文章编号：2095-2163(2023)10-0166-06 中图分类号：TP391 文献标志码：A

基于属性补全的药物与疾病关联预测

唐瑞泽[1]，玄 萍[2]
（1 黑龙江大学 计算机科学技术学院，哈尔滨 150080；2 汕头大学 计算机科学技术系，广东 汕头 515063）

摘 要：预测药物-疾病关联关系，有助于降低药物开发的成本和时间开销。先前的方法没有基于异构网络的拓扑信息对缺失属性的疾病节点进行节点属性补全，本文提出了一个新的预测方法来编码和整合多个元路径的语义，学习得到药物和疾病节点的拓扑嵌入。以节点间的拓扑关系为指导，对有属性的药物节点属性进行加权聚合，来补全没有属性的疾病节点。此外，本文还设计了一个元路径层面注意力机制和一个邻居层面注意力机制，分别融合来自多个元路径的语义信息和节点邻居的信息。采用了五倍交叉验证的方法进行评估，结果表明新的预测模型取得了比其它模型更高的预测性能。

关键词：药物-疾病关联；节点属性补全；元路径层面注意力机制；邻居层面注意力机制

Prediction of drug-disease associations based on attribute complement

TANG Ruize[1], XUAN Ping[2]
(1 School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China;
2 Department of Computer Science and Technology, Shantou University, Shantou Guangdong 515063, China)

【Abstract】Predicting drug-disease associations can help reduce the cost of drug development. Previous approaches ignore the node attribute complementation for the disease nodes with missing attributes based on topological information of heterogeneous network. In this paper, we propose a novel prediction method to encode and integrate the semantics of multiple meta-paths and to construct the topological embeddings of drug and disease nodes. By utilizing the topological relationships among nodes as the guide, the attributes of drug nodes are aggregated to complement the attributes of disease nodes. In addition, we design a meta-path-level attention mechanism and a neighbour-level attention mechanism that fuse semantic information from multiple meta-paths and information from node neighbours, respectively. In this paper, a five-fold cross-validation approach is adopted for evaluation. The experimental results show that the new prediction model achieves higher prediction performance than other compared models.

【Key words】drug-disease association; node attribute complement; metapath-level attention mechanisms; neighbour-level attention mechanisms

0 引 言

研发一个用于疾病治疗的新药需要一个漫长的过程约 10~15 年，同时还会花费 8~15 亿美元[1]。药物重新定位是为已批准的药物寻找新的治疗效果[2]。已上市的药物具有已知的安全性和药理学特征，因此药物重新定位可以将药物开发的时间缩短到 6.5 年，并把研发成本降低到 3 亿美元。

计算已批准药物的新治疗适应症，有助于在筛选现有药物进行进一步实验验证时预测候选疾病。现有的计算预测方法大致可分为 3 类，两种药物的功能越相似，就越有可能与类似的疾病相关。因此，第一类的方法主要是利用药物-疾病关联、疾病相似性和药物相似性数据进行药物-疾病关联预测。例如，Zhang 等[3-4]利用非负矩阵分解和相似性约束的矩阵分解来整合已知的药物和疾病信息，获取药物和疾病的关联概率。还有一些方法通过在药物-疾病异构网络上随机游走来预测关联分数[5-6]。Wang 等[7]构建了一个支持向量机模块(SVM)来推断药物的未知治疗效果。然而，随着药物相关数据的增加和多样化，除了考虑药物的基本靶点信息和蛋白质结构外，其他信息对预测疾病候选者也很重

基金项目：国家自然科学基金(61972135,62172143)；黑龙江省自然科学基金(LH2023F044)；汕头大学科研启动基金(NTF22032)。

作者简介：唐瑞泽(1998-)，男，硕士研究生，主要研究方向：生物信息学、深度学习；玄 萍(1979-)，女，博士，教授，主要研究方向：复杂生物网络分析、深度学习、生物信息学。

通讯作者：玄 萍 Email：pxuan@ stu.edu.cn

收稿日期：2022-11-02

要，而这些方法并没有整合这些多源数据。

第二类方法考虑使用与药物和疾病相关的多个数据源进行关联预测。已经开发了几种方法，非负矩阵分解、稀疏子空间学习或推理概率矩阵分解来预测候选药物注释。还有一些方法通过在构建的异构网络上随机游走来预测各种药物的候选疾病[8]。然而，多个数据源表现出复杂的非线性关系，整合这些数据对于探索药物与疾病的相关性至关重要。

第三类方法采用深度学习方法整合药物和疾病相关信息，以更准确地识别合适的疾病候选者。Xuan 等[9]提出了一个基于 CNN(Convolution Neural Network)和 BiLSTM(Bi-directional Long Short-Term Memory)架构的模型，用于预测药物-疾病关联分数。此外，还构建了基于卷积神经网络的模型和基于图卷积网络(GCN)的模型来推断药物的候选疾病。然而，在深度学习过程中，没有考虑以节点间的拓扑关系为指导，通过加权聚合有属性节点的属性来补全无属性节点的属性[10]。在这项研究中，本文提出了一个基于属性补全的预测模型，从不同的元路径编码和捕捉异构网络中节点的拓扑嵌入，为无属性节点进行属性补全。

1　材料和方法

为了预测特定药物的潜在适应症即候选疾病，本文提出了药物-疾病关联预测模型。首先，基于多种药物相似性、疾病相似性和药物-疾病关联构建了 3 种不同的药物-疾病异构网络；构建多个元路径，用来编码和学习药物和疾病节点的拓扑嵌入，并提出一个基于元路径层面的注意力机制，融合来自多个元路径的不同的语义信息；以融合后的药物(疾病)节点的拓扑嵌入为指导，对有属性的药物节点的属性进行加权聚合来补全没有属性的疾病节点的属性；最后，将得到的 3 个网络的药物-疾病节点对的属性通过 1 × 1 卷积融合，通过两层全连接神经网络，输出药物和疾病是否存在关联的分数。

1.1　相关数据集

本文从以往的药物-疾病关联预测工作中获得药物与疾病的关联、药物的化学亚结构、药物的靶蛋白结构域、药物的靶注释以及疾病语的语义相似性。3 051个已知的药物-疾病关联数据最初是从联合医学语言系统(UMLS)中提取的，其中包含 763 种药物和 681 种疾病之间的治疗关系。本文主要利用了 3 种药物属性，药物的化学结构是从 PubChem 数据库中提取的化学指纹，从 InterPro 数据库和 UniProt 数据库中获得了药物的靶蛋白结构域和药物的靶注释。相关的疾病命名由美国国家医学图书馆提供(MeSH)。

1.2　药物和疾病的多源数据矩阵表示

1.2.1　多种药物属性表示

基于多种药物相关的数据，本文用矩阵 $\boldsymbol{B}^p(p = chem, doma, anno)$ 分别表示药物的 3 种属性，即药物的化学子结构，药物靶蛋白的目标域和基因注释。$\boldsymbol{B}^p$ 被定义为式(1)：

$$\boldsymbol{B}^p = \begin{cases} B^{chem} \in R^{N_r * N_{chem}} \\ B^{doma} \in R^{N_r * N_{doma}} \\ B^{anno} \in R^{N_r * N_{anno}} \end{cases} \tag{1}$$

其中，N_r 表示药物的数量，N_{chem}(N_{doma}，N_{anno})是药物化学子结构(药物靶蛋白的目标域，基因注释)的数量，$N_{chem} = 623$，$N_{doma} = 1\ 426$，$N_{anno} = 447$。

如果 $B^{chem}(i, j)$ 的值为 1，表示药物 r_i 具有化学子结构 c_j，否则值为 0。同样地，如果药物 r_i 含有靶蛋白结构域 o_j(基因注释 t_j)，将 $B^{doma}(i, j)$($B^{anno}(i, j)$)的值置为 1，否则为 0。

1.2.2　多种药物相似性表示

两个药物 r_i 和 r_j 之间具有越多相同的化学子结构，通常这种情况下药物 r_i 和 r_j 在功能上具有更高的相似性；类似地，当药物 r_i 和 r_j 具有更多相同的靶蛋白域或者靶注释，r_i 和 r_j 之间也会具有更高的相似性。基于这些生物性前提，Wang 等[11]通过余弦相似性计算得到了 3 种不同的药物相似。3 个药物相似矩阵分别为 $\boldsymbol{S}^r_{chem}$，$\boldsymbol{S}^r_{doma}$，$\boldsymbol{S}^r_{anno}$。药物相似性矩阵 $\boldsymbol{S}^r_p$ 定义为式(2)：

$$\boldsymbol{S}^r_p = \begin{cases} \boldsymbol{S}^r_{chem} \in R^{N_r * N_r} \\ \boldsymbol{S}^r_{doma} \in R^{N_r * N_r} \\ \boldsymbol{S}^r_{anno} \in R^{N_r * N_r} \end{cases} \tag{2}$$

$\boldsymbol{S}^r_{chem}$ 反映了药物之间在化学亚结构方面的相似度大小，$\boldsymbol{S}^r_{doma}$($\boldsymbol{S}^r_{anno}$)表示一对药物在蛋白质结构域(靶注释)下的相似度大小，取值范围在[0,1]之间，数值越大说明两种药物就越相似。

1.2.3　疾病相似性表示

有向无环图(DAG)通常被用来表示一种疾病，该图是由多个与该疾病相关的疾病术语组成。两个疾病有越多相同的疾病术语，两个疾病之间越相似。通过余弦相似性计算得到的矩阵 $\boldsymbol{S}^d \in R^{N_d * N_d}$ 表示两种疾病之间的相似性，N_d 是疾病的数量，S^d_{ij} 的值域[0,1]，值越高，d_i 和 d_j 之间越相似。

1.2.4　药物-疾病关联表示

关联矩阵 $\boldsymbol{A}^{rd} \in R^{N_r * N_d}$ 包含了 N_r 个药物和 N_d

个疾病之间的关联。每一行和每一列分别代表一种药物和一种疾病。如果 r_i 和 d_j 之间存在关联,则 A_{ij}^{rd} 的值为 1,否则 $A_{ij}^{rd}=0$。

1.3 多个药物-疾病的异构网络

面对 3 种不同的药物相似性,构建 3 个药物-疾病异构网络 $G^p=(V,E)$。每个异构网络包含了两种类型的节点 $V=(V^r \cup V^d)$ 和 3 种类型的边 $E=(E_p^{r-r} \cup E^{d-d} \cup E^{r-d})$。每个异质网络中的节点总数是药物节点和疾病节点数量之和 ($N^{total}=N_r+N_d$),E_p^{r-r} 是基于第 p 种药物相似性建立的药物-药物相似性的边。利用已知的关联数据,建立药物-疾病的边,用 E^{r-d} 表示。如果节点 $v_i,v_j \in V$ 之间存在一个连接,那么 $e_{ij} \in E$。

1.4 多个药物-疾病双层网络的邻接矩阵

基于药物-疾病关联和多种药物相似性矩阵,本文构建了 p 个双层异构网络的邻接矩阵 $\boldsymbol{H}^p \in R^{N_{total}*N_{total}}$, 式(3):

$$\boldsymbol{H}^p=\begin{bmatrix} S_p^r & A^{rd} \\ (A^{rd})^{\mathrm{T}} & S^d \end{bmatrix} \tag{3}$$

其中, $(A^{rd})^{\mathrm{T}}$ 是 A^{rd} 的转置矩阵。

1.5 基于元路径的成对拓扑结构编码

本文构建的双层异构网络 $\boldsymbol{H}^p$, 包含药物和疾病节点。多重关系也包括在内, $r-r,d-d,r-d$ 表示药物-药物相似性,疾病-疾病相似性以及药物和疾病之间的关联关系。在异构图中,许多节点可以通过具有不同语义的路径连接,被称为元。长度为 m 的元路径定义为式(4):

$$v_1 \xrightarrow{n_1} v_2 \xrightarrow{n_2} \cdots \xrightarrow{n_p} v_{m+1} \tag{4}$$

其中, $v_1,v_2,\cdots,v_{m+1}$ 表示节点类型, $n_1,n_2,\cdots,n_p$ 表示连接 v_1 和 v_{m+1} 的边的类型。

一个元路径实例被定义为异构图中的一个节点序列。r_1 和 r_4 可以通过元路径 $r-r-r$ 和 $r-d-r$ 的方式连接。例如,目标节点 r_1 的元路径 $r_1-r_2-r_4$, 如果药物 r_1 和 r_4 都有 r_2 类似的功能,其可能是相似的;在 $r_1-d_3-r_4$ 中,这两种药物都和疾病关联,表明 r_1 可能和 r_4 相似。不同的元路径显示出不同的语义信息。考虑到药物 r_i 的直接邻居和经过两跳之后的邻居对其影响较大。因此,本文建立长度为 1 的元路径和长度为 2 的元路径 $\delta \in \{r-r,r-d,r-r-r,r-d-r,r-r-d,r-d-d\}$。同样的,对于疾病节点 d_j, 分别建立长度为 1 和长度为 2 的元路径 $\delta \in \{d-r,d-d,d-r-r,d-d-r,d-r-d,d-d-d\}$。用 $P_r^\delta(P_d^\delta)$ 表示药物(疾病)节点的元路径。

基于 H^p 结构信息,药物和疾病之间存在各种连接关系, $\varphi \in \{r-r,r-d,d-d,d-r\}$。元路径与这些关系相对应的邻接矩阵表示为 $\boldsymbol{X}^k \in R^{N_{total}*N_{total}}$, 其中 $k \in \varphi$。以 $k=r-d$ 为例,$\boldsymbol{X}^k$ 被定义为式(5):

$$\boldsymbol{X}^k=\begin{bmatrix} 0 & A^{rd} \\ 0 & 0 \end{bmatrix} \tag{5}$$

当且仅当节点 i 和 j 之间存在 $r-d$ 的关系时, $X_{ij}^{r-d}=1$, 否则为 0。

对于每条元路径,都要建立其相应的拓扑嵌入。对于元路径 k 包括在 δ 中的第 e 个关系的邻接矩阵 $\boldsymbol{X}^{k_e}$ 被归一化为 $\tilde{X}^{k_e}$, 式(6):

$$\tilde{X}_{ij}^{k_e}=\begin{cases} \dfrac{X_{ij}^{k_e}}{\sqrt{O_i}\sqrt{Z_j}} \\ 0 \end{cases} \tag{6}$$

其中, O_i 表示第 i 行元素之和, Z_j 表示第 j 列元素之和。元路径 δ 的拓扑嵌入是 T, 式(7):

$$T_= \tilde{X}^{k_1}\tilde{X}^{k_2}\cdots\tilde{X}^{k_\delta} \tag{7}$$

其中, $|\delta|$ 是长度。

例如,元路径 $r-r-d$ 的长度为 2,相对应的拓扑嵌入 $T_{r-r-d}=\tilde{X}^{r-r}\tilde{X}^{r-d}$。在不同的元路径下 $P_r^\delta(P_d^\delta)$, $T^p_{r,\delta}(T^p_{d,\delta})$ 表示药物(疾病)节点基于 $\boldsymbol{H}^p$ 下的拓扑嵌入,其中 $p \in \{chem,doma,anno\}$。

1.6 多种语义信息的融合

给定药物 r_i (疾病 d_j) 的元路径 $P_r^\delta(P_d^\delta)$, 其特定的语义表示为 $(T^p_{r,\delta})_i((T^p_{d,\delta})_j)$。每一个元路径都反映了一个特定的语义信息,对构造药物(疾病)节点的拓扑嵌入有着明显不同的贡献。因此,本文提出了一个元路径层面的注意力机制,有助于融合多种语义。以药物节点为例,元路径类型层面的注意力得分为 $(s^p_{r,\delta})_i$, 式(8):

$$(s^p_{r,\delta})_i=q^{\mathrm{T}}\tanh(M^{atte}(T^p_{r,\delta})_i+b^{atte}) \tag{8}$$

其中,tanh 表示一个非线性激活函数; $\delta \in \{r-r,r-d,r-r-r,r-d-r,r-r-d,r-d-d\}$; b^{atte} 是注意力参数; q^{T} 是可学习参数。

$(\beta^p_{r,\delta})_i$ 代表归一化的注意力权重,式(9):

$$(\beta^p_{r,\delta})_i=\frac{\exp((s^p_{r,\delta})_i)}{\sum_{s\in\delta}\exp((s^p_{r,s})_i)} \tag{9}$$

药物节点的拓扑表示 $(h_r^p)_i$ 通过元路径层面注意力机制增强后定义如下,式(10):

$$(h_r^p)_i=\sum_\delta(\beta^p_{r,\delta})_i(T^p_{r,\delta})_i \tag{10}$$

类似地，也得到了疾病 d_j 在不同元路径聚合下的拓扑表示 $(h_d^p)_j$。

1.7　基于邻居层面注意力机制的属性补全

给定一对药物和疾病节点 (r_i, d_j)，和其相对应的节点拓扑嵌入表示 $(h_r^p)_i$ 和 $(h_d^p)_j$，本文用 V_r^+ 表示所有与疾病 d_j 相关联的药物节点的集合，其中药物 r_i 具有节点属性，疾病 d_j 不具有节点属性。通过对与疾病节点 d_j 直接相连的药物节点的属性加权聚合作为疾病节点 d_j 的属性，实现对疾病节点 d_j 的属性补全，属性补全的示意图如图 1 所示。

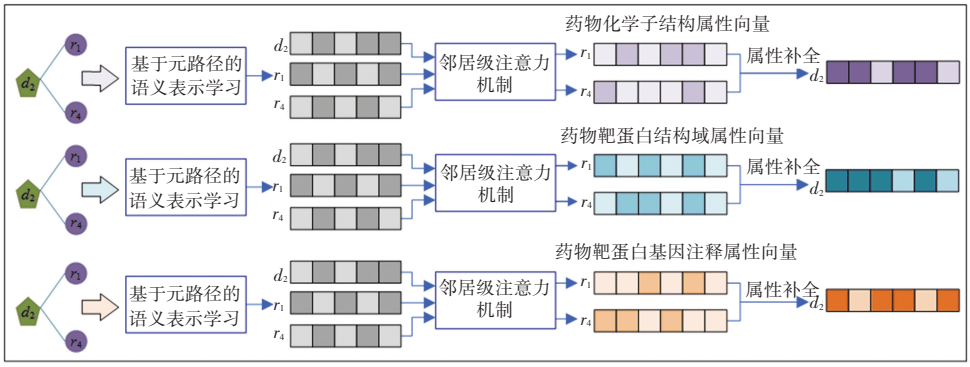


图 1　基于疾病节点属性补全的示意图

Fig. 1　Schematic of disease node attribute complement

因为局部拓扑结构不同，每个节点的邻居在属性聚合的重要性不同，也就是一个节点的邻居越多，其对每个邻居的重要性就越低。因此，本文提出一个邻居层面的注意力机制来学习节点不同邻居的重要性，式(11)：

$$e_{ij}^p = \sigma((h_d^p)_j^{\mathrm{T}} \boldsymbol{W}_c^p (h_r^p)_i) \tag{11}$$

其中，σ 是激活函数，$\boldsymbol{W}_c^p$ 是权重矩阵。

归一化注意力权重 a_{ij}^p 表示如式(12)：

$$a_{ij}^p = \frac{\exp(e_{ij}^p)}{\sum_{s \in V_r^+} \exp(e_{sj}^p)} \tag{12}$$

最终利用注意力机制对与疾病 d_j 相连的药物节点的属性加权聚合对疾病 d_j 实现了属性补全，补全后的属性向量 $\boldsymbol{X}_j^p$ 定义为式(13)：

$$\boldsymbol{X}_j^p = \sum_{i \in V_r^+} a_{ij}^p (S_p^r)_i \tag{13}$$

本文还建立了多头注意力机制，用来稳定属性补全的学习过程，式(14)：

$$X_j^p = mean\left(\sum_k^K \sum_{i \in V_{keep}^r} a_{ij}^p (S_p^r)_i\right) \tag{14}$$

由于药物节点具有 3 种药物属性，根据不同的属性分别对疾病节点进行属性补全。最后，疾病节点的属性矩阵被表示为 $\boldsymbol{X}_p^d (p \in \{chem, doma, anno\})$。为了使属性补全过程是可学习的，同时保证补全的属性的准确性，按照比例 μ 将药物节点 V^r 随机划分为两个部分，分别是 V_{keep}^r 和 V_{drop}^r，其中 $|V_{drop}^r| = \mu |V^r|$，删除掉 V_{drop}^r 中药物节点的属性，通过 V_{keep}^r 对丢掉属性的节点进行属性补全，计算得到的节点 V_{drop}^r 的重构属性定义为式(15)：

$$X_j^p = mean\left(\sum_k^K \sum_{i \in V_{keep}^r \cap V_r^+} a_{ij}^p (S_p^r)_i\right) \tag{15}$$

为了使重构的属性尽可能的接近于原始属性，通过计算原始属性和重构属性之间的欧氏距离得到属性补全的监督损失 $loss_c^p$，式(16)：

$$loss_c^p = \frac{1}{|V_{drop}^r|} \sum_{i \in V_{drop}^r} \sqrt{((\boldsymbol{X}_p^d)_i - (S_p^r))^2} \tag{16}$$

通过属性补全机制，对已有的药物节点属性和补全的疾病节点属性进行组合，得到了关于药物和疾病节点的属性矩阵 $\boldsymbol{X}_p^{new}$，式(17)：

$$\boldsymbol{X}_p^{new} = \begin{bmatrix} S_p^r \\ X_p^d \end{bmatrix} \tag{17}$$

1.8　最终整合和预测

通过属性补全机制，得到 p 个药物-疾病节点的属性矩阵 $\boldsymbol{X}_p^{new}$，其中药物节点 r_i 的属性表示为 $(X_p^{new})_{r_i}$，疾病节点 d_j 的属性表示为 $(X_p^{new})_{d_j}$。为了利用每个属性矩阵的特征，将其降维到相同的维度后上下堆叠，用 1×1 卷积进行融合，得到 $r_i - d_j$ 最终的属性向量表示 t，并将其作为全连接层的输入，以得到药物 r_i 和疾病 d_j 的关联得分。

2　实验结果与分析

2.1　评价指标

本文使用五倍交叉验证法来评估基于属性补全预测模型的性能。将所有已知的关联关系视为正例样本，并随机分为 5 组，其中 4 组用于训练，另一组

用于测试。将所有未观察到的药物-疾病相关性视为反例样本。随机选择与正例样本数同等数量的反例样本进行训练,剩余的反例样本进行测试。

评估指标包括受试者操作特征(*ROC*)曲线、*ROC* 曲线下的面积(*AUC*)、精确召回曲线(*PR* 曲线)、*PR* 曲线下的面积(*AUPR*)。真阳率(*TPR*)和假阳率(*FPR*)的计算,式(18)和式(19):

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

$$FPR = \frac{FP}{TN + TP} \tag{19}$$

其中,*TP*(*TN*)表示正确预测的正例(反例)样本数,*FP*(*FN*)表示错误预测的正例(反性)样本数,用来计算绘制 *ROC* 曲线,该曲线是以 *TPR* 为纵坐标,*FPR* 为横坐标,其曲线下方的面积表示为 *AUC* 值,用于评估模型的性能。*AUC* 值越高代表模型的性能越优秀。

精确度和召回率是评估机器学习模型性能的重要指标。精确度表示预测为正例样本中真实正例样本的比率,式(20);而召回率表示在所有正例样本的样本中被正确识别为正例样本的比率,式(21)。

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

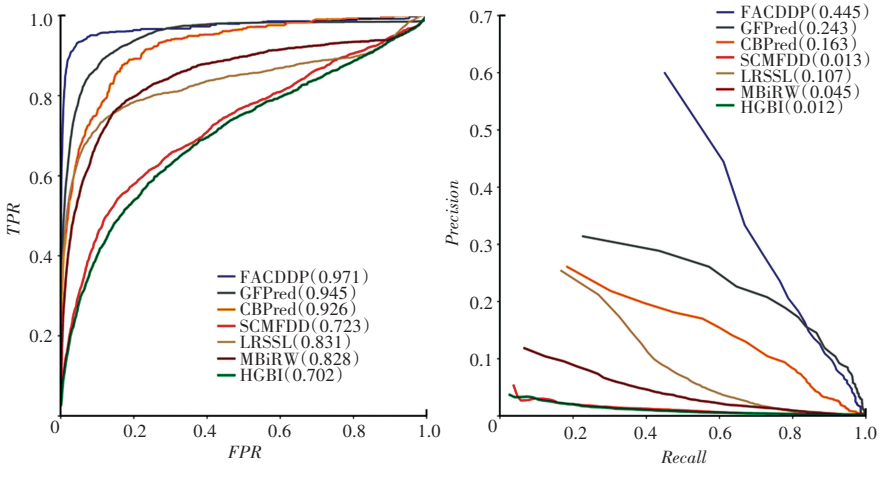
$$Recall = \frac{TP}{TP + FN} \tag{21}$$

通过绘制以 *Precision* 为纵轴、*Recall* 为横轴的曲线,可以直观地展示模型的性能。如果这条曲线处于左上角附近,那么就意味着模型的性能更佳。相反,曲线越靠近右下角则意味着模型性能越差。

2.2 与其他方法的比较

为了评估基于属性补全预测模型的性能,将本文提出的方法与 6 种最先进的有关药物疾病关联预测的方法进行比较,包括 GFPred、CBPred、SCMFDD、LRSSL、MbiRW 和 HGBI。为了使比较结果更具说服力,本文的模型和所有比较的模型在训练和测试时使用了相同的数据集,并且每种对比方法的最佳性能是通过使用各自文献中提供的最优参数设置。

在五倍交叉验证中,本文对 763 种药物进行了评估,并计算了各自的平均 *AUC* 和 *AUPR*;最终将所有 763 种药物的平均 *AUC*(或 *AUPR*)作为最终结果。不同预测模型的 *ROC* 曲线与 *PR* 曲线如图 2 所示。在所有方法中,基于属性补全的预测模型取得了最佳的性能,优于其他对比模型;GFPred 在性能上排名第二,其从多个异构网络中学习,获得药物(疾病)节点的拓扑表示,该结果表明,融合多个异构网络的信息可以提高预测性能;CBPred 考虑了节点对之间的路径信息,在性能上排名第三;尽管 LRSSL 和 MBiRW 的 *AUC* 没有明显差异,但 LRSSL 的 *AUPR* 明显更高,这是因为前者利用了多种药物的相似性,而后者只考虑一种药物的相似性。SCMFDD 和 HGBI 的性能稍差,其 *AUC* 和 *AUP* 几乎没有差别,这是因为两者都没有利用多种药物的相似性。与上述方法相比,基于属性补全的预测模型的性能提高主要是通过多个不同的路径,捕获了药物和疾病节点的多种拓扑结构表示,并基于这些拓扑信息通过注意力机制对疾病节点进行属性补全。



(a) *ROC* curves (b) *PR* curves

图 2 不同预测模型的 *ROC* 曲线与 *PR* 曲线(分图)

Fig. 2 *ROC* curves and *PR* curves of different prediction methods

(下转第 178 页)