一石二鸟:图数据无监督学习中的检测与分类协同

李思聪,王 飞*,魏子令,陈曙晖
(国防科技大学计算机学院,湖南长沙 410073)

摘 要: 现实世界中的图机器学习系统通常运行于开放环境,测试阶段不可避免地接触到与训练分布不一致的样本,这违背了传统监督学习中训练与测试同分布的假设。模型不仅需要在分布内(In-Distribution,ID)样本上保持稳定的分类性能,还需具备识别并拒绝分布外(Out-Of-Distribution,OOD)数据的能力,以避免过度自信的错误预测。由于图数据中节点属性与拓扑结构高度耦合,分布偏移往往以隐式形式发生,使得图OOD检测较欧氏数据更加复杂。现有图OOD检测方法通常依赖强监督假设,如引入预标注的异常样本,或假设辅助OOD数据与ID数据在特征空间中显著可分。然而在实际应用中,OOD数据多以无标注、与ID数据天然混杂的形式出现,例如社交网络中的跨平台用户或推荐系统中的冷启动节点。这类野生数据难以通过先验规则进行显式区分,限制了现有方法在开放环境下的适用性。针对这一问题,本文提出一种全开放训练范式,在无需任何OOD标注或分布先验的条件下,利用无标注ID/OOD混合数据联合优化图节点分类与OOD检测任务。该方法通过构建带约束的优化目标,在严格约束ID分类误差与误检率的同时,引导模型提升对潜在OOD样本的识别能力,从而刻画真实开放环境中ID与OOD分布的隐式耦合关系。在方法层面,引入基于能量函数的检测机制,将图神经网络输出映射为能量值以度量样本与训练分布的一致性。能量约束引导模型在表示空间中形成可分离的分布结构,使ID样本集中于低能量区域,而潜在OOD样本对应较高能量,从而实现有效区分。该机制避免了基于Softmax置信度方法在分布外场景下的过度自信问题,并使检测目标能够直接作用于图表示学习过程。为求解上述带约束优化问题,本文采用增广拉格朗日方法,在训练过程中动态平衡约束满足与目标优化,增强模型在混合分布下的稳定性。实验结果显示,该方法在多个真实世界图数据集上均取得显著性能提升。在Twitch数据集上,AUROC和AUPR分别达到95.97%和92.84%,较当前最优基线GNNsafe++提升超过21个百分点,同时将误报率控制在12.30%,验证了其在无强监督条件下的有效性与鲁棒性。

关键词: 分布外检测;图神经网络;节点分类;机器学习;野生数据;能量函数
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0153-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250761

Joint Detection and Classification in Unsupervised Graph Learning

LI Sicong, WANG Fei*, WEI Ziling, CHEN Shuhui
(*College of Computer Science*, *National University of Defense Technology*, *Changsha*, *Hunan* 410073, *China*)

Abstract: Real-world graph machine learning systems typically operate in open environments, where test-time data inevitably deviate from the training distribution, violating the common assumption of identical training and testing distributions in supervised learning. In this setting, models are required not only to maintain stable classification performance on in-distribution (ID) samples, but also to accurately identify and reject out-of-distribution (OOD) data to avoid overconfident erroneous predictions. Due to the strong coupling between node attributes and graph topology, distribution shifts in graph data often occur implicitly, making graph OOD detection more challenging than its Euclidean counterpart. Existing graph OOD detection methods commonly rely on strong supervision assumptions, such as the availability of pre-labeled anomalous samples or the assumption that auxiliary OOD data are clearly separable from ID data in the feature space. However, in practical applications, OOD samples typically appear in an unlabeled and naturally mixed manner with ID data, as observed in cross-platform users in social networks or cold-start nodes in recommendation systems. Such wild data are difficult to distinguish using prior rules, which limits the applicability of existing approaches in open environments. To address this issue, we propose a fully open training paradigm that jointly optimizes graph node classification and OOD detection using unlabeled ID/OOD mixed data, without requiring any OOD annotations or distributional priors. The proposed method formulates a constrained optimization objective that strictly controls ID classification error and false positive rates, while encouraging the model to improve its capability to identify potential OOD samples, thereby capturing the implicit coupling between ID and OOD distributions in real-world open settings. At the methodological level, we introduce an energy-based detection mecha-

收稿日期:2025-09-04;录用日期:2025-12-29;责任编辑:覃怀银
*通信作者:王飞

nism that maps the outputs of graph neural networks to energy values, which quantify the consistency of samples with the training distribution. The imposed energy constraints guide the model to learn separable representations, where ID samples concentrate in low-energy regions while potential OOD samples are pushed toward higher-energy regions. This design alleviates the overconfidence issue of Softmax-based confidence methods under distribution shifts and allows the detection objective to directly influence graph representation learning. To effectively solve the resulting constrained optimization problem, we adopt an augmented Lagrangian approach that dynamically balances constraint satisfaction and objective optimization during training, enhancing model stability under mixed distributions. Experimental results on multiple real-world graph datasets demonstrate significant performance improvements. On the Twitch dataset, the proposed method achieves an AUROC of 95.97% and an AUPR of 92.84%, outperforming the current state-of-the-art baseline GNNsafe++ by over 21 percentage points, while maintaining a false positive rate of 12.30%. These results confirm the effectiveness and robustness of the proposed framework under fully unsupervised and open-world conditions.

Keywords: out-of-distribution detection; graph neural networks; node classification; machine learning; wild data; energy function

0 引言

随着大数据时代的到来,图数据因能够直观且精准地描述实体间复杂关系,逐渐成为机器学习研究的热点。它广泛应用于社交网络分析[1]、生物信息学[2]、推荐系统[3]等领域,为各种任务提供丰富的上下文信息和结构化知识。然而,图数据的复杂性及特有分布特性给处理和分析带来了诸多挑战。

图神经网络(Graph Neural Networks,GNNs)近年来在图数据处理上表现出强大能力,能够学习节点嵌入并捕捉图中的隐藏模式,在节点分类和链接预测等任务中效果显著。然而,现有GNNs主要关注分布内(In-Distribution,ID)数据的性能,对于分布外(Out-Of-Distribution,OOD)数据的处理能力相对有限[4]。在实际应用中,模型不可避免地会遇到与训练数据分布不同的新样本。对于可信的人工智能系统而言,核心挑战是如何稳健识别和处理这些置信度较低的OOD数据,否则模型可能出现显著性能下降或误导性预测[5],尤其在安全关键场景如医疗诊断[6]和自动驾驶[7]中,错误预测可能带来严重后果。因此,提高模型对OOD数据的识别与处理能力至关重要[8]。

本文工作深入探讨了OOD检测在图学习模型中的应用及其重要性。尽管已有研究尝试通过引入辅助异常数据进行模型正则化以提升OOD检测性能,这些方法通常依赖一个不切实际的假设:辅助异常数据与训练分布数据完全可分[9-17]。然而,在现实场景中这一假设往往不成立[18]。

以推荐系统为例,模型通常在特定社交网络图上训练以预测用户感兴趣的内容。现有方法要么仅基于标注完备的ID数据进行封闭世界下的分类学习,要么在训练阶段引入少量已知的OOD样本,并假设其与ID数据之间具有明确的判别边界。这类设置在一定程度上简化了问题建模,但忽略了真实开放环境中ID与OOD数据天然混杂且普遍缺乏可靠标注的客观事实。

针对上述局限,本文提出了一种更贴近真实应用场景的训练范式,通过构建无标注的ID/OOD混合训练集,使模型在训练阶段持续暴露于潜在的分布外样本,并在不依赖任何显式先验分布假设的条件下,实现ID分类与广义OOD检测的联合优化,从而增强模型对动态数据分布的适应能力并提升整体鲁棒性。相关问题设定与方法动机的整体示意如图1所示。

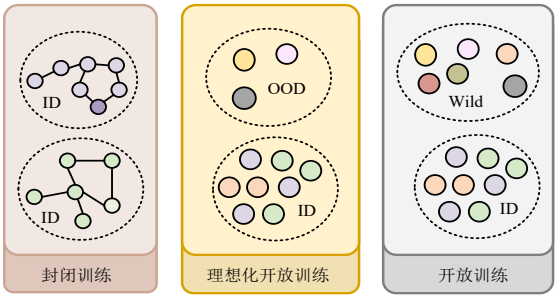


图1 不同开放性假设下的训练范式

Figure 1 Training paradigms under different openness assumptions

为提高图学习模型在数据分布变化下的准确性与可靠性,本文提出了一种基于能量函数的OOD检测框架[19]。该框架利用GNNs学习表示空间中的能量分布差异,有效区分特征分布变化与类别含义变化,并通过未标注野生数据提升模型对分布外数据的真实检测能力。进一步地,通过构建带约束的优化问题,使模型学习更加稳健的决策边界,实现高OOD检测率的同时严格控制ID数据的分类误差。实验部分验证了方法在多个真实世界数据集上的有效性,包括社交网络图和推荐系统等五个应用场景。结果显示,Open-GOD(Open-Graph Out-of-Distribution detection)在无需强假设的条件下显著提升了模型对OOD数据的检测性能,证明了其在提升图学习模型泛化能力与鲁棒性方面的潜力。

1　相关工作

1.1　图的分布外检测

OOD检测通过识别推理样本与训练分布之间的偏离,为机器学习系统的安全部署提供关键保障[20]。针对图结构数据,Cai等人[21]和Li等人[22]从理论建模、任务划分与评估协议等方面系统总结了图OOD检测的发展脉络,为该领域研究奠定了基础。在基准构建方面,Gui等人[23]提出的GOOD基准通过引入可控且具有语义意义的分布偏移,为模型鲁棒性评估提供了统一框架。在方法层面,围绕图OOD检测任务,近期研究提出了多种具有代表性的模型设计:Huang等人[24]构建了端到端的开放集半监督节点分类框架,将OOD检测与分类任务进行联合建模;Bao等人[25]从邻域塑造的角度探索了OOD节点的识别机制;Wang等人[26]则尝试在图级异常检测中建立统一建模框架。这些方法在不同任务设定下取得了积极进展,但图数据的非欧几里得结构也为OOD检测带来了独特挑战:节点属性与拓扑结构的强耦合使分布偏移路径高度复杂,传统基于特征统计的方法难以刻画其隐性关联;动态图的时序演化进一步加剧了分布漂移的不确定性;同时,GNNs的消息传递机制对局部扰动具有放大效应,细微偏移即可引发全局预测失稳。基于上述挑战,现有研究主要沿着不确定性量化与特征解耦学习两条技术路线展开探索。

在不确定性量化方面,Zhao等人[27]提出了基于图的不确定性感知学习框架,通过核狄利克雷分布估计(Gaussian Kernel Dirichlet Estimation,GKDE)预测节点级狄利克雷分布,从而实现OOD节点检测;Stadler等人[28]提出的图后验网络(Graph Posterior Network,GPN)通过显式贝叶斯后验更新建模预测不确定性,为节点分类中的OOD识别提供了新思路。

在特征解耦方向,Ma等人[29]提出的DisenGCN和Yang等人[30]提出的FactorGCN通过多通道建模将节点表征解耦至不同潜在因子空间,以期捕获与标签因果相关的信息;Bevilacqua等人[31]从全局视角识别决定类别的不变因果子图,以增强泛化能力;Fan等人[32]引入基于注意力的因果发现模块,实现节点级与图级特征的协同稳定;Ganguly等人[33]提出的半监督图特征网络则从特征层面探索了OOD检测的新范式。

尽管上述方法在受控设定下表现良好,但在全开放环境中仍存在显著局限。不确定性量化方法[27-28]高度依赖于ID数据上的置信度校准,在复杂分布偏移下容易失效;而基于因果与不变性学习的方法[29-33]通常需要预先定义多个训练环境或偏移来源,难以适用于ID与OOD样本天然混合且无任何分布先验的Wild Mix场景。这一现实挑战促使研究者探索无需强假设、能够直接从混合数据中学习稳健判别边界的新型OOD检测范式。

1.2　能量基模型

能量基模型(Energy-Based Models,EBMs)是一类通过能量函数对数据进行建模的方法,其核心思想是为输入样本分配一个标量能量值,以刻画其与训练数据分布的一致性。该模型最早由Ackley等人[34]在1985年提出,不同于显式概率分布建模方法,EBMs通过能量函数刻画数据分布,逐渐发展为一种统一的学习框架,可适用于多种概率与非概率学习任务[35-36]。在EBMs中,能量函数$E(x)$将输入空间中的样本映射为实数值,较低的能量通常对应于高概率密度区域,其与概率分布之间的关系可通过Gibbs分布进行刻画[37]。

近年来,能量基方法在OOD检测领域受到广泛关注。Cong等人[38]提出的Sneakoscope重新审视了无监督OOD检测中的能量建模基础,为后续研究提供了重要启发。在模型训练过程中,EBMs通常通过最小化负对数似然损失进行优化,使模型对分布内样本赋予较低能量值,而对分布外样本赋予较高能量值,从而实现对异常样本的有效区分。

在图数据与OOD检测的背景下,能量函数为衡量样本分布偏离程度提供了一种自然的度量方式。Xu等人[39]利用能量基模型对复杂数据分布进行建模,验证了其在高维结构化数据建模中的潜力。Grathwohl等人[40]提出的Jacobian Energy Model(JEM)通过联合建模输入与输出分布,在OOD检测任务中取得了良好效果。在实际应用中,能量函数常被直接用作OOD打分指标,用以评估输入样本与训练数据分布的一致性。

与基于Softmax置信度的OOD检测方法相比,基于能量函数的方法在理论上与样本概率密度更加一致,能够有效缓解过度自信的问题。Liu等人[41]提出的ARC(A Generalist Graph Anomaly Detector with In-Context Learning)检测器展示了能量建模在上下文学习中的应用潜力;Pan等人[42]通过引入异质性引导机制,为图欺诈检测提供了新的能量建模视角。此外,能量函数具有较强的灵活性,可直接从判别式分类模型中导出,无需复杂的生成式模型或显式密度估计,计算高效,便于在图神经网络框架中部署。

1.3　数据增强与训练策略

近年来,数据驱动方法在图学习领域取得了显著进展,相关研究主要从数据增强和训练策略两个层面提升模型的泛化能力。在数据层面,Zheng等人[43]从数据中心视角系统梳理了图机器学习的发展路径;

Feng 等人[44]提出的 GRAND(Graph Random Neural Networks for Anomaly Detection)通过随机丢弃节点特征,缓解模型对局部邻域的过度依赖;Kong 等人[45]提出的 FLAG(Fast Local Augmentation for Graph anomaly detection)在特征空间中引入对抗扰动,生成更具挑战性的训练样本;Zhao 等人[46]进一步通过可微分边预测器构造增强图,在保持模型结构不变的情况下有效提升节点级任务性能。

在训练策略层面,不变学习通过引入跨环境一致性约束提升模型稳定性。Gui 等人[47]提出的 G-adapter 为图 Transformer 提供了结构感知的参数高效迁移方案;Miao 等人[48]通过随机注意力机制增强模型的可解释性与泛化能力。此外,自监督学习方法通过设计预训练任务从无标注图数据中学习通用表征,例如 You 等人[49]的多视角对比学习框架,以及 Maskey 等人[50]提出的图提示学习方法,均在下游任务中展现出良好的迁移效果。

2 问题定义

2.1 带标签的分布内数据

在监督学习框架下,模型通常处理由特征向量和标签组成的数据对。带标签的分布内数据集可以表示为一个样本集合 $\{(x_i,y_i)\}_{i=1}^{N}$,其中 $x_i\in\mathbf{R}^d$ 是第 i 个样本的特征向量,$y_i\in Y$ 是对应的标签,N 是样本总数,而 Y 是标签空间。

假设这些数据是从某个未知的概率分布 $P_{\text{data}}(x,y)$ 中独立同分布 P_{ID} 中抽取的,其中 P_{data} 是定义在特征空间 $\mathbf{R}^d$ 和标签空间 Y 上的联合概率分布。在训练阶段,目标是找到一个模型 $f:\mathbf{R}^d\to Y$,它能够以最小的期望损失 L 映射输入特征到正确的标签:

$$f^*=\arg\min_f E_{(x,y)\sim P_{\text{data}}}\left[L(f(x),y)\right] \tag{1}$$

其中,L 指损失函数,用于衡量模型预测和真实标签之间的差异。

2.2 分布外数据

在现实世界中部署机器学习模型时,可靠的分类器不仅要准确地对已知 ID 样本进行分类,还要将模型在训练过程中没有接触过的来自不同分布的输入样本识别为“未知”。除了多类分类器 f_θ 之外,还可以通过 OOD 分类器来实现这一点。为了处理 OOD 数据,需要定义一个与 P_{data} 不同的概率分布 P_{OOD},该分布描述了在实际应用中可能遇到的与训练数据分布不同的数据。本文的目标是建模一个能够区分来自 P_{data} 和 P_{OOD} 的数据的函数。被检测为 OOD 的样本将被剔除,被检测为 ID 的样本将由 f_θ 进行分类。OOD 检测可以表述为一个二元分类问题:

$$D_{\text{OOD}}=\left\{(x_{\text{OOD}},y_{\text{OOD}})\right\}_{i=1}^{N_{\text{OOD}}} \tag{2}$$

在测试时,目标是判断测试时的输入 $x\in\mathcal{X}$ 是否来自 ID 样本。将 $g_\theta:\mathcal{X}\mapsto\{\text{in},\text{out}\}$ 作为 OOD 检测的函数映射。

OOD 检测问题可以表述为一个最小化期望损失的问题,其中模型需要在保持对 P_{data} 的低损失的同时,对 P_{OOD} 的数据具有高损失。这可以通过以下优化问题来形式化:

$$\min_f E_{x\sim P_{\text{data}}}\left[L\left(f(x),y\right)\right]+\lambda E_{x\sim P_{\text{OOD}}}\left[M\left(f(x)\right)\right] \tag{3}$$

其中,M 是一个度量模型对 OOD 数据敏感度的函数,例如可以是模型输出的概率或能量分数。参数 λ 是一个正则化系数,用于平衡对分布内数据和分布外数据的关注度。

2.3 无标签的野生数据构建

为了更真实地模拟 OOD 场景,本文利用了模型运行环境中自然产生的未标记野生数据,将这些数据从潜在的威胁转化为有价值的学习资源。

无标签野生数据集 D_{wild} 是从真实世界环境中收集的未经标注的数据样本集合。本文将自然产生的未标记野生样本 $\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_m$ 融入 OOD 检测中,这些样本被假设为从多个未知的概率分布 P_{wild_k} 中抽取,其中 k 表示不同的数据源或环境。形式上,无标签野生数据可以表示为

$$f^*=\arg\min_f E_{(x,y)\sim P_{\text{data}}}\left[L\left(f(x),y\right)\right] \tag{4}$$

具体而言,我们考虑了野生数据的一种广义特征,即它们可以被建模 ID 数据和 OOD 数据边缘分布的混合组合:

$$P_{\text{wild}}:=(1-\pi)P_{\text{in}}+\pi P_{\text{out}} \tag{5}$$

其中,$\pi\in(0,1]$,本文通过设计一个带约束的优化目标,将 P_{wild} 整体视为 OOD 数据的代理,并约束模型在纯净 ID 数据 P_{in} 上的性能,从而实现了无需先验标注即可从野生混合数据中学习分布外检测的能力。通过这种方式,无标签野生数据可以帮助模型学习到从源域到目标域的映射,其中源域数据 D_{src} 是有标签的,而目标域数据 D_{trg} 是无标签的。通过最小化源域和目标域数据的期望损失,可以训练一个在目标域上表现良好的模型:

$$\begin{aligned}f^*=\arg\min_f\ & E_{x_{\text{src}}\sim P_{\text{src}}}\left[L\left(f(x_{\text{src}}),y_{\text{src}}\right)\right]\\ &+E_{x_{\text{trg}}\sim P_{\text{trg}}}\left[L\left(f(x_{\text{trg}})\right)\right]\end{aligned} \tag{6}$$

其中,P_{src} 和 P_{trg} 分别表示源域和目标域的概率分布。通过这种方式,无标签野生数据为机器学习模型提供

了一种在未知环境中学习和适应的手段,从而提高了模型在实际应用中的有效性和鲁棒性。

2.4 学习目标

为了更清晰地刻画不同应用场景下模型的评价方式,如图2所示,本文区分了封闭世界与开放世界两种典型设定。在封闭世界中,训练与测试数据共享相同的分布,而在开放世界设定下,模型在测试阶段将不可避免地面对来自未知分布的OOD样本,且其分布可能与训练阶段所接触的OOD数据不同。

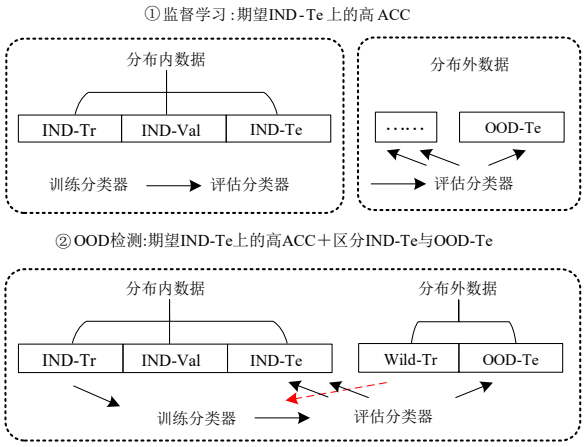


图2　不同世界设定下的模型评估场景

Figure 2　Evaluation scenarios under different world settings

本文的学习框架是通过利用 P_{in} 和 P_{wild} 的数据,建立OOD分类器 g_θ 和多类分类器 f_θ。使用共享参数集 θ^* 表示它们可能共享神经网络参数,使得以下目标函数最小化:

$$\theta^*=\arg\min_{\theta}\left(\lambda_1\cdot\text{FPR}(\theta)+\lambda_2\cdot\left(1-\text{TPR}(\theta)\right)+\lambda_3\cdot\left(1-\text{ACC}(\theta)\right)\right) \tag{7}$$

其中,λ_1、λ_2、λ_3 是用于权衡假阳性率(False Positive Rate,FPR)、真阳性率(True Positive Rate,TPR)和准确率(ACCuracy,ACC)重要性的超参数。通过选择合适的 λ 值,可以根据不同的应用场景调整模型对各个性能指标的关注程度。

在测试中,测量了以下误差:

$$\min_f E_{x\sim P_{\text{data}}}\left[L\left(f(x),y\right)\right]+\lambda E_{x\sim P_{\text{OOD}}}\left[M\left(f(x)\right)\right] \tag{8}$$

其中,$X\to\{-\}$ 为指示函数,箭头表示越高/越低越好。为符合现实开放世界的设定,将OOD训练分布 P_{out} 和OOD测试分布 $P_{\text{out}}^{\text{test}}$ 区分开来,因为在分布外检测中,可用的训练数据可能与测试时的数据不同,因此将 $P_{\text{out}}^{\text{test}}=P_{\text{out}}$ 的情况称为静态分布。

3 基于能量函数的分布外检测方法

本节提出的分布解耦机制是本文方法的核心思想。其核心在于,通过能量函数的约束优化,引导图神经网络学习一个低维表示空间,使ID样本对应的能量显著低于OOD样本,从而形成两个可分离的能量分布。这种能量层面的分离为ID与OOD的判别提供了清晰、稳定的决策边界,有效刻画复杂的分布偏移。

整体框架如图3所示,图3①描述测试阶段对混合分布数据的建模,以反映真实场景中的隐式分布偏移;图3②展示训练阶段利用野生数据对模型进行约束优化,使其持续接触未标注的OOD样本;图3③给出了最终ID与OOD能量分布的分离结果。能量阈值通过验证集确定,低能量区域对应ID数据的稳定拓扑模式,高能量区域则反映结构或属性层面的分布偏移。为求解该约束问题,本文将其集成至GNNs训练过程,并采用增广拉格朗日方法进行优化。

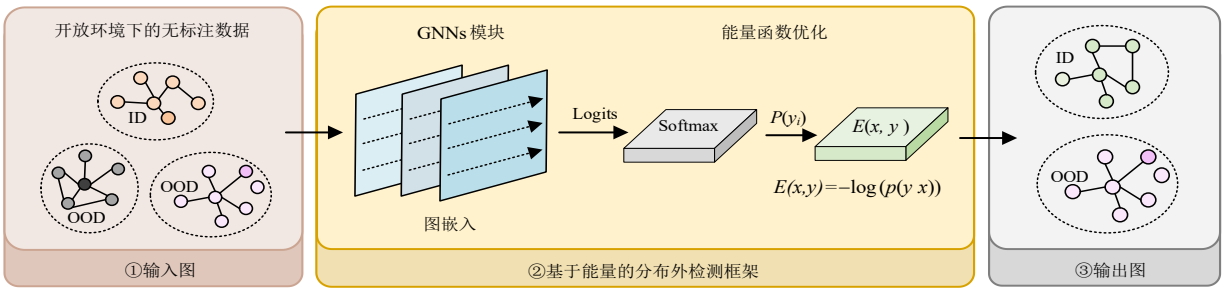


图3　基于能量函数的分布外检测框架示意图

Figure 3　Framework of the proposed energy-based out-of-distribution detection method

3.1 能量函数的引入

能量函数从神经网络的Logits计算而来,是一个将输入数据映射到单一标量值的函数,这个标量值对于观测到的数据较低,而对于未观测到的数据较高。能量函数的理论基础与概率密度有关,具有更高能量的样本可以被解释为出现概率较低的数据。

具体来说,能量函数可以表示为

$$E(x;f)=-T\cdot\log\left(\sum_{i=1}^{K}e^{f_i(x)/T}\right) \tag{9}$$

其中,$f_i(x)$ 是神经网络对于输入 x 的第 i 个类别的Log-

its 输出；K 是类别的总数；T 是温度参数，它影响能量函数的平滑程度。

在本文中，能量值由 GNNs 最后一层的 Logits 计算。通过温度参数 T 调节能量平滑度，低能量区域对应 ID 数据的拓扑特征一致性，即节点特征与图结构的高度匹配；高能量区域反映结构或属性的分布偏移，即节点特征与图结构的显著偏离。这种设计使得能量函数能够有效捕捉图数据中的分布偏移，为 OOD 检测提供了一种与图结构特性直接相关的度量方式。

尽管能量函数由 Logits 计算而得，但其与传统的最大 Softmax 概率存在本质区别，并具有理论上的优越性。MSP（Maximum Softmax Probability）输出是模型在已知类别上的归一化置信度，其设计初衷是执行分类而非检测异常，因此容易对分布外样本产生过度自信的预测[11]。相反，能量函数通过未归一化的 Logits 之和提供了一个与输入数据概率密度 $p(x)$ 直接相关的度量（由 Gibbs 分布 $p(x)=\mathrm{e}^{-E(x)/T}/z$ 定义）。这种与概率密度的直接对应关系，使能量值能够可靠刻画样本相对于训练分布的偏离程度：ID 样本位于数据流形上，对应较低能量；OOD 样本偏离流形，对应较高能量。因此，基于能量函数构建检测框架，从理论上缓解了 Softmax 预测过度自信的问题，也是本文方法性能提升的重要原因之一。

需要指出的是，尽管能量函数定义在输出层 Logits 上，但其本质与图数据的结构建模密切相关。GNNs 的 Logits 是前端消息传递与邻域聚合的综合结果，因此能量值 $E(x)$ 实际反映了节点属性与其局部拓扑上下文的联合状态。异常的高能量既可能源于节点特征偏移，也可能来自邻域结构的异常变化。通过端到端的能量约束优化，ID/OOD 区分目标被反向传播至各 GNNs 层，引导模型学习对分布偏移更敏感的图表示，从而在不显式修改网络结构的情况下，实现分类与检测的协同优化。

能量函数的负值可以被解释为样本属于已知类别的置信度。基于这个能量函数，损失函数 L_{OOD} 可以采用 Sigmoid 交叉熵的形式：

$$L_{\mathrm{OOD}}\left(g_\theta(x),\mathrm{in}\right)=\frac{1}{m}\sum_{i=1}^{m}\sigma\left(-g_\theta(x_i)\right) \tag{10}$$

其中，$g_\theta(x)$ 是基于能量函数的 OOD 检测器的输出，σ 是 Sigmoid 函数，m 是样本数量。

3.2 带约束的优化边际

能量限定学习目标是一种用于训练神经网络的损失函数，用于区分 ID 样本与 OOD 样本。它通过最小化组合损失，在 ID 与 OOD 样本之间显式建立能量差异。该损失包括标准分类损失和能量正则化项：交叉熵损失确保 ID 样本的分类性能；能量正则化项则鼓励模型为 ID 样本分配低能量、为 OOD 样本分配高能量，并通过平方铰链损失惩罚不符合预期的样本。其形式如下：

$$\begin{aligned}L_{\mathrm{energy}}=&\sum_{\mathrm{IDsamples}}\max\left(0,E(x)-\mathrm{margin}_{\mathrm{in}}\right)^2\\&+\sum_{\mathrm{OODsamples}}\max\left(0,\mathrm{margin}_{\mathrm{out}}-E(x)\right)^2\end{aligned} \tag{11}$$

其中，$\mathrm{margin}_{\mathrm{in}}$ 和 $\mathrm{margin}_{\mathrm{out}}$ 是预设的阈值，且满足 $\mathrm{margin}_{\mathrm{in}}<\mathrm{margin}_{\mathrm{out}}$。第一项惩罚那些能量值高于 $\mathrm{margin}_{\mathrm{in}}$ 的 ID 样本，将其能量值“推低”；第二项惩罚那些能量值低于 $\mathrm{margin}_{\mathrm{out}}$ 的 OOD 样本，将其能量值“推高”。通过这种双向约束，模型被强制在能量轴上拉大 ID 与 OOD 分布之间的距离，从而实现解耦。

这里，$E(x)$ 是模型对样本 x 输出的能量值，margin 是一个预先设定的阈值，用于定义 ID 样本和 OOD 样本之间的能量差距。第一项惩罚那些能量值高于 margin 的 ID 样本，第二项惩罚那些能量值低于 margin 的 OOD 样本。

能量函数提供了一种与概率密度对齐的理论方法，用于评估样本异常程度。带约束的优化目标使模型在正确分类 ID 数据的同时，也能有效识别和拒绝 OOD 样本。通过调整正则化项和能量阈值，可进一步提升模型对 OOD 样本的敏感性，使其在图结构数据中更好地适应数据分布变化。

3.3 动态邻域采样与邻居纯度验证

为提升模型在复杂图结构中的适应性与可扩展性，本文引入了动态邻域采样与邻居纯度验证机制。

（1）动态邻域采样策略。针对大规模图中节点度分布不均的问题，提出基于节点度的自适应采样策略。采样率计算公式如下：

$$p(v)=\min\left(1,\alpha\cdot\frac{\log(\deg(v)+\epsilon)}{\deg_{\mathrm{avg}}}\right) \tag{12}$$

其中，$\deg(v)$ 表示节点 v 的度；$\deg_{\mathrm{avg}}$ 为图中节点的平均度；α 为调节因子；ϵ 为平滑项。该策略确保低度节点保留更多邻居信息，同时控制高度节点的计算复杂度。

具体采样过程如算法 1 所示。

（2）邻居纯度验证模块。为增强对拓扑噪声的鲁棒性，设计了邻居纯度验证机制。邻居纯度定义为

$$\mathrm{Purity}(v)=\frac{1}{\left|\mathcal{N}(v)\right|}\sum_{u\in\mathcal{N}(v)}\mathbb{I}(\hat{y}_u=\hat{y}_v) \tag{13}$$

其中，$\mathcal{N}(v)$ 表示节点 v 的邻居集合，$\hat{y}_u$ 和 $\hat{y}_v$ 分别为节点 u 和 v 的预测标签。该纯度度量被整合到能量损失函数中：

$$L'_{\mathrm{energy}}=L_{\mathrm{energy}}+\beta\cdot\sum_{v\in V}\left(1-\mathrm{Purity}(v)\right)\cdot E(v) \tag{14}$$

算法1　动态邻域采样

输入:节点集 $\boldsymbol{V}$,邻接矩阵 $\boldsymbol{A}$,采样率参数 α
输出:采样后的邻接矩阵 $\boldsymbol{A}'$
计算每个节点的 $\deg(v)$ 和平均度 $\deg_{avg}$
for each 节点 $v \in V$ do
计算采样率 $p(v)$ ing Eq. (X)
从 v 的邻居中按概率 $p(v)$ 随机采样
更新邻接矩阵 $\boldsymbol{A}'$
end for
return $\boldsymbol{A}'$

其中,β 为平衡超参数。该设计使模型对低纯度邻域中的节点赋予更高的能量惩罚,增强了对结构异常节点的识别能力。

3.4　约束优化建模

本节将OOD检测建模为一个带约束的优化问题:

$$\begin{aligned}&\inf_{\theta} E\\ &\text{s.t. } E_{x\sim P_{\text{in}}}\left[\mathbb{I}(g_\theta(x)=\text{out})\right]\leqslant\alpha\\ &E_{(x,y)\sim P_{xy}}\left[\mathbb{I}(f_\theta(x)\neq y)\right]\leqslant\tau\end{aligned}\tag{15}$$

该优化问题依据如下。

目标函数:驱使模型最小化将野生数据错误接受为ID样本的期望,其本质是最大化对OOD样本的识别率。

约束(1):确保模型对ID数据的误检率低于阈值 α。该约束是模型可靠性的关键保障,可有效避免因过度检测而导致的系统可用性下降。

约束(2):确保模型在分布内数据上的分类错误率低于阈值 τ。从而保证OOD检测功能的引入不会破坏模型原有的分类能力,即维持其核心任务的性能。

通过求解这个约束优化问题,期望得到一个同时在分类精度、检测可靠性和OOD发现率上表现优异的模型。

学习目标是在满足特定约束的前提下,最小化从 P_{out} 中分类数据为ID的误差,这个目标可以表述为一个带有约束的优化问题。旨在优化以下目标:

$$\begin{aligned}&\inf_{\theta} E_{x\sim P_{\text{out}}}\left(X\to\left\{g_\theta(x)=\text{in}\right\}\right)\\ &\text{s.t. } E_{x\sim P_{\text{in}}}\left(X\to\left\{g_\theta(x)=\text{out}\right\}\right)\leqslant\alpha\\ &E_{(x,y)\sim P_{xy}}\left(X\to\left\{f_\theta(x)\neq y\right\}\right)\leqslant\tau\end{aligned}\tag{16}$$

然而,通常无法直接访问一个纯净的 P_{out} 数据集,为了解决这个问题,使用 P_{wild} 来代替 P_{out},等价的目标可以表示为

$$\begin{aligned}&\inf_{\theta} E_{x\sim P_{\text{wild}}}\left(X\to\left\{g_\theta(x)=\text{in}\right\}\right)\\ &\text{s.t. } E_{x\sim P_{\text{in}}}\left(X\to\left\{g_\theta(x)=\text{out}\right\}\right)\leqslant\alpha\\ &E_{(x,y)\sim P_{xy}}\left(X\to\left\{f_\theta(x)\neq y\right\}\right)\leqslant\tau\end{aligned}\tag{17}$$

在经验上,可以通过最小化来自 P_{wild} 分布且被标记为ID的样本数量 $\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_m$ 来解决优化问题。这需要满足两个条件:

(1)正确标记至少 $1-\alpha$ 的ID样本 $x_1,x_2,\cdots,x_n$。

(2)达到分类性能阈值,即在ID数据上的分类准确率达到 τ。

可以转化为以下优化问题:

$$\begin{aligned}&\inf_{\theta}\frac{1}{m}\sum_{i=1}^{m}X\to\left\{g_\theta(\tilde{x}_i)=\text{in}\right\}\\ &\text{s.t.}\frac{1}{n}\sum_{i=1}^{n}X\to\left\{g_\theta(x_i)=\text{out}\right\}\leqslant\alpha\\ &\frac{1}{n}\sum_{i=1}^{n}X\to\left\{f_\theta(x_i)\neq y_i)\right\}\leqslant\tau\end{aligned}\tag{18}$$

由于0/1损失函数的不可微性,这个问题难以直接求解。为了解决这个问题,使用了一个易于处理的代理问题来替代原始的优化问题。使用连续可微的损失函数来代替0/1损失。这个代理问题可以表示为

$$\begin{aligned}&\inf_{\theta}\frac{1}{m}\sum_{i=1}^{m}L_{\text{OOD}}\left(g_\theta(\tilde{x}_i),\text{in}\right)\\ &\text{s.t.}\frac{1}{n}\sum_{j=1}^{n}L_{\text{OOD}}\left(g_\theta(x_j),\text{out}\right)\leqslant\alpha\\ &\frac{1}{n}\sum_{j=1}^{n}L_{\text{cls}}\left(f_\theta(x_j),y_j\right)\leqslant\tau\end{aligned}\tag{19}$$

其中,L_{OOD} 是二进制OOD分类器的损失函数,L_{cls} 是分类任务的损失函数。选择Sigmoid损失函数 $\sigma(t)=\frac{1}{1+e^{-t}}$ 作为 L_{OOD},以及Hinge损失函数作为 L_{cls}。

3.5　增广拉格朗日法

为了将带约束的优化问题转化为无约束优化问题,本文引入了一种新的基于增广拉格朗日方法的训练过程[47],算法2详细描述了其迭代优化步骤。首先固定拉格朗日乘子,通过梯度下降更新主模型参数 θ 以最小化增广拉格朗日损失 $L(\theta,\lambda,\mu)$;随后再固定模型参数,根据约束违反情况动态更新乘子 λ、μ,以增强约束满足,定义拉格朗日函数 $L(\theta,\lambda,\mu)$ 如下:

$$L(\theta,\lambda,\mu)=F(\theta)+\lambda\left(G(\theta)-\alpha\right)+\mu\left(H(\theta)-\tau\right)\tag{20}$$

其中,$\lambda\geqslant 0$ 和 $\mu\geqslant 0$ 是非负的拉格朗日乘子,用于惩罚约束条件的违反。

为了处理可能的不等式约束,引入增强项 $\Psi(\beta)$,其中 β 是一个正的调节参数,用于增强对约束违反的

惩罚。增强拉格朗日函数$L(\theta,\lambda,\mu,\beta)$定义为

$$L(\theta,\lambda,\mu,\beta)=L(\theta,\lambda,\mu)+\frac{\beta}{2}\left(G(\theta)-\alpha\right)^2+\frac{\beta}{2}\left(H(\theta)-\tau\right)^2 \quad (21)$$

ALM通过迭代更新θ、λ、μ来最小化拉格朗日函数$L(\theta,\lambda,\mu)$。在每次迭代k中，ALM执行以下步骤。

(1)参数更新。在每次迭代k中，首先固定λ、μ和β，然后通过梯度下降或其他优化算法来更新模型参数θ：

$$\theta^{(k+1)}=\arg\min_{\theta} L(\theta,\lambda^{(k)},\mu^{(k)},\beta^{(k)}) \quad (22)$$

此步骤中优化的模型参数θ包含了GNNs编码器的所有权重。通过最小化融合了能量约束与分类约束的增广拉格朗日损失，梯度信号会从末层的能量与Logits计算反向传播至前面的消息传递层与特征变换层，从而端到端地塑造整个网络的表征能力，使其学习到的节点嵌入同时有利于精确的ID分类和显著的OOD检测。

(2)乘子更新。更新拉格朗日乘子λ和μ，以及增强参数β，以确保约束条件得到满足：

$$\begin{aligned}\lambda^{(k+1)}&=\lambda^{(k)}+\beta\left(G(\theta^{(k+1)})-\alpha\right)\\ \mu^{(k+1)}&=\mu^{(k)}+\beta\left(H(\theta^{(k+1)})-\tau\right)\\ \beta^{(k+1)}&=\beta^{(k)}+\rho\end{aligned} \quad (23)$$

(3)迭代。重复步骤(1)和(2)，直到满足收敛条件，例如θ、λ和μ的变化小于某个阈值，或者达到预设的迭代次数。

增广拉格朗日法的训练流程如算法2所示。

通过这个迭代过程，ALM能够逐步找到满足所有约束的模型参数θ，同时最小化目标函数$F(\theta)$。

算法2 增广拉格朗日法训练流程

输入：训练数据(x,y)，初始模型参数θ，拉格朗日乘子λ，增强参数ρ，最大迭代次数T

输出：训练后的模型参数θ^*

```
for t = 1 to T do
    # 1. 参数更新
    固定λ和ρ,通过梯度下降更新θ:
    θ ← θ − η ∇θ L(θ, λ, ρ)
    # 2. 乘子更新
    更新拉格朗日乘子λ:
    λ ← λ + ρ* (约束条件违反量)
    # 3. 增强参数更新
    更新增强参数ρ:
    ρ ← α*ρ#α为增强因子,通常 α > 1
    # 4. 检查收敛条件
    if |θ_t − θ_{t−1}| < ε or t == T then
        break
end for
return θ*
```

4 实验分析

4.1 数据集构建与评价指标

本文选取了多个广泛使用的图基准数据集进行评估，涵盖引文网络(Cora[48]、ogbn-Arxiv[49])、电商与推荐场景(Amazon-Photo[50])、学术合作网络(Coauthor-CS[51])以及开放社交网络(Twitch-Explicit[52])。各数据集在节点语义、结构特性和任务设置上具有显著差异，可全面评估模型在不同图场景下的泛化能力。具体数据集统计信息如表1所示。

表1 数据集的详细信息

Table 1 Detailed statistics of the datasets

数据集	类型	节点数	边数	特征维度	类别数
Cora	引文网络	2 708	5 429	1 433	7
Amazon-Photo	商品共购网络	77 650	238 162	745	8
Coauthor-CS	作者共作网络	8 333	163 788	6 805	15
Twitch-Explicit	社交网络	变化	变化	2 545	变化
ogbn-Arxiv	引用网络	变化	变化	128	变化

ID数据均采用半监督设置，按8∶1∶1比例划分为训练、验证和测试集。各数据集的Wild数据构建方式和混合比例如表2所示。其中，Cora、Amazon-Photo和Coauthor-CS的扰动包括结构缺失、特征缺失和标签缺失。结构缺失通过随机删除部分边或邻居信息模拟图拓扑不完整；特征缺失通过随机掩盖部分节点特征来模拟信息不全；标签缺失通过随机去除部分训练或测试标签来模拟半监督开放环境。该设置旨在评估模型在多样化分布外样本和受损数据下的泛化能力与鲁棒性。

4.2 模型性能分析实验

为了全面评估Open-GOD的性能，本文选取了三类代表性基线方法进行对比：基于置信度的后验方法(如MSP、ODIN)，直接利用分类器的Softmax输出进行检测；基于特征空间的方法(如Mahalanobis、GKDE)，在隐表示空间中进行密度估计或距离度量；依赖OOD曝光的方法(如OE、Energy FT、GNNsafe++)，其训练过程需要访问部分标注或纯净的OOD样本，通常作为性能上界的参考。通过这些设置，旨在验证Open-GOD在不依赖OOD曝光和强分布假设的条件下，是否仍能达到甚至超越强监督方法的检测能力。

表2　数据集划分策略

Table 2　Dataset splitting strategy

数据集	ID数据来源	Wild数据构建方式
Twitch-Explicit	Twitch-DE子图	其他时间片的Twitch子图
ogbn-Arxiv	2015年之前发表的论文	2017年后发表的论文
Cora	原始图	对测试集施加三种扰动
Amazon -Photo	原始图	对测试集施加三种扰动
Coauthor-CS	原始图	对测试集施加三种扰动

如表3所示,Open-GOD在多个基准数据集上均取得显著优势。在最具挑战性的Wild Mix场景中,Twitch数据集上的AUROC和AUPR分别达到95.97%和92.84%,较当前最优基线GNNsafe++提升约21个百分点,同时将FPR降至12.30%。这些结果表明,基于野生混合数据的训练范式能够有效缓解对显式OOD样本的依赖,使模型在真实开放环境下具备更强的分布外检测能力。

表3　不同基准模型下的性能对比分析　　单位:%

Table 3　Performance comparison across different baseline models　　unit: %

Model	Dataset	Twitch				Arxiv			
		AUROC	AUPR	FPR	ACC	AUROC	AUPR	FPR	ACC
MSP[11]	IN	33.59	49.14	97.45	68.72	63.91	75.85	90.59	53.78
ODIN[14]	IN	58.16	72.12	93.96	70.79	55.07	68.85	100.00	51.39
Mahalanobis[20]	IN	55.68	66.42	90.13	70.51	56.92	69.63	94.24	51.59
Energy[18]	IN	51.24	60.81	91.61	70.40	64.20	75.78	90.80	53.36
GKDE[21]	IN	46.48	62.11	95.62	67.44	58.32	72.62	93.84	50.76
GNNSAFE[21]	IN	66.82	70.97	76.24	70.40	71.06	80.44	87.01	53.39
OE[12]	OOD Expo	55.72	70.18	95.07	70.73	69.80	80.15	85.16	52.39
Energy FT[18]	OOD Expo	84.50	88.04	61.29	70.52	71.56	80.47	80.59	53.26
GNNSAFE++[21]	OOD Expo	95.36	97.12	33.57	70.18	74.77	83.21	77.43	53.50
MSP[11]	Wild Mix	75.83	74.47	84.27	67.18	62.48	79.54	94.03	52.01
OE[12]	Wild Mix	67.77	63.92	94.23	65.19	67.82	77.21	84.52	53.19
Energy FT[18]	Wild Mix	75.03	81.95	80.70	70.36	72.04	83.95	69.43	58.27
GNNSAFE++[21]	Wild Mix	74.75	71.71	92.66	53.45	72.54	76.28	90.23	50.68
Open-GOD	Wild Mix	95.97	92.84	12.30	78.46	81.02	85.39	81.91	56.30

Model	Dataset	Cora (AUROC)			Amazon (AUROC)			Coauthor (AUROC)		
		S	F	L	S	F	L	S	F	L
MSP[11]	IN	70.90	85.39	91.36	98.27	97.31	93.97	95.30	97.05	94.88
ODIN[14]	IN	49.92	49.88	49.80	93.24	81.15	65.97	52.14	51.54	51.44
Mahalanobis[20]	IN	46.68	49.93	67.62	71.69	76.50	73.25	80.46	93.23	85.36
Energy[18]	IN	71.73	86.15	91.40	98.51	97.87	93.81	96.18	97.88	95.87
GKDE[21]	IN	68.61	82.79	57.23	76.39	58.96	65.58	65.87	80.69	61.15
GNNSAFE[21]	IN	87.52	93.44	92.80	99.58	98.55	97.35	99.60	99.64	97.23
GNN[2]	OODExpo	67.98	81.83	89.47	99.60	98.39	95.39	97.86	99.04	96.04
Energy FT[18]	OODExpo	75.88	88.15	91.36	98.83	98.55	97.35	98.84	99.43	96.23
GNNSAFE++[21]	OODExpo	90.62	95.56	92.75	99.82	99.64	97.51	99.99	99.97	97.89
MSP[11]	Wild Mix	72.39	83.27	91.12	98.01	98.05	92.64	92.31	93.87	90.96
ODIN[14]	Wild Mix	40.11	40.82	40.15	91.43	93.09	87.22	94.58	94.72	93.28
OE[12]	Wild Mix	49.03	53.01	61.72	70.15	71.89	69.81	92.19	94.12	91.38
Energy FT[18]	Wild Mix	71.58	73.59	72.83	70.24	71.14	69.52	94.47	95.55	93.27
Mahalanobis[20]	Wild Mix	52.16	53.17	69.18	71.32	76.33	73.11	82.59	96.23	93.69
Open-GOD	Wild Mix	91.63	96.19	93.55	99.99	99.99	98.59	99.99	99.99	98.45

注:S表示结构缺失,F表示特征缺失,L表示标签缺失。本实验通过这三种缺失类型系统评估模型在不同OOD场景下的鲁棒性。

针对不同类型的OOD数据，Open-GOD在结构缺失、特征缺失和标签缺失场景下均表现良好。在Cora数据集上，其AUROC分别达到91.63%、96.19%和93.55%，其中在特征缺失场景中较传统能量检测方法提升超过10个百分点。在Amazon数据集的结构缺失任务中，AUROC达到99.99%，各子任务均接近理论上限，体现出较强的鲁棒性。进一步的可视化结果（图4、图5）也验证了上述结论。

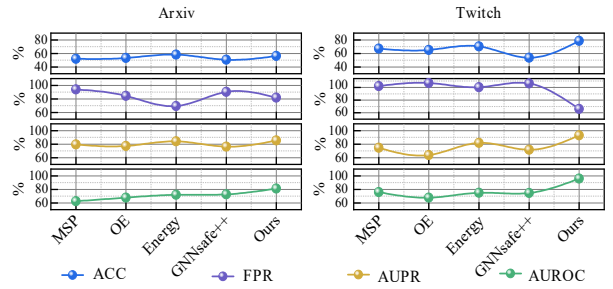


图4 Arxiv与Twitch数据集上的模型性能对比

Figure 4 Performance comparison of different methods on the Arxiv and Twitch

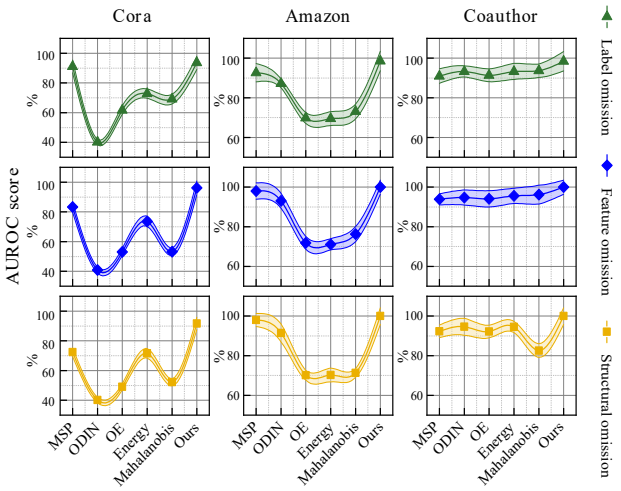


图5 模型在Cora/Amazon/Coauthor数据集上AUROC的对比实验

Figure 5 AUROC comparison of different methods on the Cora, Amazon, and Coauthor

4.3 能量分数可视化分析

为深入探究模型的决策机制与局限性，本节在能量分布可视化的基础上，系统分析了Open-GOD的能量分离效果及其典型错误模式。图6展示了Twitch和Arxiv数据集上的ID与OOD样本的能量分布对比。

可以观察到，基础能量函数产生的ID/OOD样本能量分布存在明显重叠，导致部分样本处于分类模糊区间。引入边际约束后，两类样本的能量分布分离度显著提升，重叠区域明显缩小。经过增广拉格朗日乘子优化的Open-GOD方案展现出优越的性能，其能量分布呈现出明显的双峰特征，有效扩大了ID/OOD样本间的能量差距，实现了显著的分布解耦效果。

4.4 组件消融分析与鲁棒性分析

通过系统性的消融实验（表4），分析了Open-GOD各核心组件对OOD检测性能的影响。在Twitch Wild Mix场景下，基准模型仅获得51.24%的AUROC，表明原始架构对复杂OOD模式的识别能力有限。引入Wild Mix后，AUROC提升至64.84%，但ACC下降至65.19%，说明单纯的数据混合策略会削弱模型对ID分布的判别能力。

相比之下，ER表现出良好的稳定性，单独使用即可将AUROC提升至66.82%，同时保持原有70.40%的ACC。这表明ER通过约束能量分布有效提升了模型对异常样本的敏感性。进一步地，ER与ALM联合使用时，AUROC和ACC分别提升至95.97%和78.46%，

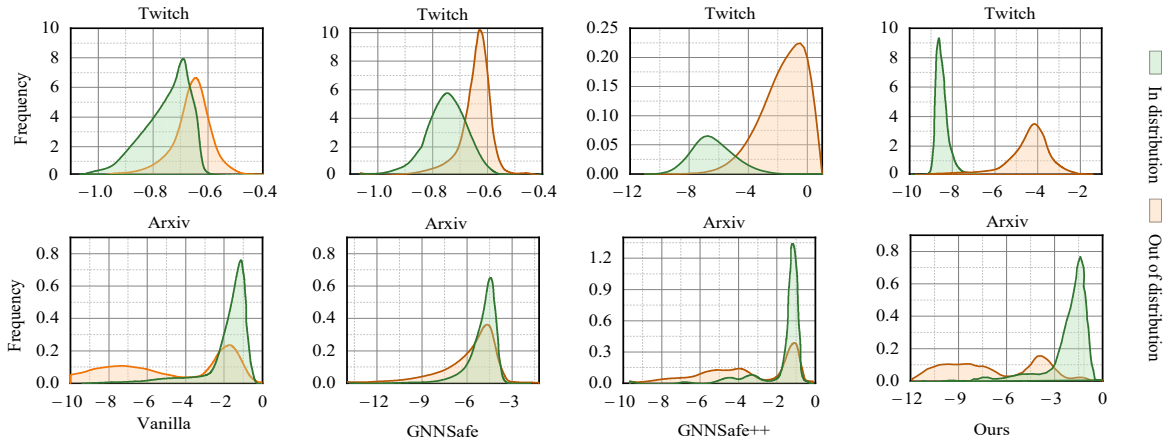


图6 Twitch和Arxiv数据集上的能量分布对比

Figure 6 Comparison of energy distributions on the Twitch and Arxiv dataset

表4　Open-GOD在Twitch数据集上的消融综合实验结果

Table 4　Ablation study results of Open-GOD on the Twitch dataset

Wild Mix	ER	ALM	AUROC	ACC
			51.24(44.73↓)	70.40
√			64.84(31.13↓)	65.19
	√		66.82(29.15↓)	70.40
		√	56.42(39.55↓)	68.44
√	√		73.91(22.06↓)	60.08
√		√	95.97	78.46
√	√	√	89.31(6.66↓)	68.50

注:ER表示能量正则化,ALM表示增广拉格朗日乘子。括号内数值(X↓)表示表示该配置的AUROC相较于最优配置(Wild Mix + ER + ALM,AUROC=95.97%)的百分比下降幅度。

体现了ALM在多目标优化中的动态平衡作用。

当三种组件同时启用时,性能出现一定回落,原因在于Wild Mix引入的噪声与ER的特征压缩约束在高度优化的特征空间中产生冲突。这一结果表明,在复杂混合数据场景下,需要更精细的数据质量控制与协同优化策略。

在计算效率方面(表5),Open-GOD表现出显著优势。在Twitch数据集上,训练速度为0.008 s/样本,推理速度为0.039 s/样本,比GNNsafe++提升约25%。效率提升得益于动态邻域采样、能量阈值批处理以及拉格朗日乘子的异步更新。在Arxiv超大规模数据集上,模型仍能保持0.073 s/样本的训练效率和0.257 s/样本的推理速度,同时GPU内存占用降低约37%。结果表明,即使引入能量正则化和增广拉格朗日优化,Open-GOD在多数场景下仍能保持与轻量基线相当的效率,结合卓越的检测性能,展现出良好的工程可实现性和实际应用价值。

4.5　参数选取综合实验

为了探究模型性能对超参数的依赖性,本节系统地分析了在Twitch数据集中传播步长K和自循环强度α的影响。实验结果如图7所示,表明了模型性能对这两个超参数的选择较为敏感,但存在明确的优化规律。当K=1时,模型性能随α增大呈先升后降趋势,最佳α=0.5(AUROC=86.55%)。但在$K\geq5$时,α=0.1始终表现最优,说明深层传播中需抑制过强的自循环。在α=0.1条件下,模型性能随K增加呈现三阶段变化:K=1→5时AUROC显著提升11.33%;K=5→12时小幅提升0.29%;$K\geq12$后趋于稳定。这表明信息传播在12步后达到充分均衡。

基于实验结果,推荐采用K=12和α=0.1的组合,既可获得97.55的峰值性能,又能保证计算效率。需特别避免$\alpha\geq0.7$且$K\leq5$的低效组合。该分析表明模型在$K\geq12$和α=0.1时具有稳定的优异表现。

表5　模型在不同数据集上每个节点的训练时间与推理时间对比

单位:s

Table 5　Comparison of per-node training and inference time across different datasets　　unit: s

Model	Twitch		Arxiv	
	TR	IN	TR	IN
MSP	0.008	0.040	0.043	0.172
OE	0.013	0.045	0.007	0.251
Energy FT	0.016	0.052	0.069	0.218
GNNsafe++	0.016	0.052	0.075	0.240
Open-GOD	0.008	0.039	0.073	0.257

Model	Cora		Amazon		Coauthor	
	TR	IN	TR	IN	TR	IN
MSP	0.008	0.021	0.015	0.026	0.058	0.181
ODIN	0.008	0.027	0.013	0.081	0.069	0.285
OE	0.015	0.015	0.019	0.025	0.124	0.180
Energy FT	0.013	0.018	0.018	0.024	0.131	0.171
Mahalanobis	0.008	0.251	0.013	0.306	0.073	1.081
Open-GOD	0.010	0.014	0.012	0.030	0.120	0.173

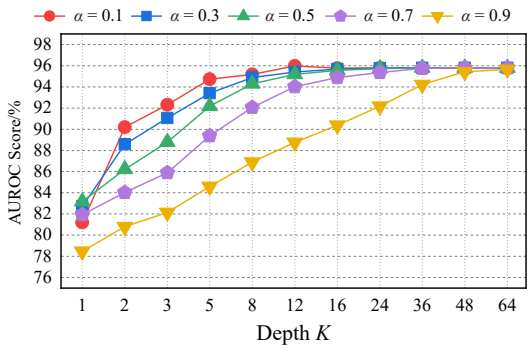


图7　Twitch数据集上传播步数K与自循环权重α的超参数敏感性分析

Figure 7　Hyperparameter sensitivity analysis of the propagation depth K and self-loop weight α on the Twitch

5　总结与未来工作

本文提出了一种基于野生混合数据的图OOD检测框架Open-GOD,通过能量正则化与增广拉格朗日约束的协同优化,实现了在无标注ID/OOD混合数据上稳定识别分布外样本。实验结果表明,该方法在多个基准图数据集上显著提升了OOD检测性能,同时保持了可靠的ID分类精度,验证了其在真实开放环境下的有效性和实用性。

展望未来,研究将主要沿两方面展开:一是扩展Open-GOD,以适应动态图、稀疏或低可分OOD场景;二是探索自适应优化与鲁棒性增强策略,以进一步提升模型在复杂开放环境中的稳定性和应用价值。

参考文献

[1] Guo Xiaoyu, Liu Yan, Gong Daofu, et al. Dual graph convolutional networks for social network alignment[J]. IEEE Transactions on Big Data, 2025, 11(2): 684-695.

[2] Zhang Xiaomeng, Liang Li, Liu Lin, et al. Graph neural networks and their current applications in bioinformatics[J]. Frontiers in Genetics, 2021, 12: 690049.

[3] Wu Shiwen, Sun Fei, Zhang Wentao, et al. Graph neural networks in recommender systems: A survey[J]. ACM Computing Surveys, 2023, 55(5): 1-37.

[4] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 427-436.

[5] Bevandić P, Šegvić S, Krešo I, et al. Discriminative out-of-distribution detection for semantic segmentation[C]//Proceedings of the 7th International Conference on Learning Representations. Munich: OpenReview.net, 2019: 18-36.

[6] Kukar M. Transductive reliability estimation for medical diagnosis[J]. Artificial Intelligence in Medicine, 2003, 29(1/2): 81-106.

[7] Dai Dengxin, Van Gool L. Dark model adaptation: Semantic image segmentation from daytime to nighttime[C]//2018 21st International Conference on Intelligent Transportation Systems. Piscataway: IEEE, 2018: 3819-3824.

[8] Nalisnick E, Matsukawa A, Teh Y W, et al. Do deep generative models know what they don't know[C/OL]//Proceedings of the 7th International Conference on Learning Representations. OpenReview.net, 2019: https://openreview.net/pdf?id=H1xwNhCcYm.

[9] 席亮, 刘涵, 樊好义, 等. 基于深度对抗学习潜在表示分布的异常检测模型[J]. 电子学报, 2021, 49(7): 1257-1265.
Xi Liang, Liu Han, Fan Haoyi, et al. Deep adversarial learning latent representation distribution model for anomaly detection[J]. Acta Electronica Sinica, 2021, 49(7): 1257-1265. (in Chinese)

[10] Fang Zhen, Lu Jie, Zhang Guangquan. Out-of-distribution detection with non-semantic exploration[J]. Information Sciences, 2025, 705: 121989.

[11] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[C/OL]//Proceedings of the 5th International Conference on Learning Representations. OpenReview.net, 2017: https://openreview.net/references/pdf?id=Hkg4TI9xl.

[12] DeVries T, Taylor G W. Learning confidence for out-of-distribution detection in neural networks[PP/OL]. V1. arXiv (2018-02-13)[2025-09-01]. https://doi.org/10.48550/arXiv.1802.04865.

[13] Hein M, Andriushchenko M, Bitterwolf J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 41-50.

[14] Hsu Y C, Shen Yilin, Jin Hongxia, et al. Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10948-10957.

[15] Sun Yiyou, Guo Chuan, Li Yixuan. ReAct: Out-of-distribution detection with rectified activations[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2021: 144-157.

[16] Bitterwolf J, Meinke A, Augustin M, et al. Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities[C]//Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022: 2041-2074.

[17] 李思聪, 王坚, 宋亚飞, 等. TriCh-LKRepNet: 融合三通道映射与结构重参数化的大核卷积恶意代码分类网络[J]. 电子学报, 2024, 52(7): 2331-2340.
Li Sicong, Wang Jian, Song Yafei, et al. TriCh-LKRepNet: A large kernel convolutional malicious code classification network for structure reparameterisation and triple-channel mapping[J]. Acta Electronica Sinica, 2024, 52(7): 2331-2340. (in Chinese)

[18] Liu Weitang, Wang Xiaoyun, Owens J D, et al. Energy-based out-of-distribution detection[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 21464-21475.

[19] Ranzato M, Boureau Y L, Chopra S, et al. A unified energy-based framework for unsupervised learning[C]//Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. San Juan: PMLR, 2007: 371-392.

[20] Liang Shiyu, Li Yixuan, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks[C]//Proceedings of the 6th International Conference on Learning Representations. OpenReview.net, 2018.

[21] Cai Tingyi, Jiang Yunliang, Liu Yixin, et al. Out-of-distribu-

tion detection on graphs: A survey[PP/OL]. V1. arXiv (2025-02-12)[2026-03-30]. https://doi.org/10.48550/arXiv.2502.08105.

[22] Li Haoyang, Wang Xin, Zhang Ziwei, et al. Out-of-distribution generalization on graphs: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(11): 10490-10512.

[23] Li Haoyang, Wang Xin, Zhang Ziwei, et al. GOOD: A graph out-of-distribution benchmark[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 068431-0150.

[24] Huang Tiancheng, Wang Donglin, Fang Yuan, et al. End-to-end open-set semi-supervised node classification with out-of-distribution detection[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022: 2087-2093.

[25] Bao Tianyi, Wu Qitian, Jiang Zetian, et al. Graph out-of-distribution detection goes neighborhood shaping[C]//Proceedings of the 41st International Conference on Machine Learning. New York: ACM, 2024: 2923-2943.

[26] Wang Yili, Liu Yixin, Shen Xu. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark[C/OL]//Proceedings of the 13th International Conference on Learning Representations. OpenReview.net, 2025: https://openreview.net/forum?id=g90RNzs8wX.

[27] Zhao Xujiang, Chen Feng, Hu Shu, et al. Uncertainty aware semi-supervised learning on graph data[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 1076.

[28] Stadler M, Charpentier B, Geisler S, et al. Graph posterior network: Bayesian predictive uncertainty for node classification[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2021: 1380.

[29] Guo Jingwei, Huang Kaizhu, Yi Xinping, et al. Learning disentangled graph convolutional networks locally and globally[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3640-3651.

[30] Yang Yiding, Feng Zunlei, Song Mingli, et al. Factorizable graph convolutional networks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 20286-20296.

[31] Bevilacqua B, Zhou Yangze, Ribeiro B. Size-invariant graph representations for graph classification extrapolations[C]//Proceedings of the 38th International Conference on Machine Learning. Cambridge: PMLR, 2021: 837-851.

[32] Fan Shaohua, Wang Xiao, Shi Chuan, et al. Generalizing graph neural networks on out-of-distribution graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 322-337.

[33] Ganguly D, Gupta D. Detecting out-of-distribution data with semi-supervised graph "feature" networks[C/OL]//Proceedings of the 11th International Conference on Learning Representations. OpenReview.net, 2023: https://openreview.net/forum?id=0OlEBibFa_g.

[34] Ackley D, Hinton G, Sejnowski T. A learning algorithm for Boltzmann machines[J]. Cognitive Science, 1985, 9(1): 147-169.

[35] Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Cambridge: JMLR, 2010: 693-700.

[36] Du Yilun, Li Shuang, Sharma Y, et al. Unsupervised learning of compositional energy concepts[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2021: 15608-15620.

[37] Cong Tianji, Prakash A. Sneakoscope: Revisiting unsupervised out-of-distribution detection[C/OL]//Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

[38] Ranzato M, Poultney C, Chopra S, et al. Efficient learning of sparse representations with an energy-based model[M]//Schölkopf B, Platt J, Hofmann T. Advances in neural information processing systems 19. Cambridge: The MIT Press, 2007: 1137-1144.

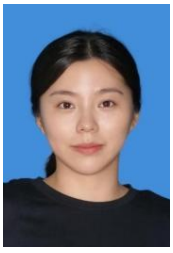
[39] Xu Pingmei, Ehinger K A, Zhang Yinda, et al. TurkerGaze: Crowdsourcing saliency with webcam based eye tracking[PP/OL]. V2. arXiv (2015-05-20) [2025-09-01]. https://doi.org/10.48550/arXiv.1504.06755.

[40] Grathwohl W, Wang K C, Jacobsen J H, et al. Learning the Stein discrepancy for training and evaluating energy-based models without sampling[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 3732-3747.

[41] Liu Yixin, Li Shiyuan, Zheng Yu, et al. ARC: A generalist graph anomaly detector with in-context learning[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. New York: Curran Asso-

ciates Inc., 2024: 1606.
[42] Pan Junjun, Liu Yixin, Zheng Xin, et al. A label-free heterophily-guided approach for unsupervised graph fraud detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(12): 12443-12451.
[43] Zheng Xin, Liu Yixin, Bao Zhifeng, et al. Towards data-centric graph machine learning: Review and outlook[PP/OL]. V1. arXiv (2023-09-20)[2026-03-30]. https://doi.org/10.48550/arXiv.2309.10979.
[44] Feng Wenzheng, Zhang Jie, Dong Yuxiao, et al. Graph random neural networks for semi-supervised learning on graphs[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 1853.
[45] Kong Kezhi, Li Guohao, Ding Mucong, et al. Robust optimization as data augmentation for large-scale graphs[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 60-69.
[46] Zhao Tong, Liu Y, Neves L, et al. Data augmentation for graph neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(12): 11015-11023.
[47] Hestenes M R. Multiplier and gradient methods[J]. Journal of Optimization Theory and Applications, 1969, 4(5): 303-320.
[48] Sen P, Namata G, Bilgic M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93-106.
[49] Hu Weihua, Fey M, Zitnik M, et al. Open graph benchmark: Datasets for machine learning on graphs[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 1855.
[50] McAuley J, Targett C, Shi Qinfeng, et al. Image-based recommendations on styles and substitutes[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2015: 43-52.
[51] Sinha A, Shen Zhihong, Song Yang, et al. An overview of microsoft academic service (MAS) and applications[C]//Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 243-246.
[52] Rozemberczki B, Sarkar R. Twitch gamers: A dataset for evaluating proximity preserving and structural role-based node embeddings[PP/OL]. V2. arXiv (2021-02-16) [2025-09-01]. https://doi.org/10.48550/arXiv.2101.03091.

作者简介



李思聪 女,2000年8月出生于陕西省西安市。现为国防科技大学计算机学院博士研究生。主要研究方向为异常检测、开放环境下的机器学习。
E-mail: lisicong@nudt.edu.cn

魏子令 男,1991年11月出生于湖南省邵阳市。现为国防科技大学副研究员。主要研究方向为网络取证和网络优化。
E-mail: weiziling@nudt.edu.cn

王 飞 女,1985年3月出生于吉林省吉林市。现为国防科技大学副研究员。主要研究方向为机器学习、网络流量异常检测。
E-mail: wangfei09a@nudt.edu.cn

陈曙晖 男,1974年7月出生于湖南省长沙市。现为国防科技大学研究员。主要研究方向为网络取证和网络流量处理。
E-mail: shchen@nudt.edu.cn