文章编号：2095-2163(2023)01-0158-06 中图分类号：R965.1;R285.5 文献标志码：A

融入异构网络特征的深度学习预测中药靶点

黄群富[1]，丁长松[1,2]
（1 湖南中医药大学 信息科学与工程学院，长沙 410208；2 湖南省中医药大数据分析实验室，长沙 410208）

摘 要：针对传统预测中药靶点相互作用，忽略了中药、成分、靶点 3 者之间的潜在联系，导致特征提取不足，模型精度不高的问题，构建中药-成分-靶点 3 层异构网络。利用重启随机游走、高斯核、信息熵算法提取网络特征，并提出了运用深度神经网络分析中药成分-靶点相互作用的模型 TCMIT-DNN。实验结果表明，GBDT、RF、SVM 模型使用 TCMIT 3 层异构网络策略后，分类性能均有所提升。TCMIT-DNN 的 *AUC*、*F*1 值、准确率分别为 96.0%、89.5%、89.5%，均优于 TCMIT-GBDT、TCMIT-RF、TCMIT-SVM 分类模型。

关键词：中药；靶点预测；异构网络；深度神经网络

Prediction of traditional Chinese medicine targets based on deep learning on heterogeneous network features

HUANG Qunfu[1], Ding Changsong[1,2]
(1 College of Information science and Engineering, Human University of Chinese Medicine, Changsha 410208, China;
2 Laboratory of traditional Chinese medicine in Hunan Province, Changsha 410208, China)

【Abstract】 The existing prediction methods of traditional Chinese medicine target interaction ignores the potential relationship among traditional Chinese medicine, ingredient and targets, resulting in insufficient feature extraction and low model accuracy. To address the problems, this paper constructs tripartite heterogeneous network of traditional Chinese medicine ingredient targets, and extracts the network features by restarting random walk, Gaussian kernel and information entropy algorithms, proposes a TCMIT-DNN model to analyze the interactions between ingredient and target of traditional Chinese medicine by deep neural network. Experimental results show that the classification performance of GBDT, RF and SVM models is improved after using TCMIT tripartite heterogeneous network strategy. The *AUC*, *F*1 and *accuracy* of TCMIT-DNN are 96.0%, 89.5% and 89.5% respectively, which are better than TCMIT-GBDT, TCMIT-RF and TCMIT-SVM classification models.

【Key words】 traditional Chinese medicine; target prediction; heterogeneous network; deep neural network

0 引 言

中医临床经验丰富、疗效显著，但对中药成分、治疗靶点的作用机制仍知之甚少，给临床精准治疗带来了极大挑战。然而，中药具有多成分、多靶点等特点，很多潜在成分与靶点间的关系尚未明确。通过生物实验，分别从中药的成分研究其作用靶点花费的时间、经济成本大且难以实现。因此，研究快速高效的中药成分-靶点相互作用预测方法亟不可待。

中药靶点发现的关键，在于探究中药多成分与多靶点的相互作用关系。现有的定量结构活性关系方法预测中药靶点方法，主要以分子指纹、分子描述符结合机器学习为主[1]，忽略了中药、成分、靶点 3 者之间的潜在联系，不利于模型的泛化调用。目前，网络分析已广泛应用于疾病分类、生物医疗、新药研发等领域，其有效性已在实践中得到验证。如：Hao 等[2]针对药物-靶点相互作用，提出一种双网络集成逻辑矩阵分解的相似性度量方法；于亚运等[3]基于分子指纹相似度构建中药成分-靶点相互作用分类模型。此类方法的准确度很大程度依赖于分子结构相似性。近年来，深度神经网络（Deep Neural

基金项目：湖南省中医药科研计划重点课题（2020002；长沙市自然科学基金（kq2202265）。

作者简介：黄群富（1998-），男，硕士研究生，主要研究方向：深度学习与计算药物分子设计；丁长松（1975-），男，博士，教授，博士生导师，主要研究方向：大数据技术与中医药信息学。

通讯作者：丁长松 Email：dingcs1175@ hnucm.edu.cn

收稿日期：2022-06-10

Network，DNN）结合传统算法已成功应用于海量、复杂的药物－靶点网络拓扑结构分析[4]。如：使用 DNN 和因子分解机实现自动学习特征的高阶及低阶表达式[5]；利用 XGBoost 确定药物分子指纹非冗余特征[6]，并通过 DNN 提高药物靶点分类模型精度等。

本文针对传统中药靶点预测忽略中药、成分、靶点之间的潜在联系和药物-靶点网络研究中存在特征提取不全、过于依赖结构相似性的问题，提出了一种基于中药－成分－靶点（Traditional Chinese Medicine-Ingredient-Target，TCMIT）3 层异构网络的中药靶点预测方法。利用结构相似性和关联矩阵构建 TCMIT 异构网络，使用数学统计算法提取网络拓扑特征，并结合深度学习建立 TCMIT-DNN 中药成分-靶点相互作用分析模型，通过异构网络从分子维度分析中药治疗疾病的现代物质基础。

1　方　法

基于成分 SMILES 相似性、靶点氨基酸序列相似性，以及已知相互作用的中药－成分、成分－靶点关系，构建 TCMIT 3 层异构网络。采用重启随机游走、高斯核算法，分别提取层内相似性网络和层间异构网络的拓扑特征；结合信息熵，分别融合成分、靶点特征矩阵，并利用 DNN 构建分类模型。本文整体框架如图 1 所示。

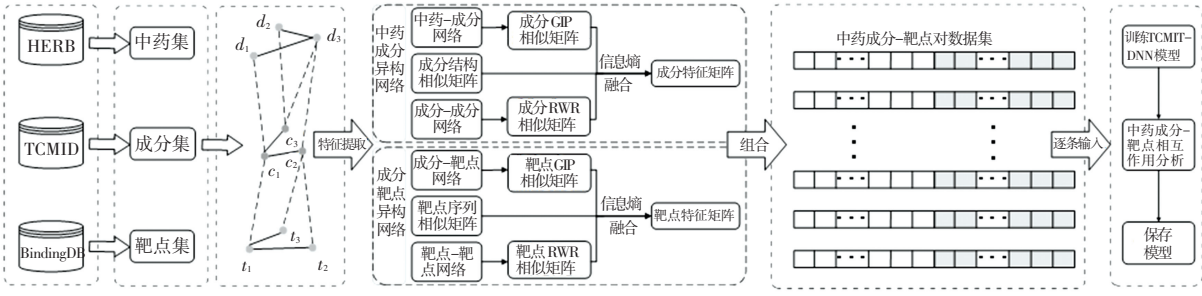

图 1　融入“中药-成分-靶点”异构网络特征的深度学习预测中药靶点框架

Fig. 1　Prediction of traditional Chinese medicine targets based on deep learning on integrated “TCM-ingredient-target” network features

1.1　结构相似性计算

令 $D=\{d_1,d_2,\cdots,d_I\}$ 表示中药集合；I 为中药总数；$C=\{c_1,c_2,\cdots,c_J\}$ 表示中药包含的成分集合；J 为成分总数；$T=\{t_1,t_2,\cdots,t_K\}$ 表示种属来源于 Human 物种的靶点集合；K 为靶点总数。

利用 Jaccard 算法[7]，分别计算成分扩展连通性指纹向量的结构相似性，构建成分相似矩阵 $\boldsymbol{SIM}_{ingre}\in\boldsymbol{R}^{J\times J}$。公式如下：

$$\boldsymbol{SIM}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j)=\frac{\boldsymbol{c}_i\cap\boldsymbol{c}_j}{\boldsymbol{c}_i\cup\boldsymbol{c}_j}\tag{1}$$

式中，$\boldsymbol{c}_i$、$\boldsymbol{c}_j$ 分别表示两种成分的指纹向量。

利用史密斯－沃特曼（Smith－Waterman）算法[8]，计算两个不等长氨基酸序列的相似性，构建靶点结构相似矩阵 $\boldsymbol{SIM}_{target}\in\boldsymbol{R}^{K\times K}$。公式如下：

$$\boldsymbol{SIM}_{target}(t_i,t_j)=\max\begin{cases}\max\ (0,(\boldsymbol{SIM}_{target}(t_{i-1},t_{j-1})+s))\\ \max\ (0,(\boldsymbol{SIM}_{target}(t_i,t_{j-1})-w))\\ \max\ (0,(\boldsymbol{SIM}_{target}(t_{i-1},t_j)-w))\end{cases}\tag{2}$$

式中，空位罚分数 w 设为 2，若当前对比的两个元素相同，则 s 为 3，否则 s 为 -3。

1.2　TCMIT 网络构建

中药、成分、靶点分别作为 3 个相似性网络的节点，根据节点间的相互作用关系，定义连接中药层与成分层的邻接矩阵 $\boldsymbol{M}\in\boldsymbol{R}^{I\times J}$、连接成分层和靶点层的邻接矩阵 $\boldsymbol{N}\in\boldsymbol{R}^{J\times K}$。当矩阵中存在相互作用关系时编码为 1，否则编码为 0。编码为 1 表示异构网络相应的节点间存在连边，编码为 0 则不存在连边，分别构建“中药-成分”、“成分-靶点”异构网络；以成分层为连接层，将“中药-成分”、“成分-靶点”异构网络融合为 TCMIT 3 层异构网络（如图 1 中第三部分所示）。

1.3　相似性网络拓扑特征提取

中药成分-靶点相互作用的预测过程，可被视为节点同时在成分层相似性网络和靶点层相似性网络随机游走的过程。重启随机游走（Random Walk with Restart，$\boldsymbol{RWR}$），对于解决具有多种异构拓扑结构的生物网络计算具有一定优势[7]，可利用相似性网络中的拓扑相似性构建 $\boldsymbol{RWR}$ 相似矩阵。以成分层网络为例：

定义成分层转移概率矩阵 $\boldsymbol{T}_C\in\boldsymbol{R}^{J\times J}$，其中 $\boldsymbol{T}_C(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 为随机游走过程中，成分节点 i 到达 j 的概率，计算公式如下：

$$\boldsymbol{T}_C(\boldsymbol{c}_i,\boldsymbol{c}_j)=\frac{\boldsymbol{SIM}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j)}{\sum_{j=1}^{J}\boldsymbol{SIM}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j)}\tag{3}$$

定义矩阵 $\boldsymbol{R}_C \in \boldsymbol{R}^{J\times J}$，其中 $\boldsymbol{R}_c^t(i,j)$ 表示成分 i 经过 t 论迭代后到达 j 的概率，$\boldsymbol{R}_c^t$ 矩阵计算公式如下：

$$\boldsymbol{R}_c^t = (1-a)\,\boldsymbol{R}_c^{t-1}\,\boldsymbol{T}_C + a\,\boldsymbol{R}_C^0 \tag{4}$$

其中，$\boldsymbol{R}_C^0$ 为初始随机游走矩阵，a 为重启概率，表示成分在每轮迭代时返回出发点的概率。

通过重启概率限制了成分游走的范围，从而扩大成分局部拓扑信息对结果的影响。随着迭代的不断进行，$\boldsymbol{R}_c^t$ 会逐渐收敛，收敛达到一定条件归一化后，将其视为成分层网络的相似矩阵 $\boldsymbol{RWR}_{ingre} \in \boldsymbol{R}^{J\times J}$。收敛条件如公式(5)所示，类似计算靶点层相似性网络的相似矩阵 $\boldsymbol{RWR}_{target} \in \boldsymbol{R}^{K\times K}$。

$$\sum_i \sum_j \mid \boldsymbol{R}_c^{t+1}(i,j) - \boldsymbol{R}_c^t(i,j) \mid < 10^{-6} \tag{5}$$

1.4 异构网络拓扑特征提取

利用药物靶点的相互作用关系，计算药物高斯核相互作用属性(Gaussian Interaction Profile，***GIP***)相似性方法[9]，同时计算“中药-成分”和“成分-靶点”异构网络间拓扑结构相似性 $\boldsymbol{GIP}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和 $\boldsymbol{GIP}_{target} \in \boldsymbol{R}^{K\times K}$。以 $\boldsymbol{GIP}_{ingre}$ 为例，计算公式如下：

$$\boldsymbol{GIP}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j) = \exp\left(-\gamma_d \parallel f(\boldsymbol{c}_i) - f(\boldsymbol{c}_j) \parallel^2\right) \tag{6}$$

$$\gamma_d = \gamma_d^{'} \Big/ \left(\frac{1}{J}\sum_{a=1}^{J} \parallel f(\boldsymbol{c}_a) \parallel^2\right) \tag{7}$$

其中，$f(\boldsymbol{c}_i)$ 表示在邻接矩阵 $\boldsymbol{M}$ 中，成分 $\boldsymbol{c}_i$ 与所有中药的对应关系；γ_d 为控制核宽度的调节参数；J 为成分集合的总数；$\gamma_d^{'}$ 的值则是根据使用高斯核的经验而设置。

1.5 特征融合

计算相似矩阵信息熵可获得其携带多少信息，信息熵越小表示该相似矩阵中随机信息越少，从而能为特征矩阵提供更大、更丰富的信息量。在异构网络中，使用信息熵算法融合各特征矩阵，降低矩阵中数据噪声的影响。以矩阵 $\boldsymbol{SIM}_{target} \in \boldsymbol{R}^{K\times K}$ 为例，信息熵计算如下：

$$E = -\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} P(t_i,t_j)\,\log_2(P(t_i,t_j))}{k} \tag{8}$$

其中，$P(t_i,t_j)$ 表示靶点节点 i 和 j 在网络中相连的概率值，计算公式如下：

$$P(t_i,t_j) = \frac{\boldsymbol{SIM}_{target}(t_i,t_j)}{\sum_{j=1}^{k} \boldsymbol{SIM}_{target}(t_i,t_j)} \tag{9}$$

$\boldsymbol{SIM}_{ingre}$、$\boldsymbol{RWR}_{ingre}$、$\boldsymbol{RWR}_{target}$、$\boldsymbol{GIP}_{ingre}$、$\boldsymbol{GIP}_{target}$ 矩阵的信息熵值计算与 $\boldsymbol{SIM}_{target}$ 矩阵类似。根据熵值确定各矩阵融合权重，分别将成分和靶点的结构信息、相似性网络拓扑信息、异构网络拓扑信息线性融合，构建成分特征矩阵 $\boldsymbol{FEA}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和靶点特征矩阵 $\boldsymbol{FEA}_{target} \in \boldsymbol{R}^{K\times K}$。融合公式如下：

$$\boldsymbol{FEA}_{ingre}(i,j) = \alpha_1\,\boldsymbol{SIM}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j) + \beta_1\boldsymbol{RWR}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j) + \gamma_1\,\boldsymbol{GIP}_{ingre}(\boldsymbol{c}_i,\boldsymbol{c}_j) \tag{10}$$

$$\boldsymbol{FEA}_{target}(i,j) = \alpha_2\,\boldsymbol{SIM}_{target}(t_i,t_j) + \beta_2\boldsymbol{RWR}_{target}(t_i,t_j) + \gamma_2\,\boldsymbol{GIP}_{target}(t_i,t_j) \tag{11}$$

其中，$\boldsymbol{FEA}_{ingre}(i,j) \in [0,1]$ 表示成分 $\boldsymbol{c}_i$ 与 $\boldsymbol{c}_j$ 经信息融合后的值，$\boldsymbol{FEA}_{target}(i,j)$ 与其类似。

1.6 TCMIT-DNN 分类模型

DNN 采用多层神经网络结构，将复杂映射分解为一系列嵌套的简单映射，以逐层抽象实现从局部特征到整体特征提取解决复杂问题。异构网络的拓扑属性可表示为节点的特征向量，利用 DNN 的非线性拟合能力构建 TCMIT-DNN 模型，预测异构网络上中药成分和靶点的相互作用。当邻接矩阵 $\boldsymbol{N}_{(i,j)}=1$ 时，表示 $\boldsymbol{c}_i$ 与 $\boldsymbol{t}_j$ 存在相互作用，则将 $\boldsymbol{c}_i$ 与 $\boldsymbol{t}_j$ 视为中药成分 - 靶点对正例样本($y=1$)，当邻接矩阵 $\boldsymbol{N}_{(i,j)}=0$ 时，则将其视为负例样本($y=0$)，样本特征向量 $\boldsymbol{v}$ 定义如下：

$$\boldsymbol{v} = \begin{cases} concat(\boldsymbol{FEA}_{ingre}(i,:),\boldsymbol{FEA}_{target}(j,:)),\ y=1\ N_{(i,j)}=1 \\ concat(\boldsymbol{FEA}_{ingre}(i,:),\boldsymbol{FEA}_{target}(j,:)),\ y=0\ N_{(i,j)}=0 \end{cases} \tag{12}$$

其中，$\boldsymbol{FEA}_{ingre}(i,:)$ 表示矩阵 $\boldsymbol{FEA}_{ingre}$ 的第 i 行，$\boldsymbol{FEA}_{target}(j,:)$ 表示矩阵 $\boldsymbol{FEA}_{target}$ 的第 j 行。因此，$\boldsymbol{FEA}_{ingre}(i,:)$ 和 $\boldsymbol{FEA}_{target}(j,:)$ 经 $concat(\cdot)$ 拼接后，生成 $(J+K)$ 维的样本特征向量 $\boldsymbol{v}$，J 和 K 分别为成分、靶点数据集总数。

TCMIT-DNN 模型由一个输入层、3 个隐含层和一个输出层组成。样本特征向量 $\boldsymbol{v}$ 由输入层神经元流向下一层神经元，通过 3 个隐含层的非线性函数运算后传递至输出层，输出 $\boldsymbol{v}$ 预测为正例和负例的概率值。

2 实验结果与分析

2.1 数据

本文采用的数据来源于中药药理学数据库和药物化学数据库。在 BindingDB 数据库(网址 http://www.bindingdb.org/)中收集所有包含 Human 物种来源的靶点，共计 2 135 个，将靶点信息在 TCMID 数据库(网址 http://www.megabionet.org/)中查询其具有相互作用的成分，共计 1 633 个，将成分信息在 Herb 数据库(网址 http://herb.ac.cn/)查询其具有

所属关系的中药，共计1 558个，并收集成分 SMILES（Simplified Molecular Input Line Entry Specification，SMILES）信息及靶点氨基酸序列信息。

2.2　实验设计

2.2.1　建立 TCMIT-DNN 分类模型

中药集合 D、成分集合 C、靶点集合 T 的数量 I、J、K 分别为 1 558、1 633、2 135，由 Jaccard 和 Smith-Waterman算法分别计算中药成分和靶点的结构相似性，构建结构相似矩阵 $\boldsymbol{SIM}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和 $\boldsymbol{SIM}_{target} \in \boldsymbol{R}^{K\times K}$，并结合中药-成分和成分-靶点的关联关系构建 TCMIT 异构网络。在 RWR 算法中，初始随机游走矩阵 $\boldsymbol{R}_C^0$ 主对角线的值为 1，其余值为 0；重启概率 a 设置为 0.5；基于成分-成分和靶点-靶点网络构建具有相似性网络拓扑特征的矩阵 $\boldsymbol{RWR}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和 $\boldsymbol{RWR}_{target} \in \boldsymbol{R}^{K\times K}$。在 GIP 算法中，调节核宽度的参数 γ_d' 和 γ_c' 设置为 1，基于中药-成分和成分-靶点网络构建具有异构网络拓扑特征的矩阵 $\boldsymbol{GIP}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和 $\boldsymbol{GIP}_{target} \in \boldsymbol{R}^{K\times K}$；分别计算 $\boldsymbol{SIM}_{ingre}$、$\boldsymbol{RWR}_{ingre}$、$\boldsymbol{RWR}_{target}$、$\boldsymbol{GIP}_{ingre}$、$\boldsymbol{GIP}_{target}$ 和 $\boldsymbol{SIM}_{target}$ 矩阵的信息熵值，并确定特征矩阵融合权重，结果见表 1。融合后生成中药成分特征矩阵 $\boldsymbol{FEA}_{ingre} \in \boldsymbol{R}^{J\times J}$ 和靶点特征矩阵 $\boldsymbol{FEA}_{target} \in \boldsymbol{R}^{K\times K}$，并将中药成分、靶点结合邻接矩阵 $\boldsymbol{N} \in \boldsymbol{R}^{J\times K}$ 拼接生成中药成分-靶点对，作为 DNN 的输入。

表 1　相似矩阵信息熵值

Tab. 1　Information entropy of similarity matrix

相似矩阵	$\boldsymbol{SIM}_{ingre}$	$\boldsymbol{RWR}_{ingre}$	$\boldsymbol{GIP}_{ingre}$	$\boldsymbol{SIM}_{target}$	$\boldsymbol{RWR}_{target}$	$\boldsymbol{GIP}_{target}$
信息熵值	10.03	10.29	10.45	11.06	11.04	10.58
权重参数	α_1	β_1	γ_1	α_2	β_2	γ_2
权重值	0.34	0.33	0.33	0.34	0.34	0.32

中药成分-靶点对存在相互作用的 38 286 条数据作为正例样本集，将中药成分和靶点随机组合，可以获取 344.816 9 万条未知标签的组合样本数据，并在未知标签的数据中随机选取与正例样本集数量相同的作为负例样本集[2]。生成正例和负例样本集后，将其混合并打乱生成 76 572 * 3 768 的样本数据，按比例 8∶1∶1 划分训练集、验证集和测试集。

本文选用 python3.7 编程语言结合 Pytorch 框架建立 TCMIT-DNN 模型。模型为 5 层网络结构，输入层神经元数为中药成分 $\boldsymbol{c}_i$ 与靶点 $\boldsymbol{t}_j$ 特征向量维数之和，共计 3 768 个；隐含层神经元数分别为 128、64、32；输出层神经元数为 2 个；采用非线性激活函数 *ReLU*，学习率设为 0.001，batch_size 一次性输入模型中的样本数设为 64，epochs 数据训练轮次设为 50；模型的损失函数 *loss* 采用交叉熵（Cross Entropy Loss），其公式如（13）：

$$L = \frac{1}{n}\sum_{i=1}^{n} -y_i\log(p_i) - (1-y_i)\log(1-p_i) \quad (13)$$

式中，n 为样本数量，y_i 表示第 i 个中药成分-靶点对的实际标签，正例为 1，负例为 0，p_i 表示样本 i 预测为正例标签的概率。

2.2.2　模型性能指标

为检验 TCMIT-DNN 模型的分类结果并对模型进行评估，遵循二分类模型评估指标，采用 *ROC* 曲线下面积（Area Under the ROC Curve，*AUC*）、准确率（Accuracy，*ACC*）和 *F*1 值（F-Measure）从不同角度评估模型性能。

2.3　对比实验分析

2.3.1　消融实验

为检验 TCMIT-DNN 模型整合异构网络拓扑特征的有效性，分别使用包含传统属性特征的 STR-DNN 模型、包含层内相似性网络拓扑特征的 RWR-DNN 模型、包含层间异构网络拓扑特征的 GIP-DNN 模型进行对比；为检验信息熵融合相似矩阵的有效性，使用相似矩阵融合权重取均值的 ENT-DNN 模型进行对比。各模型相似矩阵融合权重见表 2，实验结果见表 3。

表 2　5 种算法相似矩阵权重

Tab. 2　Weight of similarity matrix of 5 algorithms

算法	$\boldsymbol{SIM}_{ingre}$	$\boldsymbol{RWR}_{ingre}$	$\boldsymbol{GIP}_{ingre}$	$\boldsymbol{SIM}_{target}$	$\boldsymbol{RWR}_{target}$	$\boldsymbol{GIP}_{target}$
STR-DNN	1	0	0	1	0	0
RWR-DNN	0	1	0	0	1	0
GIP-DNN	0	0	1	0	0	1
TCMIT-DNN	0.34	0.33	0.33	0.34	0.34	0.32
ENT-DNN	1/3	1/3	1/3	1/3	1/3	1/3

在相同测试集、实验参数和评价标准下,使用信息熵整合网络拓扑特征的 TCMIT-DNN 模型 *AUC* 值、*F*1 值、*ACC* 值均为最高,较传统属性特征 STR-DNN 模型分别提升了 4%、5.6%、5.4%。结果表明,本文整合异构网络拓扑特征,有助于中药成分-靶点相互作用分析模型性能提升,同时信息熵算法有利于降低相似矩阵数据噪声的影响。

表 3 5 种算法性能比较

Tab. 3 Performance comparison of 5 algorithms

算法	*AUC*/%	*F*1/%	*ACC*/%
STR-DNN	92.0	83.9	84.1
RWR-DNN	93.9	87.2	87.3
GIP-DNN	94.1	87.3	87.3
TCMIT-DNN	**96.0**	**89.5**	**89.5**
ENT-DNN	95.6	88.7	88.8

2.3.2 与基线模型对比

为检验 TCMIT-DNN 模型在中药成分-靶点相互作用分析优越性,将其与近年来基于指纹相似度常用的随机森林(Random Forest, RF)模型[3]、梯度提升树(Gradient Boosting Decision Tree, GBDT)模型[10]、支持向量机(Support Vector Machine, SVM)模型[11]进行对比实验。RF、GBDT、SVM 模型采用成分和靶点结构相似性作为输入,利用网格搜索法寻找最优参数组合,TCMIT-RF、TCMIT-GBDT、TCMIT-SVM 分别为 RF、GBDT、SVM 模型在使用 TCMIT 3 层异构网络策略后的模型,其中分类模型的参数保持一致。RF 分类模型的参数为:子树的数量为 100,最大深度为 10;GBDT 分类模型的参数为:子树的数量为 50,最大深度为 5,子采样系数为 0.7;SVM 分类模型的参数为:惩罚系数 *C* 为 1,核函数为线性核函数。

表 4 常用算法性能比较

Tab. 4 Performance comparison of common algorithms

算法	*AUC*/%	*F*1/%	*ACC*/%
TCMIT-DNN	**96.0**	**89.5**	**89.5**
RF	91.3	82.4	83.1
GBDT	93.8	86.7	86.7
SVM	93.8	88.4	88.1
TCMIT-RF	95.3	88.0	88.8
TCMIT-GBDT	95.4	88.8	88.6
TCMIT-SVM	94.2	89.0	88.7

由表 4 可知,在相同测试集上的评价指标表明,TCMIT-DNN 具有最高的 *AUC*、F_1 值和准确率,TCMIT-GBDT和 TCMIT-RF 模型效果稍差,TCMIT-RF 模型效果较差。究其原因,是由于 TCMIT-DNN 模拟人脑的工作原理建立多个函数单元,以及其强大的非线性拟合能力,能很好地模拟成分和靶点的子结构,并有效处理具有空间拓扑特征的不规则数据,通过验证集调整确定网络结构参数,建立高精度判别模型。实验结果还表明,在对中药成分和靶点数据进行 TCMIT 3 层异构网络的构建和网络特征提取后,GBDT、SVM 和 RF 模型的性能均有不同程度的提升。其中,TCMIT 网络结合 DNN 模型效果最好,表明 TCMIT 异构网络策略能有效提取中药多成分、多靶点之间的潜在关联特征,从而提升中药成分与靶点相互作用的分类性能。

2.4 案列分析

以黄芪为例基于 TCMIT-DNN 模型预测中药成分和靶点的关系。将黄芪的成分信息按 TCMIT 异构网络策略编码后,输入模型得到预测靶点,并利用 Cytoscape 软件构建黄芪成分-靶点网络,由网络图的度筛选出重要潜在靶点,并通过药物化学、药理等理论分析结果,以此验证 TCMIT-DNN 模型的有效性。

利用预测结果构建黄芪成分-靶点网络,计算黄芪成分-靶点网络图的度值,并以排序前 10 的靶点作为最终的潜在靶点,最终结果见表 5。

表 5 黄芪靶点预测结果分析

Tab. 5 Analysis of target prediction results of Huangqi

靶点	基因名	度值	验证方式
P31389	HRH1	33	文献[12]
Q9H7Z6	KAT8	31	文献[13]
Q9UQB9	AURKC	31	-
P15559	NQO1	30	TCMSP、BATMAN-TCM
P08183	ABCB1	30	BATMAN-TCM
Q96GD4	AURKB	30	ETCM
Q92499	DDX1	30	文献[14]
P08588	ADRB1	30	TCMSP
Q05469	LIPE	29	文献[15]
Q76LX8	ADAMTS13	29	文献[16]

将结果进行中药化学数据库验证,以及从 DrugBank、OMIM 数据库和文献中获取靶点功能和已知药物进行分析。分析结果表明,NQO1、ABCB1、AURKB、ADRB1 均得到中药化学数据库验证,在其余靶点中,HRH1 基因大量表达于平滑肌和神经元中参与觉醒、情绪和激素分泌的控制,靶向 HRH1 有助于早期治疗一些自身免疫性疾病[12];Huai 等[13]发现,KAT8 通过促进 I 型干扰素的产生,保护

突变小鼠抗病毒感染；DDX1 基因与抗病毒免疫应答、肿瘤发生发展密切相关[14]；LIPE 的缺失会导致胰岛素抵抗、糖尿病的风险增加[15]；ADAMTS13 是一种多结构域蛋白酶，其缺陷会导致微血管过程触发血小板和内皮细胞的补体激活，从而引发血栓性微血管病[16]。上述结果体现了黄芪镇静、增强免疫力、抗突变、抗病毒、抗肿瘤、降血糖、预防周围血管病变的药理作用。

3　结束语

中药治疗通过多成分、多靶点、多环节、多途径综合调节，作用于机体治疗疾病，其作用机理具有明显的空间拓扑结构，且其拓扑结构具有明显的异构特性。现有的基于分子结构相似性分析中药成分-靶点相互作用的方法忽略了中药、成分、靶点 3 者之间的复杂关系，导致分析结果不够精准。本研究通过 TCMIT3 层异构网络建立中药、成分、靶点 3 者之间的联系，并利用数学统计结合深度学习技术预测中药靶点。实验表明所有对比模型结合 TCMIT 异构网络策略时分类性能均得到提升，其中 TCMIT-DNN 模型均优于其他常用模型，并将模型有效应用于黄芪的中药靶点预测。本研究的中药靶点虚拟筛选方法充分利用了不同特征的优势，降低了传统依赖于结构相似性特征和单一描述符的风险，同时拟合了中药多成分、多靶点之间潜在联系，减轻特征提取的影响，有望应用于分子维度分析中药的现代物质基础。

参考文献

[1] 郑一夫，孔令雷，贾 皓，等. 基于系统的化合物-靶点相互作用预测模型的消栓通络方抗脑卒中网络药理学研究[J]. 药学学报，2020，55(2)：256-264.

[2] HAO M, BRYANT S H, WANG Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization [J]. Scientific reports, 2017, 7(1): 1-11.

[3] 于亚运，刘勇国，蒋 羽，等. 基于指纹相似度的药物-靶点相互作用预测[J]. 中国中药杂志，2017，42(18)：3578-3583.

[4] ZHAO Q, YANG M, CHENG Z, et al. Biomedical data and deep learning computational models for predicting compound-protein relations[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2021, 19(4): 2092-2110.

[5] WANG J, WANG H, WANG X, et al. Predicting drug-target interactions via FM-DNN learning[J]. Current Bioinformatics, 2020, 15(1): 68-76.

[6] CHEN C, SHI H, HAN Y, et al. DNN-DTIs: improved drug-target interactions prediction using XGBoost feature selection and deep neural network[J]. bioRxiv, 2020, 16(2): 768-779.

[7] LUO H, WANG J, LI M, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random Walk algorithm[J]. Bioinformatics, 2016, 32(17): 2664-2671.

[8] WEN HUI W, SEN Y, XIANG Z, et al. Drug repositioning by integrating target information through a heterogeneous network model[J]. Bioinformatics, 2014, 30(20): 2923-2930.

[9] ZHENG Y, WU Z. A Machine Learning-Based Biological Drug-Target Interaction Prediction Method for a Tripartite Heterogeneous Network[J]. ACS omega, 2021, 6(4): 3037-3045.

[10] QIU W, LV Z, HONG Y, et al. BOW-GBDT: A GBDT Classifier Combining With Artificial Neural Network for Identifying GPCR-Drug Interaction Based on Wordbook Learning From Sequences[J]. Frontiers in Cell and Developmental Biology, 2021, 8: 1789.

[11] MAHMUD S, CHEN W, JAHAN H, et al. Dimensionality Reduction based Multi-Kernel framework for Drug-Target Interaction Prediction[J]. Chemometrics and Intelligent Laboratory Systems, 2021, 212: 104270.

[12] NOUBADE R, MILLIGAN G, ZACHARY J F, et al. Histamine receptor H1 is required for TCR-mediated p38 MAPK activation and optimal IFN-γ production in mice[J]. Journal of Clinical Investigation, 2007, 117(11): 3507-3518.

[13] HUAI W, LIU X, WANG C, et al. KAT8 selectively inhibits antiviral immunity by acetylating IRF3KAT8 acetylates IRF3 and inhibits IFN-I production[J]. The Journal of experimental medicine, 2019, 216(4): 772-785.

[14] 王 沐，侯 晋. DDX 解旋酶家族分子功能的研究进展[J]. 中国肿瘤生物治疗杂志，2020，27(10)：1162-1169.

[15] ALBERT J S, YERGES-ARMSTRONG L M, HORENSTEIN R B, et al. Null mutation in hormone-sensitive lipase gene and risk of type 2 diabetes[J]. New England Journal of Medicine, 2014, 370(24): 2307-2315.

[16] TATI R, KRISTOFFERSSON A C, STAHL A L, et al. Complement activation associated with ADAMTS13 deficiency in human and murine thrombotic microangiopathy[J]. The Journal of Immunology, 2013, 191(5): 2184-2193.