一种基于拓扑信息的预测疾病相关的MicroRNAs方法

高 鹏[1],陈智华[2]
(1.华中科技大学人工智能与自动化学院,湖北武汉 430074;2.广州大学计算科技研究院,广东广州 510006)

摘 要: 研究表明,micRNAs(miRNA)突变或者异常会导致多种疾病,鉴定出与疾病相关的microRNA可以帮助诊断和治疗相关疾病.然而,通过生物实验方式获取准确关联关系,花费大,而且周期长.因此,本文中提出了一种基于网络内部拓扑信息的机器学习方法(HNDLM)预测疾病-miRNA关联关系.HNDLM避免搭建相似网络,而是将近年提出的network embedding方法应用在生物网络上.实验结果显示,HNDLM相较于MIDPE,MIDP,WBSMDA,RLSMDA,CPTL,HDMP经典算法在准确率和AUC值上效果更好.此外,在案例学习中,HNDLM推荐的前30个候选miRNAs,基本都能通过先前实验得到证实,HNDLM能发现潜在疾病-miRNA关系,有助于进一步研究疾病发病机制,推动生物信息学发展.

关键词: miRNA;网络嵌入;异质网络;连接预测;拓扑信息;机器学习
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2020)02-0333-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.02.016

A Method for Predicting Disease-Related MicroRNAs Based on Topological Information

GAO Peng[1],CHEN Zhi-hua[2]
(1. *School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*;
2. *Institute of Computing Science and Technology, Guangzhou University, Guangzhou, Guangdong* 510006, *China*)

Abstract: Studies show that mutations or abnormalities in micRNAs can lead to many diseases, and the identification of disease-associated microRNAs (miRNAs) can help diagnose and treat related diseases. However, it is costly and long-term to obtain accurate correlations through biological experiments. Therefore, this paper proposes a machine learning method (HNDLM) that uses network topology information to predict disease-miRNA associations. HNDLM avoids the construction of similarity networks, but applies the network embedding method proposed in recent years to biological networks. Experimental results show that HNDLM performs better than MIDPE, MIDP, WBSMDA, RLSMDA, CPTL, HDMP classical algorithms in accuracy and AUC value. In case study, the top 30 candidate miRNAs recommended by HNDLM can be confirmed by previous experiments. HNDLM can discover the potential disease-miRNA relationship and help to further study the pathogenesis of the disease, promote the development of bioinformatics.

Key words: miRNA; network emmbedding; heterogeneous network; link prediction; topology information; machine learning

1 引言

MicroRNAs(miRNAs)是一类微小的内源性非编码RNA,长度约有20~24个核苷酸,通过碱基配对来控制mRNA的降解和表达.研究表明,miRNAs在很多生物活动中起着重要的作用,比如细胞的发育,分化,凋亡,代谢和信号传导等[1].同时,miRNA的突变和调控异常与多种疾病相关,发现疾病-miRNA关系有助于研究疾病发病机制,治疗方案等.但是通过实验的方法发现与疾病相关的miRNA消耗巨大,周期长,从而发展出通过计

收稿日期:2018-12-05;修回日期:2019-03-06;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61876047,No.61702200)

算的方式获取潜在的疾病-miRNA 连接[2].

miRNA-疾病连接预测的研究主要基于两个方面:(1)相似网络;(2)机器学习模型。Jiang 等人提出了第一种计算方法,该方法构建了人类疾病-miRNA 网络,通过累计超几何分布,计算相似性分数[3]. 2013 年,Xuan 等人提出了一种基于 K 最近邻的 miRNA-疾病连接预测算法,简称 HDMP[4]. Chen 等人[5]提出了 RWRMDA 模型,该方法通过在 miRNA 功能相似网络上采用随机游走来识别潜在 miRNA-疾病关联关系. 此外,Shi 等人[6]通过考虑 miRNA-目标关联,疾病-基因关系,以及蛋白质-蛋白质作用网络来改进 RWRMDA 以及 Xuan 等人提出了了 MIDP[7]算法,该方法在带重启的随机游走基础上使用不同节点特征.

机器学习类方法在 miRNA-疾病关联预测也有很多应用,Jiang 等人[8]构建了 miRNA 功能相似网络和疾病表型网络,并结合支持向量机进行预测. Xu 等人基于 miRNA 表达信息的特征,在 miRNA 靶基因失调网络上训练支持向量机,预测疾病-miRNA 关系[9]. 此外,Chen 等研究人员提出了一种基于正则化最小二乘法的半监督方法 RLSMDA[10],在无需负样本的情况下,可以对所有疾病进行预测. 但是,现有的基于相似性方法主要是依靠网络外的生物知识来计算节点间的相似性,从而对网络进行编码,尚未在基于相似性的方法中利用生物网络节点之间的丰富拓扑信息来预测疾病-miRNA 关系,而研究表明,异构网络中生物实体之间丰富的拓扑相互作用对关联预测具有重要价值[11].

本文为了探究网络拓扑信息在预测疾病相关的 miRNA 的重要性,提出了 HNDLM 方法,采用 DeepWalk 挖掘并用向量表示网络节点的拓扑信息. 深度学习揭示了大型网络顶点的特征,可以适应基于相似性或者机器学习的解决方案,以提供灵活的疾病-miRNA 预测方法. DeepWalk[12]作为一种深度学习方法,不仅可以进行大型网络特征学习,而且注重网络局部特征. 同时,为了避免搭建相似网络,考虑到已知疾病-miRNA 关联关系较少,HNDLM 在节点拓扑特征基础上,利用机器学习模型 Deep Forest[13]进行预测.

2 数据和网络

2.1 数据

本文整理了 Ding 等人研究数据[14],选择了疾病,miRNAs,基因三种节点类型. SIDD 数据库[15]整理了丰富的疾病-基因关系数据,可以获取已知的疾病-基因关联关系. MiRTarbase[16]是一个 miRNA-靶标相互作用的数据库,它收集了已经通过实验验证的 miRNA-靶目标关系,本文通过 MiRTarbase 数据库获取已知 miRNA 和基因关系. 对于初始已知的疾病-miRNA 关系来源于 HMDD[17],该数据库收集了大量可靠的人类 miRNA-疾病关联数据. 表 1 中列出了详细的数据统计结果及来源:

表 1 网络中节点和连接信息

节点类型	统计	数据来源
疾病	330	文献[16]
miRNA	1724	文献[16]
基因	6180	文献[16]
疾病-miRNA 连接	5219	HMDD v3.0 (http://www.cuilab.cn/hmdd)
miRNA-基因连接	242860	MiRTarbase (http://mirtarbase.mbc.nctu.edu.tw/) DIANA-Tarbase (http://www.microrna.gr/tarbase)
疾病-基因连接	19476	SIDD(http://mlg.hit.edu.cn/SIDD)

2.2 疾病-miRNA-基因网络

针对疾病相关的 miRNA 预测问题,常见方法是构建疾病-miRNA 双层网络进行计算. 但是,基于以下生物理论:miRNA 调控一些已知的致病基因或者与致病基因功能类似的基因,从而导致功能障碍引发疾病以及 Ding 等人的研究,本文构建了包含疾病,基因,miRNA 三种节点类型的异质网络,如图 1 所示. 圆点 D 代表疾病,G 代表基因,M 代表 miRNA,虚线表示已知的节点间关系.

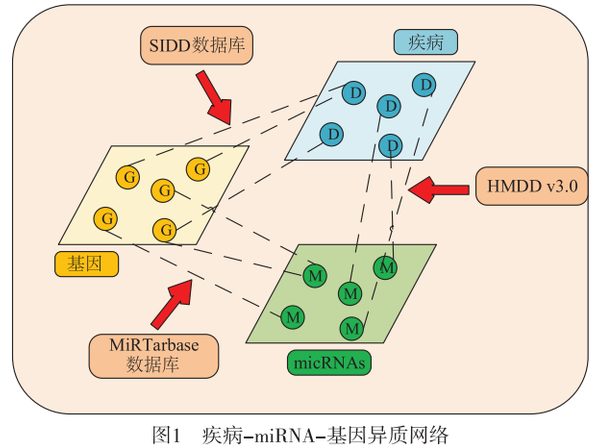


图1 疾病–miRNA–基因异质网络

3 HNDLM 算法

3.1 算法描述

HNDLM 利用网络表示学习方法—Deepwalk 来挖掘疾病-miRNA-基因异质网络中拓扑信息,使生物网络中每一个节点获得表示其拓扑信息的向量. 在预测部分,基于疾病和 miRNA 的拓扑向量构建数据集,并利用

Deep Forest 模型来进行预测.

3.2　挖掘拓扑信息

Deepwalk 作为一种深度学习方法,主要有两个部分组成:(1)随机游走[18]生成器随机均匀地选取网络节点,并生成固定长度的随机游走序列. 对于每一个节点 v_i,随机游走生成器为其生成长度为 t 的 γ 个随机游走序列;(2)对于每一次游走,采用 SkipGram 算法来更新节点的向量表示. SkipGram 算法通过最大化窗口 w 内的顶点共同出现的概率来更新节点的表示向量 $\boldsymbol{\Phi}$,如式(1)所示:

$$P_r(\{v_{i-w},\cdots,v_{i+w}\}\backslash v_i \mid \boldsymbol{\Phi}(v_i)) = \prod_{j=i-w}^{i+w} P_r(v_j \mid \boldsymbol{\Phi}(v_i)) \tag{1}$$

其中,$\boldsymbol{\Phi}$ 为节点的表示向量,大小为 $|V| \times n$,$|V|$ 为节点数量,n 为用户设定的表示向量长度,v_i 为网络中节点标号,参数 w 为 SkipGram 选择的中心节点相邻区域形成的窗口大小. 为了加速计算,可以将每个节点分配到哈夫曼树的叶子上,$P_r(v_j \mid \boldsymbol{\Phi}(v_i))$ 可以用 Hierarchial Softmax[19]来近似,此时 $P_r(v_j \mid \boldsymbol{\Phi}(v_i))$ 计算成式(2):

$$P_r(v_j \mid \boldsymbol{\Phi}(v_i)) = \prod_{l=1}^{\log|V|} 1/(1 + e^{-\boldsymbol{\Phi}(v_i)\boldsymbol{\Psi}(b_l)}) \tag{2}$$

其中,$b_l \in \{b_0, b_1, \cdots, b_{\log|V|}\}$,$\boldsymbol{\Psi}(b_l)$ 是节点 b_l 父节点的表示向量. $b_l \in \{b_0, b_1, \cdots, b_{\log|V|}\}$ 是用来标识顶点 v_j 的顶点序列,是从根节点到节点 v_j 的路径,其中 b_0 为根节点,$b_{\log|V|}$ 表示节点 v_j.

HNDLM 基于 Deepwalk 算法提取网络中拓扑特征的工作流程见图 2,将构建的疾病-基因-miRNA 网络图 G 作为输入数据,经过 Deepwalk 处理,会为网络中每个节点生成一个表示其拓扑信息的向量. 算法主要参数:窗口大小 w 为 5,表示每个节点的向量长度 n 为 128,对于每一个节点,随机游走生成的序列数量 γ 为 100,游走产生序列的长度 t 为 50. 在算法输出结果中,可以根据网络节点对应编号,分别获得 D_m 个疾病和 M_h 个 miRNA 的表示向量组成的矩阵. 实验中,Deepwalk 算法代码实现可以从 deeplearning4j library 获取(http://deeplearning4j.org/).

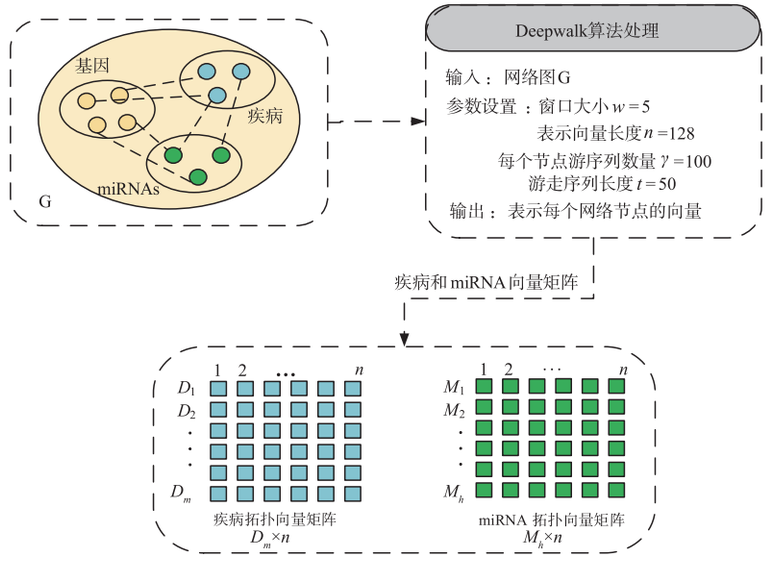


图2　Deepwalk算法处理流程

3.3　预测 miRNA-疾病连接

预测部分,基于网络中疾病,miRNA 节点拓扑特征,HNDLM 构造了适合学习的新特征,并利用 Deep Forest 模型来获得疾病-miRNA 潜在连接的可能性. Deep Forest 模型[13]具有与深度学习神经网络相当的效果,但是参数更少,训练简单,而且能够在小样本上表现出良好的效果. DeepForest 结构由两部分组成:(1) Multi-Grained Scanning,类似卷积神经网络,即将多个相邻特征进行分组处理;(2) Cascade Forest,将若干个弱分类器集成得到森林并再次集成,每层都是 4 个森林构成. 如图 3 所示,以一个维度为 N 的样本为例,首先在 Multi-Grained Scanning 部分,滑动窗口大小 W(对应了 CNN 中不同的卷积器大小),滑动采样后每个样本生成 $N-W+1$ 个 W 维的样本实例,每个实例代表了一种局部结构或者亚采样样本. 假设共有 m 个样本,则每个实例可以获得 m 个训练样本. 对于二分类问题,将这些样本实例作为随机森林输入,可以得到一个 2 维类概率分布,则所有的实例可以组成一个 $2 \times (N-W+1)$ 维的概率分布,将概率分布连在一起作为特征;Cascade Forest 部分,输入的特征在经过 level 1 的 4 个随机森林之后,转换为了 4 个维度为 2 的向量,向量每一维的含义是该样例属于该类别的概率,接下来在输入 level 2

的时将这 8 维的向量和之前的输入拼接起来进行输入,依次类推直到最后一层,把最后一层的 4 个 2 维向量对位取平均得到分类结果.

HNDLM 搭建了 Deep Forest 模型,选取的滑动窗口大小 W 为 100,分类类别 $n\_classes$ 为 2. 为了预测疾病-miRNA 关系,需要基于疾病,miRNA 表示向量构造用于训练的特征集. HNDLM 中采用的方法是将疾病,miRNA 表示向量进行拼接构造成新的特征向量,从而实现对应的疾病和 miRNA 信息相互关联,同时增加特征维度,避免信息缺失. 实验中,设定的 DeepWalk 表示向量长度为 128,则将 128 维的疾病 d 的表示向量与 128 维的 miRNA m 的表示向量拼接起来,生成一个 256 维的向量作为模型训练样本. 如果疾病 d 与 miRNA m 有已知连接,则将样本标签设为 1,否则设为 0,处理之后样本总和为 330 × 1724,具体流程见图 4. 实验中,HNDLM 将已知的 5219 个疾病-miRNA 关系作为训练数据正样本,而在未知连接中,由于潜在连接只占很少部分,所以从大量未知连接样本抽取少量作为负样本具有可行性,本文中正负样本比例为 1∶1,避免正负样本不均衡问题. 模型完成训练之后,可以对疾病和 miRNA 之间关系作出预测.

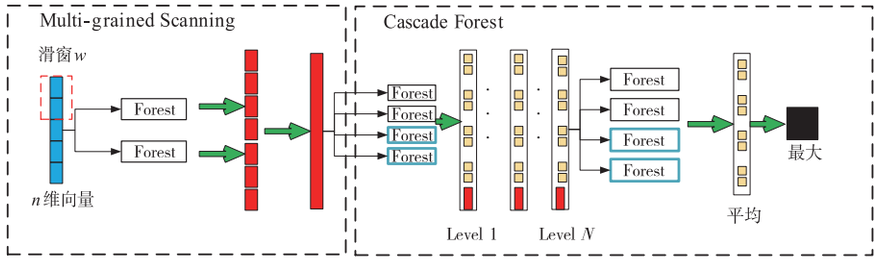


图3 Deep Forest模型

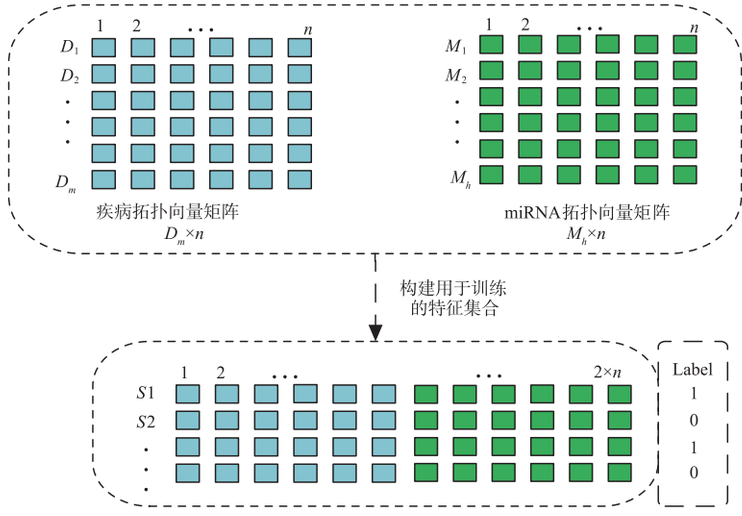


图4 预测miRNA和疾病关联

4 实验结果

4.1 评估标准

为了证明 HNDLM 方法的有效性,本文采用交叉验证的方式,即对于一种特殊的疾病,将与该疾病相关的 miRNA 分成 5 份,每次取其中一份作为测试集,其余 4 份为已知信息. 同时,本文采用三种度量方式来说明该方法在预测疾病相关的 miRNA 的表现.

AUC(ROC 曲线与坐标轴围成的面积)是一个全面评价模型表现的指标. 通过改变阈值来计算真阳率(TPR) 和假阳率(FPR),获取 ROC 曲线,具体见式(3),式(4):

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

PRE 是特定疾病的预测准确率,式(5)为所选相关样品与所选样品数的比率.

$$\mathrm{PRE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

公式中 TP 和 FP 分别是关于特定疾病的真阳性和假阳性样品的数量. 根据定义,PRE 值越大,预测精度

越好. REC 是选中疾病的召回率,见式(6):

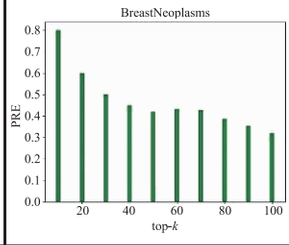
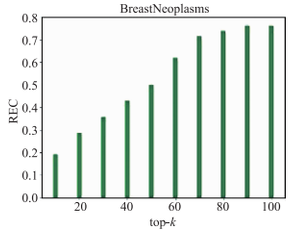
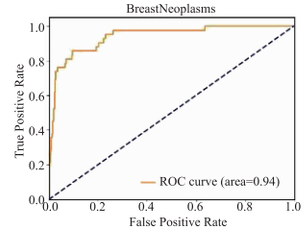
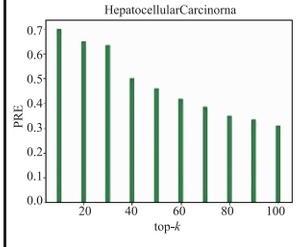
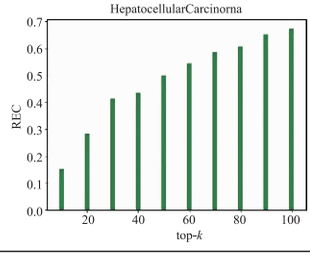
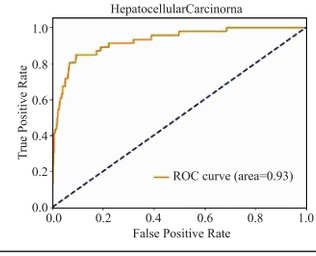
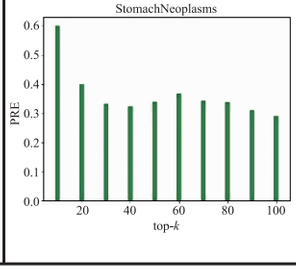
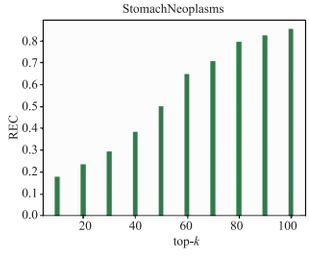
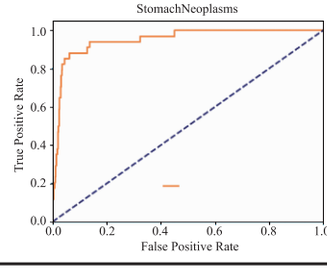
$$\mathrm{REC}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

公式中,FN 是关于特定疾病的假阴性样本的数量.

4.2　疾病-miRNA-基因网络预测效果

验证方法的有效性上,选取具有较多已知 miRNA 连接的疾病. 本文选取了三种疾病,即 Breast Neoplasm, Stomach Neoplasms, Hepatocellular Carcinoma,计算每种疾病在前 10,20,…,100 个候选 miRNAs 中的 PRE,REC 和 ROC 曲线. 其中 PRE 显示出在 top-k 个样本中,预测正确的比例,而 AUC 体现整体的预测效果. 表 2 中列出了不同疾病的 PRE,REC 和 ROC 曲线. 根据结果,可以发现本文中的方法在疾病-miRNA 的连接预测上有良好的效果,具有较高预测分数的 miRNA 更有可能与该疾病有潜在的连接.

表2　不同疾病的PRE,REC和ROC曲线

疾病名称	PRE	REC	ROC
breast neoplasm			
hepatocellular carinoma			
stomach neoplasms			

4.3　方法对比

本文选取了 6 种当前比较前沿的方法,即 MIDPE, MIDP, WBSMDA[20], RLSMDA, CPTL[21], HDMP,在相同数据集下,与本文中的方法 HNDLM 进行对比分析. 实验中 HNDLM 的具体参数为:Deepwalkw 算法部分 $w=5, n=128, \gamma=100, t=50$, Deep Forest 算法部分 $w=100$, $n\_\mathrm{classes}=2$. 对比方法中的可调参数均根据文章中作者推荐的数值进行设置,其中 RLSMDA 参数 $n_m=1, n_d=1, w=0.9$, MIDPE 算法中 $a=0.9, r=0.8$. MIDP $r_Q=0.4, r_U=0.1$, CPTL 算法参数中 $\lambda_{MM}=0.2, \lambda_{MD}=0.4, \lambda_{DD}=0.2, \lambda_{DM}=0.4$, HDMP 中 $k=10$. 为了公平比较,本文选取了 14 种已知 miRNA 连接较多的疾病,并对每种疾病求取交叉验证的平均结果. 表 3 中列出了每种疾病,在不同方法下计算得到的 AUC 数值,可以反应出不同方法的预测情况. 实验结果表明,本文方法 HNDLM 的 AUC 数值明显高于其他方法. 表中 HNDLM, MIDPE, MIDP, WBSMDA, RLSMDA, CPTL 和 HDMP 的平均 AUC 为 0.942,0.832,0.821,0.770,0.805,0.825,0.832. 可以看出,HNDLM 具有最高的平均 AUC,比其他方法高出了 11%,12.1%,17.2%,13.7%,11.7% 和 11%. 为了进一步评价 HNDLM,本文将 HNDLM 推荐分数前 50 以内的预测准确率与 MIDP 和 RLSMDA 进行对比,为了公平可靠,实际结果取 14 种疾病预测结果的平均值,从图 5 可以看出,HNDLM 在前 10,20,…,40 个候选 miRNAs 中,预测正确率高于 MIDP 和 RLSMDA. 特别是在前 10 个候选 miRNAs 中,HNDLM, MIDP 和 RLSMDA 的平均正确率为 42%,34%,28%,HNDLM 比 MIDP 高了 8%,比 RLSMDA 高了 14%. 因此,相比较其他方

法,HNDLM 不仅避免构建同质相似网络,而且预测效 果更好.

表 3 对于 14 种疾病,不同方法下的 AUC 值比较

疾病类型	HNDLM	NIDPE	MIDP	WBSMDA	RLSMDA	CPTL	HDMP
breast neoplasms	**0.909**	0.814	0.806	0.754	0.802	0.804	0.796
colorectal neoplasms	**0.951**	0.806	0.799	0.705	0.797	0.796	0.812
glioblastoma	**0.944**	0.800	0.790	0.772	0.768	0.808	0.828
heart failure	**0.942**	0.798	0.781	0.717	0.763	0.792	0.786
hepatocellular carcinoma	**0.910**	0.771	0.749	0.713	0.742	0.763	0.755
lung neoplasms	**0.974**	0.902	0.892	0.800	0.870	0.892	0.906
melanoma	**0.946**	0.820	0.812	0.759	0.801	0.822	0.823
non-small-cell lung	**0.950**	0.859	0.850	0.790	0.834	0.851	0.844
ovarian neoplasms	**0.972**	0.885	0.893	0.817	0.884	0.893	0.891
pancreatic neoplasms	**0.964**	0.901	0.882	0.807	0.859	0.895	0.891
prostatic neoplasms	**0.952**	0.838	0.830	0.827	0.805	0.830	0.852
renal cell carcinoma	**0.926**	0.814	0.804	0.766	0.784	0.803	0.810
stomach neoplasms	**0.911**	0.786	0.779	0.743	0.773	0.763	0.780
rinary bladder neoplasms	**0.936**	0.844	0.827	0.815	0.785	0.836	0.871
平均 AUC	**0.942**	0.832	0.821	0.770	0.805	0.825	0.832

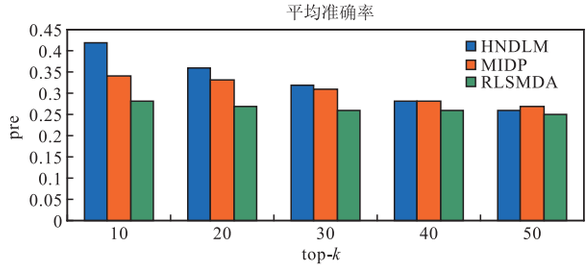


图5 HNDLM和MIDP, RLSMDA准确率比较

4.4 新的疾病-miRNA 关系预测

为了进一步证明 HNDLM 方法在预测疾病相关的 miRNA 上的有效性,本文选取了两种特殊的疾病 esophageal cancer 和 kidney neoplasms 来进行研究. 在实验中将选定的疾病当做全新未研究过的疾病看待,根据预测过程中未使用过的独立癌症数据库 dbDEMC 2.0[22],miR2Disease[23],可以对预测结果中分数较高的 miRNA 进行验证. 在本文中,选用 kidney neoplasms 作为例子进行说明,表 4 列出了预测结果中分数排名前 30 的 miRNAs. dbDEMC 是人类癌症中差异表达的 miRNA 数据库. HNDLM 推荐的前 30 个候选 miRNA 中,29 个能在 dbDEMC 得到证明. miR2Disease 是一个人工整理的数据库,目的在于提供人类疾病中 miRNA 调控异常的各种资源. HNDLM 推荐的前 30 个候选 miRNA 中有 9 个在该数据库得到证实. 根据列出的疾病实例,证明本文方法可以有效的预测疾病和 miRNA 连接.

表 4 kidney neoplasm 相关的前 30 候选 miRNAs

排名	miRNA	数据库	排名	miRNA	数据库
1	hsa-mir-155	dbDEMC	16	hsa-mir-29c	dbDEMC,miR2Disease
2	hsa-mir-34b	dbDEMC	17	hsa-mir-29a	dbDEMC,miR2Disease
3	hsa-mir-34a	dbDEMC	18	hsa-mir-26a	dbDEMC,miR2Disease
4	hsa-mir-126	dbDEMC,miR2Disease	19	hsa-mir-195	dbDEMC
5	hsa-mir-20a	dbDEMC,miR2Disease	20	hsa-mir-133a	dbDEMC
6	hsa-mir-1	Unconfirmed	21	hsa-mir-125b	dbDEMC
7	hsa-mir-146a	dbDEMC	22	hsa-mir-210	dbDEMC,miR2Disease
8	hsa-mir-200b	dbDEMC,miR2Disease	23	hsa-mir-145	dbDEMC
9	hsa-mir-221	dbDEMC	24	hsa-mir-23a	dbDEMC
10	hsa-mir-223	dbDEMC	25	hsa-mir-222	dbDEMC
11	hsa-mir-133b	dbDEMC	26	hsa-mir-16	dbDEMC
12	hsa-mir-203	dbDEMC	27	hsa-mir-122	dbDEMC,miR2Disease
13	hsa-mir-203	dbDEMC	28	hsa-mir-132	dbDEMC
14	hsa-mir-206	dbDEMC	29	hsa-mir-24	dbDEMC
15	hsa-mir-31	dbDEMC	30	hsa-mir-200a	dbDEMC,miR2Disease

5　结论

本文构建了疾病-miRNAs-基因三层网络,并提出一种新的方法 HNDLM 用于疾病-miRNA 的关联预测. HNDLM 主要思路是通过挖掘网络中节点拓扑信息作为特征来进行学习,从而对疾病相关的 miRNAs 进行预测. 为了和当前的方法进行对比,文中选取了 14 种疾病,分别在 5 折交叉验证下取平均结果,发现 HNDLM 在 AUC 值和正确率上都表现的更好. 同时,在样例学习中,以 kidney neoplasms 为例,发现预测分数排名前 30 个候选 miRNAs,其中有 29 个能在 dbDEMC,miR2Disease 数据库得到验证. 因此,可以证明 HNDLM 能有效的预测疾病和 miRNA 关联关系,有助于进一步研究疾病发病机制.

参考文献

[1] A Ambros V. The functions of animal microRNAs[J]. Nature,2004,431(7006):350 – 355.

[2] Lü L,Zhou T. Link prediction in complex networks:A Survey[J]. Physica A,2011,390(6):1150 – 1170.

[3] Jiang Q,et al. Prioritization of disease microRNAs through a human phenome-microRNAome network[J]. BMC Systems Biology,2010,4(Suppl 1):S2.

[4] Xuan P,et al. Correction:Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors[J]. PloS One,2013,8(8):e70204. 1971.

[5] Chen X,Liu M,Yan G. RWRMDA:predicting novel human mi-croRNA-disease associations[J]. Molecular BioSystems,2012,8(10):2792 – 2798.

[6] Shi H,et al. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes[J]. BMC Systems Biology,2013,7(1):101 – 101.

[7] Xuan P,et al. Prediction of potential disease-associated microRNAs based on random walk[J]. Bioinformatics,2015,31(11):1805 – 1815.

[8] Jiang Q,Wang G,Jin S,et al. Predicting human microRNA-disease associations based on support vector machine[J]. International Journal of Data Mining and Bioinformatics,2013,8(3):282 – 293.

[9] Xu J,Li C X,Lv J Y,et al. Prioritizing candidate disease mirnas by topological features in the mirna target-dysregulated network:Case study of prostate cancer[J]. Molecular Cancer Therapeutics,2011,10(10):1857 – 1866.

[10] Chen X,Yan G Y. Semi-supervised learning for potential human microRNA-disease associations inference[J]. Scientific Reports,2014,4(1):5501.

[11] Chen B,et al. Assessing drug target association using semantic linked data[J]. PLoS Comput Biol,2012a,8(7):e1002574.

[12] Perozzi B,et al. Deepwalk:Online learning of social representations[A]. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York,USA:ACM,2014. 701 – 710.

[13] ZHOU Z H,FENG J. Deep forest:towards an alternative to deep neural networks[A]. Twenty-Sixth International Joint Conference on Artificial Intelligence[C]. Melbourne,Australia:IJCAI,2017. 3553 – 3559.

[14] Ding,Pingjian,et al. Human disease MiRNA inference by combining target information based on heterogeneous manifolds[J]. Journal of Biomedical Informatics,2018,80:26 – 36.

[15] L Cheng,et al,SIDD:a semantically integrated database towards a global view of human disease[J]. PloS One,2013,8(10):e75504.

[16] C H Chou,et al. miRTarBase 2016:updates to the experimentally validated miRNA-target interactions database[J]. Nucl Acids Res,2015,44 (D1):D239 – D247.

[17] Huang Z,et al. HMDD v3.0:a database for experimentally supported human microRNA-disease associations[J]. Nucleic Acids Research,2019,47(D1):D1013 – D1017.

[18] 李敏,王晓桐,罗慧敏,孟祥茂,王建新. 随机游走技术在网络生物学中的研究进展[J]. 电子学报,2018,46(8):2035 – 2048.
LI Min,WANG Xiao-tong,LUO Hui-min,MENG Xiang-mao,WANG Jian-xin. Progress on random walk and its application in network biology[J]. Acta Electronica Sinica,2018,46(8):2035 – 2048. (in Chinese)

[19] Mnih A,Hinton. A scalable hierarchical distributed language model[A]. Proceedings of the 21st International Conference on Neural Information Processing Systems[C]. British Columbia,Canada:ACM,2008,1081 – 1088.

[20] X Chen,et al. WBSMDA:within and between score for MiRNA-disease association prediction[J]. Sci Reports,2016,6(1):21106.

[21] J. Luo,et al. Collective prediction of disease-associated miRNAs based on transduction learning[J]. IEEE/ACM Trans Comput Biol Bioinform,2016,14 (6):1468 – 1475.

[22] Yang Z. dbDEMC:a database of differentially expressed miRNAs in human cancers[J]. BMC Genomics,2010,11(Suppl 4):S5.

[23] Jiang Q,et al. miR2Disease:a manually curated database for microRNA deregulation in human disease[J]. Nucleic Acids Research,2009,37(suppl_1):D98 – D104.

作者简介

陈智华 女,1976 年出生,广西柳州人,广州大学计算科技研究院教授,研究方向智能控制,计算机控管一体化,生物计算.
E-mail:czhgd@ gzhu. edu. cn

高 鹏 男,1993 年出生,湖北随州人. 华中科技大学人工智能与自动化学院硕士,研究方向为生物信息学,机器学习.
E-mail:18202718193@ 163. com