节点分类中的不确定性传播

Zhao Xu†、Carolin Lawrence†、Ammar Shaker†、Raman Siarheyeu†
†NEC欧洲实验室，德国海德堡{Zhao.Xu, Carolin.Lawrence,
Ammar.Shaker,Raman.Siarheyeu}@neclab.eu

arXiv:2304.00918v1 [cs.LG] 3 Apr 2023

摘要——**神经网络预测不确定性的量化近年来日益受到关注。本文聚焦于衡量图神经网络（图神经网络）在节点分类任务中的不确定性。现有的大多数图神经网络都对节点间的消息传递进行建模，而消息往往是确定性的。由此自然产生一些问题：消息中是否存在不确定性？我们如何将这种不确定性与消息一起在图上进行传播？为解决这些问题，我们提出了一种贝叶斯不确定性传播（BUP）方法，将图神经网络嵌入贝叶斯建模框架，并通过预测概率的贝叶斯置信度以及消息的不确定性，对节点分类的预测不确定性进行建模。我们提出了一种受高斯模型启发的新型不确定性传播机制。此外，我们还提出了一种面向不确定性的节点分类损失函数，使图神经网络能够在学习过程中明确整合预测不确定性。因此，预测不确定性较大的训练样本将受到惩罚。我们从预测可靠性和分布外（OOD）预测两个方面展示了BUP的效果。我们还深入分析了学习到的不确定性。通过大量实验，我们考察了不确定性与图拓扑结构之间的关系，以及分布外情况下的预测不确定性。使用常见基准数据集进行的实证结果表明，所提出的方法性能优越。**

一、引言

现代神经网络凭借出色的性能，已广泛应用于各种学习任务和数据模态。然而，人们近来开始关注神经网络的预测不确定性 [1]–[7],，尤其是在医疗保健和自动驾驶等领域，因为过度自信或信心不足的预测所造成的成本和损害会带来重大影响。对于图数据，预测不确定性的建模同样是一个关键问题。在节点分类中，不确定性通常以预测概率（置信度）表示，并通过图神经网络（GNNs）计算 [8]–[10]。图 1 展示了在引文网络 *CiteSeer*中使用图卷积神经网络（GCNN） [8], 进行节点分类时所学习到的预测概率直方图。图中包含分布内（InDist）节点和分布外（OOD）节点。分布外节点是指来自训练数据中未观测类别的节点。可以发现，不同类型测试节点的推断预测概率统计特征相似。更具体地说，中间图显示，GNN 方法对其错误预测仍然很有信心。右侧图则显示了分布外情形下的类似

本研究得到 H2020 MonB5G项目（资助协议编号：871780）的支持。

DOI：10.1109/ICDM54844.2022.00167 ©2022年 IEEE

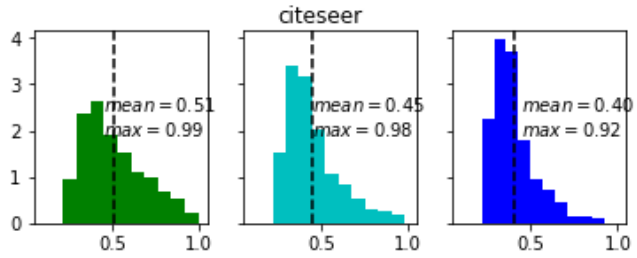


图 1. 使用 GCN [8]学习得到的预测概率直方图：分布内节点（左）、预测错误的分布内节点（中）和分布外节点（右）。横轴为预测概率，纵轴为其密度。

趋势：尽管 GNN 并未学习任何有关分布外节点（属于未观测类别）的知识，它仍会以较高置信度将这些节点分类为已观测类别。这个例子表明，在节点分类问题中，需要采用更复杂的方法来恰当地建模预测不确定性。

为应对这一挑战，我们通过对预测概率的贝叶斯置信度进行建模，研究神经网络用于节点分类时的预测不确定性。具体而言，我们基于消息不确定性学习预测概率的分布。大多数现有图神经网络会在节点之间传递消息。每个节点都关联着一条消息，并以向量表示。消息通过连接在节点之间传递，从而使整个网络能够共享信息。消息通常被假设为确定性的。图神经网络通过最小化某种损失来学习每条消息的点估计，然后根据消息（例如使用 Softmax）计算预测概率。然而，消息是不可观测变量，作为模型的一部分引入。消息的多个取值可能都与观测数据相符。因此，学习消息的分布，而非消息的固定取值，是合理的。具体而言，我们将每条消息建模为一个随机向量，例如服从多元高斯分布。通过整合消息的不确定性，我们便能学习预测概率的分布，从而以贝叶斯方式量化预测不确定性。

当一个节点的消息传递给其邻居时，消息的不确定性也应相应地传递。然而，常用的消息传递机制并不容易应用于不确定性传播。直观而言，如果节点 i 与更多节点相连，那么节点 i 的预测不确定性应低于另一个节点 j 。

因为更多证据会通过连接流向该节点。然而，消息传递机制的表现恰好相反，这会导致连接紧密的节点具有更大的不确定性，或者至少无法保证不确定性会降低。为了解决这一问题，我们探索了一种基于条件高斯模型的新型不确定性传播机制。节点 i 的不确定性取决于其邻居的不确定性。节点连接的邻居越多，其预测概率的不确定性就越低。此外，邻居对所传递消息越不确定，该节点的不确定性也会相应增加。为了更好地将预测不确定性融入模型，我们还提出了一种用于节点分类的不确定性感知损失，显式地将预测不确定性纳入训练过程。除了常用的交叉熵损失外，所提出的损失还会惩罚不确定性较高的预测。该损失与模型无关，也可应用于其他分类问题。我们通过大量实验测试所提方法在不同情况下的性能，并考察该方法的可靠性，深入分析模型学到的不确定性。

II. 相关工作

不确定性量化对于神经网络的实际应用至关重要 [11], ，相关内容将在下文中介绍。

用于模型不确定性的贝叶斯神经网络将贝叶斯建模引入神经网络（NNs）已受到文献的广泛关注。贝叶斯方法并非对神经网络参数进行点估计，而是假设参数服从特定的先验分布，并学习其后验分布以尽可能拟合数据。这种建模策略能够量化模型自身的不确定性（亦称认知不确定性），因此通常具有更好的泛化能力。鉴于贝叶斯神经网络的计算复杂度，人们提出了多种近似方法，例如 [12]–[14]。此外，贝叶斯神经网络也用于建模数据不确定性（亦称偶然不确定性）。例如， [3] 利用贝叶斯框架结合了两种不确定性类型，以处理图像。[15] 扩展了 [3] ，以处理模型不确定性、数据不确定性和分布不确定性。

用于不确定性估计的非贝叶斯方法贝叶斯神经网络推进了深度神经网络不确定性估计的最新进展，但计算复杂度仍令人担忧。另一类研究采用非贝叶斯技术计算参数分布。例如， [16] 提出了自助法， [17]引入了一种集成学习方法来估计模型不确定性。[18]提出了克罗内克因子化拉普拉斯近似，以获得神经网络参数的后验分布。

预测校准 这是频率学派关于不确定性估计的一个概念。与对参数随机性进行建模的贝叶斯框架不同，校准方法（例如 [19]–[21], ）关注推断所得预测与经验长期频率之间的偏差。 [4]通过大量实验指出，现代深度神经网络的校准效果并不理想，并分析了其中的原因。A 一种后处理校准方法——温度缩放——被用于缓解校准失准问题。 [22] 通过加入对抗样本扩展了这项工作，以区分分布内和分布外图像。 [6] 提出了一种缩放分箱校准器，结合了普拉特缩放和直方图分箱的思路，以降低样本复杂度。Chung 等人提出了一种基于分位数的校准方法 [19]。 [20]重新审视了近期图像分类模型的校准问题，并发现其中大多数模型的校准效果都名列前茅。

贝叶斯图神经网络上述方法大多聚焦于非图数据，尤其是图像。一些研究探讨了图中的不确定性 [23]–[28]。具体而言，这些研究考察了图结构的随机性。例如， [29] 结合 GCNN 和混合成员随机块模型，学习随机图（参数）与节点标签的联合后验分布。该方法对噪声连接具有更强的鲁棒性。Ma 等人研究了一种灵活的生成式图神经网络，用于建模节点特征、标签与图结构的联合分布[30]。 [31] 提出了一种图随机神经网络，用于学习分类函数的分布，并针对性地采用摊销变分推断来近似不可解析的联合后验。 [1] 引入了一种贝叶斯图神经网络，将随机正则化推广至训练过程。 [7] 将图神经网络扩展到无法获取输入图拓扑或存在噪声边的场景。 [27]提出了一种用于节点分类的图后验网络。Chen 等人 [28] 引入了多关系图高斯过程网络，以提升确定性方法的灵活性。

III. 图神经网络中的消息传递

已有大量文献研究了图神经网络中节点之间的消息传递。一个典型示例是 GCNN [8],，它定义了一个用于传播消息的核：$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}, \mathbf{K} = \underline{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}$，其中$\mathbf{A}$表示图的邻接矩阵，$\mathbf{I}$ 是对应于节点额外自环的单位矩阵。$\mathbf{D}$ 是对角矩阵，其元素为$d_{i,\ i}$，即节点 i的度，也就是说，$d_{i,\ i} = \sum_j \hat{a}_{i,\ j}$。该核定义了消息传递过程：节点 i 在第 $\ell + 1$ 层的消息，是所有与其直接相连的节点（即 i的邻居，记作 $N(i)$）在第 ℓ层的消息的加权和。消息 $\theta_i^{\ell+1}$ 更新如下：

$$\theta_i^{(\ell+1)} = a(\frac{1}{\sqrt{d_{i,i}}} \sum_{j' \in N(i)} \frac{1}{\sqrt{d_{j',j'}}} \theta_{j'}^{(\ell)} \mathbf{W}^{(\ell+1)}). \quad (1)$$

这种消息传递机制是一个确定性过程：未考虑消息的不确定性/随机性。

IV. 不确定性建模

量化深度神经网络（DNNs）的预测不确定性是一项重要但尚未解决的问题。已有文献表明，深度神经网络的预测往往过度自信 [32], [33]。在许多应用中，深度神经网络可能会以较高的预测概率预测出错误的标签。图 1 展示了图数据中的这一问题。理想情况下，分布外节点的预测概率较低。事实上，实验

表明，预测概率并不一定与之相符。然而，常用工具（例如置信区间）并不适用，因为标签是离散变量。我们如何量化节点分类的预测不确定性？

现在，我们提出一种基于贝叶斯置信度的新方法，通过对预测概率的分布进行建模来量化预测的不确定性。通常，图神经网络分类器使用以消息 θ_i 为输入的离散概率函数来学习节点 i 的类别标签 $\mathbf{y}_i$ ： $p(\mathbf{y}_i|\mathbf{X}, \mathbf{A}) = softmax(\theta_i)$ $\theta_i = f(\mathbf{X}, \mathbf{A}; \phi)$，其中 $\mathbf{y}_i \in \{0,1\}^C$ 是一个独热向量。X 是节点特征。函数 $f(\mathbf{X}, \mathbf{A}; \phi)$ 使用以 X 和 A 为输入、以 ϕ 为参数的图神经网络（例如 GCNN）定义。给定函数 f 后，预测概率 $p(\mathbf{y}_i|\mathbf{X}, \mathbf{A})$ 和消息 θ_i 都是确定性的。这里，我们引入贝叶斯置信度，以获得更大的灵活性。具体而言，我们将分类器嵌入贝叶斯框架中：对预测概率 $p(\mathbf{y}_i|\mathbf{X},\mathbf{A})$ 的概率分布进行建模，并利用该概率的置信区间推导预测不确定性：

$$p(\mathbf{y}_i|\mathbf{X}, \mathbf{A}) = \int p(\mathbf{y}_i|\theta_i)p(\theta_i|\mathbf{X}, \mathbf{A})d\theta_i \tag{2}$$

在这里，我们将消息 $\theta_i \in \mathcal{R}^C$ 建模为服从分布 $p(\theta_i|\mathbf{X}, \mathbf{A})$ 的随机变量。在贝叶斯框架中， θ_i被称为不可观测变量，每个数据样本对应一个。这些变量作为模型的一部分引入，并将从数据中学习得到。消息的多个取值都可能与观测数据相符。因此，对其随机性进行建模，即学习消息的分布，而非固定值，是合理的。由于消息 θ_i 是连续变量，我们假设其服从 C维高斯分布：

$$\begin{aligned} p(\theta_i|\mathbf{X}, \mathbf{A}) &= \mathcal{N}(\mathbf{m}_i, \Sigma_i) \\ \mathbf{m}_i = g(\mathbf{X}, \mathbf{A}; \gamma),\ \Sigma_i &= h(\mathbf{X}, \mathbf{A}; \psi) \end{aligned} \tag{3}$$

协方差矩阵 Σ_i 通常假设为对角矩阵。这主要有两方面的好处。首先，对角协方差简化了学习和推断中的计算。更重要的是，消息 θ_i 是由模型引入的潜变量。预期潜变量的各个维度彼此独立，并分别表示节点不同的隐藏属性。均值函数 $\mathbf{m}_i$ 和协方差函数 Σ_i 由图神经网络定义，因为神经网络擅长函数逼近，尤其适用于复杂数据。通过对消息不确定性进行建模，所提出的框架能够更好地捕捉预测不确定性。方差Σ_i越大，预测概率的贝叶斯置信区间就越大，因此预测的不确定性也越高。由于消息是多维的，我们可以使用平均方差或其高斯熵来量化预测不确定性。请注意，高斯熵可以视为衡量高斯椭圆体积的标准。我们在实验中展示了不确定性分数（见图3）。

在相关文献中， [34]和 [15] 针对图像数据研究了类似的不确定性，并将其命名为分布不确定性。他们使用狄利克雷分布而非高斯分布来对类别概率的分布建模。此处引入的高斯分布先验会更适用于图数据，因为它明确地定义了消息的不确定性（Σ_i），并且可以通过图神经网络定义的传播机制传递给邻居节点。

V. 不确定性传播

本节将介绍在节点之间传递消息时如何传播不确定性。一种直接的方法是：分别学习两个架构相同但参数不同的图卷积神经网络，用于估计高斯分布（3）的均值和方差，然后使用重参数化技巧为每条消息抽取多个样本，并采用蒙特卡洛梯度估计器来近似计算参数的梯度，以进行训练。然而，常用的消息传递机制在传播不确定性时效果不佳。直观而言，如果节点 i 与多个节点相连，我们可以合理地推断，该节点的预测不确定性应小于另一个连接较少的节点 j ，因为节点 i 的预测将以更多证据为条件。这一直觉与常用的消息传递机制（例如（1））并不相符，因为这些机制无法保证连接更多邻居会降低不确定性。因此，我们受高斯过程（GP） [35], 的启发，引入了一种新的传播机制，以在节点之间传递不确定性。高斯过程定义了一组随机变量，其中任意有限个随机变量都服从联合高斯分布。属性为$\mathbf{x}_*$ 的样本的条件方差计算如下： $var（y_*）= k_{*,\ *} - \mathbf{k}_*^T\mathbf{K}^{-1}\mathbf{k}_*$. $k_{i,\ j}$ 是样本 i 与 j之间的协方差。预测方差的第一项可以大致理解为该样本的先验方差（不确定性），而第二项则表示由于依赖其他样本而带来的不确定性降低。受预测方差的启发，我们引入了以下不确定性传播机制。

在每一层，我们假设所有节点的消息均服从高斯分布。对于一个节点子集，即节点 i 及其邻居$N（i）$，它们消息的每个维度均服从一个（$N_i + 1$）维高斯分布，其中 N_i 是 $N（i）$ 的基数。这里的高斯分布定义了节点之间的概率依赖关系，而高斯分布（3）则指定单个节点的消息分布。现在我们来定义协方差。

定义1。 节点 i 与其邻居 $N（i）$ 之间的协方差矩阵是一个分块矩阵

$$\begin{bmatrix} var(i) & cov(i,j_1) & cov(i,j_2) & \dots & cov(i,j_{N_i}) \\ cov(i,j_1) & var(j_1) & & & \\ cov(i,j_2) & & var(j_2) & & 0 \\ \vdots & 0 & & \ddots & \\ cov(i,j_{N_i}) & & & & var(j_{N_i}) \end{bmatrix} \tag{4}$$

$cov(i,j) = cor(i,j)\sqrt{var(i)var(j)}, cor(i,j) = 1/\sqrt{\lambda d_{i,i}d_{j,j}}$

协方差的定义忽略了邻居节点之间的连接，因为在实际应用中，这些连接通常较为稀疏。对于稠密图，可以直接将协方差$cov(j,\ j')$加入矩阵中，而不会对后续步骤的计算造成显著影响。在定义1中，相关性

函数 $cor(i,\ j)$由子图指定，因此，协方差矩阵可由节点 i 及其邻居的消息方差$var(\cdot)$完全确定。方差是一个将在训练过程中学习的潜变量。这里我们利用相关函数和学习得到的方差来计算节点之间的协方差。这将大幅减少需要学习的潜变量数量。相关函数可以是任意形式，但必须对称。当前采用的相关函数受 GCNN 的核启发，并符合这样的直觉：如果两个节点的邻居较少，它们的联合变异性可能更强，即相关性与节点度成反比$d_{i,\ i}$。 λ 是用于调整相关性的超参数。基于分块矩阵的舒尔补 引理[36]，我们可以轻松证明：定义1中的协方差矩阵是有效的（其行列式不为零）。综合以上结果，节点 i 的条件方差为：

$$\begin{aligned} &\hat{var}(i|N(i)) = var(i) - CB^{-1}C^T \\ &B = diag([var(j_1), var(j_2), \ldots, var(j_{N_i})]) \\ &C = [cov(i,j_1), cov(i,j_2), \ldots, cov(i,j_{N_i})], \end{aligned} \tag{5}$$

其中， B 和 C 是定义1中的块。根据（5），我们可以在每一层 ℓ将不确定性从邻居 N（i）传播到节点i。这种变化会在各层之间传递：

$$var^{(\ell+1)}(i) = a\left(\hat{var}^{(\ell)}(i|N(i))W_{var}^{(\ell+1)}\right) \tag{6}$$

通过不确定性传播，节点连接的邻居越多，可用证据降低的不确定性就越大，因此该节点的方差越小。

VI. 不确定性惩罚损失

理想情况下，应将学得的预测不确定性纳入损失，以指导训练过程。如果某个训练样本的预测不确定性很高，就应对该样本的损失施加惩罚。然而，常用的分类损失函数（例如期望交叉熵）并未恰当地纳入不确定性。受回归问题中基于高斯似然的损失启发，我们提出一种新损失，可直接利用消息分布，将预测不确定性明确纳入训练。该新损失基于高斯累积分布函数。具体而言，训练节点 i 与标签 $y = c_*$ 及消息变量 $\theta = [\theta_1, ..., \theta_C]$相关联。消息的每个维度 θ_c彼此独立，并服从均值为m_c 、方差为 σ_c^2的高斯分布。由于 i 的标签为 c_*，因此有： $\tau_c = \theta_c - \theta_{c_*} < 0,\ c \neq c_*.\ \tau = [\tau_1, ..., \tau_C]_{c \neq c_*}^{C}$服从均值为 $\mu = [m_1 - m_{c_*}, \ldots, m_C - m_{c_*}]$、协方差为：

$$\Lambda = \begin{bmatrix} \sigma_1^2 + \sigma_{c_*}^2 & \sigma_{c_*}^2 & \ldots & \sigma_{c_*}^2 \\ \sigma_{c_*}^2 & \sigma_2^2 + \sigma_{c_*}^2 & \ldots & \sigma_{c_*}^2 \\ & & \ldots & \\ \sigma_{c_*}^2 & \sigma_{c_*}^2 & \ldots & \sigma_C^2 + \sigma_{c_*}^2 \end{bmatrix} \tag{7}$$

给定高斯分布后，我们现在可以将示例 i 的损失 $\mathcal{NLL}$ 计算为其似然概率的负对数：

$$\mathcal{L} = p(y = c_*|\mu, \Lambda) = \int_{-\infty}^{0} \mathcal{N}_{C-1}(\tau; \mu, \Lambda) d\tau \tag{8}$$

这样，我们就可以通过对高斯分布进行反常积分来计算离散标签的似然，从而避免对非共轭的 softmax函数进行代价高昂的边缘化。为了计算反常积分，我们首先通过乔列斯基分解 $\Lambda = LL^T$ 将变量τ 变换为 ξ，使得 ξ 服从标准高斯分布： $\xi = L^{-1}(\tau - \mu)\ \sim \mathcal{N}_{C-1}(\mathbf{0},\ \mathbf{I})$。 尽管 ξ_c 彼此独立，但积分 (8) 无法以与维度无关的方式计算，因为某一维度 c 的积分上限取决于它之前的维度 $c' < c$ 。为了解决这个问题，我们可以使用基于 [37], 的准蒙特卡洛近似，大幅降低蒙特卡洛估计量带来的方差。考虑到实际计算效率，我们还可以通过忽略 Λ的非对角元素来近似似然，从而简化似然：

$$\mathcal{L} \approx \prod_{c=1, c \neq c_*}^{C} \frac{1}{2}\left[1 + erf\left(\frac{m_{c_*} - m_c}{\sqrt{2(\sigma_c^2 + \sigma_{c_*}^2)}}\right)\right] \tag{9}$$

其中， erf 表示误差函数。(9)明确纳入了消息的方差以及消息期望之间的差异。给定标签 c_* 对应的学习到的均值 m_{c_*} 比其他维度 m_c越大，模型做出正确预测的可能性就越高。然而，如果模型对预测不确定，即 σ_c 或 σ_{c_*}较大，那么差异 $m_{c_*} - m_c$ 就会受到惩罚，样本的似然也会相应变化。

VII. 实证分析

由于真实不确定性的真实值不可得，评估不确定性估计的质量是一项具有挑战性的任务。为了衡量所提出方法 BUP 的性能，我们考虑两种设置：常规设置 ，其中所有测试节点都属于训练数据中已观测到的类别；以及 分布外设置，其中测试节点可能属于某个未观测类别。

数据集 实证分析使用了常用的基准数据集 *Cora* 和 *CiteSeer*[1], [7], [8] ，其中节点代表文档，特征是文档的词袋表示。根据引文链接构建无向图。每个节点的标签表示对应文档的类别。学习任务是预测未知标签。

基线方法 我们将所提出的方法与近期工作进行比较，包括：变分图卷积网络 (VGCN) [7], 贝塔-伯努利图 DropConnect (BBGDC 和 BBDE) [1], 基于图高斯过程 (GGP) 的贝叶斯半监督学习 [38]以及图卷积神经网络 (GCNN) [8]。我们采用 VGCN (https://github.com/ebonilla/VGCN)、GGP (https://github.com/ yincheng/GGP)、BBGDC/BBDE (https://github.com/armanihm/ GDC) 和 GCNN (https://github.com/tkipf/gcn) 作者提供的最优参数设置和网络架构。

指标 我们采用三种度量来量化模型性能。除了常用的预测准确率（ACC）外，我们还使用平均校准误差（ ACE）和期望校准误差（ECE） [4], [39]来衡量方法的可靠性。在分类任务中，如果预测概率始终是

表 I 常规设置下各方法的预测性能。指标 ACC 衡量预测准确率（越高越好），而 ACE 和 ECE 量化预测的可靠性（越低越好）。与基线方法相比，所提出的方法 BUP 可靠性高得多，同时准确率更高或相当。

训练	GCNN			VGCN			BBGDC			BBDE			GGP			BUP		
	ACC	ACE	ECE	ACC	ACE	ECE	ACC	ACE	ECE	ACC	ACE	ECE	ACC	ACE	ECE	ACC	ACE	ECE
Cora 5	**69.93**	19.43	23.14	59.00	10.65	10.63	68.36	31.66	33.77	69.40	34.14	36.31	59.69	33.92	33.13	69.41	**9.78**	**7.82**
10	**75.73**	19.22	21.68	64.94	12.03	11.29	74.52	32.85	37.02	65.75	32.66	34.82	70.47	28.66	35.32	75.02	**9.45**	**6.93**
15	**77.99**	18.93	20.75	68.86	11.74	10.28	74.55	32.87	37.36	76.25	35.5	40.35	68.72	23.81	29.69	77.40	**9.90**	**6.90**
20	**79.10**	18.80	20.06	71.34	10.67	9.58	77.07	38.33	43.02	76.16	37.49	42.03	76.83	24.41	29.31	78.64	**10.03**	**6.80**
CiteSeer 5	40.14	34.21	18.81	48.21	15.12	14.84	50.76	14.97	13.45	44.59	19.50	19.65	41.56	29.01	20.77	**53.69**	**10.69**	**9.93**
10	53.75	25.63	20.58	55.62	13.22	14.20	59.38	19.12	19.55	57.14	18.97	20.28	53.94	26.56	28.64	**60.38**	**10.29**	**8.63**
15	63.66	18.78	20.84	58.91	13.66	14.60	**64.49**	20.51	23.20	58.60	21.67	22.41	59.81	24.95	31.64	63.37	**10.80**	**8.36**
20	**67.42**	17.00	20.23	61.03	13.49	14.94	66.80	20.62	21.09	60.12	21.17	23.48	65.27	23.09	33.07	64.77	**10.58**	**8.01**

真实概率，那么模型就是可靠的（也称校准良好）。ACE 和 ECE 的分数越低，这两个概率就越接近。我们期望分类器不仅准确率高，还能具备良好的可靠性。

A. Q1：所提出的方法在常规设置下能否提供更可靠的预测？

这里我们假设处于常规设置，即训练数据中包含所有类别 $\{1,\dots,C\}$。为进行全面评估，我们设置了不同规模的训练数据。具体而言，对于每个数据集，我们随机选取每类 5、10、15、20 个节点用于训练，选取 200 个节点用于验证集，另选取 2000 个节点用于测试集。我们重复实验 10 次，并报告所有重复运行结果的平均值。表 I 总结了实验结果。可以发现，所提出的方法在可靠性（ACE 和 ECE）方面表现更优，同时准确率（ACC）更好或相当。作为一种贝叶斯生成式神经网络，VGCN 的可靠性也不错，但仍不及我们的方法 BUP。VGCN 的预测准确率并不理想。其他贝叶斯方法，如 BBGDC、BBDE 和 GGP，虽然将一定的不确定性纳入模型，但可靠性似乎较低。GCNN 只在节点之间传递消息，不考虑不确定性。尽管 GCNN 在某些情况下报告的准确率略高，但其可靠性远不如我们的方法。实验结果表明，在常规设置下，我们的方法能够提供最可靠的预测，同时准确率更好或相当。学习消息不确定性是捕捉预测不确定性的有效方式。

B. Q2：所提出方法的预测结果是否能在分布外设置中恰当地报告其不确定性？

在实际应用中，分布外设置时常会出现。对于图数据，我们按以下方式创建分布外样本：从训练集和验证集中移除类别 C 的节点，并使用其余类别 $\{1,\dots,C-1\}$ 的样本训练模型。我们将标签为 $y_i \in \{1,\dots,C-1\}$ 的节点视为分布内样本（InDist），将标签为 $y_i = C$ 的节点视为分布外样本。对于分布内样本，其预测性能呈现出与表 I 类似的趋势。由于篇幅限制，详细信息见代码仓库。

此外，我们对分布外预测进行了深入分析。实验设计如下。给定一个测试节点 j，概率 $\mathbf{p}^j$ 是 $C-1$ 维的，其中$p_{j,\,c}$ 表示 j 属于 c的概率。节点 j 是

表 II

对 Cora 数据集中分布外设置下推断得到的预测概率进行分析。与分布内节点相比，分布外节点较小的$p_{j,\,c*}$ 和标准差表明，学习得到的概率质量并未集中在单一类别上，而是分布在所有可能的类别上，这意味着我们的方法能够正确捕捉预测不确定性。

训练	$p_{j,c*}$		标准差 $\mathbf{p}^j$	
	分布内	分布外	分布内	分布外
5	0.70	0.56	0.16	0.10
10	0.70	0.59	0.15	0.10
15	0.75	0.62	0.19	0.13
20	0.74	0.60	0.19	0.11

标记为 $c*$，如果 $p_{j,c*} > p_{j,c}$，对于任意 $c \in \{1,\dots,C-1\}$和$c \neq c*$。$\mathbf{p}^j$ 的标准差（std dev）计算为$\sqrt{\frac{1}{C-1}\sum_{c=1}^{C-1}\left(p_{j,c} - \frac{1}{C-1}\sum_{c=1}^{C-1} p_{j,c}\right)^2}$。。标准差越小，表示预测概率在所有维度（即类别）上的分布越均匀，这意味着模型实际上无法确定测试节点应属于哪个类别。就最大概率 $p_{j,c*}$而言，概率越小，概率质量就越分散。表 II 报告了分布内节点和分布外节点的 $p_{j,c*}$ 及 $\mathbf{p}^j$ 的均值。可以发现，与分布外节点相比，我们的模型为分布内测试节点给出的 $p_{j,c*}$ 显著更大。分布内节点的 $\mathbf{p}^j$ 标准差远大于分布外节点。这些结果表明，所提出的方法对分布内预测的置信度高于对分布外预测的置信度，并且确实捕捉到了分布外节点的不确定性。图 2 展示了分布内节点和分布外节点的 $p_{j,c*}$ 分布。两者差异显著，进一步直观地展示了我们的方法能够很好地对预测的不确定性进行建模。

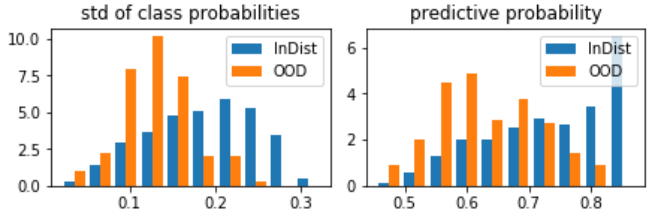


图 2. 对 Cora 数据在分布外设置下推断得到的预测概率进行分析。$p_{j,\,c*}$ 分布内节点（分布内，以蓝色标记）的分布与分布外节点（分布外，以橙色标记）的分布显著不同。

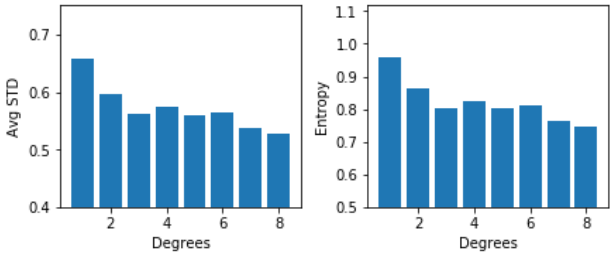


图3. Cora 数据上学习到的消息不确定性分析：节点度与不确定性。我们的方法捕捉到了这样一种趋势：随着度的增加，不确定性（通过消息分布的平均标准差和熵来衡量）会降低，因为可以通过邻居获取更多证据。

C. Q3：所提出的方法能否捕捉图结构对预测不确定性的影响？

由于不确定性没有真实值，我们设计了这项实验，以进一步验证所提出的方法能否恰当地学习预测不确定性，从而捕捉图结构与预测不确定性之间的关系。例如，如果一个节点连接了多个邻居，即节点度较高，那么它的预测不确定性更有可能较低，因为更多证据可以通过其邻居传递到该节点，用于类别预测。此外，如果一个节点距离训练节点较远，那么证据可能在传播过程中消失，因此预测不确定性可能更高。我们开展实验，检验学习到的不确定性是否符合这些直觉。为了量化不确定性，我们使用学习到的消息分布。具体而言，我们选取两个指标：平均标准差和消息分布的熵。图3显示，随着节点度增加，学习到的不确定性会降低。在代码仓库中，我们报告了更多结果，例如不确定性与测试节点到训练节点的平均最短路径长度之间的关系。结果表明，我们的 BUP 方法确实捕捉到了预测不确定性与图结构之间的关系，因此能够恰当地学习预测不确定性。

VIII. 结论

本文提出了一种基于预测概率贝叶斯置信度的新型图神经网络方法，用于显式量化节点分类中的不确定性。与大多数图神经网络的确定性过程不同，该方法对消息的不确定性进行建模，并在整个图中连同消息一起传播不确定性，从而学习预测概率的分布。大量实证分析表明，与当前最先进的方法相比，我们的方法能够提供可靠得多的预测。

参考文献

[1] A. Hasanzadeh 等人，《采用自适应连接采样的贝叶斯图神经网络》，发表于 *ICML*，2020年。[2] H. Jiang、B. Kim、M. Guan 和 M. Gupta，《该不该信任分类器》，发表于 *NeurIPS*，2018年。[3] A. Kendall 和 Y. Gal，《计算机视觉中的贝叶斯深度学习需要哪些不确定性？》，发表于 *NIPS*，2017年。

[4] C. Guo、G. Pleiss、Y. Sun 和 K. Weinberger，《现代神经网络的校准》，发表于 ICML，2017年。[5] S. Depeweg、J.-M. Hernandez-Lobato、F. Doshi-Velez 和 S. Udluft，《贝叶斯深度学习中的不确定性分解：面向高效且风险敏感的学习》，发表于 ICML，2018年。[6] A. Kumar 等人，《经过验证的不确定性校准》，发表于 NeurIPS，2019年。[7] P. Elinas 等人，《在缺少图数据及对抗性设置下对图卷积网络进行变分推断》，发表于 NeurIPS，2020年。[8] T. Kipf 和 M. Welling，《采用图卷积网络的半监督分类》，发表于 ICLR，2016年。[9] P. Velickovic 等人，《图注意力网络》，发表于 ICLR，2018年。[10] W. Hamilton、R. Ying 和 J. Leskovec，《大规模图上的归纳式表示学习》，发表于 NIPS，2017年。[11] M. Abdar 等人，《深度学习中的不确定性量化综述：技术、应用与挑战》，《Information Fusion》，2021年。[12] D. Kingma、T. Salimans 和 M. Welling，《变分 dropout 与局部重参数化技巧》，发表于 NIPS，2015年。[13] Y. Gal 和 Z. Ghahramani，《将 dropout 作为贝叶斯近似：表示深度学习中的模型不确定性》，发表于 ICML，2016年。[14] C. Louizos 和 M. Welling，《采用矩阵高斯后验的结构化高效变分深度学习》，发表于 ICML，2016年。[15] A. Malinin 和 M. Gales，《通过先验网络估计预测不确定性》，发表于 NeurIPS，2018年。[16] I. Osband、C. Blundell、A. Pritzel 和 B. V. Roy，《通过自举 DQN 进行深度探索》，发表于 NIPS，2016年。[17] B. Lakshminarayanan、A. Pritzel 和 C. Blundell，《使用深度集成进行简单且可扩展的预测不确定性估计》，发表于 NIPS，2017年。[18] H. Ritter、A. Botev 和 D. Barber，《适用于神经网络的可扩展拉普拉斯近似》，发表于 ICLR，2018年。[19] Y. Chung 等人，《超越弹珠损失：用于校准不确定性量化的分位数方法》，发表于 NeurIPS，2021年。[20] M. Minderer 等人，《重新审视现代神经网络的校准》，发表于 NeurIPS，2021年。[21] V. Kuleshov、N. Fenner 和 S. Ermon，《利用校准回归实现深度学习中的准确不确定性估计》，发表于 ICML，2018年。[22] S. Liang、Y. Li 和 R. Srikant，《增强神经网络对分布外图像检测的可靠性》，发表于 ICLR，2018 年。[23] D. Eswaran 、 S. G¨unnemann 和 C. Faloutsos，《确定性的力量：用于置信传播的狄利克雷-多项式模型》，发表于 SDM，2017年。[24] Z.-Y. Liu 等人，《用于半监督学习的感知不确定性的图高斯过程》，发表于 AAAI，2020年。[25] X. Zhao、F. Chen、S. Hu 和 J.-H. Cho，《图数据上的感知不确定性的半监督学习》，发表于 NeurIPS，2020年。[26] M. Monteiro 等人，《随机分割网络：对空间相关的偶然不确定性进行建模》，发表于 NeurIPS，2020年。[27] M. Stadler 等人，《图后验网络：节点分类中的贝叶斯预测不确定性》，发表于 NeurIPS，2021年。[28] G. Chen 等人，《采用贝叶斯高斯过程网络的多关系图表示学习》，发表于 AAAI，2022年。[29] Y. Zhang 等人，《用于半监督分类的贝叶斯图卷积神经网络》，发表于 AAAI，2019年。[30] J. Ma、W. Tang、J. Zhu 和 Q. Mei，《面向基于图的半监督学习的灵活生成式框架》，发表于 NeurIPS，2019年。[31] H. Wang 等人，《用于半监督学习的图随机神经网络》，发表于 NeurIPS，2020年。[32] A. Nguyen 等人，《深度神经网络很容易被欺骗：对无法识别图像的高置信度预测》，发表于 CVPR，2015年。[33] Y. Ovadia 等人，《你能信任模型的不确定性吗？数据集偏移下预测不确定性的评估》，发表于 NeurIPS，2019年。[34] M. Sensoy、L. Kaplan 和 M. Kandemir，《用于量化分类不确定性的证据深度学习》，发表于 NeurIPS，2018年。[35] C. Rasmussen 和 C. Williams，《机器学习中的高斯过程》。MIT Press，2006年。[36] F. Zhang（编），《舒尔补及其应用》。Springer，2005年。[37] A. Genz，《多元正态概率的数值计算》，《Journal of Computational and Graphical Statistics》，1992年。[38] Y. C. Ng、N. Colombo 和 R. Silva，《采用图高斯过程的贝叶斯半监督学习》，发表于 NeurIPS，2018年。[39] L. Neumann 等人，《松弛 Softmax：用于安全行人检测的高效置信度自动校准》，发表于 NIPS Workshop，2018年。