(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110993113 B
(45)授权公告日 2023.04.07

(21)申请号 201911146003.5

(22)申请日 2019.11.21

(65)同一申请的已公布的文献号
申请公布号 CN 110993113 A

(43)申请公布日 2020.04.10

(73)专利权人 广西大学
地址 530004 广西壮族自治区南宁市西乡塘区大学东路100号

(72)发明人 兰伟 赖德焕 陈庆锋 吴锡敏 刘锦

(74)专利代理机构 长沙市融智专利事务所(普通合伙) 43114
专利代理师 杨萍

(51)Int.Cl.
G16H 50/70(2018.01)
G16B 20/00(2019.01)
G16B 30/00(2019.01)
G16B 40/00(2019.01)

(56)对比文件
CN 109797221 A,2019.05.24
CN 108537005 A,2018.09.14
CN 107862179 A,2018.03.30
CN 106599610 A,2017.04.26
CN 108763367 A,2018.11.06
US 2016174902 A1,2016.06.23
US 2019106732 A1,2019.04.11
WO 2019173446 A1,2019.09.12
US 2017321198 A1,2017.11.09
Wei Lan.LDICDL: LncRNA-disease association identification based on Collaborative Deep Learning.《IEEE/ACM transactions on computational biology and bioinformatics》.2020,第第19卷卷(第第19卷期),第1715-1723页.
赵琪;梁丹;胡桓;张力;刘宏生;.基于随机游走算法预测lncRNAs与疾病关系的研究进展.辽宁大学学报(自然科学版).2018,第43卷(第3期),第273-280页.

审查员 张永武

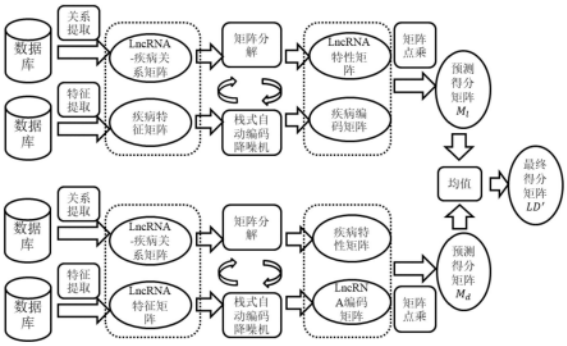
权利要求书4页 说明书10页 附图3页

(54)发明名称

基于MF-SDAE的lncRNA-疾病关系预测方法及系统

(57)摘要

本发明提出了一种基于MF-SDAE的lncRNA-疾病关系预测方法及系统。首先构建已知的lncRNA-疾病关系矩阵、lncRNA特征矩阵与疾病特征矩阵;使用矩阵分解模型来对已知的lncRNA-疾病关系矩阵进行分解得到lncRNA特性矩阵和疾病特性矩阵,栈式降噪自动编码器分别对lncRNA特征矩阵和疾病特征矩阵进行编码得到各自的编码矩阵,以最小化损失函数值为目标,迭代求解模型的最优参数。最终利用训练好的模型得到编码矩阵和特性矩阵,将它们进行矩阵乘法操作得到lncRNA-疾病关系得分矩阵。本发明简单有效,预测性能好。



CN 110993113 B

1.一种基于MF-SDAE的lncRNA-疾病关系预测方法,其特征在于,包括以下步骤:

1)构建已知的lncRNA-疾病关系矩阵LD、lncRNA特征矩阵M_{lf}与疾病特征矩阵M_{df};

所述步骤1)中,构建已知的lncRNA-疾病关系矩阵的过程如下:

构建一个矩阵LD,其每一行对应一种lncRNA,每一列对应一种疾病,若有数据库记录了第i种lncRNA与第j种疾病存在关系,则将LD中第i行第j列的元素LD(i,j)设为1;否则将LD(i,j)设为0;由此得到的矩阵LD即为已知的lncRNA-疾病关系矩阵;

构建lncRNA特征矩阵的过程如下:

构建一个矩阵M_{lf},其每一行对应一种lncRNA,每一列对应一种与lncRNA关联的信息,若有数据库记录了第i种lncRNA与第p种与lncRNA关联的信息存在关系,则将M_{lf}中第i行第p列的元素$M_{lf}(i,p)$设为1;否则将$M_{lf}(i,p)$设为0;由此得到的矩阵M_{lf}即为lncRNA特征矩阵;

构建疾病特征矩阵的过程如下:

构建一个矩阵M_{df},其每一行对应一种疾病,每一列对应一种与疾病关联的信息,若有数据库记录了第j种疾病与第q种与疾病关联的信息存在关系,则M_{df}中第j行第q列的元素$M_{df}(j,q)$设为1;否则$M_{df}(j,q)$设为0;由此得到的矩阵M_{df}即为疾病特征矩阵;

2)构建基于MF-SDAE的混合预测模型;其中,MF为矩阵分解,SDAE为栈式降噪自动编码器;所述混合预测模型包括矩阵分解模型和栈式降噪自动编码器模型;其中,矩阵分解模型用于对输入矩阵进行分解,得到两个输出矩阵;栈式降噪自动编码器模型用于对输入矩阵进行特征编码,提取高层特征;

3)利用矩阵分解模型对lncRNA-疾病关系矩阵LD进行分解,得到两个输出矩阵,即lncRNA特性矩阵L与疾病特性矩阵D;

所述步骤3)具体过程如下:

设定迭代次数T;

初始化lncRNA特性矩阵L与疾病特性矩阵D;

进行T次迭代,在每一次迭代过程中,按以下公式更新矩阵L和D:

$L(i,:)=LD(i,:)C^{i}D(\gamma' I+D^{T}C^{i}D)^{-1}$

$$D(j,:)=LD(:,j)^{T}\widetilde{C}^{j}L(\gamma' I+L^{T}\widetilde{C}^{j}L)^{-1}$$

其中,L(i,:)为矩阵L的第i行;D(j,:)为矩阵D的第j行,LD(i,:)为矩阵LD的第i行,C^{i}为第i种lncRNA对应的对角矩阵,其第j行第j列的元素值$C^{i}(j,j)=\beta_{i,j}$,$\beta_{i,j}$是偏好因子,$\beta_{i,j}=1+\theta \cdot LD(i,j)$,θ为自由参数;$\widetilde{C}^{j}$为第j种疾病对应的对角矩阵,其i行第i列的元素值$\widetilde{C}^{j}(i,i)=\beta_{i,j}$;LD(:,j)为lncRNA-疾病关系矩阵LD中的第j列;I是单位矩阵,γ′为自由参数;

T次迭代后得到的矩阵L和D即矩阵分解模型的输出矩阵;

4)对混合预测模型进行训练;

初始化混合预测模型参数;

定义损失函数;以最小化损失函数值为目标,迭代求解混合预测模型的最优参数,得到训练好的混合预测模型;

每轮训练迭代过程中,先采用混合预测模型进行以下两部分数据处理:

采用混合预测模型中的栈式降噪自动编码器模型对lncRNA特征矩阵M_{lf}进行特征编码,

得到隐藏层和输出层输出的lncRNA特征编码矩阵，分别记为$X_{encodes_l}$和X_{out_l}；

采用混合预测模型中的栈式降噪自动编码器模型对疾病特征矩阵M_{df}进行特征编码，得到隐藏层和输出层输出的疾病特征编码矩阵，分别记为$X_{encodes_d}$和X_{out_d}；

然后根据混合预测模型的输入和输出计算相应的损失函数值；

所述步骤4)中，损失函数为：

$$Loss=\sum_{i,j}\beta_{i,j}[LD(i,j)-L(i,:)\cdot D(j,:)^{T2}+\gamma(\sum_i||L(i,:)||^2+\sum_j||D(j,:)||^2)+\gamma_l(||L-X_{encodes\_l}||^2)+\gamma_d(||D-X_{encodes\_d}||^2)+\gamma_{n\_l}(||M_{lf}-X_{out\_l}||^2)+\gamma_{n\_d}(||M_{df}-X_{out\_d}||^2)+\sum_k\gamma_k||W_k||^2+\sum_k\gamma_b||W_b||^2$$

$$\beta_{i,j}=1+\theta\cdot LD(i,j)$$

其中，$||\cdot||$表示求2-范数，$\beta_{i,j}$是偏好因子；LD(i,j)为矩阵LD中第i行第j列的元素；L(i,:)为矩阵L的第i行；D(j,:)为矩阵D的第j行；θ、γ、γ_l、γ_d、γ_{n_l}、γ_{n_d}和γ_k均为自由参数；W_k和b_k分别为栈式降噪自动编码器中第k个隐藏层的权值矩阵和阈值向量；

5)利用训练好的混合预测模型对lncRNA特征矩阵M_{lf}和疾病特征矩阵M_{df}进行处理，得到相应的lncRNA特征编码矩阵X_{encods_l}和疾病特征编码矩阵X_{encods_d}；

结合X_{encods_l}与D计算得分矩阵M_l，其第i行第j列的元素$M_l(i,j)$计算方法为：

$$M_l(i,j)=X_{encods\_l}(i,:)\cdot D(j,:)^T$$

其中，$X_{encods_l}(i,:)$表示X_{encods_l}的第i行，D(j,:)表示D的第j行；

结合X_{encods_d}与L计算得分矩阵M_d，其第i行第j列的元素$M_d(i,j)$计算方法为：

$$M_d(i,j)=L(i,:)\cdot X_{encods\_d}(j,:)^T$$

其中，L(i,:)表示L的第i行，$X_{encods_d}(j,:)$表示X_{encods_d}的第j行；

求M_l和M_d的加权平均值，所得结果即为预测得到的lncRNA-疾病关系得分矩阵LD′，其第i行第j列的元素LD′(i,j)表示预测得到的第i种lncRNA和第j种疾病存在关系的可能性。

2.根据权利要求1所述的基于MF-SDAE的lncRNA-疾病关系预测方法，其特征在于，所述与lncRNA关联的信息包括与lncRNA关联的基因信息、基因功能信息和miRNA信息。

3.根据权利要求1所述的基于MF-SDAE的lncRNA-疾病关系预测方法，其特征在于，所述与疾病关联的信息包括与疾病关联的基因信息和miRNA信息。

4.根据权利要求1所述的基于MF-SDAE的lncRNA-疾病关系预测方法，其特征在于，将lncRNA特性矩阵L与疾病特征矩阵D初始化为服从0～1均匀分布的随机矩阵，即产生[0,1)上均匀分布的随机数，来填充L和D，完成L和D的初始化。

5.根据权利要求1所述的基于MF-SDAE的lncRNA-疾病关系预测方法，其特征在于，所述栈式降噪自动编码器模型包括依次连接的一个输入层、一个损坏层、三个隐藏层和一个输出层；将其第二个隐藏层的输出作为$X_{encodes_l}/X_{encodes_d}$。

6.一种基于MF-SDAE的lncRNA-疾病关系预测系统，其特征在于，包括以下四个模块：

Ⅰ.特征矩阵构建模块，用于构建已知的lncRNA-疾病关系矩阵LD、lncRNA特征矩阵M_{lf}与疾病特征矩阵M_{df}；

构建已知的lncRNA-疾病关系矩阵的过程如下：

构建一个矩阵LD，其每一行对应一种lncRNA，每一列对应一种疾病，若有数据库记录了第i种lncRNA与第j种疾病存在关系，则将LD中第i行第j列的元素LD(i,j)设为1；否则将LD(i,j)设为0；由此得到的矩阵LD即为已知的lncRNA-疾病关系矩阵；

构建lncRNA特征矩阵的过程如下：

构建一个矩阵M_{lf}，其每一行对应一种lncRNA，每一列对应一种与lncRNA关联的信息，若有数据库记录了第i种lncRNA与第p种与lncRNA关联的信息存在关系，则将M_{lf}中第i行第p列的元素$M_{lf}(i,p)$设为1；否则将$M_{lf}(i,p)$设为0；由此得到的矩阵M_{lf}即为lncRNA特征矩阵；

构建疾病特征矩阵的过程如下：

构建一个矩阵M_{df}，其每一行对应一种疾病，每一列对应一种与疾病关联的信息，若有数据库记录了第j种疾病与第q种与疾病关联的信息存在关系，则M_{df}中第j行第q列的元素$M_{df}(j,q)$设为1；否则$M_{df}(j,q)$设为0；由此得到的矩阵M_{df}即为疾病特征矩阵；

Ⅱ.混合预测模型构建模块，用于构建基于MF-SDAE的混合预测模型；其中，MF为矩阵分解，SDAE为栈式降噪自动编码器；所述混合预测模型包括栈式降噪自动编码器模型和矩阵分解模型；其中，栈式降噪自动编码器模型用于对输入矩阵进行特征编码，提取高层特征；矩阵分解模型用于对输入矩阵进行分解，得到两个输出矩阵；

Ⅲ.模型训练模块，用于对混合预测模型进行训练；方法为：

利用矩阵分解模型对lncRNA-疾病关系矩阵LD进行分解，得到两个输出矩阵，即lncRNA特性矩阵L与疾病特性矩阵D；分解过程如下：

设定迭代次数T；

初始化lncRNA特性矩阵L与疾病特性矩阵D；

进行T次迭代，在每一次迭代过程中，按以下公式更新矩阵L和D：

$L(i,:)=LD(i,:)C^iD(\gamma' I+D^TC^iD)^{-1}$

$$D(j,:)=LD(:,j)^T\widetilde{C}^jL(\gamma' I+L^T\widetilde{C}^jL)^{-1}$$

其中，$L(i,:)$为矩阵L的第i行；$D(j,:)$为矩阵D的第j行，$LD(i,:)$为矩阵LD的第i行，C^i为第i种lncRNA对应的对角矩阵，其第j行第j列的元素值$C^i(j,j)=\beta_{i,j}$，$\beta_{i,j}$是偏好因子，$\beta_{i,j}=1+\theta\cdot LD(i,j)$，θ为自由参数；$\widetilde{C}^j$为第j种疾病对应的对角矩阵，其i行第i列的元素值$\widetilde{C}^j(i,i)=\beta_{i,j}$；$LD(:,j)$为lncRNA-疾病关系矩阵LD中的第j列；I是单位矩阵，γ′为自由参数；

T次迭代后得到的矩阵L和D即矩阵分解模型的输出矩阵；

初始化混合预测模型参数；

定义损失函数；以最小化损失函数值为目标，迭代求解混合预测模型的最优参数，得到训练好的混合预测模型；其中损失函数为：

$$Loss=\sum_{i,j}B_{i,j}[LD(i,j)-L(i,:)\cdot D(j,:)^{T2}+\gamma(\sum_i||L(i,:)||^2+\sum_j||D(j,:)||^2)+\gamma_l(||L-X_{encodes\_l}||^2)+\gamma_d(||D-X_{encodes\_d}||^2)+\gamma_{n\_l}(||M_{lf}-X_{out\_l}||^2)+\gamma_{n\_d}(||M_{df}-X_{out\_d}||^2)+\sum_k\gamma_k||W_k||^2+\sum_k\gamma_b||W_b||^2$$

$$\beta_{i,j}=1+\theta\cdot LD(i,j)$$

其中，||·||表示求2-范数，$\beta_{i,j}$是偏好因子；$LD(i,j)$为矩阵LD中第i行第j列的元素；$L(i,:)$为矩阵L的第i行；$D(j,:)$为矩阵D的第j行；θ、γ、γ_l、γ_d、γ_{n_l}、γ_{n_d}和γ_k均为自由参数；W_k和b_k分别为栈式降噪自动编码器中第k个隐藏层的权值矩阵和阈值向量；

每轮训练迭代过程中，先采用混合预测模型进行以下两部分数据处理：

采用混合预测模型中的栈式降噪自动编码器模型对lncRNA特征矩阵M_{lf}进行特征编码，

得到隐藏层和输出层输出的lncRNA特征编码矩阵，分别记为$X_{encodes_l}$和X_{out_l}；

采用混合预测模型中的栈式降噪自动编码器模型对疾病特征矩阵M_{df}进行特征编码，得到隐藏层和输出层输出的疾病特征编码矩阵，分别记为$X_{encodes_d}$和X_{out_d}；

然后根据混合预测模型的输入和输出计算相应的损失函数值；

Ⅳ.预测模块，用于预测各种lncRNA与各种疾病存在关系的可能性，方法为：

利用训练好的混合预测模型对lncRNA特征矩阵M_{lf}和疾病特征矩阵M_{df}进行处理，得到相应的lncRNA特征编码矩阵X_{encods_l}和疾病特征编码矩阵X_{encods_d}；

结合X_{encods_l}与D计算得分矩阵M_l，其第i行第j列的元素$M_l(i,j)$计算方法为：

$M_l(i,j)=X_{encods\_l}(i,:)\cdot D(j,:)^T$

其中，$X_{encods_l}(i,:)$表示X_{encods_l}的第i行，D(j,:)表示D的第j行；

结合X_{encods_d}与L计算得分矩阵M_d，其第i行第j列的元素$M_d(i,j)$计算方法为：

$M_d(i,j)=L(i,:)\cdot X_{encods\_d}(j,:)^T$

其中，L(i,:)表示L的第i行，$X_{encods_d}(j,:)$表示X_{encods_d}的第j行；

求M_l和M_d的加权平均值，所得结果即为预测得到的lncRNA-疾病关系得分矩阵LD′，其第i行第j列的元素表示预测得到的第i种lncRNA和第j种疾病存在关系的可能性。

基于MF-SDAE的lncRNA-疾病关系预测方法及系统

技术领域

[0001] 本发明涉及生物信息学领域，具体涉及一种基于MF-SDAE的lncRNA-疾病关系预测方法及系统。

背景技术

[0002] 随着生物技术和计算方法的飞速发展，越来越多的非编码RNA得到了鉴定，人们对非编码RNA的了解也越来越深入，最近的研究表明非编码RNA如长链非编码RNA(Long non-coding RNA，简称lncRNA)，在许多生物过程中都发挥着至关重要的作用。研究发现表明，lncRNA的异常不仅可以引起多种疾病，而且一种疾病的发生也有可能是多种lncRNA共同调节的结果，lncRNA可以用来作为衡量很多疾病产生的早期标志物。图1显示了lncRNA与疾病的调控网络，其中三角形和圆形分别表示lncRNA和疾病。从图中可以观察到，lncRNA的突变或失调都会引发相应疾病的产生。因此，识别lncRNA与疾病的关系，已成为医学界和病理学界研究的热点问题。但尽管人们已经发现了lncRNA与疾病间存在着关联，可是要确定与某种疾病的发生与发展最可能存在关系的lncRNA仍然是分子生物学家和遗传学家们的一大挑战。目前，在人类基因组发现的9万多条lncRNA中，只有不到1％的lncRNA有相关疾病报道，大量未知的(潜在的)lncRNA-疾病关系有待挖掘。

[0003] 在预测lncRNA与疾病是否存在关系的过程中，基于生传统物实验的方法成本非常高，消耗了大量的人力和时间，所以其应用受到了一定的限制。基于相似的lncRNA可能与相似的疾病存在关系的假设，一些基于计算的lncRNA-疾病关系预测算法被提了出来，这些基于计算的预测方法则有效地解决了基于生传统物实验的方法存在的问题。但现有的基于计算的预测方法虽然在预测潜在的lncRNA-疾病关系方面已取得了巨大成功，但还存在着一些缺陷。例如随着生物数据的快速增长，某些lncRNA和疾病出现了大量的特征数据，而现有的基于多特征的lncRNA-疾病关系预测算法或模型单一，或没有很好的处理数据的噪音，导致预测效果一般。因此，急需开发一种快速有效的基于计算的lncRNA-疾病关系预测算法及系统。

发明内容

[0004] 本发明所解决的技术问题是，针对现有技术的不足，提供一种基于MF-SDAE的lncRNA-疾病关系预测方法及系统，提高了lncRNA与疾病关系预测的准确性。

[0005] 本发明的技术方案为：

[0006] 一种基于双重反馈式矩阵分解及栈式降噪自动编码器的lncRNA与疾病关系预测方法，包括以下步骤：

[0007] 1)构建已知的lncRNA-疾病关系矩阵LD、lncRNA特征矩阵M_{lf}与疾病特征矩阵M_{df}；

[0008] 2)构建基于MF-SDAE，即矩阵分解和栈式降噪自动编码器的混合预测模型；所述混合预测模型包括栈式降噪自动编码器(SDAE)模型和矩阵分解(MF)模型；其中，栈式降噪自动编码器模型用于对输入矩阵进行特征编码，提取高层特征；所述矩阵分解(MF)模型用于

对输入矩阵进行分解，得到两个输出矩阵；

[0009] 3)利用矩阵分解模型对lncRNA-疾病关系矩阵LD进行分解，得到两个输出矩阵，即lncRNA特性矩阵L与疾病特性矩阵D；

[0010] 4)对混合预测模型进行训练；

[0011] 初始化混合预测模型参数；

[0012] 定义损失函数；以最小化损失函数值为目标，迭代求解混合预测模型的最优参数，得到训练好的混合预测模型；

[0013] 每轮迭代过程中，先采用混合预测模型进行以下两部分数据处理：

[0014] 采用混合预测模型中的栈式降噪自动编码器模型对lncRNA特征矩阵M_{lf}进行特征编码，得到隐藏层和输出层输出的lncRNA特征编码矩阵，分别记为$X_{encodesl}$和X_{out_l}；

[0015] 采用混合预测模型中的栈式降噪自动编码器模型对疾病特征矩阵Md_f进行特征编码，得到隐藏层和输出层输出的疾病特征编码矩阵，分别记为$X_{encodesd}$和X_{out_d}；

[0016] 然后根据混合预测模型的输入和输出计算相应的损失函数值；

[0017] 5)利用训练好的混合预测模型对lncRNA特征矩阵M_{lf}和疾病特征矩阵M_{df}进行处理，得到相应的lncRNA特征编码矩阵X_{encods_l}和疾病特征编码矩阵X_{encods_d}；

[0018] 结合X_{encods_l}与步骤3)中得到的D计算得分矩阵M_l，其第i行第j列的元素$M_l(i,j)$计算方法为：

[0019] $M_l(i,j)=X_{encods\_l}(i,:)\cdot D(j,:)^T$

[0020] 其中，$X_{encods_l}(i,:)$表示X_{encods_l}的第i行，$D(j,:)$表示D的第j行；

[0021] 结合X_{encods_d}与步骤3)中得到的L计算得分矩阵M_d，其第i行第j列的元素$M_d(i,j)$计算方法为：

[0022] $M_d(i,j)=L(i,:)\cdot X_{encods\_d}(j,:)^T$

[0023] 其中，$L(f,:)$表示L的第f行，$X_{encods_d}(j,:)$表示X_{encods_d}的第j行；

[0024] 求M_l和M_d的加权平均值，所得结果即为预测得到的lncRNA-疾病关系得分矩阵LD′，其第i行第j列的元素LD′(i,j)表示预测得到的第f种lncRNA和第j种疾病存在关系的可能性。

[0025] 进一步地，设M_l和M_d的取值均为0.5，得到$LD'(i,j)=\frac{M_l(i,j)+M_d(i,j)}{2}$。

[0026] 进一步地，所述步骤1)中，构建已知的lncRNA-疾病关系矩阵的过程如下：

[0027] 构建一个N×M的矩阵LD，其每一行对应一种lncRNA，每一列对应一种疾病，若有数据库记录了第i种lncRNA与第j种疾病存在关系，则将LD中第i行第j列的元素LD(i,j)设为1；否则将LD(i,j)设为0；其中i＝1,2,…,N；j＝1,2,…,M；N和M分别为lncRNA和疾病的种类数；由此得到的矩阵LD即为已知的lncRNA-疾病关系矩阵；

[0028] 构建lncRNA特征矩阵的过程如下：

[0029] 构建一个N×P的矩阵M_{lf}，其每一行对应一种lncRNA，每一列对应一种与lncRNA关联的信息，若有数据库记录了第i种lncRNA与第p种与lncRNA关联的信息存在关系，则将M_{lf}中第i行第p列的元素$M_{lf}(i,p)$设为1；否则将$M_{lf}(i,p)$设为0；其中i＝1,2,…,N；p＝1,2,…,P；N为lncRNA的种类数，P为与lncRNA关联的信息的种类数；由此得到的矩阵M_{lf}即为lncRNA特征矩阵；

[0030] 构建疾病特征矩阵的过程如下：

[0031] 构建一个M×Q的矩阵M_{df}，其每一行对应一种疾病，每一列对应一种与疾病关联的信息，若有数据库记录了第j种疾病与第q种与疾病关联的信息存在关系，则M_{df}中第j行第q列的元素$M_{df}(j,q)$设为1；否则$M_{df}(j,q)$设为0；其中j＝1,2,…,M；q＝1,2,…,Q；M为疾病的种类数，Q为与疾病关联的信息种类数；由此得到的矩阵M_{df}即为疾病特征矩阵。

[0032] 进一步地，所述与lncRNA关联的信息包括与lncRNA关联的基因信息、基因功能信息和miRNA信息。

[0033] 进一步地，所述与疾病关联的信息包括与疾病关联的基因信息和miRNA信息。

[0034] 进一步地，所述步骤3)具体过程如下：

[0035] 设定迭代次数T；

[0036] 初始化一个N×R的lncRNA特性矩阵L与一个M×R的疾病特性矩阵D；

[0037] 进行T次迭代，在每一次迭代过程中，按以下公式更新矩阵L和D：

[0038] $L(i,:)＝LD(i,:)C^{i}D(\gamma'I+D^{T}C^{i}D)^{-1}$

[0039] $$D(j,:) = LD(:,j)^{T}\widetilde{C}^{j}L(\gamma^{'}I + L^{T}\widetilde{C}^{j}L)^{-1}$$

[0040] 其中，L(i,:)为矩阵L的第i行；D(j,:)为矩阵D的第j行，LD(i,:)为矩阵LD的第i行，C^{i}为第i种lncRNA对应的对角矩阵，其第j行第j列的元素值$C^{i}(j,j)＝\beta_{i,j}$，$\beta_{i,j}$是偏好因子，$\beta_{i,j}＝1+\theta \cdot LD(i,j)$，θ为自由参数；$\widetilde{C}^{j}$为第j种疾病对应的对角矩阵，其i行第i列的元素值$\widetilde{C}^{j}(i,i) = \beta_{i,j}$；LD(:,j)为lncRNA-疾病关系矩阵LD中的第j列；I是单位矩阵，γ′为自由参数(根据经验取值)；

[0041] T次迭代后得到的矩阵L和D即矩阵分解模型的输出矩阵。

[0042] 进一步地，将lncRNA特性矩阵L与疾病特征矩阵D初始化为服从0～1均匀分布的随机矩阵，即产生[0,1)上均匀分布的随机数，来填充L和D，完成L和D的初始化。

[0043] 进一步地，所述栈式降噪自动编码器模型(SDAE)包括依次连接的一个输入层、一个损坏层、三个隐藏层和一个输出层；栈式降噪自动编码器模型对lncRNA特征矩阵M_{lf}进行特征编码时，将其第二个隐藏层的输出作为$X_{encodes_l}$；栈式降噪自动编码器模型对疾病特征矩阵M_{df}进行特征编码时，将其第二个隐藏层的输出作为$X_{encodes_d}$；设栈式降噪自动编码器模型第二个隐藏层中神经元个数为R，则$X_{encodes_l}$为N×R的矩阵，$X_{encodes_d}$为M×R的矩阵。

[0044] 进一步地，所述步骤4)中，损失函数为：

[0045] $Loss＝\sum_{i,j}\beta_{i,j}[LD(i,j)-L(i,:)\cdot D(j,:)^{T}]^{2}+\gamma(\sum_{i}||L(i,:)||^{2}+\sum_{j}||D(j,:)||^{2})+\gamma_{l}(||L-X_{encodes\_l}||^{2})+\gamma_{d}(||D-X_{encodes\_d}||^{2})+\gamma_{n\_l}(||M_{lf}-X_{out\_l}||^{2})+\gamma_{n\_d}(||M_{df}-X_{out\_d}||^{2})+\sum_{k}\gamma_{k}||W_{k}||^{2}+\sum_{k}\gamma_{b}||W_{b}||^{2}$

[0046] $\beta_{i,j}＝1+\theta \cdot LD(i,j)$

[0047] 其中，||·||表示求2-范数，$\beta i_{,j}$是偏好因子；LD(i,j)为矩阵LD中第i行第j列的元素；L(i,:)为矩阵L的第i行；D(j,:)为矩阵D的第j行；θ、γ、γ_{l}、γ_{d}、γ_{n_l}、γ_{n_d}和γ_{k}均为自由参数(根据经验取值)；W_{k}和b_{k}分别为栈式降噪自动编码器中第k个隐藏层的权值矩阵和阈值向量(需要优化的参数)。

[0048] 进一步，所述步骤4)中，迭代求解混合预测模型的最优参数采用小批量梯度下降算法。

[0049] 有益效果：

[0050] 本发明提出了一种基于矩阵分解与栈式降噪自动编码器相结合的lncRNA(长链非编码RNA)与疾病关系预测方法和系统。该方法是基于相似的疾病可能与相似的lncRNA存在关系的假设实施的。首先充分利用多个lncRNA数据库及多个疾病数据库，提取lncRNA的多种特征以及疾病的多种特征，构建已知的lncRNA-疾病关系矩阵、lncRNA特征矩阵与疾病特征矩阵，以全面描述lncRNA与疾病关系；在使用矩阵分解模型来对已知的lncRNA-疾病关系矩阵进行分解得到lncRNA特性矩阵和疾病特性矩阵之后，把分解得到的特性矩阵输入到栈式降噪自动编码器，协助lncRNA特征矩阵和疾病特征矩阵进行编码得到各自的编码矩阵(即对多特征数据进行降维编码、学习更复杂的高层特征)，然后结合矩阵分解模型与栈式降噪自动编码器模型的结果，计算损失函数值，通过损失函数，利用矩阵分解模型生成的特性矩阵监督栈式降噪自动编码器的特征编码，以达到防止机器学习冷启动的效果，最终把训练好的编码矩阵和特性矩阵进行矩阵乘法操作得到lncRNA-疾病关系打分矩阵，打分矩阵中的元素值即预测得到的各种lncRNA与各种疾病存在关系的可能性。所述系统用于实现上述预测方。本发明简单有效，通过使用十折交叉验证法、De novo交叉验证法和案例分析对本发明提出的方法和系统进行测试，结果表明该方法和系统在预测潜在的(未知的)lncRNA-疾病关系方面具有较好的预测性能。

附图说明

[0051] 图1为LncRNA-疾病调控网络；其中上半部分为正常的lnRNA-疾病相互作用网络，下半部分为lncRNA突变或扰动网络；

[0052] 图2为本发明实施例流程图；

[0053] 图3为本发明实施例中矩阵分解-栈式降噪自动编码模型；

[0054] 图4为本发明(CDLLD)和其他方法基于十倍交叉验证的ROC曲线及相应的AUC值；

[0055] 图5为本发明(CDLLD)和其他方法基于De novo实验测试的ROC曲线及相应的AUC值；

具体实施方式

[0056] 如图2所示，本实施例具体实现过程如下：

[0057] 一、构建已知的lncRNA-疾病关系矩阵、lncRNA特征矩阵与疾病特征矩阵

[0058] 随着高通量测序技术的快速发展，产生了大量的生物数据，为了存储和管理方便，人们建立了标准的数据库用来存储这些生物数据。例如由马由里兰大学医学院主办创建的Disease Ontology人类疾病数据库、人类基因和遗传疾病知识库Online Mendelian Inheritance in Man(OMIM)、人类lncRNA的综合数据库LNCipedia、包含16个物种的非编码RNA数据库NONCODE、真核生物的lncRNA数据库lncRNAdb以及主要记录哺乳动物相关的非编码RNA与疾病的关联信息的数据库MNDR等。随着越来越多的lncRNA相关数据库和疾病相关数据库的建立和规范化，使基于计算的方法来预测未知的lncRNA与疾病关系成为可能。本实施例充分提取了lncRNA的多种特征以及疾病的多种特征，以全面描述lncRNA与疾病关系。

[0059] 1.已知的lncRNA-疾病关系提取

[0060] 首先对存储lncRNA信息和存储疾病相关信息的相关数据库进行数据下载，对多个数据库中记录的已知lncRNA-疾病关系进行统计、去重整理，找出已知的lncRNA-疾病关系(经传统生物实验证实的lncRNA-疾病关系)；

[0061] 本实施例通过对LncRNADisease数据库、Lnc2Cancer数据库以及GeneRIF数据库中记录的已知的lncRNA-疾病关系进行统计、去重整理，最后从中获取了240种lncRNA、412种疾病以及它们所对应的2697对已知的lncRNA-疾病关系(即已知存在关系的lncRNA-疾病对有2697个)。其中，本实施例创建了N×M的lncRNA-疾病关系矩阵LD来存储这些已知关系。如果已有记录表明第i种lncRNA与第j种疾病存在关系，则将LD(i,j)置为1，否则将LD(i,j)置为0，其中f＝1,2,…,N;j＝1,2,…,M;N和M分别为lncRNA和疾病的种类数，本实施例中N＝240,M＝412。

[0062] 2.lncRNA特征提取

[0063] 本实施例对多个数据库中与lncRNA关联的信息(包括已知的与lncRNA关联的基因信息、基因功能信息和miRNA信息)进行整合，把每一项与lncRNA关联的信息都作为一项lncRNA特征信息，得到lncRNA特征矩阵。本实施例从lncRNA2target数据库中提取得到了与lncRNA关联的基因信息，从GeneRIF数据库中提取得到了与lncRNA关联的基因功能信息，从starBase数据库中提取得到了与lncRNA关联的miRNA信息。通过去重整合后，本实施例一共获取了6066维lncRNA特征数据。为了管理这些特征数据，本实施例创建了N×P的lncRNA特征矩阵M_{lf}来存储它们，如果数据库中记录了第f种lncRNA与第p种与lncRNA关联的信息(第p维特征)存在关系，则把M_{lf}(f,j)设为1，如果还没有数据库记录证明它们有关系，则把M_{lf}(f,j)设为0，其中f＝1,2,…,N;p＝1,2,…,P;N为lncRNA的种类数，P为与lncRNA关联的信息种类数，本实施例中N＝240,P＝6066。

[0064] 3.疾病特征提取

[0065] 本实施例对多个数据库中与疾病关联的信息(包括已知的与疾病关联的基因信息和miRNA信息)进行整合，把每一项与疾病关联的信息作为一项疾病特征信息，得到疾病特征矩阵。其中我们从DisGeNet数据库中提取得到了与疾病关联的基因信息，从HMDD数据库中提取得到了与疾病关联的miRNA信息。通过去重整合后，本实施例一共获得了10621维疾病特征数据。同样的，为了存储这些特征数据，本实施例创建了M×Q的疾病特征矩阵M_{df}，如果数据库中记录了第j种疾病与第q种与疾病关联的信息存在关系(第q维特征)存在关系，则把M_{df}(j,q)设为1，否则把M_{df}(j,q)设为0，其中j＝1,2,…,M;p＝1,2,…,Q;M为疾病的种类数，Q为与疾病关联的信息种类数，本实施例中M＝412,Q＝10621。

[0066] 二、构建栈式降噪自动编码器模型

[0067] 自动编码器是一种自监督的机器学习算法，或者说是一种尽可能复现原始输入信号的神经网络。其算法的基本思想是：通过不断迭代，不断调整自编码器的参数，得到每一层中的权重，来使输出的信息尽可能与输入编码器的信息相同。为了实现这种复现，自动编码器就必须要捕捉可以代表输入数据的最重要的因素，即找到可以代表原信息的主要成分。自动编码器可用于数据压缩和从输入数据中提取有用的“高层”特征。降噪自动编码器是一类可以接受损坏数据作为输入，并通过训练来预测原始未被损坏数据作为输出的自编码器，其核心思想是能够从损坏的数据中还原原始数据的自编码器所学到的特征才是最好的。而设计多层编码器有利于获得更优秀的高层特征，所以本实施例设计了具有三层隐藏

层的栈式降噪自动编码器(SDAE)。

[0068] 本实施例使用栈式降噪自动编码器(SDAE)对lncRNA的特征信息与疾病的特征信息分别进行特征编码,提取高层特征,即将lncRNA特征信息和疾病特征信息转换成R维的高层特征(本实施例中设置R=100)。本实施例中的栈式降噪自动编码器模型如图3所示。其中,X_input是输入层,输入lncRNA或疾病的特征矩阵(M_{lf}或M_{lf}),X_input_noise是对原始数据(X_input)进行加高斯噪声处理的"损坏"层,X_1、X_encodes、X_3层是3个隐藏层(本实施例从第2个隐藏层X_encodes提取lncRNA或疾病特征信息的编码特征,即"高层"特征数据,第2个隐藏层的神经元个数设为R个,其它两个隐藏层的神经元个数大于等于R),X_out为输出层。本实施例使用小批量梯度下降算法(Mini-Batch Gradient Descent)来训练栈式降噪自动编码器模型,其中批量大小(Batch_size)设置为60(即每批包括60个样本)。

[0069] 三、矩阵分解模型

[0070] 本实施例使用了一种监督式矩阵分解模型来对已知的lncRNA-疾病关系矩阵进行分解,所谓"监督"即利用已知的lncRNA-疾病关系来反馈模型(通过损失函数实现监督功能),使模型具有一定的记忆功能,通过矩阵分解将已知的lncRNA-疾病关系矩阵分解为疾病特性矩阵以及lncRNA特性矩阵。上述部分中,定义了lncRNA-疾病关系矩阵为LD,经过矩阵分解算法后其将被分解成为对应R维"高层特性"的N×R的lncRNA特性矩阵L以及M×R的疾病特性矩阵D,矩阵L中每一行表示一种lncRNA的潜在因子向量,其中第f行表示为L(i,:),即第f种lncRNA的潜在因子向量,矩阵D中每一行表示一种疾病的潜在因子向量,其中第j行表示为D(j,:),即第j种疾病的潜在因子向量。可以通过$L(i,:)\cdot D(j,:)^T$来计算lncRNA i和疾病j存在关系的可能性得分。其损失函数定义如下:

[0071] $$l=\sum_{i,j}\beta_{i,j}[LD(i,j)-L(i,:)\cdot D(j,:)^T]^2+\gamma(\sum_i||L(i,:)||^2+\sum_j||D(j,:)||^2) \quad (1)$$

[0072] $$\beta_{i,j}=1+\theta\cdot LD(i,j) \quad (2)$$

[0073] 其中,||·||表示求2-范数,γ和θ都是一个自由参数,本实施例中都设定为100,$\beta_{i,j}$是偏好因子,目的是加强对已知的lncRNA-疾病关系在模型中的比重,监督模型以提高模型质量。

[0074] 在矩阵分解模型的每一次迭代过程(本实施例设定迭代次数T为30)中,其使用公式(4)和(5)来更新lncRNA特性矩阵L和疾病特性矩阵D。

[0075] $$L(i,:)=LD(i,:)C^iD(\gamma' I+D^TC^iD)^{-1} \quad (3)$$

[0076] 其中,C^i为第i种lncRNA对应的对角矩阵,其第j行第j列的元素值$C^i(j,j)=\beta_{i,j}$;LD(i,:)为lncRNA-疾病关系矩阵中的第i行,即第i种lncRNA与所有疾病的关系向量;I是R阶单位矩阵,γ′被设为100。

[0077] $$D(j,:)=LD(:,j)^T\widetilde{C}^jL(\gamma' I+L^T\widetilde{C}^jL)^{-1} \quad (4)$$

[0078] 其中,$\widetilde{C}^j$为第j种疾病对应的对角矩阵,其第i行第i列的元素值$\widetilde{C}^j(i,i)=\beta_{i,j}$;LD(:,j)为lncRNA-疾病关系矩阵中的第j列,即第j种疾病与所有lncRNA的关系向量;I是R阶单位矩阵,γ′被设为100。

[0079] T轮迭代过后,将更新好的lncRNA特性矩阵L和疾病特性矩阵D输出给栈式降噪自动编码器,栈式降噪自动编码器根据新的L和D来更新自身参数。

[0080] 四、构建基于双重反馈式矩阵分解-栈式降噪自动编码器的混合预测模型

[0081] 本实施例构建了基于双重反馈式矩阵分解-栈式降噪自动编码器的混合预测模型，通过该混合预测模型来预测未知的lncRNA-疾病关系。混合预测模型的损失函数由矩阵分解的损失函数以及栈式降噪自动编码器的损失函数组合构成。所谓“双重”即基于lncRNA特征信息的矩阵分解-栈式降噪自动编码(SDAE-1)以及基于疾病特征信息的矩阵分解-栈式降噪自动编码(SDAE-2)的融合，对SDAE-1预测得到的lncRNA-疾病关系打分矩阵与SDAE-2预测得到的lncRNA-疾病关系打分矩阵进行求均值来得到最终的未知lncRNA-疾病关系预测得分。所以，在混合预测模型中，这些未知的lncRNA-疾病关系预测将依赖于已知的lncRNA-疾病关系信息、lncRNA的特征信息以及疾病的特征信息，而不是单一的已知lncRNA-疾病关系信息。“反馈式”即通过损失函数，利用矩阵分解模型生成的特性矩阵影响栈式降噪自动编码器的特征编码。lncRNA特征矩阵/疾病特征矩阵经栈式降噪自动编码器训练后生成的特征编码$X_{encodes_l}$/$X_{encodes_l}$最后与矩阵分解模型得到的D/L进行矩阵乘法操作，即$X_{encods_l}(i,:)\cdot D(j,:)^T$/$L(i,:)\cdot X_{encods_d}(j,:)^T$来获得lncRNA-疾病关系预测得分$M_l(i,j)/M_d(i,j)$。栈式降噪自动编码器不仅在输出层$X_{out}$处重新构建输入$X_{input}$，而且还寻找最佳特征编码$X_{encodes}$，以便最小化损失函数。

[0082] 在具体实现时，可以采用两个混合预测模型，两个混合预测模型一起运行，一个混合预测模型进行基于lncRNA特征信息(M_{lf})的矩阵分解-栈式降噪自动编码(SDAE-1)，其损失函数可以定义为：

[0083] $$Loss_1=\Sigma_{i,j}\beta_{i,j}[LD(i,j)-L(i,:)\cdot D(j,:)^T]^2+\gamma(\Sigma_i||L(i,:)||^2+\Sigma_j||D(j,:)||^2)+\gamma_l(||L-X_{encodes\_l}||^2)+\gamma_n(||X_{input}-X_{out}||^2)+\Sigma_k\gamma_w||W_{k1}||^2+\Sigma_k\gamma_b||b_{k1}||^2 \quad (5)$$

[0084] 其中，前面两部分是矩阵分解的损失函数；第三部分是最小化栈式降噪自动编码器编码得到的lncRNA特征编码矩阵$X_{encodes_l}$与矩阵分解得到的lncRNA特性矩阵间的误差值；第四部分是栈式降噪自动编码器重构得到的lncRNA特征信息X_{out}与原输入的特征信息X_{input}(M_{lf})的误差值，其中γ_l与γ_n为自由参数，本实施例中设定它们的比值γ_l/γ_n为500。最后两部分分别为所有隐藏层和输出层权值和阈值的正则化项，其中W_{k1}为栈式降噪自动编码器中第k个隐藏层的权值矩阵，本实施例中设置三个隐藏层，即k=1,2,3；b_{k1}为栈式降噪自动编码器中第k个隐藏层的阈值向量，γ_w和γ_b为自由参数，本实施例中它们都设为200。

[0085] 另一个混合预测模型进行基于疾病特征信息(M_{df})的矩阵分解-栈式降噪自动编码(SDAE-2)，其损失函数可以定义为：

[0086] $$Loss_2=\Sigma_{i,j}\beta_{i,j}[LD(f,j)-L(i,:)\cdot D(j,:)^T]^2+\gamma(\Sigma_i||L(i,:)||^2+\Sigma_j||D(j,:)||^2)+\gamma_d(||D-X_{encodes\_d}||^2)+\gamma_n(||X_{input}-X_{out}||^2)+\Sigma_k\gamma_w||W_{k2}||^2+\Sigma_k\gamma_b||b_{k2}||^2 \quad (6)$$

[0087] 其中，前面两部分是矩阵分解的损失函数；第三部分是最小化栈式降噪自动编码器编码得到的疾病特征编码矩阵X_{encode_d}与矩阵分解得到的疾病特性矩阵间的误差值；第四部分是栈式降噪自动编码器重构得到的疾病特征信息X_{out}与原输入的特征信息X_{input}(M_{df})的误差值，其中γ_d/γ_n为自由参数，本实施例中它们的比值(γ_l/γ_n)设为500；最后两部分分别为所有隐藏层和输出层权值和阈值的正则化项，其中W_{k2}为栈式降噪自动编码器中第k个隐藏层的权值矩阵，b_{k2}为栈式降噪自动编码器中第k个隐藏层的阈值向量，γ_w和γ_b为自由参数，本实施例中它们都设为200。

[0088] 在具体实现时，也可以采用同一个混合预测模型，先后进行基于lncRNA特征信息

(M_{lf})的栈式降噪自动编码-矩阵分解(SDAE-1)和基于疾病特征信息(M_{df})的栈式降噪自动编码-矩阵分解(SDAE-2),其损失函数可以定义为:

[0089] $Loss=\sum_{i,j}\beta_{i,j}[LD(f,j)-L(f,:)\cdot D(j,:)^{T}]^{2}+\gamma(\sum_{i}||L(i,:)||^{2}+\sum_{j}||D(j,:)||^{2})+\gamma_{l}(||L-X_{encodes\_l}||^{2})+\gamma_{d}(||D-X_{encodes\_d}||^{2})+\gamma_{n\_l}(||M_{lf}-X_{out\_l}||^{2})+\gamma_{n\_d}(||M_{df}-X_{out\_d}||^{2})+\sum_{k}\gamma_{k}||W_{k}||^{2}+\sum_{k}\gamma_{b}||W_{b}||^{2}$

[0090] 本实施例使用小批量梯度下降算法来训练栈式降噪自动编码器.

[0091] 训练完毕后,先利用训练好的混合预测模型对lncRNA特征矩阵M_{lf}和疾病特征矩阵M_{df}进行处理,得到相应的lncRNA特征编码矩阵X_{encods_l}和疾病特征编码矩阵X_{encods_d};

[0092] 然后计算:

[0093] $M_{l}(i,j)=X_{encods\_l}(i,:)\cdot D(j,:)^{T}$

[0094] $M_{d}(i,j)=L(i,:)\cdot X_{encods\_d}(j,:)^{T}$

[0095] 其中,M_{l}是基于lncRNA特征信息的矩阵分解-栈式降噪自动编码(即基于SDAE-1输出的$X_{encods_l}(i,:)$)预测得到的lncRNA-疾病关系打分矩阵;M_{d}是基于疾病特征信息的矩阵分解-栈式降噪自动编码(即基于SDAE-2输出的$X_{encods_d}(j,:)$)预测得到的lncRNA-疾病关系打分矩阵;

[0096] 最后,未知的lncRNA i与疾病j关系的最终预测得分LD′(i,j)可以定义为:

[0097] $$LD'(i,j)=\frac{M_{l}(i,j)+M_{d}(i,j)}{2} \quad (7)。$$

[0098] 五、实验验证

[0099] 1.评价指标

[0100] 为了验证CDLLD方法的预测有效性,本节使用十折交叉验证法(10-Fold Cross Validation)、De novo交叉验证法来对方法进行测试。

[0101] (1)十折交叉验证法

[0102] 所谓十折交叉验证法就是把数据集中已知的lncRNA-疾病关系分成十份,每一次取一份作为测试集,其他九份作为训练集,然后进行轮转试验。因此,对于给定的第i种疾病,每一对已知的与i存在关系的lncRNA-疾病关系对会被轮流移除(LD中相应元素置为0),作为测试集,其它的已知关系作为训练集。然后,根据训练的模型对测试样本和未标记的与第i种疾病相关的lncRNA样本进行评分并按降序排列。lncRNA的排名越高,说明其与第i种疾病存在关系的可能性就越大。最后,把每一个排名当做阈值来计算真阳性概率TPR(True-positive rate)和假阳性概率FPR(False-positive rate)。本节对FPR和TPR的定义如下:

[0103] $$FPR=\frac{FP}{FP+TN} \quad (8)$$

[0104] $$TPR=\frac{TP}{TP+FN} \quad (9)$$

[0105] 其中,TP(True positive)代表排序高于阈值的正样本数量,FN(False negative)代表正样本被错误识别为负样本的数量,FP(False positive)代表排名高于阈值的负样本数量,TN(True negative)代表负样本被正确分类为负样本的数量。

[0106] 基于所有的TPR和FPR值,画出了CDLLD的ROC曲线图(受试者工作特征曲线,Receiver Operating Characteristic Curve)。其横轴代表的是假阳性概率(FPR),纵轴代表的是真阳性概率(TPR)。进一步的,计算了ROC曲线与横轴的面积即AUC值(Area Under

Curve)来衡量算法的性能。如果AUC值为0.5,则说明该算法的预测结果是随机的,相反,如果AUC的值为1,则说明该算法的预测性能是最好的。

[0107] (2)De novo交叉验证法

[0108] 在实际数据中,有很多疾病研究者们至今还未找到与之关联的任何lncRNA,即该疾病没有任何与lncRNA关联的先验信息。为了验证本发明提出的CDLLD算法在疾病没有任何已知lncRNA关系信息时,对预测未知的lncRNA-疾病关系的性能,本节将CDLLD进行了De novo测试。

[0109] 类似十折交叉验证方法,De novo测试是指每次把特定对象的所有正例样本删除作为训练集,保留其它对象的正例样本作为训练集。在完成轮转测试后,我们也计算其TPR和FPR值,并画出ROC曲线,求出AUC值。

[0110] 2.与其它方法的比较

[0111] 为了评价CDLLD的有效性,本节将其与其他两种方法(SIMLDA、MFLDA)进行比较。SIMLDA通过使用主成分分析(PCA)来提取lncRNA和疾病的主要特征向量,然后通过诱导矩阵填充来预测lncRNA-疾病关系;MFLDA主要是通过矩阵分解来预测潜在的lncRNA-疾病关系。

[0112] (2)十折交叉验证法结果分析

[0113] 十折交叉验证的结果如图4所示,从结果可以看出,CDLLD、SIMLDA以及MFLDA的AUC值分别为0.9134、0.8259、0.6430,其中CDLLD的AUC值明显高于其他两种方法,说明了本发明提出的CDLLD算法可以显著提高对潜在lncRNA-疾病关系的预测性能。

[0114] (3)De novo交叉验证法结果分析

[0115] De novo交叉验证法的结果如图5所示,从结果可以看出,CDLLD、SIMLDA和MFLDA的AUC值分别为0.8917、0.7923、0.5952。结果说明了CDLLD在疾病无任何已知lncRNA关系的先验情况下也有较好的预测性能。

[0116] (4)案例分析

[0117] 为了进一步验证CDLLD在预测未知的lncRNA-疾病关系上的性能,本节选取了由CDLLD预测得到的前10种与骨肉瘤疾病(骨肉瘤是一种常见的恶性骨肿瘤,根据报道其已成为年轻人癌症相关死亡的第二大原因)相关的lncRNA进行分析,其具体结果如表1所示。从表1可以看出,在这10种lncRNA中有9种在最近的科学文献中得到了验证,表明了CDLLD具有较高的预测准确性。

[0118] 表1.案例分析结果

[0119]

疾病	排名前 10 的 lncRNA	排序	证据
Osteosarcoma	H19	1	PMID:28975992
	PVT1	2	PMID:27813492
	GAS5	3	PMID:28519068
	NEAT1	4	PMID:29416922
	KCNQ1OT1	5	DOI:10.1039/C8RA07209D
	MIR155HG	6	
	AFAP1-AS1	7	PMID:29901121
	XIST	8	PMID:28409547
	CCAT1	9	PMID:28549102
	SPRY4-IT1	10	PMID: 28078006

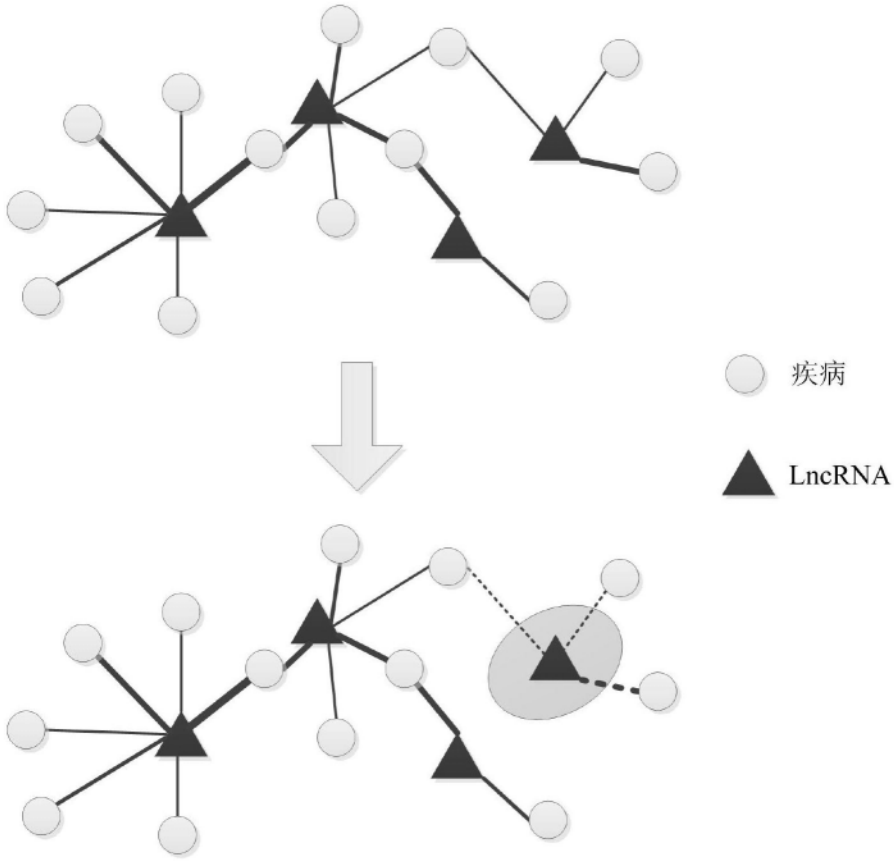

疾病
LncRNA

图1

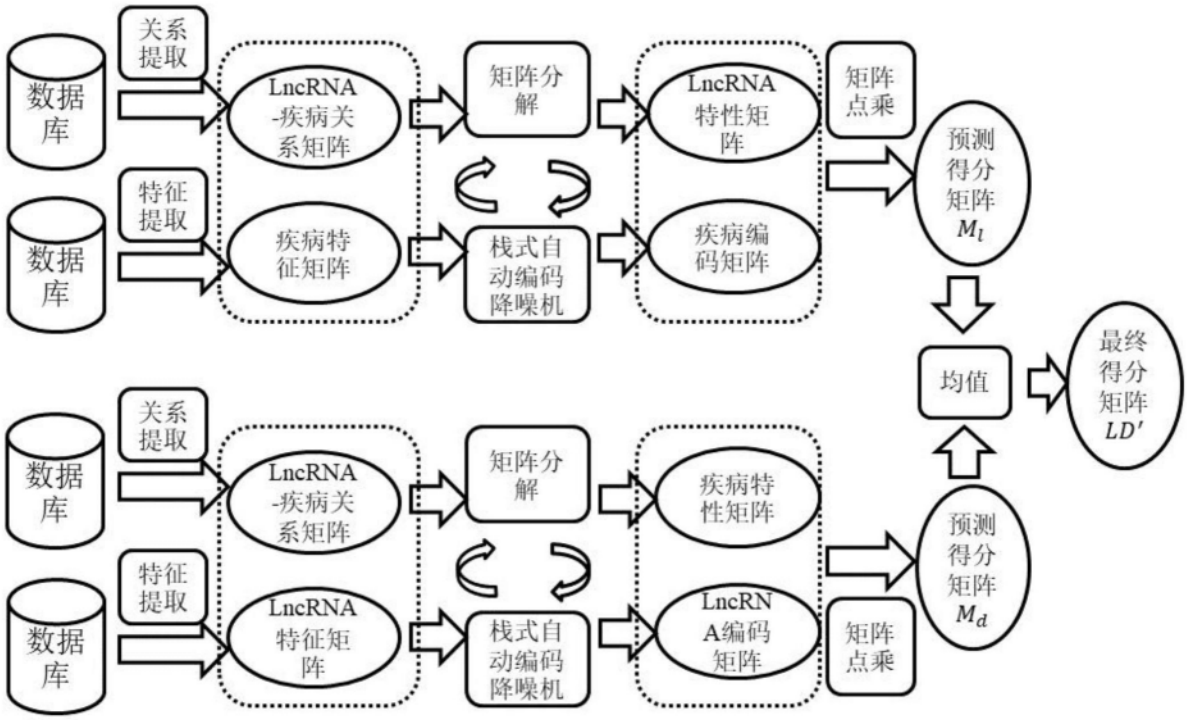

数据库
关系提取
LncRNA-疾病关系矩阵
矩阵分解
LncRNA特性矩阵
矩阵点乘
预测得分矩阵 M_l
数据库
特征提取
疾病特征矩阵
栈式自动编码降噪机
疾病编码矩阵
均值
最终得分矩阵 LD'
数据库
关系提取
LncRNA-疾病关系矩阵
矩阵分解
疾病特性矩阵
预测得分矩阵 M_d
数据库
特征提取
LncRNA特征矩阵
栈式自动编码降噪机
LncRNA编码矩阵
矩阵点乘

图2

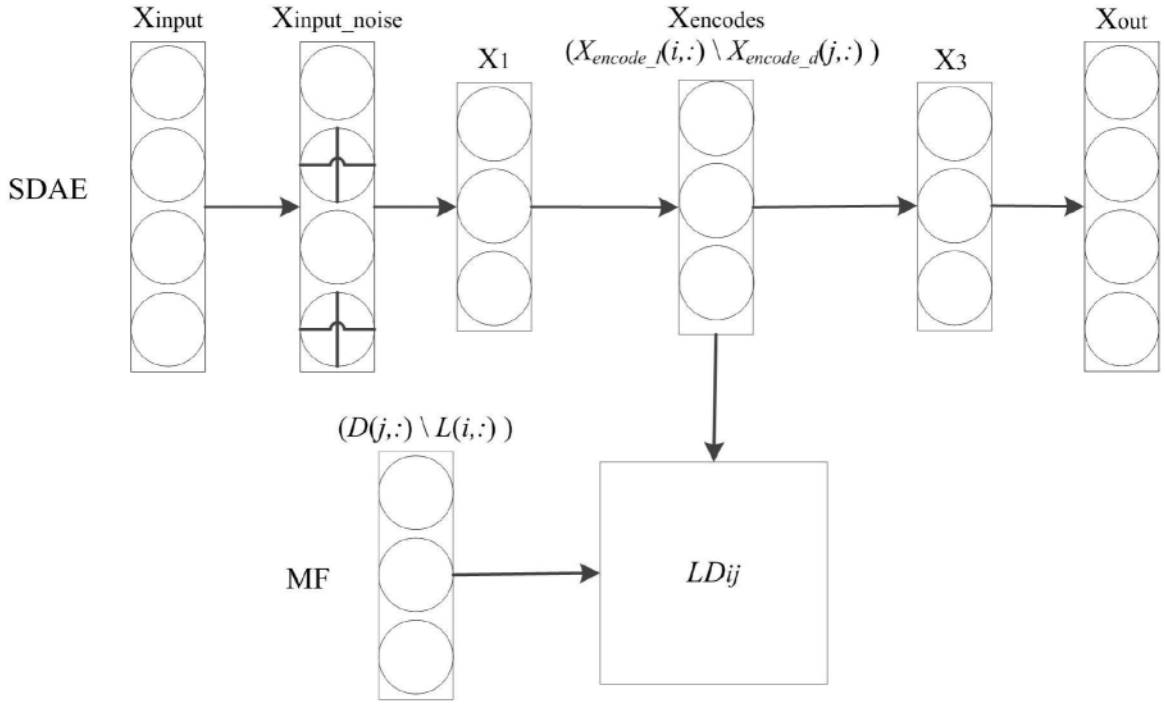

Xinput
Xinput_noise
X1
Xencodes
(Xencode_I(i,:) \ Xencode_d(j,:))
X3
Xout
SDAE
(D(j,:) \ L(i,:))
MF
LDij

图3

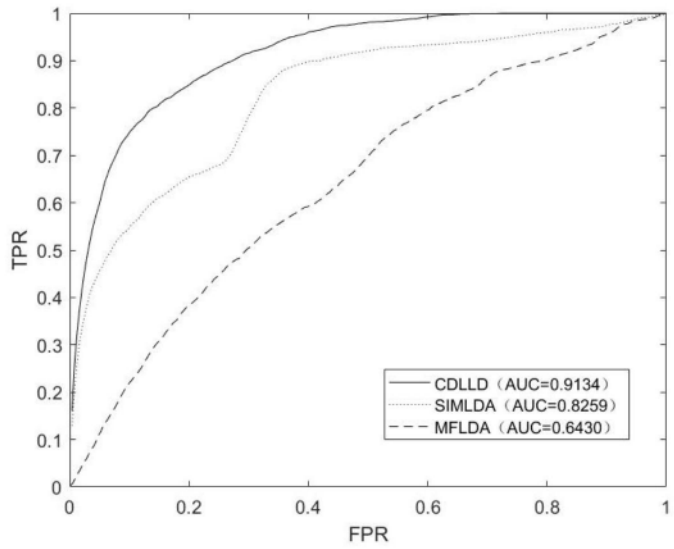

TPR
FPR
CDLLD（AUC=0.9134）
SIMLDA（AUC=0.8259）
MFLDA（AUC=0.6430）

图4

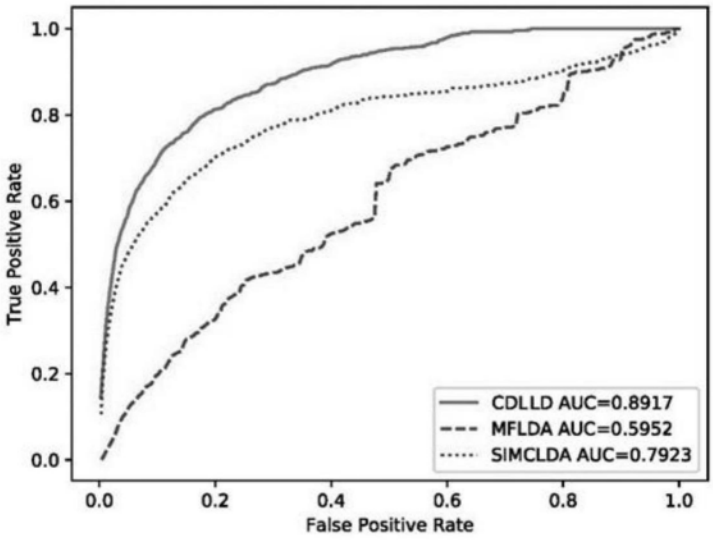

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CDLLD AUC=0.8917
MFLDA AUC=0.5952
SIMCLDA AUC=0.7923

图5