Statistics and Application 统计学与应用, 2025, 14(11), 54-66
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411310

Hans汉斯

不均衡数据集条件下基于熵的自适应BNs参数学习

刘　蓉，刘　赪

西南交通大学数学学院/统计系，四川　成都

收稿日期：2025年10月5日；录用日期：2025年10月26日；发布日期：2025年11月10日

摘　要

针对不均衡数据集条件下构建贝叶斯网络易出现零概率值问题，提出一种无需专家依赖的基于熵的自适应贝叶斯网络参数学习方法。首先，本文用熵量化数据的不均衡性，并将这种不均衡性以正态分布的形式作为先验信息，利用标准条件熵以及3σ准则构造自适应方差，最大似然估计得到均值，并通过网格搜索以及交叉验证寻找最优容许误差；然后，用正态分布作为Dirichlet分布的近似，结合最大后验概率估计计算网络参数值；最后，在不同样本量、不同网络的数据集下进行实验测试，并将本文方法与其他3种主要方法进行比较。结果表明：在不均衡数据集条件下，本文方法无需专家依赖且参数学习精度都优于其他3种方法。

关键词

不均衡数据集，熵，贝叶斯网络，参数学习

Parameter Learning of Adaptive BNs Based on Entropy under Imbalanced Dataset Conditions

Rong Liu, Cheng Liu

Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu Sichuan

Received: October 5, 2025; accepted: October 26, 2025; published: November 10, 2025

Abstract

To address the issue of zero probability values that often arises when constructing Bayesian networks from imbalanced datasets, this paper proposes an expert-independent adaptive parameter learning method for Bayesian networks based on entropy. First, entropy is used to quantify the

文章引用: 刘蓉, 刘赪. 不均衡数据集条件下基于熵的自适应 BNs 参数学习[J]. 统计学与应用, 2025, 14(11): 54-66.
DOI: 10.12677/sa.2025.1411310

degree of imbalance in the data. This imbalance is then incorporated as prior information in the form of a normal distribution. An adaptive variance is constructed using standard conditional entropy and the three-sigma rule, while the mean is derived via maximum likelihood estimation. The optimal permissible error is identified through grid search and cross-validation. Subsequently, the normal distribution is used as an approximation of the Dirichlet distribution, and network parameters are calculated by integrating maximum a posteriori estimation. Finally, experiments are conducted on datasets with varying sample sizes and network structures, and the proposed method is compared with three other major approaches. The results demonstrate that, under imbalanced dataset conditions, the proposed method achieves higher parameter learning accuracy without relying on expert knowledge compared to the other three methods.

Keywords

Imbalanced Dataset, Entropy, Bayesian Network, Parameter Learning

1. 引言

贝叶斯网络(Bayesian Networks, BNs) [1]通过有向无环图和条件概率分布，为多变量间的依赖关系提供了一种严格的概率图模型表示，其核心环节之一——参数学习直接决定了模型能否准确进行变量间的推理。然而在实际应用中，采集到的训练数据往往呈现出显著的不均衡特性，这对参数学习的鲁棒性构成了严峻挑战。以图 1 所示的“草坪湿润”网络为例，节点“下雨”(R)和“洒水”(S)作为父节点，共同影响子节点“草坪湿润”(W)。在实际观测数据中，某些父节点组合如“不下雨 - 不洒水”对应子节点 W 的概率表状态为“草坪湿润”，但由于观测样本有限以及这类情况在现实生活中出现频次很低，导致该组合状态下数据未被观测到，最终得到分布不均衡的数据。采用传统最大似然估计(Maximum likelihood Estimation, MLE) [2]方法计算条件概率会出现零概率值，直接影响 BNs 的推理性能。

针对此问题，研究人员在参数学习时融入专家给定的参数先验知识，目前研究方法主要分为两类：一是将参数学习问题转化为基于似然函数或熵函数的约束优化问题，并借助凸优化[3]或梯度法[4]求解，如文献[3]提出约束最大似然算法，将最大熵函数作为准则，与非单调性约束结合转化为凸优化问题，但当数据分布不均衡时，容易导致专家给出的约束变为空可行集，直接导致凸优化无解。二是将约束与专家先验知识融合，转化为虚拟样本的形式结合贝叶斯估计求解[5]-[11]，如文献[9]用均匀分布描述参数取值区间，结合单调性约束推断 Dirichlet 分布超参数，最后结合最大后验概率估计(Maximum a posterior, MAP) [12]求解；文献[10] [11]则用正态分布描述近似不等式约束，同样作为 Dirichlet 分布的近似求解超参数，这类方法有效解决了零概率值问题并提高了网络的参数学习精度。Dirichlet 分布是多项分布的共轭先验且能简化计算，但其超参数的确定高度依赖于专家领域知识，且当变量分类过多时往往难以确定约束类型，其适用性显著降低，因此如何在不依赖专家知识的情形下确定 Dirichlet 分布的超参数，从而求得网络参数值成为一个亟待解决的问题。

为减少对专家知识的依赖并提高参数学习精度，本文选择从数据本身出发，探索其内在结构信息以构建适用的先验信息。在参数学习时，数据的不均衡性体现为不同概率分布的不确定性存在差异。以天气预测为例，均匀分布 P (下雨) = 0.5、P (不下雨) = 0.5 的不确定性，要强于倾斜分布 P (下雨) = 0.8、P (不下雨) = 0.2 的不确定性，那么如何衡量这两种不确定性的强弱呢？在信息论中，熵[13]作为度量随机

变量的不确定性的核心指标，能够精确量化这种差异：熵值越大，表明分布越均衡，不确定性越强；熵值越小，则表明分布越不均衡，不确定性越弱。

因此，本文提出用熵刻画数据的内在结构信息并用正态分布表示，并设计一种不均衡数据集条件下自适应 BNs 参数学习方法。首先，利用熵和 3σ 准则构造自适应方差，MLE 计算均值，然后通过网格搜索和交叉验证寻找最优容许误差，得到超参数的先验正态分布，将其作为 Dirichlet 分布的近似分布，通过目标优化求解超参数，最后结合 MAP 计算网络参数值。

2. 贝叶斯网络相关理论

2.1. 贝叶斯网络

贝叶斯网络[14]由结构 $\mathcal{N}$ 和参数 Θ 两部分组成，结构 $\mathcal{N}$ 是一种有向无环图 $\mathcal{N}=(V,E)$，其中节点变量集合 $V=\{X_1,\cdots,X_n\}$，有向边集合 E 代表变量之间的直接依赖关系，$\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 是网络结构中的参数集合，描述变量对其父节点的依赖关系。每个节点都附有一个概率分布，根节点 X 所附的是它的边缘分布 $P(X)$，而非根节点 X 所附的是其条件概率分布 $P(X\mid\pi(X))$。

贝叶斯网是联合分布的分解的一种表示，基于条件独立性假设(马尔科夫假设)，把各变量所附的条件概率分布相乘就得到联合分布，即

$$P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P(X_i\mid\pi(X_i)), \tag{1}$$

其中当 $\pi(X_i)=\varnothing$ 时，$P(X_i\mid\pi(X_i))$ 为边缘分布 $P(X_i)$。

贝叶斯网络结构学习就是对于给定的数据集 $D=\{D_1,D_2,\cdots,D_m\}$，找到一个与数据集拟合最好的网络。而在有了具体的网络之后，需要找到与之对应的条件概率表(conditional probability table, CPT)即参数学习。本文就是在结构已知的情况，探讨 BNs 参数学习问题。

2.2. 最大似然估计(MLE)

根据文献[14]，考虑一个由 n 个变量 $V=\{X_1,X_2,\cdots,X_n\}$ 组成的贝叶斯网络 $\mathcal{N}$。不失一般性，设其中节点 X_i 共有 r_i 个取值 $1,2,\cdots,r_i$，其父节点 $\pi(X_i)$ 的取值共有 q_i 个组合，$1,2,\cdots,q_i$。若 X_i 无父节点，则 $q_i=1$。那么，网络的参数为

$$\theta_{ijk}=P(X_i=k\mid\pi(X_i)=j), \tag{2}$$

其中 i 的取值范围是 $1\sim n$，而对一个固定的 i，j 和 k 的取值范围分别是从 $1\sim q_i$ 及从 $1\sim r_i$。

设 m_{ijk} 是数据中满足 $X_i=k$ 和 $\pi(X_i)=j$ 的样本的数量，于是由文献[13]可知参数 θ_{ijk} 的最大似然估计值为：

$$\theta_{ijk}^{MLE}=\begin{cases}\dfrac{m_{ijk}}{\sum_{k=1}^{r_i}m_{ijk}}, & \text{若 } \sum_{k=1}^{r_i}m_{ijk}>0,\\ \dfrac{1}{r_i}, & \text{若否.}\end{cases} \tag{3}$$

2.3. 最大后验概率估计(MAP)

根据文献[12]，贝叶斯公式将先验分布和似然函数结合，得 θ 到后验分布即：

$$P(\theta\mid D)\propto P(\theta)L(\theta\mid D), \tag{4}$$

式(4)中，概率分布 $P(\theta)$ 表示 θ 的先验知识，似然函数 $L(\theta|D)=P(D|\theta)$ 表示数据集 D 的影响。在独立同分布(independent and identically distributed, i.i.d.)条件下，式(4)中 $L(\theta|D)$ 为多项式似然函数的共轭分布。本文假设先验分布 $P(\theta)$ 是 Dirichlet 分布，即：

$$P(\theta)=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1}. \tag{5}$$

则后验分布 $P(\theta|D)$ 也是 Dirichlet 分布。在观测样本下，θ 的后验分布 $P(\theta|D)$ 为：

$$P(\theta|D)\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}. \tag{6}$$

由文献[10]，当参数的后验分布取极大值时，得到该参数的贝叶斯最大后验估计值为：

$$\theta_{ijk}^{MAP}=\frac{m_{ijk}+\alpha_{ijk}}{\sum_{k=1}^{r_i}\left(m_{ijk}+\alpha_{ijk}\right)}. \tag{7}$$

本文的无信息先验 MAP 估计方法则是令 $\alpha_{ijk}=1$，即默认虚拟样本为 1。

2.4. 均匀约束先验方法

根据文献[9]，根据单调性和规范性得到每个参数的取值范围，再无任何先验参数的情况下，认为参数在区间内服从均匀分布，即设参数 $\theta\sim U(\theta_1,\theta_2)$，则

$$\begin{cases}E(\theta)=\dfrac{\theta_1+\theta_2}{2},\\ D(\theta)=\dfrac{(\theta_1-\theta_2)^2}{12}.\end{cases} \tag{8}$$

然后将该均匀分布的均值作为约束，方差作为目标函数，求出相应 Beta 分布的参数，并以该 Beta 分布作为先验参数的分布函数。

3. 基于熵的自适应 BNs 参数学习方法

3.1. 自适应方差

信息熵是刻画随机变量不确定性的经典工具，又称为香农熵或熵，根据文献[13]中 Shannon 对离散随机变量熵的描述，贝叶斯网络 $\mathcal{N}$ 中节点 X_i 的熵可定义为：

$$H(X_i)=\sum_{k=1}^{r_i}P(X_i=k)\log\frac{1}{P(X_i=k)}, \tag{9}$$

其中 r_i 表示节点 X_i 的状态数，为了消除状态数的影响，本文采用标准熵，使得不同变量间、不同网络之间的比较具备一致性。

定义 1 (**标准条件熵**)在贝叶斯网络 $\mathcal{N}$ 中，对于每个条件分布 $P(X_i=k|\pi(X_i)=j)$ 的标准条件熵 NH_{ij} 定义为：

$$NH_{ij}=-\frac{1}{\log r_i}\sum_{k=1}^{r_i}P(X_i=k|\pi(X_i)=j)\log P(X_i=k|\pi(X_i)=j), \tag{10}$$

其中 $\log r_i$ 表示均匀分布时的最大熵值，该度量将熵值规范化至[0, 1]区间。$NH_{ij}=0$ 对应于完全确定的分布(单一事件概率为 1)，$NH_{ij}=1$ 则表示最大不确定性状态。

定义 2 (**自适应方差**) Dirichlet 分布参数估计的渐近分布服从正态分布[15]，又根据文献[10] [11]用正态分布描述近似不等式约束，因此本文有理由设数据结构信息服从正态分布，即设 Dirichlet 分布超参数：

$$\alpha_{ijk} \sim N\left(\mu_{ijk}, \sigma_{ijk}^2\right), \tag{11}$$

其中 μ_{ijk} 为正态分布的数学期望，通过式(3)计算得到； σ_{ijk}^2 是利用标准条件熵以及 3σ 准则构造的自适应方差，本文希望并且对于低熵(不均衡)区域，希望在计算时更依赖于先验信息；对于高熵(均衡)区域，希望先验分布较弱，在做参数估计时更依赖于数据本身，因此其标准差定义为：

$$\sigma_{ijk} = \frac{\varepsilon_{ijk}}{3}\left(1 - NH_{ij}\right), \tag{12}$$

其中容许误差 ε_{ijk} 是一个基础参数，控制每个条件分布的全局先验强度水平，在真实数据中通过网格搜索和交叉验证得到。NH_{ij} 作为熵调节因子，实现局部自适应调节，根据每个条件分布的均衡性程度微调先验强度，熵越大代表分布越均衡，需要的先验信息越少，从而令方差越小，α_{ijk} 更集中在 μ_{ijk} 附近；而熵越小代表分布越不均衡，需要更多的先验信息，从而令方差越大，α_{ijk} 在 μ_{ijk} 附近较为分散。

根据文献[9]-[11]，Dirichlet 分布的边缘分布为 Beta 分布，且 Beta 分布为二项式分布的共轭分布族，因此可以通过最小化方差平方和确定 Beta 分布中的 2 个形参来近似表达与其相关的领域知识。

$$\begin{aligned} &\min\left(D_{N_k} - D_{B_k}\right)^2 \\ &\text{s.t.} \begin{cases} E_{N_k} = E_{B_k} \\ \alpha > 0, \beta > 0 \end{cases} \end{aligned} \tag{13}$$

其中 α, β 是 Beta 分布的两个参数， D_{B_k} 和 E_{B_k} 分别是 Beta 分布的方差和期望，其表达式为：

$$\begin{cases} E_{B_k} = \dfrac{\alpha}{\alpha + \beta}, \\ D_{B_k} = \dfrac{\alpha\beta}{\left(\alpha + \beta\right)^2\left(\alpha + \beta + 1\right)}. \end{cases} \tag{14}$$

利用式(13)和式(14)求得 Beta 分布的两个参数，得到 α, β 后，将其作为虚拟样本代入式(7)，计算网络参数 θ_{ijk} 的估计值。

3.2. 算法描述

给定样本量为 m 的数据集 $D = \{D_1, D_2, \cdots, D_m\}$，基于熵的贝叶斯网络自适应参数学习方法的具体流程如下：

步骤 1 根据式(3)计算得到每个节点的条件概率，将其作为正态分布的均值 μ_{ijk}；

步骤 2 对每个条件分布 $P\left(X_i = k \mid \pi\left(X_i\right) = j\right)$，根据式(10)计算标准熵 NH_{ij}；

步骤 3 利用网格搜索以及 K 折交叉验证选择最优 ϵ_{opt}。将观测数据集 D 划分为 K 个互斥子集 $D = \bigcup_{s=1}^{K} D^{(s)}$，其中 $D^{(i)} \cap D^{(j)} = \varnothing$， $i, j = 1, \cdots, K$， $\forall i \neq j$。对于每个候选 $\epsilon \in \mathcal{E}$，进行交叉验证：

a) 初始化平均对数似然 $\overline{LL_0}(\epsilon) = 0$， $s = 1, \cdots, K$，训练集 $D_{train}^{(s)} = D \setminus D^{(s)}$，验证集 $D_{val}^{(s)} = D^{(s)}$；

b) 使用训练集 $D_{train}^{(s)}$ 和当前 ϵ，生成自适应先验参数 $\alpha_{ijk}(\epsilon)$，然后根据式(7)得到参数 $\theta^{(s)}(\epsilon)$；

c) 验证集评估：计算验证集对数似然 $LL_s(\epsilon) = \sum_{d \in D_{val}^{(s)}} \log P\left(d \mid \theta^{(s)}(\epsilon)\right)$，其中单个样本 d 的概率计算为： $P(d) = \prod_{i=1}^{m} P\left(X_i = d_i \mid \pi_i = d_{\pi_i}, \theta^{(s)}(\epsilon)\right)$；

d) 累加对数似然： $\overline{LL}(\epsilon) \leftarrow \overline{LL_0}(\varepsilon) + LL_s(\epsilon)$，计算平均对数似然： $\overline{LL}(\epsilon) = \dfrac{\overline{LL}(\epsilon)}{K}$；

e) 选择最优 ϵ_{opt}^{*}：$\epsilon_{opt}^{*}=\arg\max_{\epsilon\in\mathcal{E}}\overline{LL}(\epsilon)$。

步骤 4 将最优 ϵ_{opt}^{*} 代入式(12)中得到自适应方差，根据式(13)和式(14)求得 Beta 分布参数，将其代入式(7)，得到最终估计参数：θ_{ijk}^{*}。

4. 实验测试

4.1. 实验内容

本文在完整样本数据集下对本文方法的参数学习精度和运行耗时进行测试，并且与 MLE、无信息先验 MAP (默认虚拟样本为 1)、均匀约束先验[9] 3 种估计方法进行比较。

本文在草坪湿润模型(Grass Wet)、Earthquake、Asia 三个不同复杂度的 BNs 上进行实验测试，其中 Grass Wet 网络如图 1 所示，该模型被广泛应用于评估各种参数学习算法如文献[7]-[10]。每个节点取值为 0 或 1，当节点取值为 0 时，表示事件不发生，取值为 1 时，表示事件发生。该 BNs 节点的真实参数如表 1 所示。

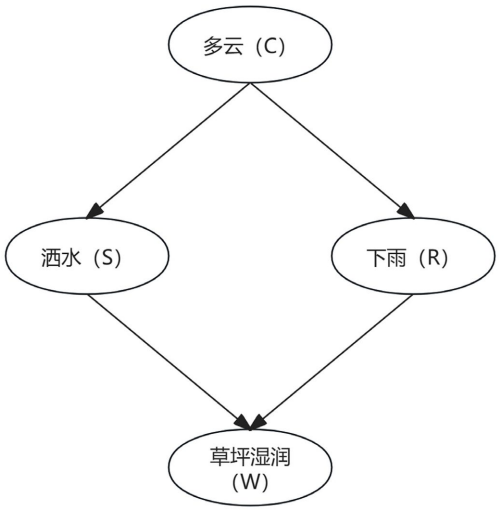


Figure 1. Structure of grass wet Bayesian network
图 1. 草坪湿润推理贝叶斯网络

Table 1. System resulting data of standard experiment
表 1. 草坪湿润网络节点参数

节点	参数值
C	$P(C=1)=0.5$，$P(C=0)=0.5$
S	$P(S=1\|C=0)=0.8$，$P(S=0\|C=0)=0.2$ $P(S=1\|C=1)=0.4$，$P(S=0\|C=1)=0.6$
R	$P(R=1\|C=0)=0.4$，$P(R=0\|C=0)=0.6$ $P(R=1\|C=1)=0.8$，$P(R=0\|C=1)=0.2$
W	$P(W=1\|C=0,R=0)=0.1$，$P(W=0\|C=0,R=0)=0.9$ $P(W=1\|C=0,R=1)=0.8$，$P(W=0\|C=0,R=1)=0.2$ $P(W=1\|C=1,R=0)=0.9$，$P(W=0\|C=1,R=0)=0.1$ $P(W=1\|C=1,R=1)=0.99$，$P(W=0\|C=1,R=1)=0.01$

在实验测试中，根据图 1 所示网络结构和表 1 所示网络真实参数值，采用随机抽样方法生成样本数据。实验测试环境为：Windows 11 系统，处理器为 Inter CPU 2.40 GHz，平台软件为 PyCharm2022.2.2。

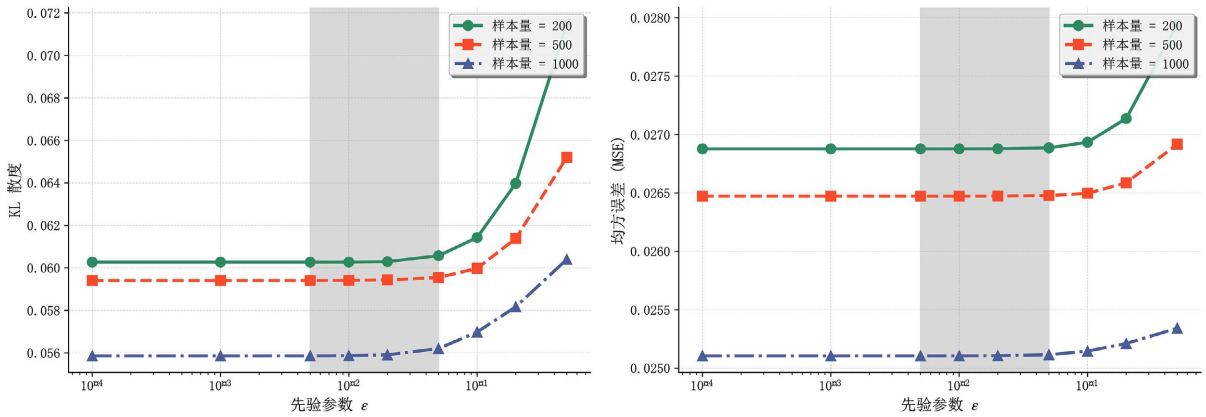
本文实验以 KL 散度和均方误差(Mean Square Error, MSE)为参数学习精度的衡量指标，差异越小说明精度越高，具体计算形式如式(15)和(16) [16]。

$$\mathrm{KL}\left(\theta_i^*\right)=\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\theta_{ijk}\log\frac{\theta_{ijk}}{\theta_{ijk}^*}. \tag{15}$$

$$\mathrm{MSE}\left(\theta_i,\theta_i^*\right)=\frac{1}{q_i r_i}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\left(\theta_{ijk}-\theta_{ijk}^*\right)^2. \tag{16}$$

4.2. 不同 ϵ 取值对 KL 散度和 MSE 的影响

以 Grass Wet 为例，探究本文方法在样本量为 200、500、1000 情况下，ϵ 取值分别为 0.0001、0.001、0.005、0.01、0.05、0.1、0.2、0.5 时对 KL 散度与 MSE 的影响，如图 2 所示。



(a) Impact of Prior Parameters ϵ on KL Divergence
(a) 先验参数 ϵ 对 KL 散度的影响

(b) Impact of Prior Parameters ϵ on MSE
(b) 先验参数 ϵ 对 MSE 的影响

Figure 2. Impact of different ϵ values on KL divergence and MSE (Grass Wet)
图 2. 不同 ϵ 取值对 KL 散度和 MSE 的影响(Grass Wet)

根据图 2 所示，可以看出，本文方法随着 ϵ 取值的增大，KL 散度和 MSE 都随之增长。在 $\epsilon=0.05$ 之前，增长速度缓慢几乎相等；在 $\epsilon=0.05$ 之后，增长速度随着 ϵ 增大而逐渐增大，KL 散度和 MSE 也显著增大，表明 ϵ 越大，与真实网络的估计偏差越大，因此需要选择合适的 ϵ 取值。

4.3. 实验结果

(1) 零概率值处理效果

根据图 1 所示 Grass Wet 网络结构和表 1 所示网络参数真实值，采用随机抽样算法随机生成样本量为 100、200、500、800、1000、1500、2000 的样本集，以 W 节点 4 个参数 $P(W=0\mid C=1,R=1)$、$P(W=0\mid C=1,R=0)$、$P(W=0\mid C=0,R=1)$、$P(W=0\mid C=1,R=1)$ 为例，计算本文方法得到的网络参数值，并与 MLE、无信息先验 MAP、均匀约束先验方法对比，结果如表 2 所示。

由表 2 可以看出，本文方法与无信息先验 MAP、基于均匀先验方法都克服了小概率事件参数漏算、误算的情况，如样本量为 100、500 的情况下，MLE 将 $P(W=0\mid C=1,R=1)$ 估计为 0.000，即在“下雨 - 洒水”同时发生的情况下“草坪不湿润”的概率为 0.000；而在样本量为 100 的情况下，MLE 将

$P(W=0|C=0,R=1)$ 估计为 0.000，即“不洒水–下雨”的情况下“草坪不湿润”的概率为 0.000，这与我们设置的真实参数估计偏差较大。而本文方法与无信息先验 MAP、基于均匀先验的方法都有效解决了零概率值问题，接下比较本文方法与其他方法的精度。

Table 2. Parameter learning results for node W
表 2. 节点 W 参数学习结果

样本量	本文方法				均匀约束先验				无信息先验 MAP				MLE			
100	0.181	0.157	0.212	0.921	0.381	0.138	0.257	0.873	0.352	0.140	0.244	0.875	0.000	0.000	0.098	0.924
200	0.108	0.125	0.193	0.914	0.329	0.129	0.193	0.872	0.300	0.130	0.192	0.871	0.100	0.014	0.107	0.895
500	0.010	0.114	0.203	0.882	0.189	0.110	0.137	0.898	0.173	0.111	0.138	0.897	0.000	0.104	0.102	0.907
800	0.010	0.106	0.200	0.897	0.122	0.102	0.131	0.893	0.115	0.103	0.132	0.892	0.008	0.100	0.110	0.899
1000	0.010	0.100	0.202	0.900	0.111	0.101	0.111	0.898	0.104	0.101	0.112	0.898	0.009	0.098	0.092	0.903
1500	0.010	0.102	0.199	0.901	0.084	0.105	0.107	0.899	0.079	0.105	0.108	0.898	0.020	0.103	0.094	0.902
2000	0.010	0.100	0.198	0.900	0.064	0.101	0.110	0.893	0.059	0.101	0.111	0.893	0.013	0.100	0.101	0.896

(2) 参数学习精度测试(KL 散度)

为了从总体上更好地说明参数学习的精度，通过计算与真实参数的 KL 散度来分析比较本文的 4 种方法。并且在不同网络复杂度的经典 BNs 上对比，网络信息如表 3 所示。由于文中数据是通过简单随机采样得到，一次实验并不能反映算法的优劣，因此本文每次实验均重复 100 次，结果如表 4、图 3 和图 4 所示。

Table 3. Simulated networks information
表 3. 仿真网络信息

网络名称	节点数	边的数目	参数个数
Grass Wet	4	4	9
Earthquake	5	4	10
Asia	8	8	18

Table 4. Contrast of average KL divergence for the four algorithms
表 4. 4 种算法的平均 KL 散度对比

网络	样本量	本文方法	均匀约束先验	无信息先验 MAP	MLE
Grass Wet	100	0.075	0.127	0.121	0.453
	200	0.066	0.102	0.099	0.233
	500	0.056	0.076	0.074	0.101
	800	0.056	0.068	0.067	0.083
	1000	0.056	0.066	0.065	0.086
	1500	0.056	0.063	0.062	0.079
	2000	0.056	0.060	0.060	0.074

续表

Earthquake	100	0.078	0.121	0.121	1.052
	200	0.067	0.101	0.101	0.799
	500	0.059	0.078	0.077	0.326
	800	0.062	0.078	0.077	0.132
	1000	0.058	0.074	0.073	0.209
	1500	0.055	0.068	0.066	0.083
	2000	0.041	0.064	0.060	0.051
Asia	100	0.101	0.133	0.128	0.378
	200	0.082	0.120	0.114	0.313
	500	0.048	0.091	0.084	0.210
	800	0.032	0.078	0.082	0.172
	1000	0.042	0.077	0.071	0.180
	1500	0.008	0.049	0.043	0.083
	2000	0.013	0.049	0.045	0.070

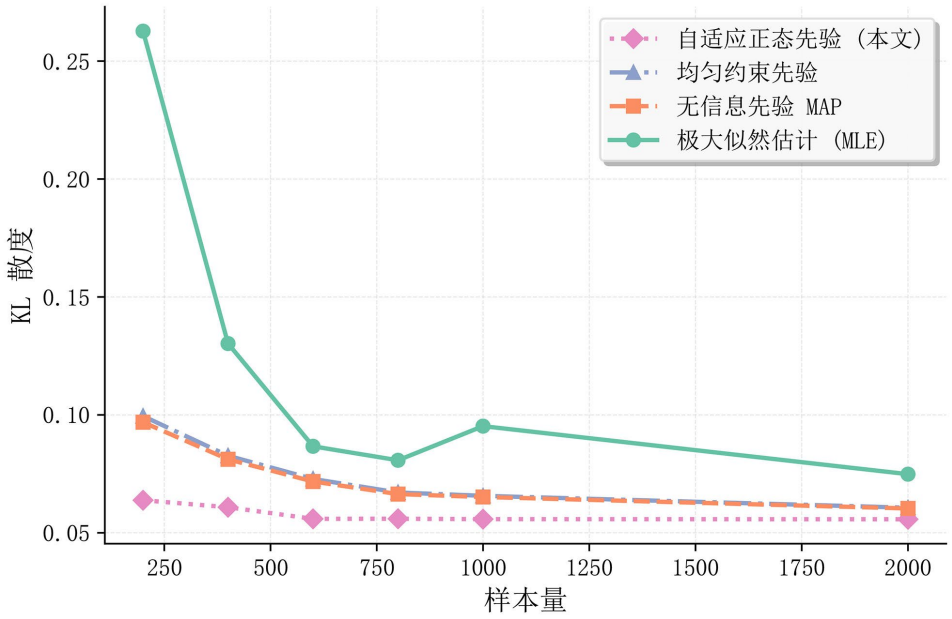


Figure 3. Contrast of average KL divergence for the four algorithms (Grass Wet)
图 3. 四种算法平均 KL 散度对比(Grass Wet)

由表 4 和图 3、图 4 的曲线可以看出，在参数精度(KL 散度)方面，4 种方法随着样本量的增长与真实参数的距离逐渐减小并趋于一致，且本文方法要明显优于均匀约束先验、无信息先验 MAP，MLE 3 种方法。如在不同的网络中，样本量为 100~500 时，本文方法得到的平均 KL 散度均没有超过 0.1。在 Grass Wet 网络中，与 MLE 相比，在样本量从 100 上升到 2000 的过程中，本文方法的平均 KL 散度从 0.075 降低到 0.013，幅度变化较小；而 MLE 的平均 KL 散度从 0.453 降低到 0.074，幅度变化较大。本文的方法相对于 MLE 至少降低了 83%。

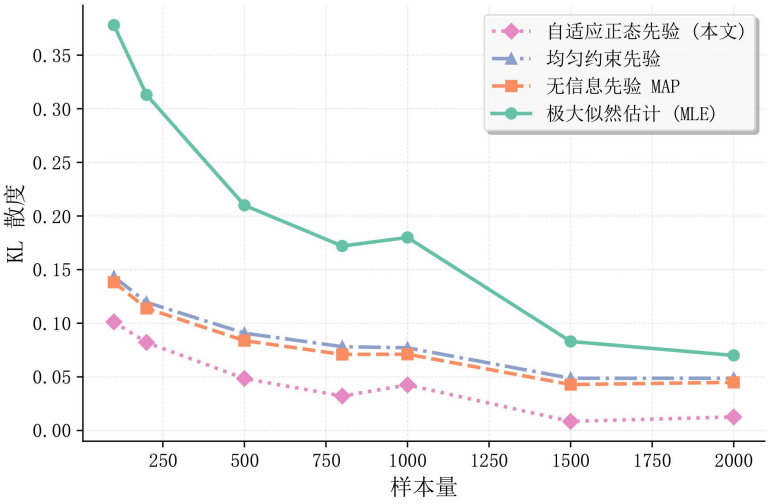


Figure 4. Contrast of average KL divergence for the four algorithms (Asia)
图 4. 四种算法平均 KL 散度对比(Asia)

(3) 参数学习精度测试(MSE)

Table 5. Contrast of MSE for the four algorithms
表 5. 4 种算法的 MSE 对比

网络	样本量	本文方法	均匀约束先验	无信息先验 MAP	MLE
Grass Wet	100	0.032	0.048	0.045	0.038
	200	0.029	0.038	0.037	0.032
	500	0.025	0.029	0.029	0.026
	800	0.025	0.027	0.027	0.026
	1000	0.025	0.026	0.026	0.026
	1500	0.025	0.026	0.026	0.026
	2000	0.025	0.026	0.026	0.025
Earthquake	100	0.029	0.046	0.046	0.059
	200	0.024	0.037	0.037	0.045
	500	0.021	0.028	0.027	0.028
	800	0.022	0.028	0.028	0.027
	1000	0.021	0.026	0.026	0.025
	1500	0.020	0.024	0.023	0.021
	2000	0.015	0.022	0.020	0.016
Asia	100	0.030	0.050	0.048	0.042
	200	0.026	0.041	0.039	0.033
	500	0.020	0.031	0.028	0.022
	800	0.013	0.026	0.023	0.019
	1000	0.017	0.026	0.024	0.019
	1500	0.004	0.015	0.012	0.010
	2000	0.005	0.016	0.014	0.015

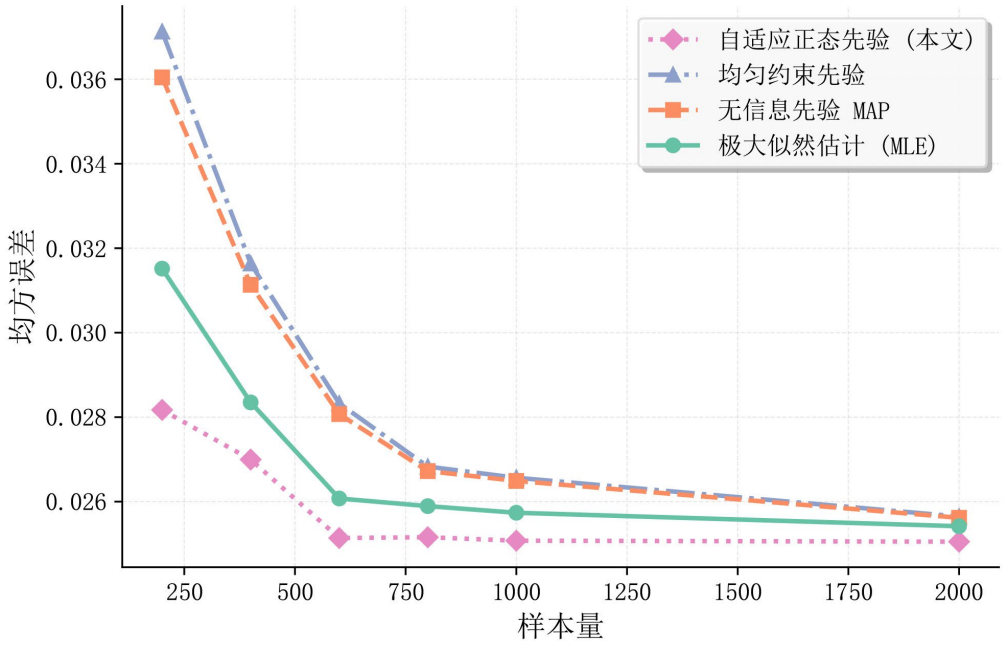


Figure 5. Contrast of MSE for the four algorithms (Grass Wet)
图 5. 四种算法 MSE 对比(Grass Wet)

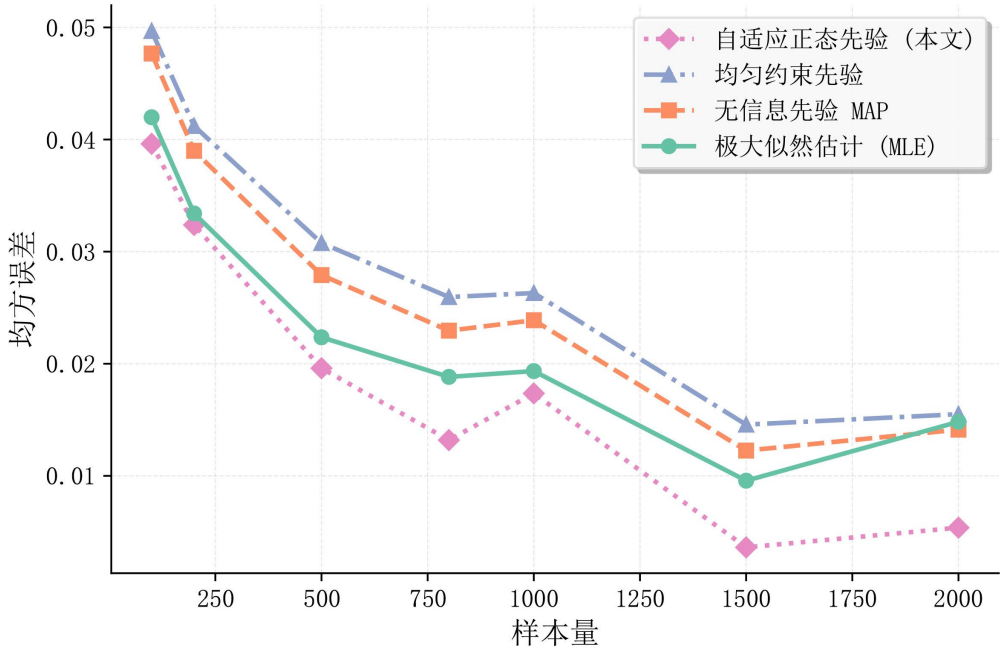


Figure 6. Contrast of MSE for the four algorithms (Asia)
图 6. 四种算法 MSE 对比(Asia)

由表 5 和图 5、图 6 可以看出，4 种方法随着样本量的增长与真实参数的距离逐渐减小，这与平均 KL 散度的趋势一致，且不同复杂度的网络其 MSE 变化也不同，但本文方法均优于其他 3 种方法。例如在 Grass Wet 网络和 Asia 网络中，四种算法在 MSE 上的优劣均表现为本文方法 ＞MLE＞ 无信息先验 MAP＞均匀约束先验方法。

(4) 运行时间测试

以 Grass Wet 网络为例，在不同的样本集下，对 4 种方法的运行时间进行测试，统计每种方法的平均运行时间，测试结果如表 6、图 7 所示。

Table 6. Contrast of running time for the four algorithms
表 6. 4 种算法的计算时间对比

网络	样本量	本文方法	均匀约束先验	无信息先验 MAP	MLE
Grass Wet	100	1.602E−01	4.352E−03	2.007E−04	9.999E−05
	200	1.609E−01	4.755E−03	1.973E−04	1.094E−04
	500	1.529E−01	5.183E−03	2.737E−04	4.762E−04
	800	1.658E−01	5.290E−03	7.021E−04	6.520E−04
	1000	1.586E−01	5.211E−03	7.004E−04	8.257E−04
	1500	1.661E−01	5.737E−03	1.048E−03	1.102E−03
	2000	1.815E−01	7.470E−03	1.712E−03	1.401E−03

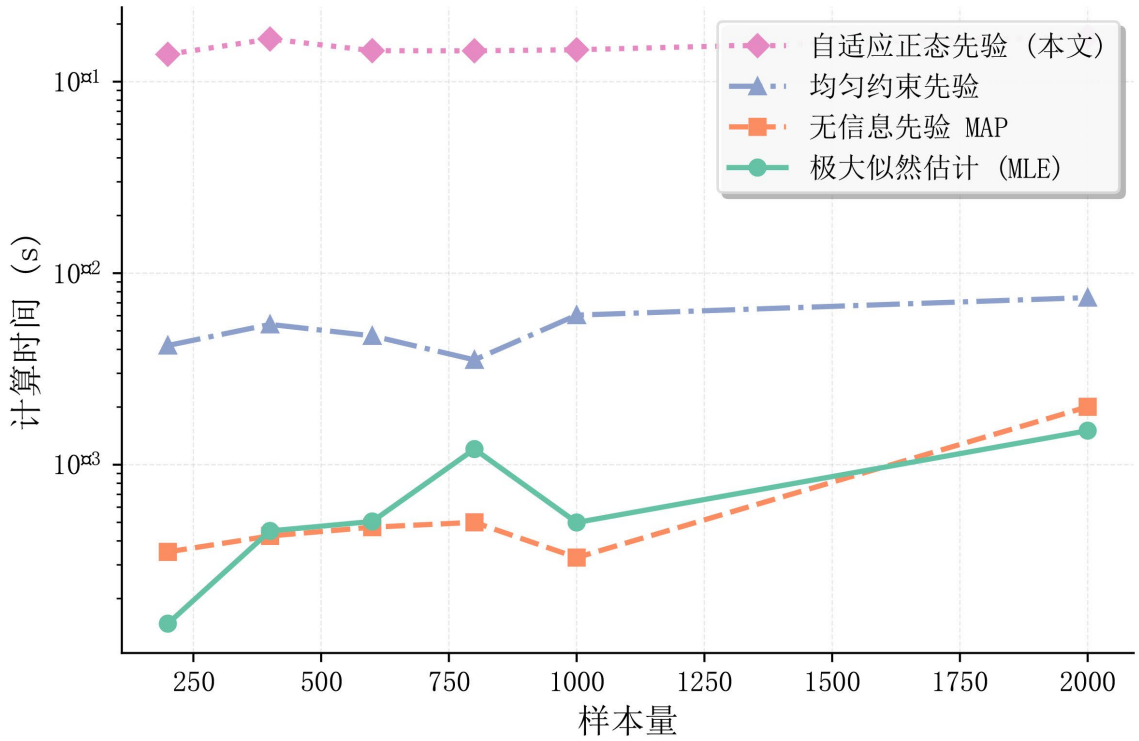


Figure 7. Contrast of running time for the four algorithms (Grass Wet)
图 7. 四种算法运行时间对比(Grass Wet)

由表 6 和图 7 可以看出，在不同的数据集上，本文方法耗时明显都高于其他三种方法，且随着样本量与网络复杂度的增长，运行时间也随之增长，但均不超过 0.3 秒。例如在样本量从 100 升至 2000 的过程中，与 MLE 相比，本文方法的运行时间从 0.1602 秒上升至 0.1815 秒，增幅较大，而 MLE 方法从 0.0009 秒上升至 0.0013 秒，增幅较小。显然，本文方法提高了参数学习精度，但增加了一定的运行时间。

5. 结论

本文针对不均衡数据集条件下建立贝叶斯网络易出现零概率值问题，提出一种基于熵的自适应参数学习方法。实验测试结果表明，该方法的精度大于基于均匀约束先验的方法、无信息先验 MAP 方法、MLE 3 种方法，虽然运行时间要高于这三种方法，但时间消耗总体不超过 0.3 秒。本文也存在不足，本文只是考虑了完整数据样本下的参数学习，后续需要进一步研究缺失数据样本下的学习问题。

参考文献

[1] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann.

[2] Li, J., Du, P., Ye, A.Y., Zhang, Y., Song, C., Zeng, H., *et al*. (2019) GPA: A Microbial Genetic Polymorphisms Assignments Tool in Metagenomic Analysis by Bayesian Estimation. *Genomics*, *Proteomics & Bioinformatics*, **17**, 106-117. https://doi.org/10.1016/j.gpb.2018.12.005

[3] de Campos, C.P. and Ji, Q. (2008) Improving Bayesian Network Parameter Learning Using Constraints. 2008 19*th International Conference on Pattern Recognition*, Tampa, 8-11 December 2008, 1-4. https://doi.org/10.1109/icpr.2008.4761287

[4] Liao, W. and Ji, Q. (2009) Learning Bayesian Network Parameters under Incomplete Data with Domain Knowledge. *Pattern Recognition*, **42**, 3046-3056. https://doi.org/10.1016/j.patcog.2009.04.006

[5] Ru, X., Gao, X., Wang, Z., Wang, Y. and Liu, X. (2023) Bayesian Network Parameter Learning Using Fuzzy Constraints. *Neurocomputing*, **544**, Article ID: 126239. https://doi.org/10.1016/j.neucom.2023.126239

[6] Hou, Y., Zheng, E., Guo, W., Xiao, Q. and Xu, Z. (2020) Learning Bayesian Network Parameters with Small Data Set: A Parameter Extension under Constraints Method. *IEEE Access*, **8**, 24979-24989. https://doi.org/10.1109/access.2020.2971099

[7] Jiang, Y., Liang, Z., Gao, H., Guo, Y., Zhong, Z., Yang, C., *et al*. (2018) An Improved Constraint-Based Bayesian Network Learning Method Using Gaussian Kernel Probability Density Estimator. *Expert Systems with Applications*, **113**, 544-554. https://doi.org/10.1016/j.eswa.2018.06.058

[8] 邸若海, 李叶, 万开方, 等. 基于改进 QMAP 的贝叶斯网络参数学习算法[J]. 西北工业大学学报, 2021, 39(6): 1356-1367.

[9] 邸若海, 高晓光, 郭志高. 基于单调性约束的离散贝叶斯网络参数学习[J]. 系统工程与电子技术, 2014, 36(2): 272-277.

[10] 柴慧敏, 赵昀瑶, 方敏. 利用先验正态分布的贝叶斯网络参数学习[J]. 系统工程与电子技术, 2018, 40(10): 2370-2375.

[11] 曾强, 黄政, 魏曙寰. 融合专家先验知识和单调性约束的贝叶斯网络参数学习方法[J]. 系统工程与电子技术, 2020, 42(3): 646-652.

[12] Zhou, Z., Lam, E.Y. and Lee, C. (2019) Nonlocal Means Filtering Based Speckle Removal Utilizing the Maximum *a Posteriori* Estimation and the Total Variation Image Prior. *IEEE Access*, **7**, 99231-99243. https://doi.org/10.1109/access.2019.2929364

[13] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[14] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006: 35-154.

[15] 梅家斌. 关于 Dirichlet 分布参数估计的渐进分布[J]. 武汉科技学院学报, 2002, 15(1): 8-12.

[16] Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**, 79-86. https://doi.org/10.1214/aoms/1177729694