多维贝叶斯网络分类器加速学习算法

傅顺开 李志强 Sein Minn

(华侨大学计算机科学与技术学院 厦门 361021)

摘 要 作为概率图模型,无限制多维贝叶斯网络分类器(GMBNC)是贝叶斯网络(BN)应用在多维分类应用时的精简模型,只包含对预测有效的局部结构。为了获得GMBNC,传统方法是先学习全局BN;为了避免全局搜索,提出了仅执行局部搜索的结构学习算法DOS-GMBNC。该算法继承了之前提出的IPC-GMBNC算法的主体框架,基于进一步挖掘的结构拓扑信息来动态调整搜索次序,以避免执行无效用的计算。实验研究验证了DOS-GMBNC算法的效果和效率:(1)该算法输出的网络质量与IPC-GMBNC一致,优于经典的PC算法;(2)在一个包含100个节点的问题中,该算法相对于PC和IPC-GMBNC算法分别节省了近89%和45%的计算量。

关键词 多维分类,贝叶斯网络,多维贝叶斯网络分类器,马尔科夫毯

中图法分类号 TP181,TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.055

Accelerated Structure Learning for General Multi-dimensional Bayesian Network Classifier

FU Shun-kai LI Zhi-qiang Sein Minn

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract General multi-dimensional Bayesian network classifier (GMBNC) is one kind of Bayesian network (BN) tailored for the application of multi-dimensional classification, hence it contains only features necessary for the prediction. To avoid global search, a novel algorithm called DOS-GMBNC was proposed. It inherits the framework of existing IPC-GMBNC, conducts a dynamic order of search by making use of the underlying topology information. Experimental studies indicate the effectiveness and efficiency of DOS-GMBNC. It outputs networks with equal quality as PC and iPC-GMBNC algorithms, and it brings considerable reduction of computation complexity, e.g. about 89% and 45% less than PC and IPC-GMBNC respectively on a 100-node network problem.

Keywords Multi-dimensional classification, Bayesian network, Multi-dimensional Bayesian network classifier, Markov blanket

1 引言

贝叶斯网络(Bayesian network,BN)(见图1(a))是一种经典概率图模型,能紧凑和直观地表示变量间的相互关系,并提供强大的推理能力。BN在诸多领域获得了应用,例如无人机决策[1]和反应堆补水系统诊断[2]。

应用BN进行建模和推理,需要先获得其模型,常见的有两种途径:(1)由领域专家利用个人专业知识经验进行人工处理,一般限于小规模问题;(2)基于专门的机器学习算法从历史样本数据集自动恢复模型。途径(2)是目前的主流方式[3],不过已被证明为NP困难问题[4],常见的学习算法只能支持数十个变量规模的应用。

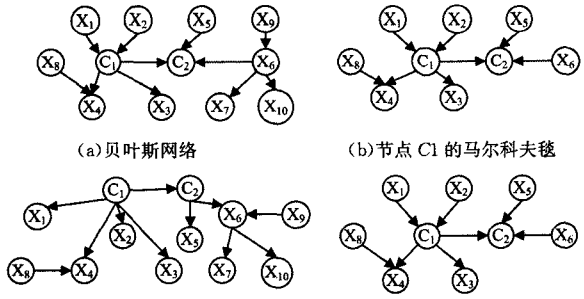
分类是机器学习和数据挖掘的基本任务,同时属于推理任务。将BN应用于单维分类时,许多经典的简化模型被提出,以降低学习代价,例如朴素贝叶斯及其各种增强版本[5,6]。虽然它们所依赖的假设在现实中不成立,但这些简化的贝叶斯网络分类器(Bayesian network classifier,BNC)取得了惊人的成功[5-7]。

BN被van der Gaag和de Waal等在2006年应用到多维分类问题[8],相应的模型取名为多维贝叶斯网分类器(Multi-dimensional Bayesian network classifier,MBNC)(见图1(c))。为了避免高昂的学习代价,MBNC被定义为两偶图(bi-partite graph),即可分解为3个独立的子图:类子图(Class sub-graph)、特征子图(Feature sub-graph)以及连接它们的桥接子图(Bridge sub-graph)。根据对类和特征子图的不同假设,MBNC又可进一步区分为〈empty〉-〈empty〉、〈polytree〉-〈polytree〉[9]、〈tree〉-〈tree〉[8]和〈DAG〉-〈DAG〉[10]等。MBNC的结构学习相应地可分解为分别独立学习这3部分子图,例如基于经典的Chow-Liu算法学习类子图和特征子图,而特征(子图)选择普遍基于wrapper或filter方法[8-10,12,18]。文献[11]首次提出基于马尔科夫毯的推导实现特征选择,从而可以有效完成特征子图的学习。然而其所依赖的学习马尔科夫毯的HITON-PC/MB算法在理论上并不正确[14]。尽管两偶图的假设无法让MBNC实现对联合分布

到稿日期:2014-07-22 返修日期:2014-10-08 本文受国家自然基金项目(61305058,61300139),福建省自然科学基金(2014J05074),中央高校基本科研基金(11J0263),厦门科技计划基金资助项目(3505Z20133027),华侨大学科研基金(11Y0274)资助。

傅顺开(1978—),男,博士,讲师,主要研究方向为数据挖掘及其应用,E-mail:fusk@hqu.edu.cn;李志强(1990—),男,硕士生,主要研究方向为概率图模型;Sein Minn(1990—),男,硕士生,主要研究方向为贝叶斯网络及其应用。

的准确建模，文献[11,12]的实验研究显示，MBNC 的分类准确率并不逊于甚至优于常见的多维分类器，例如 ML-kNN[13]、BP-MLL[17]等。



(a)贝叶斯网络　(b)节点 C1 的马尔科夫毯

(c)对应 BN 的一种可能的 MBNC　(d)对应 BN 的 GMBNC

图 1

MBNC 与生俱来的结构约束使其无法对联合分布准确建模，例如其限制类变量的父节点只可以是其他类节点。文献[14]提出的无限制多维贝叶斯网络分类器(General MBNC，GMBNC)只要求是合法的有向无环图(DAG)(见图 1(d))。GMBNC 本质上是全局 BN 关于目标变量以及其马尔科夫毯(定义 3)之上的局部 DAG。如果已知全局 BN，那么读取相关的局部 DAG 是一件容易的任务，但这将受限于现有结构学习算法所能对付的 BN 的规模。傅顺开等提出了一种仅执行局部搜索的 GMBNC 结构学习算法 IPC-GMBNC[14]，它可以在不牺牲学习效果的前提下节省可观的计算量。例如，在一个包含 200 个节点的实验中，IPC-GMBNC 比传统的执行全局搜索的 PC 算法[15]节省了高达 94%的计算量。

本文在 IPC-GMBNC 的基础上提出了新的加速学习算法 DOS-GMBNC(Dynamic Order based Search for GMBNC)。同 IPC-GMBNC，该算法仍执行局部搜索，但通过动态调整搜索计算次序来实现，计算效率显著提高。

2 理论基础

2.1 符号表示

采用粗体罗马字母表示变量集合(比如 $\boldsymbol{X}$)，非粗体的罗马字母表示单个的变量(比如 X)，小写罗马字母表示变量值(比如 $\boldsymbol{X}=\boldsymbol{x}$)。对于一个分布 ρ，采用 $\boldsymbol{X}\perp_\rho\boldsymbol{Y}|\boldsymbol{Z}$ 来表示给定条件集 $\boldsymbol{Z}$ 下，$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 是相互独立的。若条件集 $\boldsymbol{Z}=\emptyset$，用 $\boldsymbol{X}\perp_\rho\boldsymbol{Y}$ 来表示。为了简化表示，当考虑单个变量时，忽略标准的集合表示。比如，采用 $X\perp_\rho Y|Z$ 表示 $\{X\}\perp_\rho\{Y\}|\{Z\}$。类似地，使用 $\boldsymbol{X}\perp_G\boldsymbol{Y}|\boldsymbol{Z}$ 来表示在概率图模型 G 中的条件独立关系。

2.2 贝叶斯网络

无论是 MBNC 还是更一般的 GMBNC，其首先都是贝叶斯网络。给定问题域 $\boldsymbol{U}=\{X_1,\cdots,X_n\}$ 上的联合分布 ρ，对应的 BN 是一个二元组，$B=\langle G,\boldsymbol{\Theta}\rangle$；$G=\langle\boldsymbol{V},\boldsymbol{A}\rangle$ 是一个有向无环图，$\boldsymbol{V}(=\boldsymbol{U})$ 和 $\boldsymbol{A}$ 分别表示节点和边集合，包含了 BN 的定性和直观描述。对于任意的节点 $X\in\boldsymbol{V}$，与其关系紧密的父、子和配偶节点分别记为 $\boldsymbol{Pa}(X)$、$\boldsymbol{Ch}(X)$ 和 $\boldsymbol{Sp}(X)$。例如图 1(a)中的 $\boldsymbol{Pa}(C_1)=\{X_1,X_2\}$，$\boldsymbol{Ch}(C_1)=\{X_3,X_4,C_2\}$，$\boldsymbol{Sp}(C_1)=\{X_5,X_6,X_8\}$。$\boldsymbol{\Theta}$ 为参数集合，包含了诸如 $\boldsymbol{\Theta}_{x|\boldsymbol{Pa}(x)}=p(x|\boldsymbol{Pa}(x))$ 这样的条件概率分布，它定量地给出当父节点集合取值为 $\boldsymbol{Pa}(x)$ 时，它们的共同子节点 X 取 x 的条件概率。若 $\boldsymbol{Pa}(X)=\emptyset$，参数为先验概率 $p(x)$。

一个 BN 所表示的联合分布可因式分解为：

$$p(X_1,\cdots,X_n)=\prod_{i=1}^{n}p(X_i|\boldsymbol{Pa}(X_i)) \tag{1}$$

定义 1(忠实，Faithfulness)　一个贝叶斯网络 B 和一个联合分布 ρ 是相互忠实的，当且仅当任意的 $\boldsymbol{X}\perp_G\boldsymbol{Y}|\boldsymbol{Z}$ 有相应的 $\boldsymbol{X}\perp_\rho\boldsymbol{Y}|\boldsymbol{Z}$。

定义 2(d-分割，dependence-separation)　有向图 G 中的条件独立关系 $X\perp_G Y|\boldsymbol{Z}$ 表示节点 X 和 Y 被 $\boldsymbol{Z}$ 所 d-分割，相应的 $\boldsymbol{Z}$ 称作关于 X 和 Y 的 d-分割集。

当忠实性满足时，d-分割概念桥接了概率视图和图视图下的条件独立关系，即 $\boldsymbol{X}\perp_\rho\boldsymbol{Y}|\boldsymbol{Z}$ 同 $\boldsymbol{X}\perp_G\boldsymbol{Y}|\boldsymbol{Z}$ 是一一对应的。d-分割是 DOS-GMBNC 类算法的核心基础，即只要能找到任意节点对 X 和 Y 的 d-分割集，即可判断它们非相邻(也就是非直接连接)。

定义 3(马尔科夫毯)　一个变量 $X\in\boldsymbol{U}$ 的马尔科夫毯标记为 $\boldsymbol{MB}_X$，它是满足以下条件的变量集合：

$$\forall Y\in\boldsymbol{U}\backslash\boldsymbol{MB}_X\backslash\{X\}\quad X\perp_\rho Y|\boldsymbol{MB}_X \tag{2}$$

马尔科夫毯不是唯一的，而满足式(2)的最小马尔科夫毯称作马尔科夫边界(Markov boundary)。

定理 1　给定忠实性假设，贝叶斯网络 B 中的节点 X 满足：(1)$\boldsymbol{MB}_X$ 是唯一的，且 $\boldsymbol{MB}_X=\boldsymbol{Pa}(X)\cup\boldsymbol{Ch}(X)\cup\boldsymbol{Sp}(X)$；(2)$\forall Y\in\boldsymbol{V}\backslash\boldsymbol{MB}_X\backslash\{X\}$，$X\perp_G Y|\boldsymbol{MB}_X$；(3)$X\perp_G\boldsymbol{V}\backslash\boldsymbol{MB}_X\backslash\{X\}|\boldsymbol{MB}_X$。

可见，忠实假设统一了马尔科夫毯的概率和拓扑特征。图 1(b)是关于图 1(a)中 C_1 的马尔科夫毯(节点)。根据 d-分割概念，$\boldsymbol{MB}_X$ 是 X 和非马尔科夫毯节点集 $\boldsymbol{V}\backslash\boldsymbol{MB}_X\backslash\{X\}$ 之间的最小 d-分割集。但对于 $\forall Y\in\boldsymbol{U}\backslash\boldsymbol{MB}_X\backslash\{X\}$，$\boldsymbol{MB}_X$ 在多数情况下并非满足条件的最小 d-分割集，而 DOS-GMBNC 算法中 *RmvFalsePositives* 正是被设计为找到尽可能小的 d-分割集，不仅有利于提高计算效率，更有助于保证统计学习的可靠性。

2.3 限制和无限制多维贝叶斯网络分类器

定义 4(多维贝叶斯网络分类器，MBNC)[8]　MBNC 属于贝叶斯网络 $B=\langle G,\boldsymbol{\Theta}\rangle$，其中 $G=\langle\boldsymbol{V},\boldsymbol{A}\rangle$ 除了要满足 DAG 的基本约束，还包含以下限制：(1)$\boldsymbol{V}=\boldsymbol{V}_C\cup\boldsymbol{V}_X$，其中 $\boldsymbol{V}_C$ 包含所有的类节点，$\boldsymbol{V}_X$ 包含所有的特征节点，并且 $\boldsymbol{V}_C\cap\boldsymbol{V}_X=\emptyset$；(2)$\boldsymbol{A}=\boldsymbol{A}_C\cup\boldsymbol{A}_X\cup\boldsymbol{A}_{CX}$，其中：

- $\boldsymbol{A}_C\subseteq\boldsymbol{V}_C\times\boldsymbol{V}_C$，对应的子图称为类子图，记为 G_C；
- $\boldsymbol{A}_X\subseteq\boldsymbol{V}_X\times\boldsymbol{V}_X$，对应的子图称为特征子图，记为 G_X；
- $\boldsymbol{A}_{CX}\subseteq\boldsymbol{V}_C\times\boldsymbol{V}_X$，对应的子图称为桥接子图，记为 G_{CX}。

注：$\boldsymbol{A}_{CX}$ 只包含诸如 $C\rightarrow X$ 的边。

由于 MBNC 的额外约束，可称其为限制多维贝叶斯网络分类器，图 1(c)是 MBNC 的一个例子。然而，以图 1(a)中的贝叶斯网络为例，假设我们希望预测 C_1 和 C_2 最大可能的取值组合，即求解满足以下最大后验概率的 c_1 和 c_2：

$$\operatorname*{argmax}_{c_1,c_2}p(c_1,c_2|\boldsymbol{x}) \tag{3}$$

根据式(1)的分解规则和图 1(a)中 BN 的实际拓扑结构，式(3)可进一步分解为：

$$\begin{aligned}\operatorname*{argmax}_{c_1,c_2}p(c_1,c_2|\boldsymbol{x})&\propto\operatorname*{argmax}_{c_1,c_2}p(c_1,c_2,\boldsymbol{x})\\&\propto\operatorname*{argmax}_{c_1,c_2}p(x_1)p(x_2)p(x_3|c_1)p(c_1|x_1,x_2)p(x_8)\cdot\\&\quad p(x_4|c_1,x_8)p(c_2|c_1,x_5,x_6)p(x_5)p(x_9)p(x_6|x_9)\cdot\\&\quad p(x_7|x_6)p(x_{10}|x_6)\end{aligned} \tag{4}$$

观察式(4)可发现包含 c_1 和 c_2 的项有 $p(x_3|c_1)$、$p(c_1|x_1,x_2)$、$p(x_4|c_1,x_8)$ 和 $p(c_2|c_1,x_5,x_6)$，相关特征节点有 $\{X_1,X_2,X_3,X_4,X_5,X_6,X_8\}$。若从式(4)中删除其余 3 个特

征节点 X_7、X_9 和 X_{10} 相关的项，对预测计算没有影响；相应从图 1(a)中删除这 3 个节点，剩余的网络(图 1(d))保留原来的分类建模能力，而相应的结构即所谓的无限制贝叶斯网络分类器。

定义 5(无限制多维贝叶斯网络分类器，GMBNC)[14] GMBNC 是一种贝叶斯网络 $B=\langle G,\Theta\rangle$，其中 $G=\langle \boldsymbol{V},\boldsymbol{A}\rangle$：

· $\boldsymbol{V}=\bigcup_{C\in\boldsymbol{V}_C}(\{C\}\cup\boldsymbol{MB}_C)$，即特征节点只包含类节点的马尔科夫毯(其中所包含的)节点；

· $\boldsymbol{A}_X\subseteq\boldsymbol{V}\times\boldsymbol{V}$，即允许一般的关联，只要求不存在有向无环边。

GMBNC 和传统的 MBNC 在结构上存在两个显著区别：(1)GMBNC 并不包含所有特征节点；(2)GMBNC 允许存在诸如 $X\rightarrow C$ 的边。事实上，GMBNC 只是关于 $\bigcup_{C\in\boldsymbol{V}_C}(\{C\}\cup\boldsymbol{MB}_C)$ 的局部贝叶斯网络。如果已知 $\boldsymbol{U}$ 上的 BN 和 $\boldsymbol{V}_C$，则可以轻易"读取"相关的 GMBNC，但 BN 高昂的学习代价将限制应用的规模。考虑到一般情况下 $\bigcup_{C\in\boldsymbol{V}_C}(\{C\}\cup\boldsymbol{MB}_C)\subset\boldsymbol{U}$，理想的学习算法是将搜索区域局限在 $\bigcup_{C\in\boldsymbol{V}_C}(\{C\}\cup\boldsymbol{MB}_C)$ 上，这是本文工作的重点。

3 动态搜索学习算法

3.1 结构学习

关于 BN 的结构学习研究持续了近 30 年，涌现出来的算法可分成 3 类：基于约束(Constraint based，CB)的学习算法、基于评分搜索(Scoring and searching，S&S)的学习算法和混合学习算法。CB 学习算法认为结构学习是约束满足问题，通过检验条件独立性(Conditional independence，CI)来构建结构。此类算法的典型代表有 SGS 和 PC 算法，而后者的学习效率明显优于前者[15]。但是 PC 算法只能输出全局 BN，之后再进行抽取获得目标 GMBNC，故该途径可称为基于全局搜索的学习。为了避免全局搜索，诸多 MBNC 推导算法已被提出，目前已知的成果普遍属于 CB 类学习算法[8-12,18]。

IPC-GMBNC 算法[14]是首个专门针对学习 GMBNC 而提出的，它同 PC 属于 CB 类算法，即依赖一系列的 CI 统计测试。与 PC 算法不同的是，IPC-GMBNC 基于 GMBNC 的拓扑信息设计了精妙的广度优先搜索策略，并进一步限制搜索的深度，实现了"只学习我们需要学习的"。

S&S 类算法虽然在贝叶斯网络结构学习中获得了更大的成功，但鲜有应用在多维贝叶斯网络分类器的结构推导中。这可以解释为目前尚未发现全局结构评分和局部推导任务之间的关联。例如，假设增加一条不属于 GMBNC 但属于全局 BN 的边将提高全局得分，这条边在 S&S 类算法中将被选中，但显然这个步骤对目标 GMBNC 的推导没有直接帮助，其直接结果是搜索效率无法得到提升。此外，实际搜索过程中由于并不知道局部 GMBNC 的结构(节点和边)信息，无法实现局部结构的评分，这也阻碍了全局和局部学习任务的关联。这两个原因联合导致目前缺乏有效的可执行局部搜索的 S&S 类 GMBNC 学习算法。

在继承 IPC-GMBNC 的局部搜索框架基础上提出了 DOS-GMBNC，它通过利用更多的结构拓扑信息来引导搜索学习过程，进一步提高了学习效率而没有牺牲算法的正确性。同 IPC-GMBNC 和 PC 算法，DOS-GMBNC 算法也属于 CB 类。

3.2 算法描述

DOS-GMBNC 的主体框架(见表 1)同 IPC-GMBNC，通过推导每个 $C\in\boldsymbol{V}_C$ 的马尔科夫毯 $\boldsymbol{MB}_C$ 以及 $\{C\}\cup\boldsymbol{MB}_C$ 节点之间的连接性来实现 GMBNC 的结构学习。DOS-GMBNC 依赖 *SkeletonLearner*(见表 2，对应 IPC-GMBNC 的 IPC-BNC)来完成 $\{C\}\cup\boldsymbol{MB}_C$ 上的 DAG 结构主干的推导。*SkeletonLearner* 则进一步依赖 *RmvFalsePositives*(见表 3，对应 IPC-GMBNC 的 *RemoveFalsePC*)，后者是该算法的核心，执行的是后向选择(Backward selection)搜索，即从一个大候选全集开始，不断删除集合中的假正(False positive)节点(与相应关联的边)，直到无法识别更多的假正。该算法所需的 CI 测试几乎都发生在 *RmvFalsePositives* 中，因此 DOS-GMBNC 选择对该过程进行优化。

表 1 算法 DOS-GMBNC

```
输入：VC(类变量集合)，D(训练数据集)，ε(CI 测试显著性阈值)
输出：AInd(目标 GMBNC 的边集合)
1)AInd=ADel=∅;
2)FOR (C∈VC)DO
3)    SkeletonLearner(C,D,ε,AInd,ADel);
4)END FOR
5)利用文献[15]的规则对 AInd 中的未确定方向的边设定方向;
6)RETURN AInd
```

表 2 算法 SkeletonLearner

```
输入：C(类变量)，D(训练数据集)，ε(CI 测试显著性阈值)，AInd(已知属于
    GMBNC 的边)，ADel(已知不属于 GMBNC 的边)
输出：更新的 AInd 和 ADel
1)VStudied={X|(C—X)∈AInd};
2)ACan={(C—X)|∀X∈V\{X}\VStudied};
3)ACan=ACan\ADel;
4)RmvFalsePositives(C,ACan,AInd,ADel,D,ε);
5)VCanPC={X|(C—X)∈ACan};
6)VStudied=VStudied∪{C};
7)FOR (X∈VCanPC) DO
8)    ACan=ACan∪{(X—Y)|∀Y∈VStudied};
9)    RmvFalsePositives(X,ACan,ADel,ADel,D,ε);
10)   VStudied=VStudied∪{X};
11)END FOR
12)VPC={X|(C—X)∈ACan};
13)VSpouse=∅;
14)FOR (X∈VPC) DO
15)   VCanPC={Y|(X—Y)∈ACan};
16)   FOR (∀Y∈VCanPC AND ∀Y∉VStudied) DO
17)     IF(ID(C,Y|VSepset(C,Y)∪{X})>ε) DO
18)       Specify C→X←Y in ACan;
19)       VSpouse=VSpouse∪{Y};
20)     RmvFalsePositives(Y,ACan,ADel,ADel,D,ε);
21)     END IF
22)   END FOR
23)END FOR
24)Delete arcs connecting to V\VPC\VSpouse from ACan;
25)AInd=AInd∪ACan;
```

DOS-GMBNC 的设计基于这样的观察和假设：(1)越靠近节点 C 的节点越可能出现在 C 与其他节点 X 的 d-分割集中；(2)常出现在 d-分割集的节点一般较靠近节点 C；(3)这样的节点在构建 d-分割集时应被优先选择。具体实现体现在 *RmvFalsePositives*(见表 3)中，一旦发现一个 d-分割集 $\boldsymbol{S}$(第 8 行)，则更新 $\boldsymbol{S}$ 中每个变量 Y 的频率信息 $\boldsymbol{Fr}_Y$(第 13—15 行)。显然，$\boldsymbol{Fr}_Y$ 值越高，表明 Y 越靠近 C。相应的 $\boldsymbol{Fr}_Y$ 信息在 *GenSubSets*(见表 4)中被用来对候选条件集合进行排序(表 4 的第 7 行)，即综合得分 $\sum_{X\in\boldsymbol{S}}\boldsymbol{Fr}_X$ 高的 $\boldsymbol{S}$ 将在算法 *Rmv-*

FalsePositives 中被优先使用(表 3 的第 7 行),因为这样的 $\boldsymbol{S}$ 更可能成为 C 与其他节点 X 的 d-分割集。

表 3 算法 RmvFalsePositives

```
输入:C(类变量),D(训练数据集),ε(CI 测试显著性阈值),A_Can(候选和 C 直接关联的边),A_Ind(已知属于 GMBNC 的边),A_Del(已知不属于 GMBNC 的边)
输出:更新的 A_Ind 和 A_Del
1)V_NotPC=∅;
2)css=0;
3)V_CanPC={X|(C-X)∈A_Can};
4)DO
5)  FOR(∀X∈V_CanPC AND (X-C)∉A_Del) DO
6)    SS=GenSubSets(V_CanPC\{X},css,Fr);
7)    FOR(∀S∈SS) DO
8)      IF (I_D(C,X|S)≤ε) DO
9)        A_Can=A_Can\{(C-X)};
10)       A_Del=A_Del∪{(C-X)};
11)       V_NotPC=V_NotPC∪{X};
12)       V_Sepset(C,Y)=S;
13)       FOR(∀Y∈S) DO
14)         Fr_Y=Fr_Y+1;
15)       END FOR
16)       BREAK;// Skip the remaining S in SS
17)     END IF
18)   END FOR
19) END FOR
20) V_CanPC=V_CanPC\V_NotPC;
21) css=css+1;
22)WHILE(|V_CanPC|≥css)
```

表 4 算法 GenSubSets

```
输入:V(变量集合),css(目标子集大小),Fr(频率信息集合)
输出:SS(候选条件集合集)
1)SS=∅;
2)IF (css<1) DO
3)  SS=SS∪{∅};
4)ELSE
5)  SS={S|S⊆V AND |S|=css};
6)END IF
7)Sort S in SS in descending order based on Σ_{X∈S} Fr_X;
8)RETURN SS
```

3.3 算法正确性

DOS-GMBNC 和 IPC-GMBNC 的核心差异在于前者在 *RmvFalsePositives* 中动态调整 CI 测试的顺序。文献[14]中关于该过程有两个相关的基础引理,在此分别证明这两个引理依旧成立。

引理 1 在 *RmvFalsePositives* 中,C 的真实父、子节点不会被误判,即不会从候选的父、子集合 $\boldsymbol{V}_{CanPC}$ 里被删除。

证明:给定 *css*,*GenSubSets*(见表 4)能够生成完备的同 IPC-GMBNC 中一样的候选条件集合,只是候选集合的优先级按一定规则进行排序。若 C 和 X 直接相连,它们仍然可以通过表 3 第 8 行的测试;如果 C 和 X 不直接相连,且在 IPC-GMBNC 中无法通过基于条件集合 $\boldsymbol{S}$ 的测试,则它们在 DOS-GMBNC 中有可能失败于不同于 $\boldsymbol{S}$ 的条件集合,比如 $\boldsymbol{S}'$。得益于 DOS-GMBNC 所采取的动态排序策略,可能在执行第一个测试时即碰到 $\boldsymbol{S}'$,剩余的条件集合和测试随机被忽略(由于第 16 行的 BREAK);而在 IPC-GMBNC 中的场景很可能是先通过一些 CI 测试,直到碰到这样的 d-分割集为止。

引理 2 除了某些后代节点,C 的假父、子节点都将在 *RmvFalsePositives* 中被正确识别。

证明:可以参考引理 1 的方式,在文献[14]中关于引理 2 的证明的基础上完成证明,在此略之。

从引理 1 和引理 2 可以看到,DOS-GMBNC 仅仅影响 CB 类算法所需执行的 CI 测试的顺序,并不影响决策。

定理 2 DOS-GMBNC 在给定忠实假设情况下可以输出正确的无限制多维贝叶斯网络分类器。

证明:基于引理 1 和引理 2 的说明,*RmvFalsePositives* 可以实现与 IPC-GMBNC 的 *RemoveFalsePC* 一样的效果,而算法的其他部分保持不变,故文献[14]中关于 IPC-GMBNC 正确性的结论依旧成立。

3.4 算法优点讨论

IPC-GMBNC 开创了通过局部搜索完成对 GMBNC 的结构学习,而 DOS-GMBNC 继承了 IPC-GMBNC 的学习框架以及相应的优点,包括:

- 分而治之(Divide-and-conquer)。完整的结构学习依赖一系列的局部结构学习;而局部结构学习又分解为类与父、子和配偶节点之间的连接性学习。
- 局部搜索。整个学习过程本质上是广度优先搜索,但搜索被局限在与类节点半径不超过 2(假设每条边的距离相等,都为 1)的范围内。
- 前向和后向选择交叉的搜索策略。邻居节点以及相应边的推导是基于后向选择的搜索(*RmvFalsePositives*),即从候选邻居集合中不断删除假正,而假正元素的判断是基于"存在性"检查,即只要存在一个 C 与 X 的 d-分割集即,可判定 X 为假正元素。而由于检查是从空的 d-分割集开始增大,这极大限度地保证了 CI 测试的可靠性,而这是 CB 类算法已知的最大问题。

DOS-GMBNC 在执行后向选择时做了改进,选择并优先执行能带来"效果"的 CI 测试,即更可能发现假正节点的统计测试。一旦确认一个节点为假正搜索即停止,该策略可有效避免执行大量"无效"的 CI 测试,从而间接提高了学习效率。第 4 节的实验结果验证了这一点。

3.5 推导示例

为了便于理解,以经典的 Alarm 网络的 HR 和 HREK 两个节点为目标变量,在图 2 中展示了 DOS-GMBNC 算法执行关键步骤后的学习效果。

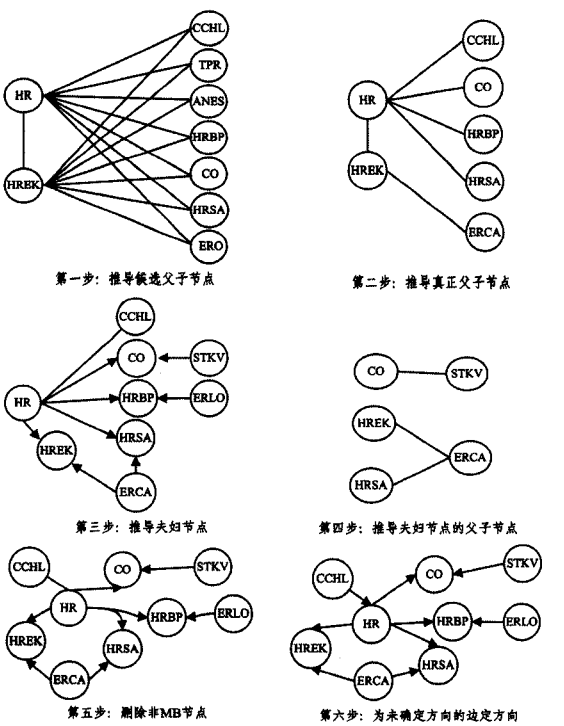


图 2 DOS-GMBNC 推导实例

可以注意到节点和边的推导同步进行;先期学习中已知不存在的边将存储在 $\boldsymbol{A}_{Del}$ 中,在未来学习中将直接忽略 $\boldsymbol{A}_{Del}$ 中的边,从而可以节省大量的计算。而早期保存在 $\boldsymbol{A}_{Del}$ 中的假正(边)则可能在未来的学习中被删除;以目标节点为中心,搜索半径局限在 2 层。

4 实验研究

4.1 实验方法

由于本研究着重在 GMBNC 的结构学习,实验选取了两个已知网络的经典问题 Asia 和 Alarm;另外,基于 Weka 的 BN 软件包随机生成了一个包含 100 个节点和一个包含 200 个节点的 BN,分别命名为 Test100 和 Test200(见表 5)。

表 5 实验用网络的基本信息

数据集	节点数目	边数目
Asia	8	8
Alarm	37	46
Test100	100	130
Test200	200	250

针对不同网络问题以及不同的学习样本集大小,分别独立生成 10 组样本集供实验需要。针对每个网络,随机选取一系列双节点($|\boldsymbol{V}_C|=2$)和三节点组合($|\boldsymbol{V}_C|=3$),分别视它们为目标节点集合,并根据定义可获得真实的网络结构用于实验结果的比较。

除了新提出的 DOS-GMBNC,还选取了 IPC-GMBNC 和 PC 算法,并从以下两个维度进行比较:

• 学习效果。这是通过度量和比较学习所得网络和真实网络之间的 Hamming 距离。Hamming 距离定义为将学习所得网络转变为真实网络所需的操作数,而允许的操作包括添加边、删除边和翻转边的方向。因为严格统计每一条边的差异,Hamming 距离是评估学习效果的理想度量。

• 学习效率。学习效率是基于 CI 测试的加权统计[16]。针对一个 CI 测试 $I_D(X,Y|\boldsymbol{Z})$,它的权重为 $2+|\boldsymbol{Z}|$,反映了统计测试复杂度和涉及的变量成正比。而一个算法的复杂度则为整个学习过程所执行的所有这种 CI 测试的权重之和。这是 CB 类算法的标准比较方法,最大的优点是机器无关性。

同文献[14],由于 MBNC 和 GMBNC 属于不同的知识发现任务,因此本文没有考虑 MBNC 的相关学习算法。算法基于 Weka 框架实现;PC 算法则直接调用 Weka 的实现版本。

4.2 学习效果

图 3 和图 4 分别展示了当面对两个($|\boldsymbol{V}_C|=2$)和 3 个($|\boldsymbol{V}_C|=3$)类变量时 3 个算法的学习效果,其中 X 轴表示样本数,从 1k 到 5k;Y 轴表示 Hamming 距离。可以观察到:(1)随着样本数目的增加,3 个算法的学习效果都得到了提高;(2)虽然 3 个算法理论上都是正确的,并且有理由相信当提供充分多的训练样本时它们都能输出理想的网络结构,但 DOS-GMBNC 和 IPC-GMBNC 提高的速度高于 PC,这显示了它们能够更充分地利用训练数据;(3)给定相同样本数时,DOS-GMBNC 与 IPC-GMBNC 的学习效果相当,这得益于 DOS-GMBNC 继承了 IPC-GMBNC 正确的算法框架,二者均优于 PC 算法;(4)IPC-GMBNC 和 DOS-GMBNC 的实际应用效果较佳,例如当面对 Test100 网络和 5000 个样本时,其输出的网络中仅有平均 5.48 条出错,占全部 130 条边的4.22%。

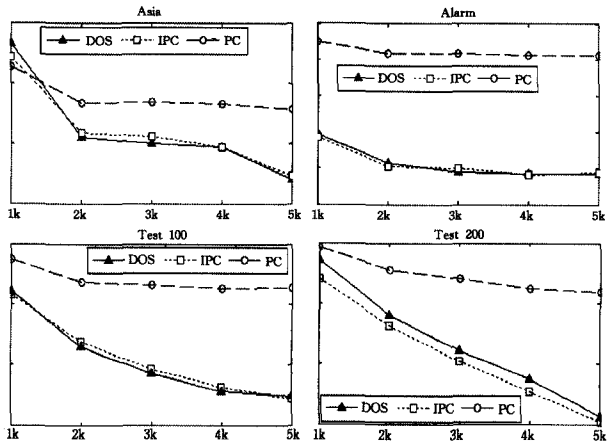


图 3 当$|\boldsymbol{V}_C|=2$时,DOS-GMBNC、IPC-GMBNC 与 PC 算法的学习效果比较

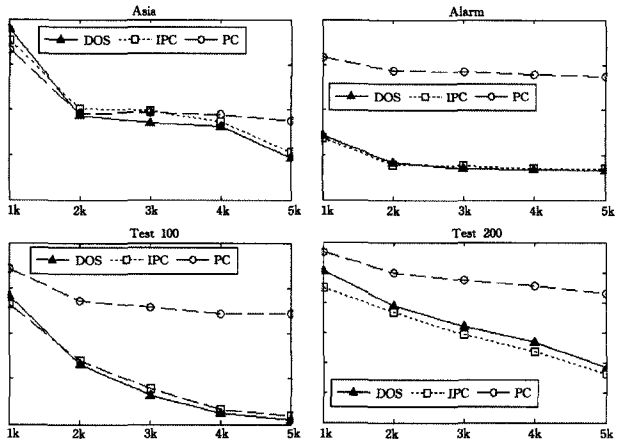


图 4 当$|\boldsymbol{V}_C|=3$时,DOS-GMBNC、IPC-GMBNC 与 PC 算法的学习效果比较

4.3 学习效率

图 5 和图 6 分别展示了当面对 2 个和 3 个类变量时 3 个算法的学习效率,其中 X 轴表示样本数,从 1k 到 5k;Y 轴表示所需的加权 CI 测试数。可以观察到:(1)随着样本数目的增加,3 个算法所执行的 CI 计算量都增加了,这得益于更多的样本包含更多的信息,可以提供更多可信赖的 CI 测试,故可以执行更"深"的搜索(对应的学习效果也获得了提升);(2)IPC-GMBNC 和 DOS-GMBNC 相对于 PC 算法节省了可观的计算量。例如在给定 Test100 网络和 5000 个样本的两类变量($|\boldsymbol{V}_C|=2$)实验中,PC、IPC-GMBNC 和 DOS-GMBNC 平均所需的加权 CI 测试数为 38752、7856 和 4344(见表 6)。IPC-GMBNC 和 DOS-GMBNC 分别比 PC 节省了 79.73% 和 88.79%的计算量;(3)由于 IPC-GMBNC 和 DOS-GMBNC 相比于 PC 相差过分悬殊,导致图中对应这两个算法的两条曲线近乎重叠。然而通过表 6 中更详细的基于 Test100 和 5000 个样本时的实验结果,可以看出 DOS-GMBNC 的动态调整策略是相当有效的,因此 DOS-GMBNC 的动态搜索策略是有效的。

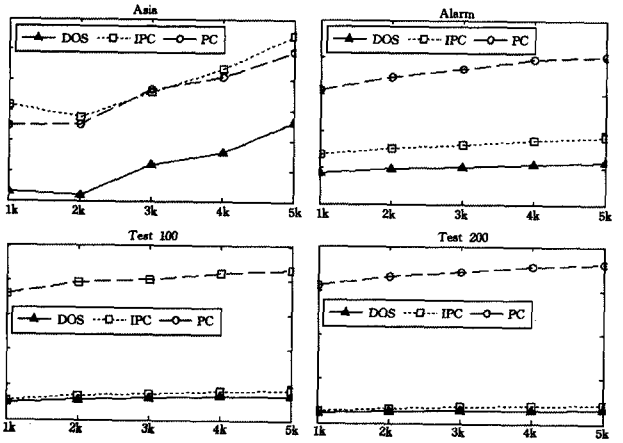


图 5 当$|V_C|=2$时，DOS-GMBNC、IPC-GMBNC 与 PC 算法的学习效率比较

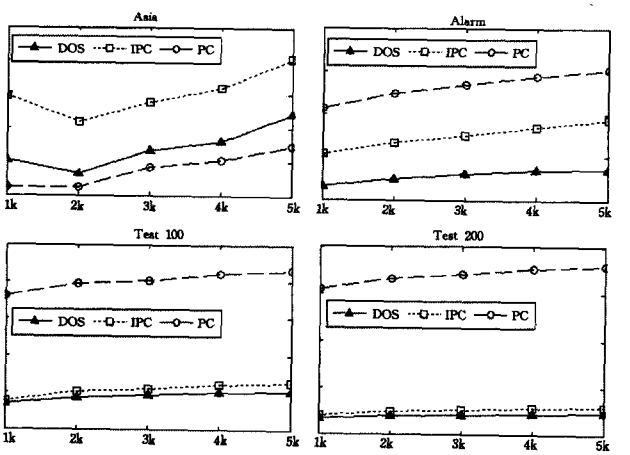


图 6 当$|V_C|=3$时，DOS-GMBNC、IPC-GMBNC 与 PC 算法的学习效率比较

表 6 Test100 网络以及$|V_C|=2$时实验中的平均加权 CI 测试数

样本集	PC	IPC	DOS
1k	31097	4609	2514
2k	34572	6282	3462
3k	35461	6787	3665
4k	37486	7511	4148
5k	38752	7856	4344

结束语 无限制多维贝叶斯网络分类器能够提供比(有限制)多维贝叶斯网络分类器关于不确定性更准确的建模能力，但传统的自动学习方法不得不先推导关于整体输入变量的完整贝叶斯网络。IPC-GMBNC 是已知第一个基于局部搜索的有效学习算法。这里提出的 DOS-GMBNC 可视为 IPC-GMBNC 的升级版本，它基于拓扑信息动态调整 CI 测试顺序，并被实验证明是有效的。未来的研究可以尝试挖掘更多的拓扑信息来进一步降低学习复杂度，另外可以尝试结合 CL 和 S&S 类实现混合的局部搜索算法。

参 考 文 献

[1] 任佳，杜文才，白勇. 基于贝叶斯网络自适应推理的无人机任务决策[J]. 系统工程理论与实践，2013，33(10)：2575-2582
Ren Jia, Du Wen-cai, Bai Yong. UAV mission decision-making based on Bayesian networks adaptive inference [J]. Systems Engineering-Theory & Practice, 2013, 33(10): 2575-2582

[2] 梁洁，蔡琦，初珠立，等. 反应堆补水系统诊断贝叶斯网络的建立和应用[J]. 原子能科学技术，2013，47(10)：1840-1844
Liang Jie, Cai Qi, Chu Zhu-li, et al. Constitution and application of reactor make-upsystem's fault diagnostic Bayesian networks [J]. Atomic Energy Science and Technology, 2013, 47(10): 1840-1844

[3] 刘建伟，黎海恩，罗雄麟. 概率图模型学习技术研究进展[J]. 自动化学报，2014，40(6)：1025-1044
Liu Jian-wei, Li Hai-en, Luo Xiong-lin. Learning technique of probabilistic graphical models: a review [J]. ACTA Automatica Sinica, 2014, 40(6): 1025-1044

[4] Chickering D M, Geiger D, Heckerman D. Learning Bayesian Network is NP-Hard[R]. Microsoft, 1994

[5] Zhang H, Liang Liang-xiao, Su Jiang. Hidden Naïve Bayes[C]// Proceedings of Canadian Artificial Intelligence Conference. 2005: 432-441

[6] Bielza C, Larranaga P. Discrete Bayesian Network Classifiers: A Survey[J]. ACM Computing Surveys, 2014, 47(1)

[7] Friedman N, Gierger D, Goldszmidt M. Bayesian Network Classifiers [J]. Machine Learning, 1997, 29: 131-163

[8] van der Gaag L C, de Waal P R. Multi-dimensional Bayesian Network Classifiers [C]// Proceedings of 3rd European Workshop on Probabilistic Graphical Models (PGM). 2006

[9] de Waal P R, van der Gaag L C. Inference and learning in multi-dimensional Bayesian network classifiers [C]// Proceedings of 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU). 2007: 501-511

[10] Zaragoza J C, Sucar E, Morales E. A Two-step Method to Learn Multi-dimensional Bayesian Network Classifiers based on Mutual Information Measures [C]// Proceedings of 24th International FLAIRS Conference. 2011

[11] Borchani H, Bielza C, Martinez-Martin P, et al. Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: An application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39) [J]. Journal of Biomedical Informatics, 2012, 45(6): 1175-1184

[12] Zaragoza J C, Sucar E, Morales E, et al. Bayesian chain classifiers for multidimensional classification [C]// Proceedings of 22nd International Joint Conference on Artificial Intelligence(IJCAI). 2011: 2092-2097

[13] Zhang Min-ling, Zhou Zhi-hua. A Lazy Learning Approach to Multi-label Learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048

[14] 傅顺开，Minn S，李志强. 多维贝叶斯网络分类器结构学习算法[J]. 计算机应用，2014，34(4)：1083-1088
Fu Shun-kai, Minn S, Li Zhi-qiang. Structure learning algorithm for general multi-dimensional Bayesian network classifiers[J]. Journal of Computer Application, 2014, 34(4): 1083-1088

[15] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search(Second Edition)[M]. The MIT Press, 2001

[16] Bromberg F, Margaritis D, Honavar V. Efficient Markov network structure discovery using independence tests[J]. Journal of Artificial Intelligence, 2009, 35(1): 449-484

[17] Zhang Min-ling, Zhou Zhi-hua. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351

[18] Borchani H, Bielza C, Toro C, et al. Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers [J]. Artificial Intelligence in Medicine, 2013, 57(3): 219-229