# 基于信息熵函数的启发式贝叶斯因果推理

刘洋[1)] 王利民[1),2)] 孙铭会[1)]

[1)](吉林大学计算机科学与技术学院, 长春 130012)

[2)](吉林大学符号计算与知识工程教育部重点实验室, 长春 130012)

**摘 要** 贝叶斯网络分类器（BNC）由于其优越的分类性能和可解释性在数据挖掘和人工智能等领域有着广泛的应用。信息论为其迅速发展奠定了坚实的数学理论基础，例如条件互信息被用来度量 BNC 拓扑结构中属性间的条件依赖关系。然而，贝叶斯网络又被称为因果网络，但目前人工智能等领域中有关贝叶斯网络因果关系的研究是一个很有争议性的课题。属性间因果性的定义远比相关性的定义复杂微妙很多。而条件互信息可能不适用于度量 BNC 整体拓扑结构对数据的拟合性，并且其表达式的对称性决定了其只能描述属性之间的无向相关性，而非有向因果性。本文从信息熵的角度对贝叶斯网络中的因果关系进行了探索性的研究，首先基于对似然函数定义了联合熵函数与贝叶斯网络拓扑结构中联合概率分布的映射关系，然后在此基础上提出了类条件熵和局部条件熵函数来识别拓扑结构中属性间的因果关系。最后提出了一种基于类标签驱动的启发式结构学习方法来构建可以兼顾有标签数据拟合和无标签数据泛化的 BNC（记为 HBN）。对美国加州大学欧文分校（UCI）机器学习数据库中 35 个数据集的实验评估表明，本文所提出算法与其它算法相比在分类性能上具有显著优势，例如 HBN 在 0-1 损失函数上明显优于 CFWNB（17 优 5 劣）、SKDB（14 优 5 劣）、AIWNB（17 优 7 劣）；在偏差上 HBN 与 CFWNB 相比 26 优 6 劣，与 SKDB 相比 10 优 5 劣，与 WAODE 相比 15 优 7 劣，与 RF 相比 29 优 4 劣，与 AIWNB 相比 22 优 6 劣。由于 CFWNB、WAODE、AIWNB 没有结构学习过程，其拓扑结构不受训练数据扰动的影响。这三种算法的方差显著低于其它算法。而 HBN 的局部拓扑结构能充分体现测试实例中隐含的因果关系，在一定程度上减轻训练数据过拟合带来的负面影响。因此，与 SKDB 和 RF 相比 HBN 的方差结果均明显占优（20 优 9 劣，26 优 3 劣）。与其他算法相比，HBN 的 0-1 损失函数和偏差结果分别平均提高了 6.06%和 12.65%。与 SKDB 和 RF 相比，HBN 的方差结果平均提高了 16.49%。HBN 为不确定性知识表示和推理提供了一种有效和可行的方法。

关键词 贝叶斯网络分类器；对数似然函数；联合熵；条件熵；交叉熵

**中图法分类号** TP18

# Heuristic Bayesian Causal Inference based on Information Entropy Function

Liu Yang[1)] Wang limin[1),2)] Sun minghui[1)]

[1)](College of Computer Science and Technology, Jilin University, Changchun 130012)

[2)](Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012)

**Abstract** Bayesian network classifier (BNC) has been widely used in the data mining, artificial intelligence and other fields due to its excellent classification performance and interpretability. Information theory has established a strong mathematical and theoretical basis for its rapid development. For example, conditional mutual information is widely used to measure the conditional dependence between attributes in the topology structure of BNC. However, Bayesian network is also called causal network, the research on causality in the Bayesian network is a controversial topic in the artificial intelligence and other fields. The definition of causality between attributes is much more complex and subtler than that of correlation. Conditional mutual information may be not suitable for measuring the

本课题得到国家重点研发计划 (No. 2019YFC1804804)、吉林省科技发展计划项目(No. 20200201281JC)资助. **刘洋**，博士研究生，主要研究领域为数据挖掘、贝叶斯网络. E-mail: liu_y18@mails.jlu.edu.cn. **王利民（通信作者）**，博士，教授，中国计算机学会(CCF)会员（19493S），主要研究领域为概率逻辑推理、贝叶斯网络. E-mail: wanglim@jlu.edu.cn. **孙铭会**，博士，副教授，中国计算机学会(CCF)会员（33008S），主要研究领域为人工智能. E-mail: smh@jlu.edu.cn.

extent to which the global topology structure of BNC fits data, and the symmetry of its expression determines that it can only describe the undirected correlation between attributes, not the directed causality. An exploratory research is carried out in the causal relationship of Bayesian networks from the perspective of information entropy. This paper firstly defines the mapping relationship between the joint entropy function and the joint probability distribution within the Bayesian networks from the perspective of the log-likelihood function, and then proposes the class conditional entropy function and local conditional entropy function based on the joint entropy function to identify the causal relationships between attributes in the topology structure. Finally, a label-driven heuristic structure learning method is proposed to build a BNC that can balance labeled data fitting and unlabeled data generalization, which is named HBN. Experimental evaluation on 35 datasets from the UCI machine learning repository shows that the proposed algorithm enjoys significant advantages in terms of classification performance over other state-of-the-art algorithms. For example, in terms of 0-1 loss function, HBN beats the algorithm of correlation-based feature weighting filter for naive Bayes (CFWNB) on 17 datasets and loses 5, beats selective $k$-dependence Bayesian classifier (SKDB) on 14 datasets and loses 5, beats attribute and instance weighted naive Bayes (AIWNB) on 17 datasets and loses 7. In terms of bias, HBN beats CFWNB on 26 datasets and loses 6, beats SKDB on 10 datasets and loses 5, beats AIWNB on 22 datasets and loses 6. Besides, when compared with ensemble algorithms, HBN also achieves significant advantages over weighted average one-estimators (WAODE: 11 wins and 2 loses in terms of 0-1 loss; 15 wins and 7 loses in terms of bias) and random forest (RF: 19 wins and 9 loses in terms of 0-1 loss; 29 wins and 4 loses in terms of bias). Variance-wise, CFWNB, WAODE and AIWNB have no structure learning and are irrelevant to the variation of training data, thus they enjoy lower variance results. The local topology of NBN can fully reflect the implicit causality in test instances, and reduce the negative impact of training data over fitting to a certain extent. Thus, HBN has significant advantages in terms of variance over SKDB (20 wins and 9 loses) and RF (26 wins and 3 loses). Compared with other algorithms, the average 0-1 loss and bias results of HBN are improved by about 6.06% and 12.65%. Compared with SKDB and RF, the average variance results of HBN is improved by about 16.49%. HBN is effective and feasible for uncertain knowledge representation and reasoning.

**Key words** Bayesian network classifier; log likelihood function; joint entropy; conditional entropy; cross entropy

## 1 引言

分类是人工智能与机器学习领域最重要的任务之一，其核心问题是从数据中学习一个分类模型或分类决策函数，建立从输入空间$X$到输出空间$Y$的映射，表达为条件概率分布$P(Y|X)$或决策函数$Y = f(X)$的形式[1]。前者表示给定输入条件下输出的概率模型，例如朴素贝叶斯（Naïve Bayes，NB）和隐马尔科夫；后者表示输入到输出的非概率模型，例如 $k$ 近邻和支持向量机。在众多分类器中，贝叶斯网络分类器（Bayesian network classifiers，BNCs）[2]为不确定性条件下的知识表示和推理提供了一种强大的工具。以有向无环概率图形式的拓扑结构描述变量间关联性，广泛应用于各种领域[3-8]。

信息论[9]为 BNC 的迅速发展奠定了坚实的数学理论基础，例如为克服 NB 的条件独立性假设，研究人员普遍采用条件互信息来度量 BNC 拓扑结构中属性之间的条件依赖关系，并由此衍生了由一阶树增广型朴素贝叶斯（Tree-augmented Naive Bayes，TAN）[10]到任意 $k$ 阶依赖的贝叶斯网络分类器（$k$-dependence Bayesian network classifier，KDB）[11]的单模型族，以及以平均一阶依赖估测器（Averaged one-dependence estimators，AODE）[12]为代表的集成模型族。然而基于信息测度（如条件互信息）识别属性间依赖关系并构建贝叶斯网络拓扑结构，并不能精准表达在不同属性取值情况下依赖关系的动态变化[13]。此外，虽然贝叶斯网络又被称为信念网络或者因果网络，但条件互信息表达式的对称性决定了其只能描述无向依赖关系（而非有向因果关系）。基于贝叶斯网络拓扑结构的有向无环特性，现有的贝叶斯网络模型基本上均人为定义弧定向策略，并不能体现真正的因果关系。

本文在实验中发现，条件互信息实质上是度量

条件依赖和条件独立两种局部拓扑结构分别编码数据集所需的平均比特数之差，而不能用于度量整体拓扑结构的数据拟合性，最终构建的 BNC 模型将是次优的。因此，本文从对数似然函数的角度来定义联合熵函数与 BNC 拓扑结构中联合概率分布的映射关系，以条件熵来识别拓扑结构中属性依赖关系并证明其合理性。针对贝叶斯网络的因果关系表达，从信息熵的角度进行了分析和解释。在此基础上，提出基于类标签驱动的启发式结构学习方法来构建可以兼顾数据拟合和泛化的 BNC。

为方便叙述与理解，首先约定文中符号的基本含义。大写字母（例如：$X_i$，$Y$）表示变量（或属性），小写字母（例如：$x_i$，$y$）表示对应的变量取值。大写黑体（例如：$\boldsymbol{X}$，$\boldsymbol{\Pi}$）表示变量的集合，小写黑体（例如：$\boldsymbol{x}$，$\pi$）表示对应的变量集合的取值。$\boldsymbol{\Pi}_i^{\mathcal{B}}$表示属性$X_i$在拓扑结构$\mathcal{B}$中的父节点属性集合。

本文的具体贡献如下:

1) 从对数似然的角度出发，提出熵函数$H_{\mathcal{B}}$的概念来度量 BNC 拓扑结构$\mathcal{B}$中所蕴含的信息量，推导并证明以条件熵$H\left(X_i\middle|\boldsymbol{\Pi}_i^{\mathcal{B}},Y\right)$来度量属性之间因果关系的合理性。

2) 提出类条件熵和局部条件熵来分别度量$\mathcal{B}$中属性（或属性值）间所涉及的一般（或特殊）因果关系。最终的分类决策由基于交叉熵的模型匹配算法实现。

3) 本文使用美国加州大学欧文分校（UCI）机器学习数据库[14]中的 35 个数据集进行实验，其中数据集规模∈[24, 299285]，类标签个数∈[2, 50]。实验结果验证了本文所提出算法在 0-1 损失函数、偏差和方差方面上的有效性和可靠性。

## 2 相关工作

一般来说，BNC 是用来表示一组属性$\boldsymbol{X}=\{X_1,X_2,\ldots,X_n\}$之间概率依赖关系的图模型$\mathcal{B}=\langle\mathcal{G},\theta\rangle$，其中$\mathcal{G}=\langle\boldsymbol{V},\boldsymbol{E}\rangle$表示有向无环图，$\boldsymbol{V}$为节点集，与拓扑结构中的变量（或属性）相对应；$\boldsymbol{E}$为有向边集，表示变量之间的依赖关系；θ为每个节点所对应条件概率表。假设已知 BNC 的拓扑结构$\mathcal{B}$，对于类标签未知的测试实例$\boldsymbol{x}=\{x_1,x_2,\ldots,x_n\}$，BNC 将基于最大后验概率准则预测$\boldsymbol{x}$最可能隶属的类标签$y^*\in Y=\{y_1,\cdots,y_m\}$，具体表达如下：

$$\begin{aligned}y^*&=\arg\max_{y\in Y}P_{\mathcal{B}}(y|\boldsymbol{x})=\arg\max_{y\in Y}\frac{P_{\mathcal{B}}(\boldsymbol{x},y)}{P_{\mathcal{B}}(\boldsymbol{x})}\\&\propto\arg\max_{y\in Y}P_{\mathcal{B}}(\boldsymbol{x},y)\\&=\arg\max_{y\in Y}P(y)\prod_{i=1}^{n}P\left(x_i|\pi_i^{\mathcal{B}},y\right)\end{aligned}\tag{1}$$

其中$\pi_i^{\mathcal{B}}$表示$\mathcal{B}$中属性$X_i$的父节点属性值集合。

利用对数似然函数来评价分类器对训练数据的拟合程度是学习 BNC 的有效方法之一[15][16]。对于训练数据集$\mathcal{D}$中的任意实例$d=\{x_1,x_2,\ldots,x_n,y\}$，对数似然函数$\log P_{\mathcal{B}}(d)$可以根据概率分布$P_{\mathcal{B}}(d)$来度量$\mathcal{B}$描述实例$d$所需的比特数[10]。从统计学角度而言，对数似然函数取值越高，说明$\mathcal{B}$对$d$拟合得越好。给定含有 $N$ 条实例的训练集$\mathcal{D}$，$LL_{\mathcal{B}}(\mathcal{D})$可以度量$\mathcal{B}$中所蕴涵的用于描述$\mathcal{D}$的信息量，其定义如下：

$$LL_{\mathcal{B}}(\mathcal{D})=\sum_{d\in\mathcal{D}}N\cdot P(d)\log P_{\mathcal{B}}(d)\tag{2}$$

其中$N\cdot P(d)$表示$d$在$\mathcal{D}$中出现的次数。如果训练数据规模足够大，则$P(d)$和$P_{\mathcal{B}}(d)$的估计值将分别逼近于真实联合概率分布$P(\boldsymbol{x},y)$和从$\mathcal{B}$中学习到的概率分布$P_{\mathcal{B}}(\boldsymbol{x},y)$。因为对于不同的 BNC, $LL_{\mathcal{B}}(\mathcal{D})$中的参数 $N$ 可以视为常量，不会影响比较结果，结合公式（1）和（2）可以得到如下所示的联合熵函数$H_{\mathcal{B}}$:

$$\begin{aligned}H_{\mathcal{B}}&=-\sum_{Y,\boldsymbol{X}}P(\boldsymbol{x},y)\log P_{\mathcal{B}}(\boldsymbol{x},y)\\&=-\sum_{Y,\boldsymbol{X}}P(\boldsymbol{x},y)\log\left\{P(y)\prod_{i=1}^{n}P\left(x_i|\pi_i^{\mathcal{B}},y\right)\right\}\\&=-\sum_{Y,\boldsymbol{X}}P(\boldsymbol{x},y)\log P(y)\\&\quad-\sum_{Y,\boldsymbol{X}}\sum_{i=1}^{n}P(\boldsymbol{x},y)\log P\left(x_i|\pi_i^{\mathcal{B}},y\right)\\&=-\sum_{Y}P(y)\log P(y)\\&\quad-\sum_{i=1}^{n}\sum_{Y,X_i,\boldsymbol{\Pi}_i^{\mathcal{B}}}P\left(x_i,\pi_i^{\mathcal{B}},y\right)\log P\left(x_i|\pi_i^{\mathcal{B}},y\right)\\&=H(Y)+\sum_{i=1}^{n}H\left(X_i\middle|\boldsymbol{\Pi}_i^{\mathcal{B}},Y\right)\end{aligned}\tag{3}$$

公式（3）表明，为了使得对数似然函数$\log P_{\mathcal{B}}(d)$取值统计极大化，需要找到能使得联合熵函数$H_{\mathcal{B}}$最小的网络拓扑结构。从信息论的角度对

$H_{\mathcal{B}}$的解释是，它可以度量对数据集 $\mathcal{D}$ 中实例进行编码所需的平均比特数，因此可以用来评估 $\mathcal{B}$ 的合理性。

如图 1 所示，NB 假设属性之间是相互独立的，任何属性对之间都不存在有向弧（即$\boldsymbol{\Pi}_i^{\mathrm{NB}}=\emptyset$），因此，NB 也是 BNC 中最简单的模型之一。NB 拓扑结构所对应的联合概率表达式和熵函数$H_{\mathcal{B}}$表达式如下所示：

$$\begin{cases} P_{\mathrm{NB}}=P(y)\prod_{i=1}^{n}P(x_i|\,y) \\ H_{\mathrm{NB}}=H(Y)+\sum_{i=1}^{n}H(X_i|\,Y) \end{cases} \tag{4}$$

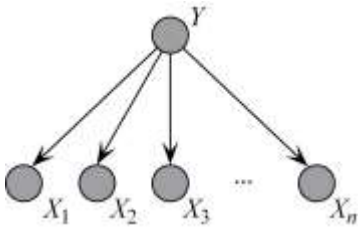



图 1 NB 示例

在许多实际应用场境中，属性之间的依赖关系并不总是完全独立的，识别显著性依赖关系以便放松 NB 的条件独立性假设将有效地提高其分类性能[17-19]。一般来说，目前现有的改进方法主要从两个方面展开：NB 拓扑结构的高阶扩展；低阶 BNC 集成。

KDB[11]将 NB 扩展到任意高阶依赖的网络拓扑结构，即允许每个属性 $X_i$ 最多有 $k$ 个非类变量的属性为其父节点。在 KDB 的学习过程中，所有属性都要按照互信息的取值进行降序排序。通过比较条件互信息，每个属性（除了根节点属性）的父节点仅能从序列中在其之前的属性集合中进行选择。KDB 网络拓扑结构所对应的联合概率表达式和熵函数$H_{\mathcal{B}}$表达式如下所示：

$$\begin{cases} P_{\mathrm{KDB}}=P(y)\prod_{i=1}^{n}P\left(x_i|\pi_i^{\mathrm{KDB}},\,y\right) \\ H_{\mathrm{KDB}}=H(Y)+\sum_{i=1}^{n}H\left(X_i|\boldsymbol{\Pi}_i^{\mathrm{KDB}},\,Y\right) \end{cases} \tag{5}$$

其中$\left|\pi_i^{\mathrm{KDB}}\right|=|\boldsymbol{\Pi}_i^{\mathrm{KDB}}|\le k$。当$k=2$时，KDB 示例如图 2 所示。KDB 对属性规模和结构复杂度的限定性为后续的改进研究提供了可能，例如綦小龙等人[20]针对基于约束的方法存在的序依赖、高阶检验等问题，采用度量信息矩阵和“偷懒”启发式策略构建高阶贝叶斯网络结构。Martínez 等人[21]以均方根误差（Root mean square error，RMSE）为目标函数，采用余一校验法自适应选择最优属性子集和依赖阶数 $k$。

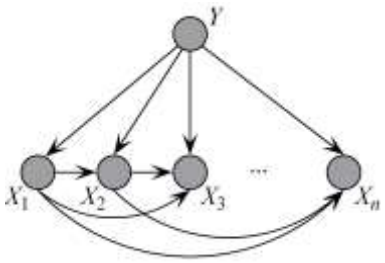



图 2 KDB 示例（$k$=2）

不同于结构扩展，通过构建多个低阶 BNC 进行集成的方式也可以有效提高 BNC 的分类性能，例如 Webb 等人[12]所提出的 AODE 是由$n$个一阶依赖的估测器（One-dependence estimators，ODE）所集成，其中$n$代表属性数目，如图 3 所示，每个 ODE 顺序选择属性$X_\alpha$和类变量$Y$为其它所有属性的超父节点，并假设非超父节点之间条件独立，其中$1\le\alpha\le n$。每个 ODE 网络拓扑结构所对应的联合概率表达式和熵函数$H_{\mathcal{B}}$表达式如下所示：

$$\begin{cases} P_{\mathrm{ODE}}^{\alpha}=P(x_\alpha,\,y)\prod_{i=1,i\neq\alpha}^{n}P(x_i|x_\alpha,\,y) \\ H_{\mathrm{ODE}}^{\alpha}=H(X_\alpha,\,Y)+\sum_{i=1,i\neq\alpha}^{n}H(X_i|X_\alpha,\,Y) \end{cases} \tag{6}$$

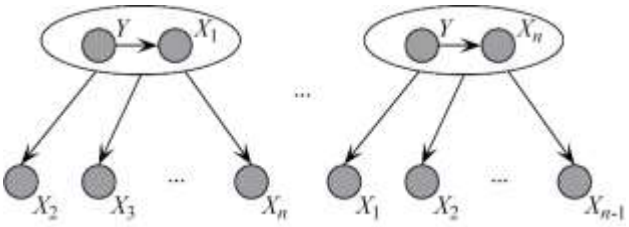


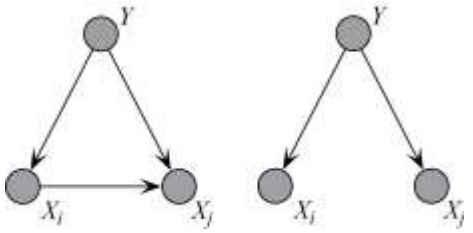

图 3 AODE 示例

AODE 采用平均加权法对每个 ODE 子模型都赋予相同的权重。然而在现实中，不同子模型的条件独立性假设相对于同一个样本的拟合程度不同，同一模型的条件独立性假设相对于不同样本的拟合程度也不同。加权法对于区分各子模型的差异性是最有效的方法之一，段智议等人[22]针对 ODE 子模型对不同样本的拟合差异性，通过权系数分配来体现属性与类变量间的相关性以及属性间的条件相关性。向忠良等人[23]基于超父变量与非超父变量的相关性计算权重。

# 3 算法介绍

## 3.1 可行性及相关概念

如图 4 所示，包含两个属性$X_i, X_j$的局部 BNC 拓扑结构可以表达为条件依赖和条件独立两种形式，其对应的条件联合概率分别为$P(x_i, x_j|y)$和$P(x_i|y)P(x_j|y)$。




(a) 条件依赖　　(b) 条件独立

图 4　属性$X_i$与$X_j$之间条件依赖(a)和条件独立(b)示例

可以将条件互信息的$I(X_i; X_j|Y)$改写为如下形式：

$$
\begin{aligned}
I(X_i; X_j|Y) &= \sum_{X_i,X_j,Y} P(x_i, x_j, y) \log \frac{P(x_i, x_j|y)}{P(x_i|y)P(x_j|y)} \\
&= \sum_{X_i,X_j,Y} P(x_i, x_j, y) \log P(x_i, x_j|y) \\
&\quad - \sum_{X_i,X_j,Y} P(x_i, x_j, y) \log P(x_i|y)P(x_j|y) \\
&= \left\{ \sum_{Y} P(y) \log P(y) \right. \\
&\quad \left. + \sum_{X_i,X_j,Y} P(x_i, x_j, y) \log P(x_i, x_j|y) \right\} \\
&\quad - \left\{ \sum_{Y} P(y) \log P(y) \right. \\
&\quad \left. + \sum_{X_i,X_j,Y} P(x_i, x_j, y) \log P(x_i|y)P(x_j|y) \right\}
\end{aligned} \tag{7}
$$

比较公式（3）和（7）所示的联合熵和条件互信息表达式可知，$I(X_i; X_j|Y)$实质上是比较条件依赖和条件独立两种局部拓扑结构分别编码数据集所需的平均比特数之差，适用于验证局部结构的相对合理性而非全局结构最优性。此外，条件互信息表达式的对称性决定了其只能描述属性之间的无向相关性，而非有向因果性。BNC 只能通过人工定向的方式构建有向无环图。以 TAN 为例，它构建的最大权重跨度树的定向方式是：通过选择一个根节点并将所有边的方向设置为向外，将这棵无向树转换为有向树。对于不同的根节点，属性$X_i$与$X_j$之间的边可能是$X_i \to X_j$或$X_j \to X_i$。很明显，$X_j \to X_i$这样的有向边只能表达属性$X_i$与$X_j$是条件依赖的，但并不意味着$X_j$是原因，$X_i$是结果。

条件熵函数$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, Y)$描述了在给定$\boldsymbol{\Pi}_i^{\mathcal{B}}$，$Y$前提下$X_i$的不确定性。如果$\{\boldsymbol{\Pi}_i^{\mathcal{B}}, Y\}$是原因，$X_i$作为结果确定发生，则很明显$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, Y)$将变小。而公式（4）~（6）则严格体现了联合熵函数与联合概率分布的对应关系，其中条件熵函数$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, Y)$可以在一定程度上表达父子变量间的有向因果关系（$\boldsymbol{\Pi}_i^{\mathcal{B}}$是原因，$X_i$是结果）和二者之间的强依赖关系。很明显，相对于条件互信息，联合熵函数在知识表达和数据拟合方面具有明显的优势。以数据集 Seer_mdl 为例（详见表 1），条件互信息$I(X_i; X_j|Y)$最大化的属性对和联合熵$H(X_i, X_j, Y)$最小化的属性对分别为$\{X_4, X_7\}$和$\{X_{10}, X_{12}\}$，二者在不同实例上对应的联合概率分布如图 5 所示，后者的数据拟合程度明显优于前者。

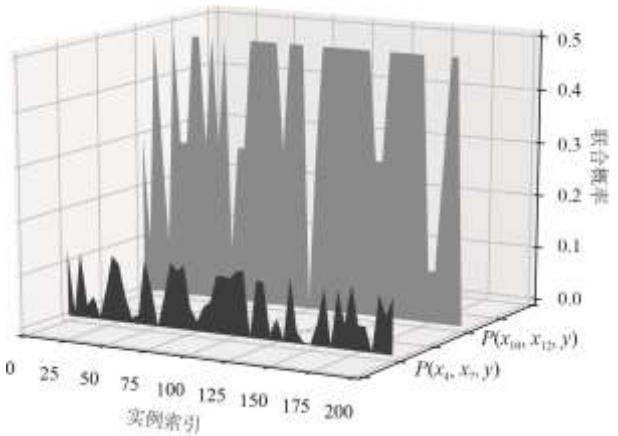
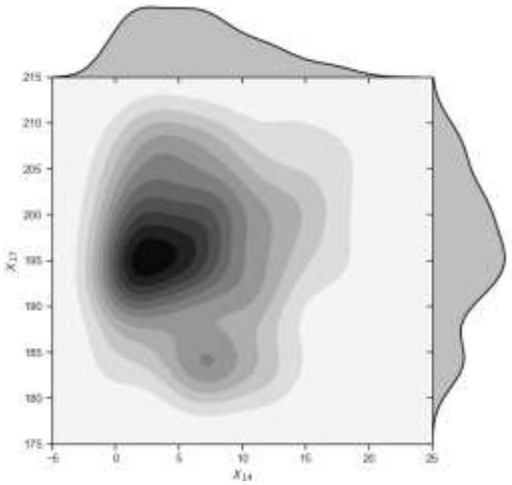
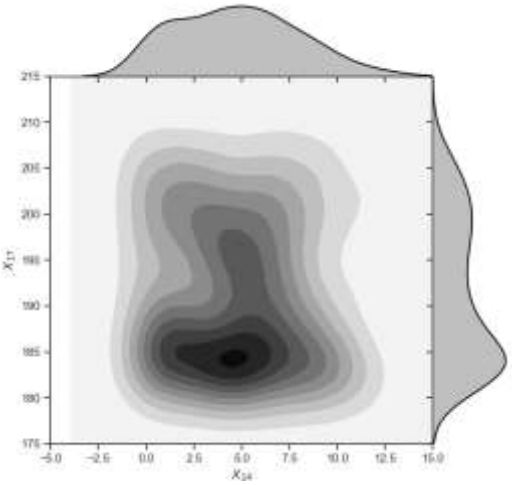



图 5　联合概率$P(x_4, x_7, y)$与$P(x_{10}, x_{12}, y)$在不同实例上的取值分布示意图

此外，类似聚类“物以类聚，人以群分”的逻辑思路，不同类的数据的特征属性关联性应存在一定程度的差异性。以数据集 Vehicle 为例（详见表 1），如图 6 所示，类条件概率分布$P(x_{14}, x_{17}|y)$在不同类标签的情况下，拟合不同样本时存在明显的差异性。可见，属性值之间的关联性相对于属性之间的关联性具备更细的粒度。识别不同样本间的差异性依赖特征将有助于提升 BNC 的图式知识表达性和分类泛化性能。



(a) 类标签$y_0$ = van



(b) 类标签$y_1$ = bus

图 6 给定类标签 $y_0$ (a)和 $y_1$ (b)条件下属性 $X_{14}$与$X_{17}$的取值分布

**定义 1.** 类条件熵函数. 给定类标签为$y$的训练子集$\mathcal{D}_y$以及其对应的 BNC 拓扑结构$\mathcal{B}$，类条件熵$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, y)$可用于度量$\mathcal{D}_y$中属性间的一般因果关系，

$$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}},y) = -\sum_{X_i,\boldsymbol{\Pi}_i^{\mathcal{B}}} P(x_i,\pi_i^{\mathcal{B}},y)\log P(x_i|\pi_i^{\mathcal{B}},y) \quad (8)$$

其中$\pi_i^{\mathcal{B}}$表示属性值$x_i$在$\mathcal{B}$中的父节点属性值集合。公式（8）中涉及的概率计算如下所示：

$$\begin{cases} P(\pi_i^{\mathcal{B}},y) = \dfrac{1}{N}\displaystyle\sum_{r=1}^{N}\delta_r(\pi_i^{\mathcal{B}},y) \\ P(x_i,\pi_i^{\mathcal{B}},y) = \dfrac{1}{N}\displaystyle\sum_{r=1}^{N}\delta_r(x_i,\pi_i^{\mathcal{B}},y) \\ P(x_i|\pi_i^{\mathcal{B}},y) = \dfrac{P(x_i,\pi_i^{\mathcal{B}},y)}{P(\pi_i^{\mathcal{B}},y)} \end{cases} \quad (9)$$

其中$\delta_r(\cdot)$为二值逻辑函数，如果属性值集合在第$r$条实例中出现，则为 1；否则为 0。

由于测试实例$\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$的所属类标签未知，预测结果$y^*$可能是$m$个类标签中的任意一个，即$y^* \in Y = \{y_1, \cdots, y_m\}$。假设测试实例$\boldsymbol{x}$属于每个类标签的概率是相等的，即$1/m$。则可以将$\boldsymbol{x}$转化为如下所示的$m$条伪训练实例$\{x_1, x_2, \dots, x_n, y_i\}$（$1 \le i \le m$）：

$$\boldsymbol{x} = \{x_1, x_2, \dots, x_n\} = \begin{cases} \{x_1, x_2, \dots, x_n, y_1\} \\ \{x_1, x_2, \dots, x_n, y_2\} \\ \vdots \\ \{x_1, x_2, \dots, x_n, y_m\} \end{cases},$$

并将其添加到训练集$\mathcal{D}$来构成伪训练集$\mathcal{P}$[13]。类似于公式（3）中的$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, Y)$，对于测试实例$\boldsymbol{x}$，有如下定义：

**定义 2.** 局部条件熵函数. 给定测试实例$\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$转化成的伪训练集$\mathcal{P}$以及其对应的 BNC 拓扑结构$\mathcal{B}$，局部条件熵$H(x_i|\pi_i^{\mathcal{B}}, Y)$可用于度量$\boldsymbol{x}$中属性值间的特殊因果关系，

$$H(x_i|\pi_i^{\mathcal{B}}, Y) = -\sum_{Y} \hat{P}(x_i, \pi_i^{\mathcal{B}}, y)\log \hat{P}(x_i|\pi_i^{\mathcal{B}}, y) \quad (10)$$

其中

$$\begin{cases} \hat{P}(\pi_i^{\mathcal{B}},y) = \dfrac{1}{N+1}\left[\displaystyle\sum_{r=1}^{N}\delta_r(\pi_i^{\mathcal{B}},y) + \dfrac{1}{m}\right] \\ \hat{P}(x_i,\pi_i^{\mathcal{B}},y) = \dfrac{1}{N+1}\left[\displaystyle\sum_{r=1}^{N}\delta_r(x_i,\pi_i^{\mathcal{B}},y) + \dfrac{1}{m}\right] \\ \hat{P}(x_i|\pi_i^{\mathcal{B}},y) = \dfrac{\hat{P}(x_i,\pi_i^{\mathcal{B}},y)}{\hat{P}(\pi_i^{\mathcal{B}},y)} \end{cases} \quad (11)$$

### 3.2 基于熵函数$H_{\mathcal{B}}$的启发式贝叶斯网络集成模型

如图 7 所示，本文提出构建基于熵函数$H_{\mathcal{B}}$的启发式贝叶斯网络集成模型（$H_{\mathcal{B}}$-based Bayesian network ensemble classifier，记为 HBN）。首先根据类标签$y$取值的不同将$\mathcal{D}$划分为$m$个训练子集。针对每个训练子集$\mathcal{D}_y$，基于类条件熵最小原则确定属性次序以及父子属性之间的因果关系，并构建子模型HBN$^y$。然后将测试实例$\boldsymbol{x}$预分配类标签并拓展成伪训练集$\mathcal{P}$，基于局部条件熵最小原则确定属性值之间的因果相关性并构建子模型HBN$^{\boldsymbol{x}}$。最终的分类决策将由这$m+1$个子模型集成实现。

对于训练子集$\mathcal{D}_y$而言，$H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, y)$的取值越小，$X_i$与$\{\boldsymbol{\Pi}_i^{\mathcal{B}}, y\}$之间的因果关系越显著。例如首先选择$H(X_i|y)$取值最小的属性$X_i$作为 BNC 拓扑中的第一个属性节点；根据$H(X_j|X_i, y)$取值最小确定第

二个属性节点$X_j$和因果弧段$\{X_i, y\} \to X_j$；根据$H(X_t|X_i, X_j, y)$取值最小确定第三个节点$X_t$和因果弧段$\{X_i, X_j, y\} \to X_t$，然后以此类推。通过这种方式可以直观地区分不同属性以及属性间因果关系在树状拓扑结构中的层次。HBN$^y$使用参数$k$来动态调整 BNC 的拓扑结构复杂度，可以使其因果依赖关系与训练数据的拟合程度更加紧密，进而获得更好的泛化能力。本文将从训练子集$\mathcal{D}_y$上学习得到的 BNC 记为HBN$^y$。具体学习过程如算法 1 所示。

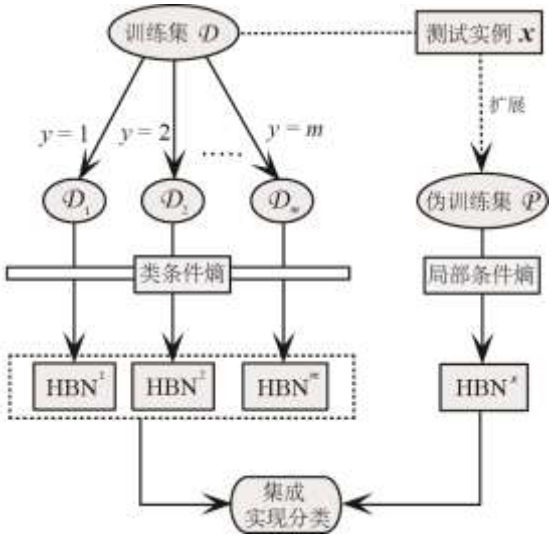



图 7 算法 HBN 的流程图

**算法 1**. 基于数据子集$\mathcal{D}_y$的 BNC 学习算法HBN$^y$.

输入：数据子集$\mathcal{D}_y = \{X_1, X_2, \dots, X_n, y\}$和参数 $k$;

输出：模型$\mathcal{B}_y$.

1. 初始化 BNC 拓扑结构$\mathcal{B} = \{y\}$，属性列表$\mathcal{L} = \{\emptyset\}$;
2. 重复以下步骤直到$\mathcal{L}$包含所有属性：
   2.1 选出不属于$\mathcal{L}$且使得$H(X_j|\boldsymbol{\Pi}_j^{\mathcal{B}}, y)$取值最小的属性$X_j$，其中$\boldsymbol{\Pi}_j^{\mathcal{B}} \in \mathcal{L}$且$|\boldsymbol{\Pi}_j^{\mathcal{B}}| = \min(|\mathcal{L}|, k)$;
   2.2 在$\mathcal{B}$中添加一个代表属性$X_j$的节点，并添加一条由$y$指向$X_j$的有向边;
   2.3 在$\mathcal{B}$中添加$|\boldsymbol{\Pi}_j^{\mathcal{B}}|$条从$\boldsymbol{\Pi}_j^{\mathcal{B}}$中属性节点指向$X_j$的有向边;
   2.4 将$X_j$添加到$\mathcal{L}$中;
3. 返回$\mathcal{B}_y = \mathcal{B}$.

算法 1 所构建的HBN$^y$模型力求充分表达训练子集$\mathcal{D}_y$中所蕴含的属性之间的因果关系。但从统计角度而言，对训练数据的过拟合将导致对测试样本的欠拟合，进而降低 BNC 的泛化性能。因此本文针对测试实例$\boldsymbol{x}$，使用局部条件熵函数来构建与HBN$^y$互补的局部模型HBN$^x$。HBN$^x$采用与HBN$^y$相同的学习策略来构建 BNC 拓扑结构。具体学习过程如算法 2 所示。

**算法 2**. 基于实例$\boldsymbol{x}$的 BNC 学习算法HBN$^x$.

输入：测试实例$\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$，伪数据集$\mathcal{P}$和参数 $k$;

输出：模型$\mathcal{B}_{\boldsymbol{x}}$.

1. 初始化 BNC 拓扑结构$\mathcal{B} = \{Y\}$，属性列表$\mathcal{L} = \{\emptyset\}$;
2. 重复以下步骤直到$\mathcal{L}$包含所有属性：
   2.1 选出不属于$\mathcal{L}$且使得$H(x_i|\pi_i^{\mathcal{B}}, Y)$取值最小的属性$X_i$，其中$\pi_i^{\mathcal{B}} \in \mathcal{L}$且$|\pi_i^{\mathcal{B}}| = \min(|\mathcal{L}|, k)$;
   2.2 在$\mathcal{B}$中添加一个代表属性$X_i$的节点，并添加一条由$Y$指向$X_i$的有向边;
   2.3 在$\mathcal{B}$中添加$|\pi_i^{\mathcal{B}}|$条从$\pi_i^{\mathcal{B}}$中属性节点指向$X_i$的有向边;
   2.4 将$X_i$添加到$\mathcal{L}$中;
3. 返回$\mathcal{B}_{\boldsymbol{x}} = \mathcal{B}$.

传统 BNC 在训练集上通过结构学习得到网络拓扑结构后，通常是将测试实例的属性值映射到该拓扑结构中计算后验概率，再使用最大后验概率准则进行分类决策。但它通常只适用于单个模型，而并不适用于类似HBN的集成模型。交叉熵（Cross Entropy）方法[24]近年来被广泛应用到许多组合优化问题的求解中，例如通信网络可靠性优化问题[25]、电力系统可靠性评估[26]和机器学习[27]等领域。交叉熵刻画了两个概率分布之间的距离。设$q(x)$和$\hat{q}(x)$是给定的两个概率分布函数，$x$ 为 $n$ 维随机变量，则$q(x)$和$\hat{q}(x)$的交叉熵可定义为：

$$H(q, \hat{q}) = -\sum_{x} q(x) \log \hat{q}(x) \qquad (12)$$

不同 BNC 拓扑结构的相似性可以根据其所对应联合概率的交叉熵函数的近似程度来体现。子模型HBN$^i$的拓扑结构描述了不同类标签$y_i$前提下属性间的一般因果关系，HBN$^x$的拓扑结构则充分描述了测试样本$\boldsymbol{x}$蕴含的属性间的特殊因果关系。根据交叉熵值最小原则，利用公式（12）选择与HBN$^x$的拓扑结构最相似的第$t$个子模型HBN$^t$，该过程共需进行 $m$ 次交叉熵匹配操作。最终分类决策将通过对最终选出的子模型HBN$^t$与HBN$^x$进行线性加权集成的方法来实现，对应的概率判别规则如下所示：

$$\hat{P}(y|\boldsymbol{x}) = \alpha \cdot P_{\mathrm{HBN}^t}(y|\boldsymbol{x}) + \beta \cdot P_{\mathrm{HBN}^x}(y|\boldsymbol{x}) \qquad (13)$$

然而，在没有任何先验知识的情况下很难确定每个子模型在处理不同样本时的动态加权系数。因此本文从效率角度出发采用了平均静态加权，即$\alpha = \beta = 1/2$。HBN的具体学习过程如算法 3 所示。

**算法 3**. HBN算法.

输入：训练集$\mathcal{D} = \{X_1, X_2, \dots, X_n, Y\}$，其中$Y = \{y_1, y_2, \dots, y_m\}$，测试实例$\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$和参数 $k$;

输出：预测类标签$y^*$.

1. 根据类标签$Y$对$\mathcal{D}$进行划分，得到$m$个训练子集$\mathcal{D}_i$，其中$1 \le i \le m$；
2. FOR $i = 1 \to m$：
   2.1 根据算法1，基于训练子集$\mathcal{D}_i$和$k$构建子模型$\mathrm{HBN}^i$；
   END FOR
3. 根据算法2，基于测试实例$\boldsymbol{x}$和$k$构建子模型$\mathrm{HBN}^{\boldsymbol{x}}$；
4. 使用公式（12）选出与子模型$\mathrm{HBN}^{\boldsymbol{x}}$的拓扑结构最相似的第$t$个子模型$\mathrm{HBN}^t$，其中
   $t = \arg \min_{\substack{i \in [1,m] \\ y \in Y}} \left\{ H\left( P_{\mathrm{HBN}^i}(y, \boldsymbol{x}), P_{\mathrm{HBN}^{\boldsymbol{x}}}(y, \boldsymbol{x}) \right) \right\}$；
5. 返回$y^* = \arg\max_{y \in Y} \frac{P_{\mathrm{HBN}^t}(y|\boldsymbol{x}) + P_{\mathrm{HBN}^{\boldsymbol{x}}}(y|\boldsymbol{x})}{2}$.

在HBN模型训练阶段，计算类条件熵的时间复杂度为$O(t(nv)^2)$，计算局部条件熵的时间复杂度为$O(tmn^2)$，其中$t$为实例个数，$n$为属性个数，$v$为属性是离散属性可能取值的个数，$m$为类标签个数。模型匹配的时间复杂度为$O(m)$。因此HBN在训练阶段所需的总时间复杂度为$O(t(nv)^2 + tmn^2 + m)$。在分类阶段，对于任意给定测试实例，HBN根据公式（13）来计算后验概率并进行分类决策，该过程的时间复杂度为$O(nmk)$。

# 4 实验与分析

## 4.1 实验数据与方法

本文使用UCI机器学习库中的35个数据集进行实验，数据集的具体描述如表1所示。所有数据集根据实例个数降序排列，其中离散型属性的缺失值使用训练数据中的众数代替，而连续型属性的缺失值使用训练数据中的平均值代替。使用最小描述长度（Minimum Description Length）方法对连续型属性进行离散化预处理[28]。

表1 数据集

| 序号 | 数据集名称 | 实例数 | 属性数 | 类标签个数 |
|---|---|---|---|---|
| 1 | Labor | 57 | 16 | 2 |
| 2 | Labor-negotiations | 57 | 16 | 2 |
| 3 | Zoo | 101 | 16 | 7 |
| 4 | Promoters | 106 | 57 | 2 |
| 5 | Lymphography | 148 | 18 | 4 |
| 6 | Wine | 178 | 13 | 3 |
| 7 | Glass-id | 214 | 9 | 3 |
| 8 | Hungarian | 294 | 13 | 2 |
| 9 | Heart-disease-c | 303 | 13 | 2 |
| 10 | Soybean-large | 307 | 35 | 19 |
| 11 | Primary-tumor | 339 | 17 | 22 |
| 12 | Dermatology | 366 | 34 | 6 |
| 13 | Musk1 | 476 | 166 | 2 |
| 14 | Chess | 551 | 39 | 2 |
| 15 | Balance-scale | 625 | 4 | 3 |
| 16 | Soybean | 683 | 35 | 19 |
| 17 | Breast-cancer-w | 699 | 9 | 2 |
| 18 | Vehicle | 846 | 18 | 4 |
| 19 | German | 1000 | 20 | 2 |
| 20 | Led | 1000 | 7 | 10 |
| 21 | Contraceptive-mc | 1473 | 9 | 3 |
| 22 | Yeast | 1484 | 8 | 10 |
| 23 | Mfeat-mor | 2000 | 6 | 10 |
| 24 | Hypothyroid | 3163 | 25 | 2 |
| 25 | Splice-c4.5 | 3177 | 60 | 3 |
| 26 | Abalone | 4177 | 8 | 3 |
| 27 | Waveform-5000 | 5000 | 40 | 3 |
| 28 | Phoneme | 5438 | 7 | 50 |
| 29 | Page-blocks | 5473 | 10 | 5 |
| 30 | Seer_mdl | 18962 | 13 | 2 |
| 31 | Magic | 19020 | 10 | 2 |
| 32 | Adult | 48842 | 14 | 2 |
| 33 | Shuttle | 58000 | 9 | 7 |
| 34 | Waveform | 100000 | 21 | 3 |
| 35 | Census-income | 299285 | 41 | 2 |

为了说明算法的可靠性和有效性，本文从以下四个方面来验证算法：1）0-1 损失函数（0-1 loss function）：比较分类器在不同数据集下的误分类率；2）偏差（bias）：衡量预测概率值与真实概率值之间的偏离关系；3）方差（variance）：度量在不同迭代阶段下预测概率值的变化波动情况；4）时间负荷：用于度量训练分类器和分类决策所需时间的差异性。实验选取如下算法与本文所提出算法HBN进行比较：

- 基于相关性的NB属性加权过滤算法（Correlation-based feature weighting filter for NB，CFWNB）[17]；
- 选择性KDB（Selective KDB，SKDB）[19]；
- 加权AODE（Weighted AODE，WAODE）[23]；
- 随机森林（Random forest，RF）[29]；
- 基于属性和实例加权的NB算法（Attribute and

instance weighted naive Bayes, AIWNB）[30]; 并使用十折交叉验证法来获得数据集上每个分类器的分类结果。AIWNB 算法有两种学习权重策略，分别为主动式学习（Eager learning）和懒惰式学习（Lazy learning）方法：前者基于训练数据学习权重并记为 AIWNB$^{E}$，后者基于测试实例学习权重并记为 AIWNB$^{L}$。由于篇幅所限，本文所涉及的所有基本实验数据可通过访问网址 http://github.com/Bayes514/HBN 获取。

### 4.2 实验结果对比与分析

网络拓扑结构复杂度随着阶数的增加而增加，拟合数据的能力也越强。相比之下，虽然 AODE 仅能表达一阶的依赖关系，但是它集成了$n$个 ODE 子模型的依赖关系。本文将 BNC 的结构复杂度限制为二阶，即在 HBN 和 SKDB 中$k=2$。而 RF 中决策子树的数目设置为 10。本文使用 Win/Draw/Loss（W/D/L）来记录在给定评估函数条件下，算法 A 与算法 B 相比表现更优/类似/更劣的数据集的数量。如果两个算法的分类性能指标的差异度小于 5%，则视为分类性能相等。

#### 4.2.1 0-1 损失函数

表 2 给出了本文所涉及算法在 0-1 损失函数上的 W/D/L 比较记录。很明显， HBN 在 0-1 损失函数上明显优于 CFWNB（17 优 5 劣）、SKDB（14 优 5 劣）、AIWNB$^{E}$（17 优 7 劣）和 AIWNB$^{L}$（17 优 8 劣）。而与集成模型 WAODE 和 RF 相比，HBN 也取得了明显优势（分别为 11 优 2 劣和 19 优 9 劣）。

表 2 W/D/L 在 0-1 损失函数上的比较结果

| W/D/L | CFWNB | SKDB | WAODE | RF | AIWNB$^{E}$ | AIWNB$^{L}$ |
|---|---|---|---|---|---|---|
| SKDB | 13/10/12 | | | | | |
| WAODE | 15/13/7 | 12/14/9 | | | | |
| RF | 12/5/18 | 9/10/16 | 9/8/18 | | | |
| AIWNB$^{E}$ | 7/24/4 | 11/9/15 | 7/12/16 | 17/7/11 | | |
| AIWNB$^{L}$ | 14/17/4 | 12/9/14 | 10/13/12 | 19/5/11 | 7/26/2 | |
| HBN | **17/13/5** | **14/16/5** | **11/22/2** | **19/7/9** | **17/11/7** | **17/10/8** |

目前社会应用中对于能处理海量数据的自适应分类模型的需求日益迫切，本文使用了净胜累计函数（Goal Difference，*GD*）[31]来评估分类器 A 和 B 在处理不同规模数据集$\mathcal{D}$时的分类性能差异性，具体定义如下：

$$GD(\mathrm{A};\mathrm{B}|\mathcal{D}) = |WIN| - |LOSS| \tag{14}$$

其中$|WIN|$和$|LOSS|$分别表示分类器 A 在给定的评估函数条件下表现优于或劣于分类器 B 的数据集的数量。图 8 给出了 HBN 与 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 在 0-1 损失函数上的净胜累计拟合曲线，其中 X 轴表示数据集序号（按实例数大小降序排序），Y 轴表示分类器 A 相对于分类器 B 在当前数据集上的净胜累计数。

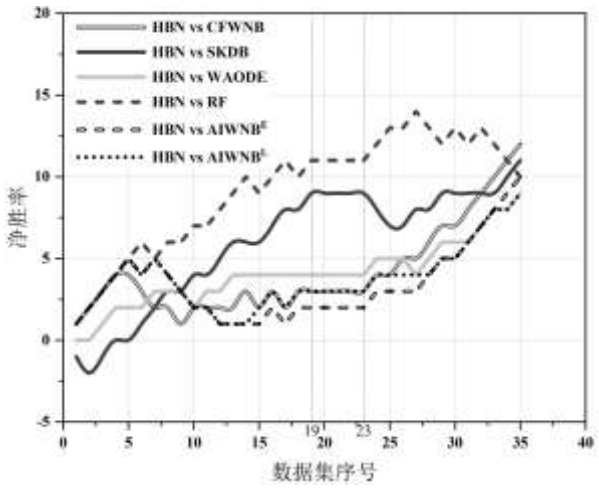



图 8 HBN 相对于 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 在 0-1 损失函数上的净胜累计拟合曲线

如图 8 所示，当数据规模大于 1000（数据集序号 19）且小于 2000（数据集序号 23）时，HBN 相对于所有比较算法的净胜累计拟合曲线较为平稳，说明各算法在该数据规模范围内的 0-1 损失函数结果没有显著性差异。当数据规模大于 2000 时，HBN 相对于 CFWNB、WAODE、AIWNB$^{E}$ 和 AIWNB$^{L}$ 的净胜累计拟合曲线与数据规模呈正相关。当数据规模小于 1000 时，HBN 相对于 SKDB 和 RF 的净胜累计拟合曲线与数据规模呈正相关。这说明基于熵函数学习因果依赖关系有助于提升 BNC 的分类性能。

#### 4.2.2 偏差和方差

基于采样统计理论，通过对 0-1 损失函数进行分解可以得到偏差和方差值[32]，分别描述分类器对训练数据的拟合能力和抗训练数据扰动的鲁棒性。低偏差往往标志着算法在处理大数据方面的优越性能[19]。就单模型而言（如 CFWNB、SKDB、

AIWNB$^{E}$ 和 AIWNB$^{L}$），拓扑结构复杂导致低偏差和高方差，而结构简单则导致高偏差和低方差。集成模型（如 WAODE 和 RF）可以通过限定每个子模型的复杂度实现低方差，并利用集成机制实现低偏差。HBN 也是基于该学习策略实现偏差和方差性能指标的全面提升。表 3 和 4 分别给出了本文所涉及算法在偏差和方差上的 W/D/L 比较记录。

表 3 W/D/L 在偏差上的比较结果

| W/D/L | CFWNB | SKDB | WAODE | RF | AIWNB$^{E}$ | AIWNB$^{L}$ |
|---|---|---|---|---|---|---|
| SKDB | 22/5/8 | | | | | |
| WAODE | 22/4/9 | 10/12/13 | | | | |
| RF | 8/0/27 | 4/2/29 | 7/1/27 | | | |
| AIWNB$^{E}$ | 9/18/8 | 7/5/23 | 8/8/19 | 27/0/8 | | |
| AIWNB$^{L}$ | 14/16/5 | 7/6/22 | 9/8/18 | 28/0/7 | 11/23/1 | |
| HBN | **26/3/6** | **10/20/5** | **15/13/7** | **29/2/4** | **22/7/6** | **22/7/6** |

表 4 W/D/L 在方差上的比较结果

| W/D/L | CFWNB | SKDB | WAODE | RF | AIWNB$^{E}$ | AIWNB$^{L}$ |
|---|---|---|---|---|---|---|
| SKDB | 5/4/26 | | | | | |
| WAODE | 11/2/22 | 25/3/7 | | | | |
| RF | 1/0/34 | 8/1/26 | 5/2/29 | | | |
| AIWNB$^{E}$ | 1/14/20 | 23/4/8 | 21/1/13 | 32/1/2 | | |
| AIWNB$^{L}$ | 4/9/22 | 24/5/6 | 21/1/13 | 32/2/1 | 10/24/1 | |
| HBN | **7/1/27** | **20/6/9** | **7/7/21** | **26/6/3** | **9/3/23** | **8/4/23** |

如表 3 所示，在偏差方面，HBN 与其它算法相比有着明显的优势，例如 HBN 与 CFWNB 相比 26 优 6 劣，与 SKDB 相比 10 优 5 劣，与 WAODE 相比 15 优 7 劣，与 RF 相比 29 优 4 劣，与 AIWNB$^{E}$ 和 AIWNB$^{L}$ 相比 22 优 6 劣。由于 CFWNB、WAODE、AIWNB$^{E}$ 和 AIWNB$^{L}$ 没有结构学习过程，其拓扑结构不受训练数据扰动的影响。因此，如表 4 所示，这四个模型的方差显著低于其它算法。而 HBN 的局部拓扑结构能充分体现测试实例中隐含的因果关系，在一定程度上减轻训练数据过拟合带来的负面影响。因此，与 SKDB 和 RF 相比 HBN 的方差结果均明显占优（20 优 9 劣，26 优 3 劣）。可见，基于交叉熵的模型匹配机制有助于充分挖掘训练数据中属性间所涉及的一般因果关系以及那些隐藏在未标记测试实例中属性值所涉及的特殊因果关系。

图 9 和 10 分别给出了 HBN 相对于 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 在偏差和方差上的净胜累计拟合曲线。如图 9 所示，随着数据集规模的增加，HBN 的偏差明显优于 CFWNB、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$。当数据规模小于 303（数据集序号 9），HBN 的偏差优于 SKDB；当数据规模大于 303 时，二者偏差结果近似相等。当数据规模大于 1000 时（数据集序号 20），HBN 的偏差优于 WAODE。

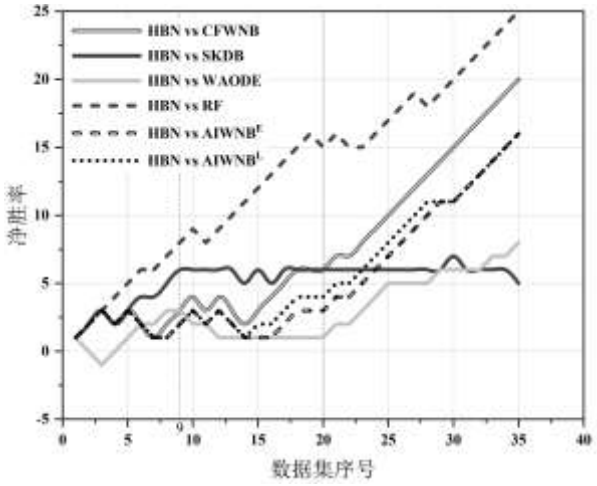



图 9 HBN 相对于 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 在偏差上的净胜累计拟合曲线

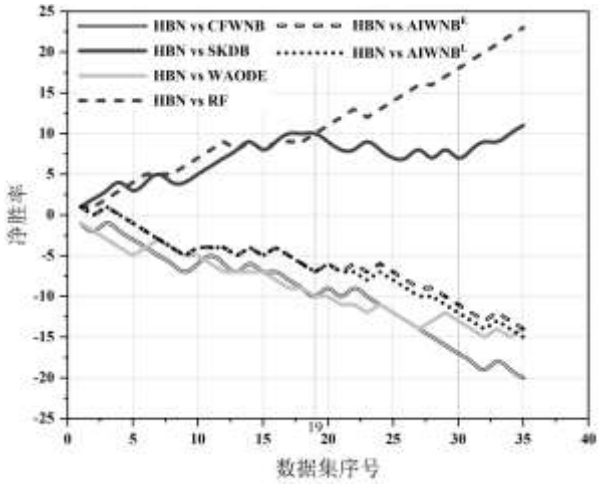



图 10 HBN 相对于 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 在方差上的净胜累计拟合曲线

如图 10 所示，HBN 相对于 RF 的方差净胜累计拟合曲线与数据规模明显呈正相关；当数据规模小于 1000（数据集序号 19）且大于 18962（数据集序号 30）时，HBN 相对于 SKDB 的方差净胜累计拟合曲线与数据规模呈正相关。因此就偏差和方差性

能指标而言，HBN 相对于 SKDB 体现了方差优势，相对于 RF 则体现了全面的优势。CFWNB、WAODE、AIWNB$^{E}$ 和 AIWNB$^{L}$ 由于其拓扑结构与训练数据无关的特性，天然具有在方差方面的优势。HBN 相对于 CFWNB、WAODE、AIWNB$^{E}$ 和 AIWNB$^{L}$ 的方差净胜累计拟合曲线与数据集规模明显呈负相关。值得指出的是，相对于其它算法，HBN 的局部结构可以更灵活、精准地表达各属性在不同情境下的因果关系，为精准知识表达提供了一个可行的思路。

### 4.3 时间结果对比与分析

图 11 和 12 分别给出了所有算法的训练和分类时间的比较结果，其中柱状图分别代表各算法在 35 个数据集上进行十折交叉验证时间结果的平均值。如图 11 所示，由于 CFWNB、WAODE、AIWNB$^{E}$ 和 AIWNB$^{L}$ 没有结构学习过程，因此其平均训练时间最少。而对于具有结构学习过程的算法而言，HBN 的平均训练时间明显小于 SKDB 和 RF。如图 12 所示，由于 HBN 仅需根据训练好的贝叶斯网络拓扑结构计算测试实例的后验概率，因此与 WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 相比取得了明显的分类时间优势，而与 CFWNB 和 SKDB 相比，HBN 的分类时间优势显著性则较小。

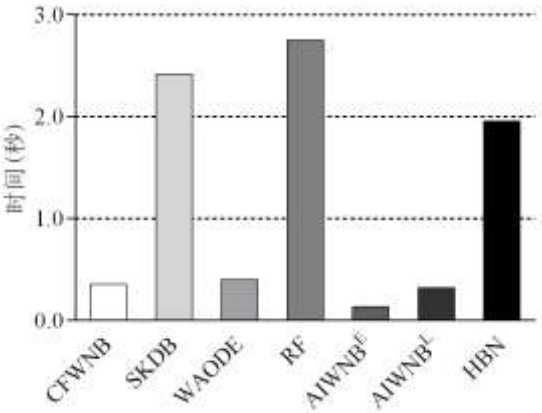



图 11 训练时间比较结果

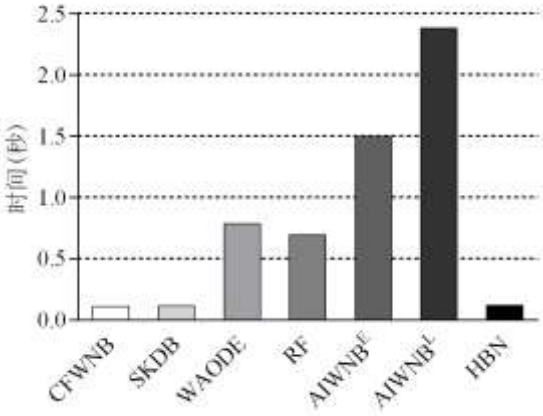



图 12 分类时间比较结果

## 5 总结与展望

条件互信息的对称性决定了其只能描述属性之间的无向相关性，而非有向因果性。这使得 BNC 只能通过人工定向的方式构建有向无环图。此外，条件互信息还无法度量整体贝叶斯网络拓扑结构对数据的拟合程度。这都将导致最终构建的 BNC 是次优的。针对上述问题，本文提出了类条件熵 $H(X_i|\boldsymbol{\Pi}_i^{\mathcal{B}}, y)$ 和局部条件熵 $H(x_i|\pi_i^{\mathcal{B}}, Y)$ 来分别构建可以兼顾属性（值）间所涉及的一般（特殊）因果关系的 BNC，最终的分类决策由基于交叉熵的模型匹配算法实现。在 35 个 UCI 开源数据集上的实验结果表明，该算法在 0-1 损失函数、偏差和方差方面与 CFWNB、SKDB、WAODE、RF、AIWNB$^{E}$ 和 AIWNB$^{L}$ 等算法相比具有明显综合优势。

本文证明了算法的合理性和有效性。一般来说，集成模型的分类效果要比单一模型更加准确，而模型集成的方法有很多种，如线性组合、投票法和提升法。如果子分类器对数据有着不同的分类精度，线性非平均组合的方法在理论上将更适合 HBN。无标注测试实例是不完全的，根据测试实例构建的子模型也不够精准，因此本文仅从效率角度出发采用了简单的平均方法，并没有提出确定线性组合中每个子分类器权系数的学习策略。因此，如何从测试实例中挖掘足够“正确”的知识来修正子分类器的权重将是我们未来工作的主要研究方向。

## 参 考 文 献

[1] Li Hang. Statistics learning method. Beijing: Tsinghua university press, 2012 (in Chinese)

(李航. 统计学习方法. 北京: 清华大学出版社, 2012.)

[2] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco,USA: Morgan Kaufmann, 1988.

[3] Yang Y, Gao X, Guo Z, et al. Learning Bayesian networks using the constrained maximum a posteriori probability method. Pattern Recognition, 2019, 91: 123-134.

[4] Yang X, Guo Y, Liu Y. Bayesian-inference-based recommendation in online social networks. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(4): 642-651.

[5] Varshney D, Kumar S, Gupta V. Predicting information diffusion probabilities in social networks: A Bayesian networks based approach. Knowledge-Based Systems, 2017, 133: 66-76.

[6] Chen Wei, Zhu Biao, Zhang Hongxin. BN-Mapping:Visual Analysis of Geospatial Data with Bayesian Network. Chinese Journal of Computers, 2016, 39(7): 1281-1293 (in Chinese)
(陈为, 朱标, 张宏鑫. BN-Mapping：基于贝叶斯网络的地理空间数据可视分析. 计算机学报, 2016, 39(7): 1281-1293.)

[7] Wang Shuangcheng, Gao Rui, Du Ruijie. Restricted Bayesian Network Classifier Based on Gaussian Copula. Chinese Journal of Computers, 2016, 39(8): 1612-1625 (in Chinese)
(王双成, 高瑞, 杜瑞杰. 基于高斯 Copula 的约束贝叶斯网络分类器研究. 计算机学报, 2016, 39(8): 1612-1625.)

[8] Wang Shuangcheng, Zhang Li, Zheng Fei. Asynchronous dynamic Bayesian network classifiers. Chinese Journal of Computers, 2020, 43(9): 1737-1754 (in Chinese)
(王双成, 张立, 郑飞. 异步动态贝叶斯网络分类器研究. 计算机学报, 2020, 43(9): 1737-1754.)

[9] Shannon C E. A mathematical theory of communication. The Bell System Technical Journal, 1948, 27(4): 623-656.

[10] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997, 29(2-3): 131-163.

[11] Sahami M. Learning limited dependence Bayesian classifiers// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. New York,USA, 1996: 335-338.

[12] Webb G I, Boughton J R, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. Machine Learning, 2005, 58(1): 5-24.

[13] Wang L M, Chen S L, Mammadov M. Target learning: a novel framework to mine significant dependencies for unlabeled data//Proceedings of Advances in Knowledge Discovery and Data Mining. Berlin,Germany: Springer, 2018: 106-117.

[14] Dua D, Graff C, UC Irvine machine learning repository, http://archive.ics.uci.edu/ml 2020,6,17 access

[15] Wang Zhongfeng, Wang Zhihai. An optimization algorithm of Bayesian network classifiers by derivatives of conditional log likelihood. Chinese Journal of Computers, 2012, 35(2): 364-374 (in Chinese)
(王中锋, 王志海. 基于条件对数似然函数导数的贝叶斯网络分类器优化算法. 计算机学报, 2012, 35(2): 364-374.)

[16] Cui Jiaxu, Yang Bo. Survey on Bayesian Optimization Methodology and Applications. Journal of Software, 2018, 29(10): 3068-3090 (in Chinese)
(崔佳旭, 杨博. 贝叶斯优化方法和应用综述. 软件学报, 2018, 29(10): 3068-3090.)

[17] Jiang L, Zhang L, Li C, et al. A correlation-based feature weighting filter for naive bayes. IEEE transactions on knowledge and data engineering, 2018, 31(2): 201-213.

[18] Ju Zhuoya, Wang Zhihai. A Bayesian Classification Algorithm Based on Selective Patterns. Journal of Computer Research and Development, 2020, 57(8): 1605-1616 (in Chinese)
(鞠卓亚, 王志海. 基于选择性模式的贝叶斯分类算法. 计算机研究与发展, 2020, 57(8): 1605-1616).

[19] Wang Shuangcheng, Gao Rui, Du Ruijie. With Super Parent Node Bayesian Network Ensemble Regression Model for Time Series. Chinese Journal of Computers, 2017, 40(12): 2748-2761 (in Chinese)
(王双成, 高瑞, 杜瑞杰. 具有超父结点时间序列贝叶斯网络集成回归模型. 计算机学报, 2017, 40(12): 2748-2761.)

[20] Qi Xiaolong, Gao Yang, Wang Hao, Song Bei, Zhou Chunlei,Zhang Youwei. A Measurable Bayesian Network Structure Learning Method. Journal of Computer Research and Development, 2018, 55(8): 1717-1725 (in Chinese)
(綦小龙，高阳，王皓，宋蓓，周春蕾，张友卫. 一种可度量的贝叶斯网络结构学习方法. 计算机研究与发展, 2018, 55(8): 1717-1725.)

[21] Martínez A M, Webb G I, et al. Scalable learning of Bayesian network classifiers. The Journal of Machine Learning Research, 2016, 17(1): 1515-1549.

[22] Duan Zhiyi, Limin Wang, Shenglei Chen and Minghui Sun, Instance-based Weighting Filter for SuperParent One-dependence Estimators. Knowledge-based Systems, Netherlands: Elsevier, 2020, 203: 106085.

[23] Xiang Z L, Kang D K. Attribute weighting for averaged one-dependence estimators. Applied Intelligence, 2017, 46(3): 616-629.

[24] Rubinstein R Y. Optimization of computer simulation models with rare events. European Journal of Operational Research, 1997, 99(1): 89-112.

[25] Hui K P, Bean N, Kraetzl M, et al. The cross-entropy method for network reliability estimation. Annals of Operations Research, 2005, 134(1): 101.

[26] Geng L, Zhao Y, Li W. Enhanced cross entropy method for composite power system reliability evaluation. IEEE Transactions on Power Systems, 2019, 34(4): 3129-3139.

[27] Joseph A G, Bhatnagar S. An online prediction algorithm for reinforcement learning with linear function approximation using cross entropy method. Machine Learning, 2018, 107(8-10): 1385-1429.

[28] Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning//Proceedings of the 13th International Joint Conference on Artificial Intelligence. Berlin,Germany: Springer, 1993: 1022-1029.

[29] Breiman L. Random Forest. Machine Learning, Netherlands: Springer, 2001, 45:5-32.

[30] Zhang H, Jiang L X, Yu L J. Attribute and instance weighted naive Bayes. Pattern Recognition. 2021, 111: 107674.

[31] Duan Z Y, Wang L M. *K*-dependence Bayesian classifier ensemble. Entropy. 2017, 19(12): 651.

[32] Kohavi R, Wolpert D H. Bias plus variance decomposition for zero-one loss functions//Proceedings of the 13th International Conference on Machine Learning. San Francisco,USA: Morgan Kaufmann, 1996: 275-83.

**Liu Yang,** Ph.D. candidate. His main research interests include data mining and Bayesian network.

**Wang Li-Min,** Ph.D., professor, PhD supervisor. His main research interests include probabilistic logic inference and Bayesian network.

**Sun Ming-Hui,** Ph.D., associate professor. His main research interest includes artificial intelligence.

**Background**

Bayesian network classifiers (BNC) are powerful tools for knowledge representation and inference under conditions of uncertainty. They have widely used in data mining and artificial intelligence due to its excellent classification performance and interpretability. Information theory has established a strong mathematical basis for its rapid development. For example, to overcome the conditional independent assumption of naïve Bayes, many researchers use conditional mutual information to measure the conditional dependence between attributes in the topology structure of BNCs, including single models (e.g., tree-augmented naïve Bayes and *k*-dependence Bayesian network classifier) and ensemble models (e.g., averaged one-dependence estimators). However, identifying the dependence relationships between attributes and constructing a BNC based on information measures (such as conditional mutual information) cannot accurately express the dynamic change of the dependence relationships when attributes take specific values. In addition, although Bayesian network is also called belief network or causal network, the symmetry of conditional mutual information expression determines that they can only describe undirected dependencies (rather than directed causality). Based on the directed acyclic characteristic of the Bayesian network, most of the existing BNCs use the artificial defined arc strategy, which cannot reflect the real causal relationship between attributes.

This paper proves that conditional mutual information is essentially a measure of the difference between two local topologies of conditional (in)dependence in terms of the entropy functions and cannot be used to measure the data fitting of the overall topology. The resulting model will be sub-optimal. Therefore, this paper defines the mapping relationship between the joint entropy function and the joint probability distribution in the Bayesian network from the perspective of the log-likelihood function, and uses conditional entropy to identify the attribute dependency relationship in the network and prove its rationality. On this basis, a label-driven heuristic structure learning method is proposed to construct a BNC that that can achieve the trade-off between fitting and generalization. Experimental evaluation on 35 datasets from the UCI machine learning repository shows that the proposed algorithm has significant advantages in terms of classification performance over other state-of-the-art algorithms, and the algorithm is effective and feasible for uncertain knowledge representation and reasoning.

This work is supported by the National Key R&D Program of China under Grant No. 2019YFC1804804 and the Scientific and Technological Developing Scheme of Jilin Province under Grant No. 20200201281JC.