基于二分网络表示学习的医学实体关系预测研究*

吴胜男[1] 吴佳辉[1] 董继宗[1] 蒋环宇[1] 王璐琦[1] 王欣瑶[2]

([1] 山西医科大学管理学院 太原 030012 [2] 山西医科大学基础医学院 太原 030012)

〔摘要〕 **目的/意义** 探讨网络表示学习与链路预测在挖掘潜在医学实体关系方面的应用，为医学知识发现研究提供新视角。**方法/过程** 从 PubMed 数据库获取文献摘要，利用“主语－行为－宾语”语义挖掘方案识别疾病与药物治疗信息，抽取药物实体与疾病实体，构建“药物－疾病”二分网络，综合运用社会网络分析、网络表示学习、机器学习方法分析网络结构及节点特征，挖掘医学实体间的潜在联系。**结果/结论** 二分网络表示学习方法能够揭示药物与疾病的关联知识，获取治疗疾病的关键药物，可为用药治疗提供可行性方案。

〔关键词〕 二分网络；“主语－行为－宾语”语义挖掘；网络表示学习；机器学习；医学知识发现

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2025.02.011

Study on Predicting Medical Entity Relationships Based on Bipartite Network Representation Learning

WU Shengnan[1], *WU Jiahui*[1], *DONG Jizong*[1], *JIANG Huanyu*[1], *WANG Luqi*[1], *WANG Xinyao*[2]

[1] *School of Management, Shanxi Medical University, Taiyuan 030012, China*; [2] *School of Basic Medicine, Shanxi Medical University, Taiyuan 030012, China*

〔**Abstract**〕 **Purpose/Significance** To discuss the application of network representation learning and link prediction in mining potential medical entity relationships, and to provide a novel perspective for medical knowledge discovery. **Method/Process** The literature abstracts are obtained from PubMed database, disease drug treatment information within these abstracts is identified using the subject－action－object (SAO) semantic mining scheme. Drug entities and disease entities are extracted, and a “drug－disease” bipartite network is constructed. The network structure and node characteristics are analyzed comprehensively using methods of social network analysis, network representation learning and machine learning, potential connections between medical entities are explored. **Result/Conclusion** The bipartite network representation learning method can reveal the association knowledge between drugs and diseases, identify key medications for treating diseases, and provide feasible treatment options.

〔**Keywords**〕 bipartite network; subject－action－object (SAO) semantic mining; network representation learning; machine learning; medical knowledge discovery

〔修回日期〕 2024－07－12

〔作者简介〕 吴胜男，博士，副教授，硕士生导师，发表论文 24 篇。

〔基金项目〕 国家自然科学基金青年项目（项目编号：71804102）；山西省高等学校哲学社会科学研究项目（项目编号：2019W040）；山西省研究生教育教学改革课题（项目编号：2021YJJG115）。

1 引言

近年来《关于进一步完善医疗卫生服务体系的意见》《关于推动疾病预防控制事业高质量发展的指导意见》等重要文件陆续发布，明确指出要进一步加强医学领域信息化支撑保障，提升药物创新能力。利用数字技术从医学数据中挖掘关键信息、利用关联信息预测以辅助临床决策更精确地诊疗对于打造智慧医疗体系、全面构建数字中国具有重要意义。

网络表示学习在生化医疗领域具有良好的应用前景，通过将网络节点信息转化为低维稠密的实数向量，在学习网络顶点潜在低维表示的同时保留网络拓扑结构、语义信息等内容[1]。网络表示学习在药物活性预测和化学物筛选方面能够对分子结构表示进行学习和性质预测，在蛋白质交互层面能够对蛋白质结构、功能以及相互作用建模和预测[2]。基于表示学习的分析方法可以容易且高效地执行网络分析任务，帮助揭示药物与疾病之间的关系，提供决策支持和指导。

因此，将网络表示学习与机器学习相结合应用于医学异构网络，能够有效进行知识发现研究，并在机器学习模型中度量多个特征的重要性，选取最相关、最具信息量的特征，有助于更全面地理解数据间的关系，提高模型准确性和泛化能力，产生更可靠有效的预测结果，为药物研发与用药指导提供一定的参考价值。

2 相关研究概述

医学知识发现依托网络分析、机器学习等数字技术，从生物信息、医学文献、电子病历等生物医学数据库中存储的大量非结构化、半结构化、结构化医学数据中抽取特定研究方向的实体、关系、属性等要素进行关联及预测分析，从而提高知识信息服务质量。

二分网络是特殊的复杂网络，包含两种类型的节点，边仅存在于不同类型的任意两个节点之间[3]。其研究方法可以归纳为将二分网络投影到单顶点网络进行分析和直接基于原始二分网络进行分析[4]。在医学知识发现领域，医学文献记载了科学严谨的实验结果，蕴含丰富的医学知识，基于其构建的二分网络可以表现为“药物 - 疾病”“药物 - 蛋白质”等，并用于预测分析，深入理解药物与疾病的作用方式，为医学研究和临床实践提供重要参考依据。

链路预测通过已知的网络结构等信息预测尚未产生连边的节点之间产生联系的可能性[5]，通常基于相似性、机器学习、矩阵、概率模型等方法分析。其中机器学习研究通常涉及二分类任务，正样本是网络中已存在的边，负样本是负采样得到的不存在的边，通过监督学习、无监督学习、强化学习的方式不断学习更新参数，减小损失函数值，得到训练好的模型，然后泛化到未见过的数据展开预测[6]，在医学实体识别及关系预测方面具有重要作用。

目前，网络表示学习模型有 DeepWalk、Node2Vec、LINE、SDNE 等，鉴于其能够兼顾网络结构信息和节点信息，提高机器学习效率，医学领域也开始关注基于网络表示学习的知识发现。Hu F 等[7]基于药物 - 属性异构网络，使用异构图卷积网络聚合相邻节点信息，嵌入药物及其属性，然后利用谱聚类算法划分嵌入信息，预测隐藏的医学关系。余黄樱子等[8]学习疾病网络内部和外部特征并将节点映射为空间向量，然后预测疾病知识间存在的关联关系。

综上，当前学术界已将网络表示学习与链路预测相结合，并应用于医学知识发现领域，然而针对以二分网络为核心的研究仍然缺乏足够的关注。基于此，本研究以二分网络为主体，以文献为载体，以药物疾病关系预测为例进行实证研究，为该领域的发展提供新思路。

3 研究方法

3.1 总体框架设计

医学实体关系预测研究的流程框架，见图 1，共包含 3 个步骤。一是获取数据及识别实体。从 PubMed 数据库获取文献摘要，利用“主语 - 行为

－宾语”（subject－action－object，SAO）语义挖掘方案识别药物实体与疾病实体。二是构建“药物－疾病”二分网络。包括利用复杂网络分析软件进行网络构建、可视化及网络结构分析。三是基于网络表示学习预测医学实体关系。利用网络表示学习算法获取节点向量，然后度量特征重要性选取特征并输入机器学习模型，实现“药物－疾病”二分网络的医学知识发现研究。

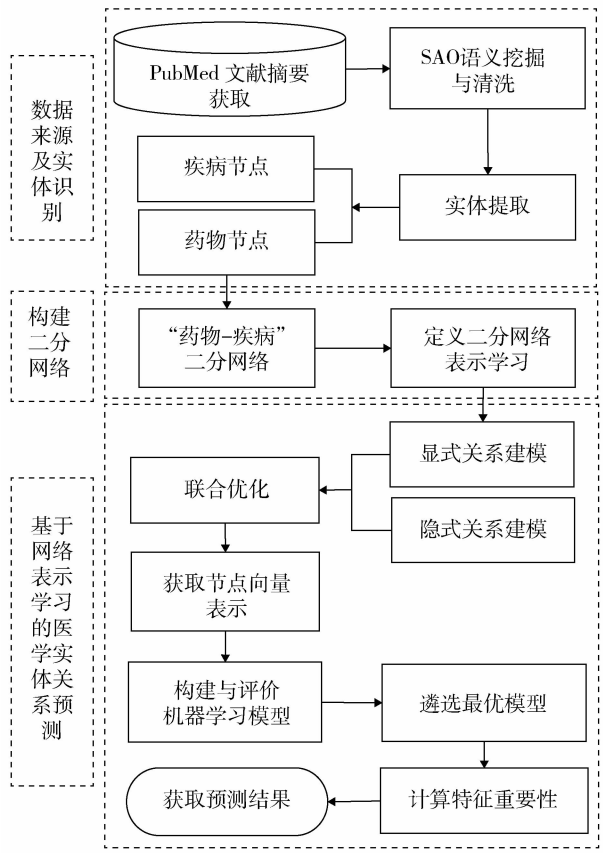

图 1 医学实体关系预测流程

3.2 数据来源及实体识别

3.2.1 数据获取 医学文献包含医学领域的科学研究、临床实践、疾病诊断和治疗，过程严谨，结论有效，分析其内容可以获取最新医学知识、了解发展前沿。文献摘要提供文献内容的关键信息，因此基于 PubMed 医学文献数据库挖掘摘要中包含的药物与疾病信息。临床上消化系统疾病较为多见，其中胃部疾病发病率高且具有多种类型。因此以胃部疾病为研究对象，以（（（（（gastric）AND（diseases））OR（gastricdiseases）OR（gastricdiseases［MeSH Terms］））AND（（medication）OR（drug therapy））））AND（humans［MeSH Terms］）AND（English［Language］）AND（“1973/01/01”［Date－Publication］：“2022/12/31”［Date－Publication］））为检索式，得到 74 030 篇文献并获取摘要。

3.2.2 SAO 语义挖掘 SAO 语义挖掘指从文本中提取“主语－行为－宾语”结构，进行深入的语义分析和应用。选取 SemRep 获取数据，该工具是语义关系抽取系统，抽取出的 SAO 结构可用于分析和处理生物医学领域文献，且抽取出的数据具有二分网络数据结构属性。SemRep 获取的数据形式为 tmp_data10001_tmp_10. txt. tx. 455 | relation | C0026256 | Mitotane | orch, phsu | phsu | | | TREATS | C0001622 | Adrenal Gland Hyperfunction | dsyn | dsyn | |。其中，“Mitotane”为药物实体，“Adrenal Gland Hyperfunction”为疾病实体，“TREATS”为二者之间的关系，即存在治疗作用。在 SemRep 语义关系中，“antb”“bacs”“elii”“imft”“inch”“irda”“orch”“phsu”均可表示药物，“acab”“anab”“comd”“dsyn”“emod”“fndg”“inpo”“mobd”“neop”“patf”“sosy”均可表示疾病，“TREATS”“DIAGNOSES”“PREVENT”“METHOD_OF”“MANIFESTATION_OF”“MEASURES”均可表示治疗。基于以上语义关系筛选出研究所需的医学实体数据。

3.3 构建“药物－疾病”二分网络

定义 1（二分网络）：$V = \{V_1, V_2, \cdots, V_p\}$ 表示药物节点集合，$U = \{U_1, U_2, \cdots, U_g\}$ 表示疾病节点集合，$G = (U, V, L)$ 表示“药物－疾病”二分网络，U、V 分别表示两类节点的集合，$L \subseteq (U \times V)$ 表示网络中连接两类节点的边的集合，如果药物对疾病有治疗作用，那么该节点对之间就有一条连边。W_{ij} 表示节点 i 和节点 j 之间的非负权值，定义边权重为节点对共现频次。

定义 2（二分网络表示学习）：给定一个二分网络 $G = (U, V, L)$，通过映射函数 f：$V \cup U \longrightarrow R^d$

将二分网络中的每个节点映射为低维向量。

3.4 基于网络表示学习的医学实体关系预测

将数据按照 8∶2 划分为训练集与测试集，同时保持节点连接关系与原始数据集的一致性，输入网络表示学习模型 BiNE，利用训练集训练网络中节点的嵌入表示，使用测试集评估训练后模型的性能和泛化能力，最后利用受试者工作特征（receiver operating characteristic，ROC）曲线衡量模型的表现。

3.4.1 显式关系建模 在“药物－疾病”二分网络中，将节点间的联系表示为显式关系。通过两个连接节点间的局部邻近性建模显式关系，计算网络中药物节点 V_i 和疾病节点 U_j 间的经验分布，其中，W_{ij} 表示边 e_{ij} 的权值。若节点以较大的权值强连接，表明药物对疾病具有治疗作用的概率也更高。

$$p(i,j) = \frac{W_{ij}}{\sum_{e(i,j)\in E} W_{ij}} \tag{1}$$

在嵌入空间中，使用 Sigmoid 函数计算药物节点 V_i 和疾病节点 U_j 间的内积，得到 V_i 与 U_j 的联合概率分布，其中，$\vec{V}_i$ 和 $\vec{U}_j$ 分别为药物节点 V_i 和疾病节点 U_j 的向量表示。

$$\hat{p}(i,j) = \frac{1}{1+\exp(-\vec{V}_i^T \vec{U}_j)} \tag{2}$$

为学习显式关系在空间中的嵌入，利用 KL 散度最小化经验分布与联合概率分布之间的差值，确保两类节点的向量表示在嵌入空间中产生聚类效果，定义最小化目标函数：

$$\begin{aligned}\text{minimize}O_1 &= \mathrm{KL}(p\|\hat{p}) = \sum_{e_{ij}\in E} p(i,j)\log\frac{p(i,j)}{\hat{p}(i,j)} \\ &\propto -\sum_{e_{ij}\in E} W_{ij}\log\hat{p}(i,j)\end{aligned} \tag{3}$$

3.4.2 隐式关系建模 隐式关系为“药物－疾病”二分网络中包含的两个同质网络内部的节点关系，两个同质网络的权重矩阵 W^V 和 W^U 为：

$$W^V = \sum_{k\in U} W_{ik}W_{jk} \tag{4}$$

$$W^U = \sum_{k\in V} W_{ki}W_{kj} \tag{5}$$

为生成真实有效的语料库，在两个同质网络上执行有偏自适应随机游走，生成学习高阶隐式关系的语料库 D^V 和 D^U，并使用 Skip－gram 模型学习顶点嵌入。定义语料库 D^V 的目标函数为：

$$\text{maximize } O_2 = \prod_{V_i\in D^V}\prod_{V_c\in\emptyset(V_i)} p(V_c \mid V_i) \tag{6}$$

其中，V_i 是给定的中心节点，$V_c \in \emptyset(V_i)$ 为节点 V_i 在节点序列中的上下文节点，U_i 同理。D^U 的目标函数为：

$$\text{maximize } O_3 = \prod_{U_i\in D^U}\prod_{U_c\in\emptyset(U_i)} p(U_c \mid U_i) \tag{7}$$

定义节点 V_c 为序列中心节点 V_i 的条件概率 $p(V_c \mid V_i)$ 以及节点 U_c 为序列中心节点 U_i 的条件概率 $p(U_c \mid U_j)$ 分别为：

$$p(V_c \mid V_i) = \frac{\exp(\vec{V}_i^T \vec{V}'_c)}{\sum_{k=1}^{|V|}\exp(\vec{V}_i^T \vec{V}'_k)} \tag{8}$$

$$p(U_c \mid U_j) = \frac{\exp(\vec{U}_j^T \vec{U}'_c)}{\sum_{k=1}^{|U|}\exp(\vec{U}_j^T \vec{U}'_k)} \tag{9}$$

节点 V、U 作为上下文时的向量用 $\vec{V}'$ 和 $\vec{U}'$ 表示，实现式（6）和（7）中定义的目标最大化，使上下文相似的节点在嵌入空间中也彼此接近。由于在式（8）中遍历同质网络节点对时计算量大且耗时，因此利用局部敏感哈希算法优化，获取高质量且多样化的负样本。令 $N_S^{ns}(V_i)$ 表示序列 S 中的中心顶点 U_i 的 ns 负样本，式（8）中条件概率 $p(V_c \mid V_i)$ 可近似为：

$$p(V_c, N_S^{ns}(V_i) \mid V_i) = \prod_{z\in|V_c|\cup Nns_{S(V_i)}} p(z \mid V_i) \tag{10}$$

$$p(z \mid V_i) = \begin{cases}\sigma(\vec{V}_i^T \vec{V}'_z), z\in\emptyset(V_i) \\ 1-\sigma(\vec{V}_i^T \vec{V}'_z), z\in N(V_i)\end{cases} \tag{11}$$

其中，σ 表示 Sigmoid 激活函数 $1/(1+e^{-x})$，$p(U_c \mid U_j)$ 同理。

3.4.3 联合优化目标函数 为同时保留显式和隐式关系嵌入二分网络，将目标函数形成一个联合优化框架，其中，α、β、γ 为将框架中不同成分组合在一起的超参数，分别代表隐式关系和显式关系对学习节点表示的影响。

$$\text{maximize } O = \alpha\log O_2 + \beta\log O_3 - \gamma O_1 \tag{12}$$

首先，最大化联合目标函数 $-\gamma O_1$，优化随机显式关系。

$$\vec{V}_i = \vec{V}_i + \lambda\{\gamma w_{ij}[1-\sigma(\vec{V}_i^T \vec{U}_j)]\vec{U}_j\} \tag{13}$$

$$\vec{U}_j = \vec{U}_j + \lambda\{\gamma w_{ij}[1 - \sigma(\vec{V}_i^T \vec{U}_j)]\vec{V}_i\} \quad (14)$$

其次，最大化联合目标函数 $\alpha\log O_2 + \beta\log O_3$，优化隐式关系。$I(Z,V_i)$ 为指示函数，确定顶点 Z 是否在 V_i 的上下文中，$I(Z,U_j)$ 同理：

$$\vec{V}_i = \vec{V}_i + \lambda\{\sum_{Z\in\emptyset(V_i)\cup N(V_i)} \alpha[I(Z,V_i) - \sigma(\vec{V}_i^T \vec{V}_Z)]\vec{V}'_Z\} \quad (15)$$

$$\vec{U}_j = \vec{U}_j + \lambda\{\sum_{Z\in\emptyset(U_i)\cup N(U_i)} \alpha[I(Z,U_i) - \sigma(\vec{U}_i^T \vec{U}_Z)]\vec{U}'_Z\} \quad (16)$$

最后，上下文向量更新公式：

$$\vec{V}'_Z = \vec{V}'_Z + \lambda\{\alpha[I(Z,V_i) - \sigma(\vec{V}_i^T \vec{V}_Z)]\vec{V}'_Z\} \quad (17)$$

$$\vec{U}'_Z = \vec{U}'_Z + \lambda\{\alpha[I(Z,U_i) - \sigma(\vec{U}_i^T \vec{U}_Z)]\vec{U}'_Z\} \quad (18)$$

3.4.4　基于机器学习的链路预测　选择逻辑回归、支持向量机、随机森林、K 近邻、决策树预测分析，将精确率、准确率、召回率、*F*1 值作为评价指标并选择最优预测模型。由于不同的特征相似度适用不同场景，因此在模型中度量余弦相似度、欧氏距离、皮尔逊相关系数、曼哈顿距离的特征重要性，最终在最优模型中输入最重要特征获取预测结果。

4　实验与结果分析

4.1　构建“药物－疾病”二分网络

将药物实体结合 BERN2 数据库与药融云数据库进行词汇验证，去除“Pharmaceutical Preparations”等无效节点；将疾病实体输入 BERN2 数据库进行词汇验证，去除“Disease Progression”等无效节点，最终得到药物节点1 227个，疾病节点1 075个，节点连边 4 172 条。将其输入 Gephi－0.10.1 筛选 K－Core≥5 的节点生成无向加权“药物－疾病”二分网络，见图 2。

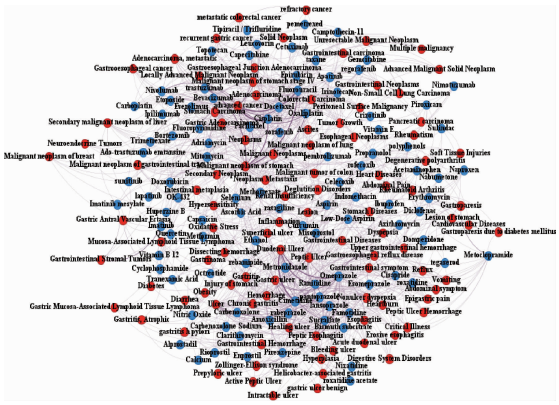

图 2　“药物－疾病”二分网络

4.2　二分网络结构分析

对二分网络加权投影得到药物网络与疾病网络，分别计算度中心性、介数中心性和特征向量中心性并获取排名前 10 位的计算结果，见表 1—2。在药物网络中，奥美拉唑、西咪替丁等与其他具有高中心性的药物紧密相连。奥美拉唑广泛应用于治疗急、慢性消化系统酸相关性疾病，包括胃食管反流、消化性溃疡等；西咪替丁能够抑制由组胺、分肽胃泌素、胰岛素和食物等刺激引起的胃酸分泌，对因化学刺激引起的腐蚀性胃炎具有预防和保护作用。在疾病网络中，胃恶性肿瘤、晚期癌症等节点具有更多连边，在传递节点信息方面具有关键作用，在临床上属于发病率高、严重程度高的胃部疾病，能够更好地帮助分析发病机制、研究并发症等。

表 1　药物节点中心性

药物	度中心性	药物	介数中心性	药物	特征向量中心性
奥美拉唑	103	奥美拉唑	211 061.771 8	西咪替丁	0.397 4
西咪替丁	99	西咪替丁	204 831.236 8	奥美拉唑	0.365 6
氟尿嘧啶	72	阿司匹林	193 911.454 9	雷尼替丁	0.314 3
顺铂	68	雷尼替丁	109 004.920 3	氟尿嘧啶	0.308 7
雷尼替丁	66	氟尿嘧啶	108 513.338 5	顺铂	0.295 2
阿司匹林	57	顺铂	84 353.268 7	阿司匹林	0.269 6
硫糖铝	41	泮托拉唑	72 158.817 3	泮托拉唑	0.239 5
兰索拉唑	38	西沙必利	68 494.447 7	硫糖铝	0.239 0
泮托拉唑	36	塞来昔布	57 787.831 4	曲妥珠单抗	0.217 0
法莫替丁	35	甲氧氯普胺	55 304.090 6	多西他赛	0.213 9

表 2　疾病节点中心性

疾病	度中心性	疾病	介数中心性	疾病	特征向量中心性
胃恶性肿瘤	275	胃恶性肿瘤	692 935.675 2	胃恶性肿瘤	1
晚期癌症	93	恶性肿瘤	136 316.483 2	晚期癌症	0.372 9
恶性肿瘤	80	胃溃疡	125 373.268 4	恶性肿瘤	0.324 8
肿瘤	67	消化性溃疡	111 088.855 1	肿瘤	0.297 6
消化性溃疡	64	晚期癌症	106 270.449 5	胃溃疡	0.273 6
胃溃疡	63	肿瘤	95 111.766 1	消化性溃疡	0.270 5
十二指肠溃疡	62	溃疡	90 418.723 4	溃疡	0.251 2
溃疡	56	病变	78 998.294 6	十二指肠溃疡	0.244 9
胃癌	50	胃炎	74 122.941 4	胃癌	0.221 0
胃炎	43	肿瘤转移	73 210.567 1	胃炎	0.194 7

4.3　基于二分网络表示学习的医学实体关系预测分析

4.3.1　获取节点向量表示　利用 PyCharm2013.1.2 运行网络表示学习模型，设置向量表示维度为 128，设置参数：$\alpha = 0.01$，$\beta = 0.01$，$\gamma = 1$，$ws = 5$，$p = 0.15$，$ns = 4$。ROC 曲线是常用的二分类模型评估指标，其能显示模型在不同阈值下的分类性能，提供综合的度量指标。ROC 曲线下面积（area under curve，AUC）通常越接近 1，表明模型性能越好。计算 AUC 值为 0.79，见图 3，表明该模型性能较好。

4.3.2　基于机器学习的链路预测分析　利用 PyCharm2013.1.2 对逻辑回归、支持向量机、随机森林、K 近邻、决策树分别计算精确率、准确率、召回率、*F*1 值，结合过采样与欠采样算法，综合比较并根据结果进行排序，见表 3。随机森林在 5 种模型预测中效果最好。因此，选取随机森林进一步分析。

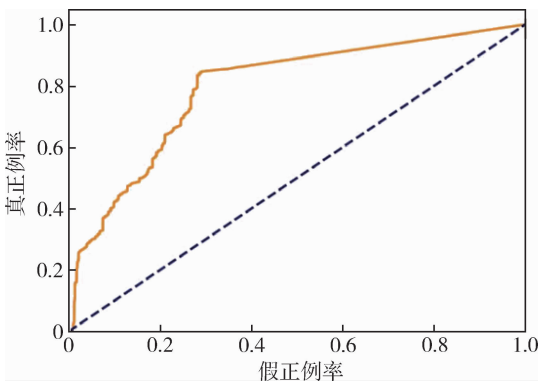

图 3　网络表示学习 ROC 曲线评价

表 3　机器学习模型评价

排序	机器学习模型	精确率	准确率	召回率	*F*1
1	随机森林	0.999 0	0.999 9	0.998 0	0.999 0
2	决策树	0.992 5	0.999 9	0.985 1	0.992 5
3	K 近邻	0.986 8	0.974 3	0.999 9	0.986 9
4	逻辑回归	0.751 4	0.807 1	0.660 7	0.726 6
5	支持向量机	0.748 3	0.895 8	0.562 0	0.690 7

在随机森林模型中，将余弦相似度、欧氏距离、皮尔逊相关系数、曼哈顿距离作为特征，利用随机森林的特征重要性度量 4 种特征。计算结果，见图 4，曼哈顿距离作为特征输入模型对预测结果的贡献更大。

取排名前 10 位的预测结果，见表 4，预测前未连接的节点对情况用 0 表示，预测后发生连接关系的情况用 1 表示。

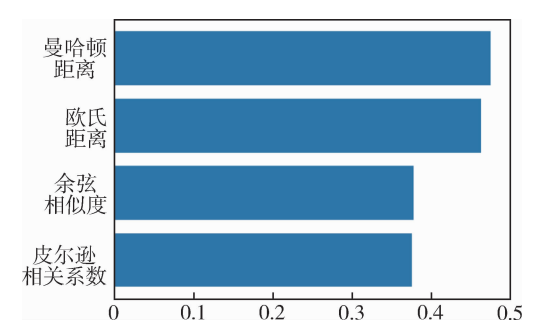

图 4　特征重要性分析

表 4 随机森林预测排名前 10 位存在潜在联系的节点对

药物实体	疾病实体	预测前连接情况	曼哈顿距离	预测后连接情况
特罗司他	高分化胰腺内分泌肿瘤	0	5.391 7	1
特罗司他	尖端扭转型室性心动过速	0	5.456 0	1
特罗司他	晚期癌症	0	5.503 1	1
盐酸西维美林	干燥综合征	0	5.538 8	1
特罗司他	创伤感染	0	5.561 3	1
托芬那酸	掌趾感觉丧失性红斑	0	5.677 3	1
特罗司他	室性心律失常	0	5.731 0	1
特罗司他	G1 级肿瘤	0	5.811 1	1
特罗司他	上消化道疾病	0	5.844 5	1
特罗司他	弥漫性溃疡	0	5.932 9	1

预测结果显示特罗司他与高分化胰腺内分泌肿瘤曼哈顿距离最短，表明在网络中连接路径比较紧密，存在较强的潜在交互关系。此外，特罗司他与尖端扭转型室性心动过速、晚期癌症、创伤感染、室性心律失常、G1 级肿瘤、上消化道疾病、弥漫性溃疡存在潜在治疗关系。DrugBank 数据库显示特罗司他是特罗司他乙酯的活性代谢物，是一种色氨酸羟化酶的抑制剂，通过抑制色氨酸羟化酶减少外周血清素的产生，用于治疗类癌综合征腹泻。类癌又称神经内分泌肿瘤，发生于消化、呼吸、循环、泌尿生殖系统等部位，是相对罕见的起源于神经内分泌特性的嗜银细胞肿瘤。类癌瘤是胃肠道和胰腺内分泌肿瘤异质性的一部分，分泌类癌相关激素如 5 - 羟色胺、缓激肽和其他介质，这些物质被认为是导致腹泻、潮红和喘息症状的原因，这些症状称为类癌综合征[9]。随着时间的推移，类癌综合征患者可能会发展为类癌性心脏病，其中包括心律失常、室性心动过速[10]。研究[11]表明，类癌与上消化道疾病、弥漫性溃疡均有关联，血清素升高与凝血功能障碍在医学上存在作用关系，而凝血功能障碍与伤口感染存在相关性[12]。晚期神经内分泌肿瘤中约 8% ~35% 的患者出现与血清素相关的类癌综合征[13]。此外，DrugBank 数据库表明盐酸西维美林可用于治疗干燥综合征，而干燥综合征在一些情况下可导致肠梗阻[14]。掌趾感觉丧失性红斑与癌症治疗中使用的化疗药物有关，属于炎症性皮肤反应，症状为刺痛感和感觉迟钝，伴随手掌和足底温度升高，可发展为灼痛、肿胀和红斑，而托芬那酸是一种非甾体抗炎药，具有镇痛、解热、抗癌的特性[15, 16]。

5 结语

本研究基于 PubMed 文献摘要数据，构建“药物 - 疾病”二分网络，通过社会网络分析发现药物网络与疾病网络中的重要节点，全面理解药物与疾病在网络中的特征和关系。然后学习节点的低维向量表示，结合机器学习思想预测药物与疾病之间是否存在潜在治疗关系。结果表明，利用医学知识网络可预测节点间未知关系，为研究人员在药物研究等方面提供更加准确的视角。

尽管本研究在预测医学实体关系方面表现出良好效果，但由于采用的数据为短文本，在分析长文本时仍存在一些不足：首先，全文数据节点数量更多，导致节点嵌入的计算复杂度与难度也相应增加，降低模型训练与推断的效率；其次，全文数据涵盖大量专业术语和知识，在预测阶段可能出现模型泛化能力不足的问题。因此采用文献全文进行验证是下一步研究方向。此外，没有考虑在动态网络结构视角下的热点与变化，未来将融合时间因素预测药物节点与疾病节点间的潜在联系。

作者贡献：吴胜男负责研究设计、论文修订；吴佳辉负责研究设计、数据下载与分析、论文撰写与修订；董继宗负责参与研究设计；蒋环宇、王璐琦、王欣瑶负责数据处理。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 王建霞，刘梦琳，许云峰，等. 异构网络表示学习方法

综述［J］. 河北科技大学学报, 2021, 42（1）: 48 – 59.
2 邹然, 柳杨, 李聪, 等. 图表示学习综述［J］. 北京师范大学学报（自然科学版）, 2023, 59（5）: 716 – 724.
3 WU Y, LAN W, FAN X, et al. Bipartite network influence analysis of a two – mode network［J］. Journal of econometrics, 2024, 239（2）: 105562.
4 张佳慧, 张婷, 吕来水, 等. 基于加权投影的二分网络的链路预测［J］. 计算机应用与软件, 2021, 38（3）: 264 – 268, 297.
5 张斌, 马费成. 科学知识网络中的链路预测研究述评［J］. 中国图书馆学报, 2015, 41（3）: 99 – 113.
6 贺圣平, 王会军, 李华, 等. 机器学习的原理及其在气候预测中的潜在应用［J］. 大气科学学报, 2021, 44（1）: 26 – 38.
7 HU F, ZHANG Y, YAN X Y, et al. An improved heterogeneous graph convolutional network for inter – relational medicine representation learning［J］. IEEE multimedia, 2023, 30（1）: 52 – 61.
8 余黄樱子, 董庆兴, 张斌. 基于网络表示学习的疾病知识关联挖掘与预测方法研究［J］. 情报理论与实践, 2019, 42（12）: 156 – 162.
9 DRUCE M, ROCKALL A, GROSSMAN A B. Fibrosis and carcinoid syndrome: from causation to future therapy［J］. Nature reviews endocrinology, 2009, 5（5）: 276 – 283.
10 RAM P, PENALVER J L, LO K B U, et al. Carcinoid heart disease: review of current knowledge［J］. Texas heart institute journal, 2019, 46（1）: 21 – 27.
11 丁炎波, 陈炳芳, 庄耘, 等. 上消化道类癌的诊断和治疗［J］. 中国医药指南, 2013, 11（8）: 592 – 594.
12 EDINOFF A N, RAVEENDRAN K, COLON M A, et al. Selective serotonin reuptake inhibitors and associated bleeding risks: a narrative and clinical review［J］. Health psychology research, 2022, 10（4）: 39580.
13 RORSTAD O. Prognostic indicators for carcinoid neuroendocrine tumors of the gastrointestinal tract［J］. Journal of surgical oncology, 2005, 89（3）: 151 – 160.
14 KAKIMOTO K, INOUE T, TOSHINA K, et al. Multiple mesenteric panniculitis as a complication of sjögren's syndrome leading to ileus［J］. Internal medicine, 2016, 55（2）: 131 – 134.
15 SANKPAL U T, GOODISON S, JONES – PAULEY M, et al. Tolfenamic acid – induced alterations in genes and pathways in pancreatic cancer cells［J］. Oncotarget, 2017, 8（9）: 14593 – 14603.
16 YANG B, XIE X, WU Z, et al. DNA damage – mediated cellular senescence promotes hand – foot syndrome that can be relieved by thymidine prodrug［J］. Genes & disease, 2023, 10（6）: 2557 – 2571.

（上接第 67 页）

10 周梅佳佳, 朱庆华. 基于 LDA 主题模型的智慧养老研究主题分析［J］. 医学信息学杂志, 2024, 45（3）: 8 – 15.
11 徐文秀. 基于 LSTM 和 LDA 的可再生能源领域主题分类研究［D］. 济南: 山东大学, 2020.
12 李散散. 基于用户行为分析和 LDA 模型的数字媒体推荐系统的设计与实现［J］. 现代电子技术, 2020, 43（7）: 146 – 149.
13 张东鑫, 张敏. 图情领域 LDA 主题模型应用研究进展述评［J］. 图书情报知识, 2022, 39（6）: 143 – 157.
14 RÖDER M, BOTH A, HINNEBURG A. Exploring the space of topic coherence measures［EB/OL］.［2024 – 06 – 25］. https: //dl. acm. org/doi/10. 1145/2684822. 2685324.
15 GRIFFITHS T L, STEYVERS M. Finding scientific topics［J］. Proceedings of the national academy of sciences, 2004, 101（S1）: 5228 – 5235.
16 陶胜阳, 许新华, 余亚烽, 等. 基于 LDA 模型的教育技术学研究主题挖掘及演化趋势分析［J］. 现代信息科技, 2023, 7（6）: 176 – 180.
17 袁永旭, 李黛, 王思源, 等. 基于 LDA 模型的我国医保政策主题建模与演化分析［J］. 预防医学情报杂志, 2023, 39（6）: 710 – 716.
18 江锐鹏, 钟广玲. 中文分词神器 Jieba 分词库的应用［J］. 电脑编程技巧与维护, 2023（9）: 87 – 89.
19 ŘEHŮŘEK R, SOJKA P. Software framework for topic modelling with large corpora［C］. Valletta: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.
20 倪维健, 孙浩浩, 刘彤, 等. 面向领域文献的无监督中文分词自动优化方法［J］. 数据分析与知识发现, 2018, 2（2）: 96 – 104.
21 GROOTENDORST M. BERTopic: neural topic modeling with a class – based TF – IDF procedure［EB/OL］.［2024 – 06 – 25］. https: //arxiv. org/pdf/2203. 05794.