Hans Journal of Computational Biology 计算生物学, 2023, 13(4), 35-44
Published Online December 2023 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2023.134004

Hans汉斯

SRF-LDA：基于堆叠集成学习的LncRNA与疾病关联预测方法

孙　捷[1]，谭者斌[2]

[1]大连交通大学理学院，辽宁　大连
[2]大连交通大学软件学院，辽宁　大连

收稿日期：2024年4月17日；录用日期：2024年5月21日；发布日期：2024年5月28日

摘　要

长链非编码RNA (lncRNA)是一类长度大于200 nt的非编码RNA，是非编码基因组的重要组成部分。大量实验证实，lncRNA与人类疾病的发生发展密不可分，但除了一小部分的lncRNA与人类疾病关系已知之外，大多数的lncRNA与人类疾病的关系仍然有待研究，因此准确识别与疾病有关的lncRNA有助于研究lncRNA在疾病中的作用机制，探索治疗疾病的新方法。在本研究中，为了提高对LDA的预测能力，我们实现了一种基于堆叠集成学习的LDA预测模型(简称SRFLDA)。在SRFLAD中，第一部分通过整合lncRNA的K-mer、疾病的高斯相互作用谱核相似性及已知lncRNA-疾病关联(LDA)三种类型的特征作为融合特征输入模型。第二部分使用堆叠集成学习策略通过组合多个不同参数的随机森林分类器作为基模型进行特征分类，并使用支持向量机作为元模型对随机森林的分类结果进行组合优化，从而得到更准确、鲁棒的LDA预测结果。第三部分通过十倍交叉验证对模型进行训练评价。结果表明该方法在预测LDA方面具有较好的性能，平均AUC的值为0.9246，平均AUPR值为0.9166，预测效果优于其他几种现有的LDA预测模型。

关键词

lncRNA，疾病，lncRNA-疾病关联，随机森林，变量重要性，特征选择，支持向量机

SRF-LDA: A Stacking-Based Ensemble Learning Model for LncRNA-Disease Association Prediction

Jie Sun[1], Zhebin Tan[2]

[1]College of Science, Dalian Jiaotong University, Dalian Liaoning
[2]College of Software, Dalian Jiaotong University, Dalian Liaoning

文章引用: 孙捷, 谭者斌. SRF-LDA: 基于堆叠集成学习的 LncRNA 与疾病关联预测方法[J]. 计算生物学, 2023, 13(4): 35-44.
DOI: 10.12677/hjcb.2023.134004

Received: Apr. 17th, 2024; accepted: May 21st, 2024; published: May 28th, 2024

Abstract

Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs larger than 200 nt in length and are an important component of the non-coding genome. A large number of experiments have confirmed that lncRNA is inseparable from the occurrence and development of human diseases, but except for a small number of lncRNAs with human diseases, the relationship between most lncRNAs and human diseases still needs to be studied, so accurate identification of lncRNAs related to diseases is helpful to study the mechanism of action of lncRNAs in diseases and explore new ways to treat diseases. In this study, in order to improve the prediction ability of LDA, we implemented an LDA prediction model based on stacked ensemble learning (SRFLDA). In SRFLAD, the first part is used to integrate three types of features of lncRNA, namely K-mer, Gaussian interaction spectral nuclear similarity of disease, and known lncRNA-disease association (LDA), as fusion features as input into the model. In the second part, the stacked ensemble learning strategy is used to classify features by combining random forest classifiers with multiple different parameters as the base model, and the support vector machine is used as a metamodel to combine and optimize the classification results of the random forest, so as to obtain more accurate and robust LDA prediction results. The third part is to evaluate the training of the model through tenfold cross-validation. The results show that the proposed method has good performance in predicting LDA, with an average AUC value of 0.9246 and an average AUPR value of 0.9166, which is better than that of several other existing LDA prediction models.

Keywords

LncRNA, Disease, LncRNA-Disease Association, Random Forest, Variable Importance, Feature Selection, Support Vector Machine

1. 引言

长链非编码 RNA (long non-coding RNA, lncRNA)是一类长度大于 200 个核苷酸[1]、广泛存在的、但不具备蛋白质编码能力的分子。大量实验证实，lncRNA 与人类疾病的发生发展密不可分[2]，可在表观遗传、顺式或反式转录及转录后水平上调控基因表达，参与 X 染色体沉默、基因组印记以及染色质修饰、转录激活、转录干扰、核内运输等生物学进程[3] [4]。大量定位在染色质上的 lncRNA 可与蛋白质相互作用，促进或抑制蛋白质在目标 DNA 区域的结合活性[5]。NONO/P54nrb 和 PSPC1 调节细胞核亚结构 paraspeckle 的形成[6] [7] [8]。此外，研究发现 lncRNA 与 12 种癌症如前列腺癌、乳腺癌、肺癌等密切相关，同时在心衰患者中有 18,480 种 lncRNA 表达模式有所差异，表明 lncRNA 在心衰类型上的反映能力优于 miRNA。与此同时，研究表明一些 lncRNA 在心衰患者接受左心室辅助装置治疗前后的表达量也有显著变化。另外，在房颤研究中发现，房颤患者与健康人群循环血中 lncRNA 表达存在较大差异，其中有 177 种 lncRNA 表达量超过 2 倍。此外，lncRNA 也与免疫系统缺陷等疾病如系统性红斑狼疮、类风湿关节炎密切相关[9] [10]。

lncRNA 疾病关联预测的计算方法大致上可以分为两类：基于传统计算方法和基于深度学习的方法。对于传统方法，Chen 等[11]提出了一种拉普拉斯正则化最小二乘模型来预测潜在的 lncRNA 疾病关联(LRSLDA)，还有随机游走算法的应用如 RWRHLD 模型[12]等。对于深度学习的方法大致上分为两大类：图神经网络(GCNs) [13]和卷积神经网络(CNN) [14] [15] [16]。文献[17]提出了一种深度学习方法 iLncRNAdis-FB。在类似疾病往往与类似 lncRNA 相关的假设下，Chen 等人将实验支持的 LDA 与 lncRNA 表达谱相结合，提出了基于拉普拉斯正则化最小二乘的 LDA 预测模型(LRLSLDA) [18]，这是该领域第一个计算模型。Xie 等人通过融合 DSS 和余弦相似度、lncRNA 表达相似度和余弦相似度，实现了基于相似核融合的 LDA 预测模型(SFK-LDA) [19]。虽然上述方法取得了优异的效果，但也存在一定的局限性。在本研究中，我们提出了一种基于堆叠集成学习的 lncRNA 与疾病关联预测方法 SRFLDA。与现有方法不同的是，我们的训练模型使用堆叠集成学习策略通过组合多个不同参数的随机森林分类器作为基模型进行特征分类，并使用支持向量机作为元模型对随机森林的分类结果进行组合优化，从而得到更准确、鲁棒的 LDA 预测结果。

2. 材料与方法

2.1. 数据集

建立一个可靠全面的 lncRNA-疾病关联数据集对于准确预测潜在的疾病-lncRNA 关联非常重要。在本文中，已知的 lncRNA-疾病关联来自 lncRNA Disease 数据库[20]。在剔除部分异常数据和重复数据后，我们最终得到 240 个 lncRNA，并且在 NCBI 中查找并截取了该 240 个 lncRNA 序列片段长度设置为 200 bp，432 种疾病，1420 个已知 lncRNA-疾病关联样本。序列长度的设置取决于序列的平均长度，若长度过长或过短都会导致一定的信息缺失和特征缺失，综合考虑选择截取长度为 200 bp 的 lncRNA 序列。我们将已知的相关样本对标记为正样本，其他的标记为未观察到的样本对。假设疾病和 lncRNA 的数量分别为 N_d 和 N_l，给定疾病 $i \in [0, N_d]$，lncRNA $j \in [0, N_l]$，则样本对 (i, j) 的关联可以用

$$A(i,j) = \begin{cases} 1, & \text{if disease } i \text{ is associated with lncRNA } j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

得到邻接矩阵 A，若实验证实疾病 i 和 LncRNA j 有关联，则 $A(i, j)$ 为 1，否则为 0。因为正样本数量远远少于未观察到的样本数量，导致关联矩阵十分稀疏，统计结果也不可靠。所以我们仿照[21] [22]所采用的策略，从这些未观察到的样例中随机抽取样本，标记为负样本进行训练。其中随机抽取的样本与正样本数量相同。

2.2. 可视化分析

为了验证本文中所建立的数据集的可靠性以及真实性，我们对收集的序列进行了独热编码，通过不同的降维方式进行可视化的数据分析；我们可以发现原本只通过序列无法进行分类，需要重新提取特征。继而可说明该数据集中的序列信息是不能线性分类的，那么利用提取到的不同的非线性的 lncRNA 与疾病的特征向量输入到五个不同参数的基分类器中进行训练，那么预测得到的结果是真实独立的，也就是说本文中所构建的独立数据集是可靠的。进行可视化数据集分析增强了数据集实用性的说服力。因此，开发有效的序列表示模型和包括深度神经网络在内的非线性建模对识别人类细胞系中的这些序列是非常必要的。在所用的四种降维方法中，横纵坐标并不代表原始数据空间中的特定特征或维度，而是新的、由降维算法生成的维度，旨在保留原始数据的重要结构或特征。特别是对于 t-SNE，这些坐标纯粹是为了可视化目的，它们并不对应于任何可解释的物理或统计属性。ICA 和 FA 的坐标代表了数据中的独立

或潜在因素，而 PCA 的坐标代表了数据中的方差方向。可视化结果如图 1 所示：

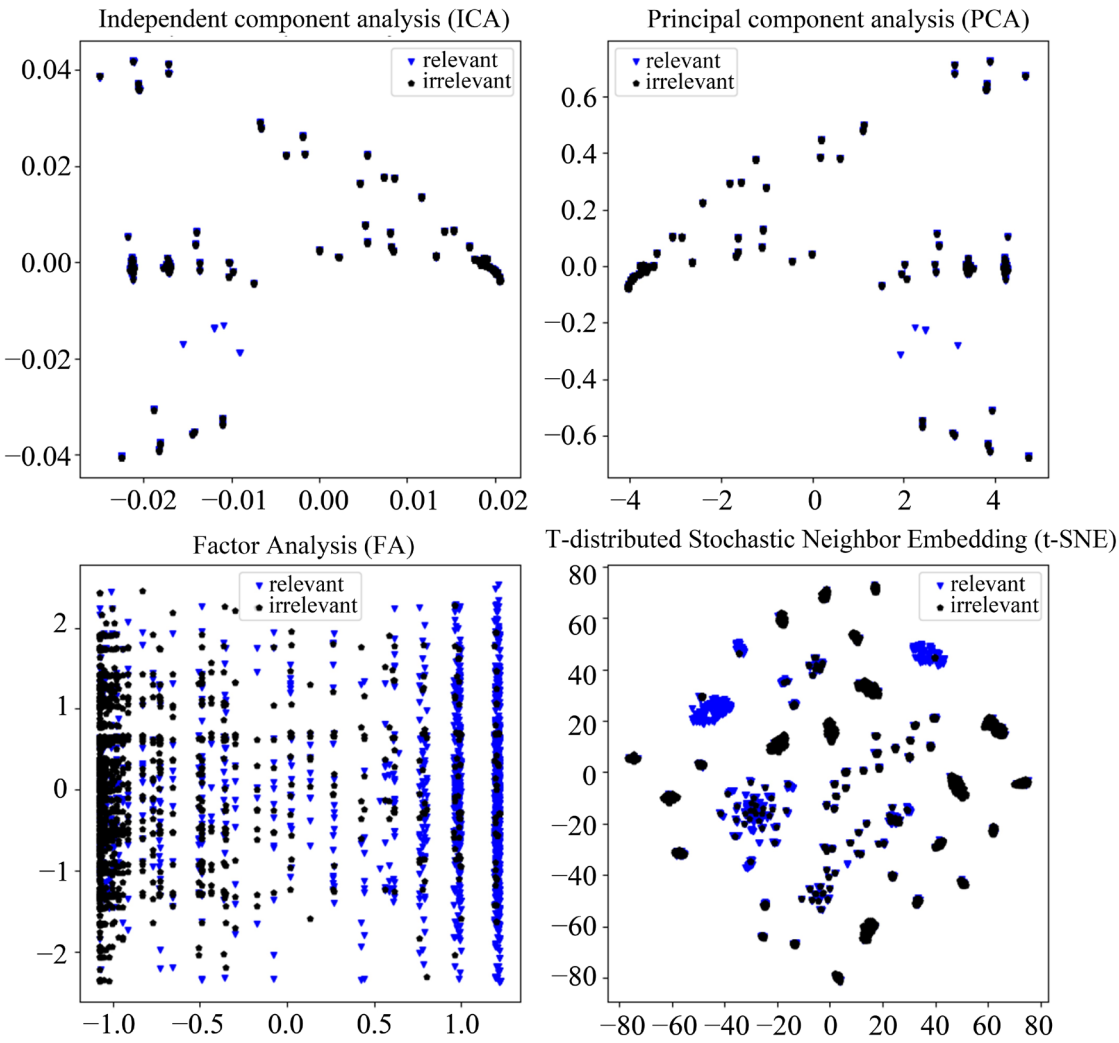

Figure 1. Dataset visualization analysis map
图 1. 数据集可视化分析图

2.3. 特征提取

特征提取是机器学习中极其重要的部分，它可以帮助机器识别模型中的相关特征，提高模型的预测性能。每个 lncRNA 序列都被认为是一个用生物语言写成的句子，这表明可以语义地解释相应的生物功能。所以在本研究中提取了 lncRNA 的 K-mer 特征集和疾病的高斯相互作用谱和相似性 K-mer 特征集是所有的 lncRNA 中的重要特征的集合，对于识别 lncRNA 之间的相似性或者计算 lncRNA 与疾病之间的关联性提供了重要的信息支持和数据支撑，疾病的高斯相互作用谱和相似性的计算是利用了 lncRNA 与疾病的已知关联特征进行计算得到的，所以疾病的高斯相互作用谱和相似性矩阵可以很好的反映所有疾病的特征，即可以利用疾病的高斯相互作用谱和相似性矩阵可以有效关联所有疾病的相似性信息。

2.4. K-mer 特征提取

K-mer 是指将 reads 迭代分成包含 K 个碱基的序列，一般长短为 L 的 reads 可以分成 L-K + 1 个

K-mers。K-mer 的用途，是用于基因组从头组装前的基因组调查，评估基因组的大小。基因组大小可以由(总 K-mer 数量)/(K-mer 期望测序深度)来估计。K-mer 是一种简单而有效的 RNA/DNA 序列特征提取方法，它代表 k 个相邻核酸的出现频率[21] [22]。该方法已成功应用于生物信息学许多方面。它采用长度为 k，步长为 1 的滑动窗口，根据滑动窗口截取 lncRNA 序列。利用下式计算 K-mer (k = 3)的特征编码：

$$f(x)=\frac{N(t)}{N}, t\in(AAA, AAC, AAG, \cdots, UUU) \tag{2}$$

(2)式中 $N(t)$ 为 t 型 K-mer 的个数，N 为核苷酸序列的长度。在本研究中，利用序列中 K-mer (k = 1,2,3)的所有可能频率来表示特征的维度。对于每一个 k，我们得到 4^K 维度特征，从而得到总共 84 个维度特征。

2.5. 疾病的高斯相互作用谱和相似性

基于相似的疾病往往表现出与 lncRNA 有相似的相互作用和非相互作用的假设，我们从已知的 lncRNA-疾病关联中构建了疾病的高斯相互作用谱核相似性。首先，我们将疾病 $d(i)$ 的相互作用谱 $IP(d(i))$ 标记为二进制向量；这个二进制向量表示疾病 $d(i)$ 与已知疾病-lncRNA 关联数据集中的每个 lncRNA 之间是否存在关联，即邻接矩阵 A 的第 i 行。然后，我们引入高斯核函数对疾病的相互作用谱进行处理。疾病 $d(i)$ 和 $d(j)$ 的核定义如下，并将其作为这两种疾病之间的相似度评分：

$$KD(d(i), d(j))=\exp(-\gamma_d \left\| IP(d(i))-IP(d(j)) \right\|^2 \tag{3}$$

3. SRF-LDA

SRF-LDA 的模型流程图如图 2 所示。首先，从各大公共数据库中整合了由实验支持的已知 LDA 以及相关 lncRNA 的序列信息，并利用已知的 lncRNA 与疾病关联数据疾病计算出疾病的高斯相互作用谱核相似度矩阵，再利用 K-mer 特征提取得到 lncRNA 特征矩阵，将疾病与 lncRN 的特征向量进行加和作为融合特征输入模型。然后，SRFLDA 训练了五个不同参数的随机森林作为基模型，并使用支持向量机算法作为元模型。最后，在基模型和元模型的基础上构造堆叠集成学习模型，将数据分别输入到各个基分类器中进行训练，再进行多次不加重复的训练之后，再进行十折交叉实验进行预测性能，将初次训练后的特征集输入到元分类器中进行二次训练及预测，得到最终的预测结果。我们在 10 倍交叉验证下对 SRFLDA 的性能进行了评价。

堆叠集成学习框架：

SRFLDA 使用堆叠集成学习框架来预测潜在的 LncRNA-疾病关联。本文提出的堆叠集成学习模型分为基分类器和元分类器两部分。堆叠集成学习算法通过将多个基分类器的预测结果作为输入，再训练一个元分类器来进行最终的预测。本研究所提出的用于 lncRNA 与疾病关联的预测器称为 SRFLDA，其中“s”代表“Stacking”，“RF”代表“Random Forest as base-classifier”。该模型使用 10 倍交叉验证进行训练并评价模型性能。在 SRFLAD 中，第一部分中，首先通过在各大公共数据库中整合实验证明的 lncRNA-疾病的已知关联和相关 lncRNA 的序列信息建立起新的数据集；其次，利用 K-mer 提取特征方法对 lncRNA 序列进行特征提取，利用已知的 lncRNA 与疾病关联信息计算出疾病的高斯相互作用谱核相似性矩阵；再次，将 lncRNA 的特征向量与疾病的高斯相互作用谱和相似性向量进行融合得到新的特征向量，将其作为融合特征输入模型。第二部分使用堆叠集成学习策略通过组合多个不同参数的基分类器进行数据集训练，得到 lncRNA 与疾病关联的所有预测可能性并进行特征分类，所有的基分类器在训练预

测之后还要进行十折交叉验证；得到的预测结果使用元模型再次进行组合优化，从而得到更准确、鲁棒的 LDA 预测结果。第三部分通过十倍交叉验证对模型进行训练评价。

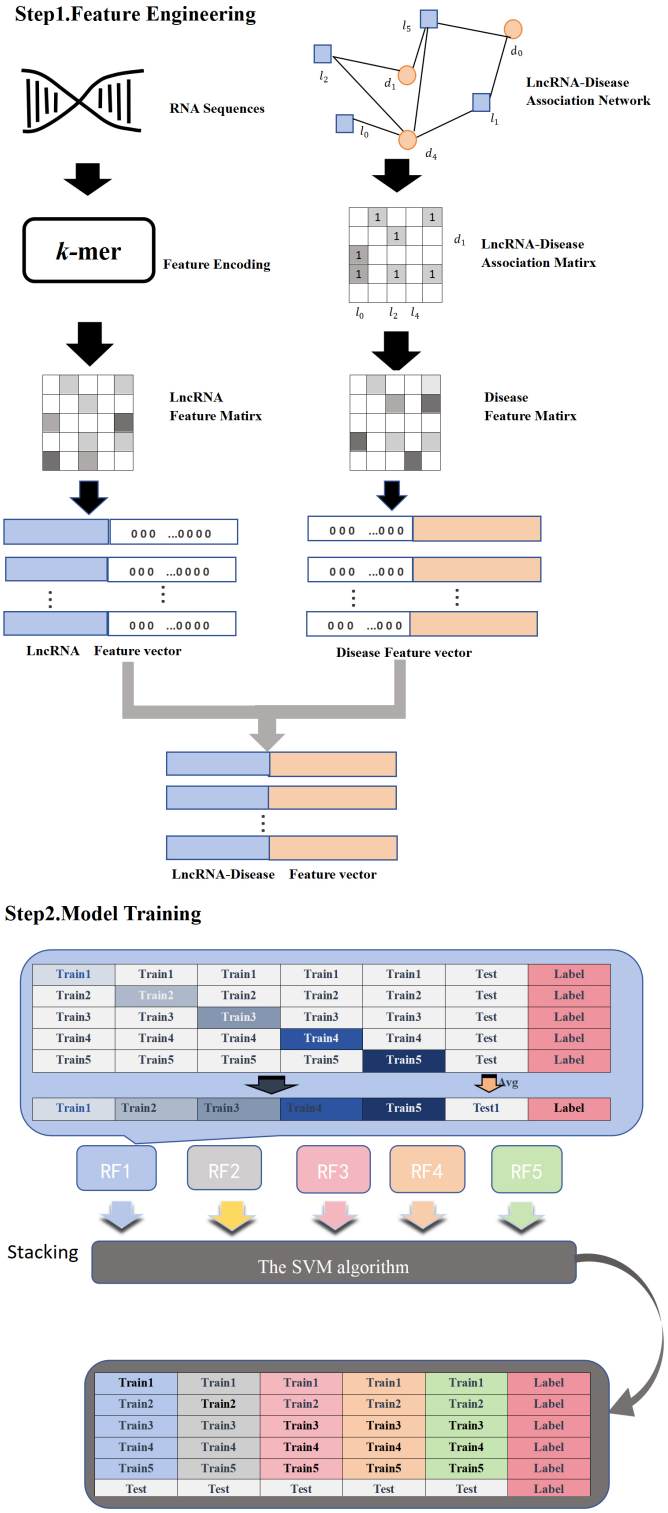

Figure 2. Structural flow chart of the SRF-LDA
图 2. SRF-LDA 的结构流程图

4. 结果讨论

4.1. 性能评估

在本文中我们采用十折交叉验证来评估计算模型的预测性能。为了评估 SRF-LDA 在我们的数据集上的有效性，1420 个已知 lncRNA-疾病关联样本被用来训练我们的计算模型，而未知的 lncRNA-疾病关联不参与我们的训练过程。我们使用了 10 倍交叉验证，这需要评估 10 个不同的测试子集的平均性能。在本研究中，我们使用受试者工作特征曲线下面积(AUC)来量化模型的性能。如图 3 所示：

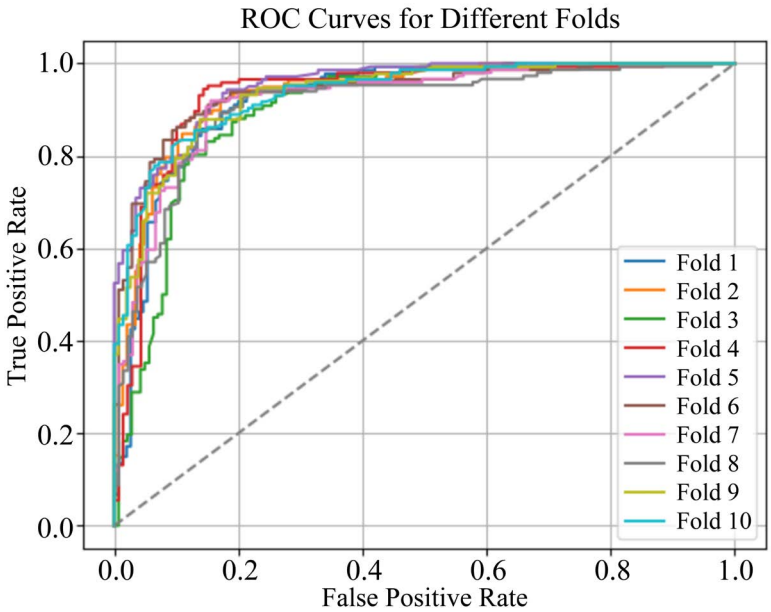

Figure 3. ROC plot of SRF-LDA
图 3. SRF-LDA 的 ROC 曲线图

4.2. 与其他方法比较

为了评估 SRFLDA 的性能，我们将其与目前最先进的 LDA 预测模型进行了比较，即 GCRFLDA、MFLDA、SIMCLDA、BiWalkLDA、BiGAN。为了提高比较实验的说服力，我们的比较模型涵盖了基于网络的方法、基于机器学习的方法和基于深度学习的方法。数据集在各模型下 10 倍 CV 中的性能显示见表 1。

Table 1. The AUC and AUPR of the different LDA prediction models
表 1. 不同 LDA 预测模型的 AUC 和 AUPR

Method	AUC	AUPR
SRF-LDA	0.9246	0.9166
GCRFLDA	0.8120	0.7806
MFLDA	0.7223	0.7895
SIMCLDA	0.7836	0.8203
BiWalkLDA	0.8435	0.8727
BiGAN	0.5246	0.5029

从表 1 可以看出，在验证数据集中，SRF-LDA 的性能优于其他五种方法。

4.3. 参数设置

为了深入了解 SRF-LDA 的性能，我们针对 lncRNA-疾病数据集上的几种最新方法对其进行了基准测试。对于 SRF-LDA 的实现，我们对随机森林中的最优参数进行了调整，设置了 $n=100$， max_ depth = 4 和随机状态 = 100。关于各算法参数调整的详细信息见表 2。

Table 2. The parameters of each machine learning algorithm
表 2. 各机器学习算法的参数

	n_estimators	max_depth	random_stat
Random Forest-base_model1	50, 60, 70, 80, 90, 100	3, 4, 5, 6	100
Random Forest-base_model2	50, 60, 70	3, 4, 5, 6	100
Random Forest-base_model3	50, 60, 70, 80, 90, 100	5, 6, 7, 8,	100
Random Forest-base_model4	50, 60, 70	5, 6, 7, 8	100
Random Forest-base_model5	80, 90, 100	7, 8, 9, 10	100
SVM-meta_model	80, 90, 100	3, 4, 5, 6	100

5. 讨论与结论

近年来，lncRNA 相关研究在国内外的热度居高不下，长非编码 RNA (lncRNA)在生命活动中发挥着重要作用，其中包括剂量补偿效应、表观遗传调控、细胞周期调控和细胞分化调控。一个典型的例子是 X 染色体剂量补偿机制，其中 Xist-lncRNA 在一条长达上亿碱基的染色体上调控着数百个基因的沉默。在表达调控方面，ceRNA 机制涉及 lncRNA、miRNA 和 mRNA 之间的交互作用，而转录因子的研究可以帮助我们了解 lncRNA 与转录因子之间的调控机制以及染色质重塑的作用。另外，一些 lncRNA，如 lncND5、lncND6 和 lncCytb，在线粒体基因组中编码，并且能够与 mRNA 相互配对，从而影响它们的稳定性。通过研究这些 lncRNA 的功能，我们可以更深入地了解它们在细胞内的重要作用。也与包括癌症在内的多种疾病息息相关，也正因此成为生命科学领域的研究热点。lncRNA 在发育和基因表达中发挥复杂精确的调控功能极大地解释了基因组复杂性之难题，同时也为人们从基因表达调控网络的维度来认识生命体的复杂性开启新的天地。但是目前对 lncRNA 的认知可以说还处在初级阶段，前路漫漫，还有许多需要探索。因此，预测 lncRNA 与疾病的新型关联将有助于生物学家对人类未知疾病的研究与探索。此外，它还有助于诊断、预防和治疗人类疾病。很多研究人员已经开发了一些计算方法来推断 lncRNA 与疾病的关联。

在这篇文章中提出了一个新的堆叠集成学习的方法。第一部分将提取的 lncRNA 特征矩阵、疾病的高斯相互作用谱和相似性矩阵和已知 lncRNA-疾病关联矩阵作为模型输入。再使用堆叠集成学习模型进行训练和分类。本文的堆叠集成学习模型分为基分类器和元分类器两部分。堆叠集成学习算法通过将预训练后的特征集输入到多个不同参数的基分类器中进行特征分类，再训练一个元分类器对基分类器的输出进行组合优化。经过对比试验，我们发现随机森林在特征分类方面具有良好的性能。第一部分，我们对随机森林设置不同的参数来构造 5 个模型作为堆叠集成学习模型的基分类器。我们使用原始训练数据集来训练这 5 个不同的基分类器，以便捕捉到数据不同方面的特征，每个基分类器都会生成对训练数据的预测结果；第二部分，通过元分类器对比实验，选择支持向量机算法作为元分类器。第三部分，模型使用十倍交叉验证进行训练评价。模型预测性能得到提高大致归因于以下几个因素，这也是我们将 K-mer

和高斯核结合起来进行潜在疾病相关 lncRNA 预测的原因。首先，可以整合已知的疾病-lncRNA 关联和 lncRNA 序列的特征矩阵，以捕获疾病与 lncRNA 之间的潜在关联。其次，将不同参数不同类型的分类器进行组合，可以显著提高分类器的预测能力。总的来说，我们的方法比传统的生物实验更具成本效益。与基于单一特征或单一分类器的模型相比，SRFLDA 显著提高了全局特征提取和特征分类的性能。我们将 SRFLDA 与现有方法的性能进行了比较和分析，结果表明 SRFLDA 在预测 lncRNA-疾病潜在关联方面比现有方法具有更好的性能。在未来的研究中，我们将考虑引入多种数据融合和深度学习方法，从 lncRNA 序列中提取更多的潜在信息，以便更好地预测 lncRNA-疾病潜在的关联。

参考文献

[1] Yang, G.D., Lu, X.Z. and Yuan, L.J. (2014) LncRNA: A Link between RNA and Cancer. *Biochimica et Biophysica Acta* (*BBA*)-*Gene Regulatory Mechanisms*, **1839**, 1097-1109. https://doi.org/10.1016/j.bbagrm.2014.08.012

[2] Wapinski, O. and Chang, H.Y. (2011) Long Noncoding RNAs and Human Disease. *Trends in Cell Biology*, **21**, 354-361. https://doi.org/10.1016/j.tcb.2011.04.001

[3] Panwar, B., Arora, B. and Raghava, G.P. (2014) Prediction and Classification of ncRNAs Using Structural Information. *BMC Genomics*, **15**, Article No. 127. https://doi.org/10.1186/1471-2164-15-127

[4] Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X. and Li, T. (2013) Computational Prediction of Associations between Long Non-Coding RNAs and Proteins. *BMC Genomics*, **14**, Article No. 651. https://doi.org/10.1186/1471-2164-14-651

[5] Saldana-Meyer, R., *et al*. (2019) RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Molecular Cell*, **76**, 412-422e415. https://doi.org/10.1016/j.molcel.2019.08.015

[6] Chen, L.L. and Carmichael, G.G. (2009) Altered Nuclear Retention of mRNAs Containing Inverted Repeats in Human Embryonic Stem Cells: Functional Role of a Nuclear Noncoding RNA. *Molecular Cell*, **35**, 467-478. https://doi.org/10.1016/j.molcel.2009.06.027

[7] Clemson, C.M., *et al*. (2009) An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Molecular Cell*, **33**, 717-726. https://doi.org/10.1016/j.molcel.2009.01.026

[8] Sasaki, Y.T., Ideue, T., Sano, M., Mituyama, T. and Hirose, T. (2009) MENepsilon/Beta Noncoding RNAs Are Essential for Structural Integrity of Nuclear Paraspeckles. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 2525-2530. https://doi.org/10.1073/pnas.0807899106

[9] Salmena, L., Poliseno, L., Tay, Y., Kats, L. and Pandolfi, P.P. (2011) A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell*, **146**, 353-358. https://doi.org/10.1016/j.cell.2011.07.014

[10] Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z. and Shen, F. (2019) Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *International Journal of Molecular Sciences*, **20**, Article No. 5573. https://doi.org/10.3390/ijms20225573

[11] Chen, X. and Yan, G.Y. (2013) Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics*, **29**, 2617-2624. https://doi.org/10.1093/bioinformatics/btt426

[12] Zhou, M., Wang, X., Li, J., *et al*. (2013) Prioritizing Candidate Disease-Related Long Non-Coding RNAs by Walking on the Heterogeneous lncRNA and Disease Network. *Molecular BioSystems*, **11**, 760-769. https://doi.org/10.1039/C4MB00511B

[13] Xuan, P., Pan, S., Zhang, T., Liu, Y. and Sun, H. (2019) Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations. *Cells*, **8**, Article No. 1012. https://doi.org/10.3390/cells8091012

[14] Xuan, P., Cao, Y., Zhang, T., Kong, R. and Zhang, Z. (2019) Dual Convolutional Neural Networks with Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes. *Frontiers in Genetics*, **10**, Article No. 416. https://doi.org/10.3389/fgene.2019.00416

[15] Zeng, M., Lu, C., Fei, Z., Wu, E., Li, Y., Wang, J. and Li, M. (2020) Dm-flda: A Deep Learning Framework for Predicting incRNA-Disease Associations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **18**, 2353-2363. https://doi.org/10.1109/TCBB.2020.2983958

[16] Zhang, Y., Ye, F. and Gao, X. (2021) MCA-Net: Multi-Feature Coding and Attention Convolutional Neural Network for Predicting lncRNA-Disease Association. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **19**, 2907-2919. https://doi.org/10.1109/TCBB.2021.3098126

[17] Wei, H., Liao, Q. and Liu, B. (2020) iLnRNADIS-FB: Identify lncRNA-Disease Associations by Fusing Biological Feature Blocks through Deep Neural Network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **18**, 1946-1957. https://doi.org/10.1109/TCBB.2020.2964221

[18] Lan, W., Li, M., Zhao, K., *et al*. (2017) LDAP: A Web Server for lncRNA-Disease Association Prediction. *Bioinformatics*, **33**, 458-460. https://doi.org/10.1093/bioinformatics/btw639

[19] Xie, G.B., Meng, T.F., Luo, Y. and Liu, Z.G. (2019) SKF-LDA: Similarity Kernel Fusion for Predicting lncRNA-Disease Association. *Molecular Therapy Nucleic Acids*, **18**, 45-55. https://doi.org/10.1016/j.omtn.2019.07.022

[20] Chen, G., Wang, Z.Y., Wang, D.Q., Qiu, C.X., Liu, M.X., Chen, X., Zhang, Q.P., Yan, G.Y. and Cui, Q.H. (2013) LncRNA Disease: A Database for Long-Non-Coding RNA-Associated Diseases. *Nucleic Acids Research*, **41**, D983-D986. https://doi.org/10.1093/nar/gks1099

[21] Fu, X., Cai, L., Zeng, X., *et al.* (2020) StackCPPred: A Stacking and Pairwise Energy Content-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency. *Bioinformatics*, **36**, 3028-3034. https://doi.org/10.1093/bioinformatics/btaa131

[22] Liang, X., Li, F., Chen, J., *et al*. (2021) Large-Scale Comparative Review and Assessment of Computational Methods for Anti-Cancer Peptide Identification. *Briefings in Bioinformatics*, **22**, bbaa312. https://doi.org/10.1093/bib/bbaa312