻辑回归,对取值为多个(大于 2)的目标变量,采用多元逻辑回归.

4.1　评价标准

本文采用 PCMB 算法所在的文献[5]里的查准率(precision)、查全率(recall)以及它们之间的欧氏距离 d 来衡量马尔可夫毯学习算法的好坏. 对于一个目标变量 T,查准率是指算法输出的 $\mathrm{MB}(T)$ 中包含正确变量的比率;查全率是指算法输出的 $\mathrm{MB}(T)$ 中正确变量的个数占实际 $\mathrm{MB}(T)$ 变量个数的比率.

$$\mathrm{precision} = \frac{\text{算法输出的 MB}(T)\text{ 包含的正确变量个数}}{\text{算法的 MB}(T)\text{ 的变量个数}},$$

$$\mathrm{recall} = \frac{\text{算法输出的 MB}(T)\text{ 包含的正确变量个数}}{\text{实际的 MB}(T)\text{ 的变量个数}}.$$

为了对上述 2 个标准进行综合评价,定义两者之间的欧氏距离为

$$d = \sqrt{(1 - \mathrm{precision})^2 + (1 - \mathrm{recall})^2}.$$

式中:d 表明算法准确率,d 越小,表明算法准确率越高.

4.2 分析与比较

针对 Alarm 网进行分析. 如图 4 所示,图中的 $\{X_{23},X_{22},X_{29},X_{21}\}$ 和 $\{X_{23},X_{22},X_{29},X_1\}$ 均构成了图 2(a)中的结构. 节点 X_{23} 的父子节点集合 PC = $\{X_{24},X_{25},X_2,X_{22}\}$,它的马尔可夫毯 MB = $\{X_{24},X_{25},X_2,X_{22},X_{27},X_{29}\}$. 当 Alarm 网中数据集大小为5 000时,利用 MMPC 算法得到的父子节点集合 PC′ = $\{X_{24},X_{25},X_2,X_{22},X_1,X_{21}\}$. 比实际网络中节点 X_{23} 的父子节点集合多余了 X_1、X_{21} 这 2 个节点. MMMB 算法中的候选马尔可夫毯为 CMB = $\{X_{24},X_{25},X_2,X_{22},X_1,X_{21},X_4,X_{15},X_{19},X_{26},X_{27},X_{29}\}$,最终返回的马尔可夫毯为 MB′ = $\{X_{24},X_{25},X_2,X_{22},X_{27},X_{29},X_1,X_{21}\}$,比节点 X_{23} 真实的马尔可夫毯多余了 X_1 和 X_{21} 这 2 个节点,即错误的父子节点会保留在马尔可夫毯内.

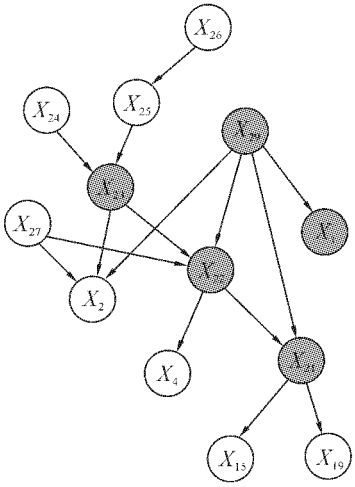


图 4 Alarm 网局部,阴影节点组成图 2(a)的结构

Fig. 4 Local Alarm network, shadow nodes form the structure in Fig. 2(a)

RA-MMMB 算法对候选马尔可夫毯 CMB 与节点 X_{23} 进行逐步后向回归,依次去掉了节点 X_{15}、X_{19}、X_1、X_{21}、X_4、X_{26},逐步后向回归去掉变量的过程如表 1 所示(左列的变量对应的概率值 p 为该变量被去掉时在回归方程中回归系数为零的概率,其余的变量对应该变量保留在最终回归方程中的 p 值). 其中最先被去掉的 2 个节点 X_{15} 和 X_{19} 是网络中节点 X_{21} 的 2 个子节点,而节点 X_{21} 是被错误引入到节点 X_{23} 的父子节点集合的节点. 接着被去掉的节点 X_1,X_{21},X_4 是节点 X_{23} 的子节点 X_{22} 的子节点,节点 X_{26} 是节点 X_{23} 的父节点 X_{25} 的父节点,剩余变量集为 $\{X_{24},X_{25},X_2,X_{22},X_{27},X_{29}\}$. 对回归分析后的剩余节点进行条件独立测试,发现这些均不是节点 X_{23} 的兄弟节点,RA-MMMB 算法返回的最终的马尔可夫毯 MB″ = $\{X_{24},X_{25},X_2,X_{22},X_{27},X_{29}\}$,跟真实的马尔可夫毯相同,去掉了 MMPC 算法引入的错误的变量 X_1 和 X_2.

表 1 节点 X_{23} 和候选马尔可夫毯 CMB 逐步后向回归过程(显著性水平 α = 0.05)

Table 1 The process of stepwise backward regression between node X_{23} and CMB (significance level α = 0.05)

节点	概率值 p	节点	概率值 p
X_{15}	0.945	X_{27}	0.035
X_{19}	0.923	X_{29}	0.030
X_1	0.749	X_2	0.025
X_{21}	0.725	X_{22}	0.022
X_4	0.631	X_{24}	0.005
X_{26}	0.482	X_{25}	0.004

以 Alarm 网中节点 X_{19} 为例,如图 5 所示,$\{X_{29},X_{21},X_{19},X_{18},X_{28}\}$ 构成了图 3 中的结构. 节点 X_{19} 的父子节点集 PC = $\{X_{20},X_{21},X_{18}\}$,它的马尔可夫毯 MB = $\{X_{20},X_{21},X_{18},X_{28}\}$. 当 Alarm 网中数据为5 000时,利用 MMPC 算法得到的父子节点集合 PC′ = $\{X_{22},X_{21},X_{18}\}$,与实际的父子节点集合相同. MMMB 算法中的候选马尔可夫毯 CMB = $\{X_{20},X_{21},X_{18},X_{14},X_{15},X_{22},X_{28},X_{29}\}$,最终返回的马尔可夫毯为 MB′ = $\{X_{20},X_{21},X_{18},X_{28},X_{29},\}$,比真实的马尔可夫毯多了节点 X_{29},即引入了节点 X_{19} 的错误的配偶节点 X_{29}.

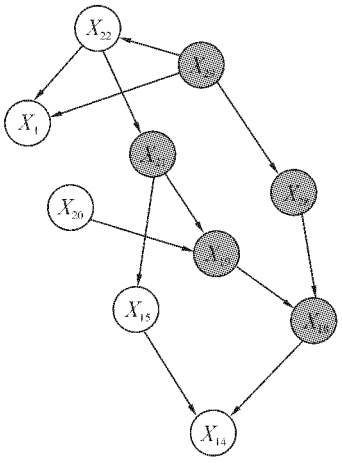


图 5 Alarm 网局部,阴影节点组成图 3 的结构

Fig. 5 Local Alarm network, shadow nodes form the structure in Fig. 3

RA-MMMB 算法对候选马尔可夫毯与节点 X_{19} 进行逐步后向回归,依次去掉了节点 X_{14}、X_{22}、X_{29},逐步后向回归去掉变量的过程如表 2 所示(左列变量含义同表 1). 其中节点 X_{14} 为节点 X_{19} 的子节点 X_{18} 的子节点,节点 X_{22} 和 X_{29} 为节点 X_{19} 的父节点 X_{21} 的父节点,剩余的变量集合为 $\{X_{20},X_{21},X_{18},X_{15},X_{28}\}$. RA-MMMB 算法接着通过条件独立性测试去掉了节点 X_{15},而节点 X_{15} 为网络中节点 X_{19} 的兄弟节

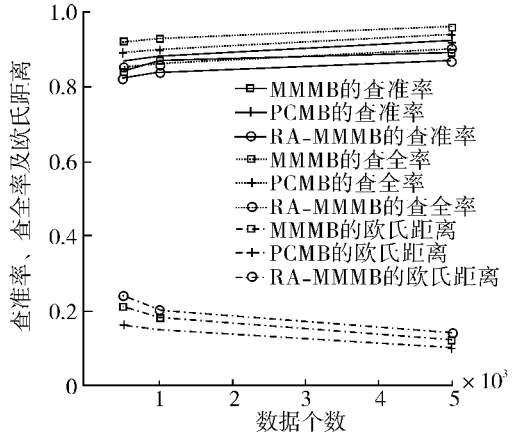
点,他们有共同的父节点 X_{21}. 得到最终的 $MB'' = \{X_{20}, X_{21}, X_{18}, X_{28}\}$,与实际马尔可夫毯相同,去掉了 MMMB 算法中引入的错误的变量 X_{29}.

表 2　节点 X_{19} 和候选马尔可夫毯 CMB 逐步后向回归过程(显著性水平 $\alpha = 0.05$)

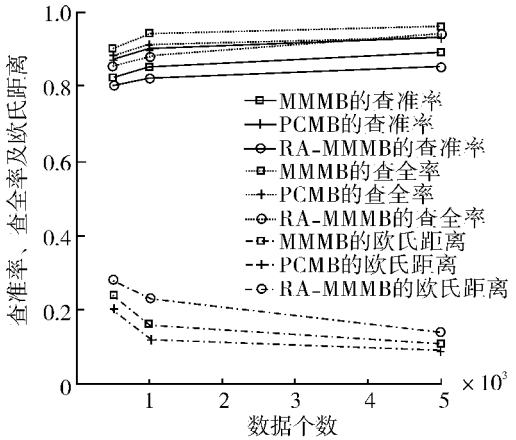
Table 2　The process of stepwise backward regression between node X_{19} and CMB(significance level $\alpha = 0.05$)

节点	概率值 p	节点	概率值 p
X_{14}	0.514	X_{28}	0.034
X_{22}	0.385	X_{18}	0.014
X_{29}	0.347	X_{20}	0.010
X_{15}	0.041	X_{21}	0.010

对 Insurance 网和 Alarm 网里各数据集的每个节点分别使用 MMMB 算法、PCMB 算法和 RA-MMMB 算法学习它的马尔可夫毯,并计算出这 3 种算法对各网络的平均查准率、平均查全率和平均欧氏距离,进行比较. 如图 6 所示.



(a) Insurance 网数据集



(b) Alarm 网数据集

图 6　Insurance 和 Alarm 数据集各算法的查准率、查全率和欧氏距离

Fig.6　Precision, recall and Euclidean distance of each algorithm in Insurance and Alarm network dataset

从图 6 可以看出,对不同的数据集,RA-MMMB 算法输出结果的查准率、查全率均比 MMMB 算法的结果高,相应的欧氏距离均比 MMMB 算法小,表明了该算法要优于 MMMB 算法;而与 PCMB 算法相比,虽然 RA-MMMB 算法查全率偏低,但查准率较高,综合评价指标欧氏距离小,体现了在整体上要优于 PCMB 算法. 同时可以看出,数据集中样本数目越多,欧氏距离就越小,算法的准确率就越高.

5　结束语

基于逻辑回归分析的马尔可夫毯学习算法,对 MMMB 算法里存在错误的父子节点和配偶节点的问题进行了分析,然后对 MMMB 算法中的候选马尔可夫毯与目标变量进行逐步后向回归,去掉了错误节点和其他与目标变量相关性弱的节点,然后进行条件独立测试去掉兄弟节点,减少了条件独立测试的次数,提高了学习马尔可夫毯的精度. 针对本算法的查全率较 PCMB 算法低的缺点,需要进一步的工作去改进.

参考文献:

[1] MARGARITIS D, THRUN S. Bayesian network induction via local neighborhoods[C]//Advances in Neural Information Processing Systems. Denver, Colorado, USA, 1999: 505-511.

[2] TSAMARDINOS I, ALIFERIS C F, STATNIKOV A. Algorithms for large scale Markov blanket discovery[C]//Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference. St Augustine, Florida, USA, 2003: 376-380.

[3] FU Shunkai, Desmarais M C. Markov blanket based feature selection: a review of past decade[C]// Proceedings of the World Congress on Engineering London, UK, 2010: 22-27.

[4] TSAMARDINOS I, ALIFERIS C F, STATNIKOV A. Time and sample efficient discovery of Markov blankets and direct causal relations[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2003: 673-678.

[5] PEÑA J M, NILSSON R, BJRKEGREN J, et al. Towards scalable and data efficient learning of Markov boundaries [J]. International Journal of Approximate Reasoning, 2007, 45(2): 211-232.

[6] ALIFERIS C F, TSAMARDINOS I, STATNIKOV A. HITON: a novel Markov blanket algorithm for optimal variable selection[C]//Proceedings of the 2003 American Medical Informatics Association Annual Symposium. Washington, DC, USA, 2003: 21-25.

[7] TSAMARDINOS I, BROWN L E, ALIFERIS C F. The

max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2006, 65: 31-78.

[8]孟晓东,袁道华,施惠丰. 基于回归模型的数据挖掘研究[J]. 计算机与现代化, 2010, 173(1): 26-28.
MENG Xiaodong, YUAN Daohua, SHI Huifeng. Research on regress-base system on data mining[J]. Computer and Modernization, 2010, 173(1): 26-28.

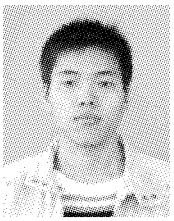
[9] SINGH S, KUBICA J, LARSEN S, et al. Parallel large scale feature selection for logistic regression[C]//SIAM International Conference on Data Mining(SDM). Sparks, Nevada, USA, 2009: 1171-1182.

[10]施朝健,张明铭. Logistic 回归模型分析[J]. 计算机辅助工程, 2005, 14(3): 74-78.
SHI Chaojian, ZHANG Mingming. Analysis of logistic regression models[J]. Computer Aided Engineering, 2005, 14(3): 74-78.

[11]高晓光,赵欢欢,任佳. 基于蚁群优化的贝叶斯网络学习[J]. 系统工程与电子技术, 2010, 32(7): 1509-1512.

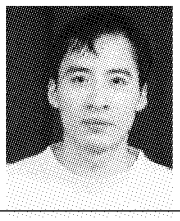
[12] SPIRTES P, GLYMOUR C, SCHEINES R. Causation, prediction, and search[M]. 2nd ed. Cambrdge, USA: The MIT Press, 2000: 23-28.
GAO Xiaoguang, ZHAO Huanhuan, REN Jia. Bayesian network learning on algorithm based on ant colony optimization[J]. Systems Engineering and Electronics, 2010, 32(7): 1509-1512.

作者简介:



郭坤,男,1987 年生,硕士研究生,主要研究方向为机器学习与数据挖掘.

王浩,男,1962 年生,教授,博士,主要研究方向为人工智能.



姚宏亮,男,1972 年生,副教授,博士,主要研究方向为机器学习与知识工程.