计算机科学
COMPUTER SCIENCE

基于自适应图自编码器的离群点检测方法

谭淇尹，于炯，陈子歆

引用本文

谭淇尹，于炯，陈子歆. 基于自适应图自编码器的离群点检测方法[J]. 计算机科学，2025，52(6)：129-138.

TAN Qiyin, YU Jiong, CHEN Zixin. Outlier Detection Method Based on Adaptive Graph Autoencoder[J]. Computer Science, 2025, 52(6): 129-138.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500092

基于自适应图自编码器的离群点检测方法

谭淇尹[1]　于　炯[1,2]　陈子歆[1]
1 新疆大学软件学院　乌鲁木齐 830000
2 新疆大学信息科学与工程学院　乌鲁木齐 830000
(tanqiyin@stu.xju.edu.cn)

摘　要　离群点检测(Outlier Detection)是通过识别数据集中不同于大多数样本的少量个体来获取数据的整体健康状态与异常信息。目前,在处理欧氏结构数据集时,大部分检测算法侧重于将数据视为独立的个体,却忽视了数据实例之间的相关性。这种信息偏向性导致了一些可能位于正常数据区域内的潜在的离群值难以被有效检测出来。针对上述问题,提出了一种基于自适应邻居的图自动编码器的深度联合表示学习算法 ANGAE(Adaptive Neighbor Graph Autoencoder)。该算法从图生成的角度构建图来捕捉数据点之间的关系,并利用结构和属性自动编码器学习数据的潜在表示。ANGAE 引入了自适应邻居构图机制,以动态更新图结构,确保在模型训练过程中对不准确的图结构进行调整和改进。通过融合结构嵌入和属性嵌入,ANGAE 实现了网络结构和节点属性之间的有效交互。实验结果表明,所提出的方法在 11 个数据集上表现优异,在保持高精度的同时展现了很好的鲁棒性,其有效性得到了充分证明。

关键词:离群点检测;深度学习;图卷积网络;图表示学习;属性网络
中图分类号　TP391.4

Outlier Detection Method Based on Adaptive Graph Autoencoder

TAN Qiyin[1], YU Jiong[1,2] and CHEN Zixin[1]
1 School of Software Engineering, Xinjiang University, Urumqi 830000, China
2 College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China

Abstract　Outlier detection involves identifying a small number of individuals in a dataset that differ from the majority of samples, thereby obtaining insights into the overall health and abnormal information of the data. Currently, in the context of Euclidean structured datasets, most detection algorithms predominantly treat data as independent entities, overlooking the correlations between data instances. This informational bias hinders the effective identification of potential outliers that might exist within the normal data regions. To address this issue, this paper proposes a deep joint representation learning algorithm named adaptive neighbor graph autoencoder(ANGAE). This algorithm constructs a graph from the perspective of graph generation to capture the relationships between data points and leverages structural and attribute autoencoders to learn latent representations of the data. ANGAE introduces an adaptive neighbor graph construction mechanism to dynamically update the graph structure, ensuring the adjustment and improvement of inaccurate graph structures during model training. By integrating structural embeddings and attribute embeddings, ANGAE facilitates effective interaction between network structure and node attributes. Experimental results demonstrate that the proposed method achieves superior performance across 11 datasets, maintaining high precision while exhibiting robust resilience, thereby substantiating the method's efficacy.

Keywords　Outlier detection, Deep learning, Graph convolutional networks, Graph representation learning, Attribute networks

1　引言

离群点检测的首要目标是辨别出与绝大多数样本显著不同的个别观测值,以便探测出不符合预期模式或统计规律的离群值。一方面,离群点的存在可能蕴含着重要信息,暗示着数据收集过程中出现了错误或者数据生成机制上发生了变化;另一方面,离群值的存在也可能导致正常数据的分布偏离预期模式或统计规律,对数据分析和模型的准确性构成干扰,进而影响对数据的深入理解和可靠推断。因此,离群点检测通过辨识存在的离群值对数据进行进一步的分析或处理,使数据更为可靠、有意义,是一项重要的任务[1]。研究人员也可以通过离群值检测探索数据中的潜在信息,拓展对数据

到稿日期:2024-05-22　返修日期:2024-09-20
基金项目:国家自然科学基金(62262064)
This work was supported by the National Natural Science Foundation of China(62262064).
通信作者:于炯(yujiong@xju.edu.cn)

本质、特征和分布的深入理解，为更好的数据决策提供支持。总体来说，离群点检测在数据分析中扮演着重要角色，并且在各行各业中具有重要的应用价值。它可以用于识别金融行业的诈骗行为[2-3]，在社交平台上筛查水军账号或者虚假新闻[4-5]，甚至在工业领域用于排查机械故障[6-7]。

依据 HawKins 提出的概念，离群点被描述为与正常数据点相差很远，以至于令人怀疑是由不同机制产生的数据点。基于这个定义，我们可以理解为，分析离群点，实质上是对其与邻近点的关系进行分析判断。目前，针对欧氏数据的离群点检测算法在处理数据时通常基于数据点的相对独立性。然而，这些方法的不足之处在于，它们忽视了同一数据集中数据之间可能存在的更深层次的关联。Li 等[8]指出数据之间具有实例相关性，数据实例可以根据它们的特征互相关联。例如，具有相似个人资料或在线行为的用户往往对广告或商品有相似的偏好；具有相似临床数据或症状的患者患类似疾病的可能性更高。忽略数据点之间的潜在关系，也就意味着忽略它们背后可能包含的复杂的依赖性和上下文信息。因此，现有的方法难以捕捉复杂的异常模式，容易忽略隐藏在正常数据区域内的离群值。为了解决这一问题，我们提出了一种新的自适应邻居的图自编码器（ANGAE）模型来检测离群点。该方法旨在从生成图的角度构建节点连接关系，从结构和属性两个角度改变数据的原始分布，并学习空间中的潜在表示。同时，为了实现节点结构和属性的联合学习，将结构嵌入和属性嵌入融合为特征解码器的最终输入，以生成节点属性。在编码器和解码器阶段之后，根据节点的重构误差来发现异常节点。本文的主要贡献如下。

1）采用动态近邻图学习方法，针对欧氏数据构建相似图。以概率方式为每个目标自适应分配最优邻域，并在学习特征的过程中实时更新图结构，有效提高了构建图的质量。

2）引入了一种深度联合表示学习框架，采用双自动编码器。这一框架旨在捕捉网络结构和节点属性之间交织复杂的互动关系，以全面评估离群情况。该方法能提高检测的准确性和鲁棒性，从而更好地识别离群点。

3）在 11 个真实世界数据集上对 ANGAE 的有效性进行了综合评估，实验结果表明了 ANGAE 在多种环境下优于其他常用的算法。

2 相关工作

2.1 离群点发展

经过多年的发展和研究，离群点检测领域积累了许多经典的方法。按照处理数据方式的不同，目前常见的离群点检测方法大致可以分为两类：基于传统方法的离群点检测和基于深度学习的离群点检测。

基于传统方法的离群点检测主要分为 4 种：基于概率统计的离群点检测、基于近邻的离群点检测、基于聚类的离群点检测和基于分类的离群点检测。基于概率统计的离群点检测方法依赖于数据分布的统计建模，通过分析数据点的概率特征和统计属性来发现离群值[9]。然而，这种方法对于高维和复杂数据模式的适应性有所不足，并且受限于对数据分布的假设。基于近邻的离群点检测方法则分为基于距离和基于密度的两种，重点分析数据点在邻域内的特征和关系。基于密度的经典离群点检测方法有：局部异常因子的 LOF[10]、基于连通性的离群因子 COF[11]和基于局部相关性的 LOCI[12]等。然而，这些方法在处理高维数据和噪声时面临着维度灾难所带来的计算成本呈指数增长的挑战，难以全面考虑整体情况。基于聚类的方法利用数据的聚类特性识别离群值，常见的 K-means[13]和 DBSCAN[14]等算法虽易实现，但对邻域大小、聚类数量等参数较为敏感。基于分类的离群点检测方法则通过建立分类器界定正常值与离群值，如 SVM[15]和 IForest[16]，但在极度不平衡数据中可能偏向于多数类别，无法有效检测离群值。

基于深度学习的离群值检测通过使用不同的神经网络模型来学习数据的复杂特征表示，从而探测出不符合正常规律的离群值[17]。根据用于提取数据特征的神经网络类型，深度离群点检测可分为基于自编码器的深度离群点检测和基于生成模型的深度离群点检测。基于自编码器的深度离群点检测通常由编码器和解码器组成，用于学习输入数据的紧凑表示。自编码器无须标记异常样本，而是通过高重构误差辨识离群值，适用于学习复杂特征的离群值检测方法。基于生成模型的深度离群值检测是利用生成对抗网络[18]（GAN）或变分自编码器[19]（VAE）等生成模型，通过学习数据的潜在分布和生成新样本的能力来检测离群点。在此思路的基础上，Liu 等[20]提出 SO-GAAL 模型，该模型通过生成器和鉴别器之间的零和博弈直接生成信息丰富的潜在离群值。为了避免模型陷入模式坍塌问题，他们将生成器的网络结构扩展为具有不同目标的多个生成器。离群点通常不能被生成模型很好地还原，因此可以被检测出来。Du 等[21]利用 GAN 的分布拟合能力生成虚假数据来扩充正常数据集，然后利用自编码器检测真实数据的重建误差，进而识别离群点。

近年来，研究者在离群点检测领域逐渐表现出对图数据的强烈兴趣。相较于传统的欧氏数据，图数据包含更丰富的属性信息，能够更有效地对复杂系统进行建模。因此，他们积极探索并应用图神经网络[22]（GNN）在图数据上进行离群点检测[23-24]，旨在充分利用这一强大工具的潜力。举例来说，Ding 等[25]使用图卷积神经网络（GCN）嵌入属性网络，利用嵌入向量对原始数据进行重构并将重构误差作为异常分数，以实现离群点检测。同时，Li 等[26]在嵌入过程中引入了拉普拉斯锐化技术，以增强正常节点与离群点之间的表示差异，更易于检测到离群点。鉴于 GNN 惊人的容量和强大的特征学习能力，这些算法表现突出。因此，研究者想要将 GNN 的强大数据处理能力应用于不同领域。然而，在大多数情况（如文本、信号等）下，图没有作为先验信息给出，这意味着现有的基于 GNN 的节点离群点检测模型无法直接应用于这些欧氏结构数据。为了解决这一问题，我们提出了一种专门用于处理欧氏结构数据的图神经网络模型，代表了对图神经网络创新型拓展应用的一种探索。

2.2 图构造

图构造在图数据分析领域中具有重要意义，其核心在于将现实世界的数据抽象为图形结构，以发现离散点之间的隐藏关系。基于欧氏距离的距离矩阵或表示数据点之间关系的

关系矩阵，按照不同的连接密度，其构图方法通常可以分为完全连接图和 k 最近邻图。

2.2.1 完全连接图

完全连接图将所有节点直接相连，通过赋予相似数据点更高的权重来凸显它们之间的联系；同时赋予不相似的数据点较小的权重，以凸显它们之间的差异。通常使用基于欧氏距离的高斯核函数构建关系矩阵，其中 σ 作为宽度参数，直接影响数据分布和关系矩阵的性能。

$$s_{ij}=\mathrm{e}^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}} \tag{1}$$

2.2.2 k 近邻图

k 近邻图与完全连接图有着相同的理念，不同的是，该方法针对每个节点，在数据集中选择其 k 个最近的邻居节点作为其直接邻居，并在图中用边连接这些邻居节点。这种构图方式在保留局部连接性的同时减小了图的密度，使得图更具稀疏性。其中，k 值的选择影响着图的稀疏程度和局部特征的捕获情况。基于 k 近邻法的关系矩阵构造如式(2)所示：

$$s_{ij}=\begin{cases}0, & j\notin i\text{ 的 }k\text{ 近邻}\\ \mathrm{e}^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}}, & j\in i\text{ 的 }k\text{ 近邻}\end{cases} \tag{2}$$

2.3 符号说明

为方便理解，表 1 对本文所用到的符号进行了详细说明。在本文中，使用粗体小写字母(如 $\boldsymbol{s}$)代表向量，用粗体大写字母(如 $\boldsymbol{S}$)代表矩阵。用 $\mathcal{G}=(\boldsymbol{A},\boldsymbol{X})$ 表示一个图结构，其中 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 表示这个图的邻接矩阵，$\boldsymbol{X}\in\mathbb{R}^{n\times d}$ 表示图的属性矩阵。属性矩阵 $\boldsymbol{X}$ 的第 i 行向量 $\boldsymbol{x}_i\in\mathbb{R}^d$ 表示第 i 个节点的属性向量。$\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 中的每个元素 $\boldsymbol{A}_{i,j}$ 表示节点 i 与节点 j 之间的连接关系。在一个有向或无向图中，矩阵 $\boldsymbol{A}$ 的行和列分别代表各个节点，而矩阵中的数值表示节点间是否存在连接或边的权重信息。若节点 i 与节点 j 之间存在连接，则 $A_{i,j}$ 为 1(或非零值)，否则为 0(或零值)。

表 1 符号及其含义

Table 1 Symbols and their meanings

符号	意义
$\boldsymbol{s}$	向量
$\boldsymbol{S}$	矩阵
$\mathcal{G}=(\boldsymbol{A},\boldsymbol{X})$	图结构
$\boldsymbol{A}\in\mathbb{R}^{n\times n}$	$\mathcal{G}$的邻接矩阵
$\boldsymbol{X}\in\mathbb{R}^{n\times d}$	$\mathcal{G}$的属性矩阵
$\boldsymbol{x}_i\in\mathbb{R}^d$	$\mathcal{G}$中第 i 个节点的属性向量
$\|\cdot\|_2^2$	$L2$ 范数的平方
$\|\cdot\|_F^2$	矩阵的 F 范数
n	$\mathcal{G}$中节点的个数
d	$\mathcal{G}$中属性的维度

3 ANGAE 模型

3.1 算法介绍

目前，基于图神经网络的图离群点检测方法利用了其优异的拓扑结构捕捉能力，在性能方面取得了显著进展。然而，这些方法的成功在很大程度上依赖于准确、完整地反映节点之间的连接关系的高质量图结构。但是，现实世界的图数据常常受到噪声或不准确性的影响，从而对图嵌入表示学习造成负面影响。为此，我们研究了基于欧氏数据的构图方法。然而，构建完全连接图的时间复杂度高，而 k 近邻图存在参数敏感性的问题；另外，这些方法构建的图结构是静态的，无法灵活适应新知识的学习，并且初始图结构可能并非最优。针对这些局限性，本研究提出了一种新颖的自适应邻居图自编码器(ANGAE)，以在统一框架下学习准确的图结构和区分性的图嵌入表示。ANGAE 框架包括 3 个主要模块：图构建模块、属性自编码器和结构自编码器。这一创新性方法的整体框架如图 1 所示。ANGAE 的引入旨在提升离群点检测的性能，降低参数敏感性的影响，并能够自适应地学习新知识，从而更精准地捕获数据特征和图结构。

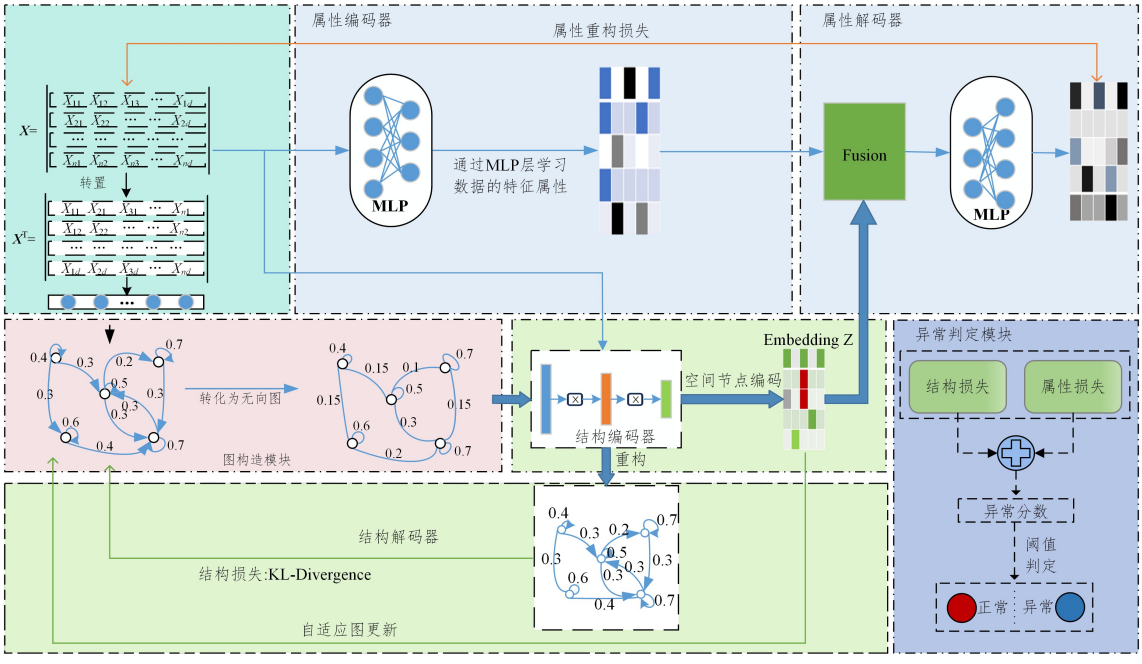


图 1 ANGAE 框架

Fig. 1 ANGAE framework

3.1.1 图构造器模块

受现有构图方法的启发，基于数据点v_i与v_j之间的相似性关联进行构图，得出了概率意义上的度量权重s_{ij}。在相似度计算方法上，为了降低原始数据中可能存在的噪声和杂质的影响，使用数据的潜在特征来计算相似度。这种方法能够更准确地反映数据点之间的关联，进而提高构图的质量和鲁棒性。在此基础上，自适应邻居构图通过设计一个优化问题并求解，使生成的图是根据数据之间的相似度自动构建的，无需人工干预。

$$d(v_i,v_j)=\| f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2 \tag{3}$$

其中，$\boldsymbol{x}_i$和$f(\cdot)$分别代表了节点的原始特征和为了使用数据的潜在特征而找的映射。本文的目标是通过这个映射找到最优的构图表示，$f(\cdot)$会在3.1.2节讨论。

$$\begin{aligned}&\min \sum_{j=1}^{n}\| f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2 s_{ij}\\&\text{s.t. } \forall i,\sum_{j=1}^{n}s_{ij}=1,s_{ij}\geqslant 0\end{aligned} \tag{4}$$

式(4)中存在能让优化问题收敛至最小值的平凡解，但这些解建立在不切实际的假设上：数据点与自身的距离为0，因此与自身的相似度值为1，而与其他数据点的相似度为0。这种情况不符合实际数据分布的特性，缺乏泛化性。因此，需要更通用和现实的相似度度量方法，以更好地反映数据之间的真实关系。为赋予式(4)意义，将正则化项引入上述问题中：

$$\min \sum_{j=1}^{n}\| f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2 s_{ij}+\gamma_i\| s_{ij}\|_2^2 \tag{5}$$

其中，γ_i扮演了一个权衡参数的角色，其调整会对节点$\boldsymbol{x}_i$的相似度向量的稀疏性产生影响。它能够调控相似度向量中非零元素的数量，即调节节点的邻居数量。通过合理设置γ_i，可以有效地控制相似度向量，确保在相似度图中，正常数据之间的相似节点具有较高的权重，而远离正常数据的离群值节点与其他节点具有较低的权重。这种调整不仅可以解决前述的平凡解问题，还有助于管理相似度分布的特性，使其更符合具体需求。将此理念应用于每个节点的邻居分配中，以获得相应的相似性表示。因此，优化的目标不再仅仅是构造邻接关系，而是要找到一种合理的方式，为所有节点分配最合适的邻居，从而提升整体表示能力。

$$\min \sum_{i=1}^{n}\sum_{j=1}^{n}\| f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2 s_{ij}+\gamma_i\| s_{ij}\|_2^2 \tag{6}$$

Nie[27]已经证明了当γ_i被设置为以下值时，优化问题(6)中的s_{ij}有精确的k个非零值。

$$\gamma_i=\frac{k}{2}d_{i,k+1}^x-\frac{1}{2}\sum_{c=1}^{k}d_{i,c}^x \tag{7}$$

其中，$d_{i,k+1}^x$表示将节点i和其他节点的距离从小到大排序后第$k+1$小的距离。当这样设置γ_i时，利用式(8)可以得到s_{ij}的精确解。

$$s_{ij}=\frac{\max(0,(d_{i,k+1}^x-d_{i,j}^x))}{d_{i,k+1}^x-\sum_{v=1}^{k}d_{i,v}^x},\forall i,j=1,2,\cdots,n \tag{8}$$

这种方法相较于直接调整γ，更容易通过离散值k来调整邻域的数量。

3.1.2 结构自编码器

在本节中，采用GCN作为映射函数$f(\cdot)$。鉴于GCN仅适用于无向图，需要根据$\boldsymbol{A}=(\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}})/2$将已构建的有向图转换为无向图。

$$\boldsymbol{H}^{l+1}=f_{\mathrm{relu}}(\boldsymbol{H}^l,\boldsymbol{A}|\boldsymbol{W}^l) \tag{9}$$

其中，$\boldsymbol{H}^l$是卷积层l的输入，$\boldsymbol{H}^{l+1}$是卷积层l的输出。把属性矩阵$\boldsymbol{X}\in\mathbb{R}^{n\times d}$作为第一层的输入，相当于$\boldsymbol{H}^0$。$\boldsymbol{W}^l$是需要在神经网络中学习特定层的可训练权重矩阵。图卷积网络的每一层都可以用函数(10)表示：

$$f(\boldsymbol{H}^l,\boldsymbol{A}|\boldsymbol{W}^l)=\sigma(\overset{\wedge}{\boldsymbol{D}}^{-\frac{1}{2}}\overset{\wedge}{\boldsymbol{A}}\overset{\wedge}{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{10}$$

其中，$\overset{\wedge}{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，代表邻接矩阵$\boldsymbol{A}$与自环矩阵$\boldsymbol{I}$相加的结果；$\overset{\wedge}{\boldsymbol{D}}$是节点度矩阵；$\sigma$是激活函数，本文采用ReLU作为激活函数。这个表示形式描述了GCN每一层如何通过邻接矩阵和权重矩阵来更新节点的特征表示，实现对高阶复杂结构信息的挖掘，学习更有判别性的图嵌入表示。

为了解决多层图卷积叠加可能导致的过度平滑问题，我们设计了一个旨在还原节点连接关系的结构解码器。与常见的0-1矩阵内积解码器不同，结构解码器的目标是重建相似矩阵。受到Sigmod函数的启发，通过节点的编码值的负距离来确定其重构权重。这种方法确保了节点间的较大相似度将对应较大的权重，而较小的相似度则对应较小的权重，与构图逻辑相吻合。因此，所设计的结构解码器能够有效地将编码后的节点连通性与图的重建保持一致融合。

$$d_{ij}=\| \boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2 \tag{11}$$

$$\bar{s}_{ij}=\frac{\mathrm{e}^{-d_{ij}}}{\sum_{j=1}^{n}\mathrm{e}^{-d_{ij}}} \tag{12}$$

因为原始图的相似矩阵$\boldsymbol{S}$和经过解码器的重构图$\bar{\boldsymbol{S}}$相似矩阵都是概率矩阵，为了最小化$\boldsymbol{S}$和$\bar{\boldsymbol{S}}$的差异，利用KL散度设计结构重构损失：

$$\mathcal{L}_s=\min KL(\boldsymbol{S},\bar{\boldsymbol{S}}) \tag{13}$$

$$KL(\boldsymbol{S},\bar{\boldsymbol{S}})=\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}\log\frac{1}{\bar{s}_{ij}} \tag{14}$$

3.1.3 自适应更新图

在第3.1.2节的结构编码器中，处理相似矩阵时将其视作固定矩阵进行训练。这种方式意味着在GCN中只能学习到低阶信息，而无法捕获到高阶信息。为了更好地模拟整个数据分布的相关性，在训练过程中需要自适应地更新相似图。具体而言，在训练初期给定了一个较小的稀疏系数k值，随着训练的进行逐渐增大k值。这种策略旨在让模型能够在训练过程中逐步适应数据的特性，并从中学习到更高阶的信息。k在达到设置的最大值时，将停止增长。鉴于离群点检测是二分类的一种极端情况，将$kmax$设置为$\min(n/2,100)$，n为样本的数量，旨在平衡不同规模数据集中的邻居数量和计算效率。步长为$kmax/iter$，$iter$为迭代次数。

3.1.4 属性自编码器

属性自动编码器更侧重于捕获节点的特征信息。因此，它仅将属性矩阵作为输入，通过两层简单的非线性多层感知器(Multilayer Perceptron，MLP)学习节点的非线性映射。这个过程将观测到的节点属性数据映射到潜在的属性嵌入中。

$$\boldsymbol{H}^1=\sigma(\boldsymbol{X}\boldsymbol{W}^0+\boldsymbol{b}^0) \tag{15}$$

$$\boldsymbol{Z}_a=\sigma(\boldsymbol{H}^1\boldsymbol{W}^1+\boldsymbol{b}^1) \tag{16}$$

其中，$\boldsymbol{H}^1$代表隐藏层输出；$\boldsymbol{W}^0$，$\boldsymbol{W}^1$和$\boldsymbol{b}^0$，$\boldsymbol{b}^1$表示对应隐藏层的可训练权重和偏置项。

为了提升模型的综合性和协同建模能力，更有效地捕捉节点的拓扑结构和节点属性，将结构编码器和属性编码器进行了融合。这一融合过程使得模型能够充分利用来自不同信息源的数据，从而提高了模型的解释性和泛化性。

$$\boldsymbol{Z}_f = Fusion(\boldsymbol{Z}_s, \boldsymbol{Z}_a) = \boldsymbol{Z}_s \oplus \boldsymbol{Z}_a \tag{17}$$

这个融合过程可以采用不同的方法，如拼接、加权相加等，本文采用对应矩阵元素相加的融合方法，将节点的拓扑结构和属性信息综合到一起。这种融合方法是通过对结构编码器和属性编码器获得的信息进行元素级相加实现的。这一综合后的表示，将更全面地反映节点的特征。

在属性解码器阶段，将融合后的信息作为属性解码器的输入。由于融合后的信息被视为更丰富的特征表示，因此解码器将这种综合表示映射回原始属性空间，这对下游任务具有重要意义。

$$\boldsymbol{H}^1 = \sigma(\boldsymbol{Z}_f \boldsymbol{W}^0 + \boldsymbol{b}^0) \tag{18}$$

$$\bar{\boldsymbol{X}} = \sigma(\boldsymbol{H}^1 \boldsymbol{W}^1 + \boldsymbol{b}^1) \tag{19}$$

其中，$\boldsymbol{W}^0$，$\boldsymbol{W}^1$和$\boldsymbol{b}^0$，$\boldsymbol{b}^1$表示属性解码器对应隐藏层的可训练权重和偏置项，解码器的最后一层的输出是重构的属性矩阵$\bar{\boldsymbol{X}}$。

$$\mathcal{L}_a = \min \| \boldsymbol{X} - \bar{\boldsymbol{X}} \|_F^2 \tag{20}$$

3.1.5 损失函数和离群点检测

为了同时降低网络结构和节点属性的损失，需要共同学习并最小化重构损失。

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \boldsymbol{s}_{ij}^2 \tag{21}$$

在离群点检测方面，使用重构误差作为评分。通常情况下，重构误差反映了节点的异常程度：

$$\boldsymbol{score} = \mathcal{L}_s + \mathcal{L}_a + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \boldsymbol{s}_{ij}^2 \tag{22}$$

3.2 算法流程

ANGAE 算法流程如算法 1 所示。

算法 1 ANGAE 算法

input：属性矩阵 **X**；初始邻居 k_0，$k = k_0$；迭代次数 iter；最大邻居数 Kmax；学习率 lr t=Kmax/iter

output：异常分数，分类结果

1. 初始化所有权重矩阵{**W**⁰，**W**¹，**b**⁰，**b**¹…}；
2. 根据初始邻居数计算初始相似矩阵 **S**
3. S 矩阵转换为对称矩阵 **A**
4. for i ← 1 to iter do
5. **Za**=σ(**W** * **X**+**b**)得到属性编码器的结果 Za
6. **Zs**=**GCN**(**A**，**X**，**W**)得到结构编码器的结果 Zs
7. **Zf**=**Zs**+**Za**
8. 获取重构的 **S** 矩阵，$\bar{\mathbf{s}}_{ij} = \dfrac{e^{-d_{ij}}}{\sum_{j=1}^{n} e^{-d_{ij}}}$
9. 得到最终编码 $\bar{\mathbf{X}} = \sigma(\mathbf{W} * \mathbf{Zf} + \mathbf{b})$
10. 获取重构的 **S** 矩阵，$\bar{\mathbf{s}}_{ij} = \dfrac{e^{-d_{ij}}}{\sum_{j=1}^{n} e^{-d_{ij}}}$
11. 计算 loss
12. if(k+t<Kmax)
13. k=k+t
14. 根据现在邻居数计算新的相似矩阵 **S**
15. S 矩阵转换为对称矩阵 **A**
16. end for
17. 得到离群分数，根据阈值得到离群点

4 实验

本章详细介绍了实验设置，包括所使用的数据集、基线方法以及评估指标。为了验证 ANGAE 算法的有效性，在 11 个数据集上进行了大量的实验，并将实验结果与 8 种对比算法进行了全面比较。通过对实验结果进行降维可视化，深入分析了实验结果。除此之外，还设计了消融实验来证实所提方法采用的各个模块和策略的有效性。这些实验的全面性和多角度分析进一步加强了我们对 ANGAE 方法的认识。

本实验的运行环境是 Windows 11，NVIDIA GeForce RTX 3070 和 Python 3.8，采用的优化器是 Adam。为了验证自适应邻居构图策略的有效性，结构编码器和属性编码器中的神经网络结构均仅含一层隐藏层，学习能力相对有限。在实验过程中，将迭代次数设置为 20～50，学习率范围设定为 0.005～0.01，初始邻居数设置为 2。为了避免构建图时出现过拟合的情况，当节点的当前邻居数达到设定的最大值时，将不再更新该节点的邻居数。

4.1 数据集与实验设置

为了展示算法的鲁棒性，本文特意选取了在数据分布、离群点占比和特征维度等方面存在差异的 11 个数据集。这些数据集来自多个领域和应用场景，主要涵盖医学、工业、生物学等领域。表 2 列出了这些数据集的基本统计数据，包括其样本数、维度、离群点个数、领域和离群率。CWRU[28]是由美国凯斯西储大学机械工程学院维护的一个广泛用于机械故障诊断和振动分析的公开数据集。本实验选取其中轴承型号为 SKF 6250、故障位置位于驱动端外圈的数据子集。其余数据集均源自近期的异常检测基准研究 ADBench[29]。

表 2 数据集概况

Table 2 Overview of datasets

数据集	样本数	维度	离群点	领域	离群率/%
Annthyroid	7200	6	534	医学	7.40
Vertebral	240	6	30	生物学	12.50
Pima	768	8	268	医学	34.80
Breastw	683	9	239	医学	34.90
Waveform	3443	21	100	通信	2.90
WPBC	198	33	47	医学	23.74
satellite	6435	36	2036	遥感	31.64
CWRU	680	23	600	工业	11.76
WDBC	367	30	10	医学	2.70
Lympho	148	18	6	医学	4.10
GLASS	214	9	9	材料科学	4.20

4.2 评估指标

在离群点检测中，数据不平衡是常见的问题。为了降低这个问题对实验结果的影响，采用 3 个关键性能指标来评估不同的检测算法，分别是 F1 分数（F1-Score）、ROC-AUC（ROC 曲线下的面积）和准确度（ACC）。F1 分数是一个综合衡量模型性能的指标，它结合了精确率和召回率两个重要的评估指标。精确率衡量了模型预测为正类的样本中实际为正类的比例，而召回率则衡量了模型成功找出的正类样本的

比例。F1 分数是这两者的调和平均数，其值越接近 1 表明模型在平衡查准率和查全率方面表现更好。ROC-AUC 是一种衡量分类器对正类和负类分类能力的指标，在处理样本不平衡时表现出比较好的鲁棒性，代表真正例率与假阳性率之间的平衡关系。ACC 是指模型正确预测的样本数量与总样本数量之比，通常用来衡量模型在整体上的预测准确程度。

4.3 对比实验

为了加强实验的可信度，在选择对比算法时，考虑了 3 种基本基线类型。第一类是传统的经典检测方法，包括 LOF，IForest，KNN 和 OCSVM。第二类是基于普通神经网络的方法，包括 AE 和 SO-GAAL。第三类是基于图神经网络的模型，包括 LUNAR[30] 和 GUIDE[31]。由于 ANGAE 结合了个体与邻居之间的关系及神经网络的特点，因此在实验中，需确保其参数设置与相关神经网络算法一致，同时最近邻的设置方式也应与其他需计算最近邻关系的算法相同。表 3 列出了实验中使用的各算法参数的设置。

1）LOF：通过计算节点属性与其邻居之间的局部密度偏差来进行离群值检测。

2）IForest：一种基于树结构的异常检测方法，通过构建树型结构来识别数据中的离群点。

3）KNN：通过观察数据点与其邻居之间的距离关系，KNN 可以识别那些远离其最近邻居的数据点并将其作为潜在的离群点。

4）OCSVM：通过学习正常数据的决策边界来有效识别离群点。

5）AE：通过训练模型来还原正常数据，然后通过比较原始数据与重构数据之间的差异来识别异常。

6）SO-GAAL：通过训练一个生成模型来捕捉正常数据的分布，并通过对抗学习的方式来尝试区分正常数据和异常数据。

7）LUNAR：引入图神经网络，将现有局部异常检测算法与图模型进行结合，从而得到一个统一的框架。

8）GUIDE：通过挖掘网络中的高阶结构模式（如子图、社区）并结合节点属性信息，识别结构和属性显著偏离常规模式的异常节点。

表 3 参数设置

Table 3 Parameter settings

算法	最近邻个数 k	学习率 lr	迭代次数 $iter$	网络层数	隔离树的数量和子样本大小	核函数	λ
ANGAE	(2,min(n/2,100))	0.005～0.01	20～50	3	—	—	0.001～10
AE	—	0.005～0.01	20～50	3	—	—	—
LOF	(2,min(n/2,100))	—	—	—	—	—	—
KNN	(2,min(n/2,100))	—	—	—	—	—	—
LUNAR	(2,min(n/2,100))	0.005～0.1	20～50	3	—	—	—
SO-GAAL	—	0.005～1	20～50	3	—	—	—
IForest	—	—	—	—	100,56～256	—	—
OCSVM	—	—		—	—	rbf	—
GUIDE	—	0.005～0.1	20～50	3	—	—	—

对于所有的对比算法，本文在已经集成这些算法的开源 Python 工具箱 pyod[32] 和 pygod[33] 中进行。

4.4 实验结果

表 4—表 6 详细列出了 ANGAE 与前文提及的其他 8 种模型在 11 个数据集上的性能指标，包括 F1-Score，AUC 和 ACC。为了直观呈现实验结果，标记了最优和次优结果。研究结果表明，ANGAE 在大多数数据集上表现优异。对于代表模型在不平衡数据集上性能的最佳指标 F1 分数，ANGAE 在 8 个数据集（Vertebral，Lympho，Pima，WDBC，Waveform，WPBC，satellite，CWRU）上获得了最高排名。对于 AUC 分数，ANGAE 在 8 个数据集上均获得了最高分数，意味着模型具有较强的区分能力。在 ACC 分数中，ANGAE 在 8 个数据集上取得了最好的效果，反映了 ANGAE 在辨别离群点和正常数据方面的可靠性和稳定性。

表 4 F1-Score 结果对比

Table 4 Comparison of F1-Score

数据集	LOF	IForest	KNN	OCSVM	SO-GAAL	LUNAR	GUIDE	AE	Ours
Annthyroid	0.20	**0.40**	0.25	0.14	0.15	0.24	0.19	0.20	**0.28**
Vertebral	0.24	0.17	0.15	0.23	0.23	0.15	**0.26**	0.19	**0.33**
Lympho	0.14	0.62	0.64	0.38	0.07	**0.65**	0.59	0.56	**0.67**
GLASS	0.25	0.19	**0.27**	0.07	0.24	**0.27**	0.18	0.12	**0.25**
Pima	0.48	0.57	**0.59**	0.49	0.36	**0.59**	0.55	0.47	**0.63**
WDBC	0.67	0.75	0.55	0.58	0.36	**0.60**	0.71	0.79	**0.90**
Breastw	0.53	0.91	0.91	0.69	0.83	**0.95**	0.87	0.87	**0.92**
Waveform	**0.20**	0.13	0.13	0.06	0.06	0.13	0.17	0.18	**0.32**
WPBC	0.54	0.57	0.54	0.49	**0.66**	0.55	0.68	0.61	**0.73**
satellite	**0.58**	0.54	0.57	0.44	0.45	0.48	0.51	0.52	**0.62**
CWRU	**0.31**	0.09	0.10	0.12	0.21	0.09	**0.27**	0.18	**0.31**

表 5 AUC 结果对比

Table 5 Comparison of AUC

数据集	LOF	IForest	KNN	OCSVM	SO-GAAL	LUNAR	GUIDE	AE	Ours
Annthyroid	0.67	**0.83**	0.69	0.53	0.56	**0.75**	0.72	0.63	**0.75**
Vertebral	0.53	0.45	0.35	0.54	0.58	0.40	**0.61**	0.50	**0.72**
Lympho	0.68	**0.99**	0.98	0.90	0.53	**0.99**	0.89	0.99	**1.00**
GLASS	0.84	0.84	**0.87**	0.44	0.83	**0.87**	0.82	0.74	**0.89**
Pima	0.59	0.71	**0.73**	0.60	0.44	**0.73**	0.63	0.58	**0.76**
WDBC	0.90	0.98	0.98	0.94	0.97	**0.99**	0.87	0.90	**1.00**
Breastw	0.52	0.95	**0.99**	0.80	0.88	0.96	0.89	0.80	**0.98**
Waveform	0.72	**0.79**	0.78	0.54	0.55	0.78	0.65	0.73	**0.80**
WPBC	0.54	0.55	0.51	0.47	**0.69**	0.55	0.54	0.46	**0.67**
satellite	0.68	0.68	**0.74**	0.61	0.69	0.67	0.67	0.69	**0.75**
CWRU	**0.80**	0.75	0.74	0.48	0.67	0.46	0.74	0.61	**0.84**

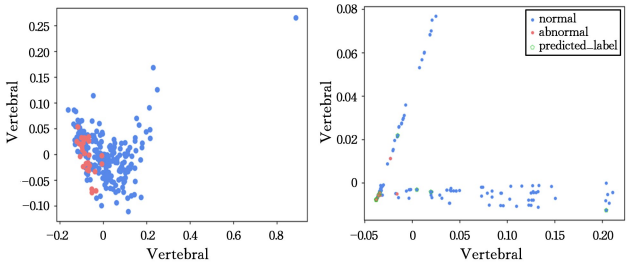
表 6 ACC 结果对比

Table 6 Comparison of ACC

数据集	LOF	IForest	KNN	OCSVM	SO-GAAL	LUNAR	GUIDE	AE	Ours
Annthyroid	0.68	0.73	0.70	0.53	0.61	**0.80**	0.69	0.65	**0.81**
Vertebral	**0.60**	0.48	0.42	0.55	**0.60**	0.47	0.55	0.51	**0.69**
Lympho	0.68	**0.95**	0.97	0.91	0.59	**0.99**	0.94	**0.99**	**0.95**
GLASS	0.81	0.76	0.83	0.52	0.87	**0.91**	0.85	0.77	**0.93**
Pima	0.62	0.68	0.69	0.60	0.72	**0.78**	0.70	0.67	**0.81**
WDBC	0.97	**0.98**	0.96	0.96	0.97	0.96	0.95	0.93	**1.00**
Breastw	0.66	0.97	0.96	0.77	0.91	**0.98**	0.87	0.88	**0.99**
Waveform	0.71	0.72	0.73	0.54	0.67	**0.79**	0.54	0.68	**0.77**
WPBC	0.51	0.57	0.51	0.47	**0.71**	0.64	0.58	0.60	**0.67**
satellite	0.57	0.70	**0.74**	0.61	0.75	0.71	0.65	0.73	**0.79**
CWRU	**0.78**	0.41	0.45	0.51	0.74	0.43	**0.78**	0.61	**0.80**

4.5 实验分析

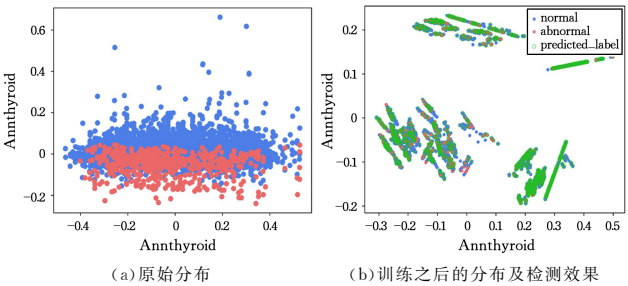
分析实验结果发现，ANGAE 在不同的数据集上表现出了不同的性能。为了更好地理解 ANGAE 的原理，采用主成分分析(Principal Component Analysis，PCA)降维技术对数据集进行针对性分析。图 2－图 12 展示了数据集的原始分布和经过 ANGAE 模型训练之后的数据分布，以及模型的检测效果。在给出的结果图中，蓝色的点代表正常样本，红色的点代表离群点，绿色菱形框代表标记的离群点。



(a)原始分布 (b)训练之后的分布及检测效果

图 2 Vertebral 数据集可视化(电子版为彩图)

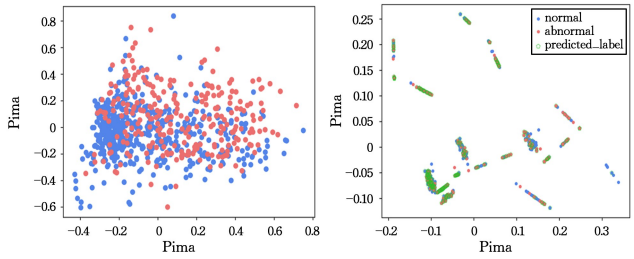
Fig. 2 Visualization of Vertebral



(a)原始分布 (b)训练之后的分布及检测效果

图 3 Annthyroid 数据集可视化(电子版为彩图)

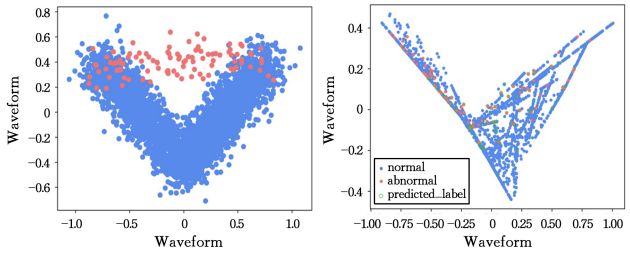
Fig. 3 Visualization of Annthyroid



(a)原始分布 (b)训练之后的分布及检测效果

图 4 Pima 数据集可视化(电子版为彩图)

Fig. 4 Visualization of Pima

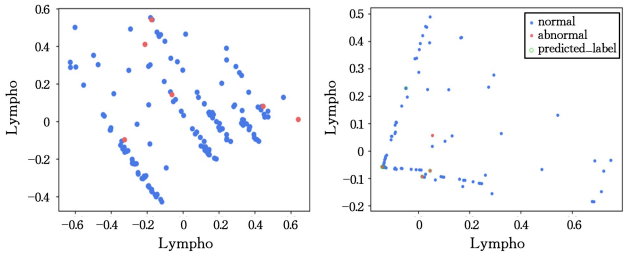


(a)原始分布 (b)训练之后的分布及检测效果

图 5 Waveform 数据集可视化(电子版为彩图)

Fig. 5 Visualization of Waveform

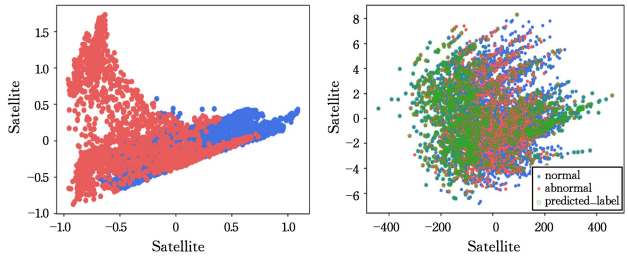
通过对数据进行 PCA 降维发现，Breastw 和 WDBC 这两个数据集的原始分布表现出相似的特性，即正常节点呈现出明显的聚类模式，而离群点则分布在其周围。针对这种正常节点和离群点之间明显的区分特性，传统方法(如基于距离大小或密度稀疏的方法)同样表现出色。在这两个数据集上，ANGAE 算法在 AUC 和 ACC 指标上与传统算法尤其是 KNN 和 IForest 的性能相当。



(a)原始分布 (b)训练之后的分布及检测效果

图 6 Lympho 数据集可视化(电子版为彩图)

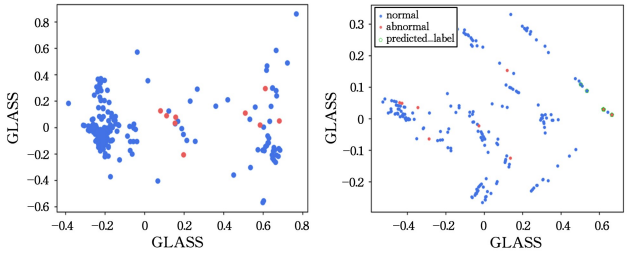
Fig. 6 Visualization of Lympho



(a)原始分布 (b)训练之后的分布及检测效果

图 7 satellite 数据集可视化(电子版为彩图)

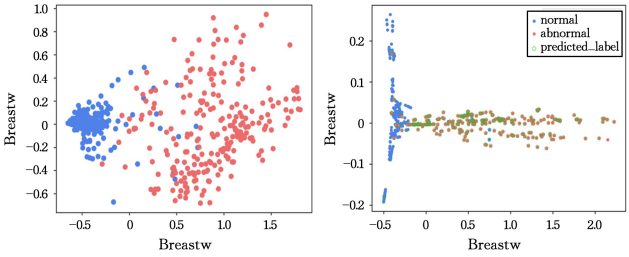
Fig. 7 Visualization of satellite



(a)原始分布 (b)训练之后的分布及检测效果

图 8 GLASS 数据集可视化(电子版为彩图)

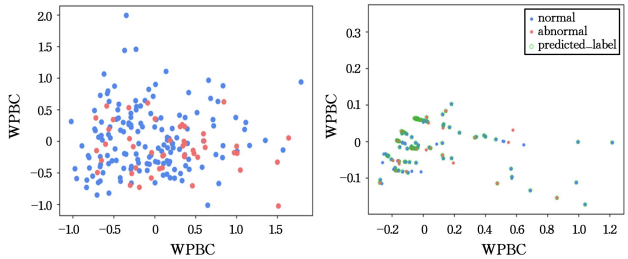
Fig. 8 Visualization of GLASS



(a)原始分布 (b)训练之后的分布及检测效果

图 9 Breastw 数据集可视化(电子版为彩图)

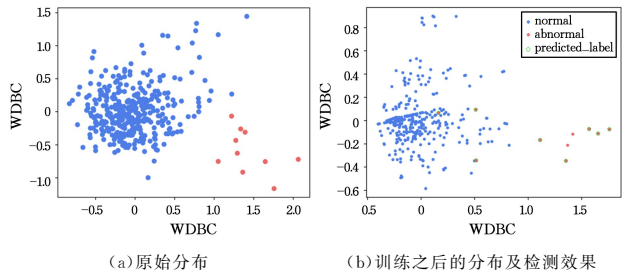
Fig. 9 Visualization of Breastw



(a)原始分布 (b)训练之后的分布及检测效果

图 10 WPBC 数据集可视化(电子版为彩图)

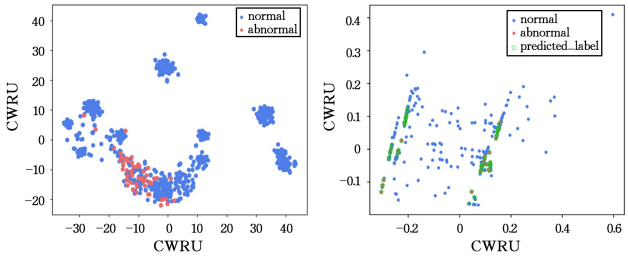
Fig. 10 Visualization of WPBC



(a)原始分布 (b)训练之后的分布及检测效果

图 11 WDBC 数据集可视化(电子版为彩图)

Fig. 11 Visualization of WDBC



(a)原始分布 (b)训练之后的分布及检测效果

图 12 CWRU 数据集可视化(电子版为彩图)

Fig. 12 Visualization of CWRU

对于 Vertebral, Lympho, Pima, GLASS, Waveform, satellite, CWRU, Annthyroid 和 WPBC 这 9 个数据集,原始数据的分布展现出正常数据和离群值部分或者完全相互混合的情况。在缺少标签信息的情况下,很难通过距离、密度或聚类等表面特征对它们进行明确区分。在这 9 个数据集上,ANGAE 表现出色,AUC 分数平均分别超过其他算法 13%,12%,10%,14%,11%,7%,15%,7.5%和 15%,F1 分数平均分别超过其他算法 13%,17%,13%,13%,22%,10%,13%,6%和 16%,这是因为 ANGAE 在处理数据时能够精准地捕捉数据的拓扑结构。通过这种方式,它提升了对数据特征的学习效果,并且能够从数据的整体关系中提取信息,使得模型更有效地理解数据的内在结构,从而优化了对离群点和正常数据的区分能力,这证明了捕获数据拓扑结构对学习数据特征的有效性。

最后,就 F1 分数而言,ANGAE 在 8 个数据集中取得了最高值,这显示出它在识别离群值方面的优秀表现。它有效地处理了数据集中正常点和离群值分布不平衡的情况,能够捕获离群值而不过度产生误报。

4.6 消融实验

本节对 ANGAE 方法进行了消融实验,通过设计 3 种变体方法来验证其中 3 个关键模块的有效性。

1)ANGAE-noAtt:为了探究双自动编码器的优越性,在完整的 ANGAE 模型中,只采用结构编码器而忽略属性编码器,将原本用于捕捉数据特征的双通道设计简化为单通道。

2)KNNGAE:为了探究自适应邻居的图构造方法的优越性,在 ANGAE 的基础上将图构造模块中使用的方法替换为 KNN 方法,同时保持其他模块不变。

3)KNNGAE-noAtt:类似于 ANGAE-noAtt,在 KNNGAE-noAtt 模型中仅使用了结构编码器,忽略了属性编码器,再次探究双自动编码器对模型的影响。

实验结果如图 13 和图 14 所示。首先，从图 13 中 ANGAE 和 ANGAE-noAtt 与 KNNGAE 和 KNNGAE-noAtt 两组实验数据可以看出，在其他条件不变的情况下，双自动编码器模型在大多数情况下的 AUC 值和 F1 分数值高于单通道模型。这主要是因为属性编码器在捕获数据特征和图结构之间的关联性方面发挥了重要的作用，其缺失会导致模型在学习图结构时失去重要的特征信息，从而影响了性能。偶尔出现单通道模型的性能与双通道模型的性能相平或者有些微优势的情况，结合 4.5 节对数据原始分布的分析，我们认为出现这种情况的原因是正常数据与离散点分布比较分散，属性编码器的信息并未对任务产生显著帮助，仅仅是单通道模型能够更好地学习到数据的关键特征。然而，这种现象可能是特例，通常双通道设计更能全面地考虑图的结构和属性信息，从而提高模型性能。

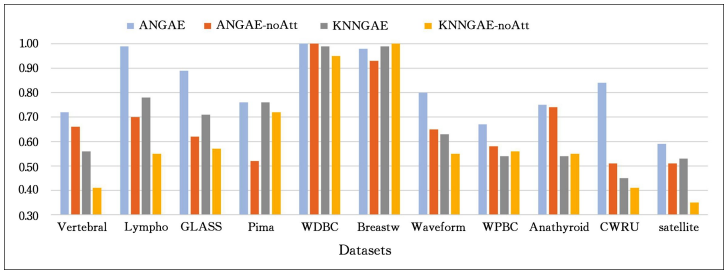


图 13 消融实验的 AUC

Fig. 13 AUC of ablation experiments

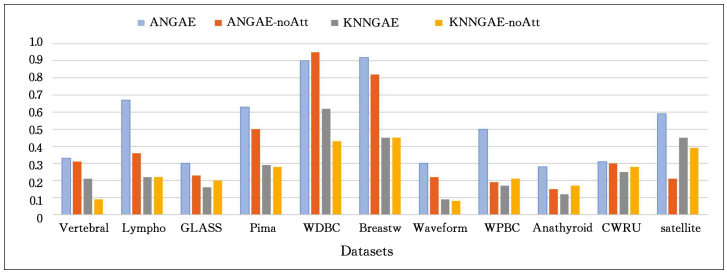


图 14 消融实验的 F1

Fig. 14 F1 of ablation experiments

其次，将 KNNGAE 的实验结果和 ANGAE 进行对比，所有数据集中的 AUC 值和 F1 分数值都有不同程度的下降。这是 KNN 方法在处理图结构时，未能很好地适应数据的复杂关系，导致了信息的丢失，且在后续的学习过程中没有利用新的特征对已经构建好的结构图进行修正或更新，进而影响了模型在学习任务上的性能表示。

结束语 针对欧氏数据，本文提出了一种名为 ANGAE 的深度联合表示学习方法。为了挖掘潜在信息，首先从图构造的角度建立了点之间的加权相似图，利用 GCN 的多层非线性变换能力，实现对非图数据的图构建和潜在信息的挖掘。由于 ANGAE 中的图结构是人为构建的，为了保持其准确性，采用自适应的邻居构图策略，以确保在后续学习中对已有不准确图结构的修正和更新。此外，为了提升模型的学习能力，引入了双自动编码器，分别从属性和结构两个方面检测离群点。最后，通过节点的属性和结构损失计算节点的离群分数，并进行排序，以寻找离群点。在 11 个公开数据集上验证了 ANGAE 方法的有效性。然而，本方法依然存在一些不足之处。双自动编码器的引入不仅增加了模型的复杂性和计算开销，也带来了过拟合问题。虽然可以通过调整属性编码器和结构编码器的重要程度来缓解这一问题，但如何有效地找到最佳平衡点仍然是一个挑战。此外，在实际应用场景中，数据可能是动态变化的，ANGAE 需要重新训练模型来适应新的数据，这对于在线数据处理和实时离群点检测可能不太适用。因此，未来需要解决的问题将集中于如何提高模型的可扩展性和实用性。

参 考 文 献

[1] PANG G, SHEN C, CAO L, et al. Deep Learning for Anomaly Detection: A Review [J]. ACM Computing Surveys, 2021, 54(2): 38: 1-38: 38.

[2] BAO Y, KE B, LI B, et al. Detecting Accounting Fraud in Publicly Traded U. S. Firms Using a Machine Learning Approach [J]. Journal of Accounting Research, 2020, 58(1): 199-235.

[3] AL-HASHEDI K G, MAGALINGAM P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019 [J]. Computer Science Review, 2021, 40: 100402.

[4] SAHOO S R, GUPTA B B. Multiple features based approach for automatic fake news detection on social networks using deep learning [J]. Applied Soft Computing, 2021, 100: 106983.

[5] ZHANG X, GHORBANI A A. An overview of online fake news: Characterization, detection, and discussion [J]. Information Processing & Management, 2020, 57(2): 102025.

[6] SAFIAN A, WU N, LIANG X. Development of an embedded piezoelectric transducer for bearing fault detection [J]. Mechanical Systems and Signal Processing, 2023, 188: 109987.

[7] YAKHNI M F, CAUET S, SAKOUT A, et al. Variable speed induction motors' fault detection based on transient motor current signatures analysis: A review [J]. Mechanical Systems and Signal Processing, 2023, 184: 109737.

[8] LI C T, TSAI Y C, CHEN C Y, et al. Graph Neural Networks

for Tabular Data Learning: A Survey with Taxonomy and Directions[J]. arXiv:2401.02143,2024.
[9] YANG X,LATECKI L J,POKRAJAC D. Outlier Detection with Globally Optimal Exemplar-Based GMM[M]// Proceedings of the 2009 SIAM International Conference on Data Mining(SDM). Society for Industrial and Applied Mathematics, 2009:145-154.
[10] BREUNIG M M,KRIEGEL H P,NG R T,et al. LOF:identifying density-based local outliers[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2000: 93-104.
[11] JIANG S Y,AN Q B. Clustering-Based Outlier Detection Method[C]// 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. 2008:429-433.
[12] PAPADIMITRIOU S,KITAGAWA H,GIBBONS P B,et al. LOCI:fast outlier detection using the local correlation integral [C]// Proceedings 19th International Conference on Data Engineering(Cat. No. 03CH37405). 2003:315-326.
[13] IKOTUN A M,EZUGWU A E,ABUALIGAH L,et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences,2023,622:178-210.
[14] DENG D. DBSCAN Clustering Algorithm Based on Density [C]// 2020 7th International Forum on Electrical Engineering and Automation(IFEEA). 2020:949-953.
[15] CERVANTES J,GARCIA-LAMONT F,RODRÍGUEZ-MAZAHUA L,et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends[J]. Neurocomputing,2020,408:189-215.
[16] LIU F T,TING K M,ZHOU Z H. Isolation Forest[C]// 2008 Eighth IEEE International Conference on Data Mining. 2008: 413-422.
[17] PANG G,CAO L,AGGARWAL C. Deep Learning for Anomaly Detection: Challenges, Methods, and Opportunities[C]// Proceedings of the 14th ACM International Conference on Web Search and Data Mining. New York: Association for Computing Machinery,2021:1127-1130.
[18] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks[J]. Communications of the ACM, 2020,63(11):139-144.
[19] GIRIN L,LEGLAIVE S,BIE X,et al. Dynamical Variational Autoencoders: A Comprehensive Review[J]. Foundations and Trends © in Machine Learning,2021,15(1/2):1-175.
[20] LIU Y,LI Z,ZHOU C,et al. Generative Adversarial Active Learning for Unsupervised Outlier Detection[J]. IEEE Transactions on Knowledge and Data Engineering,2020,32(8):1517-1528.
[21] DU X,CHEN J,YU J,et al. Generative adversarial nets for unsupervised outlier detection[J]. Expert Systems with Applications,2024,236:121161.
[22] WU Z,PAN S,CHEN F,et al. A Comprehensive Survey on Graph Neural Networks[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,32(1):4-24.
[23] KHAN W,AL E. An Exhaustive Review on State-of-the-art Techniques for Anomaly Detection on Attributed Networks[J]. Turkish Journal of Computer and Mathematics Education,2021, 12(10):6707-6722.
[24] DING K,LI J,LIU H. Interactive Anomaly Detection on Attributed Networks[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. New York: Association for Computing Machinery,2019:357-365.
[25] DING K,LI J,BHANUSHALI R,et al. Deep Anomaly Detection on Attributed Networks[C]// Proceedings of the 2019 SIAM International Conference on Data Mining(SDM). Society for Industrial and Applied Mathematics,2019:594-602.
[26] LI Y,HUANG X,LI J,et al. SpecAE:Spectral AutoEncoder for Anomaly Detection in Attributed Networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: Association for Computing Machinery,2019:2233-2236.
[27] NIE F,WANG X,HUANG H. Clustering and projected clustering with adaptive neighbors[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Association for Computing Machinery, 2014:977-986.
[28] LONSO-GONZÁLEZ M,DÍAZ V G,PÉREZ B L,et al. Bearing Fault Diagnosis With Envelope Analysis and Machine Learning Approaches Using CWRU Dataset[J]. IEEE Access,2023,11: 57796-57805.
[29] AN S,HU X,HUANG H,et al. ADBench: Anomaly Detection Benchmark[J]. Advances in Neural Information Processing Systems,2022,35:32142-32159.
[30] ODGE A,HOOI B,NG S K,et al. LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2022, 36(6):6737-6745.
[31] YUAN X,ZHOU N,YU S,et al. Higher-order Structure Based Anomaly Detection on Attributed Networks[C]// 2021 IEEE International Conference on Big Data(Big Data). 2021: 2691-2700.
[32] ZHAO Y,NASRULLAH Z,LI Z. PyOD: A Python Toolbox for Scalable Outlier Detection[J]. Journal of Machine Learning Research,2019,20(96):1-7.
[33] LIU K,DOU Y,DING X,et al. PyGOD: A Python Library for Graph Outlier Detection[J]. Journal of Machine Learning Research,2024,25(141):1-9.

TAN Qiyin, born in 2000, postgraduate. Her main research interests include machine learning and anomaly detection.

YU Jiong, born in 1965, Ph. D, professor. His main research interests include distributed computing, machine learning and data mining.

(责任编辑:柯颖)