计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(2):187−200 [doi: 10.15888/j.cnki.csa.010060] [CSTR: 32024.14.csa.010060]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于异构蛋白质网络随机游走的中药重定位模型①

李政昊[1], 徐　军[2], 陆俊滟[2], 甘　晓[2]

[1](南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
[2](南京信息工程大学 智能医学图像计算江苏高校重点实验室, 南京 210044)
通信作者: 甘　晓, E-mail: xiao.gan@nuist.edu.cn

摘　要: 中药是治疗疾病的重要药物资源, 历经数千年的临床实践与应用. 为推动中药现代化并发掘其在新适应症上的应用潜力, 本文借鉴西药领域药物重定位的研究经验, 结合近年来新兴的网络医学理论, 提出两种基于随机游走的中药-症状潜在治疗关系预测模型: M-RW 与 GO-DREAMwalk. 两种模型分别引入了中药与症状的路径信息和功能信息, 并以此指导随机游走过程, 生成节点序列后输入到异构 Skip-gram 模型, 学习节点的嵌入向量表示. 随后, 结合中药-症状关联标签与嵌入向量训练 XGBoost 分类器, 最终在肝硬化临床数据上对模型进行测试与评估. 在临床有效任务中, 两种模型的高排名预测准确率分别达到了 0.079 8 和 0.068 4, 相较于机制驱动方法 Proximity 分别提升了 145% 与 110%, 相较于数据驱动方法 node2vec 和 edge2vec, 分别提升了 40%、20%, 以及 53%、31%. 此外, 通过 Rank Aggregation 方法聚合两种模型的预测结果, 准确率分别提升了 75% 和 105%, 进一步增强了模型的预测能力. 两种模型在真实临床数据上的预测结果均具备良好的预测性能, 充分展现了其在中药重定位中的应用潜力, 有望推动中药在新适应症上的有效应用.

关键词: 中药; 随机游走; 药物重定位; 网络医学; 图表示学习

引用格式: 李政昊,徐军,陆俊滟,甘晓.基于异构蛋白质网络随机游走的中药重定位模型.计算机系统应用,2026,35(2):187–200. http://www.c-s-a.org.cn/1003-3254/10060.html

Traditional Chinese Medicine Repurposing Models Based on Random Walk of Heterogeneous Protein Networks

LI Zheng-Hao[1], XU Jun[2], LU Jun-Yan[2], GAN Xiao[2]

[1](School of Computer Science & School of Cyber Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Jiangsu Key Laboratory of Intelligent Medical Image Computing, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: As an important therapeutic resource, traditional Chinese medicine (TCM) has undergone thousands of years of clinical practice and application. To promote the modernization of TCM and explore its application potential in new indications, this study draws on research experience from drug repurposing in Western medicine and combines emerging network medicine theories to propose two random walk-based models for predicting potential therapeutic associations between TCM and symptoms: M-RW and GO-DREAMwalk. The two models incorporate path-based and functional information between TCM and symptoms to guide the random walk process. The resulting node sequences are input into a heterogeneous Skip-gram model to learn the embedded vector representations of nodes. Subsequently, an XGBoost classifier is trained by adopting TCM-symptom association labels and the learned embedded vectors. Finally, the models

① 基金项目: 国家重点研发计划 (2023YFC3402800); 国家自然科学基金 (82441029, 62171230, 62101365, 92159301, 62301263, 62301265, 62302228, 82302291, 82302352, 62401272); 江苏省教育强省建设专项资金 (1523142501042); 2024 年江苏省高等教育内涵建设与发展专项 (1311632401002); 江苏省科技厅前沿引领技术基础研究重大项目 (BK2023200)
收稿时间: 2025-07-07; 修改时间: 2025-08-01, 2025-08-15; 采用时间: 2025-08-29; csa 在线出版时间: 2025-11-17
CNKI 网络首发时间: 2025-11-19

are tested and evaluated by employing clinical data on liver cirrhosis. In the clinically effective prediction task, the top-ranking prediction precision of the two models reaches 0.079 8 and 0.068 4 respectively, improvements of 145% and 110% over the mechanism-based Proximity, 40% and 20% over the data-driven method node2vec, and 53% and 31% over the data-driven method edge2vec respectively. Furthermore, applying the Rank Aggregation method to integrate the prediction results of both models leads to precision improvements of 75% and 105%, further enhancing the predictive ability of the models. The prediction results on real-world clinical data of the two models demonstrate sound prediction performance, highlighting their potential to promote the effective application of TCM in novel indications.

Key words: traditional Chinese medicine (TCM); random walk; drug repurposing; network medicine; graph representation learning

中药, 治疗疾病的重要药物资源, 有着千年的临床经验. 然而, 由于其成分复杂及多靶点作用机制, 中药的治疗机理尚未被充分揭示, 从而限制了其在疾病治疗中的有效应用[1]. 在此背景下, 药物重定位可为挖掘传统中药在新适应症上的潜在价值提供思路.

药物重定位, 又称“老药新用”, 是指将已获批的药物应用于治疗新的疾病或适应症的过程[2], 传统上研发一种西药并投入市场是一项耗时且高成本的过程, 通常需要约 8.3±2.8 年, 并耗资约 3.741 亿美元, 且成功率不足 10%, 若将失败的成本计算在内, 平均成本会上升至 13.36 亿美元[3]. 与之相比, 药物重定位可以显著降低研发成本、缩短临床转化周期, 并有效提升研发效率[4]. 经典的案例包括西地那非, 该药物最初被研发作为抗高血压药物, 后被发现对勃起功能障碍具有显著疗效, 以及索马鲁肽, 其原适应症为 II 型糖尿病, 目前已广泛用于肥胖症治疗[4].

药物重定位的主要策略可分为实验驱动、表型筛选、网络医学与临床数据挖掘这 4 大类[5]. 随着对药物和疾病分子机制研究的不断深入, 融合系统生物学与网络科学的“网络医学 (network medicine)”成为药物重定位的重要理论和方法支持. 网络医学利用蛋白质相互作用网络 (protein-protein interaction, PPI) 揭示药物和疾病的作用模式, 其中 PPI 是由蛋白质作为节点, 蛋白质间相互作用作为边构成的网络. 通过系统分析 PPI 上药物靶标与疾病相关蛋白之间的网络关系, 有助于揭示药物的治疗机制[6], 从而推动有治疗潜力药物的发现. 基于网络医学理论, Sun 等人[7]提出的 AdaDR 模型, 已经成功预测了阿尔兹海默症和乳腺癌的候选药物. 因此, 借鉴西药领域药物重定位的研究经验, 并结合网络医学理论, 将重定位方法引入中药研究, 不仅有助于系统性挖掘中药在多种疾病中的潜在适应症, 也为中医药现代化提供技术支持. 本文的中药重定位研究聚焦于网络医学.

基于网络医学的药物重定位中, 基于药物靶标与疾病关联蛋白间网络拓扑距离的策略因具备良好的生物机制可解释性而受到广泛关注[8]. 然而, 这类方法通常仅关注局部距离信息, 难以充分捕捉网络中复杂的拓扑特征, 可能限制其预测性能[9]. 近年来, 图表示学习方法也逐渐应用于网络医学中, 其本质是将网络中的节点 (如药物、蛋白质、疾病) 转化为低维向量, 通过学习节点的结构位置、邻居关系, 提升预测的准确性[4]. 然而, 该类方法普遍存在可解释性不足的问题, 在一定程度上限制了已有生物医学知识的有效利用, 也难以揭示预测结果背后的生物机制[4,10].

当前基于网络医学的中药重定位面临机制驱动方法难以充分挖掘网络结构信息, 数据驱动方法缺乏可解释性的问题, 亟需融合机制与数据优势以提升预测性能. 为此, 本文提出两种中药-症状机制信息指导的随机游走模型: M-RW (metapath-guided random walk) 与 GO-DREAMwalk (GO-informed DREAMwalk). M-RW 模型在随机游走模型的基础上引入中药与症状的路径信息, GO-DREAMwalk 在 DREAMwalk 模型的基础上引入中药与症状的生物功能信息.

本文的主要工作如下.

1) 统计分析中药与症状间的 Proximity 路径长度, 设计代表路径规则的 MetaPath, 并基于 M-RW 模型预测中药-症状关联概率. 该模型引入药症间的路径信息, 提升了关联预测的准确性.

2) 基于 GO 术语间的语义相似性分别构建中药和症状的相似性网络, 并在该多层网络上执行 DREAM-

walk 预测中药-症状的关联概率. 该模型引入生物功能信息, 提升了关联预测的准确性.

3) 将训练好的模型应用于肝硬化临床病人的中药-症状疗效预测, 得到良好的预测结果, 证明本模型在实际应用中具备有效性.

4) 采用 Rank Aggregation 方法对 M-RW 和 GO-DREAMwalk 的预测结果进行融合, 进一步提升模型的预测性能.

本文提出的两种模型不仅保留了随机游走模型在捕捉网络结构特征方面的优势, 还进一步融合了路径信息与生物功能信息, 预测性能得到了进一步提升. 在临床有效任务中, M-RW 与 GO-DREAMwalk 在 Precision@Top1% 指标上, 相较于机制驱动的非机器学习方法 Proximity 分别提升了 145% 和 110%; 相较于数据驱动方法 node2vec 和 edge2vec, 分别提升了 40%、20% 和 53%、31%.

1 相关工作

1.1 网络医学的理论进展

2016 年, Guney 等人[11]提出 Proximity 方法, 用于衡量药物靶点蛋白与疾病关联蛋白在 PPI 中的网络距离, 揭示网络距离越近的药物-疾病对具有更高的治疗潜力. 在此基础上, Ruiz 等人[8]将 GO (gene ontology) 术语与 PPI 结合构建了多尺度相互作用网络 (multi-scale interactome, MSI). 该网络采用有偏随机游走模拟药物和疾病在网络中的扩散过程, 并以扩散图谱间的相似性作为药物-疾病关联的预测依据. 该研究表明, 药物也可以通过靶向影响相同生物功能的远距离蛋白质来治疗疾病. 在中医药领域, Gan 等人[6]将中药靶标与症状关联蛋白映射到 PPI 上并分析其拓扑关系. 通过计算两者的 Proximity 距离, 他们发现靶标与症状模块邻近程度, 预示着中药治疗该症状的有效性, 从而验证了 Proximity 方法在中医药领域中的适用性.

1.2 基于网络医学的随机游走关联预测研究进展

基于网络医学的药物-疾病关联预测是药物重定位研究中的核心任务之一, 其目标是利用网络中的结构信息, 挖掘药物与疾病之间可能存在的潜在关联. 为实现这一目标, 近年来, 越来越多的研究采用图表示学习方法, 通过节点嵌入方式将网络拓扑结构编码为可用于机器学习建模的向量表示. 其中, 随机游走 (random walk) 是一种常见的节点序列生成策略, 它通过设定的转移规则在图中遍历节点, 生成节点序列以捕捉网络拓扑特征[12], 该方法经实践证明可以用于西药-疾病关联预测任务. 下面介绍几种通过随机游走学习网络特征, 从而进行预测药物-疾病关联的方法.

Grover 等人[13]提出 node2vec 模型, 它是一种基于有偏随机游走的嵌入学习方法, 通过控制游走中的返回概率 p 和前进概率 q, 在广度优先 BFS 与深度优先 DFS 策略间灵活切换, 捕捉节点的局部与全局结构. 随后, 通过 Skip-gram 模型训练获得节点的嵌入向量表示, 用于下游的分类或预测任务. 然而, node2vec 是为同构网络设计的, 难以充分表达生物医学领域中节点类型多样、边关系复杂的图结构. 同构网络是由一种类型的节点及其连接关系组成的图结构. 为此, Dong 等人[14]提出了 metapath2vec 模型, 这是一种专为异构网络设计的表示学习方法, 旨在有效捕捉不同类型节点及其关系之间的结构特征和关联信息. 异构网络是由多种类型的节点及其连接关系组成的图结构. 该模型引入了基于领域知识构建的 meta-path, 其本质是由特定的节点类型和关系类型按顺序组成的路径模式. metapath2vec 通过 meta-path 指导的随机游走策略构建节点的异构邻域, 并结合异构 Skip-gram 模型学习节点嵌入表示. 该方法能够使随机游走严格遵循预定义的语义路径, 从而在保持语义一致性的同时生成高质量的上下文序列, 有效提升了异构网络中表示学习的表达能力. Gao 等人[15]认为 metapath2vec 需要领域知识来预设 meta-path, 而且只考虑节点类型而不考虑边类型, 于是他们提出了 edge2vec 模型, 在图表示学习中引入了边的关联信息. 通过期望最大化 (EM) 方法训练边类型转移矩阵, 并利用随机梯度下降模型, 基于该转移矩阵在异构图中学习节点嵌入. 该方法不仅降低了对经验路径设计的依赖, 还有效融合了边的关联信息, 在捕捉异构图结构特征方面表现出更高的灵活性与准确性. Bang 等人[16]指出, 传统的随机游走方法在处理复杂的网络时, 受到蛋白质节点数量庞大, 而药物和疾病节点相对稀少, 导致后者的嵌入表示效果较差. 为此, 他们提出语义信息引导的随机游走模型 DREAMwalk. 语义信息指的是药物或疾病在医学术语体系中的层级分类结构. 这些结构被用于计算药物-药物、疾病-疾病之间的语义相似性分数, 进而构建药物/疾病的相似性网络, 其中节点为药物或疾病, 边为两者之间的语义相似性关系, 边的权重由相似性分数决定. DREAMwalk

将相似性网络与 PPI 网络结合, 构建了一个多层次网络结构. 为提升药物与疾病节点的嵌入质量, 模型在多层网络上执行随机游走时引入了语义信息引导的“跳转”机制: 当游走器访问到药物或疾病节点时, 以一定概率跳转至与当前节点在相似性网络中语义相似的另一个药物或疾病节点. 这一机制提升了药物和疾病节点在游走序列中的出现频率. 随后, 模型利用异构 Skip-gram 方法学习节点嵌入, 并通过 XGBoost 分类器预测药物与疾病之间的潜在关联. 实验结果表明, 与不使用跳转或采用随机跳转策略的模型相比, DREAMwalk 显著提升了预测性能, 且能更好地捕捉具有相似治疗用途但作用机制不同的药物之间的语义关系, 使其在嵌入空间中距离更接近.

1.3 中药-症状关联预测分析

随机游走模型在西药-疾病关联预测中取得了较好的成效, 这为预测中药与症状的关联提供了参考[4]. 由于中药与西药的靶标蛋白均可映射至 PPI 上, 所以我们认为第 1.2 节中提到的经典随机游走模型 node2-vec、引入边类型转移矩阵的 edge2vec 可以直接尝试应用于中药-症状任务中. 此外, metapath2vec 利用预定义的 meta-path 指导异构网络中的随机游走, 进而生成节点的嵌入向量表示[14]. 在中药重定位任务中, 需要结合中药靶点、症状蛋白、功能等多方面的信息, 合理构建能够反映其潜在治疗路径的 meta-path. 另外, 在药物重定位任务中应用 DREAMwalk 模型需要先构建对应的语义相似性网络. 在西药研究中, Bang 等人[16]利用西药的 ATC 分类体系、疾病的 MeSH 术语等作为语义信息, 分别构建药物-药物和疾病-疾病的语义相似性网络. 然而, 中药-症状领域尚缺乏类似成熟的分层语义信息, 无法直接利用 DREAMwalk 实现中药-症状关联预测. 因此, 当前亟需利用文献挖掘、生物功能等间接方式, 构建中药与症状之间的语义相似性网络.

综上所述, 已有的方法为中药-症状关联预测任务提供了建模思路, 但仍需进一步结合中药与症状的机制信息, 设计更契合本任务的模型. 为此, 本文提出了两种模型: M-RW 与 GO-DREAMwalk, 两者分别引入了中药-症状的路径信息与功能信息.

2 方法

M-RW 和 GO-DREAMwalk 模型均由 3 个部分组成, 包括节点序列生成、异构 Skip-gram 生成节点向量表示、XGBoost 预测药症关联. 首先, 本文构建了中药-蛋白质-症状网络 G_{hps} (详见第 3.1 节数据集), 并分别将路径信息和功能信息融入随机游走为模型生成节点序列, 然后使用异构 Skip-gram 模型学习节点的向量表示, 最后训练 XGBoost 分类器预测中药-症状的关联概率. 总体工作模型图如图 1 所示. 图 1(a) 展示在限制路径中蛋白质节点数不超过 3 的条件下, 设定中药-症状的 MetaPath 及其逆路径, 并基于 MetaPath 指导的随机游走生成节点序列; 图 1(b) 展示基于中药-GO 向量矩阵构建中药相似性网络, 以及基于 GO 术语层级结构构建症状相似性网络, 进而在多层网络上执行 DREAMwalk 生成节点序列; 图 1(c) 将随机游走生成的节点序列输入异构 Skip-gram 模型, 学习节点的嵌入向量表示; 图 1(d) 将中药节点向量与症状节点向量的差值作为药症关系向量输入到 XGBoost 分类器, 训练模型, 并在测试集上测试, 输出中药-症状关联的预测概率.

2.1 节点序列生成

2.1.1 M-RW 节点序列生成

本文沿用 Dong 等人[14] meta-path 的思想, 将 MetaPath 定义为由节点类型和边类型按顺序组成的路径模式. 例如, 长度为 3 的路径对应的 MetaPath “herb→protein→protein→symptom” 表示中药通过作用于某一蛋白, 进而影响与症状相关的蛋白, 从而发挥治疗作用. 在模型设计中, 我们通过中药与症状间可能存在的 MetaPath 引导随机游走过程, 使游走器更倾向于沿着具有潜在生物意义的路径采样, 从而提升嵌入向量对药症关系的表达能力与药症关系的预测性能. 已有研究表明, PPI 网络中, 中药靶标与症状关联蛋白的 Proximity 网络距离越近, 通常预示着更高的治疗潜力[6]. MetaPath 中包含的路径信息可能与生物机制有关, 为将该类信息有效融入随机游走模型, 本文基于 Proximity 方法计算药症对之间的最短路径, 并根据其长度分布设计相应的 MetaPath.

考虑到中药-症状对样本规模较大, 为提高计算效率, 我们从 1 436 例明确治疗关系的中药-症状对中随机抽取 200 组进行统计分析. 统计结果如图 2 所示, 中药与症状在 G_{hps} 上的路径长度分布在[2, 6]之间. 其中路径长度为 3 和 4 (即经由 2 或 3 个中间蛋白连接) 数目最多. 我们将路径中除中药与症状节点外的中间蛋白节点数记为 p, 对应路径长度 $L=p+1$.

为了筛选包含丰富路径信息的 MetaPath, 本文采

用逐层递增路径上限的策略, 构建 3 组 MetaPath 集合, 记作 M_3 (p≤3)、M_4 (p≤4)、M_5 (p≤5), 并在后续的中药重定位任务中进行对比评估. 按 p 值划分的 MetaPath 如表 1 所示.

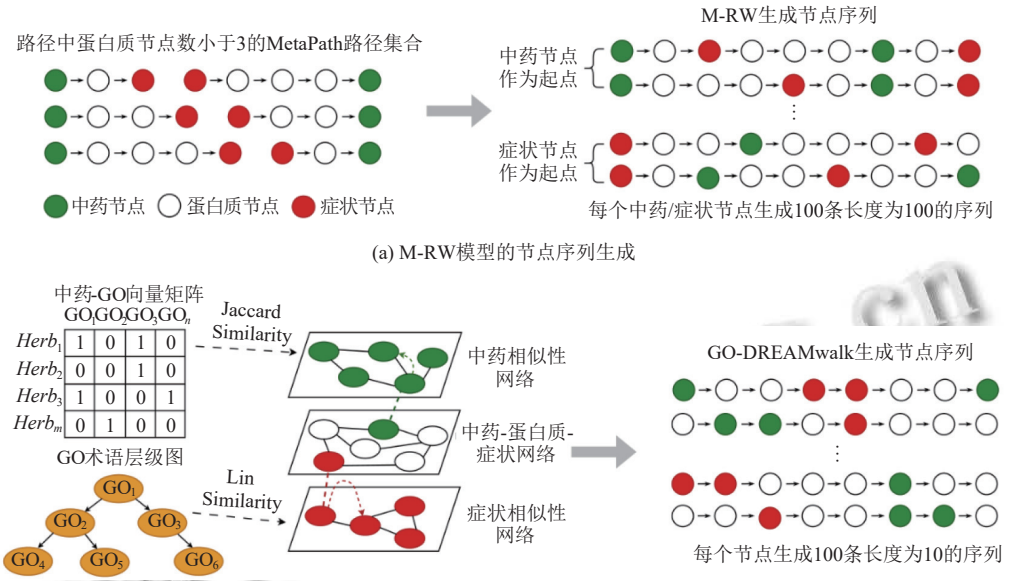


图 1 融入中药-症状路径和功能信息的随机游走模型图

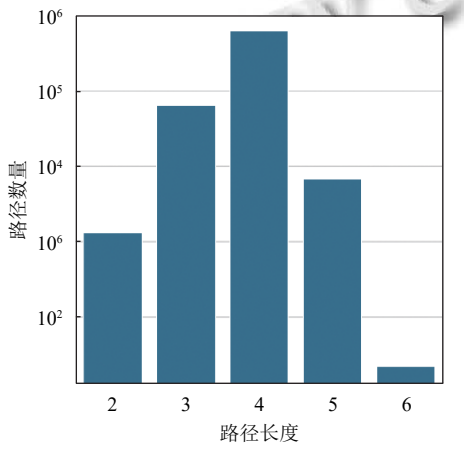


图 2 具有治疗关系的中药-症状对在 G_{hps} 网络中的最短路径长度分布

表 1 按 p 值划分的 MetaPath

p值	MetaPath
1	herb → protein → symptom
2	herb → protein → protein → symptom
3	herb → protein → protein → protein → symptom
4	herb → protein → protein → protein → protein → symptom
5	herb → protein → protein → protein → protein → protein → symptom

根据传统 metapath2vec 方法, 游走器从药物节点出发, 在 meta-path 的指导下执行随机游走, 当成功抵达症状节点时, 即完成该条 meta-path 的遍历, 游走随之终止, 生成较短的节点序列. 但 Gao 等人[15]研究中指出, 依赖于单一 meta-path 的随机游走无法在单次过程

中兼顾多条元路径, 可能导致网络中复杂信息的丢失.

针对这一问题, 我们对所有候选的 MetaPath 路径引入“逆序”路径, 并在每次随机游走中随机选择正序或逆序的路径进行采样. 其中, 逆序 MetaPath 是指从症状节点出发, 经由一系列中间蛋白最终到达中药节点的路径结构; 同时, 将正序路径的终点设为逆序路径的起点, 以此构建连续的“长序列”. 如图 1(a) 所示, 当游走器到达症状节点后, 双向采样策略会随机选择一条逆序 MetaPath 路径, 以当前症状节点作为新的起点, 继续按照逆序路径执行随机游走, 并将新生成的节点序列与前一段正序游走得到的序列进行拼接, 从而形成连贯的长序列.

这种设计既保留了基于经验知识的指导性, 又提升了路径选择的多样性, 同时为 Skip-gram 训练带来多重优势: 其一, 长序列增加了中心词数量, 使模型生成更多训练样本, 缓解短序列 (长度 2–6) 样本较少的问题; 其二, 正序与逆序 MetaPath 的集合, 使模型在窗口范围内同时学习中药-症状关联与症状-中药关联 (如症状上下文同时包含上游中药和下游中药), 融合双向语义关系; 其三, 长序列能使模型充分利用窗口内的完整上下文信息, 避免短序列对训练数据的低效利用. 为验证双向采样模型能否捕获更丰富的信息, 后续实验将对比双向采样与单向采样两种策略的预测性能差异.

此外, 在每一步游走中, 如果当前节点的邻居节点中不存在符合 MetaPath 下一跳预期类型的节点, 游走将中断. 针对这一问题, 我们设置了最大重试次数 max_retry, 若在限定重试次数内无法按照指定 MetaPath 顺利游走, 则回到该条 MetaPath 的初始节点, 重新随机选择 MetaPath, 以确保随机游走可以顺利进行, 保证了生成序列的完整性. 路径规模上, 我们为每个中药节点与症状节点执行 100 次长度为 100 的随机游走.

2.1.2 GO-DREAMwalk 节点序列生成

DREAMwalk 模型是通过构建多层网络以融合 PPI 网络与相似性网络的信息, 从而执行有偏随机游走的方法. 为了将 DREAMwalk 应用于中药-症状关联预测, 我们需要获取中药与症状的相似性信息. 受到 Ruiz 等人[8]工作的启发, 药物也可以通过靶向影响相同生物功能的远距离蛋白质来治疗疾病, 这表明药物对生物功能的调控可能是其治疗疾病的重要机制, 不同中药可能通过调控相同或相似的生物功能实现其治疗作用, 而症状的发生机制也可能与相同或相似生物功能的异常有关. 因此, 代表生物功能信息的 GO 术语可用于描述中药、症状的相似性, 本文采用 GO 术语间的语义相似性构建中药与症状的相似性网络. GO 术语间的语义相似性是一种基于 GO 术语层级图中两个术语与其祖先节点之间的结构关系, 衡量其在生物功能上相似程度的度量[17]. 以上详细介绍见第 3.1 节.

GO-DREAMwalk 生成节点序列的流程如图 1(b) 所示. 为了评估症状间的生物功能相似性, 我们采用 Lin Similarity 计算不同症状间 GO 术语的语义相似性, 相似性分数作为症状相似性网络中边的权重. Lin Similarity 是一种基于信息内容 (information content, *IC*) 的语义相似性算法, 用于衡量两个 GO 术语在层级结构中共享的信息量[18]. 其计算公式如下:

$$Sim_{\mathrm{Lin}}(c_1,c_2)=\frac{2\cdot IC(LCA(c_1,c_2))}{IC(c_1)+IC(c_2)} \tag{1}$$

其中, c_1, c_2 分别表示待比较的两个 GO 术语; $LCA(c_1,c_2)$ (lowest common ancestor, *LCA*) 表示两个术语在 GO 层级结构中最近的共同祖先; $IC(c)$表示术语 c 的信息量, GO 术语越具体, 值越高, 例如, 术语 GO:0015749 表示单糖跨膜或细胞内转运过程, 术语 GO:1904659 表示葡萄糖跨膜转运过程, 前者在功能层级上更广泛, 后者更具体, 故后者的 *IC* 值高于前者. 计算 *IC* 的表达式为:

$$IC(c)=-\log P(c) \tag{2}$$

其中, $P(c)$为术语 c 在所有 GO 数据中出现的概率.

Lin Similarity 反映了两个术语在生物功能上的相似程度. 当两种症状所关联的 GO 术语集中于相似的生物功能时, 它们之间的 Lin Similarity 会较高, 说明这两种症状可能具有相似的致病机制或调控路径[6].

由于中药的化学成分复杂、作用靶点众多, 其关联的 GO 术语数量庞大, 导致 Lin Similarity 在大规模中药样本中存在计算量大、运行时间长等问题. 为高效构建中药间的相似性网络, 本文将每种中药表示为 GO 术语的二进制向量, 并将 Jaccard Similarity 作为指标计算不同中药间的相似性, 在保留相似性信息的同时提升计算效率. 方法分为 3 个步骤.

(1) 数据预处理与 GO 术语集合扩展

1) 提取每种中药其靶标蛋白关联的全部 GO 术语;

2) 基于 GO 术语的层级结构, 追溯每个 GO 术语的所有祖先术语;

3) 结合原始术语与其祖先术语, 构建每个中药对应的扩展 GO 术语集合.

(2) 中药的 GO 向量表示

1) 汇总所有样本中出现的 GO 术语, 形成完整的 GO 术语列表;

2) 以 GO 术语为特征, 构建 N 维 (N 为 GO 术语总数) 二进制向量. 对于每种中药, 若其扩展 GO 集合中包含某一术语, 则对应位置赋值为 1, 否则为 0.

(3) 中药间相似性计算

基于中药的 GO 向量表示, 采用 Jaccard Similarity 计算中药间的相似性分数, 如式 (3) 所示:

$$Jaccard(Herb_i, Herb_j) = \frac{|Herb_i \cap Herb_j|}{|Herb_i \cup Herb_j|} \tag{3}$$

其中, $Herb_i$和$Herb_j$分别表示两种不同中药的 GO 向量表示.

将构建的相似性网络与 G_{hps} 网络整合, 构建用于 DREAMwalk 的多层网络结构. 为降低噪声并提高计算效率, 我们仅保留相似性得分排名前 60% 的边, 过滤掉低关联的连接. 在该多层网络上执行随机游走时, 步骤如下: 预设跳转概率 λ, 表示相似性信息对随机游走过程影响的强弱, λ 越大, 游走器跳转到相似性网络的动作越频繁 (本文采用 Bang 等人[16]工作中验证使用的 0.3); 当游走器遍历至中药或症状节点时, 会根据跳转概率进行选择, 以 λ 的概率将跳转到相应的相似性网络, 并根据当前节点与其相邻节点之间边的相似性权重执行遍历操作; 如果以 $1-\lambda$ 的概率选择留在 G_{hps} 网络或者遍历的节点是蛋白质节点, 游走器将沿其邻居节点继续进行随机游走. 对于每个节点, 我们执行 100 次随机游走, 每次游走长度为 10. 通过上述机制, 生成的每条游走路径能够动态融合网络拓扑信息与功能信息, 输出长度为 10 的节点序列, 作为后续嵌入学习的输入.

2.2 基于异构 Skip-gram 的节点嵌入向量生成

本文采用 Skip-gram 模型对 M-RW 和 GO-DREAMwalk 两种模型生成的节点序列进行嵌入学习, 以获取中药和症状节点的低维向量表示. Skip-gram 模型的结构如图 1(c) 所示. 训练过程中, 我们将嵌入维度设定为经验上效果较优的 128 维, 上下文窗口大小设为 4, 并以中心节点与其上下文节点构造正样本. 为增强模型对异构网络结构的建模能力, 我们采用 Bang 等人[16]提出的异构 Skip-gram 改进方法, 通过类型感知负采样策略优化训练过程. 与传统 Skip-gram 从所有节点中随机采样负样本不同, 该方法仅从与中心节点相同类型的节点中采样负样本, 从而避免了不同节点类型混合带来的干扰. 本文将负采样比例设置为 5, 即每个正样本随机采样 5 个同类型节点构成负样本集合.

Skip-gram 模型训练的目标是通过最大化中心节点与其正样本邻居之间的相似性, 并最小化其与负样本节点之间的相似性[19]. 目标函数如下:

$$L = -\left(\sum_{(u,v)\in D^+} \log\sigma(z_u^{\top}\cdot z_v) + \sum_{(u,v')\in D^-} \log\sigma(-z_u^{\top}\cdot z_{v'})\right) \tag{4}$$

其中, D^+ 表示正样本对集合: 先在网络上进行若干次随机游走生成节点序列, 随后以上下文窗口大小w (本文取$w=4$) 在序列上滑动, 对每个中心节点u与其窗口内出现的上下文节点v组成一对, 记作$(u,v)\in D^+$; 对应地, D^- 表示负样本对集合: 对每个正样本对(u,v)执行k次负采样 (本文取$k=5$), 即从与节点u不构成上下文关系的所有节点中随机抽取k个节点$v'_1, v'_2, \cdots, v'_k$, 由此构成的$(u,v')\in D^-$为负样本对, 其并集记作D^-; z_u和z_v为节点u和节点v的嵌入向量, 是通过 Skip-gram 模型训练得到的参数, 用于捕捉节点在低维向量空间中的结构特征与语义关系; $z_u^{\top}\cdot z_v$表示向量u和向量v在嵌入空间的相似度; $\sigma(\cdot)$为 Sigmoid 激活函数, 将节点间的相似度映射到 (0, 1) 区间, 表示两节点关联的概率; $\sum_{(u,v)\in D^+}\log\sigma(z_u^{\top}\cdot z_v)$表示最大化中心节点与正样本之间的相似度, $\sum_{(u,v')\in D^-}\log\sigma(-z_u^{\top}\cdot z_{v'})$表示最小化中心节点与负样本之间的相似度.

2.3 基于 XGBoost 的中药-症状关联预测

为预测中药与症状之间的关联, 本文将中药与症状嵌入向量的差值作为药症关系向量作为模型输入, 通过训练 XGBoost 二分类模型实现药症关联预测任务. XGBoost 模型图如图 1(d) 所示. 设中药节点h_i与症状节点s_j的嵌入向量分别为e_{h_i}和e_{s_j}, 我们采用元素相减的方式构造药症对的关系特征向量:

$$x_{ij} = e_{h_i} - e_{s_j} \tag{5}$$

其中, x_{ij}表示中药节点h_i与症状节点s_j的关系特征向量.

XGBoost 模型的训练目标是通过监督学习框架, 最小化预测值与真实标签之间的损失函数, 从而最大化模型对潜在药症关联的识别能力, 提升对未知药症

组合预测的准确性.

具体而言, XGBoost 的训练过程基于前一弱学习器的残差对新的学习器进行拟合. 每一轮迭代中, 模型最小化目标函数, 该函数综合了当前模型在训练集上的损失函数值以及模型复杂度正则化项, 旨在防止过拟合[20]. 设训练集中共有 n 组药症对, 第i个样本的真实标签为y_i, 预测值为$\hat{y}_i^{(t)}$. XGBoost 通过迭代地叠加回归树来更新预测结果, 形式为:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{6}$$

其中, f_t表示第t棵回归树, x_i为药症关系特征向量. 整体的目标函数形式如下:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{7}$$

其中, $l(y_i, \hat{y}_i^{(t)})$为损失函数, 表示第t轮迭代中样本i的预测值$\hat{y}_i^{(t)}$与真实值y_i之间的损失, 用于度量预测结果与真实标签间的差异 (本任务为二分类, 采用对数损失函数); $\Omega(f_k)$是第k棵树的正则化项, 用于控制模型复杂度, 具体形式为:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{8}$$

其中, T表示第k棵树中的叶子节点数量, ω_j表示第j个叶子节点的权重, γ和λ是用于控制模型复杂度的正则化系数. 该正则化项通过惩罚叶节点数量和节点权重的大小, 用来限制树的结构复杂度, 从而达到降低过拟合风险并提升模型的泛化能力的目的.

3 实验

3.1 数据集

本文使用的数据来源于多个研究成果与公开数据库. 首先, PPI 数据集来自 Barabási 实验室, 包含 18 505 个蛋白质节点以及蛋白质之间的 327 924 条相互作用构成的边[21]. 其次, 中药-症状、临床肝硬化数据均来自 Gan 等人[6]的研究. 其中, 中药-症状数据包含 798 种中药和 174 种症状, 涉及 127 759 条中药-靶标关联、12 601 条症状-蛋白关联、1 436 条中药-症状关联; 临床肝硬化数据包含 384 种中药、160 种临床症状, 涉及 79 334 条中药-靶标关联、12 446 条临床症状-蛋白关联, 中药-临床症状治疗关联有两种: 临床出现、临床有效. 临床出现定义为: 若某一中药处方被开具给患者, 则该处方中包含的所有中药视为与该患者记录的所有症状相关联[6]. 根据此定义, 经过数据处理, 共获得 4 869 组中药-临床症状关联; 临床有效定义为: 若患者使用某中药后恢复情况优于未使用该中药的患者, 则判定该中药对相关症状具有正向疗效[6]. 基于该标准, 最终获得 760 组具有疗效支持的中药-临床症状关联.

在网络构建方面, 本文将中药、症状、蛋白质作为节点, 中药与蛋白质关联、症状与蛋白质关联、蛋白质间的相互作用作为边, 构建了中药-蛋白质-症状网络, 我们用 $G_{hps}=\{V_h, V_p, V_s, E_{hp}, E_{pp}, E_{sp}\}$表示该网络的拓扑结构, 其中, V_h、V_p、V_s 分别表示中药节点、蛋白质节点以及症状节点, E_{hp}、E_{pp}、E_{sp} 则表示网络中 3 种类型的边 (中药-蛋白质、蛋白质-蛋白质、症状-蛋白质), 所有边的权重均统一设为 1, 共包含 19 578 个节点和 480 674 条边.

此外, 为进一步引入生物功能信息, 本研究使用 Python 包 pygosemsim (https://github.com/mojaie/pygosemsim) 自动下载 Gene Ontology 数据库中的蛋白质-GO 术语文件及 GO 术语间的层级关系文件. 该数据覆盖了 3 大类 GO 注释类型: 生物过程 (BP)、分子功能 (MF) 和细胞组分 (CC). GO 术语作为节点, 术语间的关系“is_a”“part_of”等作为边构成 GO 层级结构图, 该层级结构用于定义术语间的继承与组成等关系. 本文模型代码及数据集可见 https://github.com/lzh228/TCM_Repurposing.git.

3.2 实验设置

本文将 798 种中药与 174 种症状构成的 138 852 组药症对作为基础数据集, 按照 8:1:1 划分训练集、验证集、测试集, 药物与疾病存在治疗关联则标记为正例, 其中测试集包含 13 886 组药症对, 正例有 144 组. 在节点嵌入向量生成方面, 本文将 Skip-gram 模型的上下文窗口大小设置为 4, 负采样比例设置为 5. 在预测药症关联方面, 为优化超参数组合, 本文采用了网格搜索的策略. 待优化的超参数包括: 树的最大深度 (候选值为 {5, 6, 7, 8, 9})、学习率 (候选值为 {0.01, 0.05, 0.07, 0.08, 0.1}) 以及节点分裂所需的最小损失减少量 (候选值为 {0, 0.1, 0.2}). 训练样本的随机采样比例和特征的随机采样比例均固定为 0.8. 网格搜索共遍历了 75 组参数组合, 并采用 5 折交叉验证. 为应对标签分布不均衡的问题, 在模型训练过程中采用了正负样本的权重调整策略, 设置 scale_pos_weight 为负样本数与正样本

数的比值. XGBoost 模型的最大迭代轮数设为 500, 并启用 early stopping 机制以防止过拟合. 实验采用 Python 3.7 作为开发环境, 运行在 Intel Core i5-13500 处理器上, GPU 选择 NVIDIA GeForce RTX 3070, 显存大小为 8 GB.

3.3 评价指标

中药重定位的核心目标是从大规模潜在中药-症状对中筛选出有治疗潜力的药物, 因此模型在高置信度的表现尤为重要. 为此, 本文引入 Precision@Top1% 和 Recall@Top1% 两个指标. Precision@Top1% 衡量模型在预测得分排名前 1% 的中药-症状对中, 正确预测为正例的比例, 反映高分结果中候选药物的准确性; 而 Recall@Top1% 评估模型在排名前 1% 的中药-症状对成功召回真实正例的能力, 补充了 Precision@Top1% 指标可能忽略的召回性能. 为统一不同规模数据集下的评估标准, 并贴合实际应用中优先筛选潜力药物的需求, 本文采用预测得分排名的 Top1% 作为高置信度评估范围. 例如, 在包含 61 440 个样本的临床数据集中, Top1% 对应 614 组预测得分最高的中药-症状对, Precision@Top1% 和 Recall@Top1% 则以 Top1% 作为阈值计算模型的预测性能. 此外, 我们选用受试者工作特征曲线下的面积 (AUROC) 作为模型全局预测性能评估指标. AUROC 能够反映模型在所有分类阈值下区分正负样本的能力.

通过综合使用 Precision@Top1%、Recall@Top1% 和 AUROC 这 3 种评估指标, 我们在兼顾模型整体区分能力的同时, 重点考察其在高置信度区域的预测结果, 从而更加全面评估模型的预测性能.

3.4 实验结果与分析

3.4.1 M-RW 模型

为筛选具有丰富路径信息的 MetaPath 集合, 本文将候选的 3 组路径集合 M_3、M_4、M_5 应用于下游预测任务中进行性能对比. 具体而言, 分别利用每组 MetaPath 集合指导随机游走生成异构节点序列, 并输入异构 Skip-gram 模型, 学习每个中药节点与症状节点的 128 维嵌入向量, 随后, 使用这些嵌入作为特征训练 XGBoost 分类器, 并在测试集预测任务、临床有效任务以及临床出现任务上评估模型的预测性能.

实验结果如图 3(a) 所示, 在测试集任务中, M_4 的 AUROC 略低于 M_3 和 M_5, 但其在 Precision@Top1% 和 Recall@Top1% 两项高置信度指标上均优于其他两组. 在临床有效和临床出现任务中, M_4 在 AUROC、Precision@Top1% 和 Recall@Top1% 这 3 项指标上均表现最优. 综合来看, M_4 在所有预测任务中展现出最优的性能. 我们认为, 较短路径集合 M_3 可能不足以捕获复杂的路径信息, 而过长路径集合 M_5 则可能引入过多噪声和冗余信息, 影响嵌入学习效果.

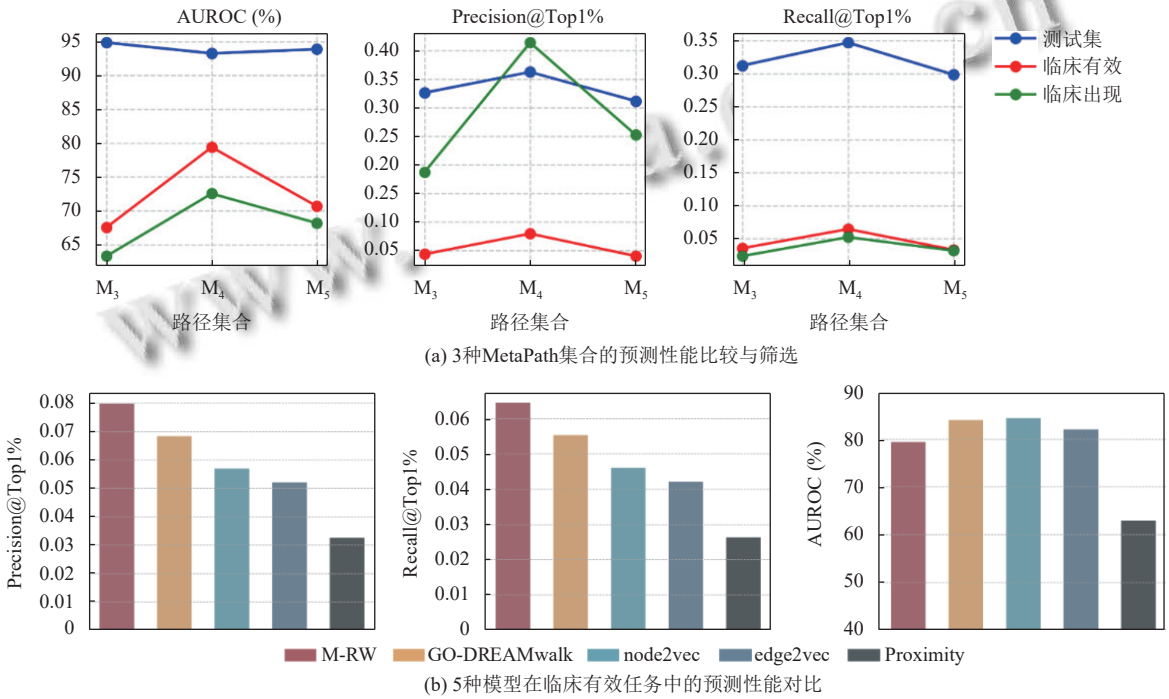
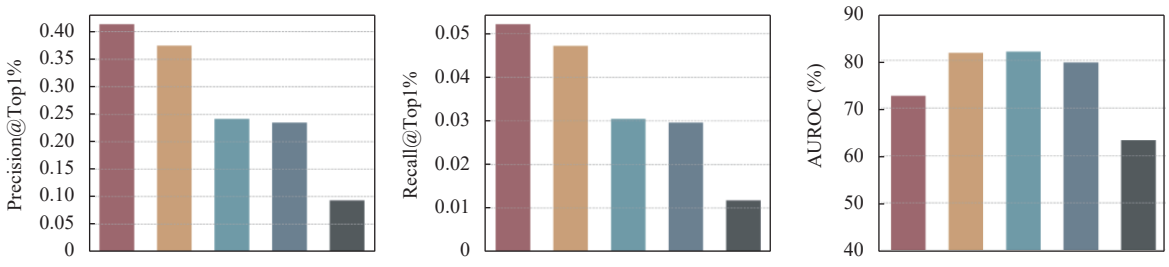
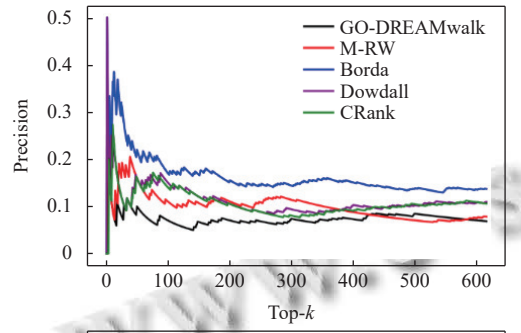
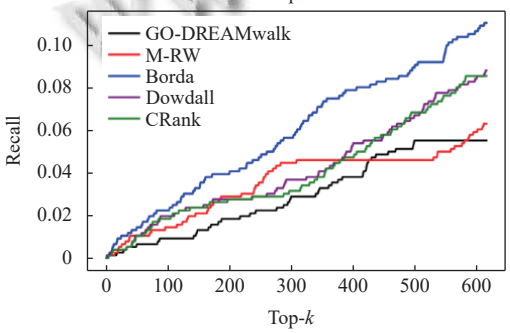


图 3 M-RW 与 GO-DREAMwalk 模型的性能评估及结果融合分析

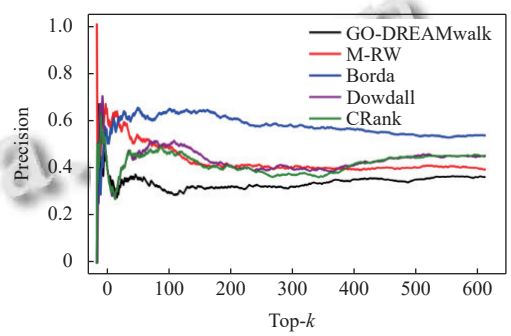


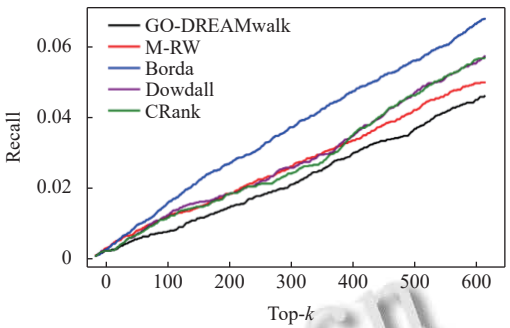
(c) 5种模型在临床出现任务中的预测性能对比





(d) M-RW与GO-DREAMwalk在不同聚合策略下的临床有效性能比较





(e) M-RW与GO-DREAMwalk在不同聚合策略下的临床出现性能比较

图 3 M-RW 与 GO-DREAMwalk 模型的性能评估及结果融合分析 (续)

上述结果已经表明, M_4 集合的 MetaPath 能够更有效地捕获 G_{hps} 中潜在的路径信息. 为进一步验证本文设计的双向采样策略能否捕获更多机制信息, 本文将其与单向采样方法进行了性能对比. 与双向采样不同, 单向采样采用动态长度的随机游走策略, 游走起点为中药或症状节点, 游走终点为首次到达的非蛋白质节点. 若终点与起点是相同类型的节点时, 系统将判断该路径是否匹配预设的 MetaPath, 若匹配, 则保留该序列, 否则重新采样. 若游走终点与起点为同类节点时, 该路径视为无效并重新开始游走. 通过该策略, 为每个起始节点生成了一系列长度不等、结构符合 MetaPath 的有效路径序列, 用于嵌入学习和关联预测.

基于上述两种采样策略, 我们进行了由 MetaPath 指导的随机游走实验. 我们统一将 M_4 路径集合作为随机游走的指导规则, 在采样规模方面, 两种采样策略均对每个中药或症状节点执行 100 次随机游走, 并在肝硬化临床数据集上评估模型性能, 实验结果如表 2 所示, 双向采样在 AUROC、Precision@Top1% 和 Recall@Top1% 这 3 项指标上均表现最佳, 说明该策略更有助于捕捉网络中丰富的机制关联, 从而提升模型的预测能力.

表 2 双向采样与单向采样预测性能比较

采样方法	预测任务	AUROC (%)	Precision@Top1%	Recall@Top1%
双向采样	测试集	93.27	0.3623	0.3472
	临床有效	79.4	0.0798	0.0645
	临床出现	72.54	0.4137	0.0522
单向采样	测试集	90.3	0.2246	0.2153
	临床有效	72.95	0.0505	0.0408
	临床出现	72.04	0.2606	0.0329

3.4.2 GO-DREAMwalk 模型

为了将 DREAMwalk 模型应用于中药重定位, 本文基于 GO 术语间的语义相似性, 构建了中药与症状的相似性网络. 其中, 症状相似性网络包含 334 种症状节点及其 33 144 条相似性关系构成的边, 中药相似性网络包含 798 种中药及其相似性关系构成的 190 802 条边, 相似性分数作为边的权重. 我们将上述两个相似性网络与 G_{hps} 整合为一个多层异构网络, 并在该网络上应用 DREAMwalk 模型进行嵌入学习与关联预测.

实验结果显示, GO-DREAMwalk 模型在测试集任务中 AUROC 值为 93.73%, Precision@Top1% 值为 0.376 8, Recall@Top1% 值为 0.361 1; 在临床有效任务中, AUROC 值为 84.03%, Precision@Top1% 值为 0.068 4, Recall@Top1% 值为 0.055 3; 在临床出现任务中, AUROC 值为 81.58%, Precision@Top1% 值为 0.374 6, Recall@Top1% 值为 0.047 2. 这一实验结果说明引入功能相似性网络的 GO-DREAMwalk 模型在高排名预测和整体预测上均有较好的性能表现, 可以有效地应用于中药-症状关联预测任务.

3.4.3 对比实验

为了验证 M-RW 和 GO-DREAMwalk 模型在筛选有治疗潜力中药上的优势, 本文在测试集上评估了两者与多种基线模型的性能差异, 并通过临床数据集上“临床有效”和“临床出现”两个任务测试模型的泛化能力. 基线方法包括: 1) 基于机制驱动的 Proximity 方法; 2) 两种基于数据驱动的随机游走模型: node2vec 和 edge2vec.

这 5 种方法在测试集上的预测性能表现如表 3 所示, 我们的模型相较于非机器学习方法 Proximity 有较大的性能提升, 而相较于 node2vec 和 edge2vec, 整体预测性能 (AUROC) 基本持平, 头部预测性能 (Precision@Top1% 和 Recall@Top1%) 得到小幅提升.

表 3 5 种模型在测试集任务中的性能表现

性能指标	Proximity	node2vec	edge2vec	GO-DREAMwalk	M-RW
AUROC (%)	70.27	93.95	93.37	93.73	93.27
Precision@Top1%	0.043 5	0.351 1	0.333 3	0.376 8	0.362 3
Recall@Top1%	0.041 7	0.336 7	0.319 4	0.361 1	0.347 2

为进一步评估模型的泛化能力, 我们基于肝硬化相关的临床数据集对上述 5 种模型进行了测试, 性能表现如图 3(b)、图 3(c) 所示. 相较于 Proximity 方法, 在临床有效任务中, M-RW 的 AUROC 提升了 26.2%, Precision@Top1% 和 Recall@Top1% 提升约 145%; GO-DREAMwalk 在 AUROC 上提升了 33.6%, Precision@Top1% 和 Recall@Top1% 提升约 110%. 在临床出现任务中, M-RW 的 AUROC 提升了 14.8%, Precision@Top1% 和 Recall@Top1% 提升约 346%; GO-DREAMwalk 的 AUROC 提升 29.1%, Precision@Top1% 和 Recall@Top1% 提升约 304%. 相较于随机游走模型, GO-DREAMwalk 在两个任务上的 AUROC 差异较小, 与 M-RW 相比两种基线方法预测性能出现了下降, 这表明我们的模型在全局预测能力方面稍弱, 但对于高置信度指标 Precision@Top1% 与 Recall@Top1%, 两种模型在临床任务中取得了较大优势. 预测性能结果如下: 在临床有效任务中, M-RW 相较于 node2vec 在 Precision@Top1% 和 Recall@Top1% 上均提升约 40%, 相较于 edge2vec 提升约 53%; GO-DREAMwalk 相较于 node2vec 提升约 20%, 相较于 edge2vec 提升约 31%. 在临床出现任务中, M-RW 相较于 node2vec 提升约 72%, 相较于 edge2vec 提升约 76%; GO-DREAMwalk 相较于 node2vec 提升约 55%, 相较于 edge2vec 提升约 60%.

上述实验结果表明, M-RW 和 GO-DREAMwalk 模型均可以增强模型对中药-症状关联的预测能力, 尤其在临床任务高排名预测中表现出更强的优越性. 考虑到这两种模型分别引入了路径信息和功能信息, 我们推测其可能在捕获网络拓扑特征上具有互补性, 有望进一步提升预测性能. 为此, 我们采用了 CRank、Borda 和 Dowdall 这 3 种 Rank Aggregation 算法[21], 对两种模型在临床任务中得到的预测结果集成排序, 并以 Precision 和 Recall 为评估指标, 绘制了 Top-k 药症对预测性能变化曲线, 如图 3(d)、图 3(e) 所示. 实验结果显示, 两种临床任务中, Borda 方法聚合排名后的综合性能最优, Top-k 的综合预测性能明显优于 CRank、Dowdall 方法和单模型的预测结果. 在 Top1% 阈值处, Borda 在临床有效任务中的 Precision 值达到 0.14, Recall 值达到 0.11, 在临床出现任务中, Precision 值达到 0.54, Recall 值达到 0.067, 预测性能实现了大幅提升. 通过上述实验, 证实了 M-RW 和 GO-DREAMwalk 模型存在互补优势, 通过聚合两种模型的预测结果, 可以有效提升中药重定位的预测性能.

3.4.4 消融实验

为了进一步验证 M-RW 和 GO-DREAMwalk 模型

路径信息和功能信息的有效性, 我们分别对 M-RW 模型和 GO-DREAMwalk 模型进行了消融实验.

对于 M-RW 模型, 比较引入与未引入 MetaPath 对模型预测性能的影响. 具体而言, 我们在网络中随机选取与中药和症状节点数相等的节点集合作为游走起点, 执行 100 次长度为 100 的随机游走, 并重复该过程 10 次以获得稳健的嵌入向量, 用于下游中药-症状关联预测任务. 预测性能在临床数据集上进行评估, 并与采用 MetaPath 指导的随机游走策略进行性能对比.

实验结果如表 4 所示, 无 MetaPath 指导的随机游走的 AUROC 略优于 MetaPath 引导策略, 说明其在捕捉全局网络拓扑信息方面具备一定优势, 但在药物重定位更为关注的 Top1% 高排名预测结果中, MetaPath 引导策略在性能上优于对照组. 在临床有效任务中, M-RW 模型的 Precision@Top1% 达到 0.079 8, 相对于对照组的 0.054 1 提升了 47.5%; 在临床出现任务中的 Precision@Top1% 更是达到 0.413 7, 相较于对照组的 0.260 3 提升了 58.9%. 我们认为, 路径约束的模型可以更好地捕获中药-症状关联信息, 提高中药重定位的准确性. 相比之下, 未引入 MetaPath 的模型虽然覆盖了更广泛的网络区域, 整体预测表现较好, 但由于缺乏路径信息的约束, 导致其高排名预测能力受限, 难以精确识别出真正具有治疗潜力的中药-症状关联.

表 4 引入与未引入 MetaPath 的模型性能比较

组别	预测任务	AUROC (%)	Precision@Top1%	Recall@Top1%
对照组	临床有效	83.21	0.054 1	0.043 7
	临床出现	80.02	0.260 3	0.032 8
M-RW	临床有效	79.40	0.079 8	0.064 5
	临床出现	72.54	0.413 7	0.052 2

对于 GO-DREAMwalk 模型, 为了评估引入相似性网络是否对中药-症状关联预测任务有积极影响, 我们比较了该模型在引入与未引入功能相似性网络两种设置下的性能差异. 实验组与对照组的预测结果如表 5 所示. 结果显示, 功能信息引导的随机游走有效提升了模型在药症关联预测中的表现, 在 AUROC、Precision@Top1% 和 Recall@Top1% 指标上均取得明显优势: 在临床有效任务中, 3 个指标分别提升了 2.4%、31.3%、31.4%, 在临床出现任务中, 3 个指标分别提升了 2.5%、59.7%、59.5%. 这表明, 功能信息在捕捉中药-症状关联信息方面具有重要作用, 同时也验证了其在真实临床预测任务中的实用价值.

表 5 引入与未引入相似性网络的模型性能比较

组别	预测任务	AUROC (%)	Precision @Top1%	Recall @Top1%
对照组	临床有效	82.04	0.052 1	0.042 1
	临床出现	79.59	0.234 5	0.029 6
GO-DREAMwalk	临床有效	84.03	0.068 4	0.055 3
	临床出现	81.58	0.374 6	0.047 2

4 结论与展望

本文提出的 M-RW 和 GO-DREAMwalk 模型, 在中药重定位任务中相较于传统的机制驱动模型和纯数据驱动的随机游走模型展现出更优的预测性能. 通过消融实验进一步验证了路径信息与功能信息的引入对于提升模型效果具有积极作用. 此外, 本文采用 Rank Aggregation 方法融合两种模型的预测结果, 融合策略进一步提升了中药重定位的预测准确性, 表明两种模型具有良好的互补性.

上述模型均在肝硬化临床数据上进行测试与评估. 表 6 展示 50 组经临床验证确认为治疗有效的中药-症状对. 其中部分关联已被《中国药典》记载, 进一步佐证了本模型预测的可靠性. 值得注意的是, 一些未被《中国药典》记载的中药-症状对, 在实际临床治疗中同样展现出良好疗效, 且成功被本模型预测出来, 显示出本模型在新药发现和药物重定位中的应用潜力. 此外, 其他高排名但尚未被《中国药典》记载的中药-症状关系, 也可作为治疗肝硬化相关症状的潜在候选药物, 通过后续的生物学实验或临床试验进一步验证. 由于本模型是围绕中药与疾病关联构建的, 方法具有良好的通用性, 未来亦可应用于其他症状或疾病的临床预测任务中.

由于中药“多成分、多靶点”的特性, 其治疗机制复杂多样, 目前难以被完全解析. 本文方法能够利用路径信息与功能信息, 有效提升预测性能, 但其局限在于无法揭示药物与症状间的作用机制. 因此, 这两种信息在预测中所发挥的具体作用, 以及二者贡献的差异, 仍有待研究. 为此, 我们计划在后续研究中引入解释性子图分析方法. 该方法通过提取与预测结果高度相关的局部子图结构, 识别其中关键节点与关键路径, 从而揭示模型预测性能提升的原因. 借助这一方法, 我们期望进一步揭示中药治疗症状的机制, 并分析出两种信息在关联预测任务中的作用及贡献差异.

本文提出的融入路径信息和功能信息的随机游走

中药重定位预测模型, 在真实临床数据上的预测结果均具备良好的预测性能, 充分展现了其在中药重定位中的应用潜力, 有望为推动中药在新适应症中的有效应用提供有力的技术支持.

表 6 预测列表头部 50 组在临床数据中显示为治疗有效的中药-症状对

中药	肝硬化症状	《中国药典》是否记载	Borda排序	中药	肝硬化症状	《中国药典》是否记载	Borda排序
黄芪	浮肿	√	2	桑白皮	浮肿	√	152
黄芩	咳嗽	√	6	山豆根	乏力	×	153
陈皮	浮肿	×	10	苍术	浮肿	√	159
川楝子	浮肿	×	11	柴胡	发热	√	161
半枝莲	浮肿	√	13	连翘	浮肿	×	172
葛根	浮肿	×	18	泽兰	腹痛	√	200
莱菔子	咳嗽	√	19	麻黄	咳嗽	√	216
车前草	浮肿	√	25	人参	腹痛	√	227
川芎	头晕	×	36	玄参	便秘	√	228
红花	浮肿	√	42	枳实	咳嗽	×	237
大腹皮	浮肿	√	50	桂枝	浮肿	√	247
枳实	腹痛	×	58	杏仁	浮肿	×	251
防风	浮肿	×	60	虎杖	浮肿	×	260
枳实	身黄	×	68	栀子	咳嗽	×	271
枳实	浮肿	×	70	西洋参	浮肿	×	274
葛根	头晕	×	78	杜仲	浮肿	×	277
杏仁	咳嗽	√	82	党参	浮肿	×	282
益智仁	浮肿	×	103	佛手	浮肿	×	287
川芎	腹痛	√	108	玄参	咳嗽	√	303
白芥子	咳嗽	√	115	地骨皮	咳嗽	√	306
北沙参	浮肿	×	121	木香	浮肿	×	309
藿香	咳嗽	√	122	金钱草	浮肿	√	314
半夏	咳嗽	√	126	附子	咳嗽	×	316
决明子	浮肿	×	141	半枝莲	咳嗽	×	319
茜草	浮肿	×	144	川芎	咳嗽	×	327

参考文献

1 Zhao L, Zhang H, Li N, *et al*. Network pharmacology, a promising approach to reveal the pharmacology mechanism of Chinese medicine formula. Journal of Ethnopharmacology, 2023, 309: 116306. [doi: 10.1016/j.jep.2023.116306]

2 Hua Y, Dai XW, Xu Y, *et al*. Drug repositioning: Progress and challenges in drug discovery for various diseases. European Journal of Medicinal Chemistry, 2022, 234: 114239. [doi: 10.1016/j.ejmech.2022.114239]

3 Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA, 2020, 323(9): 844–853. [doi: 10.1001/jama.2020.1166]

4 Perdomo-Quinteiro P, Belmonte-Hernández A. Knowledge Graphs for drug repurposing: A review of databases and methods. Briefings in Bioinformatics, 2024, 25(6): bbae461. [doi: 10.1093/bib/bbae461]

5 Xue HQ, Li J, Xie HZ, *et al*. Review of drug repositioning approaches and resources. International Journal of Biological Sciences, 2018, 14(10): 1232–1244. [doi: 10.7150/ijbs.24612]

6 Gan X, Shu ZX, Wang XY, *et al*. Network medicine framework reveals generic herb-symptom effectiveness of traditional Chinese medicine. Science Advances, 2023, 9(43): eadh0215. [doi: 10.1126/sciadv.adh0215]

7 Sun XL, Jia X, Lu ZL, *et al*. Drug repositioning with adaptive graph convolutional networks. Bioinformatics, 2024, 40(1): btad748. [doi: 10.1093/bioinformatics/btad748]

8 Ruiz C, Zitnik M, Leskovec J. Identification of disease treatment mechanisms through the multiscale interactome. Nature Communications, 2021, 12(1): 1796. [doi: 10.1038/s41467-021-21770-8]

9 Cheng FX, Desai RJ, Handy DE, *et al*. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nature Communications, 2018, 9(1): 2691. [doi: 10.1038/s41467-018-05116-5]

10 Yi HC, You ZH, Huang DS, *et al*. Graph representation learning in bioinformatics: Trends, methods and applications.

Briefings in Bioinformatics, 2022, 23(1): bbab340. [doi: 10.1093/bib/bbab340]

11 Guney E, Menche J, Vidal M, *et al*. Network-based *in silico* drug efficacy screening. Nature Communications, 2016, 7: 10331. [doi: 10.1038/ncomms10331]

12 Köhler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. American Journal of Human Genetics, 2008, 82(4): 949–958. [doi: 10.1016/j.ajhg.2008.02.013]

13 Grover A, Leskovec J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 855–864.

14 Dong YX, Chawla NV, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017. 135–144.

15 Gao Z, Fu G, Ouyang CP, *et al*. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. BMC Bioinformatics, 2019, 20(1): 306. [doi: 10.1186/s12859-019-2914-2]

16 Bang DM, Lim S, Lee S, *et al*. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. Nature Communications, 2023, 14(1): 3570. [doi: 10.1038/s41467-023-39301-y]

17 Wang JZ, Du ZD, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, 23(10): 1274–1281. [doi: 10.1093/bioinformatics/btm087]

18 Lin DK. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning. Madison: Morgan Kaufmann, 1998. 296–304.

19 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale: OpenReview.net, 2013. 3781.

20 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 785–794.

21 Morselli Gysi D, Do Valle Í, Zitnik M, *et al*. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proceedings of National Academy of Sciences of the United States of America, 2021, 118(19): e2025581118.

(校对责编: 李慧鑫)