

基于融合神经网络的LncRNA与疾病关联预测方法

李巧君, 章文, 杨伟

引用本文

李巧君, 章文, 杨伟. [基于融合神经网络的LncRNA与疾病关联预测方法](#)[J]. 计算机科学, 2023, 50(8): 226-232.

LI Qiaojun, ZHANG Wen, YANG Wei. [Fusion Neural Network-based Method for Predicting LncRNA-disease Association](#) [J]. Computer Science, 2023, 50(8): 226-232.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[类脑心智计算的科学技术和工程应用的研究与思考](#)

Thoughts on Development and Research of Science, Technology and Engineering Application of Brain & Mind-inspired Computing

计算机科学, 2023, 50(2): 364-373. <https://doi.org/10.11896/jsjcx.220500023>

[基于集成回归决策树的lncRNA-疾病关联预测方法](#)

Ensemble Regression Decision Trees-based lncRNA-disease Association Prediction

计算机科学, 2022, 49(2): 265-271. <https://doi.org/10.11896/jsjcx.201100132>

[物联网通信协议的安全研究综述](#)

Security for Communication Protocols in Internet of Things: A Survey

计算机科学, 2018, 45(12): 32-41. <https://doi.org/10.11896/j.issn.1002-137X.2018.12.005>

[基于车联网的交通管控机制设计与验证](#)

Traffic Control Mechanism Design and Verification Based on VANET

计算机科学, 2017, 44(Z6): 276-283. <https://doi.org/10.11896/j.issn.1002-137X.2017.6A.064>

[基于NAK的WISP数据传输方案](#)

Negative Acknowledgement Based Data Delivery Scheme for WISP

计算机科学, 2017, 44(Z6): 294-299. <https://doi.org/10.11896/j.issn.1002-137X.2017.6A.067>

基于融合神经网络的 LncRNA 与疾病关联预测方法

李巧君^{1,2} 章 文³ 杨 伟^{1,2}

1 河南省工业物联网应用工程技术研究中心 河南 南阳 473000

2 河南工业职业技术学院电子信息工程学院 河南 南阳 473000

3 华中农业大学信息学院 武汉 430070

摘 要 长链非编码 RNA(Long non-coding RNA, LncRNA)的异常表达与疾病的生理和病理过程密切相关,识别 LncRNA 与疾病之间的潜在关联有助于理解疾病的分子发病机制。以往的方法未能深度整合异构的多源数据以及学习高维特征表示。为此,文中提出了一种基于融合神经网络(Fusion Neural Networks, FNN)预测候选疾病相关 LncRNA 的方法 FNNLDA。FNNLDA 整合与 LncRNA、疾病和 miRNAs 相关的多种数据,采用多模型融合思想,利用栈式自编码器和融合神经网络两种深度学习模型分别学习 LncRNA-疾病对的高级特征,最后融合两个模块的预测分值来预测 LncRNA-疾病的关联性。五折交叉验证显示 FNNLDA 方法的 AUC 值比 SIMCLDA, MFLDA, CNLDA 和 LRLSLDA 分别提升了 12.5%, 15.1%, 3.4% 和 5.8%, 表明其在 LncRNA-疾病预测性能上有较大提升。基于胃癌疾病案例进行研究,结果证明 FNNLDA 能够有效识别与疾病关联的潜在 LncRNA。

关键词: LncRNA-疾病; 关联预测; 融合神经网络; 栈式自编码器

中图法分类号 TP391

Fusion Neural Network-based Method for Predicting LncRNA-disease Association

LI Qiaojun^{1,2}, ZHANG Wen³ and YANG Wei^{1,2}

1 Henan Province Industrial Internet of Things Application Engineering Technology Research Centre, Nanyang, Henan 473000, China

2 School of Electronic Information Engineering, Henan Polytechnic Institute, Nanyang, Henan 473000, China

3 College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Abstract Aberrant expression of long non-coding RNA(LncRNA) is closely associated with the physiological and pathological processes of diseases. Identifying potential associations between LncRNA and diseases are helpful to understand the molecular pathogenesis of diseases. Previous researches were scarcely integrated with heterogeneous multi-source data and seldom learned high-dimensional feature representations. In this paper, we propose a new method named FNNLDA, which based on fusion neural networks(FNN) to predict the associated LncRNAs of candidate disease. FNNLDA integrates multiple data related to LncRNAs, diseases, and miRNAs. And employs the idea of multi-model fusion to learn high-level features of LncRNA-disease pairs by using two deep learning models; stacked self-encoder and fusion neural network, separately. Finally, fusing the prediction scores of the two modules to predict the LncRNA-disease associations. Five-fold cross-validation test show that the AUC value of FNNLDA method is 12.5%, 15.1%, 3.4% and 5.8% higher than that of SIMCLDA, MFLDA, CNLDA and LRLSLDA, respectively. It indicates that this method has a significant improvement in LncRNA-disease prediction. The results of the study based on stomach cancer disease cases demonstrate that FNNLDA can effectively identify potential LncRNAs associated with disease.

Keywords LncRNA-disease, Association prediction, Fusion neural network, Stacked autoencoder

1 引言

LncRNA 是一种非蛋白编码 RNA,其长度超过 200 个核苷酸^[1]。过去,人们普遍认为 LncRNA 对基因表达几乎没有影响。目前,诸多研究表明人类身体各种复杂疾病(如癌症)的形成及发展与 LncRNA 的突变和失调的病理机制密切相关^[2]。例如,Zhuo 等^[3]发现在胃肿瘤转移和发展过程中, LncRNA GMAN 或 ephrin A1 的调节机制发挥了重要的调控

作用;Li 等^[4]发现在对癌症进行免疫治疗的过程中,被命名为 LIMIT 的癌症免疫原性 LncRNA 可能是治疗靶点;LncRNA-HOTAIR 在不同的癌细胞中扮演着致癌分子的角色,其表达水平是乳腺癌、胃癌、结直肠癌和宫颈癌等癌症诊断和治疗的潜在生物标志物^[5]。因此,识别疾病相关的 LncRNA 将有助于在 LncRNA 水平上了解人类复杂的疾病机制、疾病诊断和治疗。

近年来,随着生物技术的发展,实验发现的 LncRNA-

到稿日期:2022-10-25 返修日期:2022-12-23

基金项目:河南省科技攻关项目(212102310086)

This work was supported by the Henan Science and Technology Project(212102310086).

通信作者:李巧君(qiaojun209@163.com)

疾病关联数据逐渐增加,研究者整理这些数据并建立了 LncRNA-疾病关联数据库^[6],Lnc2Cancer^[7]和 HDMM^[8]等 LncRNA-疾病关联数据库。目前,已知的 LncRNA-疾病关联只是小部分,而生物实验验证耗时且成本昂贵。随着人工智能技术的发展和大数据技术的成熟,研究人员利用计算方法分析和处理已知数据,能够加速发现 LncRNA 和疾病之间的潜在关联。该方法已成为生物实验识别 LncRNA-疾病关联的有效补充。

目前,基于计算的 LncRNA-疾病关联预测方法可以分为两类。第一类是基于矩阵分解的方法。Fu 等^[9]开发了基于矩阵分解的预测模型 MFLDA,MFLDA 通过分配不同的权重对数据源进行融合,并利用矩阵三分解将异构数据源分解为低秩矩阵来预测 LncRNA-疾病关联。Lu 等^[10]提出了一种基于归纳矩阵补全的 LDA 预测方法 SIMCLDA,该方法通过主成分分析(Principal Component Analysis,PCA)分别从 LncRNA 相似性和疾病相似性中提取主要特征向量,计算新的 LncRNA 与其邻居之间的相互作用模式,并利用主要特征向量来实现与归纳矩阵补全的关联矩阵,构建交互模式。第二类是基于机器学习的方法。Chen 等^[11]提出了基于半监督学习框架的计算模型 LRLSLDA,LRLSLDA 计算了 LncRNA 相似度和疾病相似度,分别基于疾病空间和 LncRNA 空间中的拉普拉斯正则最小二乘法建立了两个分类算子,并将这两个分类算子合并为一个分类因子,以获得疾病与 LncRNA 之间的最终关联概率。Xie 等^[12]提出了一种利用相似核融合方法预测 LncRNA 与疾病相关性的方法,该方法简称为 SKF-LDA。SKF-LDA 对 LncRNA 和疾病采用不同类型相似性的核融合方法,LncRNA 采用表达相似度和余弦相似度,疾病采用语义相似度和余弦相似度;利用正常的拉普拉斯正则化最小二乘法计算 LncRNA 与疾病的最终关联。

深度学习作为机器学习的一个分支,已被广泛应用于生物学中的各种预测问题。Xuan 等^[13]提出了一种基于双卷积

神经网络预测与潜在疾病相关 LncRNA 的方法 CNNLDA,CNNLDA 分别在特征和关系层面构建注意力机制,以区分特征的不同贡献,学习 LncRNA-疾病关联的更多信息。Xuan 等^[14]提出了一个基于图卷积神经网络和卷积神经网络预测与疾病相关的 LncRNA 的模型 GCNLDA。Yang 等^[15]提出了一个双向生成对抗网络模型 BiGAN,模型中编码器和生成器用来学习高级特征,分类器用来预测 LncRNA-疾病关联。目前,与其他生物问题相比,被生物实验识别出来的与疾病确定存在关联的 LncRNA 数量非常有限,栈式自编码器(Stacked AutoEncoder,SAE)^[16]能够对每个成对的 LncRNA-疾病关联样本进行模型训练,解决训练样本少的问题。在生物序列特征学习中,卷积神经网络(Convolutional Neural Network,CNN)^[17]能够提取 LncRNA-疾病关联的局部关键信息,但不能学习序列不同时刻的信息以及不同特征的重要性,而长短期记忆递归神经网络(Long Short Term Memory,LSTM)^[18]能够考虑不同时刻与不同特征的重要性。

鉴于上述分析,本文提出了一种基于融合神经网络 FNN 并利用栈式自编码器的关联预测方法 FNNLDA (Fusion Neural Networks Predict LncRNA-disease Association)。FNNLDA 整合了 LncRNA 功能相似性、疾病语义相似性、LncRNA-疾病相互作用、LncRNA-miRNA 相互作用以及 miRNA-疾病关联多源数据来构建特征矩阵,并将构建的特征矩阵分别输入栈式自编码器特征提取模块和融合神经网络特征提取模块来学习 LncRNA-疾病对的潜在表示,最后将两个模块加权融合后以预测 LncRNA-疾病关联。五折交叉验证结果表明,FNNLDA 可以达到更高的 AUC 值,具有更好的性能。此外,在胃癌疾病中进行案例研究,结果表明 FNNLDA 可以准确地预测出与疾病相关的潜在 LncRNA。

2 基于融合神经网络的关联预测方法

FNNLDA 方法的整体框架如图 1 所示。

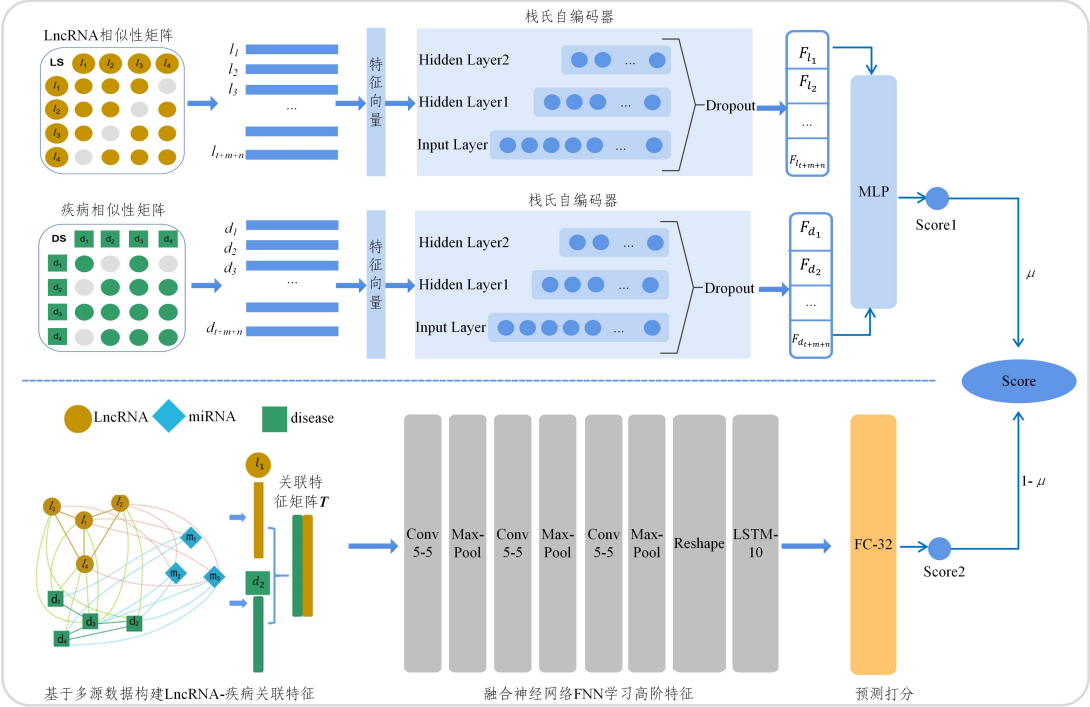


图1 FNNLDA 体系结构图

Fig. 1 FNNLDA architecture diagram

FNNLDA 方法分为两个模块:第一个模块基于 SAE 学习 LncRNA-疾病对的低维网络表示;第二个模块基于融合神经网络 FNN 进行高阶特征的提取,此模块充分利用 CNN 和 LSTM 的不同优势,学习 LncRNA-疾病时间依赖和共享参数的高级特征。两个模块都经过全连接层和 Softmax 层,得到 LncRNA-疾病的相关可能性,即关联得分。最后,赋予两个模块的预测分值不同的权值,对分值进行加权融合得到预测得分,从而筛选出与疾病相关的 LncRNA。

2.1 数据特征构建

2.1.1 LncRNA、疾病相似性特征

LncRNA-疾病的关联预测常常是利用 LncRNA、疾病的生物信息以及已知的 LncRNA-疾病关系对预测未知的 LncRNA-疾病进行关联,从而挖掘潜在的 LncRNA-疾病关联。LncRNA、疾病的生物信息常总结为相似度矩阵,Chen 等^[19]提出可以通过计算两种疾病的相似性来衡量 LncRNA 之间的功能相似性。本文采用 Chen 提出的方法,构建了 LncRNA 功能相似性矩阵 LS , $LS(i, j)$ 表示 LncRNA $l(i)$ 与 LncRNA $l(j)$ 之间的相似度,其值在 0~1 之间变化,越接近于 1 表示 $l(i)$ 与 $l(j)$ 越相似。 LS 作为栈氏自编码器特征提取模块的输入。

与某一疾病相关的所有语义术语,形成其有向无环图 (Directed Acycline Graph, DAG)。Wang 等^[20]根据疾病

DAGs 模型提取疾病之间的语义相似性。本文基于 Wang 提出的模型完成对疾病语义相似度权值的定义,构建了疾病语义相似度矩阵 DS , DS 是栈氏自编码器特征提取模块的另一个输入。 $DS(i, j)$ 表示疾病之间的语义相似度,取值范围为 0~1,越接近于 1 表示 $d(i)$ 与 $d(j)$ 越相似。

2.1.2 LncRNA-疾病关联特征

本文利用生物信息数据库中已知的 LncRNA 相似性、疾病相似性、LncRNA-疾病的相互作用关系、LncRNA-miRNA 的相互作用关系和疾病-miRNA 的相互作用关系等多种数据信息,采用嵌入策略构建 LncRNA-疾病关联特征。

LncRNA 功能相似性表示为 $LS=[LS_{ij}] \in R^{N_l \times N_l}$, 疾病的语义相似度表示为 $DS=[DS_{ij}] \in R^{N_d \times N_d}$ 。LncRNA 与疾病的相互作用关系矩阵表示为 $LD=[LD_{ij}] \in R^{N_l \times N_d}$, LncRNA 与 miRNA 的互作关系矩阵表示为 $LM=[LM_{ij}] \in R^{N_l \times N_m}$, 疾病与 miRNA 的互作关系矩阵表示为 $DM=[DM_{ij}] \in R^{N_d \times N_m}$ 。 LD, LM 和 DM 取值均为 1 或 0,若 LncRNA l_i 与 miRNA m_j 相互作用,则 $LM_{ij}=1$, 否则 $LM_{ij}=0$ 。对于 LncRNA l_i 和疾病 d_j ,如果与它们相关联的 LncRNA、疾病以及 miRNA 及其互作关系有较多的相同点,那么 l_i 和 d_j 的关联性将很大。LncRNA l_1 和疾病 d_2 构建的特征矩阵的拼接过程如图 2 所示。

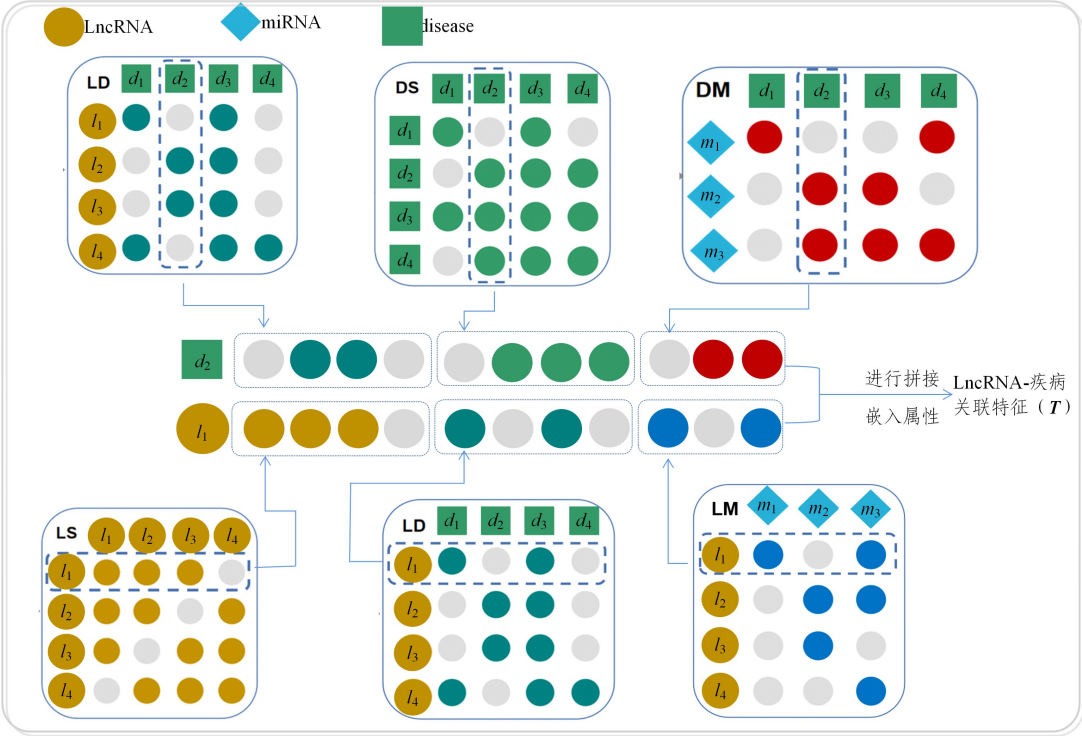


图 2 LncRNA l_1 和疾病 d_2 的关联特征

Fig. 2 Associated characteristics of LncRNA l_1 and disease d_2

本文将 LncRNA l_i 和所有 LncRNA 的相似关系记作 x_1 , l_i 与所有疾病的相互作用关系用向量 x_2 表示, l_i 与每个 miRNA 的相互作用用向量 x_3 表示;与 l_i 类似,疾病 d_j 与所有 LncRNA 的互作关系用向量 y_1 表示,疾病 d_j 与各个疾病的语义相似性记为 y_2 , 向量 y_3 表示疾病 d_j 与每个 miRNA 的互作关系。属性向量通过式(1)进行拼接,从而得到 LncRNA l_i 和疾病 d_j 的属性关联嵌入矩阵 T 。该 LncRNA 和疾病的关联嵌入

矩阵 T 就是融合神经网络的输入。

$$T = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}, T \in R^{(N_l + N_d + N_m) \times 2} \tag{1}$$

其中, $(N_l + N_d + N_m) \times 2$ 表示关联矩阵 T 的维度。

2.2 基于栈氏自编码器 SAE 的特征提取模块

栈氏自编码器是由多个自动编码器形成的深层神经网络,每层包含一定数量的神经元,并且每层的输出连接到连续

层的隐藏单元相同。在栈式自编码器 SAE 特征提取模块,将 LncRNA 相似性特征矩阵 $\mathbf{LS}$ 和疾病相似性特征矩阵 $\mathbf{DS}$ 作为输入,分别送入由多个连接层组成的两个自编码器进行特征学习,生成高阶特征向量。最后,采用多层感知机 (Multi-Layer Perceptorn, MLP) 网络模型来实现评分预估,将高阶特征向量 $\mathbf{F}$ 馈入到全连接层,如式(2)所示:

$$Z = \sigma_L(\mathbf{W}_L(\sigma_{L-1}(\mathbf{W}_{L-1} \cdots \sigma_1(\mathbf{W}_1 \mathbf{F} + \mathbf{b}_1)) + \mathbf{b}_{L-1}) + \mathbf{b}_L) \quad (2)$$

其中, L 表示多层感知机的第 L 层, $\mathbf{W}_L$ 表示 L 层的权值矩阵, $\mathbf{b}_L$ 代表 L 层的偏置向量, σ_L 为激活函数。对于多层神经网络,一般将上层节点设置为上一层输出节点,与之连接的下层节点作为下一层的输入节点,上下层之间会存在一定的函数关系即激活函数,激活函数根据应用场景进行调整。FNNLDA 方法使用 ReLU 激活函数使得模型收敛,ReLU 函数的解析式如式(3)所示:

$$\text{ReLU} = \max(x, 0) \quad (3)$$

通过最小化目标函数对模型进行训练,从而使得损失最小化。本文通过使用交叉熵代价函数 C 来获得最佳预测,如式(4)所示:

$$C = -\frac{1}{n} \sum_x \sum_t [\mathbf{y} \ln a + (1 - \mathbf{y}) \ln(1 - a)] \quad (4)$$

其中,训练样本的索引用 x 表示, t 表示不同标签的索引, $\mathbf{y}$ 表示样本的真实标签, a 用于表示样本 x 的模型输出。

在最后的三层感知机 MLP 中获得每种 LncRNA 与疾病之间关联的最终预测,当预测的每个 LncRNA-疾病样本对超过阈值时,认为该 LncRNA 与潜在的疾病具有相关性。Softmax 激活函数的功能是将未归一化的输入映射到一组指数化和归一化的概率中。该模块使用 Softmax 函数计算 LncRNA 和疾病的关联概率 Score_1 , Score_2 如式(5)所示:

$$\text{Score}_1 = \text{softmax}(\mathbf{WZ}_{\text{src}} + \mathbf{b}) \quad (5)$$

其中, $\mathbf{W}$ 代表参数矩阵, $\mathbf{b}$ 为偏置向量, $\mathbf{Z}_{\text{src}}$ 为输入特征向量。

2.3 基于融合神经网络 FNN 的特征提取模块

融合神经网络 FNN 模块利用融合神经网络对 LncRNA 和疾病构建的特征矩阵进行特征提取。在特征学习之前,为了充分学习特征信息,需要在每次输入前对输入层的周围进行补零操作,执行过程如式(6)~式(12)所示:

$$f' = \text{CNN}(f) \quad (6)$$

$$h_{t-1} = f' \quad (7)$$

$$f_t = \sigma(\mathbf{W}_{ef}e_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + \mathbf{b}_f) \quad (8)$$

$$i_t = \sigma(\mathbf{W}_{ei}e_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + \mathbf{b}_i) \quad (9)$$

$$o_t = \sigma(\mathbf{W}_{eo}e_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_{t-1} + \mathbf{b}_o) \quad (10)$$

$$c_t = f_t c_{t-1} + i_t \tanh(\mathbf{W}_{ec}e_t + \mathbf{W}_{hc}h_{t-1} + \mathbf{b}_c) \quad (11)$$

$$h_t = o_t \odot \tanh(c_t) \quad (12)$$

其中, f 是 LncRNA 和疾病构建的特征矩阵,作为 CNN+LSTM 网络的输入模块; σ (sigmoid) 是用于模型计算的损失函数; $\tanh$ 表示激活函数; $\odot$ 符号用于点乘的计算; $\mathbf{W}$ 表示权重矩阵; $\mathbf{b}$ 为偏置向量。

FNN 模块包括 CNN 卷积神经网络和 LSTM 循环神经网络两部分,其中在 CNN 卷积神经网络部分,首先使用一个具有 5 个卷积核的卷积层,然后对数据进行最大池化降维,

总共交替使用 3 个卷积层和池化层,以获得更重要、更深层次的信息。在 LSTM 循环神经网络部分使用一层 10 个神经元的双向 LSTM,再通过 ReLU 激活函数馈入到全连接层 (Fully Connected Layers, FC),统一维数为 32。

构建的特征矩阵经过融合神经网络之后,会经过一个全连接层来获取关联预测的概率分值 Score_2 。融合卷积后,不同特征图展平合并为一维向量 $\mathbf{Z}_{\text{fnn}}$ 。预测概率如式(13)所示:

$$\text{Score}_2 = \text{softmax}(\mathbf{WZ}_{\text{fnn}} + \mathbf{b}) \quad (13)$$

其中, $\mathbf{W}$ 为参数矩阵, $\mathbf{b}$ 为偏置向量, $\mathbf{Z}_{\text{fnn}}$ 为输入特征向量。

2.4 评分预测模块

FNNLDA 方法通过为基于栈式自编码器 SAE 的特征提取模块和基于融合神经网络 FNN 的特征提取模块赋予不同的权值,来对预测分值进行加权融合,以预测 LncRNA-疾病的关系。融合方法如式(14)所示:

$$\text{Score} = \mu \times \text{Score}_1 + (1 - \mu) \times \text{Score}_2 \quad (14)$$

其中, μ 表示基于栈式自编码器的关联预测模块对关联预测分值的贡献; $1 - \mu$ 表示基于融合神经网络的关联预测模型对关联预测分值的贡献; Score 是一个向量,包含 LncRNA 与疾病有关联的预测分值和无关联的预测分值。

当预测有关联的分值比预测无关联的分值大时,就可以认为该 LncRNA-疾病是有关联的,否则认为 LncRNA 和疾病之间没有关联。

3 实验对比与结果分析

3.1 实验环境与数据集

FNNLDA 的实验环境是 Python 3.8, Keras2.4.3, TensorFlow 2.3, numpy1.19.2, sklearn0.24.2。实验数据集选取 LncRNA、疾病、miRNA 以及相似和互作数据,其中 LncRNA-disease 关联数据集来源于数据库 LncRNADisease^[6] 和 Lnc2Cancer^[7], miRNA-LncRNA 互作关系以及 miRNA-disease 互作关系分别从 miRNet^[21] 和 starBase^[22] 数据库下载。所有这些下载的关联和相似性随后被编译整理为 240 个 LncRNA、412 个疾病、495 个 miRNAs 和 2 697 个已知的 LncRNA-疾病关联,如表 1 所列。

表 1 相关数据集统计

Table 1 Statistics of relevant datasets

名称	样本集	关联数量	数据来源
LncRNA-疾病	240×402	2 687	LncRNADisease, Lnc2Cancer
LncRNA-miRNA	240×495	1 002	starBase
miRNA-疾病	495×402	5 218	HDMM

3.2 评估指标

本文采用五折交叉验证法来评估 FNNLDA 方法的预测能力,并将预测与疾病关联的 LncRNA 降序排序。所有已知的 LncRNA-疾病关联被视为阳性样本,从未观察到的 LncRNA-疾病关联中选取与阳性样本同样数量的样本作为阴性样本,本文中数据集包括 5 394 个 LncRNA-疾病对 (即 2 697 个正相关和 2 697 个负相关)。将准确率 (Accuracy, ACC)、精确度 (Precision, PRE)、F1 分数 (F1-score)、召回率 (Recall, REC)、接收者操作特征 (AUC-ROC) 曲线下的面积和 PR

曲线下面积(AUPR)作为评估指标。

3.3 参数分析

对于FNNLDA模型,选择不同的超参数时其预测性能也会不同。为了获取最佳的预测性能,通过尝试不同的超参数来寻找预测LncRNA-疾病关联的最佳参数。FNNLDA模型中有4个关键参数:训练次数(*epochs*)、丢弃率(*dropout*)、学习率(*learning rate*)以及批量大小(*batch size*)。其中,训练次数 $epochs \in \{10, 20, 30, 40, 50, 60, 70\}$, 丢弃率 $dropout \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, 学习率 $learning\ rate \in \{0.01, 0.001,$

$0.0001\}$, 批量大小 $batchsize \in \{16, 32, 64, 128\}$ 。本文分别调整每个参数组合对模型性能的影响,并使用5倍交叉验证对模型进行评估,如图3所示。从图3(a)和图3(b)可以看出,训练次数 *epochs* 为40,丢弃率 *dropout* 为0.4时,模型的性能最好;从图3(c)和图3(d)可以看出,学习率 *learningrate* 为0.001,批量大小 *batchsize* 为64时,模型所有的评价指标都获得了最好的性能。因此,FNNLDA模型参数的最佳值如下:训练次数为40,丢弃率为0.4,学习率为0.001,批量大小为64。

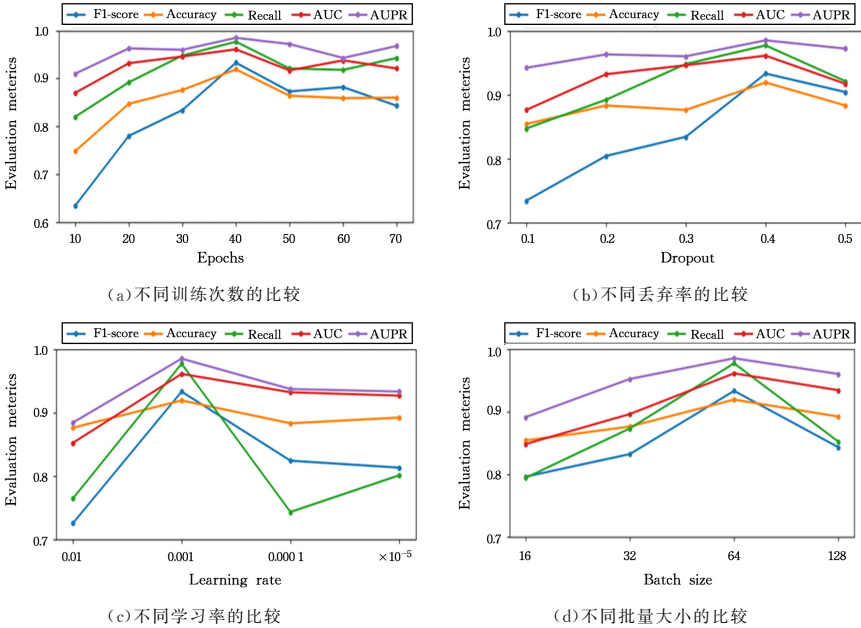


图3 参数优化结果
Fig. 3 Parameter optimization results

3.4 实验对比

本文选用SIMCLDA^[10],MFLDA^[9],CNNLDA^[13]以及LRLSLDA^[11]这4种先进方法与FNNLDA进行对比,实验结果如表2所列。其中,SIMCLDA和MFLDA是基于矩阵分解的方法,CNNLDA是基于深度学习的方法,而LRLSLDA是基于机器学习的方法,这几种方法均是近年来LncRNA-疾病关联预测方法中的典型代表。由表2可以看出,FNNLDA方法与SIMCLDA,MFLDA,CNNLDA和LRLSLDA相比,在AUC指标上分别提升了12.5%,15.1%,3.4%和5.8%;在AUPR指标上分别提升了11.7%,12.4%,4.7%和10.6%。上述两个指标结果均表明FNNLDA方法的性能优于其他4种方法。FNNLDA方法和上述4种方法的AUC值和AUPR值的ROC曲线和PR曲线如图4所示。

表2 FNNLDA与其他方法的对比

Methods	AUC	ACC	PRE	F1-Score	REC	AUPR
FNNLDA	0.961	0.919	0.901	0.933	0.977	0.985
SIMCLDA	0.836	0.742	0.684	0.764	0.900	0.868
MFLDA	0.81	0.774	0.912	0.69	0.594	0.861
CNNLDA	0.927	0.875	0.879	0.862	0.853	0.938
LRLSLDA	0.883	0.794	0.803	0.792	0.789	0.879

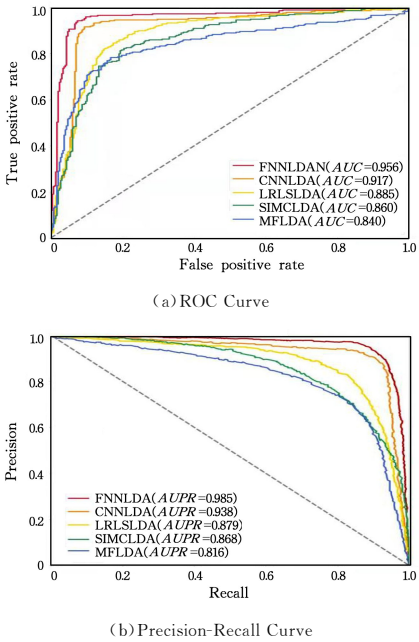


图4 FNNLDA与其他方法的ROC曲线和PR曲线
Fig. 4 ROC and PR curves of FNNLDA and other methods

除了关键指标提升之外,FNNLDA的性能也有了极大改善。首先是训练速度,原有方法的训练速度是670步每毫秒,

FNNLDA 模型只需要不到 160 步每毫秒,时间缩短为了原来的 1/4;其次是内存占用率,原有方法需要占用近 24GB 的内存,FNNLDA 模型只需要占用 4GB 内存,内存占用率大大降低。

召回率表示候选样本中成功发现正样本的比例。正确预测的正测试样本比例越高,召回率值就越大。为了进一步检验本文方法的准确性,本文计算了模型打分前 40,60,⋯,240 个样本的平均召回率,柱形图如图 5 所示。可以看出,在不同候选样本下,FNNLDA 方法超过了其他方法,其性能仍然最好。在前 40,80,120 和 160 个候选样本中,FNNLDA 方法的召回率分别为 82.3%,93.5%,98.6%和 99.1%。随着候选样本的增加,FNNLDA,CNNLDA 和 LRLSLDA 的召回率差距逐渐缩小,直到相等。

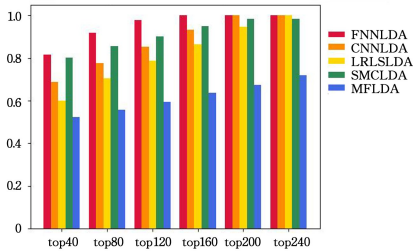


图 5 不同 K 值下的平均召回率

Fig. 5 Average recall rate with different K

4 案例研究

作为世界上非常严重的恶性肿瘤,胃癌是癌症死亡的第二大病因^[23]。为了进一步验证 FNNLDA 模型对未知 LncRNA-疾病关联的预测效果,本文将识别与胃癌相关的 LncRNA 作为案例分析。

FNNLDA 使用数据集中所有的已知 LncRNA-疾病关联建立模型,用于预测 LncRNA 和疾病之间的未知关联。FNNLDA 筛选出与胃癌相关的排名前 10 位的候选 LncRNA,并根据它们的关联分数对候选 LncRNA 进行优先排序,如表 3 所列。

表 3 预测与胃癌有关的前 10 个 LncRNA

Table 3 Top 10 stomach cancer-related candidate LncRNA

排名	LncRNA 名称	相关证据
1	MALAT1	LncRNADisease,Lnc2Cancer
2	XIST	LncRNADisease,Lnc2Cancer
3	NEAT1	LncRNADisease,Lnc2Cancer
4	KCNQ1OT1	literature ^[24]
5	TUG1	LncRNADisease,Lnc2Cancer
6	MIR155HG	unconfirmed
7	MIR17HG	literature ^[25]
8	HOTTIP	LncRNADisease,Lnc2Cancer
9	WT1-AS	LncRNADisease,Lnc2Cancer
10	PCAT1	Lnc2Cancer

经过与数据库中的数据对比,发现筛选出的前 10 个候选 LncRNA 中有 7 个能够在 Lnc2Cancer 和 LncRNADisease 数据库中得到验证,这表明这些 LncRNA 确实与疾病有关。与胃癌有关联的相关候选 LncRNA KCNQ1OT1 和 MIR17HG 在文献中得到验证,这表明它们在胃癌中的表达模式为上调

或者下调。本文对胃癌的案例进行了研究,证明了 FNNLDA 具有良好的预测能力。

结束语 本文提出了基于栈式自编码器和融合神经网络的 LncRNA-疾病关联预测算法 FNNLDA。首先,FNNLDA 利用 LncRNA、miRNA、疾病及其相互作用关系构建 LncRNA 功能相似性、疾病语义相似性以及 LncRNA-疾病数据特征,并将其作为方法的输入数据;其次,基于栈式自编码器 SAE 的特征提取模块和融合神经网络 FNN 的特征提取模块分别学习高阶特征;最后,通过赋予不同的权重计算评分来预测 LncRNA-疾病关系。利用五折交叉验证方法进行模型评估,将所提方法与 4 种其他方法进行对比,评估结果表明 FNNLDA 在 LncRNA-疾病关联预测方面的性能优异。对胃癌疾病进行案例分析,结果也表明 FNNLDA 对 LncRNA-疾病的潜在关联具有可靠的预测效果。

FNNLDA 尽管取得了较好的实验结果,但仍存在一定的局限性。由于已知 LncRNA-疾病相互作用关系矩阵稀疏,可用且可靠的数据集有限,同时存在着大量的噪声,从而影响了模型的性能,因此下一步将通过掌握更多被验证的数据源来提高数据集的可信度和完整性,改进现有模型,提升整体算法框架的预测性能。

参 考 文 献

[1] 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.

[2] 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,2017,18(4):558-576.

[3] ZHUO W,LIU Y,LI S et al. Long Noncoding RNA GMAN, Up-regulated in Gastric Cancer Tissues,Is Associated with Metastasis in Patients and Promotes Translation of Ephrin A1 by Competitively Binding GMAN-AS[J]. Gastroenterology,2019,156(3):676-691.

[4] LI G,KRYCZEK I,NAM J,et al. LIMIT is an immunogenic lncRNA in cancer immunity and immunotherapy [J]. Nat Cell Biol,2021,23(5):526-537.

[5] HAJJARI M,SALAVATY A.HOTAIR;an oncogenic long non-coding RNA in different cancers [J]. Cancer Biology & Medicine,2015,12(1):1-9.

[6] CHEN G,WANG Z,WANG D,et al. LncRNADisease;a database for long-non-coding RNA-associated diseases[J]. Nucleic Acids Research,2013,41(DI):D983-D986.

[7] NING S,ZHANG J,WANG P,et al. Lnc2Cancer;a manually curated database of experimentally supported lncRNAs associated with various human cancers[J]. Nucleic Acids Research,2016,44(DI):D980-D985.

[8] LI Y,QIU C,TU J,et al. HMDD v2. 0;a database for experimentally supported human microRNA and disease associations [J]. Nucleic Acids Research,2014,42(DI):D1070-D1074.

[9] FU G,WANG J,DOMENICONI C,et al. Matrix factorization-based data fusion for the prediction of lncRNA-disease associa-

tions[J]. *Bioinformatics*, 2018, 34(9):1529-1537.

[10] LU C, YANG M, LUO F, et al. Prediction of lncRNA-disease associations based on inductive matrix completion[J]. *Bioinformatics*, 2018, 34(19):3357-3364.

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

[12] XIE G, MENG T, LUO Y, et al. SKF-LDA;similarity kernel fusion for predicting lncRNA-Disease Association[J]. *Molecular Therapy-Nucleic Acids*, 2019, 18:45-55.

[13] XUAN P, CAO Y, ZHANG T, et al. Dual Convolutional Neural Networks with Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes [J]. *Frontiers in Genetics*, 2019, 10:416-426.

[14] 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):1012-1022.

[15] YANG Q, LI X. BIGAN;LncRNA-disease association prediction based on bidirectional generative adversarial network[J]. *BMC Bioinformatics*, 2021, 22(1):357-373.

[16] LIU G, BAO H, HAN B. A Stacked Autoencoder-Based Deep Neural Network for Achieving Gearbox Fault Diagnosis[J]. *Mathematical Problems in Engineering*, 2018, 2018(PT. 9):1-10.

[17] ZENG H, EDWARDS M D, LIU G, et al. Convolutional neural network architectures for predicting DNA-protein binding[J]. *Bioinformatics*, 2016, 32(12):i121-i127.

[18] HOCHESTER S, SCHMIDHUBER J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8):1735-1780.

[19] CHEN X, YAN C C, LUO C, et al. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity[J]. *Scientific Reports*, 2015, 5:11338.

[20] WANG D, WANG J, LU M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. *Bioinformatics*, 2010, 26(13):1644-1650.

[21] FAN Y, SIKLENKA K, ARORA S K, et al. miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis[J]. *Nucleic acids research*, 2016, 44(W1):135-141.

[22] LI J H, LIU S, ZHOU H, et al. starBase v2. 0;decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data [J]. *Nucleic Acids Research*, 2014, 42(DI):D92.

[23] FERLAY J, SHIN H R, BRAY F, et al. Estimates of worldwide burden of cancer in 2008:GLOBOCAN 2008[J]. *International Journal of Cancer*, 2010, 127(12):2893-2917.

[24] SUN X, XIN Y, WANG M, et al. Overexpression of long non-coding RNA KCNQ1OT1 is related to good prognosis via inhibiting cell proliferation in non-small cell lung cancer[J]. *Thoracic Cancer*, 2018, 9(5):523-531.

[25] BAHARI F, EMADI-BAYGI M, NIKPOUR P. miR-17-92 host gene, uderexpressed in gastric cancer and its expression was negatively correlated withthe metastasis[J]. *Indian Journal of Cancer*, 2015, 52(1):22-25.



LI Qiaojun, born in 1983, master, associate professor, is a member of China Computer Federation. Her main research interests include neural network and big data mining.

(责任编辑:柯颖)