天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于多核学习和图卷积网络的药物－疾病关联预测
作者：陈书新，李玉田，王林
DOI：10.13364/j.issn.1672-6510.20230185
收稿日期：2023-10-08
网络首发日期：2024-06-21
引用格式：陈书新，李玉田，王林．基于多核学习和图卷积网络的药物－疾病关联预测[J/OL]．天津科技大学学报. https://doi.org/10.13364/j.issn.1672-6510.20230185

2024 年 天津科技大学学报 Journal of Tianjin University of Science & Technology

DOI:10.13364/j.issn.1672-6510.20230185

基于多核学习和图卷积网络的药物–疾病关联预测

陈书新，李玉田，王　林
(天津科技大学人工智能学院，天津 300457)

摘　要：识别和预测特定药物与疾病之间的关联关系，是药物研发过程中必不可少的一部分。之前的计算方法并没有很好地整合药物和疾病的多种异源信息。本文提出了一种新的基于多核学习和图卷积网络的计算方法预测药物-疾病关联。首先，对于药物相似度，基于药物-疾病关联矩阵和药物化学结构特征信息构建多个相似度核矩阵；同样，对于疾病相似度，基于关联矩阵构建多个相似度矩阵，并结合疾病语义相似度。其次，对这些相似度矩阵使用基于中心核对齐的多核学习算法进行整合。然后，构建基于图卷积网络的模型处理相似度网络和关联网络，从而提取药物和疾病特征。最后，使用内积解码器预测药物-疾病关联。与现有的计算方法对比，验证了本预测模型可以更准确地预测药物-疾病关联。

关键词：药物；疾病；药物-疾病关联；多核学习；图卷积网络
中图分类号：TP399　　**文献标志码：**A　　**文章编号：**1672-6510 (0000)00-0000-00

Prediction of Drug-Disease Association Based on Multiple Kernel Learning and Graph Convolutional Networks

CHEN Shuxin，LI Yutian，WANG Lin
(College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China)

Abstract: Identifying and predicting the associations between specific drugs and diseases is an essential part of the drug development process. The previous computational methods did not well integrate the multiple heterogeneous information of drugs and diseases. In this article, a novel computational method based on multiple kernel learning and graph convolutional networks is proposed for drug-disease association prediction. Firstly, multiple similarity kernel matrices for drugs are constructed based on the association matrix and drug chemical structures. Similarly, multiple similarity matrices for diseases are constructed based on the association matrix, combined with disease semantic similarity. Secondly, these similarity matrices are integrated with the use of a center kernel alignment-based multiple kernel learning algorithm. A graph convolutional network model is then constructed to process the similarity network and association network, extracting features of drugs and diseases. Finally, an internal product decoder is used to predict drug-disease associations. In the experimental results, it was found that this model could predict the drug-disease associations more accurately than the state-of-the-art methods.

Key words: drug；disease；drug-disease association；multiple kernel learning；graph convolutional networks

传统药物研发是一个高成本、长时间的过程。从一开始开发新型抗病毒分子，每种药物平均花费 3.5 亿~20 亿美元，从实验室到诊所大约需要 10~15 年的时间[1]。随着药物和疾病数据的快速积累和数据处理能力的提高，已经提出了很多算法和模型对大规模的数据集进行特征提取和模式识别，以预测药物与疾病之间的潜在关联。如标签传播[2]、正则化最小二乘法[3]、矩阵分解[4-5]和矩阵补全[6]等。此外，一些采用网络分析和图论构建药物和疾病之间的网络模型也被提出，用来挖掘药物和疾病之间的

收稿日期：2023-10-08；修回日期：2024-12-29
基金项目：天津市自然科学基金重点项目(20JCZDJC00140)
作者简介：陈书新(1999—)，男，山东人，硕士研究生；通信作者：王　林，副教授，linwang@tust.edu.cn

未知关联。Luo 等[7]提出了一种新的药物重定位推荐系统，通过构建异质药物疾病相互作用网络预测未知的药物-疾病关联。还有一些方法基于已知的药物疾病关联、药物相似度和疾病相似度预测未知的药物-疾病关联。Yu 等[8]将已知的药物-疾病关联、药物-药物相似性和疾病-疾病相似性整合到异构网络中预测药物-疾病关联。尽管这些方法已经取得了一些重要的进展，但由于药物和疾病数据收集和整合的复杂性，以及药物和疾病多样性的存在，之前的方法并没有很好地对搜集到的药物和疾病信息进行良好整合，因此提出了基于多核学习和图卷积网络的预测模型更全面地对药物和疾病特征进行学习。

本研究提出了一种基于多核学习和图卷积网络(multiple kernel learning with graph convolutional networks，MKLGCN)模型预测药物-疾病关联。MKLGCN 模型以药物-疾病关联矩阵和药物分子指纹信息作为输入，输出为药物-疾病关联预测得分矩阵，具体工作流程图如图 1 所示。其中关联矩阵用 $\boldsymbol{Y}$ 表示，大小为 $m \times n$，m 为药物数量，n 为疾病数量。若药物和疾病之间存在关联则 $y_{i,j}$ 为 1，否则 $y_{i,j}$ 为 0。首先利用药物分子结构信息和药物-疾病关联特征信息，通过多个核函数计算药物之间的多种相似度，并构建多个大小相同的核矩阵；同理，对于疾病利用疾病与药物的关联特征信息构建多个大小相同的核矩阵，同时又利用疾病语义信息计算了疾病语义相似度。其次，对药物和疾病分别利用基于中心核对齐的多核学习(centered kernel alignment-based multiple kernel learning ,CKA-MKL)算法对多个相似度矩阵进行整合。然后，使用 GCN 网络模型从药物-疾病异构关联网络和药物、疾病相似网络学习药物和疾病的特征信息。最后，将学习得到的特征通过两层全连接网络，并利用内积解码器预测药物-疾病关联。基于交叉验证结果以及案例分析，验证了 MKLGCN 模型对于药物-疾病关联预测的有效性。

1 数据集

该实验使用了两个基准数据集，分别是 LRSSL 数据集[9]和 Cdataset 数据集[10]。LRSSL 数据集包含 763 种药物和 681 种疾病，经过临床证明的药物-疾病关联有 3051 个。Cdataset 中有 653 种药物，409 种疾病和 2494 个药物-疾病关联。DrugBank 数据库[11]是一个将详细的药品数据(即化学，药理学和制药)与综合药物靶点信息(即序列，结构和作用通路)相结合的生物信息学和化学信息学资源库。为了获取 LRSSL 数据集和 Cdataset 数据集中药物的化学结构信息，从 DrugBank 数据库中获取了药物 SMILES 信息，并使用开源化学信息 Python 软件包 RDKit 计算得到了药物 PubChem 分子指纹（881 维）。

2 计算方法

2.1 相似度核矩阵构建

对于药物，首先根据药物-疾病关联信息，采用 4 种核函数计算多种相似度，具体如下：

① 高斯相互作用谱(GIP)：

$$S_{GIP-d}(d_i,d_j)=\exp(-\gamma\left\|pr_{d_i}-pr_{d_j}\right\|^2) \tag{1}$$

② 余弦相似度(COS)：

$$S_{COS-d}(d_i,d_j)=\frac{pr_{d_i}^T pr_{d_j}}{\left\|pr_{d_i}^T\right\|\left\|pr_{d_j}\right\|} \tag{2}$$

③ 相关系数(Corr)：

$$S_{Corr-d}(d_i,d_j)=\frac{Cov(pr_{d_i},pr_{d_j})}{\sqrt{Var(pr_{d_i})Var(pr_{d_j})}} \tag{3}$$

④ 归一化互信息(NMI)：

$$S_{\text{NMI-d}}(d_i,d_j)=\frac{Q(pr_{d_i},pr_{d_j})}{\sqrt{H(pr_{d_i})H(pr_{d_j})}} \tag{4}$$

其中：pr_{d_i} 是 $\boldsymbol{Y}$ 的第 i 行，pr_{d_j} 是 $\boldsymbol{Y}$ 的第 j 行；γ 是高斯相互作用谱带宽；$Cov(pr_{d_i},pr_{d_j})$ 是 pr_{d_i} 和 pr_{d_j} 的协方差；$Q(pr_{di},pr_{dj})$ 是 pr_{d_i} 和 pr_{d_j} 的互信息；$H(pr_{d_i})$ 和 $H(pr_{d_j})$ 分别是 pr_{d_i} 和 pr_{d_j} 的熵 。

根据上述提供的方法，对于药物，由关联信息构建的 4 种相似度矩阵分别表示为 $\boldsymbol{S}_{\text{GIP-d}}$、$\boldsymbol{S}_{\text{COS-d}}$、$\boldsymbol{S}_{\text{Corr-d}}$、$\boldsymbol{S}_{\text{NMI-d}}$；由化学结构信息构建的 4 种相似度矩阵分别表示为 $\boldsymbol{S}_{\text{GIP-chem,d}}$、$\boldsymbol{S}_{\text{COS-chem,d}}$、$\boldsymbol{S}_{\text{Corr-chem,d}}$、$\boldsymbol{S}_{\text{NMI-chem,d}}$；对于疾病，由疾病关联信息构建的 4 种疾病相似度矩阵分别表示为 $\boldsymbol{S}_{\text{GIP-s}}$、$\boldsymbol{S}_{\text{COS-s}}$、$\boldsymbol{S}_{\text{Corr-s}}$、$\boldsymbol{S}_{\text{NMI-s}}$。

此外对于疾病相似度还利用疾病语义构建相似度矩阵。采用 Medical Subject Headings (MeSH)中的网络描述符处理疾病数据。MeSH 描述符一般被描述为疾病的有向无环图(DAG)，其中疾病由节点表示。本文采用的疾病语义相似度 S 从 LRSSL 数据集和 Cdataset 数据集中获得。

2.2　相似度矩阵整合

在本节中，将介绍对以上药物、疾病相似度矩阵整合的方法(图 1)。在整合相似度过程中，使用了基于中心核对齐的多核学习算法 CKA-MKL[12]。CKA-MKL 在一个全局特征空间上线性组合所有核矩阵，整合来自药物的多种信息，并以保留数据相似性和较好的泛化性能为目标。CKA-MKL 的优化模型如下：

$$\max_{\omega\geq 0} \frac{\left\langle \mathrm{VS}_{\mathrm{com,d}} V, YY^T \right\rangle}{\left\| \mathrm{VS}_{\mathrm{com,d}} V \right\|_F \left\| YY^T \right\|_F}$$

$$s.t.\ \mathrm{S}_{\mathrm{com,d}} = \sum_{k=1}^{8} \omega_k S_{drug}^{[k]}, \sum_{k=1}^{8} \omega_k = 1 \tag{5}$$

其中：$\boldsymbol{V}=\boldsymbol{I}_m-\boldsymbol{E}/m$，$\boldsymbol{I}_m$ 是单位矩阵，$\boldsymbol{E}\in R^{m\times m}$ 表示元素全为 1 的矩阵。$\langle\cdot,\cdot\rangle$ 和 $\|\cdot\|$ 表示矩阵的 Frobenius 内积和 Frobenius 范数。$S_{drug}^{[k]}$（k=1,2,…,8）为 2.1 节构建的药物相似度核矩阵。$\omega^{[k]}$ 是相似度矩阵的权重系数。同样的，可以利用 CKA-MKL 组合疾病的 5 种相似度核矩阵，得到融合后的疾病相似度矩阵 $S_{com,s}$。

2.3　构建异构网络

本实验所使用的药物和疾病异构网络中存在 3 种类型的边，其中一种是 m 种药物和 n 种疾病自身原始相互作用，即药物与疾病之间的关联。另外两种类型的边是相似度边，分别由上述组合得到的药物相似度矩阵和疾病相似度矩阵构建。

2.4　获取节点特征

对于药物-疾病关联异构网络，有药物和疾病两种类型的节点。利用重启随机游走(random walk with restart，RWR)[13]作用于药物相似度核矩阵和疾病相似度核矩阵提取药物和疾病的初始特征如图 1(b)所示。由 RWR 得到的药物节点表示可由以下公式计算得到：

$$P_{x,y}^{i+1}(x)=(1-\mu)e_{x,y}+\mu P_{x,y}^{i}(x)S_{com,d}(x,y) \tag{6}$$

$$P(x)=[P_{x,1}^{\infty}(x),\cdots,P_{x,m}^{\infty}(x)] \tag{7}$$

其中：$P_{x,y}^{i}(x)$ 表示经过 i 步从药物节点 x 走到 y 的概率。$e_{x,y}$ 表示从药物节点 x 走到 y 的初始取值，它是单位矩阵的元素。$S_{com,d}(x,y)$ 表示从相似度矩阵中获得的转移概率，μ 是重启概率，对于药物来说，μ 取 0.8，对于疾病来说，μ 取 0.7。将 x 与所有其他药物节点关联的概率连接起来，生成药物的节点表示。同样地，疾病节点表示 $D(x)$可以表示为：

$$D_{x,y}^{i+1}(x)=(1-\mu)e_{x,y}+\mu D_{x,y}^{i}(x)S_{com,s}(x,y) \tag{8}$$

$$D(x)=[D_{x,1}^{\infty}(x),\cdots,D_{x,n}^{\infty}(x)] \tag{9}$$

对于药物和疾病节点特征的提取，利用 GCN 作为网络模型的基础。GCN 是一种用于处理图数据的深度学习算法，可以通过多层次堆叠提取药物和疾病特征。在每个层次中，GCN 首先计算每个节点与其邻接节点的加权平均值，然后使用这些平均值更新当前节点的表示向量。首先使用 GCN 从药物疾病异构关联网络中提取药物和疾病特征。具体来说，给定一个具有相应邻接矩阵 G 的网络，GCN 层与层之间的传播方式是：

$$\mathrm{H}^{l+1}=f(H^l,G)=\sigma(D^{-\frac{1}{2}}GD^{-\frac{1}{2}}H^lW^l) \tag{10}$$

其中：H^l 是第 l 层的嵌入，$\boldsymbol{D}$ 是邻接矩阵 $\boldsymbol{G}$ 的度矩阵，$\sigma(\bullet)$ 是非线性激活函数，是第 l 层可训练的权重矩阵，$\boldsymbol{W}^l$ 表示 GCN 模型的可训练参数矩阵。

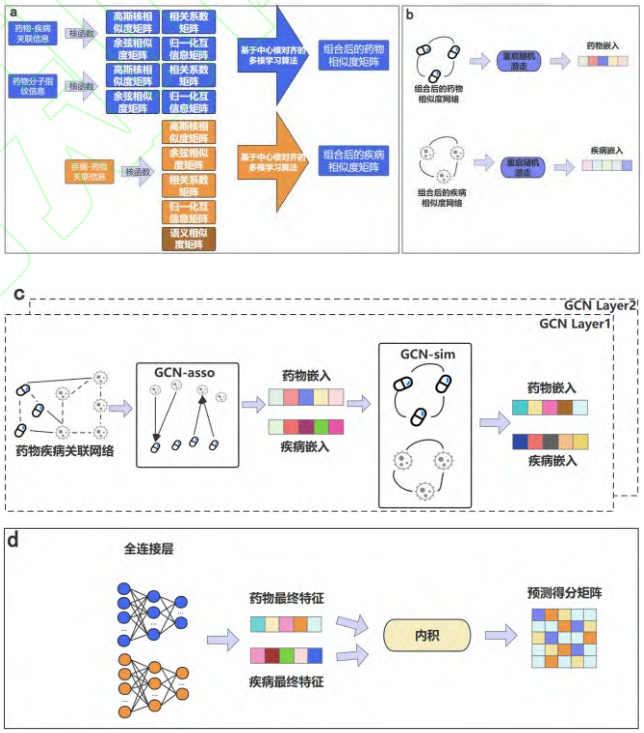


图 1　MKLGCN 的流程图

Fig.1　Flowchart of MKLGCN

为了利用 GCN 来根据药物、疾病关联信息和相似性信息获取药物和疾病的低维特征，分别在关联异构网络上和相似度网络上引入了 GCN-asso 网络和 GCN-sim 网络如图 1(c)所示。对于药物-疾病异构关联网络，采用 GCN-asso 网络提取药物和疾病节点特征。将图 G_{asso}={V_{d}，V_{s}，$E_{\mathrm{d\text{-}s}}$}作为模型的输入，其中药物和疾病节点分别用 V_{d}、V_{s} 表示，$E_{\mathrm{d\text{-}s}}$ 表示药

物和疾病之间的相互作用，将重启随机游走提取到的药物和疾病特征作为初始节点特征。在此过程中，药物节点的特征从相邻疾病节点信息聚合而来，同样，疾病节点的特征从相邻药物节点信息聚合得到。

对于药物和疾病相似度网络，采用 GCN-sim 提取药物和疾病节点特征。分别构建药物相似度网络 $G_d=\{V_d, E_{d\text{-}d}\}$和疾病相似度网络 $G_s=\{V_s, E_{s\text{-}s}\}$作为 GCN-sim 网络模型的输入，其中从药物-疾病异构关联网络获得的药物、疾病特征作为 GCN-sim 网络的初始节点特征。与 GCN-asso 处理过程不同的是，在此过程中，药物节点的特征从相邻药物节点信息聚合而来，疾病节点特征从邻居疾病特征信息中学习得到。

3 药物-疾病关联预测

从 GCN 模块中提取到药物和疾病特征分别表示为 f_{d_i}， f_{s_j}，通过两层全连接运算，可以得到最终的药物和疾病特征表示 f_{d_i}'， f_{s_j}' 如图 1(d)所示。则通过预测得到的药物和疾病之间的关联得分可通过以下公式计算得到：

$$\boldsymbol{R}_{i,j} = f_{d_i}^{'} f_{s_j}^{'T} \tag{11}$$

其中：$\boldsymbol{R}$ 是最终的预测得分矩阵。$R_{i,j}$ 中元素值越大，表明药物 d_i 与疾病 s_j 关联的可能性越大。

均方误差损失函数是一种常用于衡量模型预测值与真实值之间的差异程度的损失函数。本文采用均方误差作为损失函数，最小化预测得分矩阵 $\boldsymbol{R}$ 和标签矩阵 Y 之间差异的 Frobenius 范数。但是由于关联矩阵中未知关联数量远多于已知关联数量，为了使训练样本变得平衡，使用 θ 加强正样本训练权重[14]，θ 取值为 5：

$$\text{Loss} = \left\| \tilde{Y} - R \right\|_F^2 + \lambda \left\| W \right\|_2^2 \tag{12}$$

其中

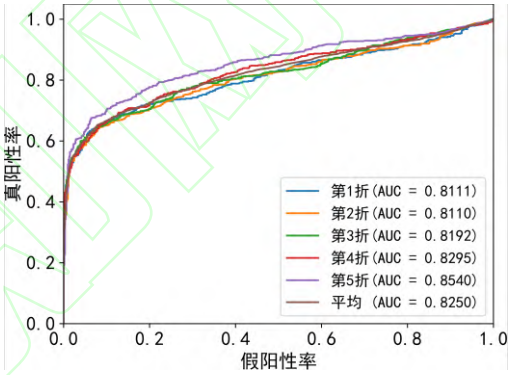
$$\tilde{\boldsymbol{Y}} = \begin{cases} 0 & \text{if } \boldsymbol{Y}(i,j) = 0 \\ \theta & \text{else} \end{cases} \tag{13}$$

其中 $\tilde{\boldsymbol{Y}}$ 是在原始关联矩阵 $\boldsymbol{Y}$ 的基础上生成的增强关联矩阵。

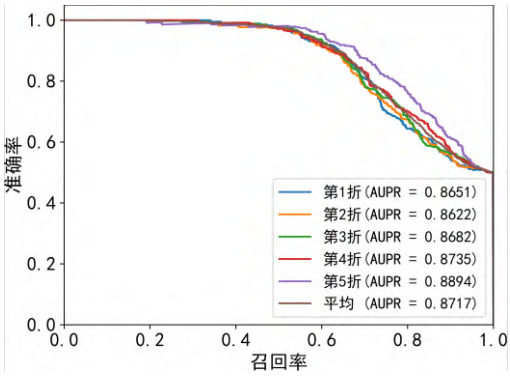
4 实验结果与分析

4.1 5 折交叉验证

在本实验中采用 5 折交叉验证验证模型，将所有已知的药物-疾病关联随机均匀的分成 5 个子集，其中 4 个子集作为训练集的正样本，选取和正样本相同数量的未知关联作为训练集的负样本。另外一个子集作为测试集的正样本，选取和正样本相同数量的未知关联作为测试集负样本。训练集负样本和测试集负样本不重复。交叉验证重复 5 次，每个子集依次作为测试集，其余 4 个子集作为训练集，使用接受者操作特征曲线(receiver operating characteristic curve，ROC)下面积 AUC (area under curve)和精确率-召回率曲线(precision- recall curve，PR)下面积 AUPR(area under the precision recall curve)作为评价指标。为了使结果更加直观，图 2、图 3 中（a）和（b）给出了在 LRSSL 数据集和 Cdataset 数据集上基于 5 折交叉验证的 ROC 曲线和 PR 曲线。



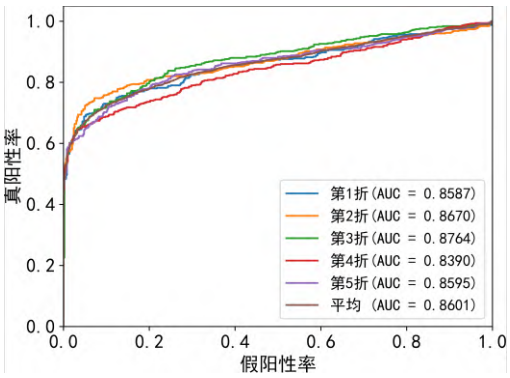
（a）



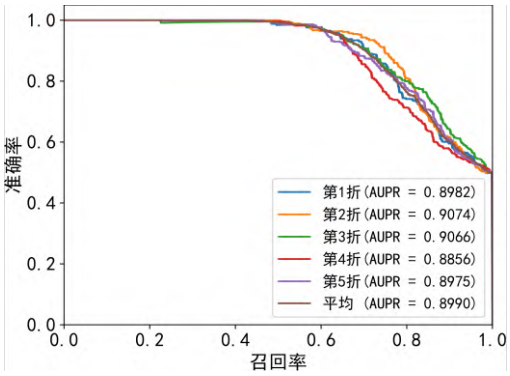
（b）

图 2 MKLGCN 模型在 LRSSL 数据集上 5 折交叉验证下的 ROC、AUC、PR 和 AUPR

Fig. 2 ROCs, AUCs, PRs, and AUPRs under five-fold cross-validation on the LRSSL datasets for MKLGCN



（a）



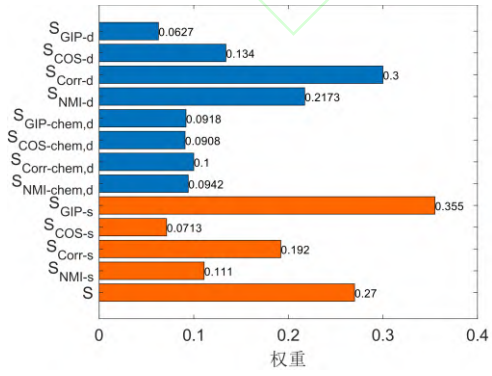
（b）

图 3　MKLGCN 模型在 Cdataset 上 5 折交叉验证下的 ROC、AUC、PR 和 AUPR

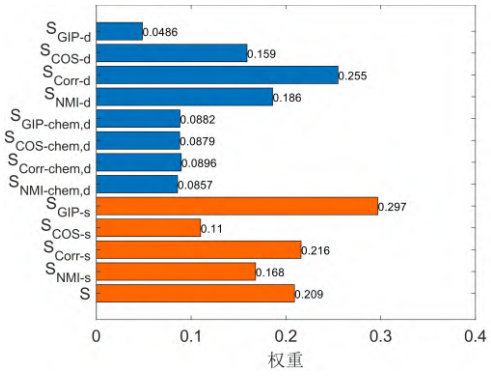
Fig.3　ROCs, AUCs, PRs, and AUPRs under five-fold cross-validation on the Cdatasets for MKLGCN

4.2　不同核矩阵对组合相似度矩阵的贡献度

在本实验中，用 CKA-MKL 算法融合相似度矩阵，CKA-MKL 的作用与特征选择类似。不同的核函数通过不同的计算方式得到药物和疾病的多种相似度，CKA-MKL 可以确定每种相似度对于关联矩阵 Y 相似度拟合的重要性程度，即权重。图 4a 和 b 分别显示了在 LRSSL 数据集和 Cdataset 数据集上每个相似度矩阵的权重。



（a）LRSSL 数据集



（b）Cdataset 数据集

图 4　LRSSL 和 Cdataset 数据集上的相似度矩阵权重

Fig. 4　Weights of the similarity matrices on the LRSSL datasets and Cdatasets

4.3　GCN 层数对模型预测性能的影响

GCN 是 MKLGCN 模型的关键部分，它可以获取来自邻居节点的特征信息。GCN 层数对模型的预测效果有十分重要的影响。表 1 展示了在 LRSSL 数据集上不同 GCN 层数对模型预测结果的影响。从表中可以看出，模型不使用 GCN 时，不能充分获取药物和疾病的节点特征，预测性能较低。当使用 GCN 捕获节点特征时，模型取得了较好的性能。当逐渐添加 GCN 层数时，模型预测效果逐渐变好，但 GCN 层数超过两层时，模型的预测性能有所下降。所以，当使用两层 GCN 时，模型的预测效果最佳。

表 1　LRSSL 数据集上 GCN 层对模型预测性能的影响

Tab. 1　Effect of GCN layers on the predictive performance of MKLGCN on the LRSSL datasets

GCN 层数	AUC	AUPR
0	0.7819	0.8413
1	0.8238	0.8693
2	**0.8250**	**0.8717**
3	0.8230	0.8707
4	0.8201	0.8679

4.4　与其他模型比较

为了验证 MKLGCN 模型预测效果的优势，与其他基线方法进行了比较。在给定的数据集中，保证训练集和测试集一致，根据原文提供的代码做了对比试验。所有方法的整体性能通过 4.1 节中指定的 5 折交叉验证评估。ROC 曲线和 PR 曲线如图 5、图 6 中 a 和 b 所示。结果显示，在相同的数据集中，MKLGCN 预测模型具有相对较大的优势。

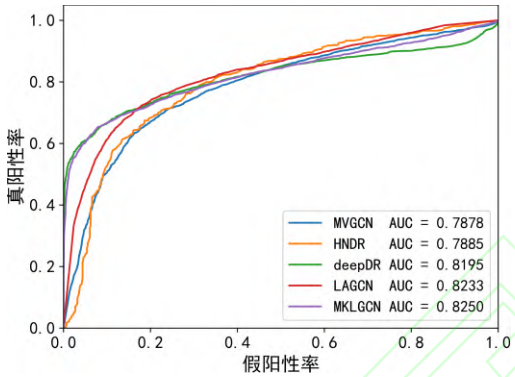
MVGCN[15]模型将药物-疾病关联网络和多种相似度网络结合起来构建成多视图异构网络，然后设计了基于 GNNs (Graph Neural Networks)的架构，将不同视图中的内部和外部信息相结合进行药物−疾

病关联预测。

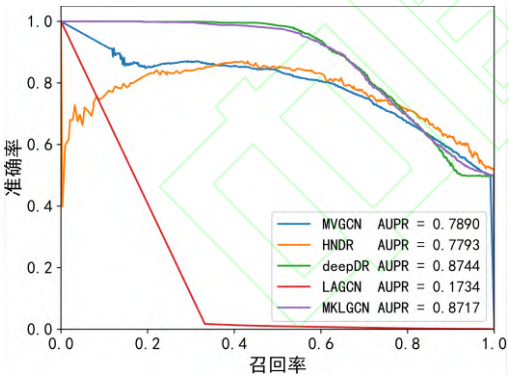
LAGCN[8]模型将已知的药物-疾病关联、药物-药物相似性和疾病-疾病相似性整合到一个异构网络中，并将图卷积操作应用于网络以学习药物和疾病的嵌入。同时使用注意力机制组合来自多个图卷积层的嵌入，并对未知的药物-疾病关联进行评分。

deepDR[16]模型计算了基于药物相关网络的PPMI (Positive Pointwise Mutual Information)矩阵作为药物特征，然后提出用于融合特征的多模式深度自动编码器和用于挖掘新关联的变分自动编码器。

HNDR[17]基于神经网络中的邻域信息聚合，结合疾病和药物的相似性、药物与疾病之间的关联性，提取药物特征和疾病特征。然后采用端到端的恢复方法，根据矩阵分解的原理预测新的药物与疾病之间的关联。



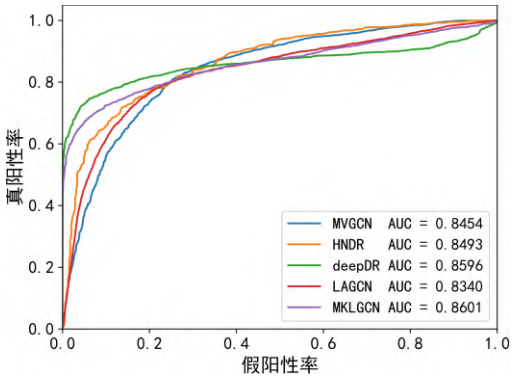
（a）AUC



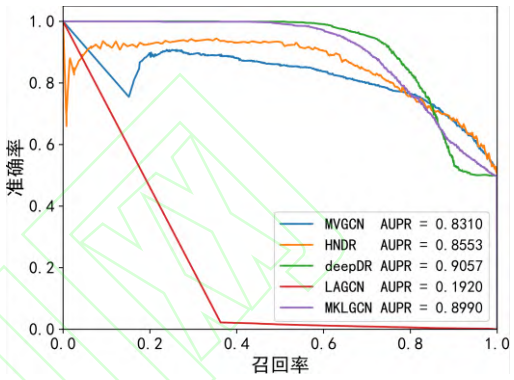
（b）AUPR

图 5　不同预测方法在 LRSSL 数据集上 AUC 和 AUPR 的比较

Fig.5　Comparison of predicting methods in terms of AUC and AUPR on the LRSSL datasets



（a）



（b）

图 6　预测方法在 Cdataset 上 AUC 和 AUPR 的比较

Fig. 6　Comparison of predicting methods in terms of AUC and AUPR on the Cdatasets

从图 5 和图 6 可以看出，本实验预测模型的 AUC 在给定数据集上明显优于其他预测模型。在 LRSSL 数据集和 Cdataset 数据集上，AUPR 值接近 deepDR 模型，但均高于其他模型。在 LRSSL 数据集和 Cdataset 数据集上进行 5 折交叉验证部分实验指标如表 2 所示，其中 F1、ACC、RE、SP、PR 分别表示 F1 分数，Accuracy、Recall、Specificity、Precision。从表中可以看出 MKLGCN 在两个基准数据集上 F1 分数均优于其他算法。这表明本实验模型可以更好对药物-疾病关联进行预测。

表 2　两个基准数据集上性能对比的实验结果

Tab. 2　Experimental results of performance comparison on two benchmark datasets

数据集	模型	F1	ACC	RE	SP	PR
LRSSL	MVGCN	0.7405	0.7310	0.7573	0.7046	0.7243
	HNDR	0.7770	0.7657	0.7873	0.7398	0.7670
	deepDR	0.7660	0.7823	0.7115	0.8531	0.8230
	LAGCN	0.0343	**0.9777**	0.3203	**0.9785**	0.0181
	MKLGCN	**0.8086**	0.7861	**0.7934**	0.8492	**0.8294**
Cdataset	MVGCN	0.7851	0.7754	0.8332	0.6976	0.7422
	HNDR	0.8005	0.7902	0.8351	0.7432	0.7687

deepDR	0.8300	0.8436	0.7628	0.9244	**0.9203**
LAGCN	0.0435	**0.9665**	0.3524	**0.9676**	0.0232
MKLGCN	**0.8735**	0.838	**0.838**	0.928	0.9192

4.5　消融实验

为了验证 CKA-MKL 组合多种药物相似度对于预测药物与疾病关联的重要性，在此部分进行了消融实验。对于药物分别只保留药物化学结构相似度(Case 1)、药物关联信息相似度(Case 2)。表 3、表 4 分别展示了在 LRSSL 数据集和 Cdataset 数据集上两种情况的 AUC 和 AUPR 值。结果表明，对于给定数据集，在去掉一部分药物相似度信息后，模型预测药物-疾病关联的性能有所下降，AUC 和 AUPR 显著降低，只有当组合两种相似度时模型预测结果最好。因此，当使用 CKA-MKL 组合多种药物相似度信息和疾病相似度信息时，模型预测效果最好。

表 3　LRSSL 数据集上的消融实验结果

Tab. 3　Results of ablation experiments on LRSSL datasets

	AUC	AUPR
Case 1 & Case 2	0.8250	0.8717
Case 1	0.8162	0.8637
Case 2	0.8130	0.8654

表 4 Cdataset 数据集上的消融实验结果

Tab. 4　Results of ablation experiments on Cdatasets

	AUC	AUPR
Case 1 & Case 2	0.8601	0.8990
Case 1	0.8516	0.8799
Case 2	0.8546	0.8869

4.6　案例分析

为了评估 MKLGCN 模型在实际中的应用，选取了 LRSSL 数据集中的肺部肿瘤和 Cdataset 数据集中阿尔茨海默病进行案例分析。根据模型预测，对得到的预测药物进行排序，对两种疾病分别选取了前 10 名候选药物进行分析，如下表 5、表 6 所示。对于选取的两种疾病，模型预测得到的前 10 名候选药物都可以在 CTD[18]数据库中得到验证。通过案例分析，MKLGCN 模型可以在实际的药物-疾病关联预测中得到应用。

表 5　肺部肿瘤相关的药物 TOP10 预测

Tab. 5　Prediction of top 10 drugs related to Lung Neoplasms

药物编号	药物名称	疾病编号	疾病名称
DB00773	Etoposide	D008175	Lung Neoplasms
DB00997	Doxorubicin	D008175	Lung Neoplasms
DB00570	Vinblastine	D008175	Lung Neoplasms
DB00563	Methotrexate	D008175	Lung Neoplasms
DB00262	Carmustine	D008175	Lung Neoplasms
DB00290	Bleomycin	D008175	Lung Neoplasms
DB01005	Hydroxyurea	D008175	Lung Neoplasms
DB01229	Paclitaxel	D008175	Lung Neoplasms
DB00541	Vincristine	D008175	Lung Neoplasms
DB00762	Irinotecan	D008175	Lung Neoplasms

表 6　阿尔茨海默病相关的药物 TOP10 预测

Tab. 6　Prediction of top 10 drugs related to Alzheimer Disease

药物编号	药物名称	疾病编号	疾病名称
DB00163	Vitamin E	D104300	Alzheimer Disease
DB00843	Donepezil	D104300	Alzheimer Disease
DB00382	Tacrine	D104300	Alzheimer Disease
DB00674	Galantamine	D104300	Alzheimer Disease
DB00989	Rivastigmine	D104300	Alzheimer Disease
DB01037	Selegiline	D104300	Alzheimer Disease
DB01043	Memantine	D104300	Alzheimer Disease
DB00313	Valproic acid	D104300	Alzheimer Disease
DB01219	Dantrolene	D104300	Alzheimer Disease
DB00413	Pramipexole	D104300	Alzheimer Disease

5　结　语

本文提出了一种基于多核学习和图卷积网络的药物-疾病关联预测模型。现有方法使用已知的药物-疾病关联、药物相似性和疾病相似性预测药物-疾病关联，但是没有对这些异源数据进行有效整合。本文针对基准数据集，对于药物和疾病利用多种核函数构建多个相似度核矩阵，并利用 CKA-MKL 算法对多个相似度矩阵进行整合；构建药物-疾病关联异构网络和相似度网络，基于图卷积网络提取药物和疾病特征，从而进行药物-疾病关联预测。在实验部分，利用 5 折交叉验证，结果显示 MKLGCN 模型的预测性能优于其他模型。在基准数据集上对肺部肿瘤、阿尔茨海默的案例分析表明 MKLGCN 模型可以在实际的药物-疾病关联预测中得到应用。

参考文献：

[1] TRIVEDI J, MOHAN M, BYRAREDDY S. Drug repurposing approaches to combating viral infections[J]. Journal of clinical medicine, 2020, 9(11):3777.

[2] ZHANG W, YUE X, CHEN Y, et al. Predicting drug-disease associations based on the known association bipartite network[C]//IEEE. IEEE International Conference on Bioinformatics and Biomedicine. New York：IEEE, 2017：

503–509.

[3] LU L, YU H. DR2DI: a powerful computational tool for predicting novel drug-disease associations[J]. Journal of computer-aided molecular design, 2018, 32：633-642.

[4] ZHANG W, YUE X, LIN W, et al. Predicting drug-disease associations by using similarity constrained matrix factorization[J]. BMC Bioinformatics, 2018, 19(1)：233.

[5] XUAN P, CAO Y, ZHANG T, et al. Drug repositioning through integration of prior knowledge and projections of drugs and diseases[J]. Bioinformatics, 2019, 35(20)：4108–4119.

[6] YANG M, LUO H, LI Y, et al. Overlap matrix completion for predicting drug-associated indications[J]. PLOS Computational biology, 2019, 15(12)：e1007541.

[7] LUO H, LI M, WANG S, et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms[J]. Bioinformatics, 2018, 34(11)：1904–1912.

[8] YU Z X, HUANG F, ZHAO X H, et al. Predicting drug-disease associations through layer attention graph convolutional network[J]. Briefings in bioinformatics, 2021, 22(4)：bbaa243.

[9] LIANG X, ZHANG P, YAN L, et al. LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning[J]. Bioinformatics, 2017, 33(8)：1187–1196.

[10] LUO H, WANG J, LI M, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm[J]. Bioinformatics, 2016, 32：2664–2671.

[11] WISHART D S, FEUNANG Y D,GUO A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018[J]. Nucleic acids research, 2018, 46(1): 1074-1082.

[12] QIAN Y, DING Y, ZOU Q, et al. Identification of drug-side effect association via restricted Boltzmann machines with penalized term[J]. Briefings in bioinformatics, 2022, 23(6)：bbac458.

[13] TONG H, FALOUTSOS C, PAN J Y. Fast random walk with restart and its applications[C]//IEEE. Sixth International Conference on Data Mining (ICDM' 06). New York: IEEE, 2006：4053087.

[14] LIU L, MAMITSUKA H, ZHU S, HPOFiller: identifying missing protein–phenotype associations by graph convolutional network[J]. Bioinformatics, 2021, 37(19): 3328–3336.

[15] FU H, HUANG F, LIU X, et al. MVGCN: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks[J]. Bioinformatics, 2021, 38(2)：426-434.

[16] ZENG X, ZHU S, LIU X, et al. deepDR：a network-based deep learning approach to in silico drug repositioning[J]. Bioinformatics, 2019, 35(24): 5191–5198.

[17] WANG Y, DENG G, ZENG N, et al, Drug-disease association prediction based on neighborhood information aggregation in neural networks[J]. IEEE Access, 2019, 7：50581–50587.

[18] DAVIS A P, WIEGERS T C, JOHNSON R J, et al, Comparative toxicogenomics database (CTD): update 2023[J]. Nucleic Acids Research，2023, 51(1): 1257-1262.