摘要

药物开发周期漫长且耗资巨大，使用计算机方法筛选先导化合物的方式可有效提升其效率。其中，使用定量构效关系建模方法来预测分子的生物活性是药物开发领域中一大研究热点，提升其预测准确率可大大加快药物研发的速度。但由于输入方式及计算能力的限制，早期使用的机器学习建模方法已不能满足药物大数据时代的需求。

本文的研究目的是构建较为可靠的分子的生物活性预测模型，目的是通过直接学习分子图特征，避免人工计算特征带来的不稳定性及不可靠性。然而在建模过程中，依然存在特征融合时无法自适应学习特征权重、数据中难易样本不均衡的问题，从而致使模型整体性能受到影响。故本文做了以下几点研究：

1）选取不同类型的数据集

本文所选用的数据集来自于公共化学数据库 PubChem。并且使用文献中的多种筛选方法对靶标等内容作出限制，选择了不同类型的几种生物活性数据集。包括 1851 靶标家族中细胞色素酶 P450 系列的 4 个数据集、两种抑制剂活性数据集和识别结合 r(CAG) RNA 重复序列的分子集。

2）研究基于边注意的图卷积网络

本文将一种基于边注意力的图卷积网络架构，应用于本文选用的生物活性数据集，直接学习分子图，从而避免人工特征工程带来的误差。并对比几种机器学习基准算法，验证了算法有效性。

3）研究特征融合方式，提出多特性融合方案

针对前人提出的模型中存在的问题：无法自适应学习边属性特征权重，本文提出了分子多特性融合的方案优化了算法模型的特征提取能力，通过注意力机制针对多个特征进行自适应融合，有效解决了这一问题，并获得了更好的预测性能。

4）研究样本不均衡问题，提出损失优化方案

针对分子生物活性数据存在的正负样本及难易样本不均衡问题，通过改进本文算法中的损失计算方案：引入了目前较优的两种损失修改方案：聚焦损失（Focal Loss）以及梯度均衡机制（Gradient Harmonizing Mechanism，GHM），实现了模型性能的进一步优化。

关键词：生物活性预测；图卷积；注意力机制；样本类别不平衡

Abstract

The drug development cycle is long and costly, and the use of computer methods to screen lead compounds can effectively improve its efficiency. Among them, using the Quantitative structure-activity relationship modeling methods to predict the biological activity of molecules is a major research hotspot in the field of drug development. Improving its prediction accuracy can greatly accelerate the speed of drug development. However, due to the limitation of input methods and computing power, the machine learning modeling methods used in the early days can no longer meet the needs of the drug big data era.

The purpose of this paper is to construct a more reliable prediction model of molecular biological activity, and to avoid the instability and unreliability caused by artificial computing characteristics by directly learning molecular graph characteristics. However, in the process of modeling, there are still problems such as the inability to adaptive learn feature weights during feature fusion and the imbalance of difficult and easy samples in the data, thus affecting the overall performance of the model. Therefore, this paper has done the following research:

1) Select different types of data sets

The data set selected in this paper is from PubChem, a public chemical database. In addition, a variety of screening methods in the literature were used to limit the contents such as targets, and several bioactivity data sets of different types were selected. Four data sets of cytochrome P450 series from the 1851 target family, two inhibitor activity data sets, and a molecular set for the recognition of the binding R(CAG) RNA repeat sequence are included.

2) Research on Graph Convolutional Networks based on edge attention

In this paper, a Graph Convolutional Networks architecture based on edge attention is applied to the biological activity data set selected in this paper to learn molecular graph directly, so as to avoid errors caused by artificial feature engineering. The effectiveness of the algorithm is verified by comparing several machine learning benchmark algorithms.

3) Research the feature fusion method and propose a multi-feature fusion scheme

Aimed at the problems existing in the model put forward by the predecessors, unable to adaptive learning edge attributes weights, molecular the scheme optimization of feature fusion is proposed in this paper the algorithm of the model feature extraction ability, through the attention mechanism for multiple characteristics of adaptive fusion, effectively solve the problem, and acquired better prediction performance.

4) Study the problem of sample imbalance and propose a loss optimization plan

In view of the imbalance of positive and negative samples and difficult samples in molecular biological activity data, the Loss calculation scheme in this paper was improved: two excellent loss modification schemes, namely Focal Loss and Gradient Harmonizing Mechanism (GHM), were introduced to further optimize the model performance.

Keywords: bioactivity prediction; Graph Convolution; Attention mechanism; the class imbalance of samples

目录

1 绪论

1.1 研究背景及意义

药物开发的过程可看作是迭代筛选，即研究者们根据化合物的分子结构与其分子性质之间存在的某种特定关系，经过层层步骤筛选出某些特定化合物。药物开发周期漫长且耗资巨大，且流失率较高[1]。临床研究中出现的药物骤减导致大量资源被损耗[2]。目前，每 10 个候选药物中就有 9 个在 I 期临床试验或监管批准时失败[1]。

除了化合物分子层面的药物发现，药物本身的研究同样重要。在 2020 年新冠肺炎疫情全球性爆发的特殊状况下，中医药抗疫的成效极其显著，再一次向世界证明了中药的优势[3]。此种情况下，加快药物研发效率就变得尤为重要。基于此，将计算机技术应用于药材功效预测研究也是一大突破口。在本文前期研究中[4]，我们利用药材和方剂数据构建了二分网络，通过网络特性分析，提出了一种基于药材网络特性的节点相似度计算方法，该方法意在描述两种药材同时存在于同一方剂中的概率。并将其嵌入于 KMeans 算法来对药材聚类，以此预测药材功效。实验结果也展现了作者改进后的算法的有效性及可用性。这些基于中药的宏观研究为我们后期对药物分子的研究打下了基石。

为改善药物发现过程效率低下的状况，以能够缩短新药研发的周期及提高研发的成功率，药物化学家们提出了定量构效关系（Quantitative Structure - Activity Relationships，QSAR）的概念[5]。QSAR 是对已知先导化合物的一系列衍生物进行定量的生物活性测定，分析衍生物的主要理化参数和生物活性的关系，建立结构与生物活性之间的数学模型，并以这种数学模型来指导药物分子设计[5]。在新药研发过程中，化学家们将 QSAR 建模作为分子生物活性预测的主要建模方法[6]。与传统的体内实验相比，QSAR 建模可以较快的筛选出人们所需的特定活性范围内的化合物，大大降低了药物研发成本，使人们能更快的研制出新药，对药物发展具有非常重要的意义。

早期阶段，机器学习方法成为化学信息学家们较为常用的建模方法。然而传统机器学习算法只能处理固定大小的输入，且特征工程需要人工生成冗长的分子描述符并进行预处理，这都是建模过程中遇到的挑战。自 1950 年以来，人工智能的概念诞生，并在各研究领域得到应用。在当下人工智能时代，深度学习作为

一种机器学习领域解决实际问题时常见的学习框架[7]，现已成功应用于图像处理、语音识别、医学研究、情感分析等各个领域[7]。由于深度学习的应用领域之广及成效极其优秀，后学者们将其用于 QSAR 领域。

本文基于前人工作，考虑到人工筛选特征及无法学习特征重要度带来的不稳定性和不可靠性，提出了基于多特性融合的注意力图卷积模型，并针对建模过程中存在的难易样本不平衡问题提出了优化方案。模型可以通过化合物的原始图形表示来自动学习分子特征，其中的注意力机制使得网络能够聚合原子自身和邻居的特征信息，有效地提升图卷积对分子图结构信息的提取能力；而其中的多特性融合方案是基于自注意力机制的特征融合方式，能够有效地让模型自适应调节多个特征张量的权重分配。

1.2 国内外研究现状

机器学习起始于 1950 年代，于 1990 年代开始蓬勃发展，并且正成为人工智能中最受欢迎的子领域。机器学习技术常分为有监督和无监督两种技术。在前者中，给定输入输出对，将会学习一个能够映射输入到输出的函数，使模型可以预测未来的情况。在后者中，模式是直接从未标记的数据中学习的。对于生物活性预测，通常使用有监督方法。由于传统机器学习方法只能处理固定大小的输入，大多数早期的 QSAR 建模都是针对不同的任务，人工生成相应的分子描述符。常用的分子描述符包括[8]：（1）分子指纹，通过一系列表示特定子结构的二进制数字对分子结构进行编码[8]；（2）一维/二维分子描述符：由统计学家和化学家处理的描述分子物理化学和微分拓扑衍生的描述符[8]。

在过去十年中，深度学习已成为各个领域的主要建模方法，尤其是医学领域。现不仅用于生物活性和物化性质的预测，还用于了药物从头设计、医学图像分析和合成预测等方向。卷积神经网络（Convolutional Neural Networks，CNN）是深度学习中的一种特殊架构，已成功解决了结构化数据（如图像）的问题[9]。但是，当图形具有不规则形状和大小、节点位置没有空间顺序且节点的邻居也与位置有关时，传统卷积神经网络则不能直接应用于图上。针对这种非欧式结构化数据，人们研究提出了图卷积网络（Graph Convolutional Network, GCN），且基于此提出了各种衍生架构。2005 年 Gori 等人[10]提出了第一个图神经网络（Graph Neural Networks, GNN），该架构基于递归神经网络学习了无向图、有向图和循环图的

体系结构。2013 年 Bruna 等人[11]基于频谱图理论提出了图卷积网络。目前，已有其他形式的 GCN，例如图注意网络（Graph Attention Network, GAT）、图自动编码器和时空图卷积等。

近几年，已有很多研究将图卷积应用于分子的生物活性预测。在化学图论中，化合物结构通常表示为氢贫化（省略氢）的分子图，每个化合物都以无向图表示，原子为节点，键为边。原子和键都包含很多属性例如原子类型、键类型等。2016 年 Kearnes 等人[12]利用节点（原子）和边（键）的属性建立图卷积模型。2017 年 Connor 等人[13]创建了原子特征向量和键特征向量，并将二者拼接形成原子键特征向量。2018 年 Pham 等人[14]提出了图记忆网络（GraphMem），这是一种记忆增强的神经网络，该网络可用于处理具有多种键类型的分子图。在这些研究中，都未对节点特征和键属性加以区分，没有关注其内部联系。但事实上，为原子对之间的各种相互作用类型赋予不同权重才是较为准确的方法。

1.3 本文主要研究内容

本文首先通过介绍药物发现研究中遇到的瓶颈解释了 QSAR 技术的意义及重要性，并介绍了机器学习在生物活性预测中的实际应用以及遇到的技术挑战。其次，针对前人工作，一种基于边注意机制的图卷积算法中存在的缺陷，首先提出了多特性融合方案，使模型不仅可以有效地学习分子图中的信息，对每条边进行权重分配，让网络能聚合自身和邻居节点的特征信息；还可以在特征融合时自适应分配特征的权重，使网络能够更充分地利用分子图属性信息提取特征。其次针对数据中存在的样本不均衡问题提出了两种优化方案。具体工作如下：

（1）提出了基于多特性融合的边注意图卷积模型，使用基于注意力机制的多特性融合方案替换 concat 拼接特征的方式，最终通过自适应分配输入特征权重来生成分子图特征，有效的避免了普通特征融合方法欠缺对重要信息关注的问题；同时可以提高模型效率和性能。

（2）针对分子生物活性数据存在的正负样本及难易样本不均衡问题，通过改进本文算法中的损失计算方案：引入了目前较优的两种损失修改方案：聚焦损失（Focal Loss）以及梯度均衡机制（Gradient Harmonizing Mechanism，GHM），实现了模型性能的进一步提升。

1.4 章节安排

本文的组织结构以及各章节的内容如下：

第1章：绪论，阐述了研究生物活性预测的意义及QSAR建模的重要性；然后介绍了深度学习在生物活性预测中的研究现状，最后简要介绍了本文主要研究内容。

第2章：介绍了与本文提出的算法有关的技术理论，包括深度学习理论中的图卷积神经网络、激活函数、数据划分方式及Softmax函数，以及生物活性预测相关理论。

第3章：介绍了本文提出的基于多特性融合的注意力图卷积算法模型（MF-EAGCN），首先将基于边注意的图卷积算法模型（EAGCN）[36]应用于本文的生物活性预测任务，并对比基准算法，验证算法的有效性；然后针对EAGCN中存在的无法自适应设置边属性特征权重致使模型性能下降的问题，提出多特性融合方案。并且经过系列实验，验证了方案的有效性。

第4章：介绍了针对分子的生物活性数据存在的正负样本及难易样本不均衡问题，本文所引入的两种基于损失函数的优化方案：Focal Loss和GHM，实现了模型预测性能的进一步优化。

第5章：综合全文进行总结与展望。针对本文中所提出的生物活性预测模型等研究内容进行了总结概况，并对其优缺点进行了分析，进而指出后期研究可以继续完善的地方。

2 深度学习与生物活性预测相关理论

本章对深度学习（包括图卷积神经网络、激活函数、数据划分方式和 Softmax 函数介绍）以及生物活性预测的相关理论（包括 QSAR 和分子不同表示形式的介绍）做出了详细阐述，为下文模型介绍奠定了理论基础。

2.1 深度学习理论

2.1.1 图卷积神经网络

卷积神经网络已成功解决了结构化数据的问题（如图像处理、自然语言处理等）。但是，当图形具有不规则形状和大小、节点位置没有空间顺序且节点的邻居也与位置有关时，传统卷积神经网络则不能直接应用于图上。以上即为非欧式（non-Euclidean）空间数据的特征，在现实生活中有大量此类数据。例如本文中所使用的化合物分子图数据。

GCN 的想法源于 Bruna 等人于 2013 年[11]引入的频谱图 CNN，他们在研究傅里叶域中的卷积操作时，需要计算拉普拉斯算子的特征分解，这在处理大型图时会产生很大的计算负担。为避免这种高负担计算，Defferrard 等人于 2016 年[15]引入了空间局部滤波器；Kipf 等人[16]利用将滤波器的操作限制于一阶邻居的方式简化了 GCN。以下内容是对 GCN 的详细阐述。

首先定义图信号 x 和滤波 g 的卷积计算公式（公式 2.1）：

$$x *_G g = U g_\theta U^T x \tag{2.1}$$

其中，g_θ 相当于 CNN 中的卷积核，我们可以根据它不同的计算方式，来选择不同的卷积核。当选择多项式滤波方式时，g_θ 的公式如下（公式 2.2）：

$$g_\theta = \sum_{i=0}^{k-1} \theta_k \wedge^k \tag{2.2}$$

公式 2.2 中，θ_k 是一个系数，$\wedge^k$ 是矩阵的特征值组成的对角矩阵。图信号 x 与卷积核进行卷积得到的公式如下（公式 2.3）：

$$y = U g_\theta U^T x = U\left(\sum_{i=0}^{k-1} \theta_k \wedge^k\right) U^T x \tag{2.3}$$

其中 θ_k 是模型自动学习的参数矩阵。为了省去卷积核中大量求特征值、特征向量的计算，我们使用 K 阶切比雪夫多项式来做了一个减少算法复杂度的近似。计算公式如下（公式 2.4）：

$$g_\theta = \sum_{i=0}^{k-1} \theta_k T_k(\tilde{\wedge}) \tag{2.4}$$

其中，$\tilde{\wedge} = \frac{2\wedge}{\lambda_{max}} - I_N$。再按照如下公式展开 K 阶切比雪夫多项式（公式 2.5）：

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \tag{2.5}$$

其中初始项$T_0(x)=1$，$T_1(x)=x$。按照公式 2.4（递归的 K 阶切比雪夫多项式）重新描述图卷积，则得到如下公式（公式 2.6）：

$$y = U\left(\sum_{k=0}^{k-1} \theta_k T_k(\tilde{\wedge})\right)U^T x = \sum_{k=0}^{k-1} \theta_k T_k(\tilde{L})x \tag{2.6}$$

其中，$\tilde{L} = \frac{2L}{\lambda_{max}} - I_N$。在公式 2.6 中，体现了如何降低运算的复杂度，如公式中所示仅需计算拉普拉斯矩阵即可。为方便表示，我们令 $K = \lambda_{max} = 2$，并简化图卷积公式（公式 2.7）：

$$y = \sum_{k=0}^{k-1} \theta_k T_k(\tilde{L})x = \theta_0 x + \theta_1(L - I_N)x \tag{2.7}$$

根据归一化的拉普拉斯矩阵公式（公式 2.8），即

$$L = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = 1 - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{2.8}$$

将公式 2.8 代入公式 2.7 可得，$y = \theta_0 x + \theta_1 D^{-\frac{1}{2}} W D^{-\frac{1}{2}x}$，若$\theta_0 = -\theta_1 = \theta$，那么有 $y = \left(I_N + D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\right)x$。

由以上定义可得，图卷积的计算方式可以定义为以下公式（公式 2.9）：

$$X^{k+1} = \delta\left(A X^k \theta^k\right) \tag{2.9}$$

其中$A = \left(I_N + D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\right)$，$X^k$为第 k 层的输入，维度为 N*C，含义为每层输入有 N 个点，而每个节点的特征维度为 C。θ^k表示第 k 层的滤波参数矩阵，维度大小为 C*F，C 为输入节点的特征维度，可得第 k+1 层的输出维度为 N*F。

在简化公式时，我们会将特征值的范围设定为 0~2。此时还需要考虑另一种情况：当累加多层卷积时，可能会出现梯度消失的现象，因为两个距离最近的矩阵可能会形成数值偏大的特征值，这很大程度上降低了结果的可信度。为解决这一负面影响，引入了重正化，即 $\tilde{A} = \tilde{D}^{-\frac{1}{2}} \tilde{W} \tilde{D}^{-\frac{1}{2}}$，其中，$\tilde{W} = \mu I_N + W$，$\tilde{D}$ 是其度

矩阵，W 是邻接矩阵。I_N是单位矩阵，其中加入了自连接。引入参数μ是为了描述自连接的重要性。那么单个节点 i 的计算公式如公式 2.10 所示：

$$X_i^{k+1} = \delta\left(\sum_{j \in N_i} \frac{1}{d_{ij}} X_i^k \theta^k\right) \tag{2.10}$$

其中，j 为节点 i 的邻居节点，N_i为节点 i 的邻居节点集合，d_{ij}为节点 i 和 j 之间连边的权重。节点 i 的更新方式如图 2-1 所示。每一个节点首先将自己变换后的特征信息发送给邻居节点，然后每个节点再将这些邻居节点的特征信息进行聚合，即进行局部信息融合，最后对聚集后的信息做非线性变换，以提高模型的表达能力。

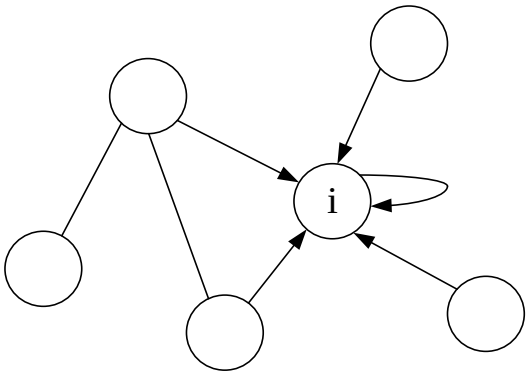


图 2-1 节点更新方式

总的来说，图卷积神经网络是由卷积架构衍生而来，且能够适用于更广泛的场景中。其中，基于谱的图卷积神经网络与卷积的操作很类似，都需要捕捉邻域相关的信息，用邻域信息来更新节点的特征。大量实验表明，使用 GCN 方法对图数据进行预测的工作中，效果明显优于其他方案。

2.1.2 激活函数

激活函数是深度学习中，为了建立一个非线性的映射关系而产生的重要步骤，它可以将线性特征无法描述的问题全部转移到一个非线性区域进行有效的多层叠加。常用的几种激活函数如下：

（1） Sigmoid 函数[17]。Sigmoid 函数也被称为逻辑函数[17]，图像大致趋势为负无穷到正无穷范围内连续单调递增，公式如下（公式 2.11）：

$$g(x) = \frac{1}{1+e^{-x}} \tag{2.11}$$

Sigmoid 函数值域为 0 到 1 之间，且关于点（0，0.5）对称。它能够将输入的实值变换成 0 到 1 范围内的输出，其结果表示某一类别的置信度。Sigmoid 函数常用于二分类问题中，但它存在很多缺点。首先，函数中的幂运算相对复杂，训练过程中会给设备增加压力，在较大的深度网络中表现尤为明显；其次，根据其图像特性，当输入达到一定范围后，函数值增长基本为 0，即梯度消失。若梯度一直未 0，这样会导致在反向传播时，参数无法正常训练，最终导致模型无法正确拟合。

（2） Tanh 函数[18]（双曲正切函数）。其常用于时序网络，表达式为（公式 2.12）：

$$tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.12}$$

双曲正切函数的值域在-1 到 1 之间，且关于点（0，0）对称，双曲正切函数可以一定程度上缓解梯度消失的问题，但并没有完全解决，且幂运算的问题也依然存在。

（3） ReLu 函数[19]（Rectified Linear Unit，修正线性单元）。表达式如下（公式 2.13）：

$$f(x)=\begin{cases}0 & x<0\\ x & x>0\end{cases} \tag{2.13}$$

ReLu 函数有以下优点：1）更有效地缓解梯度消失的问题，至少输入在正区间时，神经元不会饱和；2）ReLu 函数中只存在线性关系，所以在前向传播和反向传播过程中会减少很多的计算量，计算速度比以上两种都快；3）其收敛速度也明显优于逻辑函数和双曲正切函数。但 ReLu 函数最大的缺点是模型训练后期可能会出现神经元死亡的现象：当输入小于 0 时，反向传播的梯度为 0，此时权重无法更新，则神经元死亡。

还有其他优秀的激活函数，包括 Leaky ReLu、RReLu 等。在 Leaky ReLu 函数中，改变了 ReLu 函数中神经元会死亡的现象，并且神经元不会饱和，但是公式中涉及的 α 需要人工利用先验知识来赋值，偶然性较大。RReLu 函数在 Leaky ReLu 函数基础上，将 α 设置为从一个高斯分布中随机出来的值，然后再进行修正。

2.1.3 数据划分方式

数据划分是机器学习建模中必不可少且较为重要的一个环节，它将数据集划分为训练集和测试集。选择一个合适的数据划分方式对模型性能有很大的帮助。这里将介绍几种常用的数据划分方式，包括留出法（Hold-Out）、交叉验证（Cross Validation）和分层交叉验证（Stratified k-fold cross validation）。

1. 留出法（Hold-Out）

留出法直接将数据集按比例随机划分为互斥的两个集合：训练集和测试集。训练集用于模型学习，测试集用于评估模型，作出对模型误差的评分。留出法一般会采用若干次随机划分方式，因其单次结果不够稳定可靠。

这种方法简单快速，但具有较大的偶然性，容易出现局部极值。

2. k 折交叉验证（Cross Validation）

k 折交叉验证将数据集划分为 k 个大小相同的互斥子集 D_i，$i \in \{1,2,\ldots,k\}$。它的特点是每次训练都会进行新一轮的划分，然后在训练时，每次选用一个子集 D_i 作为测试集，另外 k-1 个子集作为训练集，这样每次训练都是在不同数据集上进行的。共训练 k 次，最终返回 k 组测试结果的均值。

此方法通过多次划分多次训练，大大降低了数据划分中的偶然性，提高了模型的泛化能力；并且对数据的使用率也更高。但其也可能出现极端情况，当划分后的某一类全部为一类数据，就可能导致模型学习时没有学习到此类数据的特征，导致预测得分很低。

3. 分层交叉验证（Stratified k-fold cross validation）

分层交叉验证在交叉验证基础上加入了分层思想，使得每一折都保持与原始数据相同的类别比例。它先将数据分类，然后在每个类别中按折数划分。这样就避免了随机划分会产生的情况，使得结果更加可信，但在复杂模型中却大大增加了计算复杂性。

在本文中，为减小因样本划分方式不同而导致的模型性能差别，统一使用 k 折交叉验证方式。这是由于其相较于留出法，降低了数据划分的偶然性，较为稳定可靠性，且相较于分层交叉验证，在深度学习模型中没有极其复杂的计算量，训练效率更高，这对于深度学习来说也是极其重要的。

2.1.4 Softmax 函数

逻辑回归模型是二分类中常用的分类器，当二分类任务推广到多分类任务时，Softmax 函数则起到很大作用[20]。Softmax 函数又称为归一化指数函数[20]，它是一种激活函数，为了将多分类的结果以概率形式输出。

1. 何为 Softmax？

Softmax 基于两个点来进行结果转换：一是概率值为非负数；二是概率值和为 1。于是 Softmax 将原本范围在$(-\infty,+\infty)$的结果按照这两步来转化为概率。Softmax 函数的公式如下（公式 2.14）：

$$\mathrm{Softmax}(z_i)=\frac{e^{z_i}}{\sum_{c=1}^{C}e^{z_c}} \tag{2.14}$$

其中，z_i是第 i 个输出节点的预测值，C 为输出节点的个数。公式可分为两个步骤：1）将结果转化为非负数。使用指数函数$y=exp(x)$，将输出值取值范围映射到$(0,+\infty)$区间。这样保证了概率非负。2）使得结果和为 1。此时再将结果归一化处理，即将 1）中转化后的结果除以所有结果之和。这样就得到了每个输出节点转化后占总数的百分比。

2. 为什么要用 Softmax？

简单来说，Softmax 函数可以将多分类结果映射成区间在（0，1）的值，再进行归一化，即使得所有输出累加等于 1。其实很多函数也可以完成将输出归一化的功能，那为什么会使用 Softmax 呢？归根结底是其使用了指数函数进行结果的映射，这么做可以让结果中大的值更大，小的值更小，增加了输出间的区分对比度，使得模型学习效率更高；其次，我们都知道深度学习的兴起是由于其反向传播机制开始的，通过梯度下降，每次优化一个固定大小的梯度。而深度学习的计算量很大，在这个过程中，研究者们希望尽量简化计算过程，加快学习效率。由于 Softmax 连续可导的特性，深度学习反向传播中的梯度求导过程变得更加方便，因为其消除了拐点，计算简单，在计算损失时计算量小，这些特性在深度学习的梯度下降中非常有必要。在本文中，使用 Softmax 函数将边权重转化为范围在（0，1）且和为 1 的概率值，是为了在权重矩阵生成过程中，使其权重值在分子图的不同边中具有可比性。

2.2 生物活性预测相关理论

2.2.1 定量构效关系

在前期先导药物筛选阶段，需要化合物的生物活性在特定范围内才能够进入下一步。随着计算机技术的逐渐成熟，化学家们不断地将各种建模方法应用于生物活性预测中，这种化学领域的建模方法也被称为定量构效关系（Quantitative Structure - Activity Relationships，QSAR）[21]。QSAR 建模是化学数据分析中公认的较好的计算方法。它是对已知先导化合物的一系列衍生物进行的定量生物活性测定，分析衍生物的主要理化参数和生物活性的关系[21]。QSAR 是基于这样的基本假设：化合物的生物活性依赖于以其核外电子分布、物质的疏水性及其空间构象等共同描述的分子结构[21]。也就是说，它假定了具有相似结构的分子也将会具有类似的活性或性质。因此，通过 QSAR 方法，人们可以建立生物活性与分子结构之间的量化关系，提高药物开发效率及降低成本，这对药物研发具有重要的意义。

一般来说，QSAR 建模的核心步骤包括[22]：1）从数据库或文献等中搜索各种化合物的活性数据（例如毒性、物化性质、甜度等），建立分子数据库；2）用统计的方法构建结构和活性数学关系模型，从中找到结构和性质之间的定量关系；3）利用这一定量关系指导药物分子设计。例如，Sardari 等人[23]使用 Dragon 开源软件生成的近 1400 个分子描述符，使用 K 最近邻算法（KNN）、k-means 聚类算法以及神经网络算法的融合模型进行建模。近几年，深度学习方法如卷积神经网络逐渐被应用于 QSAR 建模。Duvenaud 等人[24]提出了一种使用神经网络生成的指纹描述符来代替传统分子描述符，这种指纹被称为神经指纹。Eguchi 等人提出一种基于分子相似性的图卷积网络模型[25]，以预测生物碱的生物活性。

2.2.2 分子的不同表现形式

分子的生物活性研究中，输入数据是 QSAR 研究的基础，不同的算法模型所使用的分子输入数据形式也是不同的。分子的表示形式常见的有以下几种：分子标识符、分子描述符。

1. 分子标识符

第一种是基于文本的标识符，例如简化分子线性输入规范（Simplified Molecular Input Line Entry System，SMILES）[26]和国际化学标识符(InChl)[27]。

SMILES 是用一组有序规则和专门语法将三维化学结构编码的文本字符串[26]，是一种用于存储化学信息的语言结构。例如二氧化碳（CO_2）的 SMILES 标识符为 O = C = O。SMILES 是目前 QSAR 建模中较常使用的标识符。

国际化学标识符 InChl 用不同的化学信息层（连通性，立体化学，同位素和互变异构体）来表达化学结构[27]。但后期多项研究发现，其复杂的数字公式会导致预测性能下降，因此并未在深度学习中经常使用。

2. 分子描述符

第二种分子描述符是早期 QSAR 研究的基础，传统机器学习模型无法识别及处理分子结构，将分子的物理化学性质或分子结构相关参数，利用各种算法推导出模型可以处理的数值。

目前，用于分子描述符的计算工具有很多种，包括各种开源或商业软件及各种开源库。可以生成的分子描述符已接近 10000 个，包括 1D、2D、3D 描述符以及一些指纹描述符等。近些年，常用的分子描述符计算软件有 Dragon[28]、alvaDesc[29]、Gaussian[30]、Padel-Descriptor[31]、OpenBabel[32]等。其中，经典的 Dragon 软件已迭代到 7.0 版本，可以计算几千种分子描述符，但不幸的是已经停产，进而代替它的是 alvaDesc。alvaDesc 可计算 5305 种分子描述符（包括 Dragon 7 中可用的所有描述符），以及一些特殊描述符例如 MACCS 指纹的计算。常用的化学库有 RDkit[33]等。RDkit 是非常著名的开源化学信息软件包，提供了 Python 和 C++语言的 API 接口，不仅可以计算各种分子描述符，还可以进行分子可视化及化学分析等工作，适用性极好。

2.3 本文选用数据集

本文所选用的数据集来自于一个公共化学数据库 PubChem[34]。我们选用了文献中的多种分析筛选方法[35]，选择了相同类型和不同类型的生物活性数据集，对筛选的靶标等作出了限制，例如筛选了细胞色素 P450 酶的多个系列。最终本文选用了 1851 靶标家族中细胞色素酶 P450 系列的 4 个数据集、两种抑制剂和识别结合 r（CAG） RNA 重复序列的分子系列。表 2-1 列出了所选用的数据集的相关信息以及筛选条件：

表 2-1 本文所使用的来源于 PubChem 数据库的分类数据集信息表

PubChem AID	筛选条件	有活性分子数	无活性分子数
1851(1a2)	Cytochrome P450, family 1, subfamily A, polypeptide 2	5997	7242
1851(2c19)	Cytochrome P450, family 2, subfamily C, polypeptide 19	5905	7522
1851(2d6)	Cytochrome P450, family 2, subfamily D, polypeptide 6,isoform 2	2769	11127
1851(3a4)	Cytochrome P450, family 3, subfamily A, polypeptide 4	5265	7732
492992	Identy inhibitors of the two-pore domain potassium channel (KCNK9)	2097	2820
651739	Inihibition of Trypanosoma cruzi	4051	1326
652065	Identify molecules that bind r(CAG) RNA repeats	2969	1288

2.4 本章小结

本章对深度学习（包括图卷积神经网络、激活函数、数据划分方式和 Softmax 函数介绍）以及生物活性预测的相关理论（包括 QSAR 和分子不同表示形式的介绍）做出了详细阐述。并介绍了本文所选用的 PubChem 数据库中的不同类别数据集。为本文的后续研究奠定了理论基础。

3 基于分子多特性融合的注意力图卷积网络

本文基于前人提出的基于边注意的图卷积神经网络算法[36]（Edge Attention Graph Convolutional Network, EAGCN），针对其存在的问题提出了分子多特性融合的方法优化了算法模型的特征提取能力，通过注意力机制针对多个特征进行自适应融合，并将其应用于第二章提到的多种生物活性数据集，根据最终的衡量指标验证算法的有效性。

3.1 基于边注意的图卷积

基于边注意的图卷积网络（EAGCN）[36]在不同层次和不同边缘属性上学习不同的注意力权重，从而构建一个分子的注意力矩阵。该算法提出了一个边缘注意层来评估分子中每条边的权重：预先构建了一个属性张量，经过注意层处理后，生成多个注意权重张量，其中每个都包含数据集中（分子图）一个边属性的所有可能的注意权重。然后，通过查找该权重张量中分子的每个键的值来构建注意力矩阵。这种方法使得不同分子可对应不同的注意力矩阵。

3.1.1 分子建图方式

在化学图论中，化合物结构通常表示为氢贫化（省略氢）的分子图，每个化合物都可以以一个无向图表示，原子为节点，键为边。其中，分子的属性信息包括原子属性和键属性[35]，具体描述见表 3-1 和表 3-2。这些属性对于描述两个原子之间的键合强度、芳香性或键合共振等特征非常重要。如果将不同的边属性进行注意层处理，则不同的边属性对应于不同的边注意矩阵。

表 3-1 原子（节点）属性表述

原子属性	描述	值类型
原子序号	原子在元素周期表中的位置	Int
相连的原子个数	邻居节点的个数	Int
相邻氢原子个数	氢原子数量	Int
芳香性	是否具有芳香性	Boolean
形式电荷个数	形式电荷个数	Int
环状态	是否在环内	Boolean

表 3-2 键（边）属性表述

键属性	描述	值类型
原子对类型	键连接的原子类型定义	Int
键序	单键/双键/三键/芳香键	Int
芳香性	是否具有芳香性	Int
共轭性	是否共轭	Boolean
环状态	是否在环内	Boolean
占位符	原子之间是否存在键	Boolean

在图 3-1 中，将分子咖啡因（Caffeine）分成若干边色图（以 3 个为例），以 3 种边属性为例：“原子对类型”、“键序”和“环状态（键）”，每个属性表示化合物中原子对之间的特定关系。对于“原子对类型”的边属性：C-N 连边为绿色，C-O 连边为红色，C-C 连边为蓝色。对于“键序”属性：单键为深蓝色，双键为紫色。对于“环状态”属性：处于环中为橙色，否则为青色。对于每个化合物，这些颜色（实际上是权重）都是唯一的。

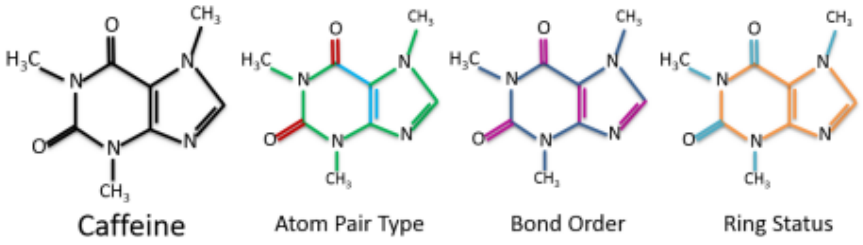


图 3-1 分子 Caffeine 被分成几个不同边颜色的图，图源自论文[36]

基于以上思想，这里提出以下定义：

定义 1 图使用 $G=(V,E)$ 表示，V 为节点的有限集，$|V|=N$，即 N 为节点数，并且 $E\subseteq V\times V$ 是边的有限集合。

定义 2 G 的邻接矩阵 A 是一个方形矩阵，维度为 $N*N$。$a_{ij}=1$ 代表节点 i 和 j 之间有连边，反之 $a_{ij}=0$ 则代表节点间无连边。

定义 3 为 G 构建一个节点特征矩阵 $H^l\in R^N\times R^F$，F 为每个节点的特征总数。第 i 行表示节点 i 的特征和一系列边属性，这里令 K 为边属性个数。

定义 4 假设对于边属性 i，有 d_i 种可能的结果（类型）。

定义 5 为 G 构造一个分子属性张量 $M \in R^{N_{atom}*N_{atom}*N_{features}}$（$N_{features}$ 即为定义 3 中的 F），作为注意层输入。

该算法利用分子的原子和键属性，为每个分子构建 1 个邻接矩阵 A、1 个节点特征矩阵 H^l 和 1 个分子属性张量 M 用于模型训练。

（1） 特征矩阵构造

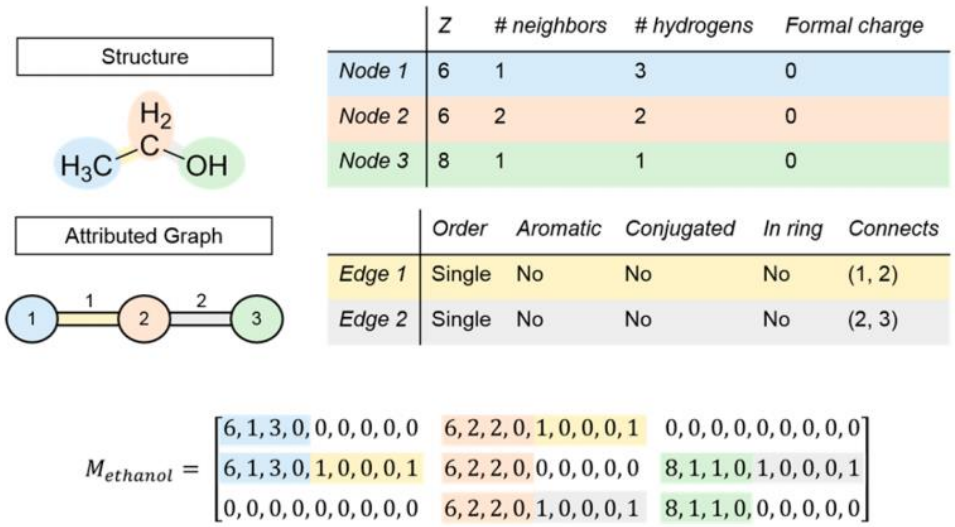
特征矩阵中包括边属性：原子对类型、键类型、芳香性、共轭性、环状态（键）；和节点属性：原子序号、相连原子数、相连氢原子数、芳香性、形式电荷个数、环状态（原子）。其中，某些属性如键类型、原子序号等可能存在多种值（即有两个以上类型）。对于输入的 SMILES 字符串，使用 RDKit 处理，获取到分子的以上属性信息后，将原子属性与键属性线性拼接，最终生成维度为 $N*F$ 的特征矩阵 H^l。

（2） 分子属性张量构造

我们为每个分子，定义一个分子属性张量 $M \in R^{N_{atom}*N_{atom}*N_{features}}$，$N_{features}$ 为原子和键属性的组合个数，即为 F（公式 3.1）。M 三维方向的最后的索引仅表示键存在的标志，因此，子张量片 $M_{i,j,end}$ 即为邻接矩阵 A。

$$M_{ij} = \begin{cases} [F_i, 0] & i = j \\ [F_i, F_{ij}] & a_{ij} = 1 \\ [0,0] & otherwise \end{cases} \tag{3.1}$$

其中，F_i 为原子特征向量，F_{ij} 为键特征向量。这里的键属性向量还加入了一个占位符，以表示节点 i 和 j 之间是否存在边。这里以乙醇为例（图 3-2），选用原子序数，氢原子数和形式电荷作为原子特征，并以键序，芳香性，共轭，环状态和占位符 1 位作为键特征，将原子和键特征根据以上公式线性组合，得到乙醇的分子属性张量 $M_{ethanol}$。



	Z	# neighbors	# hydrogens	Formal charge
Node 1	6	1	3	0
Node 2	6	2	2	0
Node 3	8	1	1	0

	Order	Aromatic	Conjugated	In ring	Connects
Edge 1	Single	No	No	No	(1, 2)
Edge 2	Single	No	No	No	(2, 3)

$$M_{ethanol}=\begin{bmatrix}6,1,3,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 0,0,0,0,0,0,0,0,0\\ 6,1,3,0,1,0,0,0,1 & 6,2,2,0,0,0,0,0,0 & 8,1,1,0,1,0,0,0,1\\ 0,0,0,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 8,1,1,0,0,0,0,0,0\end{bmatrix}$$

图 3-2 乙醇的简单原子和键属性提取（原子和键以不同颜色编码），图源自论文[13]

3.1.2 图注意网络原理

在第二节中，本文介绍了图卷积神经网络的原理及推导，图注意网络也是一种处理图结构数据的模型。GCN 网络可以有效地对节点的一阶邻居进行信息提取，且避免了复杂的矩阵运算。然而，GCN 较为依赖图结构，在一种特定图结构上训练的模型通常不具有普适性，不能直接用于其他图结构中，即不可处理不固定大小的图。针对这一问题，Cucurull 等人[37]提出了可以接受不同图结构的网络——图注意网络（Graph Attention Networks，GAT）。在 GAT 中，可以为图的每个节点分配不同的权值，而此权值是根据其邻居节点的特征得来的。GAT 是一种基于注意力的节点分类网络，可以有效地用于图的归纳学习（Inductive Learning）。归纳学习[38]简单来说就是先从训练数据中学习到一定的知识或规律，然后利用其对测试数据进行预测。这也是机器学习的主要思想。接下来介绍图注意网络的主要架构。

首先以单个图注意力层为例。注意力层的输入是一个节点特征向量集，假设为特征向量集为 $h=\{\vec{h}_1,\vec{h}_2,\ldots,\vec{h}_N\},\vec{h}_i\in R^F$，其中，N 表示节点个数，F 表示节点的特征向量维度。图注意力层所做的事就是经过一系列算法处理得到一个新的节点特征向量集 $h_i'=\{\vec{h}_1',\vec{h}_2',\ldots,\vec{h}_N'\},\vec{h}_i'\in R^{F'}$，这时，$F'$ 已经是新的特征向量的维度，可以与 F 不相等。注意力算法处理过程可以用公式 3.2 表示：

$$e_{ij}=a\left(W\vec{h}_i,W\vec{h}_j\right) \tag{3.2}$$

其中，a 是一个映射关系，也是其中用到的注意力机制，W 是一个权重矩阵。e_{ij} 表示在不考虑图结构性信息的情况下，节点 j 相对于节点 i 的重要性，称为注意力因子。Cucurull 等人[37]通过使用 masked attention 方法引入注意力机制，即仅计算节点 i 的邻节点集N_i的注意力权重，其中 $j \in N_i$。这里的注意力机制 a 经过近几年的发展已存在很多种，Cucurull 等人[37]使用的注意力层是一个单层的前馈神经网络，使用 LeakyReLU 进行非线性激活。为使不同节点的注意力因子有可比性，引入了 softmax 进行正则化（公式 3.3）：

$$a_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_i k)} \tag{3.3}$$

扩展之后得到总的计算公式为（公式 3.4）：

$$a_{ij} = \frac{exp\left(LeakyReLu(\vec{a}^T \left[W\overrightarrow{h_i} \| W\overrightarrow{h_j}\right])\right)}{\sum_{k \in N_i} exp\left(LeakyReLu(\vec{a}^T \left[W\overrightarrow{h_i} \| W\overrightarrow{h_k}\right])\right)} \tag{3.4}$$

其中，$\vec{a}^T$ 为前馈神经网络的参数。再连接图卷积层就可以得到输出特征向量 $\vec{h_i}'$ 了（公式 3.5）：

$$\vec{h_i}' = \sigma\left(\sum_{j \in N_i} a_{ij} W\vec{h}\right) \tag{3.5}$$

GAT 方法具有很多优秀的特性：

（1）一是算法复杂度较低，GAT 中没有特征值分解等复杂的矩阵运算。相对来说较为高效。

（2）二是不同于 GCN 的机制，GAT 给每个节点赋予了不同的重要性，能够更有效率的学习，有更强的表示能力。

（3）三是注意力机制在图中的所有边都是共享的，因此 GAT 类似于 GCN 也是一种局部模型。简单来说就是注意力机制只与相邻节点有关，这样就无需访问整张 graph。根据这一性质，可知 GAT 也可以应用于有向图。

（4）四是具有更好的鲁棒性。GAT 模型建立在相邻节点上，无需假设任何节点顺序，对于扰动更加鲁棒。

3.1.3 模型流程

模型总流程如图 3-3 所示。如图所示，整个模型将分子图作为输入，处理分分子图中的边属性后得到边属性张量，one-hot 编码后分别经过 GAT 层得到五个

图卷积特征，再经过 concat 拼接方式获得总张量特征，以此作为下一层 GAT 层的输入。最后使用两层 dense 层输出结果。

接下来将模型中的注意层和图卷积层分开进行详细介绍。

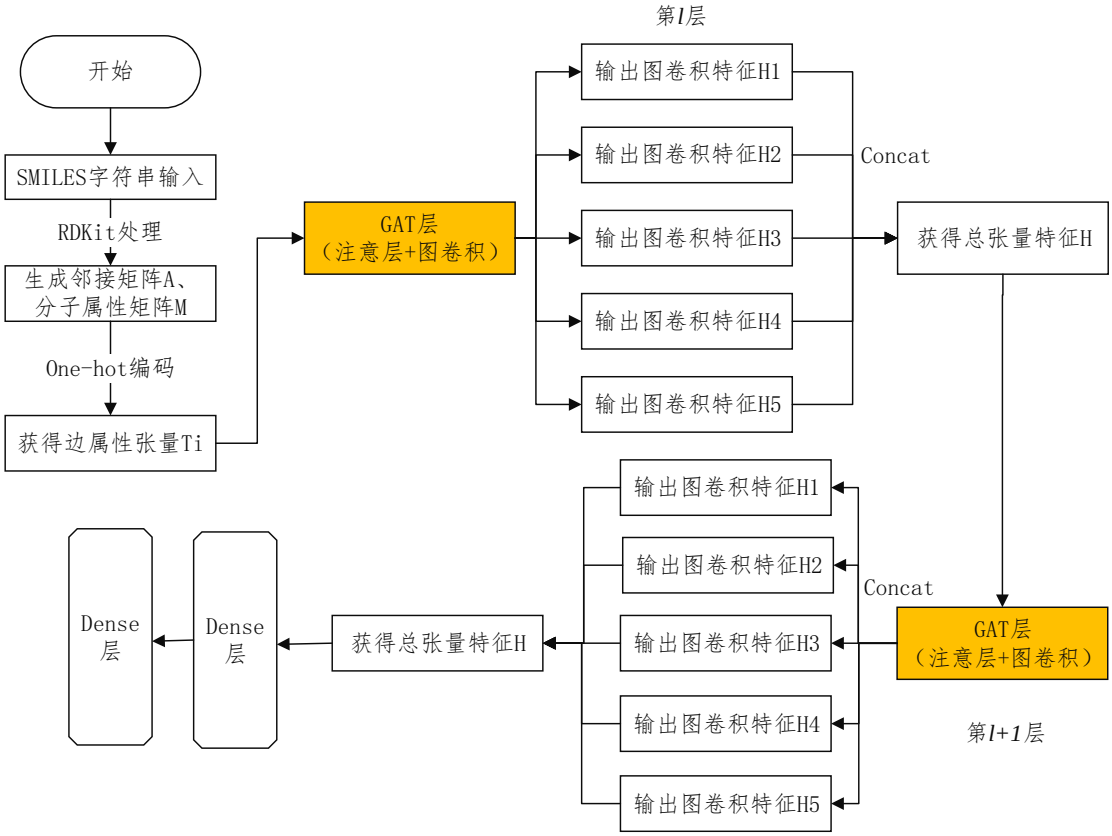


图 3-3 基于边注意的图卷积（EAGCN）模型流程图

（1） 注意层——$A_{att,i}$（图的注意力矩阵）构建过程

上文介绍了本算法中的分子图属性特征选择，由以上可得邻接矩阵 A 和分子属性张量 M。为了计算得到每种属性中不同边的权重，我们将分子属性张量 M 按 K 个边属性拆分（本文选用表 3.2 中的前五个边属性，即 K=5）。由于每个边属性含有不同的状态值，因此 one-hot 编码，会生成 5 个维度为 $N*N*d_i$ 的邻接张量 $T_i^l \in R^{N*N*d_i}$，d_i 为边属性类型数。然后，通过边属性张量 T_i 获得 $A_{att,i}^l$ 中的权重 $a_{i,j}^l$。（公式 3.6）

$$A_{att,i}^l =< T_i^l,\ D_i^l > \tag{3.6}$$

这里，$T_i^l \to A_{att,i}^l$ 的处理过程为：

1）先通过具有 d_i 个输入通道和 1 个输出通道的卷积处理，使用尺寸为 1*1* d_i 的过滤器D_i^l，以 1 为步长移动。其中，l 表示在第 l 层边注意层。

2）其次为了使权重在不同边中具有可比性，使用 softmax 函数对权重进行归一化（公式 3.7）。softmax 函数得到的输出值互相关联，它可以将其量化到 0~1 范围内，且输出值总和为 1。

$$\left(\tilde{A}_{att,i}\right)_{s,t}=\frac{exp\left(A_{att,i}^{l}\right)_{s,t}}{\sum_{t=1}^{M}exp\left(A_{att,i}^{l}\right)_{s,t}} \tag{3.7}$$

简单来说，Attention 层会通过二维卷积层（kernel_size 为 1*1）和 softmax，对属性邻接张量T_i^l进行操作，为每个边属性得出一个边权重矩阵$A_{att,i}^{l}$。

（2） 图卷积层——与图卷积网络的连接

在一个邻接矩阵中，图卷积只关注一个节点相邻节点的信息，即只取局部信息。在每个图卷积层中，对所有一阶邻居进行节点信息聚合，然后进行线性变换。通过公式 3.8 计算图卷积特征：

$$H_i^{l+1}=\sigma\left(\tilde{A}_{att,i}H^lW_i^l\right) \tag{3.8}$$

其中，i 的范围为：1≤i≤K，K 为边属性个数。σ 为激活函数（这里使用 ReLu）。每个边属性 i 会生成 $\tilde{A}_{att,i}$ 里的一个值，因此 $\tilde{A}_{att,i}H^l$ 可以看作是节点特征的加权总和。接下来将不同属性张量得到的图卷积特征进行拼接，这里直接使用 concat 方式，最终形成总特征张量 $H^{l+1}=\left\{H_i^{l+1}\in R^N\times R^{F_i'}\mid 1\le i\le K\right\}$。

重复上述步骤，再将H^{l+1}进行下一层的图卷积特征提取。经过两层 GAT 层处理后，最后再衔接两个全连接层进行分子模型的分类计算，得出分类置信度。

3.2 基于多特性融合的注意力图卷积

3.2.1 EAGCN 存在的问题

本文将 EAGCN 用于本文收集的不同种类的生物活性预测数据集，得到了比传统机器学习更好的模型性能。EAGCN 模型的很多特性是使得其在大部分生物活性数据集上性能较优的原因：

1）其直接对分子图进行学习，可以很好地避免人工筛选特征带来的误差，一定程度上避免了模型的鲁棒性和可靠性受到影响；

2）其生成的注意权重矩阵取决于一个节点的领域特性，而不是全局特性；且权重在所有图中共享，这样可以通过共享的特征来实现提取数据的局部特性。

但是在建模过程中发现，在分子属性矩阵作为注意层的输入时（包含分子信息和边信息），EAGCN 虽然使用注意层对属性信息进行处理，得到了每条边的权重，但是在特征融合时，普通的 concat 方式只是进行简单的维度拼接，不仅导致这里的多个属性信息没有区分度，增加维度还可能会降低后面模型的计算效率，影响模型性能。于是我们提出了基于注意力的多特性融合方案，可以进一步的利用分子图的边属性信息进行特征提取，进而提升模型性能。

在原始模型中，权重张量经过图卷积处理得到特征后，整合特征图信息时使用 concat 的方式合并通道。concat 方式经常用于将特征联合、多个算法框架提取的图特征融合又或是将输出层的信息进行融合，将融合后的特征作为下一个网络层的输入。concat 较为常用，但也存在一些问题：concat 是简单的特征张量的维度拼接，相当于只是通道数的增加。这只是增加了图像本身的特征，对于多特征的重要度分析并没有起到太大作用。于是本文提出使用多特性融合的方式替换 concat 方法。在 EAGCN 中，注意力机制被用于从邻居节点那里学习节点对之间边的交互强度，简单来说是为了得知边在整个图中的重要性。为了更有针对性的知道每种边属性特征的重要性，且能够有效地让模型自适应调节多个特征张量的权重分配，本文提出了多特性融合的方法进行算法优化。这是基于自注意力机制（Self-Attention）[40]的特征融合方案，它可以对输入的每个元素赋予不同的权重参数，从而“挑出”每种特征中较为重要的信息，抑制但不丢失其他信息。其最大的优势就是能一步到位的考虑全局联系和局部联系，可以进一步提高模型的学习效率。

3.2.2 注意力机制

1. Attention 的来源及本质

人类视觉注意力机制是人类大脑在处理视觉所传送的信息时的一种特有机制[39]。简单来说，就是在扫描全局图像后，人们会自动关注那些需要重点关注的区域，忽略其他较不重要的信息。注意力机制可以看作是仿照人类行为的一种算法[39]，是数据处理的一种方法，目的是要更高效的分配模型有限的资源。其

核心思想是利用“有限的注意力”资源从大量信息中快速获取对当前目标任务具有更高价值的信息。使用注意力机制可以使信息处理的效率和准确性更高。

2. Attention 的理解

Attention 是始于一个文本处理领域里常用的框架——Encoder-Decoder，但是 Attention 并不依赖于此框架，它是一种通用思想，可用于绝大部分领域。Encoder-Decoder 框架放在文本处理里，就是先利用 Encoder 对输入的句子 Source 进行编码，经过一系列非线性变换过程即编码过程，生成中间语义；然后 Decoder 根据输入语句的中间语义表示和之前生成的一些历史信息，来生成某时刻要生成的单词，最终会生成目标句子 Target。此时若 Source 为中文，Target 为英文，那么此框架就可以用来机器翻译。若 Source 是一个问题，Target 是一个答案，那么此框架就变成了智能问答系统等类似应用的框架。Attention 机制的提出是为了使解码器能选择性的关注编码后的信息。

总的来说将 Attention 机制分为两大类：

1）聚焦式注意力（focus attention）：自上而下，主动注意（有预定目标的主动去聚焦某一对象）。

2）显著性注意力（saliency-based attention）：自下而上，被动注意（与任务无关的，不需要主动干涉）。

3. Attention 的计算

准确来说，注意力机制更像是一种方法论，它根据具体的目标任务对需要关注的方向和模型进行调整，但是没有严格的数学定义。Attention 是将一个查询（Query，Q）和键值对（Key-Value，K-V）映射到输出的方法，Q、K、V 均为向量，可以将上面的 Source 中的组成元素看成是<Key, Value>数据对，Query 就是 Target 中的某个元素，Attention 的作用就是计算 Query 和各个 Key 的相似性，这个相似性就是 Value。对于每一个 Query 来说，每个 Key 都会经过 Attention 计算得到注意力权重并进行归一化，输出的向量是 Value 的加权求和。类比于人类看图像的聚焦过程，这里注意力权重的计算就是如此，权重越大就是告诉模型此信息的重要程度，Value 就是对应的信息。

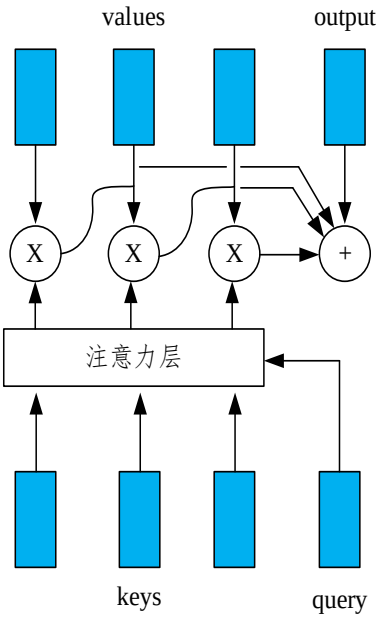


图 3-4 注意力机制框架流程图

其计算流程图如图 3-4 所示。注意力机制的步骤可以总结如下：

步骤 1：输入 $X=[x_1,\ldots,x_N]$，N 表示输入信息的数量；

步骤 2：注意力打分：$a_i=softmax(s(key_i,q))=softmax(s(X_i,q))$。其中，$s(x_i,q)$ 称为注意力打分机制；

步骤 3：信息加权平均：$att(q,X)=\sum_{i=1}^{N}a_iX_i$。用这种方法对输入信息 X 进行编码，得到注意力权重。

步骤 2 中的 a_i 即为 key_i 对应的权重系数，可以引入不同的打分机制来计算相似度。常见的方法有向量点积计算、Cosine 相似性或额外引入其他基于神经网络的模型来计算。这也是最为关键的一个步骤。

4. 自注意力机制（Self-Attention model）

目前主流的一种注意力机制为自注意力机制（Self-Attention model）[40]。循环神经网络（Recurrent Neural Network, RNN）经常用于序列处理，但它存在一个问题：不易被平行化，即不容易并行运算。这时有人提出用卷积神经网络（Convolutional Neural Network, CNN）来代替 RNN 来解决并行化问题，但也并没有真正解决问题。于是人们又提出了 Self-Attention layer 来取代 RNN[40]。

Self-Attention 一开始是用来处理机器翻译任务的，但后来由于其良好的可移植性等，逐渐应用于大量领域。

Self-Attention 的结构如图 3-5 所示。其计算公式如下：

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.9}$$

为了更好的说明，这里以机器翻译为例。Self-Attention 用在文本处理时，核心思想就是为了给输入的句子中的每个单词学习一个权重。每个单词有 3 个向量，即上节提到的 Q、K 和 V 向量。这里假设输入为 X，则 X 是句子的词嵌入向量。Q、K 和 V 是由词嵌入向量 X 与三个不同的权值矩阵 W^Q,W^K,W^V 相乘得到。之后的步骤如下：

1）得到三个向量后，为每个向量计算一个分数（score）： $score=Q\bullet K$；

2）对 score 进行归一化：除以 $\sqrt{d_k}$；

3）使用 softmax 函数对 score 进行归一化处理；

4）为输入向量的分数加权，用 3 中得到的值点乘 V 得到 v；

5）累加后得到最终权重： $\sum v$。

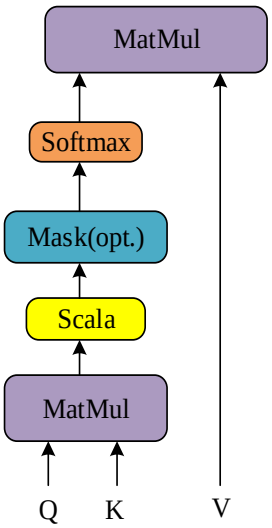


图 3-5 Self-Attention 结构图

3.2.3 多特性融合

在上一节中，本文介绍了 Self-Attention 机制用于机器翻译时的算法流程。本文所提出的多特性融合同样用到其去关注较高权重的边属性。在 EAGCN 中，

经过实验可知“原子对类型”这一边属性对整个模型性能影响较大，因此在设置网络通道数参数时，本文将为原子对类型的特征矩阵设置更高的通道数，相当于使用人工设置偏向权重的方法。而多特性融合是基于自注意力机制，这是为了能让模型自适应学习不同边属性的权重，不受人工影响，关注重要特征且不丢失信息，提升模型学习效率。

在 3.1 章节中，EAGCN 为每张图生成了分子属性张量 $M \in R^{N_{atom}*N_{atom}*N_{features}}$，为了计算得到每种属性中不同边的权重，将分子属性张量 M 进行 one-hot 编码，再将 5 个属性张量输入注意层，进而得到 5 个权重矩阵 $A_{att,i}^{l}$。经过图卷积层的处理后（如图 3-6）得到 5 个图卷积特征矩阵 H^{l+1}。

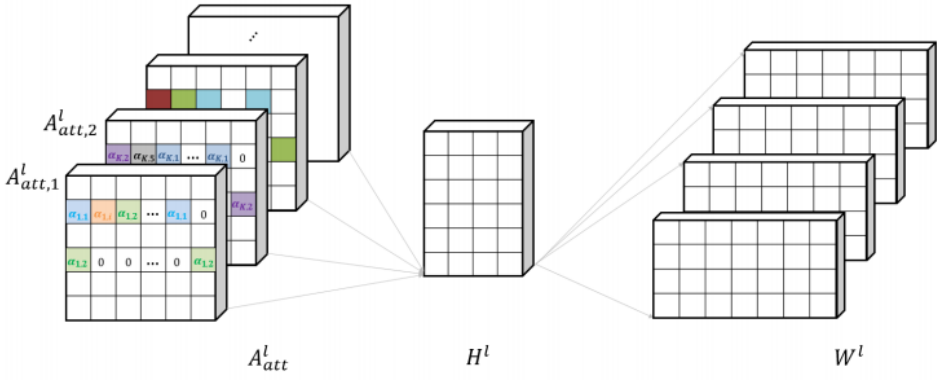


图 3-6 EAGCN 中图卷积层矩阵运算示意图，图源自论文[36]

1. 为每个输入生成 Q、K、V 权重张量

在得到图卷积特征 H^{l+1} 后，这里将这 5 个图特征张量作为输入。H^{l+1} 的维度根据模型中设置的通道数而变化。以一个维度为 N*30 的图特征矩阵为例，先为每个特征张量设置 3 个不同的权重张量，分别为查询 Q、键 K、值 V 张量，长度默认为 64。这三个张量是需要输入特征与权重矩阵相乘得到的。计算示例如图 3-7 所示。W^{Q},W^{K},W^{V} 是三个不同的权重矩阵（三个矩阵维度相同，都为 30*64），用特征张量 H_i^{l+1}（维度为 N*30）与它们相乘，得到对应的 Q、K、V 张量。上述过程在计算时其实是基于矩阵运算的，即运算时是将输入张量合并计算的。

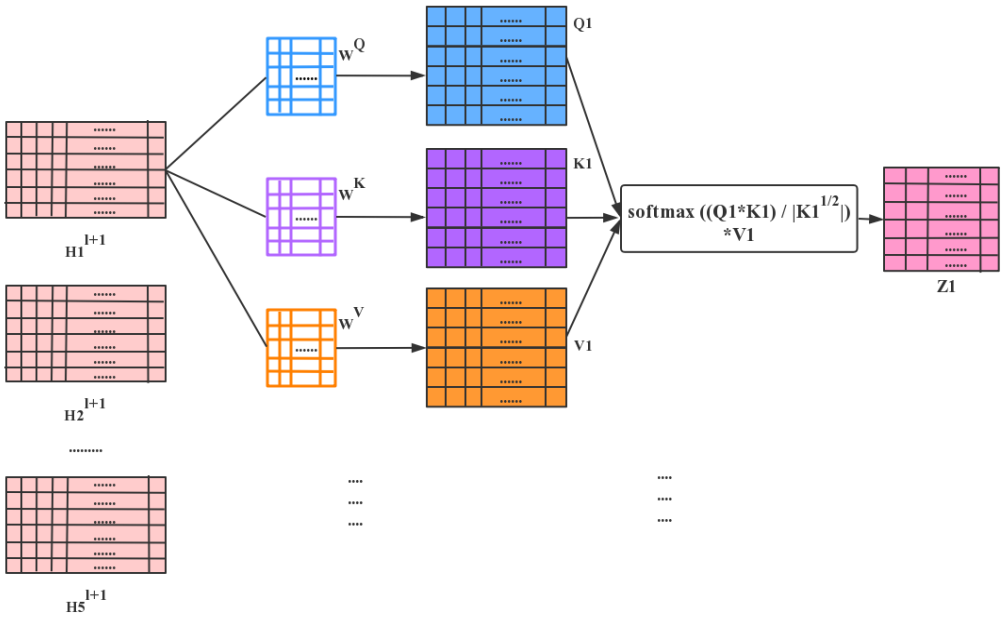


图 3-7 多特性融合方式流程图

2. 计算得分。将每个特征的键向量和查询向量进行点积运算，得到其分数：$score = Q \bullet K$；

3. 为了让梯度更稳定，将 2 中计算得到的 score 除以 $\sqrt{d_k}$ 进行归一化。这里 d_k 为 K 向量的模长，即 64。

4. softmax 归一化。这里使用 softmax 对所有特征张量的 score 进行归一化，使得得到的 score 都为正且和为 1。这一步的目的是初步得到每个边属性对于整个图的权重。

5. 将值向量 V 与 softmax 分数点乘，得到加权的每个输入张量（图卷积特征）的评分 v；

6. 对加权值向量求和，得到输出：$z = \sum v$，$z \in R^N * R^{30}$，即为权重矩阵 z。

经过多特性融合处理，我们可以更清晰的得知 K 个边属性特征张量中，哪些属性对于特征提取更有效果，使得模型性能更优。

3.3 模型性能分析

3.3.1 实验设置

本节实验首先将 EAGCN 应用于本文选用的不同类型生物活性分类数据集，然后将基于多特性融合的注意力图卷积应用于同样的数据集中。本节设计实验的

目的是：1）验证基于边注意的图卷积模型相较于传统机器学习方法（例如随机森林、深度神经网络等）确实更能提升对生物活性数据的分类性能，且由于数据的多样性，模型在生物活性预测问题中也具有一定的普适性；2）验证本文针对特征融合方式进行优化得到的模型——基于多特性融合的注意力图卷积模型，在生物活性预测任务中的性能提升；3）发现模型中依然存在的问题，并在下文中对其提出优化方案。

在传统机器学习中，需要使用计算生成的分子描述符，因此本文在设计实验前，对于分子 SMILES 数据，使用 RDKit（开源化学计算软件包）生成一维分子描述符作为基准模型的特征；同时使用 RDKit 提取出分子的原子、边属性特征用于本文算法。

为了降低传统留一法划分数据法中的偶然性，提高泛化能力，使得数据使用率高，且考虑到算法复杂度，模型的数据集划分选用八折交叉验证法，然后用不同的随机种子执行 3 次。同样，这里得到的结果均为 3 次运行的平均值，并列出了标准偏差。

1. 基准实验设置：

本文使用的基准方法为随机森林（Random Forest，RF）、支持向量机（Support Vector Machines，SVM）及深度神经网络（Deep neural networks，DNN）三种。使用 RDKit 生成的 200 个一维分子描述符建立模型。

针对三种模型，如表 3-3 所示，设置了超参数列表进行模型调参。同样的，数据集划分选用八折交叉验证法，然后用不同的随机种子执行 3 次。这里得到的结果均为 3 次运行的平均值，并列出了标准偏差。

表 3-3 各模型超参数设置表

超参数	值区间	参数意义
Random Forest		
Ntrees	(50,100,150,...,500)	树的个数
max_depth	(1, 5,10,...,45,50)	每棵树最大树深度
max_features	(1, 5,10,...,45,50)	划分时的最大特征数
Support Vector Machines		
Kernel	RBF	核函数

C	(1,10,100)	惩罚系数
γ	(0.1,0.001,0.0001,0.00001,1,10,100)	影响数据映射到新特征空间的量
Deep neural networks		
Epoch	100	迭代次数
Batch size	100	最小训练样本数
Hidden layers	(2,3,4)	隐层数
Number neurons	(10,50,100,500,700,1000)	每层神经元个数
Activation function	ReLu	神经元激活函数
Loss function	binary_crossentropy	损失函数

2. EAGAN 与 MF_EAGCN 算法的实验设置

在 EAGCN 建模时根据分析得到，原子对类型这一属性的权重设置较大时，模型性能会较好，于是在该算法中我们人工的将原子对类型的 GCN 层输出通道数设置的偏大，为了更好地学习此特征，做出了人工干涉。在我们优化的 MF_EAGCN 中，会去自行关注较高权重的边属性，即可以自适应的学习不同的边属性权重。本文设置的实验参数如下表所示。

表 3-4 EAGCN 与 MF_EAGCN 模型超参数设置表

超参数	**值区间**	**参数意义**
EAGCN		
Batch size	64	单次训练样本数
Epoch	100	迭代次数
weight_decay	0.00001	权重衰减率
dropout	0.5	随机失活率
Activation function	ReLu	激活函数
Loss function	binary_crossentropy	损失函数
kernel_size	1	卷积核大小
stride	1	卷积核滑动步长
n_sgcn1	(30, 10, 10, 10, 10)	多特征图卷积层输出通道数
MF_EAGCN		

Batch size	64	单次训练样本数
Epoch	100	迭代次数
weight_decay	0.00001	权重衰减率
dropout	0.5	随机失活率
Activation function	ReLu	激活函数
Loss function	binary_crossentropy	损失函数
kernel_size	1	卷积核大小
stride	1	卷积核滑动步长
n_sgcn1	(20, 20, 20, 20, 20)	多特征图卷积层输出通道数

3. 评价指标：本文使用两种评价指标：准确率（Accuracy，ACC）和平衡 F 分数（BalancedScore，F1-score）。

其中准确率（ACC）是分类预测中较为常用的评价指标（公式 3.10）。

$$\text{accuracy} = (TP+TN)/(P+N) \tag{3.10}$$

其中，TP、TN 分别为被正确地划分为正例、负例的个数；P、N 为实际样本中正例、负例的个数。总的来说，ACC 就是被分对的样本数占所有的样本数的比例，ACC 指标值越高，分类器性能越好。

平衡 F 分数 F1-score 也是生物活性分类任务中常用来衡量模型精确度的指标（公式 3.11）。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.11}$$

F1-score 同时考虑到了模型的精确率（precision）和召回率（recall），只有在两个值都高时，F1 的值才会更高，模型性能越好。

3.3.2 算法性能分析

表 3-5 显示了在几种数据集上，不同基准模型的 ACC、F1-score 指标结果。

表 3-5 在本文算法和 EAGCN 及三种基准方法下，七种数据集的预测结果

Task	**ACC**					**F1-score**				
Dataset	**RF**	**SVM**	**DNN**	**EAGCN**	**EAGCN_MF**	**RF**	**SVM**	**DNN**	**EAGCN**	**EAGCN_MF**
1851(1a2)	0.824	0.8	0.835	**0.85**	**0.859**	0.792	0.78	0.8	**0.83**	**0.841**
	±0.005	±0.02	±0.015	**±0.01**	**±0.012**	±0.01	±0.008	±0.007	**±0.012**	**±0.01**
1851(2c19)	0.776	0.75	0.79	**0.802**	**0.815**	0.8	0.77	0.823	**0.84**	**0.852**
	±0.01	±0.009	±0.002	**±0.007**	**±0.003**	±0.004	±0.005	±0.01	**±0.01**	**±0.008**
1851(2d6)	**0.849**	0.83	0.84	0.843	**0.851**	0.828	0.8	0.82	**0.83**	**0.834**
	±0.006	±0.007	±0.002	±0.005	**±0.003**	±0.013	±0.004	±0.003	**±0.01**	**±0.006**
1851(3a4)	0.77	0.737	0.792	**0.817**	**0.825**	0.73	0.701	0.74	**0.791**	**0.807**
	±0.006	±0.004	±0.008	**±0.006**	**±0.01**	±0.003	±0.006	±0.01	**±0.008**	**±0.005**
492992	0.713	0.705	0.745	**0.757**	**0.762**	0.683	0.674	0.692	**0.74**	**0.75**
	±0.004	±0.006	±0.005	**±0.01**	**±0.01**	±0.005	±0.006	±0.009	**±0.01**	**±0.009**
651739	0.753	0.753	0.814	**0.83**	**0.843**	0.80	0.776	0.88	**0.882**	**0.891**
	±0.004	±0.006	±0.014	**±0.006**	**±0.003**	±0.003	±0.009	±0.006	**±0.007**	**±0.002**
652065	0.75	0.70	0.755	**0.77**	**0.774**	0.73	0.67	**0.796**	0.787	**0.792**
	±0.004	±0.005	±0.015	**±0.006**	**±0.005**	±0.008	±0.009	**±0.012**	±0.01	**±0.01**

从实验结果（表 3-5）可以看出，在这些数据集中，基于图卷积的 EAGCN 展现出了比传统机器学习方法更好的分类性能，其 ACC 指标均比基准学习模型高出 2%-8%个点，F1-score 指标比基准学习模型高出 1%~5%个点。可见直接从分子图学习而不是从预先计算的特性中获得的信息使得模型性能更优。少部分数据集中，DNN 的性能能与 EAGCN 方法性能基本持平或稍微高于其性能，RF 的性能有时可以与 EAGCN 持平。可见，EAGCN 的性能还有很多优化空间。而基于多特性融合的 MF_EAGCN 模型，展现出了更好的分类性能，这也证实了多特性融合方案能够更充分地利用边属性信息进行特征提取，使得模型预测性能提升。其 ACC 指标均比 EAGCN 算法高出 1%~2%个点，F1-score 指标比 EAGCN 模型高出约 1%个点。

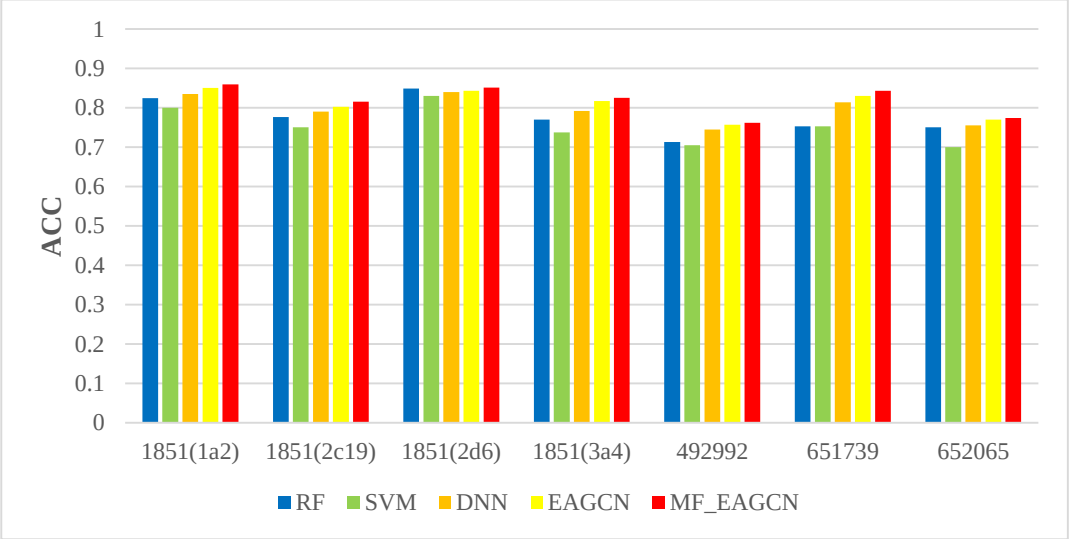


图 3-8 用于表现七种生物活性数据集在五种分类器中性能的 ACC 指标分布

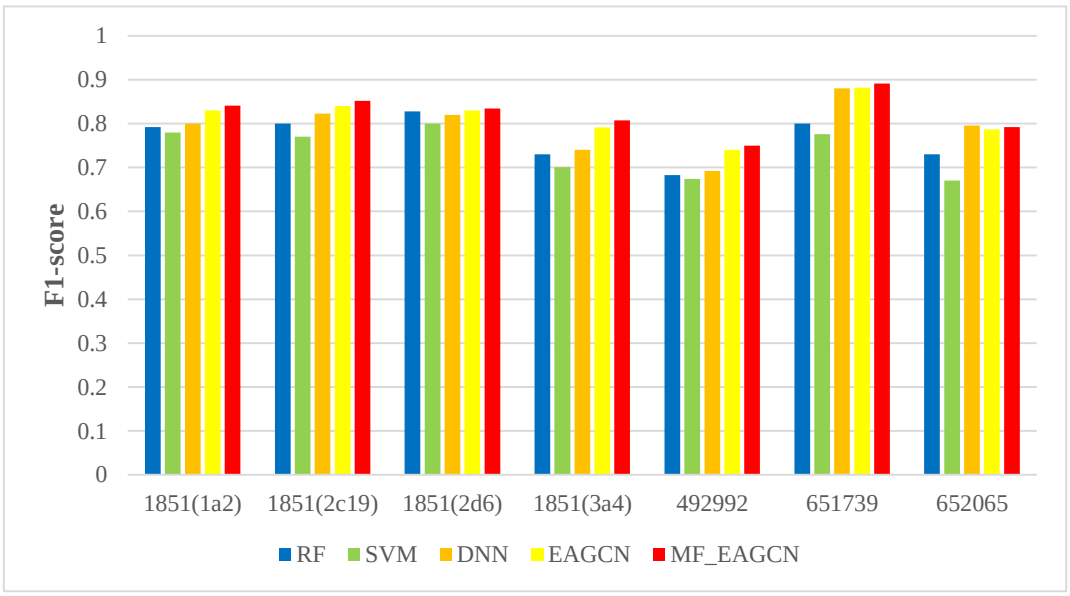


图 3-9 用于表现七种生物活性数据集在五种分类器中性能的 F1-score 指标分布

图 3-8 和图 3-9 分别展示了本文提出的 MF_EAGCN、基准算法 EAGCN 以及传统机器学习方法五种分类器，分别应用于七种生物活性数据集中的 ACC 指标和 F1-score 指标分布对比，柱状图的条目从左到右依次是 RF、SVM、DNN、EAGCN 和 MF_EAGCN 模型。在 ACC 指标分布图中，可以看到数据集 1851(2d6) 在 EAGCN 模型上的效果并不显著，分析其原因有两点，一是数据量相比较而言更大，在模型融合特征阶段对特征重要度分配不均，导致对重要信息的忽略，进而致使模型预测性能降低；二是其正、负样本比例为 1:5 相对较不均衡，也会对模型性能提升有一定的限制。

3.4 本章小结

本章主要介绍了前人提出的基于边注意的图卷积神经网络算法，经过大量实验表明，该算法在分子的生物活性预测任务中表现良好。但模型依然存在着某些问题，本节提出的问题是：权重张量经过图卷积层处理得到特征后，整合图特征信息时使用 concat 的方式合并通道，concat 只是简单的特征张量的维度拼接，对多特征的重要度分析并没有起到太大作用。为了更有针对性的知道每种边属性特征的重要性，且能够有效地让模型为特征自适应分配权重，本文提出了多特性融合的方法进行优化。这是一种基于自注意力机制（Self-Attention）的方案，它可以让模型自适应学习不同边属性特征的权重。实验结果表明，提出的多特性融合方案能够有效地解决 EAGCN 中存在的边属性权重不能自适应的问题，且获得了较好的预测性能。

4 基于样本类别不平衡的损失函数研究

本章内容是针对分子生物活性数据存在的正负样本及难易样本不均衡问题，通过改进本文算法中的损失计算方案：引入了目前较优的两种损失修改方案：聚焦损失（Focal Loss）[46]以及梯度均衡机制（Gradient Harmonizing Mechanism，GHM）[47]，实现了模型性能的进一步优化。

4.1 样本类别不平衡问题

样本类别不平衡是指在数据集中，某一类的样本远少于其他类样本[41]。这种数据集在实际业务中十分常见，如分子活性预测、金融诈骗检测、疾病医学诊断等，而其中的少数类才是真正对目标任务有较大作用的、我们需要重点关注的数据。样本类别不均衡主要体现在两方面[41]：正负样本不均衡（正负样本比例悬殊）和难易样本不均衡（简单样本主导整体损失）。例如部分分子活性数据集就存在正负样本不均衡的问题，分类器会偏向样本数较多的类，这就对模型性能产生了影响。在疾病医学诊断中，样本不平衡问题极为严重，如一些罕见病例的患者病例较少则其样本数远少于正常样本，而此时若将患者误诊为正常，后果将不堪设想。一般在分类任务中会尽量保持正负样本的比例均衡，如若正负样本比例偏差过大会导致模型不能有很好的召回能力。

难易样本不均衡问题有区别于正负样本数量不平衡。对于一个样本，如果它能很容易地被正确分类，那么这个样本对模型来说就是一个简单样本，模型很难从这个样本中得到更多的信息；而对于一个分错的样本，它对模型来说就是一个困难的样本，它更能指导模型优化的方向。对于一般的分类器来说，简单样本的数量非常大，他们产生的累计贡献在模型更新中占主导作用，而这部分样本本身就能被模型很好地分类，所以这部分的参数更新并不会改善模型的判断能力，这反而会导致整个训练变得低效。

在分子生物活性分类任务中，样本包括以下类别（根据正、负、难、易，样本一共可分为以下四类，如图 4-1）：

	难	易
正	正难	正易
负	负难	负易

图 4-1 样本分类图

- 正样本：具有/激活某种生物活性的分子。
- 负样本：不具有/抑制某种生物活性的分子。
- 易分正样本：容易被正确分类的正样本（被正确分类的正样本），一般来说该类在整体样本中占比较高，单个样本的损失值较小，但累计的损失值会主导整体损失值。
- 难分正样本：不易被分类的正样本（被错误分类的正样本），该类在整体样本中占比较低，单个样本的损失函数值会较高但不会主导整体损失。
- 易分负样本：易被正确分类的负样本（被正确分类的负样本），在整体样本中占比较高，同样单个样本损失低但累计损失占比。
- 难分负样本：不易被正确分类的负样本（被错误分类的负样本），在整体样本中占比较低，同样单个样本损失高但累计损失占比小。

解决这种样本不平衡问题主要有两种思路：从数据或算法角度来解决。

（1）数据角度，尽量修改及平衡各类别分布情况，最常见的方法如扩大数据集、均衡采样等。采样主要分为过采样和欠采样，过采样会把小类样本复制多份，且在生成数据点时加入轻微的随机扰动（如噪声等）可以防止过拟合[42]；欠采样中经典的 EasyEnsemble 算法会去除一些大类样本[42]。数据扩充是利用已有样本生成更多的样本，这种方法在疾病医学诊断中很常用。经典方法如 SMOTE，根据小类样本之间的相似性来生成新样本[43]；ADASYN 根据数据分布比例来为不同的小类样本生成不同数量的新样本[44]。

（2）算法角度，修改算法模型或损失函数等方法。从算法模型方面来说，Faster RCNN 等算法可以通过将大量大类样本和简单样本过滤掉[45]，来缓解正负、难易样本不均衡问题。从损失函数方面来说，通过增大困难样本与小类样本的损失权重，来增大损失所占整体损失值的比例，来缓解不平衡问题，这也是本文内容优化的出发点。

样本类别不均衡问题是一些交叉学科领域中是常见且非常重要的问题。当数据不均衡时，极易导致模型性能下降。这种分类器不能很好的解决实际问题，无法真正的实现数据挖掘。因此我们针对本文算法中存在的样本不均衡问题，提出了两种优化方案，下文将进行详细介绍。

4.2 模型优化

本文提出的算法模型，其预测能力已经达到了一定的精度，但生物活性数据的易分样本、难分样本类别不平衡，仍然是影响其精度提升的主要因素。于是本文通过可以降低易分样本权重的损失函数来解决此问题。优化的思想是将训练的重点放在较少的难分样本上。为了评估优化方案的有效性，我们设计并训练了基于 MF_EAGCN 模型的不同损失函数的实验。结果表明经过优化损失函数后，模型能够获得更好的性能。

4.2.1 聚焦损失（Focal Loss）

Focal Loss[46]的引入主要是为了解决难易样本数量不平衡问题，实际使用的范围非常广泛。在实际模型训练中，大量样本都是分类模型中的易分样本。不仅要考虑正负样本平衡，也许考虑难易样本的平衡问题。这些样本的损失很低，但是由于数量极不平衡，易分样本的数量相对来讲太多，最终主导了总的损失。但是如果将简单样本的权重降低就可以更有效的平衡难易样本，接下来 Focal Loss 也是通过这个方向来优化的。

首先，在分类任务中常用的损失函数——交叉熵损失（Cross Entropy loss, CE）的公式如下（公式 4.1）：

$$L=-y\log y'-(1-y)\log(1-y')=\begin{cases}-\log y' & y=1\\ -\log(1-y') & y=0\end{cases} \tag{4.1}$$

这里以本文中的二分类标签为例，y 值包含 0 和 1。其中，y' 表示模型得到的该样本的预测概率，y 表示样本的标签。由公式可以看出，当 y 为 1 时，预测概率值 y' 越贴近于 1 则损失越小；而标签为 0 时，预测概率值 y' 越贴近于 0 则损失越小。但是交叉熵函数中所有样本的权重一致，若出现正负/难易样本不均衡问题，大量负样本/易分样本占主导地位，少量正样本/难分样本起不到作用时，就会导致精度变差。

首先通过降低易分样本的权重，来使模型在训练时更加专注于难分样本的学习，从而缓解难易样本比例不均衡导致的问题。Focal Loss 使用以下公式进行更正（公式 4.2）：

$$L_{fl}=\begin{cases}-(1-y')^{\gamma}\,log\,y' & y=1\\ -y'^{\gamma}\,log(1-y') & y=0\end{cases} \tag{4.2}$$

公式中，首先在 Sigmoid 激活函数上加入一个调制因子 γ，其中 γ>0，这一步的目的是为了使易分类样本的损失在整体目标函数中所占的比重降低，γ 越大易分样本的损失贡献会越低。例如令 γ 为 2，如果预测置信度（confidence）为 0.95，那么该样本肯定是简单样本，所以（1-0.95）的 γ 次方就会非常小，这时这个简单样本在损失函数中所占的比重就会变得更小。而预测概率为 0.3 的样本，很明显这个样本没那么简单预测出来，属于困难样本，所以其在目标函数中所占的损失比重相对就比较大。

其次，进一步添加了平衡参数 α，α 的主要目的是用来解决训练数据中正负样本的比例不均衡问题（公式 4.3）：

$$L_{fl}=\begin{cases}-\alpha(1-y')^{\gamma}\,log\,y' & y=1\\ -(1-\alpha)y'^{\gamma}\,log(1-y') & y=0\end{cases} \tag{4.3}$$

α 的取值区间为[0, 1]，用来平衡正、负样本的权重。只添加 α 仅仅可以平衡正负样本的比例，但是对于分类器来说简单数据与困难数据的损失权重平衡问题并没有解决。参数 γ 主要就是用来解决分类器中简单数据与困难数据的损失权重平衡问题的，当 γ 为 0 时即为交叉熵损失函数，当 γ 增加时，调整因子的影响也在增加。

4.2.2 梯度均衡机制（Gradient Harmonizing Mechanism，GHM）

Focal Loss 虽然有很好的效果，但缺点之一就是公式中的两个超参不是自适应的，需要人工精细的调整。除此之外，它也是一个不会随着数据分布变化的静态损失。同时，样本中那些将其难拟合的样本称为离群点，当样本中有很多离群点（outliers）时，如果让模型强行去关注如何更好地学习这些离群样本也会存在问题，因为会导致其他大量的样本分类错误，最终导致模型性能出现问题。而 Focal Loss 对这些离群点并没有使用相关策略来处理。梯度均衡机制（Gradient

Harmonizing Mechanism，GHM）[47]的思想就是不过多关注易分样本，但也不过多关注这些离群点。

首先引入一个统计学概念：梯度模长 g（Gradient Norm）（公式 4.4）。

$$g=\left|p-p^{*}\right|=\begin{cases}1-p & p^{*}=1\\ p & p^{*}=0\end{cases} \tag{4.4}$$

其中 p 是模型得到的预测概率，p^{*}是真实标签。梯度模长 g 的范围为[0, 1]，对它和样本数量的关系可进行描述：g 接近于 0 时的样本数量（易分样本）较多，g 逐渐增大的过程中样本数量是迅速减少的趋势，但当 g 快到达 1 时（特别难分样本），样本数量会增多。由 4.1 知识可知，g 越大则预测难度越大。为了解决梯度分布不均匀问题，GHM 基于 g 提出了一个重要概念——梯度密度，使得 GHM 可以同时衰减易分样本和特别难分样本。

难易样本不平衡的实质就是梯度分布的不平衡。这里的梯度分布即为梯度密度（Gradient Density），这一概念可以类比物理学中密度的定义（单位体积内的质量），定义为单位梯度模长 g 内分布的样本数量。梯度密度函数的公式是 GHM 中最重要的公式[47]（公式 4.5）。

$$GD(g)=\frac{1}{l_{\varepsilon}(g)}\sum_{k=1}^{N}\delta_{\varepsilon}(g_k,g) \tag{4.5}$$

其中，$\delta_{\varepsilon}(g_k,g)$表示在 N 个样本中，梯度模长分布在$\left(g-\frac{\varepsilon}{2},g+\frac{\varepsilon}{2}\right)$区间内的样本数量，$l_{\varepsilon}(g)$表示此区间的长度。为了归一化整个梯度的分布，于是基于梯度密度的公式构造了一个密度协调参数 β[47]（公式 4.6）：

$$\beta=\frac{N}{GD(g)} \tag{4.6}$$

当$GD(g)$值较大时，β 变小；反之 β 变大。对于易分和特别难分这两类样本，其分布都很密集，即$GD(g)$的值都很大，而参数 β 刚好可以用来降低这两部分权重，同时可以提高其他样本的权重。这里使用 β 对样本的损失进行加权。

基于 β 参数，将 GHM 的思想用于分类情况时得到新的分类损失函数 GHM-C，定义如公式 4.7：

$$\begin{aligned} L_{GHM-C} &= \frac{1}{N}\sum_{i=1}^{N}\beta_i L_{CE}\left(p_i, p_i^*\right) \\ &= \sum_{i=1}^{N}\frac{L_{CE}\left(p_i, p_i^*\right)}{GD\left(g_i\right)} \end{aligned} \tag{4.7}$$

GHM-C 的公式即在交叉熵损失函数中引入了参数 β。在 GHM-C 损失函数的影响下，大量简单样本的权重被降低，而离群样本的权重也被稍稍降低，即同时解决了不平衡问题和离群点问题。同时，GHM-C 中的梯度密度每次迭代时会更新，不像 Focal Loss 中样本损失的权值是固定的，反之它是自适应的，即 GHM-C 是具有动态特性的，这使得模型更加可靠和有效。

梯度密度的提出理论上是完全有效可行的，但在具体实施时，其计算方式存在一定的问题。常规方法计算所有样本的梯度密度值时，公式 4.7 中求和有一个 N，每次求 $GD\left(g_i\right)$ 时都会遍历所有样本，因此时间复杂度为 $\mathrm{O}\left(N^2\right)$。即使并行计算，每个计算单元也有 N 的计算量。较优的算法会首先通过一个复杂度为 $\mathrm{O}\left(N\,log\,N\right)$ 的梯度正则对样本排序，然后用一个队列去扫描样本，此时得到梯度密度的时间复杂度为 $\mathrm{O}\left(N\right)$。但是当数据量较大时，仍然非常耗时。于是作者提出了近似的求样本梯度密度的方法[47]：

$$\tilde{GD}(g) = \frac{R_{ind}(g)}{\varepsilon} = ind(g)m \tag{4.8}$$

$$\hat{\beta}_i = \frac{N}{\tilde{GD}(g_i)} \tag{4.9}$$

1）将 g 空间划分成 $M = \frac{1}{\varepsilon}$ 个独立的单元区域

2）r_j 代表第 j 个区域，R_j 代表 r_j 内的样本数量，$ind\left(g\right) = t$ 代表 g 属于哪个区域。由公式 4.8 可以发现，在同一个区域内的样本的梯度密度值是相同的，那么此时计算所有样本的梯度密度的时间复杂度就为 $\mathrm{O}\left(MN\right)$。同时使用并行计算的话，每个计算单元复杂度为 $\mathrm{O}\left(M\right)$，这样相对来说计算会比较高效。

综上所说，Focal Loss 是从置信度 p 入手来逐步衰减损失，GHM 是从一定范围置信度 p 内的样本数量的角度，即梯度角度来衰减损失。它们都对易分样本的损失有很好的抑制作用，但 GHM 对特别困难样本的损失有更好的抑制效果。

4.3 实验分析

4.3.1 实验设置

为了分析基于以上两个优化方案在生物活性预测中的性能，本文基于第 3 章中的模型，在 MF_EAGCN 上进行优化，将原本使用的交叉熵损失函数替换成 Focal Loss 和 GHM-C。针对两种模型，如表 4-1 所示，模型其余参数与 3.3.1 节中相同，表里列出了 Focal Loss 和 GHM-C 需要单独设置的超参数列表。同样的，数据集划分选用八折交叉验证法，然后用不同的随机种子执行 3 次。这里得到的结果均为 3 次运行的平均值，并列出了标准偏差。

其中，GHM-C 中是通过以下两个机制来近似求解梯度密度：

1）将梯度取值区间切割为 bin 个区间（对应参数 bins），然后统计不同 bin 区间内的梯度数目 R 即为梯度密度（公式 4.5）。

2）使用一个系数来进行指数加权移动平均计算，以此来近似总样本下的梯度密度。实现代码中使用 momentum 来作为此系数，称为动量部分系数，由于 Li 等人[47]经过分析，得到模型对于 momentum 参数不敏感，因此本文设置为 0.1。

表 4-1 基于两种损失函数的 MF_EAGCN 模型的超参数设置表

超参数	值区间	参数意义
MF_EAGCN_FL		
γ	(0.5, 1, 2, 5)	调制因子
α	(0.1, 0.2, 0.5, 0.7)	平衡参数
MF_EAGCN_GHM		
bins	(1, 2, 3,...,10)	区间个数
momentum	0.1	动量部分系数

本文针对不同的数据集训练了基于取值区间内所有超参数的模型，目的是希望为每个数据集找到适合自己的超参数设置。这里以 1851 靶标家族中细胞色素酶 P450 系列之一的数据集 1851(2d6)为例，其正负样本比例接近 1:5。根据表 4-1 及表 3-5 中 MF_EAGCN 模型的参数设置，分别设计基于 Focal Loss 和 GHM-C 损失函数的 MF_EAGCN 模型。

在基于 Focal Loss 的 MF_EAGCN 模型中，γ 和 α 参数的作用是不同的，γ 主要用来调节难易样本，α 用来调节正负样本。首先调节 α：设置 γ 为 1，α 取值

为(0.1, 0.2, 0.3, 0.5, 0.7)。由实验结果可知，当 α 为 0.3 时，模型性能较好。然后调节 γ：设置 α 为 0.3，γ 取值为(1, 2, 5)。如图 4-2 所示，可以看到 γ 对 Focal Loss 的调节。当 γ 为 1 时，模型性能较优。

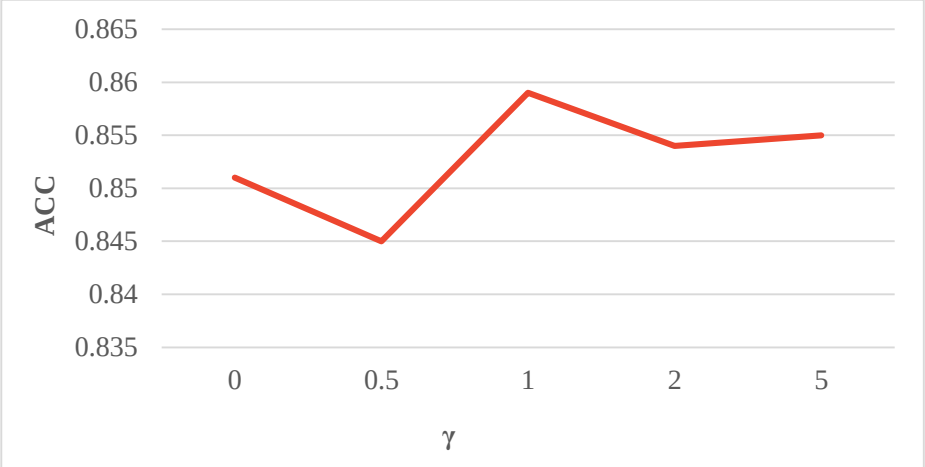


图 4-2 不同 γ 取值下的 ACC 变化趋势

在基于 GHM-C 的 MF_EAGCN 模型中，设置 bins 取值为(1, 2, ..., 10)，即 1~10 内以 1 为公差递增。经过实验得到，当 bins 取值为 5 时，模型性能较优。

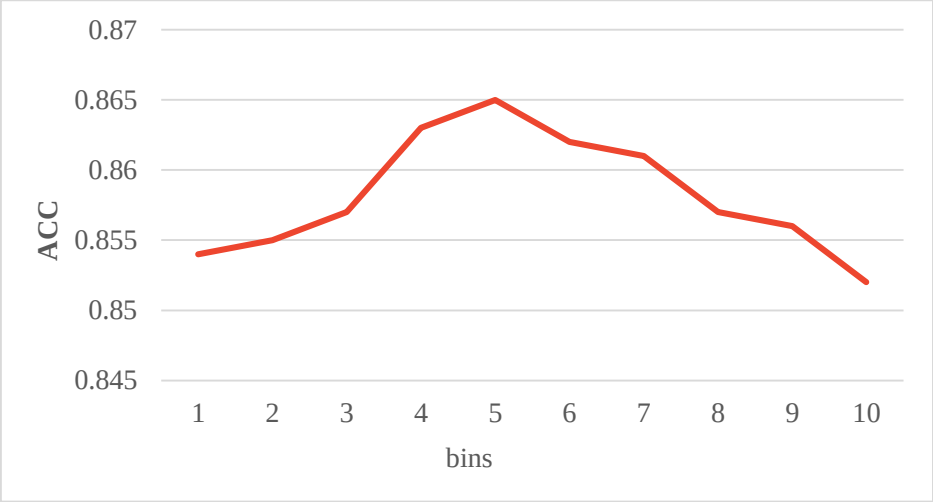


图 4-3 不同 bins 取值下的 ACC 变化趋势

4.3.2 性能分析

表 4-2 显示了在几种数据集上，基于两种不同损失函数的 MF_EAGCN 模型的 ACC、F1-score 指标结果，并与 EAGCN 进行了对比。

表 4-2 在基于两种不同损失函数的 MF_EAGCN 模型下，七种数据集的预测结果

Task	**ACC**				**F1-score**			
Dataset	**EAGCN**	**MF_EAGCN**	**MF_EAGCN_FL**	**MF_EAGCN_GHM**	**EAGCN**	**EAGCN_MF**	**MF_EAGCN_FL**	**MF_EAGCN_GHM**
1851(1a2)	0.85	0.859	**0.861**	**0.87**	0.83	0.841	**0.845**	**0.861**
	±0.01	±0.012	**±0.003**	**±0.004**	±0.012	±0.01	**±0.004**	**±0.003**
1851(2c19)	0.802	**0.815**	0.81	**0.828**	0.84	0.852	**0.856**	**0.862**
	±0.007	**±0.003**	±0.008	**±0.002**	±0.01	±0.008	**±0.009**	**±0.01**
1851(2d6)	0.843	0.851	**0.86**	**0.865**	0.83	0.834	**0.853**	**0.859**
	±0.005	±0.003	**±0.007**	**±0.01**	±0.01	±0.006	**±0.004**	**±0.008**
1851(3a4)	0.817	**0.825**	0.825	**0.839**	0.791	**0.807**	0.805	**0.817**
	±0.006	**±0.01**	±0.004	**±0.006**	±0.008	**±0.005**	±0.008	**±0.003**
492992	0.757	0.762	**0.764**	**0.776**	0.74	0.75	**0.758**	**0.763**
	±0.01	±0.01	**±0.008**	**±0.002**	±0.01	±0.009	**±0.01**	**±0.004**
651739	0.83	0.843	**0.848**	**0.859**	0.882	0.891	**0.897**	**0.904**
	±0.006	±0.003	**±0.005**	**±0.01**	±0.007	±0.002	**±0.004**	**±0.002**
652065	0.77	0.774	**0.778**	**0.782**	0.787	0.792	**0.796**	**0.801**
	±0.006	±0.005	**±0.004**	**±0.006**	±0.01	±0.01	**±0.005**	**±0.004**

这里着重标注了性能较优的前两个模型。从实验结果（表 4-2）可以看出，在这些数据集中，基于两种优化方案的 MF_EAGCN 展现出了比 EAGCN 和 MF_EAGCN 更好的分类性能，基于 Focal Loss 的 MF_EAGCN（MF_EAGCN_FL）的分类器的 ACC 指标比原先两个模型高出 1%个点左右，F1-score 指标高出 1%个点左右。可见 Focal Loss 损失函数对于模型中存在的样本不均衡问题起到了一定的缓解作用。但在少部分数据集中，MF_EAGCN 的性能与 MF_EAGCN_FL 方法性能基本持平或稍微高于其性能，这其实是由于 Focal Loss 本身的一些限制导致的，正如之前所提到的需要人工精细调参和对离群点样本的特别关注，体现在本文就是模型性能提升上出现的瓶颈。而基于 GHM-C 的 MF_EAGCN 模型（MF_EAGCN_GHM）展现出了更好的分类性能，其 ACC 指标均比 EAGCN 算

法高出 1%~3%个点，F1-score 指标高出约 2%~3%个点。相对于 MF_EAGCN，ACC 指标也是高出 1%~2%个点。

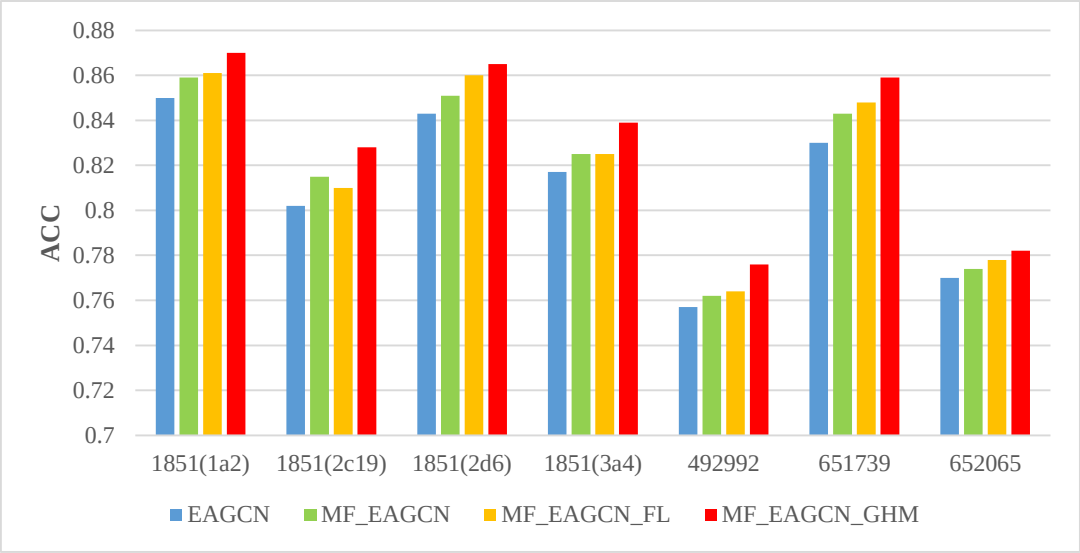


图 4-4 基于不同损失函数的 MF_EAGCN 模型与基准模型在七种生物活性数据集中的 ACC 指标分布

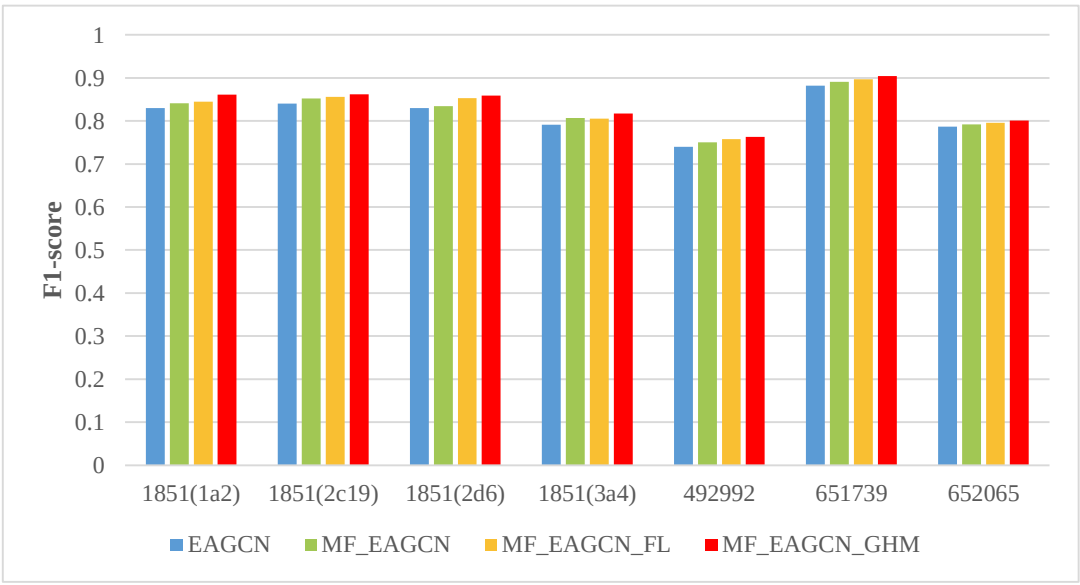


图 4-5 基于不同损失函数的 MF_EAGCN 模型与基准模型在七种生物活性数据集中的 F1-score 指标分布

图 4-4 和图 4-5 分别展示了基于不同损失函数的 MF_EAGCN 模型和 EAGCN 模型在七种生物活性数据集中的 ACC 指标和 F1-score 指标分布对比，柱状图的条目从左到右依次是 EAGCN、MF_EAGCN、MF_EAGCN_FL 和 MF_EAGCN_GHM 模型。综上，GHM-C 在分子的生物活性预测任务中，较大程度上缓解了数据中存在的样本类别不均衡问题，且比 Focal Loss 更能有效解决此问题。

4.4 本章小结

本章首先对前期模型进行了分析。虽然 MF_EAGCN 模型相对来说提升了预测精度，但生物活性数据的易分样本、难分样本类别不平衡，仍然是影响其精度的主要因素。于是本章针对生物活性数据集中普遍存在的问题：样本类别不均衡问题，引入了基于损失函数优化的两种方案：聚焦损失以及梯度均衡机制，通过可降低易分样本权重的损失函数来解决此问题。为了评估优化方案的有效性，我们设计并训练了基于 MF_EAGCN 模型的不同损失函数的实验。并且实验结果表明经过优化损失函数后，模型获得了更好的性能。

5 总结与展望

5.1 总结

本文以构建较为可靠的生物活性预测模型为目的，针对基于边注意机制的图卷积网络模型从两个不同方面进行了有效优化。文章首先介绍了研究生物活性预测的意义以及机器学习在其中的应用现状。同时介绍了与本文相关的算法及生物活性预测相关理论，为接下来对生物活性的建模研究奠定了基础。然后深入研究一种基于边注意力的图卷积的算法，并针对其存在的两大问题，提出了解决方案。详细研究内容如下：

1）选取不同类型的数据集

本文所选用的数据集来自于公共化学数据库 PubChem。并且使用文献中的多种分析筛选方法对靶标等内容作出了限制，选择了不同类型的几种生物活性数据集。包括 1851 靶标家族中细胞色素酶 P450 系列的 4 个数据集、两种抑制剂活性数据集和识别结合 r(CAG) RNA 重复序列的分子系列。

2）研究基于边注意的图卷积网络

本文将一种基于边注意力的图卷积网络架构，应用于文中选用的不同种类的生物活性预测任务，从而避免了人工特征工程带来的误差，并对比几种机器学习基准算法，验证了算法有效性。

3）研究特征融合方式，提出多特性融合方案

针对前人提出的模型中存在的问题：无法自适应设置边属性特征权重，本文提出了分子多特性融合的方案优化了算法模型的特征提取能力，通过自注意力机制针对多个特征进行自适应融合，有效地解决了这一问题，并且获得了更好的预测性能。

4）研究样本不均衡问题，提出损失优化方案

针对分子生物活性数据存在的正负样本及难易样本不均衡问题，通过改进本文算法中的损失计算方案：引入了目前较优的两种损失修改方案：聚焦损失（Focal Loss）以及梯度均衡机制（Gradient Harmonizing Mechanism，GHM），实现了模型性能的进一步优化。

5.2 展望

本文研究并构建了较为可靠的基于图卷积的生物活性预测模型，其中包括对多特征注意和样本不均衡问题两大问题的优化，但未来仍存在改进空间：

1）本文使用的数据集偏向数据量较小的数据集，未来可以将其扩展到数据量更大的数据集以及其他生物活性预测任务上。在应用于较大数据集时，模型可以针对性的对不同任务作出优化，这样可以提高模型的泛化性能，提升模型稳定性。

2）本文提出的基于注意力的图卷积模型以及多特性融合方案，虽然可以有效提升模型的性能，但仍有可以改进之处。SMILES 字符串可以看作是一种文本序列，那么在使用注意力机制提取特征时，可以使用基于序列处理的算法如长短期记忆模型，来代替本文使用的卷积神经网络，并且可以与自注意力机制结合组成新的多特性融合方案，也就是未来可以从另一种角度来探究生物活性预测的模型性能提升。

参考文献

[1] Drews J. Drug discovery: a historical perspective[J]. Science, 2000, 287(5460): 1960-1964.

[2] Zhu H. Big data and artificial intelligence modeling for drug discovery[J]. Annual Review of Pharmacology and Toxicology, 2020, 60: 573-589.

[3] 余亚茹, 鲁鹏飞, 王红霞, 等. 中药防治新型冠状病毒肺炎概述[J]. 药学实践杂志, 2020, 38(3): 202-206.

[4] Tan L, Gu S, Wu C, et al. K-Means Clustering Method Based on Node Similarity in Traditional Chinese Medicine Efficacy[C]//2020 39th Chinese Control Conference (CCC). IEEE, 2020: 742-747.

[5] Devillers J. Neural networks in QSAR and drug design[M]. Academic Press, 1996.

[6] 李晓, 孔德信. 化合物成药性的预测方法[J]. 计算机与应用化学, 2012, 29(8):999-1003.

[7] 邓力, 俞栋著. 深度学习方法及应用[M]. 谢磊译. 北京: 机械工业出版社, 2015.

[8] Sun M, Zhao S, Gilvary C, et al. Graph convolutional networks for computational drug development and discovery[J]. Briefings in bioinformatics, 2020, 21(3): 919-935.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[10] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains[C]//Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. IEEE, 2005, 2: 729-734.

[11] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.

[12] Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions: moving beyond fingerprints[J]. Journal of computer-aided molecular design, 2016, 30(8): 595-608.

[13] Connor W Coley, Barzilay R, Green W H, et al. Convolutional embedding of attributed molecular graphs for physical property prediction[J]. Journal of chemical information and modeling, 2017, 57(8): 1757-1772.

[14] Pham T, Tran T, Venkatesh S. Graph memory networks for molecular activity prediction[C]//2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 639-644.

[15] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in neural information processing systems. 2016: 3844-3852.

[16] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[17] Yin X, Goudriaan J A N, Lantinga E A, et al. A flexible sigmoid function of determinate growth[J]. Annals of botany, 2003, 91(3): 361-371.

[18] Fan E. Extended tanh-function method and its applications to nonlinear equations[J]. Physics Letters A, 2000, 277(4-5): 212-218.

[19] Li Y, Yuan Y. Convergence analysis of two-layer neural networks with relu activation[C]. Long Beach, CA, USA: Advances in Neural Information Processing Systems. 2017: 597-607.

[20] Brown P F, Della Pietra V J, Desouza P V, et al. Class-based n-gram models of natural language[J]. Computational linguistics, 1992, 18(4): 467-480.

[21] Katritzky A R, Maran U, Lobanov V S, et al. Structurally diverse quantitative structure− property relationship correlations of technologically relevant physical properties[J]. Journal of chemical information and computer sciences, 2000, 40(1): 1-18.

[22] Tropsha A. Best practices for QSAR model development, validation, and exploitation[J]. Molecular informatics, 2010, 29(6-7): 476-488.

[23] Sardari S, Kohanzad H, Ghavami G. Artificial neural network modeling of antimycobacterial chemical space to introduce efficient descriptors employed for drug design[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 130: 151-158.

[24] Duvenaud D K, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints[C]//Advances in neural information processing systems. 2015: 2224-2232.

[25] Eguchi R, Ono N, Morita A H, et al. Classification of alkaloids according to the starting substances of their biosynthetic pathways using graph convolutional neural networks[J]. BMC bioinformatics, 2019, 20(1): 380.

[26] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.

[27] Heller S, McNaught A, Stein S, et al. InChI-the worldwide chemical structure identifier standard[J]. Journal of cheminformatics, 2013, 5(1): 1-9.

[28] Mauri A, Consonni V, Pavan M, et al. Dragon software: An easy approach to molecular descriptor calculations[J]. Match, 2006, 56(2): 237-248.

[29] Mauri A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints[M]//Ecotoxicological QSARs. Humana, New York, NY, 2020: 801-820.

[30] Frisch M J, Trucks G W, Schlegel H B, et al. Gaussian 03, Revision C. 02. Wallingford, CT: Gaussian[J]. Inc. [Google Scholar], 2004.

[31] Yap C W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints[J]. Journal of computational chemistry, 2011, 32(7): 1466-1474.

[32] O'Boyle N M, Banck M, James C A, et al. Open Babel: An open chemical toolbox[J]. Journal of cheminformatics, 2011, 3(1): 33.

[33] Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling[J]. 2013.

[34] Bolton E E, Wang Y, Thiessen P A, et al. PubChem: integrated platform of small molecules and biological activities[M]//Annual reports in computational chemistry. Elsevier, 2008, 4: 217-241.

[35] Dahl G E, Jaitly N, Salakhutdinov R. Multi-task neural networks for QSAR predictions[J]. arXiv preprint arXiv:1406.1231, 2014.

[36] Shang C, Liu Q, Chen K S, et al. Edge attention-based multi-relational graph convolutional networks[J]. arXiv, 2018: arXiv: 1802.04944.

[37] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[38] 杨善林, 倪志伟. 机器学习与智能决策支持系统[M]. 科学出版社, 2004.

[39] Mnih,Volodymyr,Heess,et al.Recurrent models of visual attention[J].arXiv preprint arXiv:1406-6247,2014.

[40] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[41] Chen X S, Kang Q, Zhou M C, et al. A novel under-sampling algorithm based on iterative-partitioning filters for imbalanced classification[C]//2016 IEEE International Conference on Automation Science and Engineering (CASE). IEEE, 2016: 490-494.

[42] Liu X Y, Wu J, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 39(2): 539-550.

[43] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[44] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008: 1322-1328.

[45] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[46] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[47] Li B, Liu Y, Wang X. Gradient harmonized single-stage detector[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8577-8584.

攻读硕士学位期间的科研成果

1. 论文：

Tan L, Gu S, Wu C, et al. K-Means Clustering Method Based on Node Similarity in Traditional Chinese Medicine Efficacy[C]//2020 39th Chinese Control Conference (CCC). IEEE, 2020: 742-747.

2. 专利：

一种基于多任务深度神经网络的中药功效预测方法(发明专利实审状态,专利号：202010141041 .8）第一作者.

一种基于节点相似度的 Kmeans 中药材功效聚类方法（发明专利实审状态，专利号：202010140751 .9）第一作者.

致谢

光阴荏苒，两年半的硕士生涯转瞬即逝。回望过去的这日日夜夜，有过开心、有过忧虑，但更多的是成长。值此毕业论文完成之际，对那些曾经帮助过我的老师及同学表示最衷心的感谢。

首先，要感谢我的两位导师--周银座老师和黄剑平老师。在过去的两年里，两位老师给予了我受益一生的帮助。两位老师拥有极强的专业素养、严谨的科学态度以及对学术研究的敏锐洞察力，您们的这些品质潜移默化的影响着我。在科研路上，两位老师给了我莫大的帮助。从开始选题到最后的论文撰写，每一过程无不精心讲解，对科研细节精益求精，解决我遇到的每一个问题，让我在科研道路上硕果累累。真的很荣幸在攻读硕士期间，得到这样两位老师的教导。在此真心地向两位老师表示我最诚挚的敬意与感谢。

特别感谢校外指导导师周杰老师，在前期的中药功效相关研究中，给予了我新颖的科研思路和专业的学术指导，非常感谢您的无私帮助和悉心指导。

接下来，要感谢实验室已毕业的崔智颖、程亮、张亚东、李智猛等师兄师姐，你们的优秀是督促我前行的动力，也感谢你们在我困惑时对我的引导及支持。感谢实验室的丰仕琦、吴银豪和李达等每一位伙伴，正是由于有你们的支持与陪伴，这两年多才变得更加有意义。在此祝愿我们都前程似锦！

感谢我的室友马佳秀、刘霜霜、高嘉利，感谢你们在我生活上的关心与帮助，你们是我最坚强的后盾，给我的硕士生活带来了无数快乐。祝愿我们会越来越好！

然后，还要感谢我的父母、弟弟们，特别感谢父母的养育与教育之恩，感谢你们给予我无条件的爱。还要感谢我的朋友张鑫鑫，这四年一路走来的互相信任、互相支持、互相理解。对于家人的感激无以言表，唯有用自己的努力和行动来回报。

未来人生还长，我会谨记老师们的谆谆教诲，带着身边伙伴的美好祝福，继续奋斗，勇敢向前，成为更好的自己。