软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):3094−3102 [doi: 10.13328/j.cnki.jos.005351]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

miRNA 与疾病关联关系预测算法*

郭茂祖[1], 王诗鸣[2], 刘晓燕[2], 田 侦[2]

[1](北京建筑大学 电气与信息工程学院,北京 100044)

[2](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 刘晓燕, E-mail: liuxiaoyan@hit.edu.cn

摘 要: microRNAs(miRNAs)在生命进程中发挥着重要作用.近年来,预测 miRNAs 与疾病的关联关系成为一个研究热点.当前,计算方法整体上可以分为两大类:基于相似度度量的方法和基于机器学习的方法.前者通过度量网络中节点之间的关联强度预测 miRNA-疾病关联,但需要构建高质量的生物网络模型;后者将机器学习相关算法应用到这个问题中,但需要构建高可信度的负例集合.基于以上困难和不足,提出了一种计算模型 BNPDCMDA,用于预测 miRNAs-疾病关联关系.该方法首先构建 miRNA-疾病双层网络模型,然后利用 miRNA 的功能相似度对其进行基于密度的聚类,进而将二分网络投影应用于聚类后的 miRNAs 及疾病集合构成的 miRNA-疾病双层子网中,最终完成对 miRNA 与疾病关联关系的预测.实验结果表明,采用留一交叉验证法得到的 AUC 值可达 99.08%,明显优于当前其他高效方法.最后,采用 BNPDCMDA 方法对某些常见疾病所关联的 miRNAs 进行预测,实验结果获得了文献的支持,进一步表明了该方法的有效性.

关键词: microRNA;疾病;关联分析;二分网络投影;聚类
中图法分类号: TP18

中文引用格式: 郭茂祖,王诗鸣,刘晓燕,田侦.miRNA 与疾病关联关系预测算法.软件学报,2017,28(11):3094−3102. http://www.jos.org.cn/1000-9825/5351.htm

英文引用格式: Guo MZ, Wang SM, Liu XY, Tian Z. Algorithm for predicting the associations between MiRNAs and diseases. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):3094−3102 (in Chinese). http://www.jos.org.cn/1000-9825/5351.htm

Algorithm for Predicting the Associations Between MiRNAs and Diseases

GUO Mao-Zu[1], WANG Shi-Ming[2], LIU Xiao-Yan[2], TIAN Zhen[2]

[1](School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: MicroRNAs (miRNAs) play an important role in the process of life. In recent years, predicting the associations between miRNAs and diseases has become a hot topic in research. Existing computational methods can be mainly divided into two categories: methods based on similarity measurement, and methods based on machine learning. The former approaches predict miRNA-disease associations by measuring similarity of nodes in the biological networks, but they need to build high quality biological networks. The latter approaches apply machine learning algorithms to this problem, but they need to build a negative collection of high credibility. To address those shortcomings, this paper presents a novel computational model called BNPDCMDA (bipartite network projection based on density clustering to predict miRNA-disease associations) to predict miRNAs-disease associations. First, a miRNA-disease double-layer network model is constructed. Then, similarity of miRNAs is used to perform density clustering. Next, bipartite network projection is

* 基金项目: 国家自然科学基金(61571163, 61532014, 61671189, 61402132); 国家重点基础研究发展计划(973)(2016YFC0901902)

Foundation item: National Natural Science Foundation of China (61571163, 61532014, 61671189, 61402132); National Program on Key Basic Research Project of China (973) (2016YFC0901902)

本文由复杂环境下的机器学习研究专刊特约编辑胡清华教授、张道强教授、张长水教授推荐.

收稿时间: 2017-05-15; 修改时间: 2017-06-16; 采用时间: 2017-08-23

applied to miRNA-disease double-layer composed of density clustered miRNAs and disease sets. Finally, predictions for miRNA-disease association are performed. Further experimental results show that the proposed approach achieves AUC of 99.08% by using the leave-one-out cross-validation test, which demonstrates better predictive performance of BNPDCMDA than other methods. Moreover, certain miRNAs associated common diseases are predicted by BNPDCMDA.

Key words: microRNAs; disease; association analysis; bipartite network projection; clustering

MicroRNAs(miRNAs)是一类长度约为 22 个核苷酸的内源性非编码 RNA,通过碱基配对与其靶向的 mRNA3′端非编码区结合,导致靶 mRNA 降解或翻译抑制,从而在转录后水平上调控基因表达[1−3].Lee 等人[4]于 1993 年发现第 1 个 miRNA,即存在于秀丽隐杆线虫中的 lin-4.研究结果表明[5,6],miRNAs 在免疫反应、转录、增殖分化等多种生物过程中起着重要作用.miRNAs 与其靶 mRNAs 的功能失调会导致各种疾病.因此,识别 miRNAs 与疾病的关联关系至关重要.早期研究采用生物学实验法,但实验周期漫长、成本高.因而,计算生物学方法分析和预测 miRNAs 与疾病关联问题成为当前的研究热点,其主要分为相似度度量和机器学习两类.

研究发现,功能相似的 miRNAs 调控的疾病也比较相似,反之亦然[7].2009 年,Jiang 等人[8]构建了功能相关 miRNA 网络和人类疾病表型-miRNA 网络,预测 miRNA-疾病关联,但构建网络所用信息有限.2010 年,Jiang 等人[9]用朴素贝叶斯模型融合多种来源的数据构建模型预测基因间的功能相关性,预测效果有所提高.2013 年,Xuan 等人[10]提出了基于加权最相似 k 近邻的方法 HDMP,但无法对未知相关 miRNAs 的疾病进行预测.Shi 等人[11]通过将疾病基因和 miRNA 靶基因映射到 PPI(protein-protein interaction,蛋白质-蛋白质互作)网络上预测 miRNA-疾病关联.Chen 等人[12]将随机游走算法应用到疾病相似网络,由于没有考虑 miRNA 功能相似性,该方法性能较低.Liu 等人[13]于 2016 年通过融合多种来源的数据构建一个异构网络(heterogeneous network),对其应用随机游走算法预测 miRNA-疾病关联关系.Shi 等人[14]在构建网络模型时又加入了蛋白质互作信息、基因本体信息、miRNA-靶基因调控信息等,使数据来源更为全面.此外,近期还涌现出了更多的方法,如 You 等人[15]提出的基于路径的计算模型 PBMDA 及 Li 等人[16]提出的迭代的矩阵计算模型 MCMDA 等,均取得了较好的结果.

机器学习方法通过训练分类模型预测 miRNA-疾病关联.2010 年,Xu 等人[17]从 miRNA-疾病网络中提取特征,训练 SVM 分类器预测 miRNA-疾病关联.2013 年,Jiang 等人[18]构建不同的特征集,包括 miRNA 信息特征集和疾病表型信息特征集,得到相近的结果.然而,这些方法都将未标记的样本当作反例集,显然不合理.因此,Chen 等人[19]于 2014 年提出了半监督全局化方法 RLSMDA,在没有反例集的情况下预测 miRNA-疾病关联,然而没有考虑 miRNA-疾病关联网络的拓扑信息.Zou 等人[20]于 2016 年提出了基于社会网络分析的预测方法 KATZ 和 CATAPULT,前者在已知 miRNA-疾病关联、miRNA-miRNA 功能相似度、疾病-疾病相似度的前提下,将社会网络分析方法与机器学习相结合来预测 miRNA-疾病关联;后者是一种监督的机器学习方法,用引导聚合算法(bootstrap aggregating algorithm)训练带偏置的 SVM 分类器来预测 miRNA-疾病关联.

功能相似的 miRNAs 调控的疾病也比较相似,基于这一假设,本文提出了 BNPDCMDA 算法.首先构建 miRNA-疾病双层网络;其次,用 miRNA 的功能相似度对其进行基于密度的聚类;最后,将聚类后的 miRNA 与疾病构成的 miRNA-疾病双层子网应用二分网络投影预测 miRNA-疾病关联.本文第 1 节描述算法所涉及的相关概念.第 2 节介绍 BNPDCMDA 算法.第 3 节介绍实验过程及结果分析.第 4 节总结本文工作.

1 二分网络投影介绍

Zhou 等人[21]于 2007 年提出了二分网络投影法.Sun 等人[22]在此基础上提出了基于网络拓扑相似度的改进方法 NTSMDA.对于二分网络 $G(M,D,E)$,节点包括 miRNAs 集合 $M=\{m_1,m_2,\ldots,m_p\}$和疾病集合 $D=\{d_1,d_2,\ldots,d_q\}$,E 为边集,边只存在于集合 M 与 D 间,不存在于 M 或 D 内.选定待预测疾病 d_i 作为种子节点,二分网络投影以所有 miRNAs 与种子的关联信息构成的 p 维二进制向量作为输入,以所有 miRNAs 与种子的关联强度值构成的 p 维向量作为输出,过程分为两步.

- 第 1 步,资源从 M 流向 D,疾病 d_j 接收到的资源 $f(d_j)$见公式(1).

$$f(d_j)=\sum_{i=1}^{p}\frac{a_{ij}f(m_i)}{k(m_i)} \tag{1}$$

- 第 2 步:资源从 D 流回 M,miRNA m_i 接收到的资源 $f'(m_i)$见公式(2).

$$f'(m_i)=\sum_{j=1}^{q}\frac{a_{ij}f(d_j)}{k(d_j)} \tag{2}$$

其中,$f(x)$为节点 x 的初始资源;$k(x)$表示节点 x 的度;a_{ij} 是邻接矩阵的第 i 行第 j 列值,见公式(3).

$$a_{ij}=\begin{cases}1, & m_i d_j \in E\\ 0, & \text{if not}\end{cases} \tag{3}$$

向量$\{f'(m_1),f'(m_2),\ldots,f'(m_p)\}$为输出向量,$f'(m_i)$代表 miRNA m_i 与种子的关联强度.

2 BNPDCMDA 算法

2.1 算法描述

我们提出一种基于密度聚类的二分网络投影算法 BNPDCMDA 来预测 miRNA-疾病关联,过程如图 1 所示.

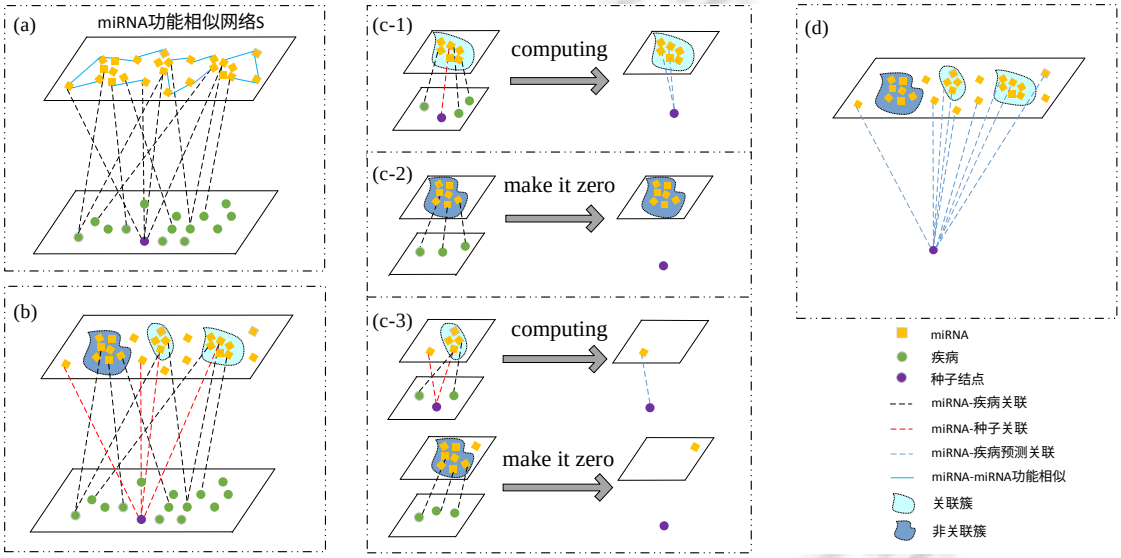

Fig.1 Illustration of the process of the algorithm

图 1 算法过程描述

(1) 构建 miRNA-疾病双层网络,包括 miRNA 功能相似网络 S、miRNA-疾病关联网络 A,其中,miRNA 集合 $M=\{m_1,m_2,\ldots,m_p\}$,疾病集合 $D=\{d_1,d_2,\ldots,d_q\}$,选择待预测疾病作为种子节点,如图 1(a)所示.

(2) 通过 miRNA m_i 与 m_j 的功能相似度 $S[i][j]$计算其距离 $Distance_{ij}$,见公式(4).

$$Distance_{ij}=1-S[i][j] \tag{4}$$

用 DBSCAN 算法对 miRNAs 做基于密度的聚类,得到 n 个簇 $C_1,C_2,\ldots,C_n$ 及若干孤立点,若簇中有 miRNA 与种子关联,则定义该簇为种子的关联簇;若簇中没有 miRNA 与种子关联,则定义该簇为种子的非关联簇,如图 1(b)所示.

(3) 对于种子的每个关联簇,簇中所有 miRNAs 构成一个 miRNA 功能相似子网,子网中,miRNAs 关联的疾病构成一个疾病子集合,从 miRNA-疾病双层网络中获取 miRNA 功能相似子网与疾病子集合构成的 miRNA-疾病双层子网,对该双层子网应用二分网络投影得到种子与关联簇中每个 miRNA 的关联强度,如图 1(c-1)所示;对于种子的每个非关联簇,将种子与该簇内 miRNAs 的关联强度置 0,如图 1(c-2)所示;对于每个孤立点 m_i,选择与其相似度均值最大的簇,并与该簇构成一个 miRNA 子集合 M_Temp_i,见公式(5).

$$M\_Temp_i = \underset{C_k \in \{C_1, C_2, ..., C_n\}}{\arg\max} \left(\sum_{m_j \in C_k} S[i][j] / |C_k| \right) \cup \{m_i\} \quad (5)$$

若 M_Temp_i 关联的疾病子集合中不包含种子,则将种子与 m_i 的关联强度置 0;若包含,则对 M_Temp_i 与疾病子集合构成的 miRNA-疾病双层子网应用二分网络投影得到种子与 m_i 的关联强度,如图 1(c-3)所示.

(4) 整合种子与所有簇及孤立点的关联强度值,设置阈值,关联强度大于阈值的 miRNA 与种子有关联关系,如图 1(d)所示.

2.2 算法伪代码

算法. BNPDCMDA.

输入:miRNA 集合 $M=\{m_1,m_2,\ldots,m_p\}$,疾病集合 $D=\{d_1,d_2,\ldots,d_q\}$,miRNA 相似度矩阵 S,miRNA-疾病邻接矩阵 A,簇内最少数目 *MinPts*,聚类半径 *Eps*,种子节点 *d_seed*,阈值 *cutoff*.

输出:所有 miRNA 与 *d_seed* 关联的 0-1 向量 *Asso*.

1. 定义矩阵 *Distance* 表示 miRNA 间的距离,*Distance*[*i*][*j*]=1–*S*[*i*][*j*];
2. 用 DBSCAN 算法对 miRNAs 聚类,得到 n 个簇 $C_1,C_2,\ldots,C_n$ 及若干孤立点;
3. 关联簇=*d_seed* 与簇中某 miRNA 有关联的簇;
4. 非关联簇=*d_seed* 与簇中任何 miRNA 都不关联的簇;
5. For 每一个关联簇
6. 集合 *D_Temp*=该簇关联的所有疾病;
7. 对该簇及 *D_Temp* 构成的双层子网做二分网络投影,得到 *d_seed* 与簇中 miRNAs 的关联强度;
8. End For
9. For 每一个非关联簇
10. *d_seed* 与簇中 miRNAs 关联强度为 0;
11. End For
12. For 每一个孤立点 m_i
13. 集合 $M\_Temp_i = \underset{C_k \in \{C_1, C_2, ..., C_n\}}{\arg\max} \left(\sum_{m_j \in C_k} S[i][j] / |C_k| \right) \cup \{m_i\}$
14. 集合 $D\_Temp=M\_Temp_i$ 关联的所有疾病;
15. IF *d_seed*∈*D_Temp*
16. 对 M_Temp_i 及 *D_Temp* 构成的双层子网做二分网络投影,得到 *d_seed* 与 m_i 的关联强度;
17. Else *d_seed* 与 m_i 的关联强度为 0;
18. End For
19. For i=0 to p
20. IFmiRNA m_i 与 *d_seed* 的关联强度大于 *cutoff*
21. *Asso*[*i*]=1;
22. Else *Asso*[*i*]=0;
23. End For
24. Return *Asso*;

3 实验相关及结果分析

3.1 实验数据介绍

实验用到人类 miRNA-疾病数据库 HMDD[23]及 miRNA 功能相似得分数据库 MISIM[24],前者收录经实验验证的 miRNA-疾病关联关系,后者提供 miRNAs 之间的功能相似得分.两者被广泛应用于 miRNA-疾病关联网络模型的构建[12−16,25].对于在 HMDD 中出现而在 MISIM 中没有出现的 miRNAs,本文将其在 HMDD 中的关联对

筛选掉,最终得到 5 526 个关联对用于构建 miRNA-疾病关联网络,并作为实验的标准数据集,包括 296 个 miRNAs 和 375 个疾病,网络特征见表 1.

Table 1 Global characteristic of the miRNA-disease association network

表 1 miRNA-疾病关联网络全局特征

miRNAs 个数	疾病个数	关联关系个数	miRNA 度均值	疾病度均值	miRNA 度最大值	miRNA 度最小值	疾病度最大值	疾病度最小值
296	375	5 526	18.669	14.736	124	1	180	1

3.2 参数对实验结果的影响

用 DBSCAN 算法对 miRNAs 聚类,*MinPts* 与 *Eps* 不同,留一交叉验证结果也不同.分别取 *MinPts* 为 2,3,*Eps* 为 0.27,0.3,0.315,0.33,获得 8 组聚类结果,对应 8 个 AUC 值,见表 2.分析可知:当 *MinPts* 相同时,*Eps* 越小,簇内平均点数越少,AUC 值就越大;当 *Eps* 相同时,*MinPts* 越小,簇内平均点数越少,AUC 值就越大.这是因为 *MinPts* 越小,代表可形成的簇个数越多;*Eps* 越小,代表簇内 miRNA 相似度越高.故选择较小的 *MinPts* 与 *Eps*,会使形成的簇有簇内相似度更高、簇间 miRNA 相似度更低的特点.根据功能相似的 miRNA 调节功能相似的疾病这一假设,对这样的聚类结果应用 BNPDCMDA 预测 miRNA-疾病关联,会使构建的 miRNA-疾病双层子网的资源流失的程度最小,从而最大程度地在双层子网内部流动,故预测结果更好,AUC 值更高.

Table 2 Effect of *Eps* and *Minpts* on AUC value

表 2 *Eps*,*Minpts* 对 AUC 值的影响

Minpts	*Eps*	簇个数	孤立点	簇内平均点数	AUC 值
2	0.270	20	135	8.05	0.990 819
	0.300	21	111	8.95	0.989 838
	0.315	17	99	11.59	0.988 478
	0.330	12	88	17.33	0.978 552
3	0.270	13	159	10.54	0.989 190
	0.300	13	138	12.15	0.987 821
	0.315	10	128	16.80	0.984 491
	0.330	9	107	21.00	0.977 853

3.3 与其他方法的比较

采用留一交叉验证法验证结果,在标准数据集中,每次取一个关联对为测试对象,其余的作为 BNPDCMDA 算法的输入,用于验证算法的性能.取 *Minpts*=3,*Eps*=0.33,将 BNPDCMDA 与 NTSMDA[22],RLSMDA[19]和 Chen 的方法[12]进行比较,方法性能用 ROC 曲线描述,横坐标为假阳性率,纵坐标为真阳性率,如图 2 所示.

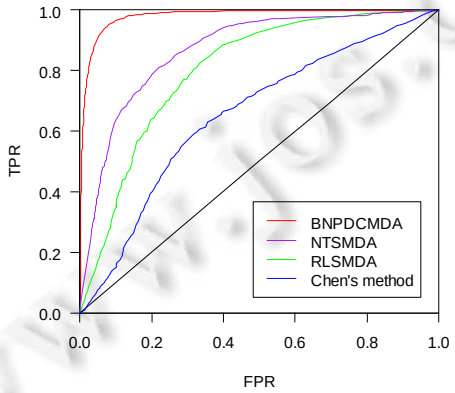

Fig.2 ROC curve

图 2 ROC 曲线

BNPDCMDA 的 AUC 为 97.79%,NTSMDA,RLSMDA,Chen 的方法的 AUC 分别为 86.82%,80.15%,65.25%.除此之外,我们将较为常见的 15 种疾病的预测结果采用留一交叉法验证,并与其他 3 种方法比较,见表 3.

Table 3 Comparison of AUC values on 15 common diseases by BNPDCMDA and the other methods

表 3 BNPDCMDA 和其他方法对 15 种常见疾病的 AUC 值的比较

疾病	关联 miRNAs 个数	AUC			
		BNPDCMDA	NTSMDA	RLSMDA	Chen's method
Carcinoma, hepatocellular	180	0.979 210	0.794 265	0.642 514	0.542 302
Breast neoplasms	179	0.983 082	0.886 653	0.763 529	0.629 856
Stomach neoplasms	161	0.949 696	0.838 562	0.706 574	0.616 371
Colorectal neoplasms	140	0.975 919	0.829 257	0.625 549	0.529 631
Lung neoplasms	139	0.976 362	0.882 935	0.786 392	0.537 713
Melanoma	126	0.976 615	0.889 856	0.836 859	0.625 514
Ovarian neoplasms	117	0.968 336	0.846 958	0.818 635	0.645 817
Prostatic neoplasms	111	0.963 349	0.851 663	0.760 210	0.685 249
Pancreatic neoplasms	102	0.976 913	0.825 214	0.725 278	0.645 214
Glioblastoma	98	0.980 420	0.809 638	0.714 425	0.639 475
Carcinoma, renal cell	93	0.965 307	0.781 456	0.706 395	0.593 214
Urinary bladder neoplasms	82	0.988 659	0.778 825	0.695 521	0.574 128
Hiv	11	0.987 962	0.829 963	0.840 663	0.584 726
Acute coronary syndrome	8	0.996 006	0.774 258	0.823 541	0.752 510
Heart diseases	4	0.957 112	0.946 314	0.925 146	0.835 219

分析可知,无论所选疾病关联 miRNA 数多或少,BNPDCMDA 均能取得较高的 AUC 值.BNPDCMDA 对于这 15 种疾病的预测取得的 AUC 均值为 97.50%,NTSMDA,RLSMDA,Chen 的方法分别为 83.77%,75.81%,62.91%.可见,BNPDCMDA 效果较好.

3.4 预测新**miRNA**-疾病关联关系

为了展现 BNPDCMDA 对新关联的预测功能,我们选择乳腺癌和肺癌这两种常见疾病,分别在不同来源的 3 个数据库 dbDEMC2.0[26],PhenomiR2.0[27]和 miRCancer[28]中验证 BNPDCMDA 预测出的关联关系,这些关联关系都是标准数据集中不存在的.

dbDEMC 数据库通过获取微阵列数据计算 miRNA 的差异表达值,从而得到与癌症相关的 miRNAs,dbDEMC2.0 对其进行更新,加入更多通过表达数据得到的与癌症相关的 miRNAs;PhenomiR2.0 数据库通过系统研究 miRNA 在疾病和生物过程中的异常调节(deregulation),人工收集 miRNAs 与疾病的关联关系;miRCancer 在收集数据的过程中使用文本挖掘的方法,提供综合的 miRNA 表达数据集,主要涉及人类的癌症数据.

结果表明,乳腺癌的潜在关联中得分大于 0.25 的 miRNA 共有 25 个,这 25 对关联均可在以上数据库中得到证实,见表 4.肺癌的潜在关联中得分大于 0.25 的 miRNA 共有 33 个,这 33 对关联亦均可得到证实,见表 5.取排名前 20 的潜在关联对进行验证,见表 6,其中有 14 对关联真实存在,有 6 对目前暂未获得相关文献或数据库的支持.

Table 4 Validation results of the potential associations of breast neoplasmsin databases

表 4 乳腺癌的潜在关联在数据库中的验证结果

排名	miRNAs	数据库	排名	miRNAs	数据库
1	hsa-mir-19b	dbDEMC2.0 PhenomiR2.0	14	hsa-mir-212	dbDEMC2.0 PhenomiR2.0 miRCancer
2	hsa-mir-106a	dbDEMC2.0 PhenomiR2.0	15	hsa-mir-520e	dbDEMC2.0 PhenomiR2.0 miRCancer
3	hsa-mir-142	miRCancer PhenomiR2.0	16	hsa-mir-181d	dbDEMC2.0 PhenomiR2.0
4	hsa-mir-150	dbDEMC2.0 PhenomiR2.0 miRCancer	17	hsa-mir-433	dbDEMC2.0 PhenomiR2.0
5	hsa-mir-99a	dbDEMC2.0 PhenomiR2.0 miRCancer	18	hsa-mir-130a	dbDEMC2.0 PhenomiR2.0 miRCancer
6	hsa-mir-185	dbDEMC2.0 PhenomiR2.0 miRCancer	19	hsa-mir-184	dbDEMC2.0 PhenomiR2.0
7	hsa-mir-92b	dbDEMC2.0	20	hsa-mir-219	dbDEMC2.0 PhenomiR2.0
8	hsa-mir-98	miRCancer PhenomiR2.0	21	hsa-mir-95	dbDEMC2.0 PhenomiR2.0
9	hsa-mir-30e	miRCancer PhenomiR2.0	22	hsa-mir-153	dbDEMC2.0 PhenomiR2.0 miRCancer
10	hsa-mir-196b	dbDEMC2.0 PhenomiR2.0	23	hsa-mir-449a	dbDEMC2.0 PhenomiR2.0 miRCancer
11	hsa-mir-15b	dbDEMC2.0 PhenomiR2.0	24	hsa-mir-331	PhenomiR2.0
12	hsa-mir-186	dbDEMC2.0 PhenomiR2.0	25	hsa-mir-144	miRCancer PhenomiR2.0
13	hsa-mir-32	dbDEMC2.0 PhenomiR2.0	–	–	–

Table 5 Validation results of the potential associations of lungneoplasmsin databases

表 5 肺癌的潜在关联在数据库中的验证结果

排名	miRNAs	数据库	排名	miRNAs	数据库
1	hsa-mir-92a	dbDEMC2.0 PhenomiR2.0 miRCancer	18	hsa-mir-302c	dbDEMC2.0 PhenomiR2.0
2	hsa-mir-199a	PhenomiR2.0	19	hsa-mir-194	dbDEMC2.0 PhenomiR2.0 miRCancer
3	hsa-mir-106b	PhenomiR2.0	20	hsa-mir-149	dbDEMC2.0 PhenomiR2.0
4	hsa-mir-19b	PhenomiR2.0	21	hsa-mir-129	dbDEMC2.0 PhenomiR2.0 miRCancer
5	hsa-mir-141	dbDEMC2.0 PhenomiR2.0 miRCancer	22	hsa-mir-302a	dbDEMC2.0 PhenomiR2.0
6	hsa-mir-218	miRCancer PhenomiR2.0	23	hsa-mir-99a	dbDEMC2.0 PhenomiR2.0 miRCancer
7	hsa-mir-133a	dbDEMC2.0 PhenomiR2.0 miRCancer	24	hsa-mir-92b	dbDEMC2.0 PhenomiR2.0 miRCancer
8	hsa-mir-16	miRCancer PhenomiR2.0	25	hsa-mir-423	PhenomiR2.0
9	hsa-mir-429	dbDEMC 2.0 miRCancer	26	hsa-mir-367	dbDEMC2.0 PhenomiR2.0
10	hsa-mir-20b	dbDEMC2.0 PhenomiR2.0	27	hsa-mir-409	miRCancer PhenomiR2.0
11	hsa-mir-24	miRCancer PhenomiR2.0	28	hsa-mir-452	PhenomiR2.0
12	hsa-mir-296	PhenomiR2.0	29	hsa-mir-15b	miRCancer PhenomiR2.0
13	hsa-mir-204	dbDEMC2.0 PhenomiR2.0 miRCancer	30	hsa-mir-196b	dbDEMC2.0 PhenomiR2.0
14	hsa-mir-195	miRCancer PhenomiR2.0	31	hsa-mir-302d	dbDEMC2.0 PhenomiR2.0
15	hsa-mir-122	dbDEMC2.0 PhenomiR2.0	32	hsa-mir-520d	dbDEMC2.0
16	hsa-mir-302b	dbDEMC2.0 PhenomiR2.0 miRCancer	33	hsa-mir-193b	dbDEMC2.0 PhenomiR2.0
17	hsa-mir-125b	miRCancer PhenomiR2.0	–	–	–

Table 6 Validation results of the potential associations in top 20 in databases

表 6 排在前 20 名的潜在关联在数据库中的验证结果

排名	miRNA	疾病	数据库
1	hsa-mir-21	Lupus vulgaris	–
2	hsa-mir-31	Heart failure	–
3	hsa-mir-16	Salivary gland neoplasms	–
4	hsa-mir-92a	Melanoma	PhenomiR2.0
5	hsa-mir-92a	Stomach neoplasms	–
6	hsa-mir-17	Carcinoma, renal cell	–
7	hsa-mir-143	Carcinoma, hepatocellular	miRCancer PhenomiR2.0
8	hsa-miR-20a	Carcinoma, renal cell	dbDEMC2.0
9	hsa-mir-143	Ovarian neoplasms	PhenomiR2.0
10	hsa-mir-125b	Muscular disorders, atrophic	–
11	hsa-mir-21	Adrenocortical carcinoma	miRCancer
12	hsa-mir-200b	Breast neoplasms	dbDEMC2.0 PhenomiR2.0 miRCancer
13	hsa-mir-92a	Lung neoplasms	dbDEMC2.0 PhenomiR2.0 miRCancer
14	hsa-mir-210	Ovarian neoplasms	miRCancer PhenomiR2.0
15	hsa-let-7b	Stomach neoplasms	miRCancer
16	hsa-mir-210	Stomach neoplasms	miRCancer PhenomiR2.0
17	hsa-mir-199a	Lung neoplasms	PhenomiR2.0
18	hsa-mir-29a	Pancreatic neoplasms	miRCancer PhenomiR2.0
19	hsa-mir-125b	Muscular disorders, atrophic	PhenomiR2.0
20	hsa-mir-181a	Muscular disorders, atrophic	PhenomiR2.0

4 总 结

本文提出一种基于密度聚类的二分网络投影算法 BNPDCMDA 来预测 miRNA-疾病关联.实验结果表明, *Eps* 及 *MinPts* 越小,簇内平均点数越少,算法性能就越好.BNPDCMDA 方法得到的 AUC 明显高于 NTSMDA, RLSMDA 和 Chen 的方法,特别是对于一些常见的疾病,BNPDCMDA 的实验结果更加突出、有效.此外,利用 BNPDCMD 算法对乳腺癌和肺癌这两种常见疾病关联的 miRNAs 成功地进行了预测.综上所述,BNPDCMDA 具有较好的性能.

References:

[1] Bartel DP. MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell, 2004,116(2):281–297. [doi: 10.1016/S0092-8674(04)00045-5]

[2] Groβhans H, Chatterjee S. MicroRNases and the regulated degradation of mature animal miRNAs. In: Proc. of the Regulation of microRNAs. Springer-Verlag, 2010. 140–155. [doi: 10.1007/978-1-4419-7823-3_12]

[3] Ambros V. The functions of animal microRNAs. Nature, 2004,431(7006):350−355. [doi: 10.1038/nature02871]

[4] Lee RC, Feinbaum RL, Ambros V. Elegans heterochronic gene lin-4 encodes small RNAS with antisense complementarity to lin-14. Cell, 1993,75(5):843−854. [doi: 10.1016/0092-8674(93)90529-Y]

[5] Karp X, Ambros V. Developmental biology. Encountering microRNAs in cell fate signaling. Science, 2005,310(5752):1288−1289. [doi: 10.1126/science.1121566]

[6] Xu P, Guo M, Hay BA. MicroRNAs and the regulation of cell death. Trends in Genetics, 2005,20(12):617−624. [doi: 10.1016/j.tig.2004.09.010]

[7] Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An analysis of human MicroRNA and disease associations. PLOS ONE, 2008,3(10):No.e3420. [doi: 10. 1371/journal.pone.0003420]

[8] Jiang Q, Hao Y, Wang G, Juan L, Zhang T, Teng M, Liu Y, Wang Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Systems Biology, 2010,4(Suppl 1):No.S2. [doi: 10.1186/1752-0509-4-S1-S2]

[9] Jiang Q, Wang G, Wang Y. An approach for prioritizing disease-related microRNAs based on genomic data integration. In: Proc. of the 2010 3rd Int'l Conf. on Biomedical Engineering and Informatics (BMEI). IEEE, 2010. 2270−2274. [doi: 10.1109/BMEI.2010.5639313]

[10] Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z, Huang Y. Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors. PLOS ONE, 2013,8(8):No.e70204. [doi: 10.1371/journal.pone.0070204]

[11] Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, Zhao Z, Jiang W, Guo Z, Li X. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Systems Biology, 2013,7(1):1−12. [doi: 10.1186/1752-0509-7-101]

[12] Chen H, Zhang Z. Prediction of associations between OMIM diseases and microRNAs by random walk on OMIM disease similarity network. Scientific World Journal, 2013,2013(10):No.204658. [doi: 10.1155/2013/204658]

[13] Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2016,14(4):905−915. [doi: 10.1109/TCBB.2016.2550432]

[14] Shi H, Zhang G, Zhou M, Cheng L, Yang H, Wang J, Sun J. Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations. PLOS ONE, 2016,11(2):No.e0148521. [doi: 10.1371/journal.pone.0148521]

[15] You ZH, Huang ZA, Zhu Z, Yan GY, Li ZW, Wen Z, Chen X. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. Plos Computational Biology, 2017,13(3):No.e1005455. [doi: 10.1371/journal.pcbi.1005455]

[16] Li JQ, Rong ZH, Chen X, Yan GY, You ZH. MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget, 2017,8(13):21187−21199. [doi: 10.18632/oncotarget.15061]

[17] Xu J, Li CX, Lü JY, Li YS, Xiao Y, Shao TT, Huo X, Li X, Zou Y, Han QL, Li X, Wang LH, Ren H. Prioritizing candidate disease miRNAs by topological features in the mirna target-dysregulated network: Case study of prostate cancer. Molecular Cancer Therapeutics, 2011,10(10):1857−1866. [doi: 10.1158/1535-7163.MCT-11-0055]

[18] Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. Int'l Journal of Data Mining and Bioinformatics, 2013,8(3):282−293.[doi: 10.1504/IJDMB.2013.056078]

[19] Chen X, Yan GY. Semi-Supervised learning for potential human microRNA-disease associations inference. Scientific Reports, 2014,4:No.5501. [doi:10.1038/srep05501]

[20] Zou Q, Li J, Hong Q, Lin Z, Wu Y, Shi H, Ju Y. Prediction of MicroRNA-disease associations based on social network analysis methods. Biomed Research Int'l, 2015,2015(10):No.810514. [doi: 10.1155/2015/810514]

[21] Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. Physical Review E Statistical Nonlinear and Soft Matter Physics, 2007,76(2):No.046115. [doi: 10.1103/PhysRevE.76.046115]

[22] Sun D, Li A, Feng H, Wang M. NTSMDA: Prediction of miRNA-disease associations by integrating network topological similarity. Molecular Biosystems, 2016,12(7):2224−2232. [doi: 10.1039/C6MB00049E]

[23] Li Y, Qiu C, Tu J, Geng B, Yang JC, Jiang TZ, Cui QH. HMDD v2.0: A database for experimentally supported human microRNA and disease associations. Nucleic Acids Research, 2014,42(Database Issue):1070−1074. [doi: 10.1093/nar/gkt1023]

[24] Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA- associated diseases. Bioinformatics, 2010,26(13):1644−1650. [doi: 10.1093/bioinformatics/btq241]

[25] Chen X, Jiang ZC, Xie D, Huang DS, Zhao Q, Yan GY, You ZH. A novel computational model based on super-disease and miRNA for potential miRNA-disease association prediction. Molecular Biosystems, 2017,13(6):1202−1212. [doi: 10.1039/C6MB00853D]

[26] Cui HL, Zhang YD, Ren F. dbDEMC2.0: A database of differentially expressed miRNAs in human cancers v2.0. China Journal of Modern Medicine, 2014,24(3):77−79.

[27] Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis F. PhenomiR: A knowledgebase for microRNA expression in diseases and biological processes. Genome Biology, 2010,11(1):1−11. [doi: 10.1186/gb-2010-11-1-r6]

[28] Xie B, Ding Q, Han H, Wu D. miRCancer: A microRNA-cancer association database constructed by text mining on literature. Bioinformatics, 2013,29(5):638−644. [doi: 10.1093/bioinformatics/btt014]

郭茂祖(1966－),男,山东夏津人,博士,教授,博士生导师,CCF 专业会员,主要研究领域为机器学习,生物信息学.

刘晓燕(1963－),女,博士,副研究员,主要研究领域为生物信息学,数据挖掘.

王诗鸣(1994－),女,博士生,主要研究领域为机器学习,生物信息学.

田侦(1987－),男,博士生,主要研究领域为机器学习,生物信息学.