Hans Journal of Computational Biology 计算生物学, 2021, 11(3), 37-47
Published Online September 2021 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2021.113005

Hans汉斯

利用深度稀疏自动编码器预测miRNA与疾病的关联关系

张树尧，刘立伟

大连交通大学，辽宁 大连

收稿日期：2022年3月12日；录用日期：2022年4月12日；发布日期：2022年4月20日

摘　要

microRNA (miRNA)是一类具有调控功能的内源性非编码RNA，在各种生物的生命发展过程中发挥着关键作用。许多生物学实验研究证明，miRNA与人类疾病密切相关，包括疾病的发生、流行、传播、诊断和治疗。但是生物实验既昂贵又耗时。因此，有效的计算模型变得越来越重要。在这项研究中，我们将稀疏性嵌入到现有的自动编码器中，形成一个新的计算框架(DSAEMDA)。首先，通过两种方式计算疾病语义相似度得到两个疾病语义相似度矩阵，计算miRNA功能相似度得到miRNA功能相似度矩阵，分别同疾病和miRNA的高斯相互作用谱核相似度矩阵融合。然后嵌入高维空间提取疾病和miRNA高维表达，利用已被证明的miRNA-疾病关联数据训练我们的深度稀疏自动编码器。此外，通过计算未知关系对的重建误差可用于预测某些疾病相关miRNA的相关值。实验结果表明，DSAEMDA可以有效地预测疾病相关的miRNA且准确率高。

关键词

microRNA，疾病，关联预测，嵌入学习，深度稀疏自编码器

Predicting miRNA-Disease Associations through Deep Sparse Autoencoder

Shuyao Zhang, Liwei Liu

Dalian Jiaotong University, Dalian Liaoning

Received: Mar. 12th, 2022; accepted: Apr. 12th, 2022; published: Apr. 20th, 2022

Abstract

MicroRNA (miRNA) is a series of endogenous non-coding RNAs with regulatory functions that take

文章引用: 张树尧, 刘立伟. 利用深度稀疏自动编码器预测 miRNA 与疾病的关联关系[J]. 计算生物学, 2021, 11(3): 37-47. DOI: 10.12677/hjcb.2021.113005

a key part in the life development of various organisms. Many biological experimental studies have proved that miRNA is closely allied to human diseases, including the occurrence, prevalence, transmission, diagnosis and treatment of diseases. But biological experiments are expensive as well as time-consuming. Therefore, efficient computational models to avoid the above problems are becoming increasingly necessary to identify potential miRNA-disease associations. In this study, we add sparsity to the existing autoencoder to form a new computational framework named DSAEMDA (deep sparse autoencoder miRNA-disease association). First, two disease semantic similarity matrices were obtained by computing disease semantic similarity in two ways, and miRNA functional similarity matrices were obtained by computing miRNA functional similarity, which were fused with the kernel similarity matrix of Gaussian interaction spectrum of disease and miRNA respectively. Then, disease and miRNA high-dimensional expressions were extracted by embedding in high-dimensional space, and our deep sparse autoencoder was trained by using proven miRNA-disease association data. In addition, the reconstruction errors of unknown relationship pairs can be used to predict the correlation values of some disease-related miRNAs. The experimental results showed that DSAEMDA could effectively predict disease-related miRNA with high accuracy.

Keywords

microRNA, Disease, Association Prediction, Embedding Study, Deep Sparse Autoencoder

1. 引言

MicroRNAs (miRNAs)广泛存在于哺乳动物细胞中。它们是长度约为 22 nt 的内源性非编码 RNA。miRNA 参与转录后基因表达的调控，从而调控细胞生长和组织分化，与生命过程中的发育和疾病有关[1] [2] [3]。研究表明，最初的 miRNA 是在 20 年前被发现。从那时起，成千上万的 miRNAs 从一系列物种中被发现[4] [5]。此外，越来越多的研究表明，miRNA 在生物生命发育过程的多个阶段中发挥着关键作用[6]，例如细胞生长、增殖[7]、发育[8]，分化[9]，凋亡[10]，老化[3]等。研究人员使用生物实验方法来建立疾病和 miRNA 之间的关联，这些方法成本高、周期长且容易失败。因此，我们需要开发一个全新的计算模型框架来识别 miRNA 与疾病之间的关联。

有许多算法通过构建复杂的异构网络来预测 miRNA 与疾病的关联，Jiang 等人集成多种算法和生物数据，以及神经网络机器学习[11]。构建 miRNA 功能相似性网络矩阵和已知人类疾病-miRNA 网络矩阵，然后计算网络矩阵中节点的相似度得分，得分越高，与疾病相关的可能性就越大。Shi 等人[12]考虑了多种因素，将疾病基因与 miRNA 靶基因之间的功能联系添加到蛋白质相互作用(PPI)网络中，建立 miRNA-疾病关联网络，采用随机游走法构建 miRNA-疾病关联预测方法。Mørk 等人[13]提出了一种 miRPD 方法来获得疾病蛋白质与 miRNA 通过组合三个关联，miRNA 与蛋白质之间的关联评分矩阵，蛋白质与疾病之间的关联评分矩阵，miRNA 与疾病之间的共享蛋白质评分矩阵。Chen 等人[14]提出第一个基于全局网络相似度的计算模型 RWRMDA，是基于经过验证的 miRNA 与疾病之间的关联信息和人类 miRNA 功能相似性信息，采用随机游走方法，RWRMDA 通过对几种关键癌症的交叉验证实现了出色的预测性能。然而，它有局限性，即它不能用于 miRNA 与疾病之间的未知关联。Chen 等人[15]还提出了一种名为 WBSMDA 的方法，通过整合 miRNA 的功能相似性、疾病的语义相似性、miRNA 的高斯核相似性和疾

病以及 miRNA 与疾病之间的已知相关性来计算最终相关性评分。WBSMDA 可以有效地识别未知疾病-miRNA 关联。Chen 等人[16]又开发了一种新的算法模型 HGIMDA，其性能优于上述四种计算算法(WBSMDA、RLSMDA、RWRMDA 和 HDMP)。

深度学习、机器学习和神经网络也广泛应用于生物信息学的预测和判别实验。Xu 等人[17]基于miRNA-靶标相互作用提取特征，提出 miRNA-靶标失调网络(MTDN)并使用 SVM 分类器区分阳性或阴性样本。Chen 和 Yan [18]提供这 RLSMDA 揭示疾病与 miRNA 之间的关系，RLSMDA 可用于没有已知相关 miRNA 的疾病。此外，它是一种半监督(不需要负样本)和全局的方法(同时优先考虑所有疾病关联)。Chen [19]为了进一步提高 miRNA-疾病关联(RBMMMDA)的预测性能，开发了受限玻尔兹曼机器模型，它可以有效地预测不同类型的 miRNA 和疾病。

本研究试图将嵌入式学习功能与从未知关联中查找关联相结合。我们的方法由三部分组成：疾病嵌入模型、miRNAs 嵌入模型和深度稀疏自动编码模型。我们使用深度学习算法来构建这些模型。首先，我们尝试训练我们的疾病嵌入模型，同时训练 miRNA 嵌入模型来学习它们在高维空间中的表示。然后我们通过疾病和 miRNA 的嵌入模型将已知的已验证的 miRNA-疾病关联起来，然后通过深度稀疏自动编码器模型来学习潜在的特征。我们研究的主要贡献如下：

1) 我们向自动编码器添加稀疏性，该方法向重构误差添加稀疏性惩罚，以限制并非隐藏层中的所有单元在任何时候都被激活。

2) 我们实施了一种嵌入式学习方法来提取疾病和 miRNA 的表示。通过将疾病语义相似度与疾病高斯相互作用谱核相似度、miRNA 功能相似度和 miRNA 高斯相互作用谱核相似度相结合，自动学习高维密集向量特征来表示疾病和 miRNA。

2. 数据准备与计算框架

2.1. 人类 miRNA 与疾病的关联

我们采用人工收集和编译的 miRNA-疾病关联数据的数据库作为基准数据集。直接下载已知的人类 miRNA 和疾病的相关数据(http://www.cuilab.cn/static/hmdd3/data/hmdd2.zip)。

2.2. 疾病语义相似性

从国家医学网络图书馆(http://www.nlm.nih.gov)下载包含人类疾病的 C 类 MeSH。构建有向无环图(*DAG*)旨在计算疾病的语义相似度[20]。对于某些疾病节点 D，定义计算公式为：$DAG(D)=(D,T(D),E(D))$，其中 $T(D)$ 包括疾病节点 D 及其前辈的节点集，$E(D)$ 表示子节点与父节点直接链接的边集。疾病 D 的语义值定义如下：

$$DV1(D)=\sum_{d\in T(D)} D1_D(dis) \tag{1}$$

$$D1_D(dis)=\begin{cases}1 & \text{if } dis=D\\ \max\{\Delta * D1_D(dis')\,|\,dis'\in \text{children of } dis\} & \text{if } dis\neq D\end{cases} \tag{2}$$

其中，$D1_D(dis)$ 表示 $DAG(D)$中每个节点 D 对疾病 D 的语义价值贡献。在 $DAG(D)$中，疾病 D 是对自身最具体的描述，其语义贡献应该最大，设置为 1；距离疾病 D 较远的节点是疾病 D 的更一般描述，因此对疾病 D 的语义价值贡献较小。Δ 为对语义贡献值的衰减因子($0<\Delta<1$，在本研究中，Δ 设置为 0.5)。假设当任何两种疾病的 *DAG* 中有更多重叠时具有更高的语义相似度，计算疾病 dis_i 与 dis_j 之间的语义相似度得分定义如下：

$$SS1_{dis}\left(dis_i,dis_j\right)=\frac{\sum_{t\in T(dis_i)\cap T(dis_j)}\left(D1_{dis_i}(t)+D1_{dis_j}(t)\right)}{DV1\left(dis_i\right)+DV1\left(dis_j\right)} \tag{3}$$

DAG(*D*)中节点 *D* 对疾病 *D* 语义值的定义如下：

$$D2_D\left(dis\right)=-\log\frac{\text{包含节点}D\text{的疾病}DAGs\text{数目}}{\text{所有疾病的}DAGs\text{数目}} \tag{4}$$

同理，定义疾病 *D* 的语义值 *DV*2(*D*)以及疾病与疾病的相似度如下：

$$DV2\left(D\right)=\sum_{d\in T(D)}D2_D\left(dis\right) \tag{5}$$

$$SS2_{dis}\left(dis_i,dis_j\right)=\frac{\sum_{t\in T(dis_i)\cap T(dis_j)}\left(D2_{dis_i}(t)+D2_{d_{isj}}(t)\right)}{DV2\left(dis_i\right)+DV2\left(dis_j\right)} \tag{6}$$

结合两种疾病的语义相似度计算结果，疾病的语义相似度如下：

$$SS_{dis}=\frac{SS1_{dis}+SS2_{dis}}{2} \tag{7}$$

2.3. miRNA 功能相似性

基于具有相似功能的 mirna 更倾向于与相似的疾病表型相关的假设，我们使用 Wang 等人提出的方法[20]计算 miRNA 的功能相似性。

直接从 http://www.cuilab.cn/files/images/cuilab/misim.zip 下载。功能相似性矩阵(FS)包含 383 × 383 个 miRNA， $FS\left(mir_i,mir_j\right)$ 表示 miRNA mir_i 与 mir_j 之间的相似度值。

2.4. miRNA 与疾病高斯相互作用谱核相似性

基于已被验证的 miRNA-疾病关联矩阵，引用了 miRNA 的高斯相互作用谱核相似度和疾病的高斯相互作用谱的核相似度。 $KM\left(mir_i,mir_j\right)$ 为 miRNA mir_i 与 mir_j 之间的高斯相互作用谱核相似度。类似地，$KD\left(dis_i,dis_j\right)$ 表示疾病 dis_i 与 dis_j 之间高斯相互作用谱核相似性。

2.5. 计算框架 DSAEMDA

DSAEMDA 由三个主要部分组成：疾病模型、miRNA 模型和深度稀疏自动编码器模型，其中自动编码器包括三层编码器(用于在高维空间中编码已知的 miRNA-疾病关联)和三层解码器(用于重建计算误差)。经过验证的 miRNA-疾病已知关联用于训练深度稀疏自动编码器。DSAEMDA 工作流程如图 1 所示。

3. 特征表示

在基于神经网络的计算方法中，疾病和 miRNA 的正确表示非常重要，对模型的预测有很大影响。Peng 等人引入基因层来计算 miRNA 基因网络和疾病基因网络中的关联分数，以生成疾病(或 miRNA)和基因的 Pearson 相关性为载体，表达疾病(或 miRNA) [21]。Xuan 等人结合 miRNA 和疾病相似性及其关联形成特征表征[22]，miRNA 矩阵的表示和疾病矩阵的表示，通过整合 miRNA 功能相似性矩阵、已知的 miRNA-靶基因相互作用矩阵以及经过验证的 miRNA-疾病关联矩阵来提取 miRNA 特征和疾病特征[23]。

与上述方法不同，我们应用学习算法通过整合两种疾病相似性和疾病的高斯相互作用谱核相似性来直接提取疾病特征表示。miRNA 的表示来源于整合 miRNA 的功能相似性和 miRNA 的高斯相互作用谱

核相似性。我们将疾病映射到高维疾病向量空间，同时将 miRNA 映射到高维 miRNA 向量空间，构建 miRNA 回归模型和疾病回归模型，通过在高维空间中的距离来学习和表示这些准确且信息丰富的向量。

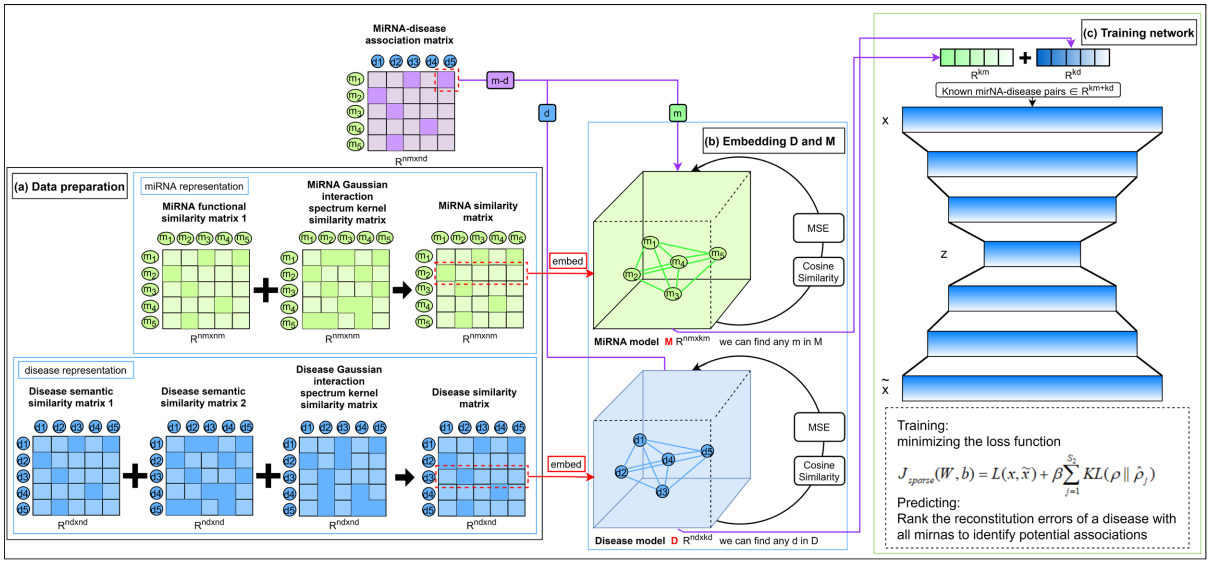

Figure 1. DSAEMDA model framework for predicting potential miRNA-disease associations. It includes three parts: (a) data preparation contains disease data and miRNAs data; (b) data embeddings to learn miRNA and disease representation features in high-dimensional spaces; (c) a deep sparse autoencoder for predicting the associations between miRNAs and diseases

图 1. DSAEMDA 模型框架用于预测潜在的 miRNA-疾病关联。包含三个部分：(a) 数据准备阶段包含疾病数据和 miRNA 数据；(b) 数据嵌入学习 miRNA 和疾病表示在高维空间中的特征；以及(c) 用以预测潜在 miRNA-疾病潜在关联的深度稀疏自动编码器

3.1. 疾病特征表示

神经语言处理(NPL)可以自动学习用向量来表示不同维度的单词或句子[24] [25] [26]，我们训练两个高维模型来学习疾病的和 miRNA 的表示。

我们首先为每种疾病定义唯一的数字来表示特定的疾病，例如，dis_i 表示矩阵 D 的第 i 行的密集向量是从矩阵 D 的嵌入中检索出来的。在训练之前，D 中的元素是不确定的，我们随机初始化 D 中的元素。经过多次学习迭代、最小化误差和反向传播，D 被训练为高维空间中的有效疾病表达。D 的定义如下：

$$D = \left[dis_1, dis_2, \cdots, dis_{nd}\right]^{\mathrm{T}}, dis_i \in R^{kd} \tag{8}$$

其中 dis_i 代表第 i 个疾病的高维空间向量。nd 是疾病个数，kd 是疾病映射在高维空间中的向量大小，所以疾病矩阵在高维空间的表示为 $D \in R^{nd\times kd}$ 。

单独使用疾病语义相似度矩阵很难获得很好的预测性能。同时，仅使用疾病高斯相互作用谱核相似性来预测已知的 miRNA-疾病关联计算也是不够准确的。因此，为了获得良好的预测性能，需要将疾病高斯相互作用谱核相似度矩阵 KD 与疾病语义相似度矩阵 SS 结合起来。通过对两个矩阵进行加权整合，我们将其表示为 SD。最后，我们利用矩阵 SD 来学习高维空间中疾病的向量 D。疾病矩阵如下：

$$SD\left(dis_i, dis_j\right) = \alpha SS_{dis}\left(dis_i, dis_j\right) + \left(1-\alpha\right) KD\left(dis_i, dis_j\right) \tag{9}$$

SD、SS_{dis}、KD 维度相同；换句话说，是介于 0 和 1 之间的权重。疾病的高斯相互作用谱核相似度矩阵 KD 和疾病语义相似度矩阵 SD 中的每个元素都在(0, 1)的范围内。我们将 $SD\left(dis_i, dis_j\right)$ 视为任意两

种疾病 dis_i 和 dis_j 之间的高维空间距离。根据余弦相似度的计算特性，通过两个向量之间的夹角余弦来衡量两个向量之间的相似度[27] [28]。对余弦函数的公式进行算术调整，保证任意两种疾病的计算结果在(0, 1)内。疾病模型的余弦相似度 SD'由以下公式给出：

$$SD'\left(dis_i, dis_j\right) = \frac{1}{2} + \frac{1}{2}\frac{dis_i \cdot dis_j}{\left\|dis_i\right\|\left\|dis_j\right\|} \tag{10}$$

我们的疾病模型通过构建回归模型来学习高维空间中疾病的特征。任意两种疾病在高维空间向量的分值越高，意味着它们在高维空间中具有更高的相似性。疾病计算模型最小化所有疾病之间的损失函数定义如下：

$$\min \frac{1}{N_{dis}} \sum^{N_{dis}} \left\| SD\left(dis_i, dis_j\right) - SD'\left(dis_i, dis_j\right)\right\|^2 \tag{11}$$

$N_{dis} = \left(nd-1\right) * nd/2$ 是我们训练样本的总数。在每次训练迭代中，以均方损失为准则，通过反向传播的独立自适应学习率(Adam)算法更新疾病矩阵 D。

3.2. miRNA 特征表示

类似地，代表 495 个 miRNA 的高维空间矩阵 M 定义如下：

$$M = \left[mir_1, mir_2, \cdots, mir_{nm}\right]^{\mathrm{T}}, mir_i \in R^{km} \tag{12}$$

其中 mir_i 代表第 i 个 miRNA 的高维空间向量。nm 是 miRNA 个数，km 是 miRNA 映射在高维空间中的向量大小，所以 miRNA 矩阵在高维空间的表示为 $M \in R^{nm \times km}$。同样，我们通过以下等式学习矩阵 M：

$$\min \frac{1}{N_{mir}} \sum^{N_{mir}} \left\| SM\left(mir_i, mir_j\right) - SM'\left(mir_i, mir_j\right)\right\|^2 \tag{13}$$

其中 $N_{mir} = \left(nm-1\right) * nm/2$ 是训练样本总数，SM 和 SM'的定义公式如下：

$$SM\left(mir_i, mir_j\right) = \beta FS\left(mir_i, mir_j\right) + \left(1-\beta\right) KM\left(mir_i, mir_j\right) \tag{14}$$

$$SM'\left(mir_i, mir_j\right) = \frac{1}{2} + \frac{1}{2}\frac{mir_i \cdot mir_j}{\left\|mir_i\right\|\left\|mir_j\right\|} \tag{15}$$

4. 基于自动编码器的关联预测器

4.1. 自动编码器

Autoencoder 简称“AE”，AE 是 Hinton [29]在 1980 年代提出的一种无监督聚类算法。它已广泛应用于特征提取、数据压缩、特征降维、异常检测和模型生成。神经网络中的自动编码器是一种无监督学习算法，也就是说，它不需要分类标签。它使用反向传播最小化损失函数算法使目标输出值无限接近原始输入值。基本的自动编码器是三层神经网络模型，第一层是数据输入层，第二层是隐藏层，第三层是输出重建层。它也是一种无监督学习模型。我们的七层稀疏自动编码器模型主要由两个元素组成：编码器(用于数据压缩和特征提取)和解码器(用于重构输入)。编码器的任务是压缩数据并从高维模型 D 和 M 中提取特征，解码器的任务是恢复数据并从隐藏码中重构输入。我们的神经网络模型由七个完全连接的层组成。图 2 显示了我们的 7 层深度稀疏自动编码器 DSAEMDA 的架构。

miRNA 与疾病之间的相关性由重建误差表示。重构误差的程度直接反映了 miRNA 与疾病的相关性。我们的稀疏自编码器输入是疾病 dis 和 miRNA mir 之间的链接向量，链接向量主要由两部分组成，一部

分是查询疾病模型 D 中的疾病高维向量 dis，另一部分是查询 miRNA 模型 M 中的 miRNA 高维向量 mir，接下来，向量 dis 和向量 mir 链接(dis, mir)作为输入向量到我们的深度稀疏自动编码器。

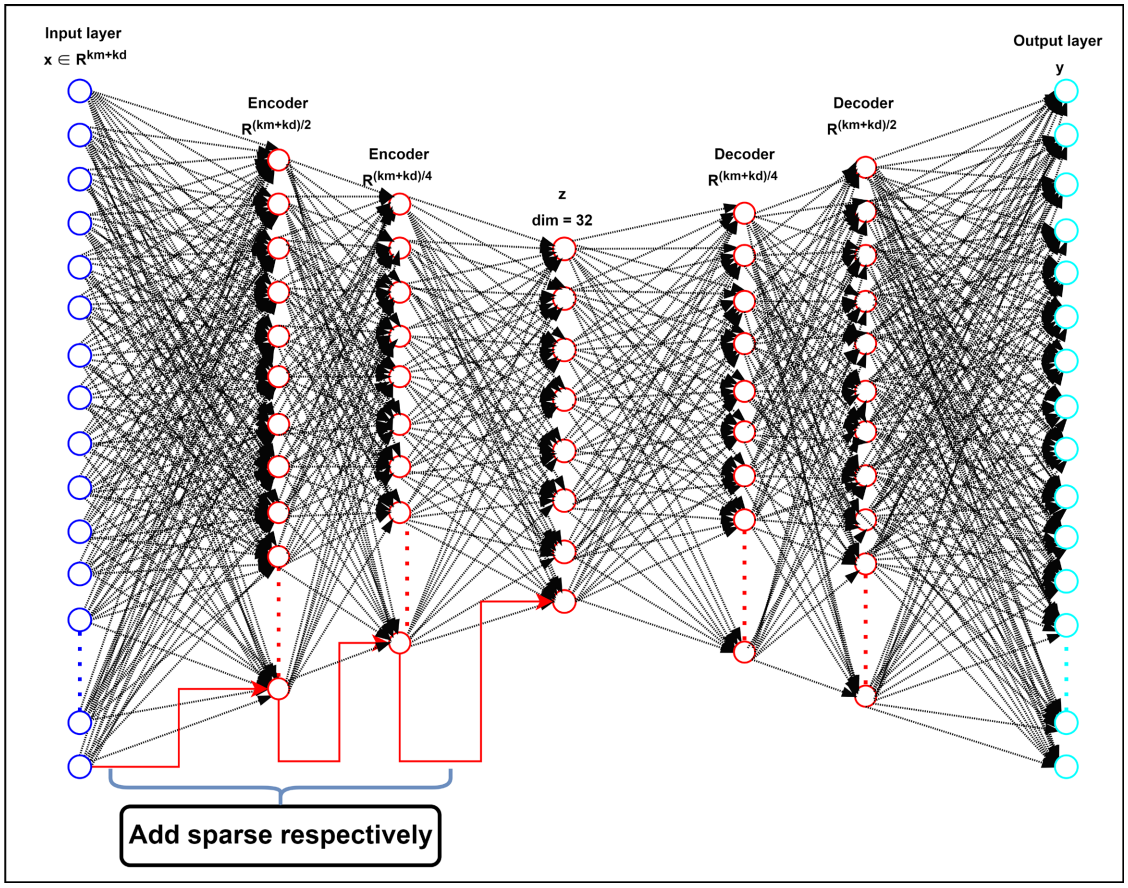

Figure 2. The framework of sparse autoencoder
图 2. 稀疏自动编码器框架

第 i 个训练样本定义如下：

$$x_i = [dis, mir] \in R^{kd+km} \tag{16}$$

样本 x_i，编码器通过以下公式压缩数据并提取潜在代码的特征 z_i：

$$h_{(e)i}^{(l)} = f_{(e)}\left(W_{(e)}{}^{l} h_{(e)i}^{(l-1)} + b_{(e)}{}^{l}\right) \tag{17}$$

$$z_i = W_{(e)}{}^{L} h_{(e)i}^{(L-1)} + b_{(e)}{}^{L} \tag{18}$$

$l = \{1, \cdots, L\}$，$L = 3$ 表示编码器的隐藏层数为 3。$h_{(e)i}^{(l)}$ 表示 l 隐藏层，$h_{(e)i}^{(0)}$ 表示输入 x_i。权重矩阵 $W_{(e)}{}^{l}$ 和偏差 $b_{(e)}{}^{l}$ 是 l 层的参数。z_i 是编码器层的输出，代表 x_i 的潜在表示。通过非线性激活函数 $f_e(\cdot)$ (ReLU) 校正线性元素[30]。

解码器旨在从编码器的潜在表示 z_i 中尽可能多的重建输入 x_i。下面的公式定义了解码器：

$$h_{(d)i}^{(l)} = f_{(d)}\left(W_{(d)}{}^{l} h_{(d)i}^{(l-1)} + b_{(d)}{}^{l}\right) \tag{19}$$

$$\tilde{x}_i = g_{(d)}\left(W_{(d)}{}^{L} h_{(d)i}^{(L-1)} + b_{(d)}{}^{L}\right) \tag{20}$$

其中 $h_{(d)i}^{(l)}$ 表示 l 层隐藏层，解码层的第一层为中间隐藏层 $h_{(d)i}^{(0)}=z_i$ 。，$L=3$ 表示解码器的隐藏层数为 3，权重矩阵 $W_{(d)}{}^{l}$ 和偏差 $b_{(d)}{}^{l}$ 是解码器 l 层的参数。我们模型中最后一个解码器的输出 $\tilde{x}_i$ 是对输入 x_i 的重构。将 $f_d(\cdot)$ 和 $g_d(\cdot)$ 分别设置为 ReLU 和 TANH。

4.2. 稀疏性

受机器学习领域的启发[31]。为了给自编码器添加稀疏性，我们选择的方法是添加 *KL* 散度，*KL* 散度就是相对熵 = 交叉熵 − 信息熵，在重构误差中添加稀疏性惩罚，以限制隐藏层中的所有单元在任何时候都被激活。

我们用 $a_j^{(2)}(x)$ 来表示在给定输入 x 下自编码神经网络中隐藏神经元 j 的激活程度。隐藏神经元 j 的平均激活度(在训练集上平均)表示为：

$$\hat{\rho}_j=\frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}\left(x^{(i)}\right)\right] \tag{21}$$

稀疏性限制可以理解为最小化隐藏层神经元的平均激活度，可以表示为 $\hat{\rho}_j=\rho$，其中 ρ 是稀疏性参数，通常是接近于 0 的较小值，我们在神经网络优化的原始目标函数中加入稀疏性限制作为额外的惩罚因子，我们可以选择如下形式的惩罚因子：

$$\sum_{j=1}^{S_2}\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{22}$$

其中 S_2 为在隐藏层中神经元的数量，索引 j 依次代表隐藏层中的每一个神经元。也可以将其描述为相对熵，表示为：

$$\sum_{j=1}^{S_2}KL\left(\rho\,\|\,\hat{\rho}_j\right) \tag{23}$$

$KL\left(\rho\,\|\,\hat{\rho}_j\right)=\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}$ 是两个伯努利随机变量之间的相对熵，均值为 ρ 和 $\hat{\rho}_j$。

现在，我们的总成本函数可以表示为：

$$J_{sparse}(W,b)=L(x,\tilde{x})+\beta\sum_{j=1}^{S_2}KL\left(\rho\,\|\,\hat{\rho}_j\right) \tag{24}$$

$L(x,\tilde{x})$ 定义如下：

$$L(x,\tilde{x})=\sum_{i=1}^{N}\left\|x_i-\tilde{x}_i\right\|^2+\lambda\left\|J_h(x_i)\right\|^2 \tag{25}$$

N 是已知已验证的 miRNA-疾病关联的总数。前面的第一项是平方损失，后面的第二项是正则化的 Frobenius 范数[32]，λ 和 β 是两个超参数。训练我们的深度稀疏自编码器的目标是尝试最小化上述损失函数并迭代更新神经元参数。

5. 数据实验与结果分析

5.1. 实施细节

在我们的环境中使用的是一个开源机器学习框架 PyTorch，版本为 1.7.1。我们的实验是在配备 NVIDIA1080TI 图形处理器的 Windows10 平台上进行的。

DSAEMDA 根据以下程序进行训练。我们首先使用均方误差(MSE)作为损失函数来训练疾病模型和

miRNA 模型。在 100 个 epoch 之后，我们获得疾病 D 和 miRNA M 的两个高维空间矩阵模型，然后将 miRNA 与疾病的已知关联嵌入到 D 和 M 中来训练深度稀疏自编码器。我们的深度稀疏自编码器通常会在 100~150 epoch 收敛。

HMDDv2.0 数据库用于训练深度稀疏自编码器，设置疾病数量 nd 为 383，miRNA 数量 nm 为 495。权重初始化时，当初始化时，两个模型的权重都均匀地分布在−0.1 和 0.1 之间。然后利用 PyTorch 中实现两个嵌入学习模型。最后生成疾病矩阵模型 D 和 miRNA 矩阵模型 M，分别表示疾病和 miRNA 的高维模型。疾病模型通过带有反向传播的 Adam 方法进行端到端训练。Adam 算法用于优化上述模型[33]。$\alpha = 0.3$，$\beta = 0.2$，$\rho = 0.02$，batch = 128。初始学习率 = 1e−4，如果损失没有改善，每 4 步学习率降低到原来的十分之一。最后，最小化损失函数迭代优化得到最佳模型。在实验参数设置中，设置了 $kd = km$。相同的参数设置被用于训练 miRNA 高维空间模型。

对于深度稀疏自编码器模型，模型的输入是已知已验证的 miRNA 与疾病关联的串联高维空间向量，表明输入层初始神经元个数为$(kd + km)$。第一层编码器的输出为$(kd + km)/2$ 作为第二层编码器的输入，第二层编码器的输出为$(kd + km)/8$ 作为第三层的输入编码器，中间隐藏码的大小为 32。解码器的结构与编码器的结构相反。

5.2. 评价结果

在我们的能力范围内，为了得到一个可靠和稳定的模型，我们使用五折交叉验证(5-CV)来评估我们模型的推理能力。已知的 miRNA-疾病关联被用来训练我们的模型，而未知关联不参与我们的训练过程。

在 5-CV 中，将已知的 miRNA-疾病关联被随机分成 5 组，一组作为测试样本，其余四组作为训练样本。该步骤重复五次作为一个完整的循环。为了获得公平公正的数据，我们为每一轮训练样本训练了一个深度稀疏自动编码器。接下来，分别计算未知 miRNA-疾病关联的重建误差和测试样本的重建误差。最后，我们根据所有重建误差进行排名。重复这些步骤 5 次的目的是为了减少随机样本分割带来的偏差，并将平均排名作为我们模型实验的最终结果。

此外，AUC 定义为 ROC 曲线下的面积。AUC 值用作我们模型的评估标准。因此，DSAEMDA 获得的 5-CV 方案的 AUC 为 0.9412 (表 1)。为了进一步证明 DSAEMDA 模型预测的优越性，我们还将 DSAEMDA 与 AEMDA [34]、GRL21NMF [35]、ICFMDA [36]、SACMDA [37]和 IMCMDA [38]进行比较。在 5-CV 方案下，AEMDA、GRL21NMF、ICFMDA、SACMDA，和 IMCMDA 模型的 AUC，分别为 0.9383、0.9276、0.9045、0.8763 和 0.8330，表明 DSAEMDA 的性能优于其他模型。可以说 SAEMDA 模型在预测 miRNA-疾病潜在关联方面取得了显著进展。

Table 1. Comparison between DSAEMDA and other prediction methods
表 1. DSAEMDA 与其它预测方法比较

预测方法	DSAEMDA	AEMDA	GRL21NMF	ICFMDA	SACMDA	IMCMDA
五折交叉验证的 AUC	0.9412	0.9383	0.9276	0.9045	0.8763	0.8330

6. 结论

在我们的研究中，我们为自动编码器添加了稀疏性，并为预测 miRNA-疾病潜在关联创建了一个新的预测框架，称为 DSAEMDA。在疾病方面，将两种不同疾病的语义相似度矩阵与疾病的高斯相互作用谱核相似度矩阵相结合。在 miRNA 方面，将 miRNA 的功能相似矩阵与 miRNA 的高斯相互作用谱核相似矩阵相结合。将已知的 miRNA-疾病关联嵌入高维空间，训练两个模型以提取疾病和 miRNA 在高维空

间中的表示。然后，提出了一个 7 层深度稀疏自编码模型，用于从已知的 miRNA-疾病关系中挖掘潜在的 miRNA-疾病关系。基于交叉验证的实验和比较其他案例研究的结果表明，DSAEMDA 是有效和可靠的，并且优于几种最先进的方法。

几个关键因素促成了 DSAEMDA 的卓越性能。首先，DSAEMDA 可以通过整合疾病相似性的学习算法提取疾病的高维特征表示，并将其嵌入到高维空间中。通过将疾病的两种语义相似度与疾病的高斯相互作用谱核相似度相结合，通过反向传播最小化损失函数来学习疾病的高维密集向量特征。miRNA 的功能相似性结合 miRNA 的高斯相互作用谱核相似性和通过反向传播最小化损失函数学习到的 miRNA 的高维密集向量特征。其中，我们使用随机梯度下降(Adam)程序寻找最优解，以保证 miRNA 特征向量和疾病特征向量的可靠性。其次，我们为自编码器添加了稀疏性，在重构误差上添加了稀疏性惩罚，以限制隐藏层中每一层的所有神经元都被激活，这有利于特征提取。鉴于上述情况，与之前提出的方法相比，DSAEMDA 确实提高了预测精度。

对于 DSAEMDA，还有进一步改进的空间，例如网络层次结构和网络中的神经元数量。未来，随着更多样本的可用，我们将整合来自 HMDDv3.2 数据的更多 miRNA-疾病关联来训练 DSAEMDA。我们将结合各种神经网络，使用具有更深层次或更复杂模型结构的神经网络架构来实现更好的性能。

参考文献

[1] Ambros, V. (2001) microRNAs: Tiny Regulators with Great Potential. *Cell*, **107**, 823-826. https://doi.org/10.1016/S0092-8674(01)00616-X

[2] Ambros, V. (2004) The Functions of Animal microRNAs. *Nature*, **431**, 350-355. https://doi.org/10.1038/nature02871

[3] Bartel, D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**, 215-233. https://doi.org/10.1016/j.cell.2009.01.002

[4] Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: Integrating microRNA Annotation and Deep-Sequencing Data. *Nucleic Acids Research*, **39**, D152-D157. https://doi.org/10.1093/nar/gkq1027

[5] Jopling, C.L., Yi, M.K., Lancaster, A.M., Lemon, S.M. and Sarnow, P. (2005) Modulation of Hepatitis C Virus RNA Abundance by a Liver-Specific MicroRNA. *Science*, **309**, 1577-1581. https://doi.org/10.1126/science.1113329

[6] Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The *C. elegans* Heterochronic Gene lin-4 Encodes small RNAs with Antisense Complementarity to lin-14. *Cell*, **89**, 1828-1835. https://doi.org/10.1016/0092-8674(93)90529-Y

[7] Cheng, A.M., Byrom, M.W., Shelton, J. and Ford, L.P. (2005) Antisense Inhibition of Human miRNAs and Indications for an Involvement of miRNA in Cell Growth and Apoptosis. *Nucleic Acids Research*, **33**, 1290-1297. https://doi.org/10.1093/nar/gki200

[8] Karp, X. and Ambros, V. (2005) Encountering microRNAs in Cell Fate Signaling. *Science*, **310**, 1288-1289. https://doi.org/10.1126/science.1121566

[9] Miska, E.A. (2005) How microRNAs Control Cell Division, Differentiation and Death. *Current Opinion in Genetics & Development*, **15**, 563-568. https://doi.org/10.1016/j.gde.2005.08.005

[10] Xu, P.Z., Guo, M. and Hay, B.A. (2004) MicroRNAs and the Regulation of Cell Death. *Trends in Genetics*, **20**, 617-624. https://doi.org/10.1016/j.tig.2004.09.010

[11] Jiang, Q., Hao, Y. and Wang, G. (2010) Prioritization of Disease microRNAs through a Human Phenome-microRNAome Network. *BMC Systems Biology*, **4**, S2. https://doi.org/10.1186/1752-0509-4-S1-S2

[12] Shi, H., Xu, J. and Zhang, G. (2013) Walking the Interactome to Identify Human miRNA-Disease Associations through the Functional Link between miRNA Targets and Disease Genes. *BMC Systems Biology*, **7**, Article No. 101. https://doi.org/10.1186/1752-0509-7-101

[13] Mørk, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J. and Jensen, L.J. (2014) Protein-Driven Inference of miRNA-Disease Associations. *Bioinformatics* (*Oxford*, *England*), **30**, 392-397. https://doi.org/10.1093/bioinformatics/btt677

[14] Chen, X., Liu, M.X. and Yan, G.Y. (2012) RWRMDA: Predicting Novel Human microRNA-Disease Associations. *Molecular BioSystems*, **8**, 2792-2798. https://doi.org/10.1039/c2mb25180a

[15] Chen, X., Yan, C.C., Zhang, X., You, Z.H., Deng, L., Liu, Y., Zhang, Y. and Dai, Q. (2016) WBSMDA: Within and

between Score for MiRNA-Disease Association Prediction. *Scientific Reports*, **6**, Article No. 21106. https://doi.org/10.1038/srep21106

[16] Chen, X., Yan, C., Zhang, X., You, Z., Huang, Y. and Yan, G. (2016) HGIMDA: Heterogeneous Graph Inference for MiRNA-Disease Association Prediction. *Oncotarget*, **7**, 65257-65269. https://doi.org/10.18632/oncotarget.11251

[17] Xu, J., Li, C.X., Lv, J.Y., Li, Y.S., Xiao, Y., Shao, T.T., Huo, X., Li, X., Zou, Y., Han, Q.L., Li, X., Wang, L.H. and Ren, H. (2011) Prioritizing Candidate Disease miRNAs by Topological Features in the miRNA Target-Dysregulated Network: Case Study of Prostate Cancer. *Molecular Cancer Therapeutics*, **10**, 1857-1866. https://doi.org/10.1158/1535-7163.MCT-11-0055

[18] Chen, X. and Yan, G.Y. (2014) Semi-Supervised Learning for Potential Human microRNA-Disease Associations Inference. *Scientific Reports*, **4**, Article No. 5501. https://doi.org/10.1038/srep05501

[19] Chen, X. (2015) Predicting lncRNA-Disease Associations and Constructing lncRNA Functional Similarity Network Based on the Information of miRNA. *Scientific Reports*, **5**, Article No. 13186. https://doi.org/10.1038/srep13186

[20] Wang, D., Wang, J., Lu, M., Song, F. and Cui, Q. (2010) Inferring the Human microRNA Functional Similarity and Functional Network Based on microRNA-Associated Diseases. *Bioinformatics*, **26**, 1644-1650. https://doi.org/10.1093/bioinformatics/btq241

[21] Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., Shang, X. and Wei, Z. (2019) A Learning-Based Framework for miRNA-Disease Association Identification Using Neural Networks. *Bioinformatics* (*Oxford*, *England*), **35**, 4364-4371. https://doi.org/10.1093/bioinformatics/btz254

[22] Xuan, P., Sun, H., Wang, X., Zhang, T. and Pan, S. (2019) Inferring the Disease Associated miRNAs Based on Network Representation Learning and Convolutional Neural Networks. *International Journal of Molecular Sciences*, **20**, 3648. https://doi.org/10.3390/ijms20153648

[23] Fu, L. and Peng, Q. (2017) A Deep Ensemble Model to Predict miRNA-Disease Association. *Scientific Reports*, **7**, Article No. 14482. https://doi.org/10.1038/s41598-017-15235-6

[24] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the* 26*th International Conference on Neural Information Processing Systems*, Volume 2, 3111-3119.

[25] Bahdanau, D., Cho, K.H. and Bengio, Y. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. 3*rd International Conference on Learning Representations*, *ICLR* 2015, San Diego, 7-9 May 2015.

[26] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*Long and Short Papers*), Volume 1, 4171-4186.

[27] Tan, P.-N., Steinbach, M. and Kumar, V. (2005) Introduction to Data Mining. Addison-Wesley Longman Publishing, Boston.

[28] Manning, C.D., Raghavan, P. and Schutze, H. (2008) Introduction to Information Retrieval. Cambridge University Press, Cambridge.

[29] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[30] Nair, V. and Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the* 27*th International Conference on Machine Learning*, Haifa, 21-24 June 2010, 807-814.

[31] Ng, A. (2011) Sparse Autoencoder. CS294A Lecture Notes, Vol. 72, 1-19.

[32] Rifai, S., Vincent, P., Muller, X., Glorot, X. and Bengio, Y. (2011) Contractive Auto-Encoders: Explicit Invariance during Feature Extraction. *Proceedings of the* 28*th International Conference on Machine Learning*, *ICML* 2011, Bellevue, 28 June-2 July 2011, 833-840.

[33] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization.

[34] Ji, C., Gao, Z., Ma, X., Wu, Q. and Zheng, C. (2021) AEMDA: Inferring miRNA-Disease Associations Based on Deep Autoencoder. *Bioinformatics*, **37**, 66-72. https://doi.org/10.1093/bioinformatics/btaa670

[35] Gao, Z., Wang, Y.-T., Wu, Q.-W., Ni, J.-C. and Zheng, C.-H. (2020) Graph Regularized L2,1-Nonnegative Matrix Factorization for miRNA-Disease Association Prediction. *BMC Bioinformatics*, **21**, Article No. 61. https://doi.org/10.1186/s12859-020-3409-x

[36] Jiang, Y., Liu, B., Yu, L., Yan, C. and Bian, H. (2018) Predict MiRNA-Disease Association with Collaborative Filtering. *Neuroinformatics*, **16**, 363-372. https://doi.org/10.1007/s12021-018-9386-9

[37] Shao, B., Liu, B. and Yan, C. (2018) SACMDA: MiRNA-Disease Association Prediction with Short Acyclic Connections in Heterogeneous Graph. *Neuroinformatics*, **16**, 373-382. https://doi.org/10.1007/s12021-018-9373-1

[38] Chen, X., Wang, L., Qu, J., Guan, N.N. and Li, J.Q. (2018) Predicting miRNA-Disease Association Based on Inductive Matrix Completion. *Bioinformatics*, **34**, 4256-4265. https://doi.org/10.1093/bioinformatics/bty503