计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100132

基于集成回归决策树的 lncRNA-疾病关联预测方法

任首朋[1]　李　劲[1]　王静茹[1]　岳　昆[2]
1 云南大学软件学院　昆明 650091
2 云南大学信息学院　昆明 650091
(supi2012212@qq.com)

摘　要　长链非编码 RNA(long non-coding RNA,lncRNA)在各种人类复杂疾病中起着重要作用。采用计算方法推断 lncRNA-疾病间的潜在关联关系不仅有助于理解疾病的致病机理,还有助于疾病诊断、预防和治疗。文中提出了一种基于集成回归决策树的 lncRNA-疾病关联预测方法。首先,利用已知的 lncRNA-疾病关联信息分别构建 lncRNA、疾病相似矩阵、lncRNA-疾病关联矩阵;其次,基于 lncRNA、疾病相似矩阵、lncRNA-疾病关联矩阵,从不同视角进一步构建 lncRNA、疾病特征向量;然后,使用主成分分析方法对 lncRNA、疾病特征进行特征提取;最后,使用回归决策树作为预测模型,并进一步采用集成学习的平均策略将多个决策树集成,从而获得最终的预测模型。留一交叉验证实验表明,该方法的预测结果优于现有方法,在 3 个真实的 lncRNA-疾病数据集上 *AUC* 值分别达到了 0.9055,0.8969 和 0.9129,与现有方法相比,分别提升了 6.46%,5.4%和 6.02%。此外,对乳腺癌、肺癌、胃癌 3 种疾病进行了案例分析,进一步验证了所提方法的准确性和有效性。

关键词:lncRNA-疾病;关联预测;特征提取;CART 决策树;集成学习
中图法分类号　TP391

Ensemble Regression Decision Trees-based lncRNA-disease Association Prediction

REN Shou-peng[1],LI Jin[1],WANG Jing-ru[1] and YUE Kun[2]
1 School of Software,Yunnan University,Kunming 650091,China
2 School of Information Science & Engineering,Yunnan University,Kunming 650091,China

Abstract　Long non-coding RNA (lncRNA) plays an important role in various complex human diseases. The development of effective prediction methods to infer the potential associations between lncRNA and diseases will not only help biologists understand the pathogenesis of diseases,but also contribute to the diagnosis,prevention,and treatment of human diseases. In this paper,an ensemble regression decision tree-based lncRNA-disease association method (ERDTLDA) is proposed to solve the lncRNA-disease association problem. First,ERDTLDA uses the open-source data of lncRNA to construct lncRNA,disease similarity matrix,lncRNA-disease association matrix respectively. Then,we obtain lncRNA,disease feature representations from these matrices. Principal component analysis is further exploited for feature extraction. Finally,a CART regression decision tree is used to yield association scores. An ensemble strategy for multiple decision trees is proposed to further improve the accuracy of our model. The results of LOOCV experiments show that the *AUC* of our method on three real lncRNA-disease datasets are 0.9055,0.8969 and 0.9129 respectively,which are 6.46%,5.4% and 6.02% higher than the existing methods,respectively. Additionally,breast cancer,lung cancer,and gastric cancer are also used as case studies to further verify the accuracy and effectiveness of ERDTLDA.

Keywords　lncRNA-disease,Association prediction,Feature extraction,CART decision tree,Ensemble learning

到稿日期:2020-11-18　返修日期:2021-05-30　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金云南联合基金项目(U1802271);云南省基础研究杰出青年项目(2019FJ011);云南省应用基础研究计划重点项目(201901BB050052)
This work was supported by the Foundation of National Natural Science Foundation of China United Yunnan Province(U1802271),Foundation of Outstanding Youth Project of Basic Research in Yunnan Province(2019FJ011) and Foundation of Key Project of Basic Research in Yunnan Province(201901BB050052).
通信作者:李劲(lijin@ynu.edu.cn)

1 引言

越来越多的研究表明，长链非编码 RNA(long non-coding RNA，lncRNA)不仅在许多生物进程中发挥着其特定的功能，对人类疾病也有重要影响[1]。因此，确定 lncRNA 与疾病之间的关联关系有助于了解疾病的发生原因、发展过程及其作用机制，进而为临床诊断、疾病预防、预后与个性化治疗提供帮助[2]。

目前，lncRNA 的相关研究刚刚起步，对 lncRNA 的功能及作用机制的理解还不够全面深入，使用生物学实验的方式研究 lncRNA 与疾病的互作用机制面临着实验周期较长、实验费用高、难以大规模进行操作等问题。近年来，随着 lncRNA 与疾病互作用生物学实验数据的积累，使得通过计算方法预测潜在 lncRNA-疾病关联关系成为可能。基于计算的 lncRNA-疾病关联关系预测可以帮助生物实验者进行有目的的筛选，提高实验效率，缩短实验耗费的时间，减少费用。因此，基于计算的 lncRNA-疾病关联预测方法的研究吸引了很多研究者的关注，产生了一系列富有成效的研究成果[3-14]。

目前，基于计算的 lncRNA-疾病关联关系预测方法主要分为两种[3]，即基于生物信息网络方法和基于机器学习方法。首先，lncRNA 与疾病关联关系的一个基本假设是[3]："功能相似的 lncRNA 所调控的疾病也是相似的，反之亦然。"基于 lncRNA 相似信息、疾病相似信息，以及已知的 lncRNA-疾病之间的关联信息可构建异构网络，基于异构网络也可进行有效的 lncRNA-疾病关联预测。文献[4]在相继构建了 lncRNA-疾病关联网络、疾病相似性网络和 lncRNA 功能相似网络后，RWRlncD 通过在 lncRNA 功能相似网络上进行随机游走重启来预测潜在的 lncRNA-疾病关系，然而此方法不能应用于没有任何已知相关的 lncRNA 和疾病预测。文献[5]提出了一种在全局网络上进行随机游走的计算方法 GrwLDA，该算法通过整合异质分子数据来预测潜在的 lncRNA-疾病关联。然而，该方法仍面临着参数选择困难的难题。随后，文献[6]提出了 Lap-BiRWRHLDA 模型，该模型将拉普拉斯标准化后的 lncRNA、疾病矩阵对应的网络进行整合，获得了一个异构网络，随后在该网络上使用双随机游走来实现 lncRNA 和疾病关联关系的预测。最近，文献[14]提出了一种基于 HeteSim 来预测 lncRNA-疾病关联关系的方法。该方法基于路径约束将元路径两端的节点随机游走到中间节点相遇的概率作为疾病与 lncRNA 之间关联关系的得分，以发掘潜在的疾病与 lncRNA 之间的关联关系。然而，与机器学习预测方法相比，该方法是一种基于异构网结构的预测方法，无法充分利用 lncRNA 或疾病本身的数据特征进行预测。

另外，随着已确定的 lncRNA 数量的持续增长，经生物学实验确定的 lncRNA 与疾病关联关系越来越多，如在 LncRNADisease[15] 和 Lnc2Cancer[16] 数据库中，收集了与普通疾病或癌症关联的 lncRNA 信息，使得通过数据驱动的机器学习方法来确定 lncRNA 与疾病关联关系成为可能。文献[9]提出了基于朴素贝叶斯分类器方法的计算方法，通过整合基因组、调控组和转录组数据，来预测潜在的与疾病相关的 lncRNA。但是该方法的模型训练需要负样本，该方法将数据集中未获得关联的 lncRNA-疾病对作为负样本，这将对模型的预测性能产生一定影响。文献[7]提出的 LRLSLDA 对 lncRNA-疾病关联使用拉普拉斯正则最小二乘，通过整合已知的 lncRNA-疾病关联与 lncRNA 表达谱来实现潜在的 lncRNA-疾病关联预测。在 LRLSLDA 的基础上，文献[8]将 lncRNA 之间的功能相似信息与 LRLSLDA 预测模型相结合，进一步提升了 lncRNA-疾病关联的预测准确率。

然而，上述两类方法均存在不足之处。具体而言，基于生物信息网络的预测方法虽然具有好的预测效果，但是其预测依赖于 lncRNA-疾病异构网，当网络结构发生变化时，需要重新进行随机游动，以获取新的关联评分。显然，现有方法对于孤立 lncRNA、孤立疾病或者新 lncRNA、新疾病无法进行有效处理。基于机器学习的方法的预测性能主要取决于 lncRNA、疾病的特征表示以及关联预测模型。现有方法直接选取 lncRNA 功能相似网邻接矩阵、疾病语义相似网邻接矩阵中的行(或列)分别作为 lncRNA、疾病特征向量，没有充分利用 lncRNA、疾病相似网蕴含的丰富的拓扑结构信息来生成特征表示。此外，基于机器学习的预测模型的预测准确率也有待进一步提高。

基于此，本文提出了一种新的基于集成回归决策树的 lncRNA-疾病关联预测方法(Ensemble Regression Decision Tree-based lncRNA-Disease Association，ERDTLDA)。首先，以 lncRNA 表达谱数据和已知的 lncRNA-疾病关联数据为基础，分别构建 lncRNA 功能相似矩阵、疾病语义相似矩阵。其次，基于 lncRNA 功能相似矩阵、疾病语义相似矩阵以及经生物学实验验证的 lncRNA-疾病关联数据构建 lncRNA-疾病异构网络，进而利用异构网络蕴含的丰富的拓扑结构语义信息，从多种不同视角分别构建 lncRNA 和疾病的特征表示。然后，利用主成分分析(PCA)分别对初始的 lncRNA、疾病特征向量进行特征提取，保留重要特征信息。最后将提取的特征表示输入到基于 CART 回归决策树的关联预测模型中，计算关联预测评分，从而获得预测结果。同时，为了降低数据噪声对模型训练的影响，本文提出了一个基于集成决策树的预测模型，进一步提高了模型的预测准确性。

为验证本文方法的有效性，在 3 个真实数据集上采用留一交叉验证法(LOOCV)和 5 折交叉验证法(5-FCV)对本文方法进行了验证。实验结果表明，本文方法在 3 个数据集上的预测性能均优于已有的代表性预测方法。同时，为进一步验证本文方法的可靠性和实用性，分别以乳腺癌、肺癌、胃癌 3 种疾病进行案例分析。案例结果表明，本文方法具有较高的关联预测准确度，是一种可靠的 lncRNA-疾病关联预测模型。

2 数据预处理

2.1 lncRNA-疾病关联矩阵

$L=\{l_1, l_2, \cdots, l_n\}$ 表示 lncRNA 集合，$D=\{d_1, d_2, \cdots, d_m\}$ 表示疾病集合。利用生物信息数据库中已知的 lncRNA-疾病关联关系构建 lncRNA-疾病关联矩阵 $\boldsymbol{Y}$，若疾病 d_i 与 lncRNA l_j 相关，则 $\boldsymbol{Y}(i,j)$ 为 1，否则为 0。

2.2 lncRNA 相似矩阵

lncRNA 相似矩阵由以下两种相似信息集成得到。首先，从文献[7,12]提供的链接下载了 lncRNA 表达水平和基因表达水平，其中包含了 22626 种在 22 类人体组织中的表达谱和 60245 种在 16 类人体组织中的基因表达水平。根据文献[7,12]中的方法，计算 l_i，l_j 的 Spearman 相关系数作为 lncRNA 表达相似性。令 $SPC(l_i,l_j)$ 表示 l_i，l_j 之间的表达谱相似性值。

然后，基于功能相似的 lncRNA 所调控的疾病也是相似的假设，从已知的 lncRNA-疾病关联数据中计算出 lncRNA 的高斯相互作用谱内核相似性[7,12]。具体计算方法如下：

$$KL(i,j)=\exp(-\gamma_l \| \boldsymbol{Y}(i,*)-\boldsymbol{Y}(j,*) \|^2) \tag{1}$$

其中，γ_l 控制核带宽参数，其计算式如下：

$$\gamma_l=\gamma_l' \Big/ \left(\frac{1}{n} \sum_{i=1}^{n} \| \boldsymbol{Y}(i,*) \|^2 \right) \tag{2}$$

与文献[7,12]一致，$\gamma_l'=1$。

按照式(3)组合 lncRNA 表达相似性和高斯相互作用谱内核相似性，获得 lncRNA 相似性。

$$SL(i,j)=\begin{cases} \lambda \cdot SPC(i,j)+(1-\lambda) \cdot KL(i,j), & i,j \in L_1 \\ KL(i,j), & \text{otherwise} \end{cases} \tag{3}$$

其中，λ 是权重系数，设为 1/2。

最后，通过拉普拉斯标准化得到 lncRNA 的拉普拉斯归一化矩阵，具体如式(4)、式(5)所示：

$$LL'(i,j)=\begin{cases} \dfrac{SL(i,j)}{\sqrt{\sum_i SL(i,j) \sum_j SL(i,j)}}, & SL(i,j) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$LL(i,j)=\begin{cases} \dfrac{LL'(i,j)}{\sum_j LL'(i,j)}, & SL(i,j) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

2.3 疾病相似矩阵

与 lncRNA 类似，疾病的高斯相互作用谱内核相似性构造方法如下：

$$KD(i,j)=\exp(-\gamma_d \| \boldsymbol{Y}(*,i)-\boldsymbol{Y}(*,j) \|^2) \tag{6}$$

其中，γ_l 控制核带宽参数，其计算式如下：

$$\gamma_d=\gamma_d' \Big/ \left(\frac{1}{m} \sum_{i=1}^{m} \| Y(*,i) \|^2 \right) \tag{7}$$

与文献[17]一致，$\gamma_d'=1$。

应用逻辑回归函数变换(logistic function transformation)来计算疾病的相似性，其计算式如下：

$$SD(i,j)=\frac{1}{1+\exp(c \cdot KD(i,j)+d)} \tag{8}$$

与文献[7,12]一致，$c=-15$，$d=\log(9999)$。与式(4)、式(5)类似，同理可计算疾病的拉普拉斯归一化矩阵。

3 构建 lncRNA、疾病特征向量

基于第 2 节介绍的 lncRNA、疾病数据，构建 3 种类型的 lncRNA、疾病特征表示，构建方法如表 1 所列。

表 1 lncRNA、疾病特征向量构造过程

Table 1 Constructions of lncRNA, disease feature representations

特征类型	名称	描述
1 类型特征 (lncRNA/disease)	n. obs	l_i 或 d_j 在关联矩阵 $\boldsymbol{Y}$ 中第 i 行或第 j 列的观测值的数量
	ave. sim	相似矩阵中 l_i 或 d_j 所在行或列的平均值
	hist. sim	相似矩阵中 l_i 或 d_j 相似值的直方图向量
	num. nb	相似矩阵对应图中节点的邻居数量
	k. sim	相似矩阵对应图中节点的 k 个最近邻居
2 类型特征 (lncRNA/disease)	k. ave. f	1 类型特征中 k-最近邻的平均值
	k. w. ave. f	1 类型特征中 k-最近邻权重的平均值
3 类型特征 (lncRNA-disease 对)	com	相似图的中介中心性、近邻中心性、特征向量中心性和 PageRank 值
	mf	矩阵分解关联矩阵 $\boldsymbol{Y}$ 获得 lncRNA/disease 潜在向量
	m. com	关联图的中介中心性、近邻中心性、特征向量中心性和 PageRank 值

首先，1)n. obs 定义了 l_i 或 d_j 在关联矩阵 $\boldsymbol{Y}$ 中第 i 行或第 j 列的观测值的数量，可以理解为某一 lncRNA 关联疾病的数量或某一疾病关联 lncRNA 的数量；2)ave. sim 定义了相似矩阵中 l_i 或 d_j 所在行或列的平均值，即某 lncRNA 关联所有疾病的平均值或某一疾病关联所有 lncRNA 的平均值；3)hist. sim 定义了相似矩阵中 l_i 或 d_j 相似值的直方图向量，即将 lncRNA相似矩阵和疾病相似矩阵中的数据按照数值大小分为 5 组，0～0.2 为一组，0.2～0.4 为一组，0.4～0.6 为一组，0.6～0.8 为一组，0.8～1.0 为一组，每组元素个数占总元素数量的比例构成的向量为一组直方图向量，对于 lncRNA 和疾病，我们分别从中获取其中的 5 组直方图向量。最后，将上述 3 种 1 类型特征按 lncRNA 和疾病进行拼接，构成 lncRNA 或疾病的 1 类型特征向量。

其次，为 lncRNA 相似性网络和疾病相似性网络中的每一个节点定义以下特征：1)num. nb 代表相似矩阵对应图中每个节点的邻居数量，这里我们取与 lncRNA 或疾病相似矩阵中元素值大于相似矩阵元素平均值的有关联的元素节点作为每个节点的邻居；2)k. sim 表示相似矩阵对应图中节点的 k 个最近邻居。本文取 $k=10$，即取每个节点邻居中排名前十的邻居构成特征向量；3)k. ave. f 表示 lncRNA 或疾病的 1 类型特征的 k-最近邻的平均值。本文取 $k=10$，具体过程为：从相似矩阵中取出某一行中元素值排名前十的 lncRNA 或疾病，对这前十个 lncRNA 或疾病的 1 类型的每个特征求平均值；4)k. w. ave. f 表示 lncRNA 或疾病的 1 类型特征的 k-最近邻的平均值权重，这里某个 1 类型特征的 k-最近邻的平均值权重被定义为排名前十的邻居对应元素值与相似矩阵中某行或某列排名前十的对应 lncRNA 或疾病的 1 类型特征做点乘后除以 10 获得的值；5)com 中包含了 lncRNA 和疾病相似网络中节点的中介中心性、接近中心性、特征向量中心性和 PageRank 值。

然后，为每一个 lncRNA-疾病对定义以下特征：1）mf 为对关联矩阵 $\boldsymbol{Y}$ 进行矩阵分解获得的 l_i 和 d_j 隐含向量；2）m. com表示lncRNA-疾病关联网络中lncRNA-疾病关联对的中介中心性、邻近中心性、特征向量中心性和 PageRank 值。将上述定义求出的特征向量拼接，获得 lncRNA 和疾病的 3 类型特征向量。

最后，将上述 3 类 lncRNA 特征进行拼接，获得 lncRNA 特征向量，并记其特征矩阵为 $\boldsymbol{L}$。将疾病的 3 类型特征拼接，获得疾病特征向量，其特征矩阵为 $\boldsymbol{D}$。在此基础上，通过随机选择 lncRNA 和疾病特征，并利用参数 $r(0<r\leqslant 1)$ 调节特征子集的大小来构建 lncRNA 特征 $\boldsymbol{L}^{(1)}$ 和疾病的特征 $\boldsymbol{D}^{(1)}$，再利用主成分分析（PCA）对 lncRNA 和疾病的特征子集进行特征提取，保留前 10 个特征，获得最终的 lncRNA 特征矩阵 $\boldsymbol{L}^{(2)}$ 和疾病特征矩阵 $\boldsymbol{D}^{(2)}$。

4 lncRNA-疾病关联预测方法

本文提出了一种基于集成回归决策树的方法，用于实现 lncRNA-疾病关联预测，具体包含以下 3 个步骤：

（1）构建训练样本集 T。因为每个数据集中拥有的正样本数最多占该样本集样本总数的 2.4%，所以本文选择所有的正样本和与其数量相同的负样本，来构建本文模型的训练集。这些负样本是从未知关联的样本中随机地选择出来的，$P=\{(l_i,d_j)|\boldsymbol{Y}(l_i,d_j)=1\}$ 表示正样本集，$U=\{(l_i,d_j)|\boldsymbol{Y}(l_i,d_j)=0\}$ 表示未知样本集，集合 $N(N\in U)$ 表示随机选择的负例集合，集合 N 和集合 P 的元素个数相同，$|N|=|P|$，集合 $T=N\cup P$ 是基础学习的训练样本集。

（2）训练回归决策树 DT。对于训练集 T 中的样本，将 lncRNA 特征矩阵 $\boldsymbol{L}^{(2)}$ 和疾病特征矩阵 $\boldsymbol{D}^{(2)}$ 拼接为样本的特征向量，用作回归决策树的输入向量。生成回归决策树的训练集表示为 $T'=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$，其中 $\boldsymbol{x}_i=(\boldsymbol{x}_i^{(1)},\boldsymbol{x}_i^{(2)},\cdots,\boldsymbol{x}_i^{(z)})$ 是 z 维的输入向量，代表邻接矩阵 $\boldsymbol{Y}$ 中第 i 个样本的特征向量，y_i 表示邻接矩阵 $\boldsymbol{Y}$ 中第 i 个样本的观测值，n 是训练集中的样本数。对于回归决策树 DT，使用平方误差最小准则的 CART 算法[18]构建回归树模型，递归地建立二叉决策树的过程就是生成回归树的过程，具体过程如下：

选择特征值 $x^{(j)}$ 来切分特征空间 R，j 和 $s(x^{(j)}=s)$ 分别是切分变量和切分点，两个特征子空间被定义为：

$$\begin{aligned}R_1(j,s)&=\{x|x^{(j)}\leqslant s\}\\R_2(j,s)&=\{x|x^{(j)}>s\}\end{aligned}\tag{9}$$

回归树可以被定义为：

$$DT(x)=c_k,x\in R_k,k=1,2\tag{10}$$

其中，c_k 表示子空间 R_k 的输出值，通过最小化平方误差 $\sum_{x_i\in R_k}(y_i-DT(\boldsymbol{x}_i))^2$ 来求解其最优解：

$$\hat{c}_k=\frac{1}{N_k}\sum_{x_i\in R_k(j,s)}y_i,\boldsymbol{x}\in R_m,m=1,2\tag{11}$$

其中，N_k 是子空间 R_k 输入向量的个数。为了选择最优的切分变量 j 与切分点 s，求解式(12)：

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]\tag{12}$$

在获得最优切分变量 j' 和最优切分点 s' 后，使用 (j',s') 根据式(9)划分特征空间，并根据式(10)和式(11)计算特征空间的输出值。然后，分别在子空间 R_1 和 R_2 中寻找新的最优切分变量和切分点，特征空间被分为 4 个部分，分别在 4 个子空间中计算新的输出 $\hat{c}_k(k=1,2,3,4)$。重复该过程直到无法切分子空间为止。最后，将特征空间划分为 K 个子空间，最终的回归树表示如下：

$$DT(x)=c_k,x\in R_k,k=1,2,\cdots,K\tag{13}$$

简而言之，将构建好的训练集 $T'=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 用于训练一棵回归决策树 DT，决策树中的输出 $\hat{c}_k$ 代表 lncRNA-疾病关联对的关联评分，每个 lncRNA-疾病关联对在决策树中均有对应评分，评分分数越大，表示其潜在关联概率就越大，反之亦然。

（3）构建基于平均策略的决策树集成模型。集成学习方法旨在将多个基础学习器通过一定策略组合为一个性能更强的学习器，本文通过上述 3 个步骤获得了集成所需的基础学习器，即回归决策树 DT，重复上述 3 个步骤 M 次获得 M 个基础学习器，并通过简单平均策略来获得最终的预测模型。集成的回归决策树表示如下：

$$Ensemble\_DT=\frac{1}{M}\sum_{i=1}^{M}DT_i\tag{14}$$

ERDTLDA 算法的算法描述如算法 1 所示。

算法 1 ERDTLDA 算法

输入：lncRNA 特征矩阵 $\mathbf{L}^{(2)}$，疾病特征矩阵 $\mathbf{D}^{(2)}$，正样本集 P，未知样本集 U，基学习器的数量 M

输出：集成的回归决策树 Ensemble_DT

1. For i=1 to M:
2. N_i = 从集合 U 中随机选择的负样本集（$|N_i|=|P|$）
3. 训练集 $T=P\cup N_i$
4. 训练集 $T'=[\mathbf{L}_i^{(2)},\mathbf{D}_j^{(2)},\mathbf{Y}]$
5. DT_i = 用训练集训练回归决策树模型
6. Return $Ensemble_{DT}=\frac{1}{M}\sum_{i=1}^{M}DT_i$

5 实验与结果

本节对本文方法进行实验验证和预测结果分析，包括不同参数对预测结果的影响，本文方法和已有方法的对比。为了进一步说明本文方法的有效性，使用本文方法获得的预测结果对乳腺癌、肺癌和胃癌 3 种特定疾病进行了案例分析。

实验基于 Python3.7，numpy 1.18.5，NetworkX 2.4，sklearn 0.21，pandas 1.0.5 实现。本实验采用 NVIDIA GeForce GTX 1080Ti 显卡、Intel(R) Xeon(R) Gold 5118 cpu 2.3GHz 2.29GHz（2 处理器）以及 128GB RAM（运行内存），采用 windows 10 操作系统。本文代码和数据可从网址[1)]获得。

5.1 评价方法与指标

本文使用留一交叉验证法和 5 折交叉验证法来评估本文方法的性能。留一交叉验证是将样本集分为 n 份（n 是所有样本的总数），每次选择一个样本用于验证，将剩余的 $n-1$ 个

1) https://github.com/ljatynu/ERDTLDA

样本用于测试，每个样本被拿出来做一次验证后，分类器最终的性能评价指标就是使用这 n 次验证得到的分类准确率的平均值。类似地，5 折交叉验证就是将样本集分为 5 份，每次选择一份样本集用于验证，将剩余的 4 份样本集用于测试，验证 5 次之后，对 5 次验证得到的分类准确率取平均作为 5 折交叉验证下该分类器的性能指标。

5.2 数据集

从文献[7,12]的补充文件中下载了 3 个 lncRNA 疾病关联数据集，该数据集包含 2012 年 10 月来自 LncRNADisease 数据库中的 167 种疾病和 118 种 lncRNA 之间的 293 种经实验证实的 lncRNA-疾病关联关系；2016 年 4 月来自 LncRNA-Disease 数据库中的 162 种疾病和 187 种 lncRNA 之间的 454 种已知的 lncRNA-疾病关联关系，以及 2016 年 7 月来自 Lnc2Cancer 数据库中的 79 种疾病和 310 种 lncRNA 之间的 594 种 lncRNA-疾病关联关系。实验数据集的具体信息如表 2 所列。

表 2 实验数据集概况

Table 2 Statistics of experimental dataset

数据集	lncRNA 数量	疾病数量	已知 lncRNA-疾病关联数量
数据集 1	118	167	293
数据集 2	187	162	454
数据集 3	310	79	594

5.3 最优参数选择

为了得到更准确的预测结果，通过实验分析不同参数对预测结果的影响，在使用平均集成策略时需要确定参数 M，也就是基学习器的个数，参数 M 选择不同，预测结果的准确性也会有所不同。因此，为了确定最优的参数 M，本文设置 M 的取值为 10，30，50，70，90，对每个取值都测试一次，测试过程使用 5 折交叉验证方法。训练集由已知的正样本和随机选择的与正样本等量的负样本构成。本文对每一组参数都使用训练集进行 5 折交叉验证，对从本文的 3 个数据集上获得的 AUC 值进行了比较，实验结果显示，M 从 10 到 50 范围内变化时，AUC 值显著提升，超过 50 之后变化不再明显，考虑到算法的时间复杂度，生成的基础学习器过多将导致本文提出的算法构建模型的时间过长，影响计算模型的总体性能，因此我们权衡算法预测的准确性和时间复杂度之后选择的最优参数 M 的取值为 50，参数 M 的数值对预测结果的影响具体表现如表 3 所列。

表 3 参数 M 不同取值对应的 AUC 值

Table 3 AUC value corresponding to the parameter M

参数 M	AUC 值(数据集 1)	AUC 值(数据集 2)	AUC 值(数据集 3)
10	0.8042	0.7972	0.8182
30	0.8111	0.8091	0.8186
50	0.8546	0.8397	0.8607
70	0.8557	0.8415	0.8623
90	0.8563	0.8473	0.8648

5.4 与其他方法的对比结果

将本文方法分别与基于生物信息网络和基于机器学习的典型预测方法进行对比。具体地，基于生物信息网络的方法如下：1)Lap-BiRWRLDA[6] 分别在 lncRNA 和疾病相似性网络上随机游走，最后将两个网络上的随机游走加权平均值作为 lncRNA-疾病关联的预测指标；2)LRLSLDA[7] 根据已知的 lncRNA-疾病关联、lncRNA 表达谱计算疾病和 lncRNA 的高斯相互作用谱内核相似性，然后利用拉普拉斯正则化最小二乘法框架来预测潜在的关联；3)GrwLDA[4] 在全局网络上通过随机游走模型来预测潜在的 lncRNA-疾病关联。本文方法分别在数据集 1、数据集 2、数据集 3 上与这些基准方法进行了预测性能的比较，结果如图 1(a)－图 1(c)所示。

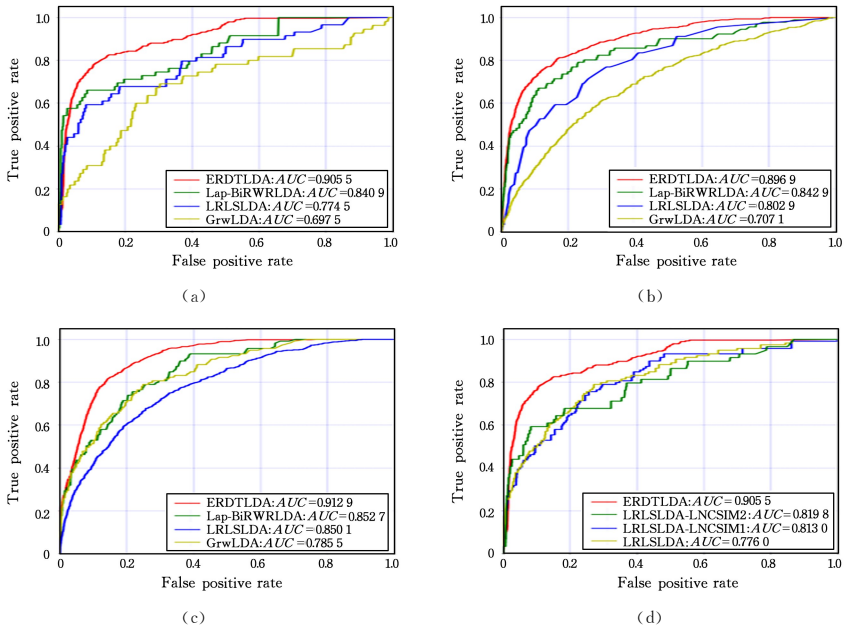

图 1 与基准方法的预测性能比较

Fig. 1 Performance comparison with other baselines

可以看到，本文方法的预测性能在 3 个数据集上均优于基准方法。基于机器学习的方法，在数据集 1 上，本文方法与 3 种不同的基于机器学习的方法 LRLSLDA[7]，LRLSLDA-LNCSIM1[8]，LRLSLDA-LNCSIM2[8] 分别进行了比较，结果

如图 1(d)所示,同样本文方法优于这 3 种机器学习的基准方法。

与现有的两类方法相比,本文既充分利用了生物信息网络的结构信息来获得 lncRNA、疾病的特征表示,又基于集成决策树,通过机器学习的方法来训练预测模型,因此取得了优于两类现有方法的预测效果。

5.5 集成学习的效果

本文方法使用集成学习中的平均集成策略,在获得多个基础学习器后,使用集成策略将多个基础学习器相结合,构建最终的分类模型,本节在 3 个真实的数据集上使用 5 折交叉验证比较了使用平均集成策略的模型获得的 *AUC* 值与仅使用基础分类器(即单棵回归决策树)模型获得的 *AUC* 值,实验结果如表 4 所列。

表 4 5 折交叉验证下单棵回归决策树和集成回归决策树的 *AUC* 值

Table 4 *AUC* value of single decision tree and ensemble decision tree under 5-fold cross-validation

模型	*AUC* 值(数据集 1)	*AUC* 值(数据集 2)	*AUC* 值(数据集 3)
单棵回归决策树	0.7706	0.7594	0.7819
集成回归决策树	0.8616	0.8433	0.8728

实验结果表明,使用平均的集成策略将基学习器组合起来的预测模型的性能明显优于仅使用单棵回归决策树的预测模型,这是因为我们在选择训练集时,采用了随机选择负例的策略,导致我们仅使用单棵回归决策树作为预测模型时,负例的选择差异较大,使得构建的模型之间的预测效果偏差较大;当采取集成策略之后,使用了更多的负例来构建模型,可以减小预测的偏差,从而提高预测的准确性。

5.6 实际案例

为了进一步验证本文方法的可靠性和实用性,分别对乳腺癌、肺癌和胃癌 3 种特定的疾病进行案例分析。我们对 ERDTLDA 预测出来的结果进行整理,对于某一个疾病,我们的预测结果中与该疾病关联的 lncRNA 按照其与该疾病的关联得分从大到小进行排序,排名前十五的 lncRNA 被认为是与该疾病有潜在关联的 lncRNA。接下来,通过从 LncRNADisease 数据库和 Lnc2Cancer 数据库中人工查找疾病和 lncRNA 关联信息,以验证本文预测的与该疾病关联的 lncRNA 是否得到了生物医学文献的证实。

乳腺癌的致病机制被认为是环境因素与遗传易感宿主之间相互作用的结果[19]。很多 lncRNA 通过上调或下调乳腺癌基因与乳腺癌疾病关联。本文方法能够识别与乳腺癌相关的潜在 lncRNA,并且在 LncRNADisease 数据库和 Lnc2Cancer 数据库中,我们预测出的与乳腺癌关联分数排名前十五的 lncRNA 中有 13 个获得了生物医学文献的证实,具体信息请见 OSID 码中的表 1。

将本文方法预测出的所有与肺癌有关联得分的 lncRNA 筛选出来,按照得分大小从大到小排序,排名前十五的被认为是与肺癌相关的 lncRNA。经验证,本文预测结果中的 lncRNA有12个在相关数据库中被证实与肺癌关联,预测结果请见 OSID 码中的表 2。

使用本文方法可以预测可能与胃癌相关的 lncRNA,预测结果与胃癌关联分数最高的前 15 个 lncRNA 中有 11 个已被证实与胃癌相关。与胃癌有关的 lncRNA 被验证的情况请见 OSID 码中的表 3。在胃癌的案例分析中发现,即使我们的数据集中原本没有胃癌与预测出的 lncRNA 之间的关联关系,我们仍然能够预测其关联的 lncRNA,且有较高的准确性。

通过上述案例研究可以得出以下结论,本文方法 ERDTLDA 的预测结果具有较高的准确性,同时也对预测 lncRNA-疾病关联具有很高的可靠性。

结束语 预测新的 lncRNA-疾病关联有助于从 lncRNA 的角度促进对疾病发病机制的了解,并有益于疾病的治疗。本文提出的基于决策树的计算模型可以成为发现 lncRNA 与疾病关联的实验方法的有利补充,研究人员能够用这种方式筛选用于实验的对象,使得生物医学实验更加富有成效。

本文提出的集成回归决策树 lncRNA-疾病关联预测方法的效果良好,但是也有不足之处和可以进一步优化的部分。例如,本文模型需要负样本,即无关联的 lncRNA-疾病关联对,我们随机抽取了一些没有已知关联的 lncRNA-疾病关联对作为负样本进行模型训练,每次随机选取不同的负样本实际上会对方法的性能产生一定影响。对此,我们将在以后的研究中努力开发新的方法,策略性地选择负样本,以提升模型的预测性能。

参 考 文 献

[1] HUTTENHOFER A, SCHATTNER P, POLACEK N. Non-coding RNAs: hope or hype? [J]. Trends in Genetics, 2005, 21(5): 289-297.

[2] GEISLER S, COLLER J. RNA in unexpected places: Long non-coding RNA functions in diverse cellular contexts[J]. Nature Reviews Molecular Cell Biology, 2013, 14(11): 699-712.

[3] CHEN X, YAN C C, ZHANG X, et al. Long non-coding RNAs and complex diseases: From experimental results to computational models[J]. Briefings in Bioinformatics, 2016, 18(4): 558-576.

[4] SUN J, SHI H B, WANG Z Z, et al. Infering novel lncRNA-disease asociations based on a random walk model of a lncRNA functional similarity network[J]. Molecular BioSystems, 2014, 10(8): 2074-2081.

[5] GU C, LI X Y, CAI L J, et al. Global network random walk for predicting potential human lncRNA-disease associations[J]. Sci. Rep., 2017, 7(1): 12442-12453.

[6] WEN Y, HAN G, ANH V. Laplacian normalization and bi-random walks on heterogeneous networks for predicting lncRNA-disease associations[J]. BMC Systems Biology, 2018, 12(9), 122-131.

[7] CHEN X, YAN G Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles[J]. Bioinformatics, 2013, 29(20): 2617-2624.

[8] CHEN X, YANG C G, LUO C, et al. Constructing lncRNA

functional similarity network based on lncRNA-disease associations and disease semantic similarity[J]. Scientific Reports, 2015,5:11338-11350.

[9] ZHAO T T,XU J Y,LIU L,et al. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features[J]. Molecular BioSystems,2014,11(1):126-136.

[10] XUAN P,PAN S,ZHANG T,et al. Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations[J]. Cells, 2019, 8(9), 1012:1-16.

[11] WANG M N,YOU Z H,WANG L. LDGRNMF:LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization[J]. Neurocomputing,2021,424:236-245.

[12] LIU J X,CUI Z,GAO Y L,et al. WGRCMF:A Weighted Graph Regularized Collaborative Matrix Factorization Method for Predicting Novel LncRNA-Disease Associations[J]. IEEE Journal of Biomedical and Health Informatics,2021,25(1):257-265.

[13] WEI H,LIAO Q,LIU B. iLncRNAdis-FB:identify lncRNA-disease associations by fusing biological feature blocks through deep neural network[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics/IEEE,ACM,2020(99):1-13.

[14] MA Y,GUO X L,SUN Y T,et al. Prediction of Disease Associated Long Non-Coding RNA Based on HeteSim[J]. Journal of Computer Research and Development,2019,56(9):1889-1896.

[15] CHEN G,WANG Z Y,WANG D Q,et al. LncRNADisease:A database for long-non-coding RNA-asociated diseases[J]. Nucleic Acids Research,2012,41(D1):D983-D986.

[16] NING S,ZHANG J,PENG W,et al. Lnc2Cancer:A manualy curated database of experimentaly supported lncRNAs asociated with various human cancers[J]. Nucleic Acids Research,2015,44(D1):D980-D985.

[17] PENG H,LAN C W,LIU Y S,et al. Chromosome preference of disease genes and vectorization for the prediction of non-coding disease genes[J]. Oncotarget,2017,8(45):78901-78916.

[18] BREIMAN L,FRIEDMAN J,STONE C J,et al. Classification and regression trees[M]. CRC Press,1984:1-18.

[19] FRIENDENSON B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers[J]. BMC Cancer,2007,7(1):152-162.

REN Shou-peng, born in 1997, master. His main research interests include bioinformatics and machine learning.

LI Jin, born in 1975, Ph.D, professor. His main research interests include machine learning and bioinformatics.

(责任编辑:喻藜)