DOI: 10.11992/tis.201805043
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180716.1134.006.html

一种预测 miRNA 与疾病关联关系的矩阵分解算法

刘晓燕[1], 陈希[1], 郭茂祖[1,2], 车凯[1], 王春宇[1]
（1. 哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001; 2. 北京建筑大学 电气与信息工程学院, 北京 100044）

摘　要: 越来越多的证据表明 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。然而, 现有的方法大多数是基于已知的 miRNA-疾病关联, 对没有任何关联信息的 miRNA 或疾病的效果是很不理想的。本文提出了一种矩阵分解的方法 LMFMDA(least squares optimization matrix factorization method for mirna-disease association) 对 miRNAs 和疾病的关联关系进行预测。LMFMDA 基于 miRNAs 相似度矩阵、疾病相似度矩阵和 miRNAs-疾病关联关系矩阵, 用迭代最小二乘法求解 miRNAs 和疾病的表达向量, 最终利用 miRNAs 和疾病的表达向量完成对 miRNA 与疾病关联关系的预测。与常规做法不同的是, 我们引入了辅助的 miRNAs 和疾病变量, 来保证在优化时能够收敛到最优解。实验结果表明,采用留一交叉验证法得到的 AUC 值可达 0.820 6, 明显优于当前其他方法,尤其在没有任何关联信息的 miRNA 和疾病上, LMFMDA 算法比最新的算法有了极大的提升。

关键词: microRNAs; 疾病; 关联预测; 矩阵分解; 迭代最小二乘
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)06−0897−08

中文引用格式: 刘晓燕, 陈希, 郭茂祖, 等. 一种预测 miRNA 与疾病关联关系的矩阵分解算法[J]. 智能系统学报, 2018, 13(6): 897–904.
英文引用格式: LIU Xiaoyan, CHEN Xi, GUO Maozu, et al. A matrix factorization method for predicting miRNA-disease association[J]. CAAI transactions on intelligent systems, 2018, 13(6): 897–904.

A matrix factorization method for predicting miRNA-disease association

LIU Xiaoyan[1], CHEN Xi[1], GUO Maozu[1,2], CHE Kai[1], WANG Chunyu[1]
(1. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; 2. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: There are increasing evidences that microRNAs (miRNAs) play an important role in life processes. In recent years, predicting the association between miRNAs and diseases has become an active topic. However, most of the existing methods are based on known miRNA-disease associations and are not ideal for miRNAs and diseases without any known associations. This paper presents a least squares optimization matrix factorization method for miRNA-disease association (LMFMDA) prediction. The LMFMDA, which is based on miRNAs similarity matrix, disease similarity matrix, and miRNAs-disease relationship, uses the iterative least squares method to solve the expression vectors of miRNAs and disease and approximates the existing associations between miRNAs and diseases by the expression vector of miRNA and disease. Different from the conventional approach, we introduce auxiliary miRNAs and disease variables to ensure that these variables converge to the optimal solution during optimization. The experiments show that the AUC obtained by applying the leave-one-out cross-validation method is 0.820 6, which is obviously better than other current methods. Especially in the miRNA and disease without any associated information, the LMFMDA algorithm significantly outperforms the latest algorithm.

Keywords: microRNAs; disease; association prediction; matrix factorization; iterative least squares

MicroRNAs(miRNAs) 是一类很小的内源性非编码 RNA, 长度约为 20 ~ 24 个核苷酸, 通过碱基配对与其靶向的 mRNA 的 3'端非编码区相结合,

收稿日期: 2018−05−27.　**网络出版日期:** 2018−07−17.
基金项目: 国家自然科学基金项目 (61671189, 61571163, 61532014, 91735306); 国家重点研发计划课题 (2016YFC0901902).
通信作者: 郭茂祖. E-mail: guomaozu@bucea.edu.cn.

导致靶 mRNA 的降解或翻译抑制，从而在转录后水平上调控基因表达[1-3]。越来越多的证据表明，miRNA 在免疫反应、转录、增殖、分化、信号传导和胚胎发育等[4-7]生物过程中起着重要的作用，miRNA 突变、miRNA 的生物合成和 miRNA 与其靶 mRNA 的功能失调可能会导致各种疾病。因此，识别 miRNA 与疾病之间的互作关系至关重要。早期研究采用生物学实验方法确定 miRNA 与特定疾病的关系，然而生物学实验方法实验周期长、成本高。因此计算生物学方法分析、预测 miRNAs 和疾病的关联问题成为了当前的研究热点。

1　相关工作

目前，miRNA 和疾病的关联预测主要分为基于网络拓扑结构的方法和机器学习的方法。

基于网络拓扑结构的研究方法建立在“功能相似的 miRNA 调控的疾病也比较相似，反之亦然[8-9]”这个假设基础上，文献[10-19]就此展开了一系列研究工作。2010 年，Jiang 等[10]首次提出一种计算方法，构建功能相关 miRNA 网络和人类疾病表型-miRNA 网络，将人类的 miRNA 组按照与疾病关联得分的大小排序，预测 miRNA 与疾病的关联。这是以前用基于网络的方法预测与疾病相关的编码蛋白基因的合理延伸。2010 年，Jiang 等[11]又提出一种基于基因组数据融合的新方法，用朴素贝叶斯模型融合多种来源的数据，构建一个模型预测基因之间的功能相关性。分别用两个向量表示疾病与基因之间的关联、miRNA 与靶基因之间的关联。对于给定的疾病，计算其与每个 miRNA 的相似得分，并从高到低排序，最高得分为与该疾病相关的 miRNA。Chen 等[12]将随机游走算法应用到 miRNA-miRNA 功能相似网络，在给定的种子结点处开始，将已知的关联关系的大小作为转移概率，模拟网络中当前结点扩散到其邻结点的过程，以此来挖掘网络中可能潜在的关联关系。Chen 等[13]在 2013 年又提出一种基于相似度的方法，分为 3 个策略：基于 miRNA 的相似度推断 (miRNA-based similarity inference, MBSI)、基于表型的相似度推断 (phenotype-based similarity inference, PBSI) 和基于网络一致性的推断 (network-consistency-based inference, NetCBI)；Shi 等[14]于 2013 年提出一种基于可重启的随机游走 (random walk with restart, RWR) 算法的新方法，将疾病基因和 miRNA 靶基因映射到蛋白质−蛋白质互作 (protein-protein interaction, PPI) 网络上，设置不同的种子应用 RWR 算法；Xuan 等[15]后又提出名为 HDMP 的方法——基于加权最相似 k 近邻的方法，预测与疾病相关的 miRNA；Xu 等[16]主要通过比对 miRNA 与 mRNA 表达谱融合多种疾病的表型关联，预测与癌症相关的 miRNA；2013 年，Mork 等[17]提出一种蛋白质介导的预测方法，通过 miRNA 与蛋白质之间的关联、蛋白质与疾病之间的关联预测 miRNA 与疾病之间的关系；2016 年，Sun 等[19]提出了基于已知的 miRNA-疾病网络拓扑相似性，以挖掘更多潜在的与疾病相关的 miRNA，利用二分投影的方法，来完成 miRNA 与疾病的关联预测工作。

到目前为止，基于网络拓扑结构的研究方法处理 miRNAs 和疾病的关联预测问题上，更多的倾向于基于已知的关联关系来挖掘其中潜在的关系，而对缺少已知关联信息的 miRNAs 和疾病，其结果往往呈现随机化。

在机器学习方法研究上。2012 年，Xu 等[20]首先使用机器学习方法预测 miRNA 与疾病之间的关系。这种方法旨在从大规模的反例中分辨出正例关联，核心是从 miRNA-疾病网络中提取特征，训练一个 SVM 分类器。2013 年，Jiang 等[21]又通过构建不同于 Xu 的特征集——一个关于 miRNA 信息的特征集和一个关于疾病表型信息的特征集，应用此方法得到相近的结果。2014 年，Chen 等[22]提出一种半监督的全局化方法 (regularized least squares for mirna-disease association, RLSMDA)，在没有负例集的情况下预测 miRNA 与疾病的关联。用正则化最小二乘法构建一个连续的分类函数，表示每个 miRNA 与给定疾病相关的概率，对于未知相关 miRNA 的疾病，该方法也适用。

基于机器学习的方法能够取得与“基于网络拓扑结构方法”相近或者更好结果，有的甚至很好地处理未知 miRNA 的疾病，例如 RLSMDA。而机器学习主要受制于 miRNAs 与疾病特征的表示，以及对如何处理有正样本数据的模型设计。

基于矩阵分解的算法用高维空间的向量解决了特征表示的问题，算法同时构建 miRNAs 和疾病在高维空间的表示，并以此为基础获得其关联关系，用迭代最小二乘法求解出最终的 miRNA-疾病关联关系的概率。这个求解思路来源于推荐系统中当前所流行的矩阵分解方法，对解决类似的关联关系预测问题在近年来也被证明非常有效。Shen[23]在 2017 年首次提出基于矩阵分解的方法对 miRNAs 和疾病的关联关系进行预测，并取得了比 Chen[22]更好的效果，但在其迭代求解的

过程中, 受到其损失函数的影响无法使用最小二乘法, 导致其每个变量都需要迭代求解, 这在同时要求多个变量迭代求解的情况下, 其结果很大程度上依赖于初始解的选择, 在很多的情况下甚至无法收敛, 算法的稳定性难以保证。

本文提出的 LMFMDA 算法, 首先构建 miRNAs 相似性网络、疾病相似性网络和 miRNA-疾病关联网络; 进而构建矩阵分解算法模型, 算法在利用迭代最小二乘法优化求解的过程中, 通过引入辅助 miRNAs 和疾病变量的方法, 提高计算速度, 解决收敛结果最优的问题, 确保算法的稳定性。

2　实验数据

在本节介绍 LMFMDA 算法所使用的数据和处理方法。数据来源如表 1 所示。

表 1　数据材料及其来源表题

Table 1　Data materials and the sources

数据库	描述	网址
MISIM	miRNA 相似性网络	http://www.mirbase.org/
MeSH	医学主题词表, 疾病描述	https://www.ncbi.nlm.nih.gov/mesh
HMDD	人类 miRNA-疾病关系	http://www.cuilab.cn/hmdd

2.1　miRNAs 功能相似度网络

直接从 MISIM 数据库获得 miRNAs 的功能相似度网络 **MS**, 网络中 miRNA 之间的相似度被表示为[0, 1]的实数。

2.2　疾病语义相似性网络

疾病的语义相似性通过 MeSH 得到, 计算方法来自 Wang[24], 假设疾病 t 是疾病 d 的一个祖先, 或者 d=t, 令:

$$C_d(t)=\begin{cases}1, & t=d\\ \max\{0.5\times C_d(t')|t'\in \text{childrenoft}\}, & t\neq d\end{cases} \tag{1}$$

疾病 1 和疾病 2 之间的语义相似性 $\mathbf{DS}(d_1, d_2)$ 即

$$\mathrm{DS}(d_1,d_2)=\frac{\sum_{t\in T(d_1)\cap T(d_2)}(C_{d_1}(t)+C_{d_2}(t))}{\sum_{t\in T(d_1)}C_{d_1}(t)+\sum_{t\in T(d_2)}C_{d_2}(t)} \tag{2}$$

2.3　miRNAs-疾病关联关系网络

在 HMDD 数据库下载了现有的 miRNAs-疾病关联关系网络。网络包含了 378 个疾病、571 个 miRNAs 及其构成的 10 381 个关联关系。关联矩阵 $\boldsymbol{R}$ 中, 如果 miRNA $m(i)$ 和疾病 $d(j)$ 被认为有关, 则 $\boldsymbol{R}(m(i)、d(j))$ 为 1, 否则, 为 0。

2.4　数据融合

将上述 3 个数据库的数据进行融合, 最终得到了重合的 446 个 miRNAs 和 322 个疾病, 和已经确认的 5 152 条 miRNAs-疾病关联关系。

在疾病上的分布如图 1 所示。

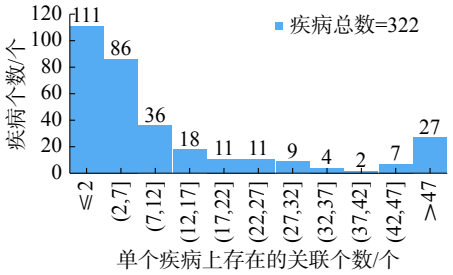

图 1　miRNAs-疾病关联关系在疾病中的分布图

Fig. 1　Distribution map of the miRNAs-disease association in diseases

在 miRNA 上的分布如图 2 所示。

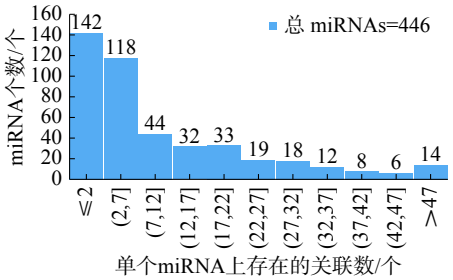

图 2　miRNAs-疾病关联关系在 miRNA 中的分布图

Fig. 2　Distribution map of the miRNAs-disease association in miRNAs

3　LMFMDA 算法模型

3.1　损失函数

本文中, 引入了矩阵分解的思想来解决 miRNAs-疾病关联关系预测问题。

首先, 通过整合 miRNAs 功能相似度网络和疾病语义相似性网络得到最终的 miRNAs 相似度矩阵 **MS** 和疾病相似度矩阵 **DS**, 以及已经被实验验证的 miRNAs-疾病关联网络 $\boldsymbol{R}$。

首先, 对每个 miRNA 和疾病, 给定它们在固定长度为 k 的维度空间的初始化投影向量, 并以其内积来表示 miRNAs 和疾病的关联关系, 可以用式 (3) 表示:

$$\boldsymbol{R}'=\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D} \tag{3}$$

式中: $\boldsymbol{M}$ 是由 m(本文中 m=446) 个 k 维列向量组成的 k 行 m 列的矩阵, 同样的, $\boldsymbol{D}$ 是 k 行 d 列 (本文中 d=322) 的矩阵。我们的目标即是通过求解合适的 $\boldsymbol{M}$ 和 $\boldsymbol{D}$ 来最小化 $\boldsymbol{R}'$和真实关系 $\boldsymbol{R}$ 的距离, 即

$$\min \lambda_1\left\|\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}-\mathbf{MS}\right\|_F^2+\lambda_2\left\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}-\mathbf{DS}\right\|_F^2 \tag{4}$$

考虑到这样的函数是二次的形式, 在迭代优

化时很难化简为不含有自身变量的等式，这会使得在迭代的过程中无法取得最优解，我们引入了辅助矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 来进行优化，式 (4) 可以变形为

$$\begin{aligned}\min \lambda_1\left\|\boldsymbol{M}^{\mathrm{T}}\boldsymbol{X}-\mathbf{MS}\right\|_F^2+\mu_1\|\boldsymbol{M}-\boldsymbol{X}\|_F^2+\\ \lambda_2\left\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Y}-\mathbf{DS}\right\|_F^2+\mu_2\|\boldsymbol{D}-\boldsymbol{Y}\|_F^2\end{aligned}\tag{5}$$

经验性地，我们对需要约束的 $\boldsymbol{M}$、$\boldsymbol{D}$ 加入二范数的约束，以防止模型陷入过拟合。最终的损失函数如式 (6) 所示：

$$\begin{aligned}\boldsymbol{L}=\left\|\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D}-\boldsymbol{R}\right\|_F^2+\lambda_0\left(\|\boldsymbol{M}\|_F^2+\|\boldsymbol{D}\|_F^2\right)+\\ \lambda_1\left\|\boldsymbol{M}^{\mathrm{T}}\boldsymbol{X}-\mathbf{MS}\right\|_F^2+\mu_1\|\boldsymbol{M}-\boldsymbol{X}\|_F^2+\\ \lambda_2\left\|\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Y}-\mathbf{DS}\right\|_F^2+\mu_2\|\boldsymbol{D}-\boldsymbol{Y}\|_F^2\end{aligned}\tag{6}$$

3.2　优化

我们采用迭代最小二乘的方式来优化这个问题，先固定 $\boldsymbol{D}$、$\boldsymbol{X}$、$\boldsymbol{Y}$，求解 $\boldsymbol{M}$。对 $\boldsymbol{M}$ 求导，有

$$\begin{aligned}\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{M}}=2\cdot\boldsymbol{D}\cdot\left(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D}-\boldsymbol{R}\right)^{\mathrm{T}}+2\cdot\lambda_0\cdot\boldsymbol{M}+\\ 2\cdot\lambda_1\cdot\boldsymbol{X}\cdot\left(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{X}-\mathbf{MS}\right)^{\mathrm{T}}+2\cdot\mu_1\left(\boldsymbol{M}-\boldsymbol{X}\right)=\\ 2\cdot\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{M}-2\cdot\boldsymbol{D}\boldsymbol{R}^{\mathrm{T}}+2\cdot\lambda_0\cdot\boldsymbol{M}+2\cdot\lambda_1\cdot\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}-\\ 2\cdot\lambda_1\cdot\boldsymbol{X}\cdot\mathbf{MS}^{\mathrm{T}}+2\cdot\mu_1\cdot\boldsymbol{M}-2\cdot\mu_1\cdot\boldsymbol{X}\end{aligned}\tag{7}$$

令 $\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{M}}=0$，有：

$$\begin{aligned}\boldsymbol{M}=\left(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}+(\lambda_0+\mu_1)\cdot\boldsymbol{I}_k+\lambda_1\cdot\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right)^{-1}\cdot\\ \left(\boldsymbol{D}\cdot\boldsymbol{R}^{\mathrm{T}}+\lambda_1\cdot\boldsymbol{X}\cdot\mathbf{MS}+\mu_1\cdot\boldsymbol{X}\right)\end{aligned}\tag{8}$$

同样，固定其他参数，分别求解 $\boldsymbol{D}$、$\boldsymbol{X}$、$\boldsymbol{Y}$，有：

$$\begin{aligned}\boldsymbol{D}=\left(\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}+(\lambda_0+\mu_2)\cdot\boldsymbol{I}_k+\lambda_2\cdot\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\right)^{-1}\\ \left(\boldsymbol{M}\cdot\boldsymbol{R}+\lambda_2\cdot\boldsymbol{Y}\cdot\mathbf{DS}+\mu_2\cdot\boldsymbol{Y}\right)\\ \boldsymbol{X}=\left(\lambda_1\cdot\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}+\mu_1\boldsymbol{I}_k\right)^{-1}\left(\lambda_1\cdot\boldsymbol{M}\cdot\mathbf{MS}+\mu_1\boldsymbol{M}\right)\\ \boldsymbol{Y}=\left(\lambda_2\cdot\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}+\mu_2\boldsymbol{I}_k\right)^{-1}\left(\lambda_2\cdot\boldsymbol{D}\cdot\mathbf{DS}+\mu_2\boldsymbol{D}\right)\end{aligned}\tag{9}$$

3.3　关联关系预测

利用得到的 $\boldsymbol{M}$ 和 $\boldsymbol{D}$，用其内积得到新的关联关系矩阵 $\boldsymbol{R}'=\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D}$，其 i 行 j 列即为第 i 个 miRNA 和第 j 个疾病的被预测的关联关系。事实上，它的值只有在和矩阵中其他值进行比较时才具有相对的意义，值越大表示关联关系出现的概率越大，但其与关联关系出现的概率并不完全等价。

3.4　算法框架

具体算法步骤如下：

1) 初始化 miRNAs 和疾病的向量矩阵 $\boldsymbol{M}$、$\boldsymbol{D}$，以及辅助向量 $\boldsymbol{X}$、$\boldsymbol{Y}$，并构建损失函数；

2) 用迭代最小二乘法求解 $\boldsymbol{M}$ 和 $\boldsymbol{D}$；

3) 根据 $\boldsymbol{M}$ 和 $\boldsymbol{D}$ 预测 miRNAs-疾病的关联关系。

$$\boldsymbol{R}'=\boldsymbol{M}^{\mathrm{T}}\boldsymbol{D}$$

算法框架如图 3 所示。

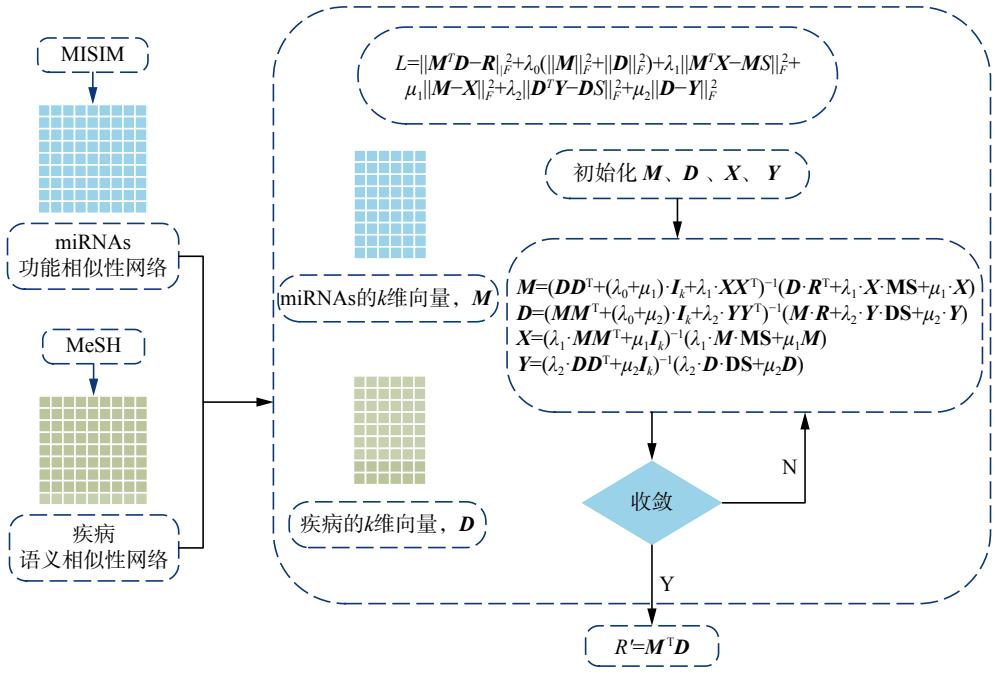

图 3　LMFMDA 算法模型框图

Fig. 3　The flow chat of LMFMDA algorithm model

3.5　复杂度分析

时间复杂度上，对于每次迭代，我们以 $\boldsymbol{M}$ 的求解为例，$\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$的时间复杂度为$O(k^2d)$，$(\lambda_0+\mu_1)\cdot\boldsymbol{I}_k$ 的时间复杂度为$O(k^2)$，$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$时间复杂度为$O(k^2m)$，这 3 项加法时间复杂度为$O(k^2)$，求逆复杂度为$O(k^3)$；$\boldsymbol{D}\boldsymbol{R}^{\mathrm{T}}$的时间复杂度为$O(kdm)$，$\lambda_1\cdot\boldsymbol{X}\cdot\boldsymbol{MS}$的时间复杂度为$O(km^2)$，$\mu_1\cdot X$时间复杂度为$O(m^2)$，这 3 项加法时间复杂度为$O(km)$；最后的乘法时间复杂度为$O(k^2m)$；综上，求解 $\boldsymbol{M}$ 的时间复杂度为$O(\max(k^2m,k^2d,kdm,km^2))$，事实上，通常有

$k \ll m, k \ll d$, 于是求解 $\boldsymbol{M}$ 的时间复杂度为 $O(\max(kdm, km^2))$。

同样地, 求解 $\boldsymbol{D}$、$\boldsymbol{X}$、$\boldsymbol{Y}$ 的时间复杂度分别为 $O(\max(kdm, kd^2))$, $O(m^3)$, $O(d^3)$。单次迭代的时间复杂度为 $O(\max(m^3, d^3))$。LMFMDA 的时间负责度即 $O(t\max(m^3, d^3))$, t 为迭代次数。

空间复杂度上, LMFMDA 要求 **MS**、**DS**、$\boldsymbol{R}$、$\boldsymbol{M}$、$\boldsymbol{D}$、$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的存储空间, 其空间复杂度为 $O(\max(m^2, d^2))$。

4　实验结果

实验采用留一交叉验证方式进行, 对每个关系, 将同一疾病下的未知关联视为负例, 当前关联视为正例, 最终得到的 AUC 作为评价结果。

4.1　实验参数

通过对不同的参数进行实验对比, 得到了以下参数组合: $k = 100$, $\lambda_0 = 6.0$, $\lambda_1 = 0.8$, $\lambda_2 = 0.8$, $\mu_1 = 3.0$, $\mu_2 = 3.0$。

miRNAs 与疾病的向量矩阵 $\boldsymbol{M}$ 与 $\boldsymbol{D}$ 初始化为取值在[0, 1]上的随机向量, $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 分别初始化为等同于 $\boldsymbol{M}$ 和 $\boldsymbol{D}$。

4.2　结果评价

在第 1 节得到的 446 个 miRNAs 和 322 个疾病上分别实验了 RWRMDA[13]、RLSMDA[22]、CMFMDA[23]以及本文提出的 LMFMDA 算法。实验结果如图 4 所示, LMFMDA 的效果明显好于其他 3 种方法。

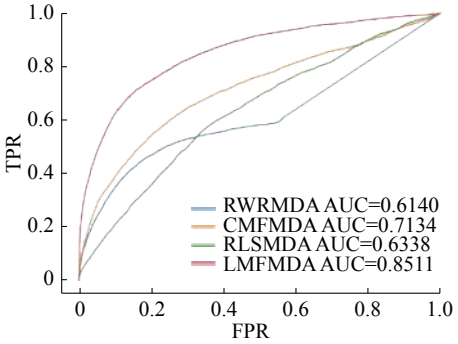

图 4　RWRMDA、CMFMDA、RLSMDA 和 LMFMDA 的 AUC 结果

Fig. 4　The AUC results of RWRMDA, CMFMDA, RLSMDA and LMFMDA

4.3　分析

我们分别记录了已知关联数>60 的 21 个疾病的实验结果 (见表 2、表 3), 以及已知关联数=1 的部分疾病的实验结果。已知关联数为 1 的疾病在进行留一法实验时, 会将唯一一个已知的关联 miRNA 抹去, 此时其已知关联数变为 0, 可以用于考察算法在新疾病中的应用效果。

表 2　高关联疾病在不同算法下的 AUC 结果

Table 2　The AUC results of high association diseases on different algorithm

疾病名称	关联个数	LMFMDA	RWRMDA	CMFMDA	RLSMDA
Carcinoma, Hepatocellular	209	0.770 559 224	0.802 276 903	0.590 942 755	0.567 805
Breast Neoplasms	188	0.830 346 921	0.827 897 392	0.707 031 51	0.581 754
Stomach Neoplasms	166	0.800 211 245	0.792 262 639	0.698 839 765	0.600 399
Colorectal Neoplasms	143	0.816 687 98	0.815 944 101	0.694 280 402	0.584 037
Melanoma	133	0.841 232 579	0.830 354 129	0.763 908 98	0.635 358
Lung Neoplasms	125	0.905 461 206	0.896 782 455	0.844 726 231	0.599 347
Heart Failure	118	0.808 275 684	0.807 732 613	0.635 453 525	0.572 331
Neoplasms	116	0.928 867 412	0.928 671 900	0.865 145 547	0.673 341
Ovarian Neoplasms	113	0.885 412 621	0.881 679 824	0.848 687 83	0.635 192
Prostatic Neoplasms	111	0.859 754 131	0.832 764 592	0.796 275 501	0.633 915
Carcinoma, Renal Cell	100	0.849 356847	0.829 757 803	0.775 121 532	0.615 241
Glioblastoma	99	0.832 863 611	0.836 865 732	0.740 363 199	0.598 405
Pancreatic Neoplasms	98	0.906 984 212	0.899 958 171	0.888 816 382	0.640 774
Carcinoma, Non-Small-Cell Lung	92	0.869 874 132	0.859 251 012	0.813 473 715	0.603 895
Urinary Bladder Neoplasms	89	0.853 023 601	0.834 512 166	0.830 017501	0.633 516
Colonic Neoplasms	82	0.866 789 314	0.865 347 844	0.808 090 055	0.642 180

续表 2

疾病名称	关联个数	LMFMDA	RWRMDA	CMFMDA	RLSMDA
Carcinoma, Squamous Cell	78	0.859 687413	0.5	0.833 772 833	0.596 178
Glioma	73	0.878 932 151	0.864 338 837	0.864 829 853	0.648 836
Esophageal Neoplasms	68	0.781 536 412	0.767 331 361	0.725 580 306	0.572 707
Leukemia, Myeloid, Acute	67	0.872 459 673	0.871 399 804	0.792 075 146	0.623 066
Head and Neck Neoplasms	63	0.847 238 105	0.5	0.836 495 898	0.665 183

表 3　新疾病在不同算法下的 AUC 结果

Table 3　The AUC results of new diseases on different algorithm

疾病名称	关联个数	LMFMDA	RWRMDA	CMFMDA	RLSMDA
Distal Myopathies	1	1	0.5	0.995 505 618	0.993 258
Hypopharyngeal Neoplasms	1	1	0.5	0.811 235 955	1
Hepatitis C, Chronic	1	1	0.5	1	1
Adenoma	1	1	0.5	1	1
Aortic Aneurysm, Abdominal	1	1	0.5	1	1
Carcinoma, Ductal, Breast	1	1	0.5	1	1
Colitis	1	1	0.5	0.970 786 517	0.997 753
Neuroma, Acoustic	1	1	0.5	0.146 067 416	1
Creutzfeldt-Jakob Syndrome	1	1	0.5	0.997 752 809	0.997 753
Eczema	1	1	0.5	0.912 359 551	0.997 753
Hepatitis B, Chronic	1	1	0.5	1	1
Hepatitis	1	1	0.5	1	1
Granulosa Cell Tumor	1	1	0.5	0.939 325 843	1
Graft vs Host Disease	1	1	0.5	1	1
Gerstmann-Straussler-Scheinker Disease	1	1	0.5	1	0.997 753
Gastritis, Atrophic	1	1	0.5	0.244 943 82	0.997 753
Encephalomyelitis, Autoimmune, Experimental	1	0.997752809	0.5	0.982 022 472	0.995 506
Moyamoya Disease	1	0.995505618	0.5	0.970 786 517	0.995 506
Cystic Fibrosis	1	0.995505618	0.5	0.013 483 146	0.997 753
Focal Epithelial Hyperplasia	1	0.995505618	0.5	0.224 719 101	1

可以看到，不论是在关联数较多的疾病或关联数极少的疾病上，LMFMDA 均表现出了优异的效果。

5　讨论

在提出 LMFMDA 的损失函数前，曾试图对每个 miRNA 和疾病标注一个先验关联值，作为第 k+1 维，也是不参与运算的常数维。即：

$$\boldsymbol{R}' = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{D} - \boldsymbol{e}m - \boldsymbol{e}\boldsymbol{d}^{\mathrm{T}}$$

然而其 k 值和 AUC 关联关系如图 5 所示。

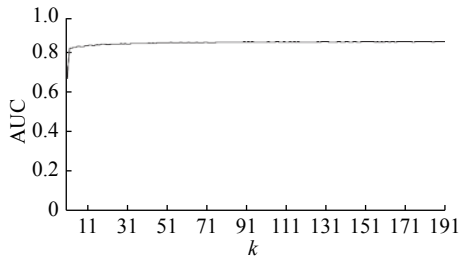

图 5　带常数维模型中 k 与 AUC 关系图

Fig. 5　The relation diagram of k and AUC in a model with constant dimensional

可以看到, 在 k>100 时, AUC 值基本趋于稳定。而对 k=100 维这样的子空间来说, 单独的常数维并不会对结果有很大的影响, 于是删除了假设的先验关联值, 最终确定了预测模型。

6 结论

本文基于矩阵分解和迭代最小二乘的方法(LMFMDA) 对 miRNAs 和疾病的关联关系进行预测。首先对 miRNAs 相似度矩阵、疾病相似度矩阵和 miRNAs-疾病关联关系进行数据融合, 采用迭代最小二乘法求解 miRNAs 和疾病的表达向量, 最后利用 miRNAs 和疾病的表达向量完成对 miRNA 与疾病关联关系的预测。同时, 通过引入辅助 miRNAs 和疾病变量的方法, 解决了收敛结果的最优问题。实验显示, LMFMDA 在高关联疾病和新疾病预测中相对于其他方法均取了较优的结果。

综上, 本文提出的 miRNA 与疾病关联预测算法 LMFMDA, 一方面可以处理未知相关 miRNAs 的疾病、或者未知相关疾病的 miRNAs; 另一方面, 实验结果也表明, LMFMDA 算法在 miRNAs 和疾病的关联关系预测上相较其他算法有更好的效果。

参考文献:

[1] WANG Qianghu, SUN Jie, ZHOU Meng, et al. A novel network-based method for measuring the functional relationship between gene sets[J]. Bioinformatics, 2011, 27(11): 1521–1528.

[2] LV Sali, LI Yan, WANG Qianghu, et al. A novel method to quantify gene set functional association based on gene ontology[J]. Journal of the royal society interface, 2012, 9(70): 1063–1072.

[3] HRISTOVSKI D, FRIEDMAN C, RINDFLESCH T C, et al. Exploiting semantic relations for literature-based discovery[J]. AMIA annual symposium proceedings, 2006, 2006: 349–353.

[4] KARP X, AMBROS V. Encountering microRNAs in cell fate signaling[J]. Science, 2005, 310(5752): 1288–1289.

[5] CHENG A M, BYROM M W, SHELTON J, et al. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis[J]. Nucleic acids research, 2005, 33(4): 1290–1297.

[6] MISKA E A. How microRNAs control cell division, differentiation and death[J]. Current opinion in genetics and development, 2005, 15(5): 563–568.

[7] XU Peizhang, GUO Ming, HAY B A. MicroRNAs and the regulation of cell death[J]. Trends in genetics, 2004, 20(12): 617–624.

[8] YOU Zhuhong, HUANG Zhian, ZHU Zexuan, et al. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction[J]. PLoS computational biology, 2017, 13(3): e1005455.

[9] SHI Hongbo, ZHANG Guangde, ZHOU Meng, et al. Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations[J]. PLoS one, 2016, 11(2): e0148521.

[10] JIANG Qinghua, HAO Yangyang, WANG Guohua, et al. Prioritization of disease microRNAs through a human phenome-microRNAome network[J]. BMC systems biology, 2010, 4(S1): S2.

[11] JIANG Qinghua, WANG Guohua, WANG Yadong. An approach for prioritizing disease-related microRNAs based on genomic data integration[C]//Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics. Yantai, China, 2010: 2270–2274.

[12] CHEN Xing, LIU Mingxi, YAN Guiying. RWRMDA: predicting novel human microRNA–disease associations[J]. Molecular biosystems, 2012, 8(10): 2792–2798.

[13] CHEN Hailin, ZHANG Zuping. Similarity-based methods for potential human microRNA-disease association prediction[J]. BMC medical genomics, 2013, 6: 12.

[14] SHI Hongbo, XU Juan, ZHANG Guangde, 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: 101.

[15] XUAN Ping, HAN Ke, GUO Maozu, et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors[J]. PLoS one, 2013, 8(8): e70204.

[16] XU Chaohan, PING Yanyan, LI Xiang, et al. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles[J]. Molecular biosystems, 2014, 10(11): 2800–2809.

[17] MØRK S, PLETSCHER-FRANKILD S, PALLEJA CARO A, et al. Protein-driven inference of miRNA–disease associations[J]. Bioinformatics, 2014, 30(3): 392–397.

[18] PASQUIER C, GARDÈS J. Prediction of miRNA-disease associations with a vector space model[J]. Scientific reports, 2016, 6: 27036.

[19] SUN Dongdong, LI Ao, FENG Huanqing, et al. NTSMDA: prediction of miRNA–disease associations by integ-

rating network topological similarity[J]. Molecular biosystems, 2016, 12(7): 2224–2232.

[20] LI Xia, XU Juan, LI Yongsheng. Prioritizing candidate disease miRNAs by topological features in the miRNA-target dysregulated network[M]//AZMI A S. Systems Biology in Cancer Research and Drug Discovery. Netherlands: Springer, 2012: 289–306.

[21] JIANG Qinghua, WANG Guohua, JIN Shuilin, 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.

[22] CHEN Xing, YAN Guiying. Semi-supervised learning for potential human microRNA-disease associations inference[J]. Scientific reports, 2014, 4: 5501.

[23] SHEN Zhen, ZHANG Youhua, HAN K, et al. miRNA-disease association prediction with collaborative matrix factorization[J]. Complexity, 2017, 2017: 2498957.

[24] WANG Dong, WANG Juan, LU Ming, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13): 1644–1650.

作者简介：

刘晓燕，女，1963 年生，副研究员，博士，主要研究方向为生物信息学、数据挖掘。

陈希，男，1995 年生，硕士研究生，主要研究方向为生物信息学。

郭茂祖，男，1966 年生，教授，博士生导师，博士，主要研究方向为机器学习、智慧城市、生物信息学。