一种基于结构分解和因子分析的贝叶斯网络隐变量发现算法

姚宏亮 王秀芳 王 浩

(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 隐变量是观察不到或虚拟的变量,直接利用数据驱动的学习方法难以有效地发现隐变量,因而需要结合概率图结构分析的方法。针对基于结构分析的隐变量发现方法中难以确定隐变量个数和位置的问题,提出一种基于结构分解和因子分析的隐变量发现算法(S-FAHF)。S-FAHF算法利用联合树算法生成具较强依赖关系的变量子集,利用因子分析思想,通过求变量子集的特征值和累积贡献率确定变量子集中隐变量的个数,利用负荷矩阵确定隐变量的位置,最后利用打分函数测试所发现的隐变量的有效性。通过算法比较和实验结果表明,该方法能准确地确定贝叶斯网络中隐变量的个数及位置。

关键词 隐变量发现,贝叶斯网络,因子分析,BIC打分函数,S-FAHF算法

Hidden Variable Discovering Algorithm of Bayesian Networks Based on Structural Decomposition and Factor Analysis

YAO Hong-liang WANG Xiu-fang WANG Hao

(Department of Computer Science and Technology, Hefei University of Technology, Hefei 230009, China)

Abstract Hidden variables are unobservable or virtual variables, and the hidden variables cannot be effectively discovered by directly using the learning methods of data driven. The structure analysis methods are used to find hidden variables. Bcause the number and location of hidden variables are difficult to be determined, a learning algorithm(S-FAHF) of hidden variables was presented based on structural decomposition and factor analysis. The S-FAHF algorithm obtains the variables sets(Cliques) by junction tree algorithm, and the variables in a set have stronger dependence relationships. Then, the factor analysis method is inducted to discriminate the number and location of hidden variables for cliques; finally, the BIC scoring function is used to test validity of hidden variables. The results of algorithm comparison and experiment show that S-FAHF algorithm can effectively determine the number of hidden variables and their location.

Keywords Hidden variable discovering, Bayesian networks, Factor analysis, BIC scoring function, S-FAHF algorithm

贝叶斯网络(Bayesian Networks, BNs)[1]是概率论和图论相结合的产物,利用有向无环图表示随机变量之间的概率依赖关系。近年来,基于数据驱动的贝叶斯网络学习一直是一个活跃的研究领域,主要有两类学习方法:基于观察数据驱动的被动学习方法,以及联合观察数据和扰动数据的主动学习方法。基于观察数据的被动学习方法可以在数据完备,如最大似然估计方法(MLE)、K2学习算法等[2]的情况下进行网络学习;也可以在数据不完备或缺省,如EM算法、SEM算法等[3]的情况下进行网络学习。为了学习网络中的因果关系,研究者又提出了联合观察数据与扰动数据的主动学习方法,如Borchani给出了利用不完备观察数据和干扰数据进行因果关系的学习方法[4];Jianxin Yin给出了干扰情况下局部结构的因果学习方法[5]等。

然而在实际情况中会存在一些变量是不能被观察到的,但它们是实际存在的且可能包含了关于系统的重要信息;也有可能是一些虚拟的变量,能帮助研究者更好理解系统或简化模型表示的复杂性。将这些实际存在但观察不到或虚拟的变量称为隐变量(Hidden variables)。近年来,贝叶斯网络中隐变量发现问题成为机器学习领域中的一个研究热点,研究者也提了一些隐变量发现方法,如Silva等人提出学习线性结构的隐变量模型的方法[6];Yi Wang等人基于层次聚类方法构建含有隐变量的树形模型,并给出一种抽样推理算法[7]。隐变量的发现具有重要的意义,但由于隐变量是观察不到的或是虚拟的,使得直接利用基于数据驱动的被动学习和主动学习方法都不能有效地发现隐变量。

如何准确确定隐变量的个数和位置是隐变量发现研究中的关键性问题。GalElidan提出一种基于半团(semi-cliques)结构的隐变量学习方法[8],在给定的网G中寻找半团并在半团中插入隐变量。这种方法没有利用变量之间的相关性,也没有讨论隐变量存在的理论依据,且在半团中插入隐变量的个数也难以确定;在基于半团结构的隐变量发现方法的基础之上,利用了表示变量间相关性的联合树建构算法,王双成提

到稿日期:2011-03-15 返修日期:2011-05-25 本文受国家自然科学基金(61070131),国家重点基础研究发展计划(973项目)(2009CB326203)资助。

姚宏亮(1972-),男,博士,副教授,主要研究方向为机器学习与数据挖掘,E-mail:lhy_y@sohu.com;**王秀芳**(1985-),女,硕士生,主要研究方向为机器学习与数据挖掘;**王 浩**(1962-),男,博士,教授,主要研究方向为人工智能与数据挖掘。

出基于概率图模型的联合树(Junction Tree)的隐变量发现算法(JTHF)[9],JTHF算法仍然不能有效地解释团中应该插入隐变量的个数与位置。由于隐变量又可看成是对系统中某些变量产生重要影响的隐因素,因而可以考虑利用多变量统计分析中的因子分析[10](Factor Analysis,FA)方法来确定隐变量的个数与位置。

因子分析是在具有较强相关性的变量集合中提取共性因子的统计技术,可用于描述隐藏在该变量集合中的一些更基本的,但又无法直接观测到的隐性变量或隐性因素。可以直接观测的变量可能只是描述了系统的一些表象,而表象通常是由隐性变量直接决定的。隐性变量是因,而表象是果。目前,因子分析已经被广泛应用于Web文本特征提取、心理分析和教育评估等诸多领域[11]。因子分析可以避免人为确定主因子的随意性,但要求可观察变量之间有较强的相关性。

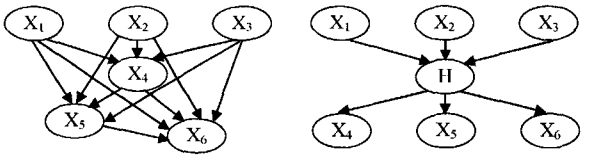
由于联合树算法是基于变量之间的依赖关系对贝叶斯网络进行分解,因此可以生成具有较强相关性的结点集;同时,在这种有较强相关性的结点集上,因子分析算法能更准确地判别隐变量的存在性,能更有效地确定隐变量的个数与位置。结合结构分解和因子分析的优缺点,提出一种发现贝叶斯网络中隐变量的个数及其位置的有效算法(S-FAHF)。利用因子分析思想,通过求变量子集的特征值和累积贡献率确定变量子集中隐变量个数,利用负荷矩阵确定隐变量的位置,利用BIC标准及数据拟合度函数测试S-FAHF算法所发现的隐变量的有效性。

1 贝叶斯网络和隐变量

贝叶斯网络(Bayesian Networks,BNs)是变量之间概率依赖关系的一种图形表示形式,结点间的边表示结点间存在概率依赖关系,且依赖的程度是一个概率参数。贝叶斯网络由网络结构与条件概率表两部分组成,其具体描述如下。

BNs是一个二元组 $B=(B_s,B_p)$。$B_s=(X,E)$是一个有向无环图,其中 $X=\{X_1,\cdots,X_n\}$为随机变量集,E是有向边集。$B_p=\{P(X_i|Pa(X_i)):X_i\in X\}$是条件概率分布集,在各结点取离散值的情况下,$B_p$为条件概率表集;$Pa(X_i)$为$X_i$的父结点集,$P(X_i|Pa(X_i))$为$X_i$的条件概率分布。BNs结构由如下的一组条件独立性假设决定:$P(X_i|X_1,\cdots,X_{i-1})=P(X_i|Pa(X_i))(i=1,\cdots,n)$。变量集$X$的联合概率分布可表示成各个局部模型的因式形式:

$$P(X)=\prod_{i=1}^{n}P(X_i|Pa(X_i)) \tag{1}$$



(a) 没有隐变量的网络　　(b) 含有隐变量H且有相同概率分布的网络

图1 贝叶斯网络与隐变量

直接基于数据驱动学习方法,难以有效地发现能揭示系统本质特性的隐变量;另外,观察数据所体现的变量之间的依赖关系比较复杂,需要大量的训练数据来建立网络结构及参数估计,这样的网络会导致数据的过度拟合,使推理过于复杂。引入隐变量能更好地解释变量之间的因果关系,且会使网络结构更为简单,网络参数减小,使知识表示和推理更加有效。如图1所示,分别表示插入隐变量前后的贝叶斯网络,插入隐变量后的贝叶斯网络具有与原贝叶斯网络同样的分布信息,且有更强的泛化能力。根据奥卡姆剃刀准则(Occam's Razor criterion)[12],具有同样分布信息的网络越简单越好。

2 基于变量依赖性的网络结构分解

由于复杂的概率图模型的联合概率分布难以计算,变量之间的依赖关系难以有效表示,因而需要分解贝叶斯网络,概率图模型的可分解性和联合树构建算法可以将相关性较强的变量聚到一起构成团(Cliques),并简化联合概率分布的计算。

下面先给出图的可分解性相关定义和性质[13]。

定义1(规范(Moral)图)　规范图是指将图中有共同孩子的所有父结点用无向边连接的图。生成规范图的操作称为规范化。

定义2(带弦图)　带弦图G是指一个无向图,图中任一长度大于3的环都是带弦的。

定义3(团,Clique)　团是指一个最大无向完备图,其中最大完备图是指图中每个结点和图中其它结点都相连。带弦图中的每个不再被其他完备子图包含的完备子图都对应一个团。

定义4(概率图模型的可分解性)　图$G(V,E)$的分解树是一个二元组$D=(S,T)$,其中$S=\{X_c|c\in C\}$是G中结点子集的集合,$T=(C,F)$是一个树,树中的结点是S中的元素,且满足以下3个条件:

(1)$\bigcup_{c\in C}X_c=V$;

(2)对于每条边$(v,w)\in E$,有一个子集$X_c\in S$包含了v和w;

(3)对于每个结点$v\in V$,结点集$\{c|v\in X_c\}$构成了T的一个联接子树。

定义5(联合树)　一个联合树(Junction Tree,JT)定义为一个二元组$T=(\Gamma,\Delta)$,其中,Γ是团结点的集合,Γ中的两个团是通过Δ中团结点相连的。对任意一对相邻团C_i、$C_j\in\Gamma$,$S_k\in\Delta$为C_i和C_j之间的一个分割团。

概率图模型G的联合树构建算法(Construction Junction Tree,CJT)可表示如下。

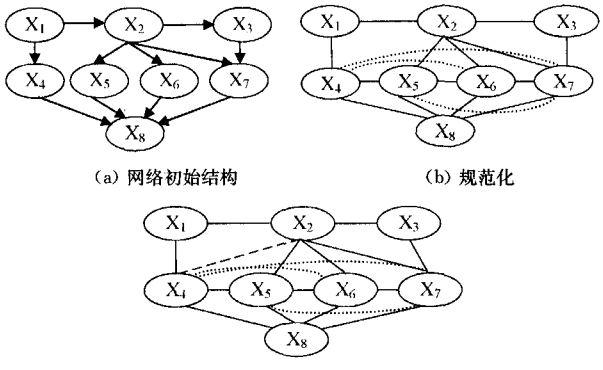
(1) 生成Moral图:道德图是将有向无环图G(如图2(a)所示)的有向边转变为无向边,并将所有具有共同子结点的节点用无向边相连,生成规范图。图2(b)是一个Moral图,图中虚线边为新添加的边,称为Moral边。

(2) Moral图的三角化:对包含4个及以上结点数的环,增加一条无向边将环中两个非相邻结点连接起来,完成对Moral图的三角化。图2(c)是对Moral图进行三角化处理后的结果,其中短划线边为新添加的边。

(3) 识别所有的团:在三角化图中,确定团节点,将具有包含关系的团合并,并计算每个团所含的结点个数,每个团都是一个完全子图且是无向图的子图。

任意一个团对$C_s\in\Gamma$,用$|C_s|$表示C_s中含有的结点个数,对于给定的正整数阈值T(T可取3、4或5,根据实际情况而定),将满足$|C_s|\geqslant T$的团输出,并考虑邻接的团是否需要合并。图2(c)是对规范图进行三角化后得到的带弦图,其中

结点个数大于等于 3 的团共有 3 个：$C_1=\{X_2,X_4,X_5,X_6,X_7,X_8\}$，$C_2=\{X_1,X_2,X_4\}$，$C_3=\{X_2,X_3,X_7\}$。



(a) 网络初始结构　　(b) 规范化

(c) 三角化

图 2　团的生成

3　因子分析模型

设有可观察的变量集 $X=\{X_1,X_2,\cdots,X_p\}$，变量集 X 对应的公共因子（或称为隐变量）集合为 $H=\{H_1,H_2,\cdots,H_m\}$。任一个变量 $X_i\in X$ 与公共因子之间的关系可表示为：

$$X_i=a_{i1}H_1+a_{i2}H_2+\cdots+a_{ij}H_j+\cdots+a_{im}H_m+U_i \quad (2)$$

式中，a_{ij} 表示变量 X_i 与公共因子 H_j 的相关系数，又称为因子负荷（factor loadings）；U_i 表示变量 X_i 与公共因子之间无关的因子（称为特殊因子），相当于统计分析中的残差，且 U_i 之间无不相关；$i=1,\cdots,p;j=1,\cdots,m$。变量集 X 与 H 之间的关系可表示成矩阵的形式：

$$X=AH+U \quad (3)$$

式中，$A=\begin{bmatrix}a_{11}&\cdots&a_{1m}\\ \vdots&\cdots&\vdots\\ a_{p1}&\cdots&a_{pm}\end{bmatrix}$，称为负荷的矩阵。

4　基于结构分解和因子分析的隐变量发现

基于结构分解和因子分析的隐变量发现算法（S-FAHF）的主要思想是：利用联合树构建算法来生成团（团中的变量之间具有较强的相关性），进而利用因子分析方法对给定团中隐变量的存在性、个数及位置进行分析。

4.1　S-FAHF 算法的主要步骤

(1)对于给定团 $C_s\in\Gamma$ 的样本进行标准化处理

为了便于计算，需要标准化处理 C_s 中变量的样本，标准化的实质是使变量的均值为 0，方差为 1。设团 C_s 中有 m 个变量，其对应的样本有 n 个，则 $C^s=\begin{bmatrix}x_{11}&\cdots&x_{1m}\\ \vdots&\cdots&\vdots\\ x_{n1}&\cdots&x_{nm}\end{bmatrix}$。

对矩阵 C^s 的列进行标准化有：

$$x_{ij}'=(x_{ij}-\bar{x_j})/\sigma_j,(1\leqslant i\leqslant n;1\leqslant j\leqslant m) \quad (4)$$

式中，$\bar{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$\sigma_j^2=\sum_{i=1}^{n}(x_{ij}-\bar{x_j})^2/n$，标准化矩阵仍记为 C^s。

(2)计算团 C_s 中变量的相关系数矩阵

相关系数矩阵描述了变量之间的相关性，有助于判断是否有必要进行因子分析。设 C_s 的相关系数矩阵为 R^s，其元素 r_{ij} 表示变量 X_i 与 X_j 的相关系数，则有 $R^s=\begin{bmatrix}r_{11}&\cdots&r_{1m}\\ \vdots&\cdots&\vdots\\ r_{m1}&\cdots&r_{mm}\end{bmatrix}$，$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x_i})(x_{kj}-\bar{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x_i})^2\sum_{k=1}^{n}(x_{kj}-\bar{x_j})^2}}$

$(1\leqslant i,j\leqslant m)$

(3)计算团 C_s 相关系数矩阵的特征值及特征向量

特征值反映了公共因子与 C_s 中变量的相关性，其值越大越好。通过特征方程 $|R^s-\lambda E|=0$，可求得特征值 $\lambda_i(1\leqslant i\leqslant m)$，并将特征值排序有 $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m$。

通过求解 $(\lambda_iE-R)X=0$，可求出对应于特征值 λ_i 的特征向量 e_i，且有 $\|e_i\|=1$，即 $\sum_{j=1}^{m}e_{ij}^2=1$，其中 e_{ij} 表示向量 e_i 的第 j 个分量。

(4) 计算团 C_s 中公共因子的贡献率和累积贡献率

定义 6(贡献率)　团 C_s 中第 j 个公共因子贡献率记为 β_j，有 $\beta_j=\frac{\lambda_j}{\sum_{i=1}^{m}\lambda_i}$。

定义 7(累积贡献率)　在 m 个公共因子中，称前 q 个主成份的贡献率之和为前 q 个主成份的累积贡献率，记为 α_q，即 $\alpha_q=\sum_{j=1}^{q}\lambda_j/\sum_{i=1}^{m}\lambda_i$，其中 $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m$。贡献率是衡量主成份重要性的关键指标，贡献率越高说明前 q 个主成份的重要性越高。

取团 C_s 累积贡献率大于 60%的前 K 主成份生成负荷矩阵 A^s，有 $A^s=\begin{bmatrix}a_{11}&\cdots&a_{1k}\\ \vdots&\cdots&\vdots\\ a_{m1}&\cdots&a_{mk}\end{bmatrix}$，其中 a_{ij} 是因子负荷，其计算公式是 $a_{ij}=\sqrt{\lambda_i}\,e_{ij}$，$(1\leqslant i\leqslant m,1\leqslant j\leqslant k)$。

(5) 确定隐变量的个数

利用特征值准则和累积贡献率准则确定隐变量的数目。特征值准则就是选取特征值大于 1 的主成份作为公共因子，而放弃特征值小于等于 1 的主成份，因为特征值小于 1 的因子其贡献会较小。累积贡献率准则，通常认为累积贡献率达到 60%才能符合要求，因为它们已经表达了绝大部分的相关信息。

将以上两个准则结合起来以确定隐变量的个数。首先是选取特征值大于 1 的成份；然后计算这些特征值大于 1 的主成分的累积贡献率，当前 K 个主成分的累积贡献率超过 60%时，停止计算并取这前 K 个主成分作为公共因子，即隐变量的个数是 K 个。

(6) 确定隐变量的位置

定义 8(公共因子方差)　因子负荷矩阵 A 中第 i 行元素的平方和称为公共因子方差，记为 $h_i^2=\sum_{j=1}^{m}a_{ij}^2$，表示变量 X_i 与所有公共因子之间的关系，其中 $1\leqslant i\leqslant p$。由于变量 X_i 的方差由公共因子方差 h_i^2 和特殊因子方差 u_i^2 两部分组成，变量 X_i 在标准化后方差为 1，即 $h_i^2+u_i^2=1$。因而，h_i^2 越大表明 X_i 对公共因子的依赖程度越大。

定义 9(方差贡献)　因子负荷矩阵 A 中第 j 列元素的平方和称为公共因子 H_j 对原始变量 X_i 的方差贡献，记为 $g_j^2=\sum_{i=1}^{p}a_{ij}^2$，其表示公共因子 H_j 对变量集 X 的总方差，是衡量某个公共因子重要性的指标，其中 $1\leqslant j\leqslant m$。

根据定义 8 和定义 9，如果一个变量在某个主因子（隐变量）上有较大的负荷，说明该变量与这个主因子有较强的相关

性，即将这个主因子(隐变量)设为其父结点。

对图 2(c)中的团 $C_1=\{X_2,X_4,X_5,X_6,X_7,X_8\}$ 进行因子分析，若发现 C_1 只含有一个隐变量 H，且 H 与团中所有的结点都有较强的相关性，那么 H 将与团中的每一个结点相连，同时原来指向团中结点的边都指向插入的隐变量 H，并将团中原有的所有边删除。将 H 插入网络后的结构如图 3 所示。

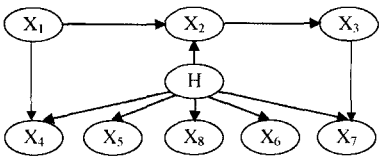


图 3　插入隐变量后的网络结构

4.2　S-FAHF 算法描述

基于结构分解和因子分析法的 S-FAHF 算法描述如下。

(1) 输入：一个贝叶斯网络结构 G_0

(2) 输出：插入隐变量后的网络结构 G

(3) CliqueList←empty list

(4) 建立网络结构图 G_0 的 *Moral* 图

(5) 对 Moral 图进行三角化

(6) 确定所有的团(Cliques)，团用 C_v 表示($v=1,2,\cdots,i$)

(7) 设 T 是阈值(T 的取值根据实际结点个数而决定)，$|C_v|$ 是团 C_v 中含有的节点个数

(8)　for $v=1$ to i do

(9)　　if $|C_v|>T$ then

(10)　　将团 C_v 输出到 CliqueList

(11)　　end if

(12)　end for

(13) 依次对 CliqueList 中的团进行因子分析，并插入隐变量

(14)　if 插入隐变量后的网络打分不比原网络有所提高 then

(15)　　取消以上做出的所有操作

(16)　end if

(17) return

5　实验

在 Matlab7.0 和 SPSS 软件环境下[14]，根据网站 http://www.norsys.com 提供的 Alarm 网和 Insurance 网生成用于实验的 1000 个数据。将 S-FAHF 算法(即因子分析方法)与文献[9]的 JTHF 算法(即依赖结构方法)、贪心爬山算法(即不插入隐变量方法)进行了比较。

5.1　评价标准

运用数据拟合度 Logloss 和 BIC 打分函数来评估所插入的隐变量的有效性，只要所插入的隐变量能够提高 BIC 打分值和数据拟合度 Logloss 的值，就保留此隐变量；否则，将其丢弃，以减少虚假隐变量的产生。

数据拟合度 Logloss 是贝叶斯网络学习算法的一种评分标准，计算公式是：

$$Logloss=\frac{1}{M}\sum_{i=1}^{M}\mathrm{Log}(X_i|B)$$

式中，M 为样本数，Logloss 是数据和结构拟合程度的度量。

BIC 打分函数在 BDe 打分的基础上增加了对复杂网络规模的惩罚，避免了过度拟合现象的产生从而可以有效地评估网络结构的质量，BIC 打分公式是：

$$BIC(B|D)=LL(B|D)-\frac{1}{2}*\log M*Dim(B)$$

式中，M 为样本数，$LL(B|D)$ 是似然对数，$Dim(B)$ 表示需要参数的个数。

5.2　实验分析与比较

针对 Alarm 网进行实验分析。利用贪心爬山算法生成无隐变量网络 G_0；然后，取结点个数最大的两个团记 $C1=\{X_2,X_{12},X_{13},X_{27},X_{30},X_{31}\}$ 及 $C2=\{X_5,X_{16},X_{30},X_{31}\}$ 进行因子分析。对团 $C1=\{X_2,X_{12},X_{13},X_{27},X_{30},X_{31}\}$ 因子分析的结果如表 1 和表 2 所列。

表 1　因子提取过程表

因子	因子提取前			提取因子旋转前			提取因子旋转后		
	特征值	贡献率%	累积贡献率%	特征值	贡献率%	累积贡献率%	特征值	贡献率%	累积贡献率%
1	2.776	46.268	46.268	2.776	46.268	46.268	2.209	36.814	36.814
2	1.183	19.714	65.982	1.183	19.714	65.982	1.664	27.741	64.555
3	1.006	16.764	82.746	1.006	16.764	82.746	1.091	18.191	82.746
4	0.600	9.997	92.743						
5	0.336	5.604	98.347						
6	0.099	1.653	100.00						

表 1 中左边第一栏为各因子(Factor)的序号，每个观察变量有一个因子，有 6 个因子。第二大栏共由 3 栏构成：特征值、贡献率和累积贡献率。在特征值栏中特征值按由大到小排列，值超过 1 的共有 3 个。贡献率和累积贡献率表示的含义可见定义 6 和定义 7。第三大栏为因子提取的结果，与第二大栏的前三行完全相同，即把特征值大于 1 的 3 个因子单独列出。特征值越大的因子与观察变量的相关性越强。第四大栏为旋转后的因子，与旋转前相比，旋转后的特征值相对集中；但旋转前、后的总特征值没有改变，最后的累计方差百分比也没有改变。由此可见，在特征值大于 1 和主因子累积贡献率大于 60%的条件下提取主因子，前两个主因子与观察变量之间有更强的相关性。

当多个变量同时在几个未旋转的因子上有较大的负荷时，将使得解释起来比较困难，因此查看旋转后的结果能较好地解决这个问题。常用的旋转法是正交旋转法(使每个因子上具有最高负荷变量数最少，即因子旋转过程中，使因子对应轴相互正交)，旋转后每列或每行的元素平方值向 0 和 1 两极分化，如表 2 所列。表 2 中各变量根据负荷量的大小进行了排列，旋转后的负荷量向 0 和 1 两极分化了。如果一个变量在某个因子上有较大的负荷，就说明可以把这个变量纳入该因子(上表中用黑体数字标出的变量分属不同的因子)。由表 2 可知，最后一个因子只有一个变量，包含的变量不多，因此删除这个因子可能更为合适。

表 2　旋转后的因子载荷矩阵

观察变量	主因子		
	1	2	3
V27	**0.892**	−0.176	0.131
V30	**0.884**	−0.276	−0.047
V31	**−0.748**	0.059	0.144
V2	−0.201	**0.896**	−0.155
V13	−0.171	**0.864**	0.294
V12	−0.037	0.057	**0.970**

对团 $\{X_5,X_{16},X_{30},X_{31}\}$ 进行因子分析后，其实验结果如表 3 和表 4 所列。

从表 3 可知，特征值大于 1 的因子只有一个，其贡献率是 43.703%。由于该结点集中只提取一个主因子，一定不会出现多个变量同时在几个未旋转的因子上有较大的负荷的情

况，因此不需要进行旋转。该主因子的因子负荷矩阵如表 4 所列，表中每一列表示一个主因子与观察变量线性组合的系数，由其系数可知表中的观察变量都与主因子有很强的相关性。

表 3　因子提取过程表

因子	因子提取前			提取因子旋转前		
	特征值	贡献率%	累积贡献率%	特征值	贡献率%	累积贡献率%
1	1.748	43.703	43.703	1.748	43.703	43.703
2	1.000	24.995	68.698			
3	0.981	24.520	93.219			
4	0.271	6.781	100.000			

表 4　因子载荷矩阵

观察变量	主因子
	1
V31	0.929
V30	−0.652
V5	0.502
V16	−0.456

因而可知，Alarm 网的团 $C1$ 和 $C2$ 中共含有 3 个隐变量。在结点集 $\{X_5, X_{16}, X_{30}, X_{31}\}$ 中插入隐变量 $H1$，在结点集 $\{X_{27}, X_{30}, X_{31}\}$ 中插入隐变量 $H2$，在结点集 $\{X_2, X_{13}\}$ 中插入隐变量 $H3$。文献[9]的 JTHF 算法在团 $C1$ 和 $C2$ 中各含有一个隐变量。

下面将基于因子分析方法（即 S-FAHF 算法）分别与不插入隐变量方法（即贪心爬山算法）、基于依赖结构方法（即文献[9]的 JTHF 算法）进行实验比较。实验针对不同样本规模，并分别利用数据拟合度 Logloss 和 BIC 打分对算法的有效性进行验证。

对最大团中含有的隐变量情况，及整个网络含有的隐变量情况分别做了实验。在 Alarm 网中，隐变量插入网络后和未插入隐变量时网络结构的数据拟合度（值越高越好）情况如图 4 所示。

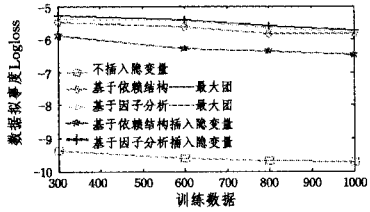


图 4　Alarm 网插入和未插入隐变量的数据拟合度情况比较

在 Insurance 网中具有不同数量的数据集中进行具有隐变量的学习，将隐变量插入网络后和未插入隐变量时网络的数据拟合度情况如图 5 所示。

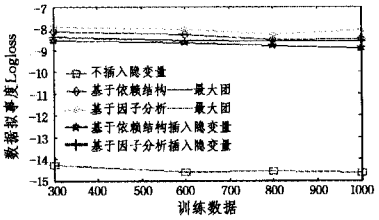


图 5　Insurce 网插入和未插入隐变量的数据拟合度情况比较

由图 4 和图 5 可知，对于不同数量的数据集基于因子分析方法（即 S-FAHF 算法）均优于基于依赖结构的方法（即 JTHF 算法），导致差距的主要原因是基于依赖结构的方法未能准确确定隐变量的个数，导致网络信息丢失从而使网络结构与数据集拟合得不够好。

将隐变量插入 Alarm 网络后，其 BIC 函数打分情况如图 6 所示。

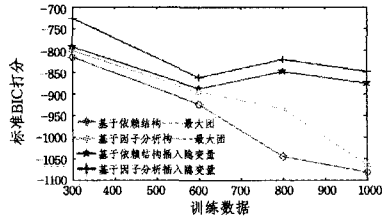


图 6　Alarm 网基于依赖结构与基于因子分析法插入隐变量后 BIC 打分的情况比较

将隐变量插入 Insurance 网后，其 BIC 打分情况如图 7 所示。

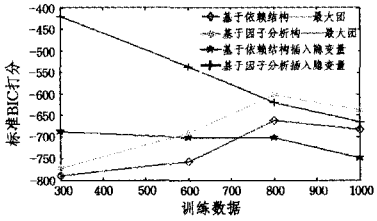


图 7　Insurance 网基于依赖结构与基于因子分析法插入隐变量后 BIC 打分的情况比较

由图 6 和图 7 可知，样本数越多 BIC 打分值越小，产生这种现象的原因是基于本文提出的因子分析方法（即 S-FAHF 算法）通过分析在网络中插入 3 个隐变量；而基于文献[9]的 JTHF 算法未进行分析便认为每个团中仅含一个隐变量，即网络中共插入两个隐变量。因此，基于本文提出的方法构建的网络结构要比基于文献[9]方法构建的网络复杂，但 BIC 打分函数适合对结构相对简单的网络进行打分。

结束语　本文提出了一种基于结构分解和因子分析的隐变量发现算法（S-FAHF），首先基于贪心爬山算法通过训练数据集建立初始网络结构；然后，运用分割团的方法找出依赖性较强的团，再用因子分析方法探索每个团中含有的隐变量个数及确定其位置；最后，依据 *BIC* 打分函数和数据拟合度 Logloss 评价所插入的隐变量的有效性。实验结果证明，该方法有效减少了虚假隐变量的产生，并能够准确确定隐变量的位置，同时也为未来的研究提供了一个很好的思路。

参 考 文 献

[1] Pearl J. Probabitistic Reasoning in Intelligent Systems: Networks of Plausible Inference[M]. San Mateo, USA: Morgan Kauf—mann, 1988: 383-408

[2] Eaton D, Murphy K. Exact Bayesian structure learning from uncertain interventions[C]// AI & Statistics. 2007: 107-114

[3] Ordonez C, Omiecinski E. Accelerating EM clustering to find high-quality solutions[J]. Knowledge and Information Systems, 2005, 7: 135-157

[4] Borchani H, Chaouachi M. Learning Causal Bayesian Networks from Incomplete Observational Data and Interventions[C]// Symbolic and Quantitative Approaches to Reasoning with Uncertainty 9th European Conference. 2007, 4724: 17-29

[5] Yin J, He Y Z. Partial orientation and local structural learning of

causal networks for prediction[C]// JMLR W&CP, WCCI2008 workshop on causality. Hong Kong, 2008, 3: 93-104

[6] Silva R, Scheines R. Learning the Structure of Linear Latent Variable Models[J]. Journal of Machine Learning Research, 2006, 7(2): 191-246

[7] Wang Yi, Zhang N L, Chen Tao. Latent Tree Models and Approximate Inference in Bayesian Networks[Z]. AAAI, 2008: 879-900

[8] Elidan G, Lotner N, Friedman N, et al. Discovering Hidden Variables: A Structure-based Approach[C]// Advances in Neural Information Processing System 13. MIT Press, 2000: 479-485

[9] 王双成. 混合贝叶斯网络隐藏变量学习研究[J]. 计算机学报, 2005, 28(9): 1564-1569

[10] 吴德辉, 李辉, 刘青松, 等. 基于因子分析信道失配补偿的 SVM 话者确认方法[J]. 人工智能与模式识别, 2010, 23(1): 59-64

[11] Yin Shou-chun, Rose R, Kenny P. A Joint Factor Analysis Approach to Progressive Model Adaptation in Text Independent Speaker Verification[J]. IEEE Trans on Audio, Speech and Language Processing, 2007, 15(7): 1999-2010

[12] Domingos P. The role of Occam's razor in knowledge discovery [J]. Journal of Data Mining and Knowledge Discovery, 1999, 3 (4): 409-425

[13] Kjaerulff U. Reduction of computational complexity in Bayesian networks through removal of weak dependences[C]// UAI-94. 1994: 374-382

[14] www. ai. nit. edu/murphyk/Software/BNT/bnt. html

(上接第 226 页)

分析：公理 29 包含公理 30，因此公理库中存在包含性。

4.4.2 验证方法

依次取出两条公理进行检查，看其中一条公理是否包含另一条公理。

4.5 公理的潜在不一致性

4.5.1 定义

公理的潜在不一致性定义：若公理库中至少存在这样一条公理，它与公理库中的其它公理之间存在矛盾，则公理库存在潜在不一致性。

公理 31 领导(s1,a1) & 员工(s1,a2) ——> 领导员工(s1,a1,a2)

公理 32 领导(s1,a1) & 员工(s1,a2) ——> ～领导员工(s1,a1,a2)

分析：上面两条公理的结论是互相矛盾的，因此公理库中存在潜在不一致性。

4.5.2 验证方法

采用归结法进行公理的潜在不一致性分析。步骤如下：

(1)由于公理是用一阶谓词逻辑表示的，可以把公理库中的公理经过一系列变换形成子句集 S。

(2)采用一定的归结策略，对 S 中可归结的子句做归结。

(3)若归结产生空子句，则进入步骤(4)，若没有产生空子句，将归结式加入子句集 S 中，返回步骤(2)。

(4)得到空子句，说明其中一条公理与其它公理之间存在矛盾，即原公理库中存在公理的潜在不一致性。

结束语 本课题一方面由于我们的公理完全是对社会群体角色知识的表示，因此保证了公理库与知识库的一致性；另一方面，对获取的公理进行了验证，保证了公理之间没有矛盾，从而保证了公理之间的一致性。但如何保证一致性的公理是完备的？我们只能保证特定群体中的公理完备性，因为不可能将所有群体中的公理都加入公理库。

在对公理进行验证时，是针对已总结出的公理类型的，但可能有的公理类型尚未总结出来，因此不能保证发现所有的错误。角色关系公理的获取限定在一个个独立的群体里面，如果把几个群体的角色融合在一起，角色关系公理也将发生一定的变化。

如果采用动态描述逻辑来表示引入角色动作之后的角色关系公理，公理将发生哪些变化？

本课题只选择了 300 多个有代表性的社会群体来构建社会群体角色本体，实际上可以增加更多的社会群体来丰富本体。本文只讲述了群体、角色、角色关系、角色关系公理的获取，而社会群体角色包含的其它内容，如情绪、动作、信念、期望、意图等，都是需要研究的，若能把这些内容作为本体元素加入到所建立的社会群体角色本体之中，则社会群体角色本体的定义就更加明确了。这都是下一步需要做的工作。

参 考 文 献

[1] 王智明, 杨旭, 平海涛. 知识工程及专家系统[M]. 北京：化学工业出版社, 2006: 12-20

[2] 陆汝钤. 世纪之交的知识工程与知识科学[M]. 北京：清华大学出版社, 2001: 9-190

[3] Lu R Q. New Approaches to Knowledge Acquisition[M]. Singapore: World Scientific Publishers, 1992: 50-78

[4] 曹存根. 面向专家的知识获取[M]. 北京：科学出版社, 1998: 20-56

[5] Cao Cun-gen. Extracting and Sharing Knowledge from Medical Texts[J]. Journal of Computer Science and Technology, 2002, 17(3): 295-303

[6] 曹存根, 李良君, 等. 智能动画创作系统 PNAI 的研究进展[J]. 系统科学与数学, 2008, 28(11): 1407-1431

[7] 高茂庭, 王正欧. ontology 及其应用[J]. 计算机应用, 2003, 23 (Z2): 31-34

[8] Wu Ping, Su S Y W. Rule Validation On Based On Logical Deduction[C]// Proceedings of the Second International Conference on Information and Knowledge Management(CIKM '93). New York: ACM, 1993: 164-174

[9] Loebe F. An analysis of roles: towards ontology-based modeling [D]. Leipzig: University of Leipzig, 2003

[10] Odell J, Noding M. A Metamodel for Agents, Roles, and Groups [C]// Proceedings of the Agent Oriented Software Engineering Workshop, 2004. Berlin: Springer-Verlag, 2005: 78-92

[11] 杜小勇, 李曼, 王珊. 本体学习研究综述[J]. 软件学报, 2006, 17 (9): 1837-1847

[12] 顾芳. 多学科领域本体的设计方法研究[D]. 北京：中国科学院研究生院, 2004

[13] 陆钟万. 面向计算机科学的数理逻辑[M]. 北京：科学出版社, 1998: 12-135

[14] 丛晓青. 一种动态描述逻辑及其应用[D]. 北京：中国科学院研究生院, 2008