基于图注意力网络的环状 RNA 与疾病关联关系预测

张瀚元[1],赵博伟[1],胡　伦[1*],王　磊[2*],尤著宏[3]
(1. 中国科学院大学 中国科学院新疆理化技术研究所,新疆 乌鲁木齐 830011;
2. 广西科学院 大数据与智能计算研究中心,广西 南宁 530007;
3. 西北工业大学 计算机学院 大数据存储与管理工业和信息化部重点实验室,陕西 西安 710072)

摘　要:环状 RNA 是一种具有环状结构并且表达水平与多种疾病有关的非编码 RNA 分子,挖掘环状 RNA 与疾病之间的内在关联关系在生命医学研究中具有重要意义。基于图注意力机制,该文提出了一种由图注意力网络(GAT)、编码器-解码器(AE)和全连接神经网络(DNN)结构组合的端到端深度学习模型 GATECDA 来预测潜在的环状 RNA 与疾病的关联关系。在包含 739 个关系的 CircR2Disease 数据集上,GATECDA 模型五折交叉验证实验取得了 ROC 曲线下面积 AUC 为 0.961 8,AUPR 为 0.903 2,衡量在非平衡数据上性能 MCC 指标达到了 0.757 6 的优异结果,综合性能在同领域预测模型中表现出色。表明基于深度学习图表示学习的策略有助于提升环状 RNA 与疾病关联关系预测模型的综合性能,同时端到端的学习模型更易于训练与泛化到其他问题中。在预测的结果得到的前 30 个环状 RNA 与疾病的关联关系中,有 25 个在最近医学文献中有支持。表明人工智能方法可以为医学研究筛选与疾病相关的标志物提供新的角度。

关键词:环状 RNA/CircRNA;疾病;关联关系预测;图注意力网络;深度学习
中图分类号:TP399　　**文献标识码**:A　　**文章编号**:1673-629X(2023)11-0126-09
doi:10.3969/j.issn.1673-629X.2023.11.019

Prediction of Circ RNA-Disease Associations Based on Graph Attention Networks

ZHANG Han-yuan[1],ZHAO Bo-wei[1],HU Lun[1*],WANG Lei[2*],YOU Zhu-hong[3]
(1. Xinjiang Technical Institute of Physics and Chemistry,Chinese Academy of Sciences,Urumqi 830011,China;
2. Big Data and Intelligent Computing Research Center,Guangxi Academy of Sciences,Nanning 530007,China;
3. MIIT Key Laboratory of Big Data Storage and Management,School of Computer Science,
Northernwestern Polytechnic University,Xi'an 710072,China)

Abstract:Circular RNA (CircRNA) is a kind of expressed RNA transcript with loop structure and its expressed level related to other diseases. It is of great significance to explore the internal correlation between CircRNA and Disease in life medicine research. Based on the graph attention mechanism,GATECDA,an end-to-end deep learning model consisting of graph attention network (GAT),AutoEncoder (AE) and deep neural network (DNN),is proposed to predict the candidate associations between CircRNA and Disease. It achieved 5-fold cross-validation on AUC at 0.961 8 and AUPR at 0.903 2,MCC index at 0.757 6 on CircR2Disease data set including 739 associations between CircRNA and Disease. The measurement result means the model performed well on the imbalanced benchmark. Hereby,we believed the strategy by integrating graph attention network embedding into the deep learning model would improve the performance of prediction CircRNA-Disease association. At top 30 of the predicted association of CircRNA and Disease,we retrieved 25 of them with published paper supporting. As we thought that the AI tech. would boost the work of discovering biomarkers related with disease.

Key words:Circular RNA/CircRNA;disease;association prediction;graph attention networks;deep learning

收稿日期:2022-11-25　　**修回日期**:2023-03-28
基金项目:新疆科技厅青年自然基金(2019D01C212);科技部国家 2030-"新一代人工智能"重大项目(2018AAA0100100)
作者简介:张瀚元(1988-),男,讲师,博士研究生,CCF 会员((L7826M),研究方向为生物信息、图神经网络、机器学习、智能信息处理;通讯作者:胡　伦(1985-),男,研究员,博士,研究方向为机器学习、复杂网络分析、图神经网络、生物信息;通讯作者:王　磊(1982-),男,研究员,博士,研究方向为机器学习、生物数据挖掘。

0　引　言

环状 RNA 是一类收尾相连具有环状结构的转录 RNA,它产生于 DNA 转录过程或转录后的修饰[1-2],具体的产生机制还在研究中。虽然细胞内的 RNA 主要是以线性结构为主,但环状 RNA 也大量存在,并且发现环状 RNA 往往会高表达转录。近年来随着高通量测序技术的发展,环状 RNA 能够通过被反向比对的双端(two-paired)短序列识别出。数据分析表明,它们在癌症等多种复杂疾病组织与正常组织的比对中有显著的转录差异,这些有差异的环状 RNA 被认为与该疾病发生和发展有关系[3]。比如,Hsa_circ_0046430 在最近研究中参与 miR-6785-5p/SRCIN1 的 ceRNA 调控网络促进结肠癌的生长[4],CircRNA DDX21 则参与 miR-1264/QKI 的 ceRNA 调控网络以弱化三阴性乳腺癌的生长[5],而利用环状 RNA 基因表达数据则可以挖掘出新的胃癌标志物[6]等等。然而,通过实验手段发现的环状 RNA 与疾病的关联关系毕竟有限,研究人员希望通过现有的研究信息和生物知识,借助机器学习和人工智能的方法,预测和挖掘环状 RNA 与疾病的关联关系[7]。

1　研究背景

建立生物基因型与表型性状的关联关系一直是生命科学研究的重要问题[8]。研究人员已经通过计算手段来挖掘这种关联关系,如小 RNA(microRNA)与疾病[9]、非编码 RNA(LncRNA)与疾病[10]、环状 RNA(Circular RNA, CircRNA)与疾病[11-13]的关联关系。由于已有知识的局限,以及不同生物分子对应的疾病特征不同,目前多数有效的环状 RNA 与疾病的预测方法是通过链路预测(Link prediction)对已知的环状 RNA 与疾病关系的补全,关联关系(Association prediction)预测可以认为是链路预测的一种特例[14]。主要关于链路预测的方法都有尝试在环状 RNA 与疾病关系预测问题上进行研究,比如 KATZHCDA 方法通过 KATZH 图信息指标对环状 RNA 与疾病的关系进行预测。KATZH 指标是一种通过节点间链路个数来衡量节点间关系程度并用于链路关系的预测[15]。iCircDA-MF 通过矩阵分解的方法整合环状 RNA 与疾病的相关信息进行链路预测[16]。也有通过深度学习模型构建分类器进行相关关系的预测,如 MSFCNN 方法通过融合多源信息后利用两层卷积网络进行关系预测[17]。GIS-CDA 也是一种采用了图注意力机制的模型,但主要是利用数据融合的技术和归纳式矩阵补全[12]。以上关于图链路预测的传统方法都有应用在环状 RNA 与疾病关系的预测中。AE-DNN 方法通过构建编码器(AutoEncoder)和深度神经网络(Deep Neural Network)进行关系预测[18]。AANE-SAE[19]利用属性网络编码算法(AANE)获得浅层特征,并利用堆叠的自动编码器(SAE)获得深层特征,最后利用 XGboost 分类器进行预测。一般来说利用信息指标进行链路预测只局限于部分结构,无法利用到全面的图结构信息。单纯利用传统的机器学习模型虽然也能取得较好的训练效果,但是在验证中相对来说具有较高的假阳性率,不利于生物实验的验证。矩阵分解方法的结果假阳性率低,但是偏重于已有知识的强化,发现新知识的能力较弱。

为了能够提高预测的能力,就需要引入更多生物知识及其关系网络来提取特征信息,比如构建生物知识的异构网络等[20]。随着近年来图表示学习(graph represent learning)算法的发展,图表示学习在人类社会网络链路预测的相关问题上取得了较好的结果[21]。一些图表示学习方法被用于环状 RNA 与疾病关联关系的预测,如 Lei 通过随机游走算法实现特征的提取,并利用 K 邻接聚类的方法实现环状 RNA 与疾病关联关系的预测[22];本课题组发表的 iGRLCDA 通过因子图卷积网络(factor Graph Convolution Network)在异构图上提取特征[23],利用随机森林分类器实现环状 RNA 与疾病关联关系的预测,取得了较好的结果。理论上,图卷积网络也可以直接做链路预测[14],但是不容易训练成功。考虑到环状 RNA 与疾病的关系中大部分关系未知,所以 iGRLCDA 利用因子图卷积网络在主要的图结构上对节点分类(node classification)。依据节点分类模型提取出所有节点的特征,最后依据分类器实现链路关系预测。在 iGRLCDA 的设计过程中,发现对传统机器学习方法进行调优的过程比较费时且需要一定技巧,希望设计一种具有自适应且综合性能良好的模型来实现环状 RNA 与疾病关联关系的预测。深度学习模型无疑具有较好的自适应性,但目前对于环状 RNA 与疾病关系预测深度学习 AE-DNN 模型[18]部分性能并不出众,反映非平衡数据性能的 MCC 指标为 0.58,低于 iGRLCDA[23]的 0.714 6。此外,在验证集上 AE-DNN 模型的 AUC 为 0.85,也低于 iGRLCDA[23]的 0.928 7。在实现自动编码器(AutoEncoder, AE)与深度全连接神经网络(Deep Neural Network, DNN)的基础上,嵌入图注意力机制(Graph Attention Network, GAT)[24],实现了 GAT-AE-DNN 结构的端到端的深度学习模型 GATECDA,在环状 RNA 与疾病预测的 CircR2Disease 数据集中[25],其综合性能 AUC 得分为 0.961 8,MCC 关系为 0.757 6。GATECDA 采用端到端的 GAT-AE-DNN 深度学习模型,具有自适应性、易于泛化和拓展等特点,训练过程也更容易。

2 基于图表示学习方法的预测

基于图表示学习方法进行特征提取并预测关联关系的基础在于从图中学习相应的知识并将图结构信息融合入图中节点的特征。相较于传统上只利用节点内部的信息,图表示学习可以利用节点有联系的不同节点的特征来强化自身以反映与相关节点的联系。以环状 RNA 参与的 ceRNA 调控网络为例,如果只考虑其自身的序列信息,那么可能在表示中无法反映出环状 RNA 通过吸附 miRNA 来调节 LncRNA 的关系。但利用图表示学习方法提取特征后,所提取的特征来源于环状 RNA 自身,但也能把现有的调控关系反映出来。

目前,主要的图表示学习方法有矩阵分解的方法、随机游走的方法、图神经网络的方法等。其中图注意力网络(Graph attention networks,GATs)也是图神经网络中一种主要的方法[21,24],在多个同质数据集上的链路预测中取得了较好的性能。研究中首先建立异构的环状 RNA 与疾病关系的网络。所谓异构是因为环状 RNA 或疾病在各自向量空间内存在关系图,如图 1 所示,需要在不同向量空间表述的节点关系中挖掘关联关系。比如关系图 $G=(u,v)$,其中的 u 与 v 分别表示不同类型的节点,它们各自在自身的向量空间存在不同的维度 u_feature 和 v_feature。已经知道部分 u 与 v 之间存在联系,因此构成了异构关系图。图表示学习方法实质就是在考虑异构关系图 G 的结构上把 u_feature 和 v_feature 映射到同一个空间成为 node_feature,该 node_feature 可以区分整体关系图 G 中不同节点的类别。

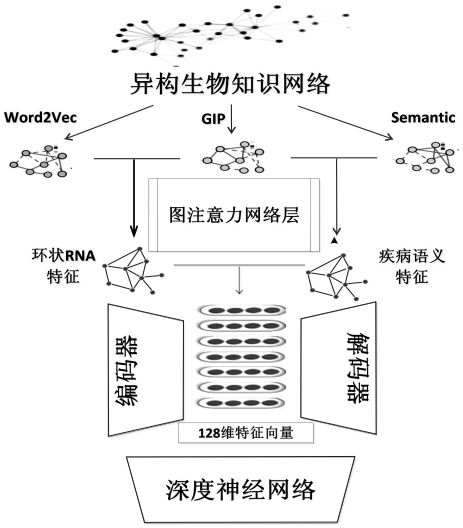

图 1 GATECDA 环状 RNA 与疾病关联关系预测模型流程

随后,u 与 v 之间的已知关系(u,v)->R 为预测的正样本集,随机产生的关系(u,v)->R^* 为预测的负样本,正负样本具有相同的大小 N($N=739$)并一同作为大小为 $2N$ 的训练集。在训练集上采用五折交叉验证。此外,为了验证不同模型的性能,从训练集中拿出 n($n=50$)个关系作为验证集。最后,将提取的节点特征联系起来利用分类器进行预测。图 1 展示了 GATECDA 的整体流程,从异构生物知识中获得环状 RNA 与疾病的特征,并用深度模型预测关联关系。

3 实验结果与分析

3.1 实验环境及参数设置

研究工作在一台双路 Intel 至强 2365V2 处理器的工作站上实现,内存为 96 GB。在实现过程中,实际使用内存不超过 16 GB,主要在属性节点的特征提前上花费较多。GATECDA 模型采用 python 3.7 语言实现,模型利用 tensorflow 2.7 张量流计算框架和 keras 深度学习框架封装构建,GAT 层的实现采用了 dgl 图神经网络工具包。

3.2 数据集

考虑通过环状 RNA 的序列信息相似性,疾病关系的语义信息相似性和由已知的环状 RNA 与疾病关系信息相似性来建立异构网络。其中,环状 RNA 序列信息源自 circBase[26] 数据库中基于 hg19 基因组的推测的环状 RNA 选择性剪切序列。疾病关系的语义信息采用引用字典 Mesh 的关系获得[27]。环状 RNA 与疾病关系信息由 CircR2Disease 数据库[25] 中经过实验验证的关系获得。部分因数据库环状 RNA 的 id 对应不上的序列也可以由 CircR2Disease 数据库[25] 提供的基因组位置或对应的基因 Symbol 获得。一共获得 739 个环状 RNA 与疾病关系作为正样本集,涉及到 661 个环状 RNA 和 100 种疾病。在这个关系中,还存在 65 261个未标注的环状 RNA 与疾病的随机关系,随机从里面取得 739 个作为负样本集。最后从 1 478 个正负样本关系中取出 50 个关系作为验证集,剩余的 1 428个关系作为训练集。

3.3 环状 RNA 与疾病的特征提取

根据获得的数据信息,可以构建三组节点间相似关系信息,包括环状 RNA 与疾病、环状 RNA 与环状 RNA、疾病与疾病。

(1)环状 RNA 与疾病关联:所有从 CircR2Disease[25] 的 739 个环状 RNA 与疾病关系,涉及到 661 个环状 RNA 和 100 种疾病,可以构成 661×100 的关系矩阵 RD,其中有关系为 1,否则为 0。从该关系矩阵就可以通过 Gaussian Interaction Profile(GIP)方法获得单个环状 RNA 或疾病的特征向量。GIP 方法也是药物与疾病关系等预测中常使用的方法[28],可以通过函数 SE($p(i)$,$p(j)$) 从关系矩阵中两个表示为 0-1 向量 $V(p)$ 获得节点 i 与 j 的相似性,如公式(1)。

$$SE(p(i),p(j)) = \exp(-\Theta V(p(i)) - V(p(j))^2) \quad (1)$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} V(p(i))^2 \quad (2)$$

其中, $V(p(i)) - V(p(j))$ 表示两个 0-1 向量间的差异,通过 L2 范式获得差异的距离,乘以归一化因子 θ 后获得 e 指数的幂。最后,通过幂指数函数 SE 可以获得 0-1 关系矩阵 RD 中任意两个节点间的相似性,进而原来稀疏的 0-1 关系矩阵就转化为稠密关系。其中环状 RNA 或疾病可以获得 761 个维度的特征。

(2)环状 RNA 与环状 RNA 相似性:可以获得环状 RNA 的序列信息,并通过序列相似性获得环状 RNA 与环状 RNA 的 661×661 的相似矩阵 CC。由此,可生成单个环状 RNA 的特征向量。这里的环状 RNA 的相似性由 skip-gram 结构的 word2vec 生成[29]。因为 RNA 序列结构的复杂性,RNA 序列的作用区域可能局限于内部的短序列片段中,直接获取两条 RNA 序列的相似性不能反映它们相互作用的关系[30]。word2vec 模型在自然预言处理中广泛使用,它通过一个单词在上下文中的出现关系来挖掘其特征表示。在生物序列的挖掘中,定义 6-mer,如"ACCATC"为一个单词 w。

$$\arg\max_{\Theta}\prod_{S\in T}\prod_{w\in S} P(w \vee S,\Theta) \quad (3)$$

word2vec 在该任务中是寻找参数 Θ 使得所有属于语料 T 中每个句子 S 内单词 W 的联合概率乘积最大。在训练中语料 T 包括 13 000 条环状 RNA 序列。

$$P(w \vee S,\Theta) = \frac{\theta(w)\sum_{ws\in S}\theta(ws) - 1}{\sum_{ws\in S}\theta(ws)\sum_{ws\in S}\theta(ws) - \mathrm{Num}(w,S)} \quad (4)$$

定义某一个 6-mer 的单词 W 可以表示为 128 维的特征向量 θ。如果两个单词在语义上近似,那么它们的特征乘积 θ 也会更大。在一个句子 S 内所有单词 W 的累加和可以表示为该句子的特征 $\sum_{ws\in S}\theta(ws)$。那么所有句子与句子间的关系可以表示为 $\sum_{ws\in S}\theta(ws)\sum_{ws\in S}\theta(ws) - \mathrm{Num}(w,S)$,可以理解为背景的语料 T 的特征。一个单词与其所在句子间的关系为 $\theta(w) * \sum_{ws\in S}\theta(ws) - 1$。在公式(4)中,希望让每个单词在考虑所有语料关系 T 后,其在序列 S 中的作用最大化。

(3)疾病与疾病相似性:建立疾病与疾病 100×100 的相似关系,就可以获得单个疾病 100 个维度的特征信息。疾病与疾病的相似关系源自 MeSH 数据库。作为医学引用词典,MeSH 数据库通过分析大量医学论文的引用关系提供了医学主题词关系。利用医学主题词关系,基于王等人[31]发表的方法,可以获得关于疾病间的相似关系。医学主题词关系构建了有向无环图(DAG)。可以记某一疾病 d 参与的 $DAG(d) = (d,N(d),E(d))$,其中 $N(d)$ 表示与某一疾病相关的所有节点,包括疾病或者症状; $E(d)$ 表示与之涉及的所有边。如果在 $DAG(d)$ 中还存在另一疾病 s,那么可以通过如下公式计算疾病 d 与疾病 s 的关系:

$$\begin{cases} D_d(s) = 1, \text{if } s = d \\ \max\{\mu \cdot D_d(s) \mid D_d(s) \in \text{childrenodes}\}, \text{if } s \neq d \end{cases} \quad (5)$$

在公式(5)中如果疾病 d 与疾病 s 关联,那么它们的关系为 1,否则找出疾病 d 到疾病 s 所有共同关联的子节点数量,作为它们之间的关系。在复杂疾病中,疾病 d 的影响力为所有与之有关疾病的关系的累加和,定义如下:

$$DI(d) = \sum_{s\in N_d} D_d(s) \quad (6)$$

有了以上(6)的信息,可以定义两个疾病间的互信息 SS_1:

$$SS_1(d(i),d(j)) = \frac{\sum_{s\in N_{d(i)}\cap N_{d(j)}}(D_{d(i)}(s) + D_{d(j)}(s))}{DI(d(i)) + DI(d(j))} \quad (7)$$

在公式(7)中,两两疾病间的相似关系可以理解为与它们相关所有节点的关系除以两个疾病的整体影响。但是有些疾病可能影响的节点少,但它却很重要,于是设计了另一个指标 $DC_d(s)$:

$$DC_d(s) = -\log\left(\frac{\mathrm{num}(\mathrm{contain}(DAG(d),s))}{\mathrm{num}(\mathrm{diseases})}\right) \quad (8)$$

其中, $\mathrm{num}(\mathrm{contain}(DAG(d),s))$ 表示 $DAG(d)$ 图中包含疾病 s 的数量, $\mathrm{num}(\mathrm{diseases})$ 表示所有的疾病。这样关联数量少的疾病 DC 的分就越高。于是,有了第二个衡量疾病关系的互信息 SS_2:

$$SS_2(d(i),d(j)) = \frac{\sum_{s\in N_{d(i)}\cap N_{d(j)}}(DC_{d(i)}(s) + DC_{d(j)}(s))}{DI(d(i)) + DI(d(j))} \quad (9)$$

最后,将 SS_1 与 SS_2 共同考虑得到 $SS = 0.5 * SS_1 + 0.5 * SS_2$,作为最后疾病之间的语义相似关系。

3.4　GATECDA 模型的实现

在 GATECDA 的实现如图 2 所示。首先,构建了环状 RNA 与疾病的初始特征,计算环状 RNA 与疾病之间关联关系的相似性,疾病的语义相似性和环状 RNA 的序列相似性。其次,GATECDA 加入了图注意力网络(Graph attention networks, GATs)提取环状 RNA 与疾病异质关系图中的特征表示。最后,将得到

的环状 RNA 与疾病的特征表示送入 AE-DNN 深度学习模型进行关系预测,其中包含了自动编码器(AutoEncoder, AE)和深度神经网络(Deep Neural Network,DNN)。笔者认为 GAT 起到了特征提取与融合的作用,AE 起到了特征降维的作用,DNN 起到了分类器的作用。单层图注意力网络 GAT 也是由数个神经元组成的单元,一般不超过三层,比图卷积网络更容易达到训练效果[24]。相比图卷积网络是一种浅层的神经网络结构,因为本身属于神经网络,所以可以嵌入到深度学习模型中。

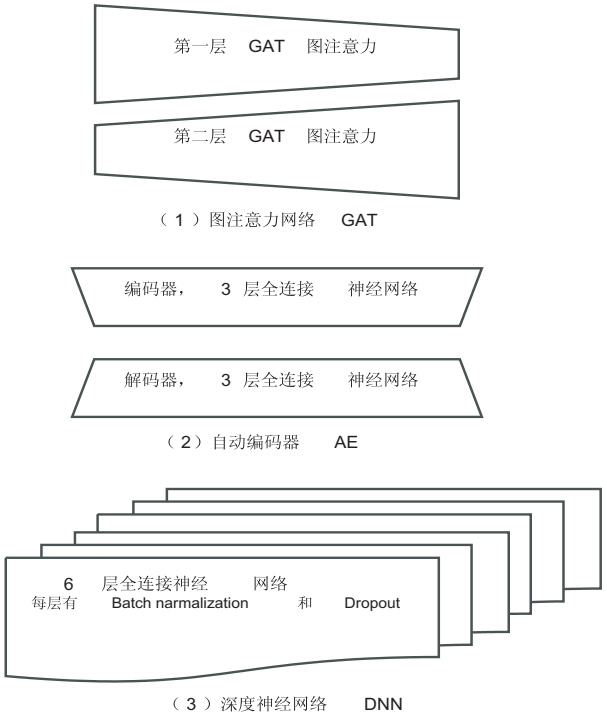

图 2 GATECDA 模型深度学习模型的结构

模型首先接受生物知识图 G 及其节点特征。图 G 可以认为是一个 $M * N$ 的二部图(bipartie graph)。M 可以认为是所有的环状 RNA,而 N 为疾病,同时 M 和 N 各自的特征也被作为参数。图注意力网络在接受数据后完成了以下工作:

$$W_{\text{updatenode}} = [\text{sigmoid}(X * [F(j), F(i)])] \tag{10}$$

$$W_{\text{updateall}} = \sum_{n=1}^{j} W_{\text{updatenode}} \tag{11}$$

$$\alpha = \text{LeakReLU}\left(\frac{W_{\text{updatenode}}}{W_{\text{updateall}}}\right) \tag{12}$$

$$F^*(i) = \text{LeakyReLU}(\alpha * F(i)) \tag{13}$$

其中,j 表示 i 节点的所有邻接节点。$W_{\text{updatenode}}$ 构成了输入层的神经网络,$X * [F(j), F(i)]$ 为该层输入的数据,其中 X 为自定义特征矩阵,$[F(j), F(i)]$ 表示 i 和 j 的联合特征向量。在学习一遍所有节点后,希望单个节点更新后在整体中起到最大作用,这里用 α 体现特征的更新,F^* 是更新后的特征。此外,作为一种随机过程,每更新一轮被认为是 1 个头(head)的注意力,更新 k 次为多个头(k-heads)的注意力,在 GATECDA 中 k 为 8。最后,所有 1 至 k 次的特征更新都被均方和作为最后的特征,如公式(14):

$$F^{(i)} = \frac{1}{n}\sum_{n=1}^{k} * F^k(i) \tag{14}$$

注意力的思想与 word2vec 一致,就是每个节点都朝着在整体背景中最显著去改变。而多头的概念与主成分分析(PCA)的概念相似。所以认为多头注意力网络起到了特征提取与融合的作用。随后的 AE-DNN 模型由自动编码器(AutoEncoder,AE)和深度神经网络 DNN(Deep Neural Network)组成,是深度学习中的经典模型,在很多机器翻译任务中都有较为出色的表现。AE 层接受稀疏的数据,在不断收窄的多层网络中实现信息的融合、压缩与标准化,之后又以多层变宽的网络压缩后的数据还原回输入数据。AE 具有降维的作用,在 GATECDA 中,如图 2(2)把两层 GAT 网络得到的 1 522 维的特征压缩为 128 维的特征。经过 AE 处理过的数据又被送入深度神经网络 6 层神经网络构建的 DNN 进行关联关系的分类预测,如图 2(3)。在所有的 AE-DNN 层中,都使用了 Batch normalization 和 dropout 机制。Batch normalization 是一种归一化方法,可以减小异常数据的干扰。dropout 机制是在每一层反馈梯度时,只更新一定比例的神经元,该模型训练时 dropout 的值为 0.3。Batch normalization 和 dropout 机制都是为了防止模型过拟合,提高模型泛化能力。

3.5 评估指标

在取得对预测结果评估矩阵的真阳性率(True Positive,TP)、真阴性率(True Negative,TN)、假阳性率(False Positive,FP)、假阴性率(False Negative,FN)后,采用了准确率(Acc.)、敏感度(Sen.)、精准率(Pre.)、F1 打分(F1)和 Matthews 关系(MCC)来较全面地评估模型的性能,这些也是机器学习领域的主流评价方法,如下:

$$\text{Acc.} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{15}$$

$$\text{Sen.} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{Pre.} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{18}$$

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{19}$$

在五折交叉验证的测试下,衡量受试者工作特征曲线(ROC)下面积(AUC)也是机器学习领域里衡量模型性能的主要指标。通过模型在逐一增长的测试集上预测结果真阳性率(TPR)与假阳性率(FPR)的平面坐标位置,就可以做出 ROC 曲线。

3.6　模型能力评估

为了评估 GATECDA 模型的能力,在 CircR2Disease 数据集上进行五折交叉验证,即将训练集划分为 5 等份,进行五次训练。每次以其中四份进行训练,一份进行测试(285 个样本)。图 3 展示了 GATECDA 模型的训练过程的 ROC 曲线及 AUC 值。GATECDA 模型的五折交叉验证平均 AUC 值为0.961 8,每次的 AUC 值分别为 0.947 6,0.952 0,0.963 7 和 0.979 5。其综合性能在表 1 中体现,平均准确率为 87.53%,敏感度为 93.62%,精准度为 83.80%,F1 打分为 88.35%,MCC 关系为 0.757 6,精准度-召回曲线下面积 AUPRC 为 0.903 2,ROC 曲线下面积 AUC 为 0.961 8。

表 1　GATECDA 在 CircR2Disease 数据集上五折交叉验证

fold	Acc./%	Sen./%	Pre./%	MCC	F1/%	AUPRC	AUC
1	88.46	91.95	86.71	0.769 8	89.25	0.914 3	0.947 6
2	87.06	92.70	82.47	0.747 4	87.29	0.893 3	0.952 0
3	89.86	92.67	88.54	0.797 2	90.55	0.925 2	0.967 3
4	81.75	94.33	75.14	0.657 2	83.65	0.861 4	0.963 7
5	90.53	96.48	86.16	0.816 4	91.03	0.922 0	0.979 5
平均	87.53	93.62	83.80	0.757 6	88.35	0.903 2	0.961 8
+标准差	+3.49	+1.82	+5.32	+0.062	+3.00	+0.026	+0.013

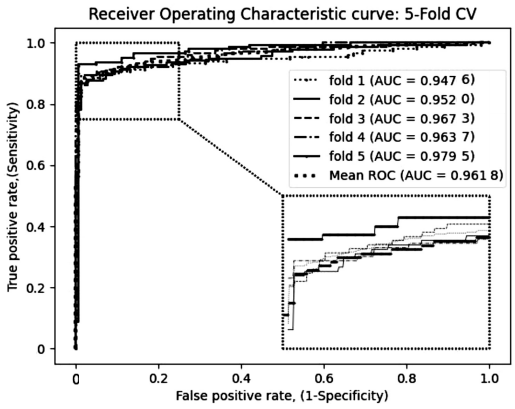

图 3　GATECDA 模型在 CircR2Disease 数据集生成的 ROC 曲线

3.7　不同预测模型比较

比较了已经发表的环状 RNA 与疾病关联关系预测的几种方法在 CircR2Disease[25] 数据集上五折交叉验证中的 AUC 值,见表 2。它们包括基于图表示学习方法 GATECDA、iGRLCD[23] 和 GIS-CDA[12],深度学习模型 AE-DNN[18] 与 AANE-SAE[19],以上模型在文中研究背景中均有介绍。通过比较可以看出,GATECDA 在五折交叉验证中平均的 AUC 为 0.961 8,高于 iGRLCDA[23] 的 0.928 7 和 AE-DNN[18] 的 0.930 3。对于衡量不平衡数据集上性能的 MCC 值,GATECDA 的 0.757 6,也高于 AE-DNN 的 0.583 6 和 iGRLCDA 模型的 0.714 6。其中 GIS-CDA 与 GATECDA 模型都采用了图注意力机制,不过 GIS-CDA 是先用编码器融合不同维度的特征后再使用图注意力机制,GATECDA 模型首先使用图注意力机制而不是进行编码的信息融合,因而比 GIS-CDA 模型的 AUC 略高。笔者认为在设计异构网络模型时,越能完整和直接地利用图结构信息,越有利于模型的预测。GATECDA 不足在于实现的图注意力机制(CAT)是一种浅学习[14,24],对于以后更大规模数据集或知识图谱上能力提升空间不如图卷积网络(GCN)模型[21]

表 2　不同预测模型的比较

指标	图表示学习			深度学习	
	GATECDA	iGRLCDA	GIS-CDA	AE-DNN	AANE-SAE
AUC	0.961 8	0.928 7	0.930 3	0.939 2	0.88
MCC	0.757 6	0.714 6	N/A	0.583 6	N/A
F1/%	88.35	85.11	N/A	59.52	N/A

3.8　不同分类器比较

比较 GATECDA 和不同分类器模型在验证集上的预测能力。其中 KNN、RF、XGboost 和 SVM 为 iLearnPlus 工具[32] 封装好的分类器。GATECDA 是该文提出的端到端图注意力网络、自动编码器与深度神经网络结合的深度学习模型(GAT_AE_DNN),其中

AE 是自动编码器加输出层的分类器,DNN 是深度神经网络分类器。SVM 是支持向量机(Support Vector Machine),KNN 是 K 邻接分类器(K - nearest Neighbor),RF 是随机森林分类器(Random Forest),XGboost 是极限学习分类器(Extreme Gradient boost)。以上所有模型都在 1 428 个正负关系构成的训练集上加以训练,并在独立划分出的 50 个样本的验证集上做性能比较。从图 4 中可以看出,在验证集样本上 GATECDA 的 AUC 最高为 0.972 6, XGboost 的 AUC 值为 0.895 0,KNN 为 0.733 3,RF 为 0.640 8, SVM 为 0.667 2。

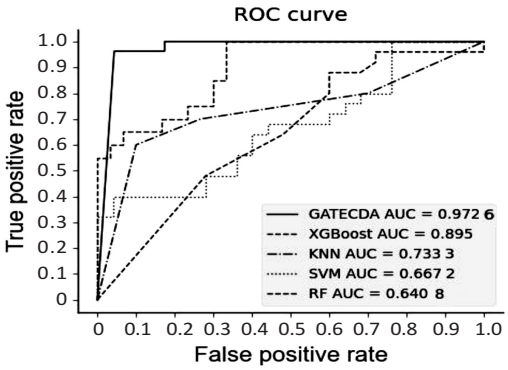

图 4 不同分类器模型在验证集上的 ROC 曲线

3.9 特征消融实验

为了分析图结构的已有知识信息与节点属性信息对模型能力的贡献,设计了特征消融实验,见表 3。研究中,GATECDA 模型既使用已有知识构建图 G,也采用节点属性特征,得到的预测结果 AUC 为 0.961 8, AUPR 为 0.903 2。GATECDA-F是GATECDA模型只包含图结构信息,得到的预测结果 AUC 为 0.582 7, AUPR 为 0.785 7。GATECDA-G 是 GATECDA 模型只包含节点属性特征,得到的预测结果 AUC 为 0.491 5,AUPR 为 0.732 8。最后为该结果符合预期,图注意力网络在考虑图结构和节点属性特征时可以强化特征信息。

表 3 特征消融实验

特征组合模型	AUC	AUPR	Acc./%
GATECDA	0.961 8	0.903 2	87.53
GATECDA-F	0.582 7	0.785 7	59.35
GATECDA-G	0.491 5	0.732 8	47.36

4 案例研究

通过 GATECDA 从 661 个环状 RNA 和 100 种疾病的 65 261 个未标注潜在组合中预测 3 743 个关联关系,约占未标注总数的 5.7%。表 4 列出预测结果排名前 30 的关联关系,并且通过文献检索查到相关 CircRNA 或其所在基因在以前的生物实验中有发现与相关疾病存在联系。在预测的结果得到的前 30 个环状 RNA 与疾病的关联关系中,其中有 25 个关联能够在最近医学文献中被发现存在关联。预测结果可以帮助研究人员缩小筛查范围,尽快找到与疾病相关的关键标志物。实验中获得的差异信息很多,一般的方法是做富集分析或是在基因共表达网络寻找关键基因。如果结合已有知识对环状 RNA 与疾病的关联关系预测可以为寻找关键基因和疾病标志物提供一种新的角度。

表 4 预测排名前 30 个环状 RNA 与疾病的关系及文献检索

Rank	CircRNA_ID	Disease_Name	Gene_Symbol	Pubmed ID
1	mmu_circ_0000375	Colorectal cancer	Hectd1	35611198
2	circ0817/hsa_circ_0024169	Breast cancer	CUL5	14641918
3	hsa_circ_0003146	Pancreatic cancer	EHD2	23283488
4	hsa_circRNA 102049	Colorectal cancer	TADA2A	32799891
5	hsa_circ_0003707	Gastric cancer	CD44	28639908
6	mmu_circRNA_30664	Colorectal cancer	C3	33765255
7	hsa_circ_0020397	Breast cancer	DOCK1	34771489
8	hsa_circ_0084021	Liver cancer	PLEKHA2	N/A
9	circHLA-C	Gastric cancer	HLA-C	19883394
10	rno_circRNA_006508	Lung cancer	LOC681740	N/A
11	hsa_circ_0000893	Breast cancer	DHPS	28744405
12	circRNA_010567	Breast cancer	N/A	N/A
13	circETFA	Breast cancer	ETFA	29221160
14	hsa_circ_0026143	Cervical cancer	TROAP	35722431
15	hsa_circ_0020397	Bladder cancer	DOCK1	30983072
16	hsa_circ_0001721	Breast cancer	CDK14	36103813
17	hsa_circRNA_100918	Breast cancer	PICALM	30979686

续表 4

Rank	CircRNA_ID	Disease_Name	Gene_Symbol	Pubmed ID
18	hsa_circ_0006404	Colorectal cancer	FOXO3	34549306
19	chr11:34060062|34073206	Pancreatic cancer	MORC3	34839357
20	hsa_circRNA_102101	Colorectal cancer	CDC27	28900500
21	hsa_circ_0029067	Triple negative breast cancer	CLIP1	32194644
22	hsa_circ_0009910	Lung cancer	MFN2	35582383
23	hsa_circ_0000554	Esophageal cancer	LIN52	33470617
24	hsa_circ_0000069	Colorectal cancer	STIL	28003761
25	hsa_circ_0089974	Esophageal cancer	NHS	27465405
26	hsa_circRNA_100258	Lung cancer	N/A	N/A
27	hsa_circ_0007006	Gastric cancer	DYM	N/A
28	hsa_circ_0005529	Gastric cancer	VPS33B	31105852
29	circRNA0003906	Cervical cancer	ZNRD1-AS1	26328261
30	hsa_circ_0037911	Colorectal cancer	GSPT1	33819920

5　挑战与发展

笔者认为,目前采用图表示学习提取特征进行环状 RNA 与疾病关联关系预测的方法比其他方法能获得较好的综合性能。针对目前取得的进展,一方面需要利用更丰富的生物网络知识,即利用复杂异构网络实现对任意环状 RNA 与疾病的预测,同时保持验证中较低的假阳性率。从这一点上看,GATECDA 的基础在于已有知识的补全,因而更适合于降低假阳性率的新知识的挖掘。另一方面,研究环状 RNA 与疾病关系的预测最初也是想实现环状 RNA、调控分子、生物过程、生物性状到疾病完整链路的预测,但相关的知识和计算方法以前达不到一定的积累。随着图神经网络、图表示学习和生物信息等方法在相关方面的进展,关联关系预测方法与生物知识的不断积累,图表示学习的方法能够在与大规模知识图谱不断结合与发展。利用 GATECDA 多头注意力机制和易于训练的特点,在多目标的二部图(bipartite graph)结构中应当会比较适用,挖掘出中间的调控过程,实现链路预测。

6　结束语

环状 RNA 与疾病关联关系预测模型在利用图表示学习机制后性能有所提升,图神经网络与深度学习结合的模型更易于训练与泛化。笔者认为利用人工智能技术挖掘已有生命科学知识进行相关的预测,其结果可以有助于解释在高通量实验中发现的大量异常信息,为研究人员推荐出与研究背景相关的关键信息,这将加快和提高相关领域的研究进展。

参考文献:

[1] MEMCZAK S, JENS M, ELEFSINIOTI A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency[J]. Nature, 2013, 495(7441): 333-338.

[2] MENG S, ZHOU H, FENG Z, et al. CircRNA: functions and properties of a novel potential biomarker for cancer[J]. Molecular Cancer, 2017, 16(1): 94-102.

[3] KRISTENSEN L S, JAKOBSEN T, HAGER H, et al. The emerging roles of circRNAs in cancer and oncology[J]. Nature Reviews. Clinical Oncology, 2022, 19(3): 188-206.

[4] HAN X, LI J, WANG Y, et al. Hsa_circ_0046430 promotes the progression of colorectal cancer by targeting miR-6785-5p/SRCIN1 axis as a ceRNA[J]. Medicine, 2023, 102(8): e33064.

[5] TANG W, FU K, SUN H, et al. CircRNA microarray profiling identifies a novel circulating biomarker for detection of gastric cancer[J]. Molecular Cancer, 2018, 17(1): 137-143.

[6] HUANG R, YANG Z, LIU Q, et al. CircRNA DDX21 acts as a prognostic factor and sponge of miR-1264/QKI axis to weaken the progression of triple-negative breast cancer[J]. Clinical and Translational Medicine, 2022, 12(5): e768.

[7] ZHAO Y, WANG C C, CHEN X. Microbes and complex diseases: from experimental results to computational models[J]. Briefings in Bioinformatics, 2021, 22(3): bbaa158.

[8] 陈文海. 关于基因型—表型相关问题的统计遗传学及计算生物学分析[D]. 上海:复旦大学, 2014.

[9] 李政伟, 李佳树, 尤著宏, 等. 基于异质图注意力网络的 miRNA 与疾病关联预测算法[J]. 电子学报, 2022, 50(6): 1428-1435.

[10] 何艳新. 基于生物信息学方法探讨模块网络及 ceRNA 调控网络在胃癌中的相关作用机制研究[D]. 郑州:郑州大学, 2020.

[11] WANG L, YOU Z H, LI Y M, et al. GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm[J]. PLoS Computational Biology, 2020, 16(5): e1007568.

[12] 张　奕, 王真梅. 图自动编码器上二阶段融合实现的环状

RNA-疾病关联预测[J]. 计算机应用, 2022, DOI: 10.11772/j. issn. 1001-9081. 2022050727: 1-9.

[13] WANG L, YOU Z H, HUANG D S, et al. MGRCDA: meta-graph recommendation method for predicting CircRNA-disease association [J]. IEEE Transactions on Cybernetics, 2023, 53(1): 67-75.

[14] WU H, SONG C, GE Y, et al. Link prediction on complex networks: an experimental survey[J]. Data Science and Engineering, 2022, 7(3): 253-278.

[15] FAN C, LEI X, WU F X. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks[J]. International Journal of Biological Sciences, 2018, 14(14): 1950-1959.

[16] WEI H, LIU B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization[J]. Briefings in Bioinformatics, 2020, 21(4): 1356-1367.

[17] FAN C, LEI X, PAN Y. Prioritizing CircRNA-disease associations with convolutional neural network based on multiple similarity feature fusion[J]. Frontiers in Genetics, 2020, 11: 540751.

[18] DEEPTHI K, JEREESH A S. An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network[J]. Gene, 2020, 762: 145040.

[19] YANG J, LEI X. Predicting circRNA-disease associations based on autoencoder and graph embedding[J]. Information Sciences, 2021, 571(4): 323-336.

[20] DENG D, CHEN X, ZHANG R, et al. XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties[J]. Journal of Chemical Information and Modeling, 2021, 61(9): 4820-4822.

[21] 刘 杰, 尚学群, 宋凌云, 等. 图神经网络在复杂图挖掘上的研究进展[J]. 软件学报, 2022, 33(10): 3582-3618.

[22] LEI X, BIAN C. Integrating random walk with restart and k-nearest neighbor to identify novel circRNA-disease association[J]. Scientific Reports, 2020, 10(1): 1943.

[23] ZHANG H Y, WANG L, YOU Z H, et al. iGRLCDA: identifying circRNA - disease association based on graph representation learning[J]. Briefings in Bioinformatics, 2022, 23(3): bbac083.

[24] VELIKOVI P, CUCURULL G, CASANOVA A, et al. Graph attention Networks[C]//International conference on learning representations. Vancouver: [s. n.], 2018.

[25] FAN C, LEI X, FANG Z, et al. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases[J]. Database, 2018, 2018: bay044.

[26] GLAŽAR P, PAPAVASILEIOU P, RAJEWSKY N. circBase: a database for circular RNAs[J]. RNA, 2014, 20(11): 1666-1670.

[27] MAO Y, LU Z. MeSH now: automatic MeSH indexing at PubMed scale via learning to rank[J]. Journal of Biomedical Semantics, 2017, 8(1): 15-24.

[28] VAN LAARHOVEN T, NABUURS S B, MARCHIORI E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. Bioinformatics, 2011, 27(21): 3036-3043.

[29] CHEN Q, SOKOLOVA M. Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts[J]. SN Computer Science, 2021, 2(5): 1-11.

[30] DENG L, LIU Y, SHI Y, et al. Deep neural networks for inferring binding sites of RNA-binding proteins by using distributed representations of RNA primary sequence and secondary structure[J]. BMC Genomics, 2020, 21(Suppl 13): 866-876.

[31] WANG L, YOU Z H, HUANG Y A, et al. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network[J]. Bioinformatics, 2020, 36(13): 4038-4046.

[32] CHEN Z, ZHAO P, LI C, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization [J]. Nucleic Acids Research, 2021, 49(10): e60.