DOI:10.3969/j.issn.1005-202X.2023.06.018

医学生物信息

基于图神经网络协同过滤的药物疾病关联预测

陈浩[1,2],秦玉芳[1,2],陈明[1,2]
1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306

【摘要】新药开发需要耗费很高的成本,建立高效且高准确度的药物新适应症预测方法非常重要,提出一种基于图神经网络协同过滤的药物疾病关联预测方法,获取药物与疾病治疗关系中的信息并结合药物相似性获得更好的预测表现。首先通过图神经网络提取药物-疾病治疗关系数据中的协作信号细化药物嵌入,然后利用嵌入计算药物之间的治疗关系相似性,再结合药物化学结构、蛋白质和副作用相似性预测药物的新作用。与现有的协同过滤方法在相同数据集上进行对比,本文方法获得了较高的预测精确率(0.664 8)。所提出的获取药物-疾病治疗关系中的潜在信息并结合相似性进行药物疾病关联预测的策略是有效的,有助于发现药物的新适应症,为药物开发提供帮助。

【关键词】药物;药物新适应症;药物疾病关联;图神经网络
【中图分类号】R318;R95 【文献标志码】A 【文章编号】1005-202X(2023)06-0780-08

Collaborative filtering based on graph neural network for drug-disease association prediction

CHEN Hao[1,2], QIN Yufang[1,2], CHEN Ming[1,2]
1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Shanghai 201306, China

Abstract: Development of new drugs takes a long time and is high-cost. Hence it's critical to have efficient and precise methods for predicting new indications for drugs. A drug-disease association prediction method based on graph neural collaborative filtering is proposed in an attempt to obtain information in drug-disease treatment relationships and combine with drug similarities for obtaining better prediction performance. The proposed method firstly capture collaborative signals in drug-disease treatment relationships through graph neural network to refine drug embeddings, then use the drug embeddings to calculate the similarities in drug-disease treatment relationships between drugs, and finally combines with the similarities in drug chemical structures, proteins, and side effects to predict drug repurposing. Compared with the existing collaborative filtering methods on the same data set, the proposed method achieves a higher prediction accuracy (0.664 8). The proposed strategy to obtain potential information in drug-disease treatment relationships and combine with similarities for drug-disease association prediction is effective and helps to discover new indications for drugs and provides assistance in drug development.

Keywords: drug; new indication for drug; drug-disease association; graph neural network

前 言

药物疾病关联预测表示为通过已知药物找到新治疗途径的过程[1],是网络药理学的重要应用领域之一[2],它相比于新药开发更能节省时间和成本[3],能够帮助制药研究人员进行更有针对性的实验[4]。相关研究表明,传统方法开发一种药物需要10~15年,预计花费8~10亿美元甚至25亿美元[4-5]。而带来的产出与投入不成比例,同时,即使研究出了新药,药物毒性等一系列检测也需要花费很长的时间[6],因此药物新作用开发获得了极大的关注。随着生物医学和制药数据成倍地增长,以及机器学习技术的快速发展,药物新作用开发已成为世界关注的热点[7]。

机器学习方法经过十几年的发展,利用其超强的学习能力寻找潜在的药物-疾病相互作用变得至关重要[8]。在基于机器学习的药物疾病关联预测研究

【收稿日期】2023-01-08
【基金项目】国家自然科学基金(61702325);上海市科技创新计划项目(20dz1203800);广东省重点领域研发计划项目(2021B0202070001)
【作者简介】陈浩,硕士在读,主要从事机器学习和生物信息方向研究,E-mail: chwww140416@163.com
【通信作者】秦玉芳,副教授,主要从事机器学习和生物信息方向研究,E-mail: yfqin@shou.edu.cn

中，不少模型已经使用了协同过滤方法，该方法多应用于推荐系统，在药物疾病关联预测中使用不同样本中的基因表达等数据表示历史趋势，从而预测药物的新适应症[9]。Napolitano 等[10]使用基因表达特征、化学结构和分子靶标多个药物相关数据集计算相似度并组合成药物相似矩阵，使用其训练多类支持向量机（SVM）分类器，最后利用协同过滤预测新的药物适应症。Zhang 等[11]使用药物相似性（蛋白质、化学结构和副作用），疾病相似性（疾病基因和疾病表型）与药物疾病关联构建药物-疾病网络，将网络看作优化问题并使用块坐标下降（Block Coordinate Drop, BCD）进行计算。Yang 等[12]使用药物靶点、通路、通路基因和疾病基因关联来构建因果网络，提出一种因果推理-概率矩阵分解（CI-PMF）方法识别分类药物-疾病关联，将得到的药物-疾病关联排序得分和预测类别识别新的药物-疾病关联。Lim 等[13]使用药物化学结构和靶蛋白数据建立双正则化单类协同过滤模型，该模型推断药物靶点相互作用从而预测药物新作用。Ozsoy 等[14]使用帕累托优势和协同过滤推荐药物-疾病关联。上述方法从药物之间的相关数据计算相似性，或使用药物-疾病关联建立图关系进行预测，很少关注隐藏在已知药物-疾病关联中的协作信号，图神经网络（Graph Neural Networks, GNN）在捕获潜在协作信号有较好的表现[15]，同时它在最近的生物医学网络分析中拥有较好的性能[16]，通过图节点之间的消息传递捕捉图的结构信息，在 miRNA-疾病关联预测[17]、多药副作用预测[18]和 miRNA-耐药性关联预测[19]中都有应用。

笔者提出一种使用 GNN 增强协同过滤的药物疾病关联预测方法。本文方法通过引入高阶连通性提取药物-疾病治疗关系中的协作信号细化药物嵌入向量，使用嵌入向量计算药物治疗关系亲和度，并将得到的亲和度与其他药物相关数据（蛋白质、化学结构和副作用）计算的相似度进行融合，使用基于帕累托优势的方法选择出药物的相似邻居。最后根据邻居药物预测新的药物-疾病关联。本文方法在预测性能上有较大的提升。

1 数　据

1.1 数据集及预处理

本文使用 Zhang 等[20]提供的黄金数据集，Li 等[1]和 Ozsoy 等[14]的研究中也使用该数据集，数据集包含 3 种药物相关数据和药物-疾病相互作用数据，药物相关数据包括化学结构、蛋白质和副作用信息。

化学结构数据从 PubChem 中提取，PubChem 是一个广泛使用的化学分子及其生物活性的数据库。化学结构数据包含 1 007 种药物的 881 个化学结构信息，共 122 022 个药物-化学子结构之间的关联。每个药物都由一个 881 维的二元载体表示，分别以 1 或 0 编码表示其化学结构存在或不存在。数据集的稀疏度为 86.25%。

蛋白质数据从 DrugBank 中提取，DrugBank 是一个药品信息公共数据库。数据集包含 1 007 种药物的 775 个靶蛋白特征载体，包括 3 152 个已知的药物-靶蛋白关联。每种药物的蛋白质信息由一个 775 维的二元载体表示，其中药物靶蛋白关系存在和不存在分别由 1 和 0 表示。数据集的稀疏度为 99.59%。

副作用数据从 SIDER 数据库中提取，SIDER 是一个关于药物副作用的分散公众信息的集合。数据集包含 888 种药物和 1 385 种副作用信息，共 61 102 个药物-副作用关联。每种药物被表示为一个 1 385 维的二元载体，其中副作用的存在或不存在分别用 1 或 0 编码。数据集的稀疏度为 95.03%。

药物-疾病相互作用数据从 National Drug File-Reference Terminology（NDF-RT）中检索，在化学结构数据检索的 1 007 种药物中有 799 种药物存在于药物靶标中，因此构建了包含 799 种药物和 719 种疾病之间的相互作用关系二元矩阵，其中行标为药物，列标为疾病，治疗关系存在和不存在分别用 1 和 0 编码。共包含 3 250 个药物-疾病相互作用。数据集的稀疏度为 99.43%。

笔者发现，药物-疾病相互作用数据中的 799 种药物并非都在药物相关数据（化学结构、蛋白质和副作用）中列出，将药物名称一一对应之后得到 781 种药物和 719 种疾病的 3 179 个相互作用关系。同时在药物-疾病相互作用数据中，存在无关联药物疾病或关联药物数量唯一的疾病，关联药物少的疾病不利于药物治疗关系亲和度的计算，首先排除可治疗药物只有一种的疾病，再排除无疾病关联的药物，最终得到 774 种药物和 430 种疾病的 2 961 种相互作用，用于药物治疗关系亲和度的计算。

2 方　法

2.1 方法设计

本文模型主要由两个过程组成，第一个过程用于药物-药物亲和度计算，第二个过程用于药物-疾病相互作用推荐。模型结构如图 1 所示。在第一个过程中，首先初始化药物及其相互作用疾病嵌入向量的嵌入层，再通过高阶连通信号注入来不断优化药物嵌入向量的传播层，最后集合所有嵌入向量计算药物-药物亲和度。第二个过程中，根据药物之间的亲和度和相似度选择邻居药物，并输出推荐列表。

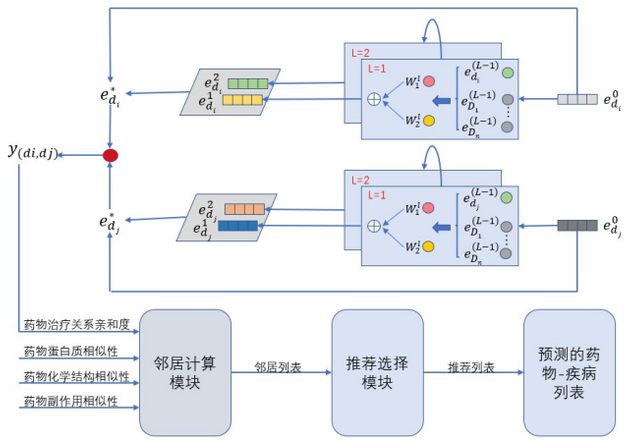


图1 方法设计

Figure 1 Method design

2.1.1 Embedding层　笔者定义Embedding(嵌入)向量$e_d \in R^s$描述药物，$e_D \in R^s$描述疾病，其中s表示嵌入大小[21]。接着构建一个嵌入表：

$$E = \left[e_{d_1}, \cdots, e_{d_M}, e_{D_1}, \cdots, e_{D_N} \right] \tag{1}$$

这个嵌入表作为药物嵌入和疾病嵌入的初始状态。笔者通过在药物-疾病交互图上传播嵌入向量来不断优化改进，并将协作信号注入嵌入向量。在之后的计算中，$e_d^{(0)} = e_d, e_D^{(0)} = e_D (e_d, e_D \in E)$。

2.1.2 传播层　传播层建立在GNN的消息传递框架上[22]，利用图结构的高阶连通性获取协作信号优化药物和疾病的嵌入。如图2所示，目标药物为d_1，左侧为药物-疾病相互作用交互图，右侧为由d_1展开的树状结构图。高阶连通性表示路径长度l大于1的任意节点到达d_1的路径。例如$d_1 \leftarrow D_2 \leftarrow d_2$可以表示目标药物$d_1$和药物$d_2$的相似性，因为二者都与疾病$D_2$关联；而相比于$d_3$，$d_2$与$d_1$更亲和，因为$d_2$有$d_1 \leftarrow D_2 \leftarrow d_2$和$d_1 \leftarrow D_3 \leftarrow d_2$两条路径；在寻找药物新作用策略中，某些相似的药物可以治疗同一疾病[23]，所以在药物-疾病关联的推荐上，更倾向于将d_2的关联药物D_5推荐给目标药物。笔者设计嵌入传播层，通过叠加多个传播层，从高阶连通性中获取协作信号，聚合关联疾病嵌入来优化目标药物的嵌入向量。

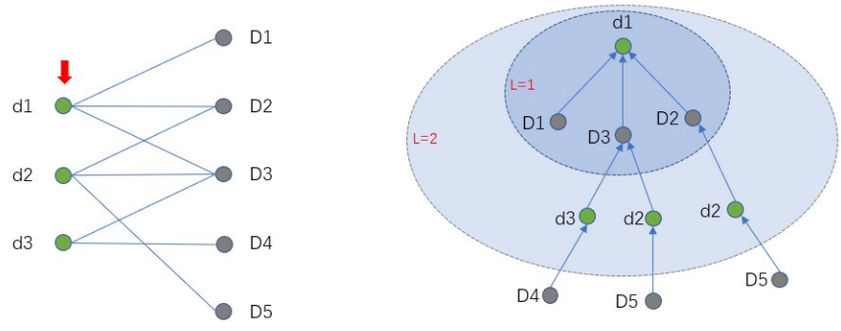


图2 高阶连通性

Figure 2 High-order connectivity

在计算药物治疗关系亲和度时，药物可以治疗某个疾病被视为该药物的一个特征，可以用来衡量两个药物的相似性[24]。那么在此基础上堆叠L个嵌入传播层，药物可以通过高阶连通性获取到L跳邻居传播的消息(编码后的协作信号)，这对最后计算药物治疗关系亲和度十分重要。每个传播层，将连接的药物和疾病之间执行嵌入传播主要有两个过程，包括消息构造和消息聚合。

L层的消息构造部分，对于连接的药物-疾病(d, D)，笔者将从疾病(D)到药物(d)的消息定义为：

$$m_{D \to d}^{(L)} = f\left(e_d^{(L-1)}, e_D^{(L-1)}, p_{(d,D)} \right) \tag{2}$$

其中，$m_{D \to d}^{(L)}$表示要传播的消息向量，f是以$e_d^{(L-1)}$和$e_D^{(L-1)}$为输入的消息编码函数，使用系数$p_{(d,D)}$控制每次传播的衰减因子。具体计算公式如下：

$$m_{D \to d}^{(L)} = \frac{1}{\sqrt{\left|N_d\right|\left|N_D\right|}} \left(\boldsymbol{W}_1^{(L)} e_D^{(L-1)} + \boldsymbol{W}_2^{(L)} \left(e_D^{(L-1)} \odot e_d^{(L-1)} \right) \right) \tag{3}$$

其中，$\boldsymbol{W}_1^{(L)}$和$\boldsymbol{W}_2^{(L)}$表示为可训练权重矩阵，用来提取有用的传播信息。$e_d^{(L-1)}$表示上一嵌入传播层消息传递生成的药物d的向量，$e_D^{(L-1)}$表示上一嵌入传播层消息传递生成的疾病D的向量，它有来自$L-1$跳邻居的消息。在这里，笔者不仅考虑了e_d的贡献，还将e_d和e_D的相互作用编码入消息传递当中[25]，$\odot$表示元素乘积。在图卷积网络的基础上[16]，将系数$p_{(d,D)}$设置为图拉普拉斯范数$1/\sqrt{\left|N_d\right|\left|N_D\right|}$，其中$N_d$和$N_D$表示药物$d$和疾病$D$第一跳的邻居。$p_{(d,D)}$反映疾病对药物偏好的贡献程度，同时从消息传递角度，考虑到传播的消息应该随着路径长度而衰减，$p_{(d,D)}$也可表示为折扣因子。

L层的消息聚合部分，笔者通过从药物(d)邻居传入的消息进行聚合来改进嵌入向量，具体公式如下所示：

$$e_d^{(L)} = \text{LeakyReLU}\left(m_{d\to d}^{(L)} + \sum_{D\in N_D} m_{D\to d}^{(L)}\right) \tag{4}$$

其中，$e_d^{(L)}$表示经过L层嵌入传播层之后得到的药物(d)的嵌入向量。LeakyReLU激活函数允许消息对正信号和小负信号进行编码[26]。同时笔者还考虑了药物(d)的自连接：

$$m_{d\to d}^{(L)} = \boldsymbol{W}_1^{(L)} e_d^{(L-1)} \tag{5}$$

这里与式(3)共享权重矩阵$\boldsymbol{W}_1^{(L)}$。同样地笔者也可以通过传播疾病(D)的邻居信息来获得疾病的嵌入向量$e_D^{(L)}$。总之嵌入传播层的功能是显式地利用高阶连接信息来关联药物和疾病的向量表示。

深度学习模型在具有较强表达能力的同时，往往存在过拟合问题，Dropout技术是防止过拟合的有效方法[27]。笔者使用Dropout技术在嵌入传播层以概率p随机丢弃传出的消息，也就是在L层传播时，笔者以概率p随机丢弃式(2)、式(3)传播的消息，只有部分消息帮助嵌入向量优化。

2.1.3 亲和度计算　在L层嵌入传播层计算之后，笔者可以得到药物的多个嵌入向量表示。不同层中获得的嵌入向量传递消息通过不同的连接，因此它们有着不同的贡献，将它们进行连接计算获得最终的药物嵌入向量表示：

$$e_d^* = e_d^{(0)} \text{II} \ldots \text{II} e_d^{(L)} \tag{6}$$

其中，II表示连接操作。最后计算内积来估计药物之间的亲和度：

$$y_{(d_i,d_j)} = e_{d_i}^{*\top} e_{d_j}^* \tag{7}$$

2.1.4 邻居选择　在之前的步骤中，已经计算出药物治疗关系亲和度，对于3种药物相关数据(化学结构、蛋白质和副作用)，笔者使用3种相似度计算方法：余弦相似度、Jaccard相似度和Smith-Waterman序列比对相似度。邻居选择则需要选择出目标药物(待推荐)的邻居，即最相似的药物。笔者使用基于帕累托优势的方法进行选择，对所有目标药物的待选择药物，分别比较它们的相似度值，如果某个药物的相似度值至少有一个比其他药物高，并且没有比其他药物低的相似度值，那么这个药物没有被其他药物支配，即选择为邻居。举个例子，d_1的亲和度和3个相似度值分别为<0.8,0.7,0.8,0.6>，d_2为<0.7,0.4,0.5,0.6>，分别比较得出(0.8>0.7,0.7>0.4,0.8>0.5,0.6=0.6)发现d_1的亲和度和两个相似度高于d_2，且没有低于d_2的相似度值，那么d_1支配d_2，选择d_1为邻居。具体计算公式如下：

$$\text{Dom}_{(d_i,d_j)} = \begin{cases} 1, \forall \text{sim}\ d_i(\text{sim}) \geq d_j(\text{sim}) \text{ and} \\ \quad \exists \text{sim}\, d_i(\text{sim}) > d_j \text{sim} \\ 0, \text{ other} \end{cases} \tag{8}$$

其中，sim为某个相似度值。

在邻居选择过程中，可能会出现第一次选择少于设定邻居个数，因此笔者进行多次邻居选择，将已找到的邻居从待选择表中删除，继续进行邻居选择，直到选择到数量满足设定的邻居个数为止。

2.1.5 推荐输出　通过收集到的邻居列表选择目标药物的关联疾病，使用计算分数的方法选择推荐的疾病，计算公式如下：

$$\text{score}_{(d,D)} = \sum_{n\in \text{Nei}} \text{sim}(d,n) \tag{9}$$

其中，d为目标药物，Nei表示邻居列表，sim为目标药物和邻居药物的相似度值。也就是说，待推荐疾病有多个关联药物在邻居列表中，且关联药物与目标药物相似性较大，则疾病可被推荐给目标药物。

2.2 评估指标

对于模型，笔者使用留一法来评估性能。为了有效地评估方法的预测能力，使用了精确率(Precision)、召回率(Recall)和F1分数3个指标，计算公式如下所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

其中，TP为真阳性，表示预测正确的药物-疾病关系；FP为假阳性，表示预测错误的药物-疾病关系；FN为假阴性，表示预测错误但实际标注的药物-疾病关系；TN为正阴性，表示预测正确实际未标注的药物-疾病关系。

3 结　果

3.1 邻居个数和输出列表大小对预测结果的影响

首先，笔者计算了邻居个数N和输出列表大小K对于预测性能的影响。在相似度组合、传播层数和丢弃概率参数不变的情况下，N和K的大小有1、4、8、12、16和20共6个设置，图3展示的是在不同邻居个数下，精确率和召回率随输出列表大小K的变化折线图。从图3可以看出，较小的K值有较好的精确率，而召回率随着K值的增大而增大。当N>8时，精确率和召回率随着输出列表大小K的变化折线图没有较大区别。邻居个数N的大小关系到获取的信息的多少，N太小则目标药物的相似药物少，可进行预测的疾病较少，虽然可以获得较高的精确率，但召回率会

受到影响；N太大，则目标药物的相似药物中会存在干扰项，但对于预测性能没有较大影响。因此，笔者认为，邻居个数N≥8，本文方法可以获取到足够的邻居信息。

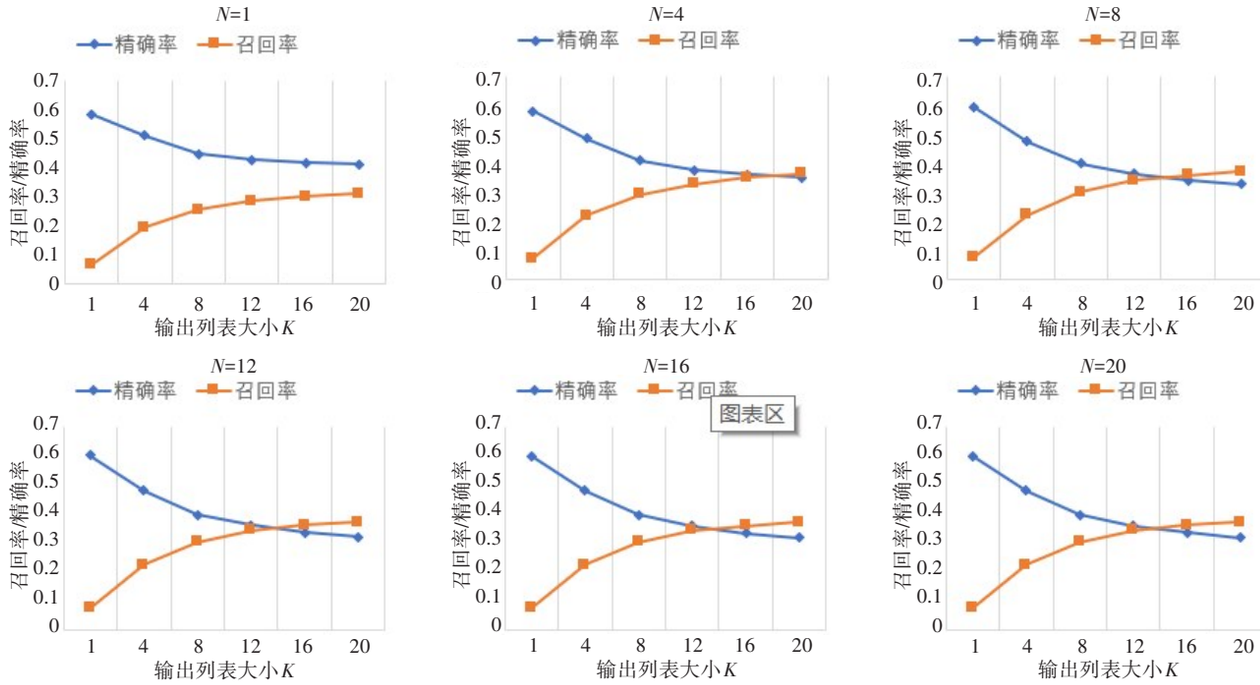


图3 不同邻居个数(N)下，精确率和召回率随输出列表大小K的变化折线图

Figure 3 Variations of accuracy rate and recall rate with the output list size (K) under the constant neighbor size (N)

图4展示的是在输出列表K不同的情况下，精确率和召回率随着邻居个数N的变化折线图。图4可以看出，当邻居个数发生改变时，精确率和召回率没有较大变化。当K≤12时，精确率始终大于召回率，当K>12时，两条曲线出现相交，召回率随着N的增大而增大，精确率随着N的增大而减小。输出列表大小则直接关系预测性能，每个药物的可治疗疾病数目是不同的，如果可治疗疾病数目小于输出列表大小K，那么即便预测出全部可治疗疾病，精确率始终会<1，相反召回率则=1(假设可治疗疾病数为5，K=10，预测出全部可治疗疾病，则精确率=0.5，召回率=1)。K值采用较为宽松的方式，预测结果个数≤K即可。

3.2 层数的影响

为了研究药物治疗关系亲和度能否从多个传播层中获取更多的信息，笔者设置不同的传播层数进行实验，假设传播层数的取值范围是1~4。图5a展示了不同传播层数下，本文方法的性能表现。从图中可以看出1层传播层和2层传播层的结果相差不大，1层传播层只考虑了一阶邻居，可以获取药物可治疗疾病的相关信息，而2层传播层可以获取治疗相同疾病的其他药物的相关信息。当进一步叠加到3层传播层和4层传播层时，性能有明显下降，这可能是由于应用过深的传播层，给学习带来噪音导致精确率下降。因此选择适合的传播层数量既可以获得足够有用的信息又可以避免学习带来的噪音对结果的影响，协同药物间相似度由二阶连通性承载。在后续实验中，笔者采用2层传播层进行实验。

3.3 不同丢弃概率的影响

为了防止深度学习模型在提高表达能力的同时带来的过拟合问题，笔者采用了丢弃法。传播层与传播层之间随机丢弃传播的消息概率为p。图5b展示了在其他参数不变的情况下，针对不同的评估标准，丢弃概率对结果的影响，其中丢弃概率分别取0.1、0.3、0.5、0.7。在精确率性能方面，丢弃概率p=0.1、0.3、0.7时精确率有相似的性能表现；而丢弃概率p=0.5时，精确率最高，可以达到0.66。在召回率方面，丢弃概率p=0.1、0.5时，召回率性能相似，而p=0.3时的召回率表现略低，p=0.7时召回率表现最低。在F1分数上，丢弃概率p=0.1、0.3时F1分数表现相似，在p=0.7时F1分数表现最低，p=0.5时F1分数最高。从上面结果来看，丢弃概率太低，则无法在防止过拟合问题上有较好表现；丢弃概率太高，则无法获取足够信息，因此丢弃概率p=0.5时表现最好。在进行其他参数实验时，取丢弃概率p=0.1。

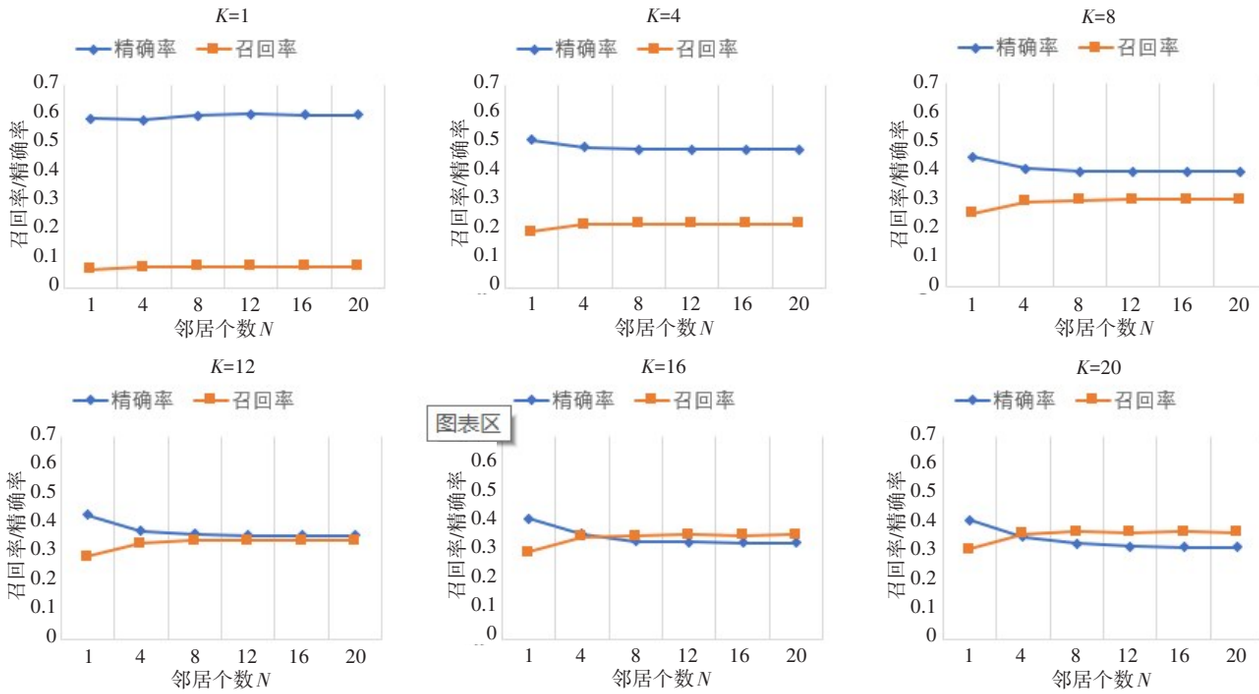
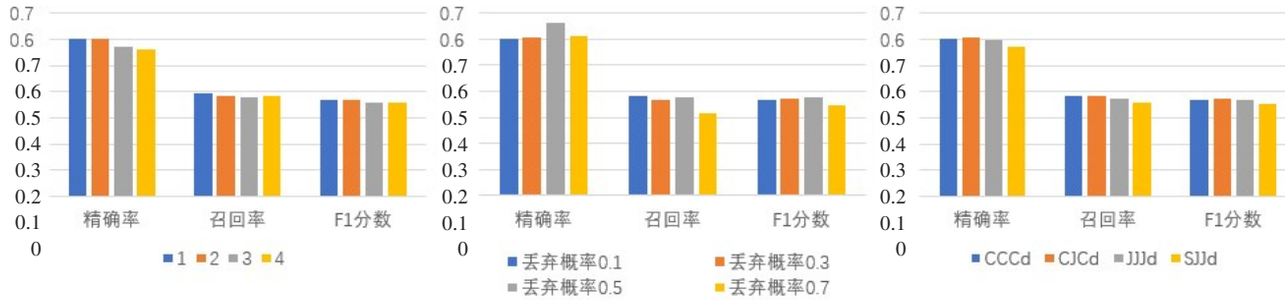


图4 输出列表K不同的情况下,精确率和召回率随着邻居个数N的变化折线图

Figure 4 Variations of accuracy rate and recall rate with the neighbor size (N) under the constant output list size (K)



a:不同传播层数的影响　　b:不同丢弃概率的影响　　c:不同相似度计算方法的影响

图5 参数对性能的影响

Figure 5 Effects of parameters on performance

3.4 相似度计算方法的影响

笔者使用不同的相似性度量来计算3种不同药物相关信息的相似性。对于相似性度量,笔者使用了余弦相似度、Jaccard相似度和基于Smith Waterman序列排列相似度,药物的相关信息有蛋白质特征、化学结构和副作用特征。第1种设置(CCCd)中,笔者使用余弦相似度来计算3个药物相关信息,并结合之前计算的药物间亲和度。第2种设置(CJCd)中,笔者使用余弦相似度计算蛋白质特征和副作用特征,使用Jaccard相似度计算化学结构特征,并与计算的药物间亲和度结合。第3种设置(JJJd)中,笔者全部使用Jaccard相似度计算,结合药物间亲和度。第4个设置(SJJd)中,化学和副作用特征采用Jaccard相似度,蛋白质特征采用基于Smith Waterman序列排列相似度,并结合药物亲和度。性能结果如图5c所示,在计算过程中使用不同的相似性度量方式对性能的影响不大。CCCd设置和CJCd设置略优于其他设置。在其他参数实验中,采用CCCd设置进行实验。

3.5 性能对比

将本文方法和文献中提出的方法进行比较。为了检验本文方法的预测性能,将本文方法和Li等[1]的方法、Zhang等[20]的方法和Ozsoy等[14]的方法在相同的数据集上进行对比实验。如表1所示,本文方法采用相似度组合CJCd时,拥有0.664 8的精确率,均优于其他方法,这表明本文方法更能精确地对药物是否能治疗疾病做出预测。在召回率上,Li等[1]的方法有着最好的召回率(0.770 0),本文方法在召回率上则较低,说明相比于其他方法本文无法预测所有的药物-疾病治疗关系。图

6展示了本文方法和Ozsoy等[14]方法精确率和召回率随着输出列表K的变化曲线，相比于其他方法，本文方法在精确度上有较大的提升，而召回率只有较小的差距，这反映Ozsoy等[14]方法可以预测许多药物疾病治疗关系，但也同时预测了较多的错误药物-疾病相互关系。在使用的数据集上，笔者使用的数据集有2 961个正相关关系，329 859个负相关关系，是正负样本不平衡的数据，正样本少，负样本多，也就是说许多药物和疾病是没有治疗关系的。所以笔者认为，精确率相比于召回率更重要，也就是说相比于找到更多的药物-疾病治疗关系，更精确地找到治疗关系更为关键。对比其他方法，本文方法在损失较小召回率性能下，获得了更高的精确率。

表1 方法性能对比

Table 1 Comparison of method performance

方法	精确率	召回率	F1分数
Li等[1]	-	0.770 0	-
Zhang等[20]	0.345 0	0.651 0	0.451 0
Ozsoy等[14]-CCCd	0.189 4	0.401 7	0.257 5
Ozsoy等[14]-SJJd	0.481 0	0.083 7	0.142 6
本文方法-CCCd	0.347 6	0.378 2	0.362 3
本文方法-CJCd	0.664 8	0.078 6	0.140 7

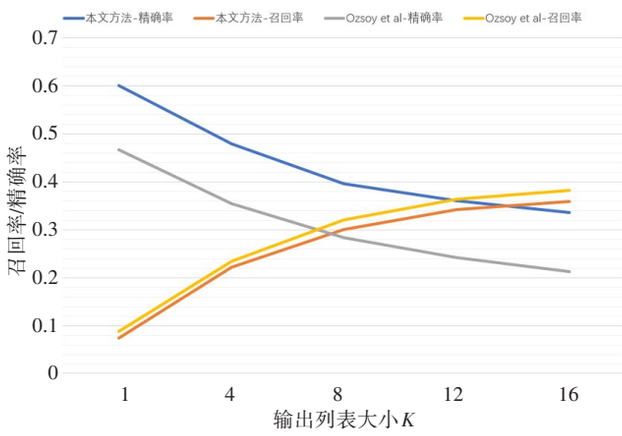


图6 方法性能对比

Figure 6 Comparison of method performance

药物疾病关联预测领域判别模型的优劣需要正确识别药物-疾病的治疗关系，但是预测新的药物-疾病相互作用（假阳性）也很有必要，不仅需要正确发现药物的新作用，也需要给予临床实验方向。同时，笔者使用的数据集缺乏近期的药物-疾病关系的信息。为了进一步验证本文方法的可靠性，笔者利用ClinicalTrials.gov网站和DrugBank网站将本文预测的药物-疾病相互作用假阳性和已经批准的新临床实验研究进行比较，证明了本文方法识别未知药物-疾病相互作用的能力。在本文方法中，精确率最高的是CJCd，预测了244个正确药物-疾病治疗关系（真阳性）和123个未知药物-疾病治疗关系（假阳性），将123个未知关系与ClinicalTrials.gov网站和DrugBank网站中的临床实验进行比较，笔者发现47个未知药物-疾病治疗关系（占比40%）已经进行了临床实验。这表明方法预测的可靠性。例如在ClinicalTrials.gov中，药物罗哌卡因（ropivacaine）和疾病疼痛（pain）存在治疗关系，本文方法成功预测出它们之间的关系，但在数据集中这两者并不存在相关作用。本文方法预测了比卡鲁胺（bicalutamide）对癌细胞（carcinoma）的作用，同时DrugBank中的相关实验也证明了它们存在治疗关系，但是数据集中没有进行标注。表2中列出了预测的47个未知药物-疾病治疗关系中20个关系以及其临床来源。

表2 预测成功示例

Table 2 Examples of successful predictions

药物名	疾病名	临床来源
扑热息痛（acetaminophen）	痛经（dysmenorrhea）	DrugBank
阿莫西林（amoxicillin）	克雷伯氏菌感染（klebsiella infections）	ClinicalTrials.gov
阿莫沙平（amoxapine）	精神失调症（psychotic disorders）	ClinicalTrials.gov
美罗培南（meropenem）	败血症（septicemia）	DrugBank
咪康唑（miconazole）	口腔念珠菌症（candidiasis oral）	DrugBank
磺胺甲醚唑（sulfamethoxazole）	沙眼（trachoma）	ClinicalTrials.gov
罗哌卡因（ropivacaine）	疼痛（pain）	DrugBank
水杨酸（salicylic acid）	疼痛（pain）	ClinicalTrials.gov
维诺瑞林（vinorelbine）	子宫内膜瘤（endometrial neoplasms）	ClinicalTrials.gov
比卡鲁胺（bicalutamide）	癌细胞（carcinoma）	DrugBank
头孢克肟（cefadroxil）	葡萄球菌感染（staphylococcal infections）	ClinicalTrials.gov
克罗尼丁（clonidine）	抑郁症（depressive disorder）	DrugBank
地塞米松（dexamethasone）	炎症（inflammation）	ClinicalTrials.gov
依托咪酯（etomidate）	overdose	ClinicalTrials.gov
磷苯妥英（fosphenytoin）	癫痫（epilepsy）	ClinicalTrials.gov
膦甲酸（foscarnet）	巨细胞病毒视网膜炎（cytomegalovirus retinitis）	ClinicalTrials.gov
伊洛前列素（iloprost）	肺高血压（hypertension pulmonary）	DrugBank
苯妥英（phenytoin）	癫痫状态（status epilepticus）	DrugBank
三甲氧嘧啶（trimethoprim）	恶性疟原虫病（malaria falciparum）	ClinicalTrials.gov
孕酮（progesterone）	肾细胞癌（carcinoma renal cell）	ClinicalTrials.gov

4 讨 论

本文提出了一种GNN增强协同过滤的药物疾病关联预测方法,该方法通过利用GNN嵌入传播层提取潜在的药物-疾病治疗关系中的协作信号来优化嵌入向量,更直观的得到隐藏的药物间的相似性。并且该方法能够有效的预测药物-疾病关联,可以帮助临床药物研究减少时间和金钱的成本,提供未来可能的研究方向。

【参考文献】

[1] Li J, Lu Z. A Network approach for computational drug repositioning[C]//Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology. IEEE, 2012: 83.

[2] 张永祥, 程肖蕊, 周文霞. 药物重定位-网络药理学的重要应用领域[J]. 中国药理学与毒理学杂志, 2012, 26(6): 779-786.
Zhang YX, Cheng XR, Zhou WX. Drug repositioning-an important application area in Network pharmacology[J]. Chinese Journal of Pharmacology and Toxicology, 2012, 26(6): 779-786.

[3] Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations[J]. Nat Rev Drug Discov, 2019, 18 (1): 41-58.

[4] Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning [J]. Brief Bioinform, 2011, 12(4): 303-311.

[5] Cost of developing a new drug[EB/OL]. [2014-11-18]. http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study.

[6] Drug repositioning[EB/OL]. [2022-02-08]. https://en.wikipedia.org/wiki/Drug_repositioning.

[7] Li J, Zheng S, Chen B, et al. A survey of current trends in computational drug repositioning[J]. Brief Bioinform, 2017, 17(1): 2-12.

[8] Chen HM, Engkvist O, Wang YH, et al. The rise of deep learning in drug discovery[J]. Drug Discov Today, 2018, 23(6): 1241-1250.

[9] Shao YW, Zhang J. Computational drug repositioning using collaborative filtering *via* multi-source fusion[J]. Expert Systems with Application, 2017, 84: 281-289.

[10] Napolitano F, Zhao Y, Moreira VM. Drug repositioning: a machine-learning approach through data integration[J]. J Cheminform, 2013, 5(1): 30.

[11] Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity[J]. AMIA Annu Symp Proc, 2014: 1258-1267.

[12] Yang Jh, Li Z, Fan X, et al. Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization[J]. J Chem Inf Model, 2014, 54(9): 2562-2569.

[13] Lim H, Poleksic A, Yao Y, et al. Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing[J]. PLoS Comput Biol, 2016, 12(10): e1005135.

[14] Ozsoy MG, Özyer T, Polat F, et al. Realizing drug repositioning by adapting a recommendation system to handle the process[J]. BMC Bioinformatics, 2018, 19(1): 136.

[15] Wang X, He XN, Wang M, et al. Neural graph collaborative filtering[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieva. 2019: 165-174.

[16] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint, arXiv: 160902907, 2016.

[17] Huang YA, Hu PW, Chan KC, et al. Graph convolution for predicting associations between miRNA and drug resistance[J]. Bioinformatics, 2020, 36(3): 851-858.

[18] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks[J]. Bioinformatics, 2018, 34(13): i457-i466.

[19] Li j, Zhang S, Liu T, et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction[J]. Bioinformatics, 2020, 36(8): 2538-2546.

[20] Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources[C]//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013: 579-594.

[21] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009: 452-461.

[22] Xu K, Li C, Tian Y, et al. Representation learning on graphs with jumping knowledge networks[C]// Proceedings of The International Conference on Machine Learning. 2018: 5453-5462.

[23] Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions[J]. J Cheminform, 2020, 12(1): 46.

[24] Xue F, He XG, Wang X, et al. Deep Item-based collaborative filtering for Top-N recommendation[J]. arXiv preprint, arXiv:1811.04392, 2019.

[25] Ying R, He RN, Chen KF, et al. Graph convolutional neural networks for web-scale recommender systems[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery. 2018: 974-983.

[26] Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models[C]//Proceedings of the 30th International Conference on Machine Learning. 2013.

[27] Berg RV, Kipf TN, Welling M. Graph convolutional matrix completion [J]. arXiv preprint, arXiv: 170602263, 2017.

(编辑:薛泽玲)