摘 要

MiRNA(MicroRNA)是一类在真核细胞中普遍存在的长约 18-25nt 的单链小RNA，miRNA 在细胞的增殖、分化和凋亡的过程中起到了十分重要的调节作用。近年来，随着人们研究的深入，越来越多的证据表明 miRNA 与疾病的发生之间有着密切的关系因此，探究 miRNA 与疾病的关联已经成为许多研究者的研究主题。近年来，随着计算机技术以及生物信息学的发展，通过计算机的方法进行仿真模拟从而计算出 miRNA 和疾病的关联已经成成为目前研究的焦点，通过这种方法可减轻生物实验人员的工作量，也可以尽早将 miRNA 用于疾病判断，开发 miRNA 靶向药物用于临床治疗。

本文通过多个数据库中的数据得到疾病相似性网络、基因-基因关联网络、miRNA 相似性网络、miRNA-基因网络以及疾病-基因网络，并通过以上五个网络得到 miRNA-基因-疾病三层网络。接着从 miRNA-基因-疾病三层网络中提取 miRNA-基因、疾病-基因特征，这一步的总体思路是基于一种创新的回归模型求得不同 miRNA 之间的相似性得分和不同疾病之间的相似性得分，接着求得 miRNA(或疾病)与不同基因之间的相似性得分。使用堆栈自编码器对得到的特征向量进行降维和去噪处理。最后使用卷积神经网络对特征向量进行分类处理，采用十折交叉验证法得到最终分类处理的结果，采用多种评估标准进行评估并与其它模型的实验结果进行比较。实验结果表明，本文提出的方法可以有效且快速的预测 miRNA 与疾病之间的关联。

关键词：miRNA 自编码器 卷积神经网络 预测 miRNA-疾病关联

ABSTRACT

MiRNA (MicroRNA) is a type of single-stranded small RNA of about 18-25 nts that is ubiquitous in eukaryotic cells. MiRNA plays a very important role in regulating cell proliferation, differentiation, and apoptosis. In recent years, with the deepening of people's research, more and more evidence shows that there is a close relationship between miRNA and disease occurrence. Therefore, exploring the relationship between miRNA and disease has become the research topic of many researchers. In recent years, with the development of computer technology and bioinformatics, the simulation of computer methods to calculate the relationship between miRNA and disease has become the focus of current research. This method can reduce the workload of biological experimenters. It is also possible to use miRNA for disease judgment as soon as possible, and to develop miRNA targeted drugs for clinical treatment.

In this thesis, we obtain the disease similarity network, gene-gene association network, miRNA similarity network, miRNA-gene network and disease-gene network from the data in multiple databases, and obtain the three-layer miRNA-gene-disease network through the above five networks . Then extract the miRNA-gene and disease-gene characteristics from the three-layer network of miRNA-gene-disease. The general idea of this step is based on an innovative regression model to find the similarity score between different miRNAs and between different diseases. Similarity score, and then find the similarity score between miRNA (or disease) and different genes. Using a stack autoencoder, the obtained feature vectors are processed for dimensionality reduction and denoising. Finally, the convolutional neural network is used to classify the feature vectors. The ten-fold cross-validation method is used to obtain the final classification processing results. Various evaluation criteria are used to evaluate and compare with the experimental results of other models. Experimental results show that the proposed method can effectively and quickly predict the association between miRNA and disease.

Keywords: miRNA autoencoder convolutional neural network prediction miRNA-disease association

ABSTRACT

目 录

第一章 绪论

1.1 研究背景及意义

21 世纪可以称得上是生命科学的时代，同时它也是信息的时代。随着近年来生物信息量的“大爆炸”，面对如此庞大的信息量，如何尽可能快速全面的分析这些信息成了目前我们需要解决的一大问题。计算机具有计算速度快，逻辑能力强以及存储容量大等特点，因此在生命科学的研究过程中，将计算机作为为工具对生物信息进行储存，并且对信息进行检索、分析变得越来越常见。对于 miRNA 和疾病关联这一问题，我们可以利用现有的一些数据库，运用某种算法，通过计算机进行仿真模拟，就能在短时间内找到 miRNA 和疾病的潜在关联，为生物实验的工作人员提供重要的参考信息，同时可以大大减少他们的工作量。

MiRNA 是一种存在于真核细胞中的物质，miRNAs 是一类单链小 RNA，它们的长度约 18-25nt，由细胞内源产生的发卡结(约70个碱基大小的单链 RNA 前体)经过 Dicer 酶加工处理后生成的[1]。microRNA 是在1993年被首次发现，由维克托·安布罗斯和加里·鲁夫昆分别领导的实验室在线虫中发现了一种名为 lin-4 的基因[2]。由于这种基因并不编码蛋白，而是表达一种长度为 22nt 的小 RNA，因此这一发现并未得到重视，只是当成一种特例的存在，直到在 2001 年 10 月，托马斯·图施勒、大卫·巴特尔和维克多·安布罗斯三人发文，将这种单链小 RNA 命名为 microRNA(微小核糖核酸)，简称 miRNA[3]。近年来，随着生物学的发展，人们逐渐认识到 miRNA 在细胞的生命活动比如增殖、分化以及凋亡的过程中有着十分重要的调节作用，越来越多的人们投入到了对 miRNA 的研究当中，到目前为止，已被发现的 miRNA 从 2008 年的 218 个增加到 2018 年的 38589 个，而且依然不断有新的 miRNA 被人们发现。miRNA 行使功能主要依靠转录抑制和 mRNA 的切割、降解抑制下游基因表达，通过前面三种方法减弱或者消除下游基因的功能，从而实现对疾病等状态的调节。具体体现在 miRNA 在肿瘤、癌症、神经系统疾病、免疫系统疾病以及心血管系统疾病的诊断、治疗以及预后方面都有着十分重要的作用。比如 miRNA 和癌症的关联，如果 miRNA 与抑制癌症

的基因转录 mRNA 特异性结合，会抑制抑癌基因的表达，导致癌症的发生，反之，如果 miRNA 与癌症基因转录 mRNA 特异性结合，会抑制癌症基因的表达，从而可以抑制癌症的发生[4]。因此，在疾病的诊断时，我们可以通过观察 miRNA 表达谱的变化从而实现通过 miRNA 诊断疾病的发生。同样，当疾病发生时，人们可以对异常含量的 miRNA 进行人工干预，起到对疾病的治疗作用，此外，利用 miRNA 也可以进行药物的疗效判断和疾病的预后判断。因此，我们可以看出 miRNA 在疾病的预测、治疗等方面都有十分重要的作用，然而 miRNA 数量多、形式复杂，尽管人们已经意识到预测 miRNA 和疾病的关联十分重要，但是我们仍然很难建立 miRNA 和疾病关联的完整结构。主要原因是传统的生物实验方法，如合酶链式反应[5]和微阵列技术[6]等生物实验周期较长，且需要大量的人力、物力支持，生物实验结果也存在假阳性问题，通过这种方法，我们很难构建完整的 miRNA 和疾病关联网络，而计算机具有计算速度快，逻辑能力强以及存储容量大等特点，因此，我们可以通过计算机的方法进行仿真模拟，利用现有数据计算出 miRNA 和疾病的关联，这样可减轻生物实验人员的工作量，也可以尽早将 miRNA 进行疾病判断，开发 miRNA 靶向药物用于临床治疗。

本文提出了一个基于监督学习的机器学习方法的模型用于预测 miRNA 和疾病的关系，采用基因作为介导，这是因为 miRNA 主要通过抑制靶向基因的表达从而减弱或者消除靶向基因的功能，从而抑制或者引起疾病的发生。该方法可以在较短时间内预测大量 miRNA 和疾病的关联，运用这种方法，可以告诉生物实验人员该 miRNA 和该疾病是否关联，从而大大减少生物实验人员的工作量。

1.2 国内外研究现状及发展趋势

近些年，科学家们发现 miRNA 的异常表达和许多疾病的发生相关联，因此，科学家们努力通过生物实验找到 miRNA 与疾病的关联关系，目前生物实验人员通过各种生物实验已经发现了上万个 miRNA 和疾病的关联关系对，其中，记载最全面且使用最广泛的是 HMDD 数据库，目前已有 3.0 版本。HMDDv3.0 数据库已从 17412 篇论文中收集了 32281 个实验支持的 miRNA-疾病关联条目，涵盖了 1102 个 miRNA 基因和 850 种疾病。然而，仅通过生物实验去发现 miRNA-

疾病的关联耗费成本高、试验周期长、其结果也存在着假阳性的问题。为了克服以上的的问题，在近十年中，研究人员提出了用计算机模型去解决预测 miRNA-疾病关联的问题，并取得一定成果。

目前，用来预测 miRNA 与疾病关联的计算机方法，主要分为三类，第一种是建立 miRNA-疾病评分模型的方法，即利用 miRNA 相关信息和疾病相关信息的统计或分布特征，通过评分函数得到对疾病-miRNA 对的得分。基于有相似功能的 miRNA 可能与具有相似表现型的疾病有关的假设，Jiang[7]基于 miRNA 相似性网络、疾病表型网络和经过实验验证的 miRNA-疾病对提出了一个计算 miRNA 与疾病关联性得分的超几何分布评分模型，然而由于该模型预测的结果有较高的假正例率，因此该模型的实验结果并不理想。Mork[8]结合了 miRNA 与蛋白质的关联以及蛋白质与疾病的关联提出了一种预测模型(MiRPD)，该模型将蛋白质作为介导，通过预测 miRNA 与蛋白质的得分以及疾病与蛋白质的得分从而得到 miRNA 和疾病的关系，这一方法依赖于 miRNA 与蛋白质关系以及疾病与蛋白质关系的准确性。基于假设表现型相似的疾病拥有相似的分子作用机制，Xu[9]通过提取基因层面的特征作为介导得到了一种预测 miRNA 与疾病关系的模型，这种方法利用了实验支持的疾病与基因之间的关系和推测出的 miRNA 与基因之间的关系，因此实验结果并不准确。此外，陈兴[10]等人基于 miRNA 和疾病的功能相似性提出名为 WBSMDA 的模型，这个模型利用实验验证的 miRNA-疾病关联数据与未经实验验证的数据进行整合，通过计算两个得分的平均数最后 miRNA 与疾病对之间的得分。第二种预测 miRNA 和疾病关联的方法是基于复杂网络的算法模型，陈兴[11]等人提出了名为 RWRMDA 的算法模型，该模型在 miRNA 功能相似性网络上利用已知 miRNA-疾病关联采取可重启的随机游走算法预测 miRNA 与疾病关联，但是该方法不能预测未出现过的疾病与 miRNA 关联。Xuan[12]提出了一个随机游走模型 MIPD 预测 miRNA 与疾病关联，该模型聚集了 miRNA 功能相似性网络、疾病语义相似性网络以及 miRNA-疾病的拓扑网络，通过整合标记节点和未标记节点的拓扑结构预测疾病与 miRNA 的潜在关联。此外，陈兴[13]等人基于异构网络提出了名为 HGIMDA 的模型，该模型首先构建了一个异构网络，接着在该网络上对优化函数进行迭代更新从而预测 miRNA 和疾病的潜在关

联，该模型比他此前提出的 WBSMAD 模型更加高效。Zeng[14]提出了一个基于结构扰动的模型 SPM，该模型将 miRNA 相似性网络、疾病相似性网络与 miRNA 与疾病关联这三个网络聚集成了一个双层网络结构，通过该网络的结构一致性预测链路的关联性，从而得出 miRNA 与疾病的关联。第三种用来预测 miRNA 和疾病关联的方法是基于机器学习的方法，同样基于假设具有类似功能的 miRNA 倾向于与具有相似表型的疾病相关[15]，Xuan[16]等人提出一种基于 K 最近邻算法的计算模型，该模型包含了疾病的语义相似性信息和疾病表型相似性信息，通过这两种信息对 miRNA 功能相似信息进行计算，HMPD 通过聚簇相似的 miRNA 信息从而提高预测的准确性，但这这个模型无法处理未知的疾病。因为针对该预测难以找到反例，陈兴等人[17]提出了基于规则化最小二乘法的学习模 RLSMDA 预测 miRNA 与疾病的关联，该模型可以在不使用反例的情况下预测潜在疾病与 miRNA 的关联。Pasquier 等人[18]利用基于空间向量的奇异值分解方法提出了名为 MiRNA 的模型，该模型利用高维空间向量表示 miRNA 与疾病的分布特征并利用向量相似性去度量 miRNA 与疾病的关系。此外，陈兴等人[19]建立了名为 RKNNDA 的模型，该模型结合了 K 最近邻算法以及支持向量机分类算法。Chen 等人[20]还提出了基于拉普拉斯正则化稀疏子空间学习的方法，首先，将 miRNA 和疾病的理论和统计特征投射到公共子空间中，接着用拉普拉斯正则化维持训练样本的局部结构，最后用 L1 范数约束选择对预测贡献较大的特征。

目前，已经有很多国内外研究学者提出预测 miRNA 和疾病关联的方法，也取得了不错的成果，但同时目前的研究也存在着一些问题，就是这些方法大多是基于网络而不是使用标记信息的非监督方法，主要原因是没有足够的标记数据来训练监督模型以及无法直接取得反例，因此，如何基于目前数据来快速并准确的预测 miRNA 和疾病的关联是我们需要解决的问题。

1.3 本文的研究内容

到目前为止，预测 miRNA 与疾病关联的机器学习方法已经有很多，但是没有足够的标记数据来进行模型训练。近年来，越来越多的人关注到这一问题，目前已经一些数据库可以给我们提供所需要的数据。基于此，本文提出了一个监督

模型预测 miRNA 和疾病的关联，与以前大部分人直接使用网络信息去计算关联分数不同，本文通过网络提取基因特征作为特征预测 miRNA 和疾病的关联，因为 miRNA 可以通过控制疾病的靶基因的表达从而引发疾病，比如 miRNA-21 通过调控 PTEN 基因的表达控制肝癌的发生[21]，因此我们将基因信息作为特征向量预测 miRNA 与疾病的关联，此外，本文不仅考虑了与 miRNA 和疾病直接关联的基因，还考虑了与 miRNA 和疾病存在潜在关联的基因，具体方法如下：

(1) 通过多个数据库包括疾病与基因关联的数据库 DisGeNET，miRNA 与疾病关联的数据库 miRWalk2.0，基因与基因关联的数据库 HPRD 等中的数据得到疾病相似性网络、基因-基因关联网络、miRNA 相似性网络、miRNA-基因关联网络以及疾病-基因关联网络，通过以上五个网络可以得到 miRNA-基因-疾病三层网络。

(2) 从 miRNA-基因-疾病三层网络中提取 miRNA-基因、疾病-基因特征，总体思路是基于一种创新的回归模型求得不同 miRNA 之间的相似性得分和不同疾病之间的相似性得分，接着通过皮尔逊相关系数求得 miRNA（或疾病）与不同基因之间的相似性得分。

(3) 使用堆栈自编码器对得到的特征向量进行降维和去噪处理。

(4) 使用卷积神经网络对特征向量进行分类处理，采用十折交叉验证法得到最终分类处理的结果，采用多种评估标准进行评估并与其它模型的实验结果进行比较。

1.4 本文组织结构

本文总体分为五个部分，每一部分的具体内容安排如下：

第一章是绪论部分，主要介绍了本文研究的背景及意义，介绍了 miRNA 的相关信息以及 miRNA 与疾病发生的关系，接着介绍了国内外研究现状，目前预测 miRNA 和疾病主要有三大方法，并分析了目前研究方法的不足，最后介绍了本文的研究方法和创新点。

第二章是课题研究的理论基础部分，这一部分首先介绍了 miRNA 参与转录的过程，接着介绍了 HMDD 数据库，该数据库可以得到经实验验证的 miRNA 和

疾病关联数据，接着分别介绍了疾病-基因关联网络、miRNA-基因关联网络、基因网络、miRNA 相似性网络和疾病相似性网络，由这五个网络我们可以构建 miRNA-基因-疾病三层网络。

第三章是基于卷积神经网络方法的 miRNA-疾病关联预测部分，这一部分介绍了一种创新的回归模型求不同 miRNA 之间的相似性得分和不同疾病之间的相似性得分，接着通过皮尔逊相关系数求得 miRNA（或疾病）与不同基因之间的相似性得分，最后使用 softmax 归一化处理。将处理过后的数据用堆栈自编码器进行降维处理最后通过卷积神经网络进行分类。

第四章是实验结果分析部分，这一部分主要采用十折交叉验证法得到模型预测的准确率、精确率、召回率、F1 得分、ROC 曲线和 PR 曲线的 AUC 值作为评估指标，并与支持向量机作为分类器的实验结果进行对比，最后再将本文实验结果与其他论文实验结果进行对比。

第五章是本文的结论部分，该部分主要总结了本文的模型的实现方法与该方法对这预测 miRNA 与疾病关联这一问题的贡献，并总结了该模型可能存在的问题与未来的研究方向。

第二章 课题研究理论基础

目前，miRNA 与疾病的关联机制仍然不清晰，传统的生物实验方法耗费大量的时间和人力、物力。为了加快我们对疾病与 miRNA 关联机制的了解，生物实验人员通过生物实验等方法构建了很多数据库，这些数据库为我们通过计算机模型来预测 miRNA 和疾病的关联提供了很好的理论依据，本文所用到的数据库就是科学家通过生物实验得到的，包括 miRNA 与疾病关联的数据库 HMDD[22]，疾病与基因关联的数据库 DisGeNET[23]，miRNA 与疾病关联的数据库 miRWalk2.0, 基因与基因关联的数据库 HPRD[24]，下面会对这些数据库进行介绍。

2.1 miRNA 介绍

MiRNA 是一种存在于真核细胞中的物质，miRNAs 是一类单链小 RNA，它们的长度约 18-25nt [25]。miRNA 在细胞的各项生命活动如增殖、分化和凋亡的过程中起到了十分重要的调节作用。如下图所示是 DNA 到表型的具体生物过程，miRNA 参与转录后调控这一过程，miRNA 通过与基因转录 mRNA 特异性结合，从而调控蛋白质的表达，进而影响功能和表型。

图 2.1 从 DNA 到表型的生物过程

2.2 MiRNA-疾病的关联关系

经过长期的研究，研究人员已经建立了一些 miRNA 和疾病关联的数据库，本文使用的 HMDD v2.0[26](Human MicroRNA Disease Database)数据库，该数据库中所有 miRNA-疾病的关联关系都是经过生物实验验证存在的。目前，HMDD v2.0 已超过 3000 篇论文中收集了多于 10000 条实验支持的 miRNA-疾病关联条目，涵盖了超过 600 个 miRNA 基因和超过 400 种疾病。为了构建最终的训练集和测试集数据，本文从 HMDD 数据库中通过构建邻接矩阵 Y 来描述 miRNA 与疾病之间的关联，具体方法是如果 miRNA m(i)已知与疾病 d(j)之间存在关联，那

么 Y(i, j) = 1, 否则 Y(i, j) = 0，我们最终使用了 HMDD 数据库中 1901 条已知的 miRNA 和疾病关数据，并从 Y=0 的位置中随机产生了等量的数据作为反例，最终包含了 243 种 miRNA 和 201 种疾病。

2.3 疾病-基因关联网络的构建

本文使用 DisGeNET[27]数据库中的信息来构建疾病-基因关联网络，DisGeNET 是一个专门收录人类疾病相关基因信息的数据库，该数据库整合了已有的基因与疾病关联的数据库并通过基于 NLP 的机器学习方法从这些数据库中获得对应关联信息，最终构建出一个统一的基因与疾病关联的数据库。从该数据库中我们可以得到疾病的产生与哪些基因有关，同时该数据库也提供了相应标准衡量某一疾病与某一基因的关联程度。

2.4 基因-基因关联网络的构建

本文使用 HPRD[28](Human Protein Reference Database)数据库得到基因-基因关联网络，HPRD 是一个专门存储人类蛋白质相互作用信息的数据库，和其他同类数据库相比，该数据库的蛋白质信息数量上有明显优势，此外冗余信息较少。这里我们通过蛋白质相互作用网络(PPI)可以直接得到基因-基因关联网络，该数据可以在 HPRD 网站上直接下载，共包含 9386 个基因及 36504 种相互作用关系，在本文计算的过程中，我们将删除与疾病或者 miRNA 无关的基因。

2.5 MiRNA-基因关联网络的构建

本文使用 miRWalk2.0[29]数据库中的信息来构建 miRNA-基因关联网络，miRWalk2.0 数据库提供了预测的以及实验验证的 miRNA 靶标相互作用集合，是目前这一领域最大的数据库，miRWalk2.0 中记录了 11,748 个 miRNA，308,700 个基因和 68,460 个 lncRNA 之间的大约 9.49 亿次相互作用。从该数据库中我们的得到某一 miRNA 和哪些基因相关联，同时该数据库也提供了相应标准衡量某一 miRNA 与某一基因的关联程度。本文使用该数据库中经过实验验证的 miRNA 和基因的关系，且使用的基因与 miRNA 和疾病均有关联。

2.6 MiRNA 相似性网络的构建

2.6.1 MiRNA 功能相似性网络构建

已知具有相似功能的基因通常与相似的疾病相关，研究表明，miRNA 也有类似的性质。Wang[30]提出了一个构建 MiRNA 功能相似性网络的方法，就是通过度量两个 miRNA 格各自关联的疾病相似性从而得到两个 miRNA 之间的相似性，具体方法如下：

步骤一：计算某一疾病的语义贡献值。首先，将某个疾病与疾病相关语义用一个有向无环图(DAG 图)表示，图 2.2 为乳腺癌的 DAG 图。

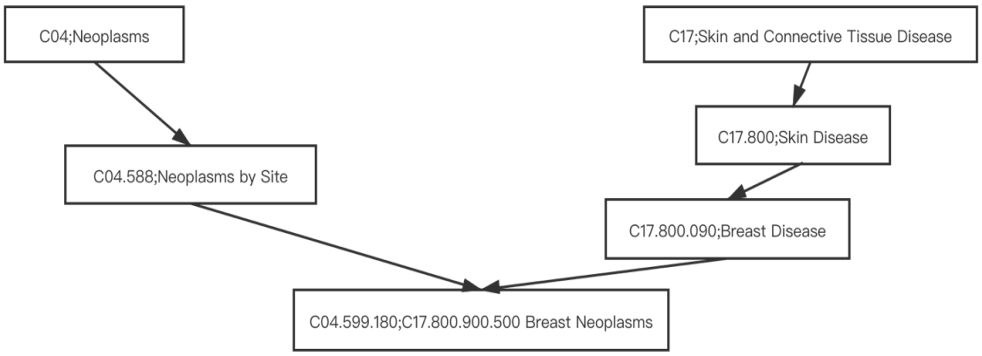

图 2.2 乳腺癌 DAG 图

例如：疾病 A 的 DAG 图用$\mathrm{DAG}_A = \{A, T_A, E_A\}$表示，其中$T_A$表示与疾病 A 有关的语义节点的集合，包括 A 本身。E_A表示这些节点之间所有边的集合。由于距离最下层节点疾病 A 距离越远，该节点对疾病 A 的贡献越小，用$D_A(t)$表示疾病t对在DAG_A中对疾病 A 的语义贡献值，公式如下：

$$\begin{cases} D_A(A) = 1 \\ D_A(t) = \max\left\{\Delta * D_A\left(t'\right) \mid t' \in \text{ children of } t\right\} & \text{if } t \neq A \end{cases} \tag{2-1}$$

其中，Δ为语义贡献因子，当 t =A 时，可以看出疾病 A 对自身的语义贡献值为 1，我们可以得到，Δ的值在 0 到 1 之间，在经过大量实验，发现$\Delta = 0.5$的效果最好，因此在这里使用$\Delta = 0.5$。该公式可以理解为在 DAG 图中，某一种疾病对于疾病 A 的语义贡献值为0.5^{l-1}，l为该疾病所在 DAG 图的层数。在基于公式(2-1)的基础上，用$DV(A)$表示疾病 A 的语义值：

$$\mathrm{DV}(A)=\sum_{t\in T_A} D_A(t) \tag{2-2}$$

步骤二：计算两种疾病之间的语义相似性。已知，两种疾病的 DAG 图越相似，两种疾病的语义相似性越高。因此，可以得下面的公式计算两种疾病的语义相似性：

$$SS(A,B)=\frac{\sum_{t\in T_A\cap T_B}\left(D_A(t)+D_B(t)\right)}{\mathrm{DV}(A)+\mathrm{DV}(B)} \tag{2-3}$$

其中，$\mathrm{DV}(A)$为疾病 A 的语义贡献值，$\mathrm{DV}(B)$为疾病 B 的语义贡献值，两种疾病的语义相似性由共同作用于疾病 A 和 B 的疾病数量决定的。

步骤三：求得 miRNA 功能相似性。假设我们求 hsa-mir-103 与 hsa-mir-151 两种 miRNA 之间的功能相似性，其中DT_1表示与 hsa-mir-103 相关的疾病集合，DT_2表示与 hsa-mir-151 相关的疾病集合。DT_1包含 m 种疾病，DT_2包含 n 种疾病，可以用如下公式计算 hsa-mir-103 与 hsa-mir-151 两种 miRNA 之间的功能相似性, FS的值越大，表示 M1 和 M2 的功能相似性越大：

$$FS(\mathrm{M1},\mathrm{M2})=\frac{\sum_{1\le i\le m} S\left(dt_{1i},DT_2\right)+\sum_{1\le j\le n} S\left(dt_{2j},DT_1\right)}{m+n} \tag{2-4}$$

其中，$S(dt,DT)$表示某个疾病 dt 与疾病的集合$DT=\{dt_1,dt_2,...,dt_k\}$之间相似性的最大值，用如下公式计算：

$$\mathrm{S}(dt,DT)=\max_{1\le i\le k}(SS(dt,dt_i)) \tag{2-5}$$

2.6.2 miRNA 功能相似性网络稀疏性问题

尽管可以用上述方法得到 miRNA 功能相似性网络，但是该网络存在稀疏性问题，You[31]提出了一种解决网络稀疏性问题的方法，就是用高斯相互作用属性核相似性来计算两个 miRNA 之间的核相似性。已知两个有相似功能的 miRNA 与相似的疾病关联，因此可以用已知的 miRNA-疾病关联网络得到两个 miRNA 之间的相似度。在 miRNA-疾病关联网络中，如果 miRNA m(i)已知与疾病 d(j)之间存在关联，那么$Y(i,j)=1$，否则$Y(i,j)=0$。对于一个给定的 miRNA m(i),它的

$IP(m(i))$被定义为邻接矩阵 Y 的第 i 行，接着可以按照如下公式计算 miRNA m(i)和 miRNA m(j)之间的高斯相互作用属性核相似性：

$$KM(m(i),m(j))=\exp\left(-\gamma_m\| IP(m(i))-IP(m(j))\|^2\right) \tag{2-6}$$

$$\gamma_m=\gamma_m^{'}/\left(\frac{1}{nm}\sum_{i=1}^{nm}\| IP(m(i))\|^2\right) \tag{2-7}$$

其中，γ_m是控制内核带宽的参数，本文我们取 $\gamma_m=1$。

我们可以用如下公式的方式补全 miRNA 功能相似性网络，这样就解决了 miRNA 功能相似性网络的稀疏性问题。

$$S_m(m(i),m(j))=\begin{cases}FS(m(i),m(j)) & m(i)\text{和}\mathrm{m}(j)\text{ 之间有功能相似性}\\ KM(m(i),m(j)) & \text{其它}\end{cases} \tag{2-8}$$

2.7 疾病相似性网络的构建

在求得 miRNA 功能相似性网络的过程中，我们已经得到两种疾病之间的语义相似性$SS(d(i),d(j))$，在这里可以直接使用由公式(2-3)求得的疾病语义相似性网络。但是，该网络依然存在稀疏性问题，因此，本文按照解决 miRNA 相似性网络稀疏性问题的方法解决疾病功能性网络稀疏性问题。其中，$IP(d(i))$表示 miRNA-疾病关联网络 Y 的第 i 列，γ_m是控制内核带宽的参数，本文我们取$\gamma_m=1$。

$$KD(d(i),d(j))=\exp\left(-\gamma_d\| IP(d(i))-IP(d(j))\|^2\right) \tag{2-9}$$

接着，我们可以用公式(2-10)计算出疾病之间的相似性。

$$S_d(d(i),d(j))=\begin{cases}SS(d(i),d(j)) & d(i)\text{ 和 }d(j)\text{ 之间具有语义相似性}\\ KD(d(i),d(j)) & \text{其它}\end{cases} \tag{2-10}$$

2.8 模型的评估指标

2.8.1 交叉验证法

本文使用 k 折交叉验证法作为模型的训练方法，具体来说就是将训练集随机等分为 k 份并随机取其中一份为验证集评估模型，其余 k-1 份为训练集训练模

型。我们将实验重复做 k 次，在每一次实验中，我们都随机的取一份做为测试数据，剩下的 k-1 份数据作为训练集，同时我们要保证在经历了 k 次实验后，每份的数据都做过测试集。最后将 k 次的实验结果进行平分就可以得到最终结果。

在本文中，我们将已知的 miRNA-疾病关联作为正例，并将它们随机的划分为 10 等份，因为我们无法从目前的数据库中得到反例，因此我们将未知的 miRNA-疾病关联作为反例，从中随机产生与正例数目相同的反例。

2.8.2 模型效果的评估指标

本文用准确率，精确率，F1 得分，召回率、ROC 曲线以及 PR 曲线衡量模型的性能，在这些衡量指标的公式中，分类结果的混淆矩阵如表 2.1 所示：

表 2.1 分类结果混淆矩阵

真实情况	**预测结果**	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

准确率(Accuracy)：准确率是用来衡量所有样本被分类正确的比例，计算公式如下：

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2-11}$$

精确率(Precision)：也叫查准率，衡量正样本的分类准确率，具体就是描述被预测为正样本的样本中有多少是真的正样本，一个分类器的精度越高，那么它的假正类错误率越低，计算公式如下：

$$precision = \frac{TP}{TP + FP} \tag{2-12}$$

召回率(Recall)：表示分类正确的正样本占所有正样本的比例，当一个分类器的召回率很高时，该分类器的特性是很少会将正样本分为负样本，计算公式如下：

$$recall = \frac{TP}{TP + FN} \tag{2-13}$$

F1 得分：F1 得分表示精确率和召回率的调和平均，当一个模型的 F1 得分较高时，表示模型的精度以及召回率都比较高，计算公式如下：

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + TN} \tag{2-14}$$

ROC 曲线：ROC 曲线的横坐标为假正率(FPR)，假正率表示本身是负样本但是被分为了正类，纵坐标为真阳性率(TPR)，真阳性率表示本身是正样本但是被分为了负类，ROC 曲线可以在图上表示，ROC 曲线越接近左上角，该分类器的性能越好，计算公式如下：

$$TPR = \frac{TP}{TN + FP}, FPR = \frac{FP}{TN + FP} \tag{2-15}$$

可以通过 ROC 曲线计算得到 AUC（Area Under Curve）的值，AUC 表示由 ROC 曲线和坐标轴包围的区域，并且该区域的值小于或等于 1。AUC 越接近 0.5，就越意味着该模型基本上没有正确分类的能力。ROC 曲线无法直观地表明模型的分类是否有效。其相对应的 AUC 作为一个值，可以直观的反应模型的分类效果，相应的 AUC 越大，模型的分类效果越好。

P-R 曲线：PR 曲线是以精准率和召回率这两个为变量而做出的曲线，其中召回率为横坐标，精确率为纵坐标。当数据不平衡时，PR 曲线是敏感的，PR 会随着正负样本比例的变化而变化。相比之下 ROC 曲线对数据是否平衡是不敏感的，无论数据平衡与否其曲线能够基本保持不变。

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \tag{2-16}$$

根据 PR 曲线下方的面积大小 AUC 可以来评估模型的好坏，AUC 越大，表示模型的性能越好。也可以平衡点或者 F1 值来评估模型，平衡点(BEP)是 P=R 时的取值，如果这个值较大，则说明该的性能较好，同样，F1 值越大，我们可以认为该模型的性能较好。

2.9 本章小结

网络构建方面：本章主要介绍了如何构建 miRNA-基因-疾病三层网络，通过 DisGeNET[32] 数据库中的信息构建疾病-基因网络、通过 miRWalk2.0[33]数据库构建 miRNA-基因网络、通过 HPRD[34]数据库构建基因-基因关联网络，并通过 Wang[35]的论文中的方法构建了 miRNA 相似性网络和疾病相似性网络，主要基于具有相似功能的 miRNA 可能与具有相似表现型的疾病有关，以及 miRNA 通过与基因转录 mRNA 特异性结合，从而影响功能和表型这两个大原理。并详细介绍了论文中网络构建方法的具体实现细节，通过以上五个网络可以得到 miRNA-基因-疾病三层网络。

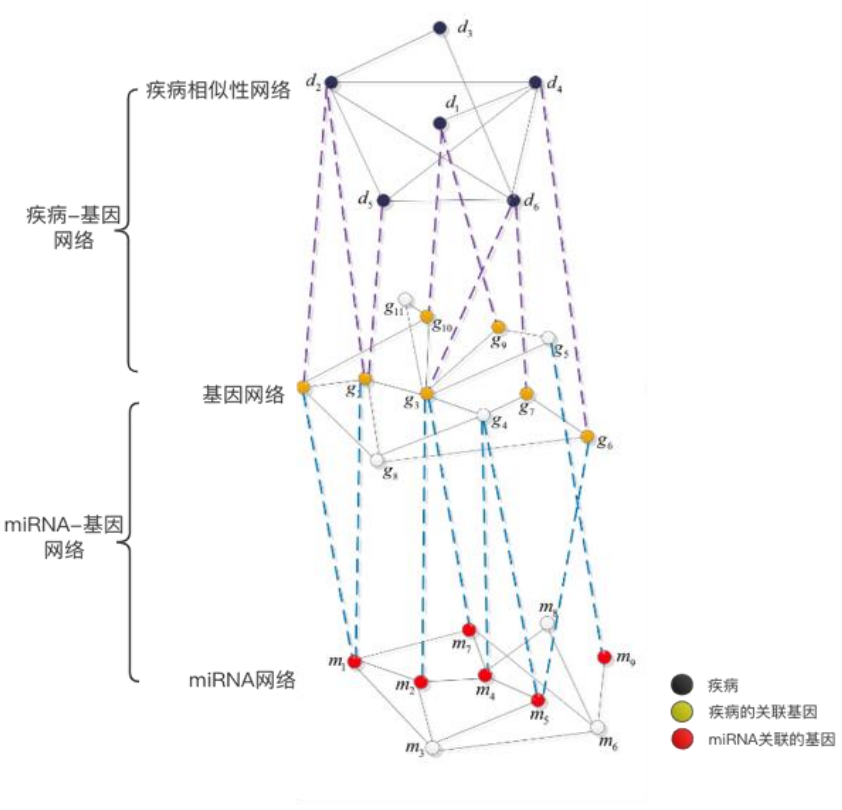

图 2.3 三层网络示意图

数据集创建方面：本章介绍了 HMDD 数据库，同时介绍了如何通过该数据库得到实验所需的训练集。

模型评估指标介绍方面：介绍了本文模型评估的几个指标的定义及其计算方法，包括准确率，精确率，F1 得分，召回率、ROC 曲线以及 PR 曲线。

第三章 基于卷积神经网络方法的 miRNA-疾病关联预测

3.1 引言

MiRNA 在真核细胞中普遍存在，长约 22nt 的单链小 RNA，近年来，人们逐渐认识到 miRNA 在细胞的生命活动如增殖、分化和凋亡的过程中起到了十分重要的作用，因此越来越多的人们投入到了对 miRNA 的研究当中，但是如何快速大量的获取 miRNA 与疾病的关联对于传统的生物实验方法是基本不可能实现的，因此越来越多的人使用生物信息学的方法预测 miRNA 和疾病的关联，目前关于 miRNA-疾病的研究，目前主要分为基于相似性度度量的方法和基于机器学习的方法，基于机器学习的方法又有监督学习和无监督学习之分，本文提出了一种基于监督学习的机器学习方法的 miRNA-疾病关联预测，即通过卷积神经网络预测 miRNA 与疾病的关联。该模型主体由三个部分组成，分别是特征提取部分、基于对讲自编码器的降维部分以及通过卷积神经网络分类的部分，该模型整体如图 3.1 所示：

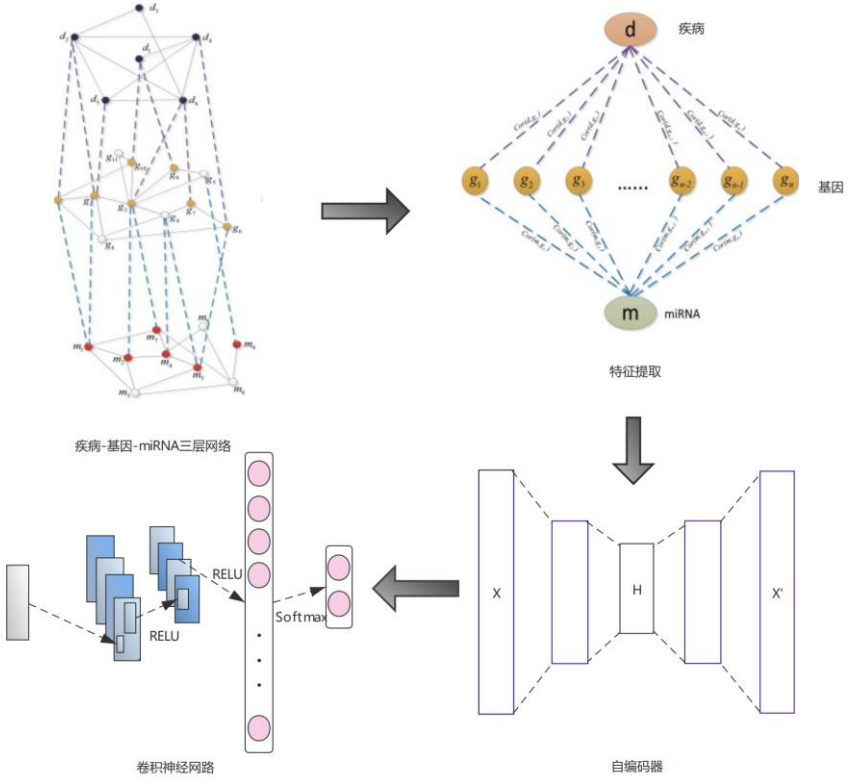

图 3.1 模型总体实现步骤

3.2 基于疾病-基因-miRNA 三层网络的特征提取

3.2.1 关联得分的计算

由 2.1-2.5 我们可以得到疾病相似性网络、基因网络、miRNA 相似性网络、疾病-基因网络以 miRNA-基因网络，如下图所示：

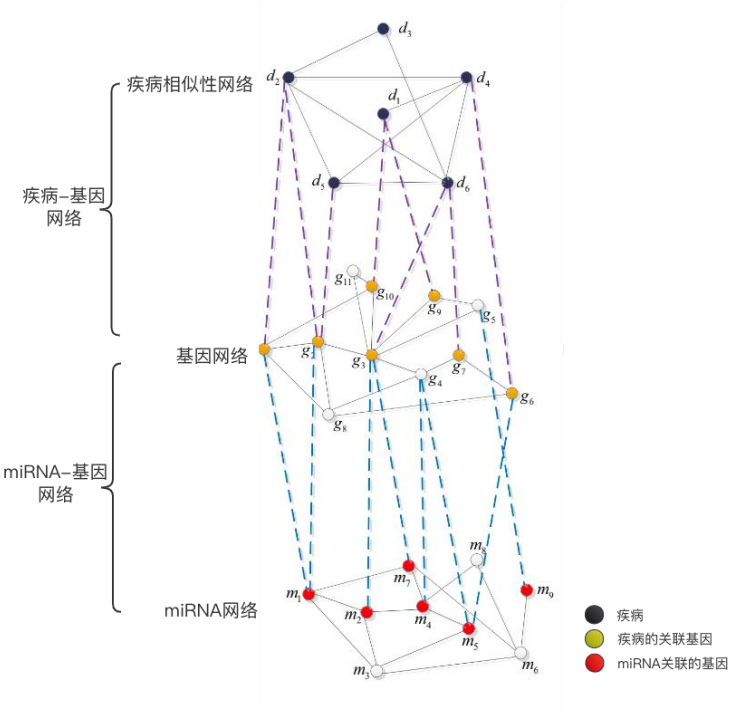

图 3.2 疾病-基因-miRNA 三层网络

但这一网络无法很好的度量 miRNA 和基因的关系，因此本文为一个 miRNA 和一个基因求出该 miRNA-基因对的关联性得分，并对所有的 miRNA-基因对分别求的其关联性得分，这样我们可以从三层网络中提取 miRNA-基因特征。同时，对于已经得到的疾病-基因网络，我们无法很好的度量疾病与基因的关系，本文按照提取 miRNA-基因特征的方法可以得到疾病-基因的特征。具体的特征提取方法参考了李稍在 2008 年论文[36]中提出的计算相似性得分的方法，具体计算方法分为以下几个步骤，对于 miRNA-基因特征(疾病-基因特征计算方法相同)：

步骤一：

G_m为 miRNA 网络，G_g为基因相似性网络，G_{m-g}表示 miRNA 与基因关联网络。无向图$G_m = (m, E_m)$,其中$m = \{m_1, m_2, ..., m_n\}$表示$G_m$中 miRNA 的集合，$E_m \in M * M$表示上文求得的 miRNA 对之间的相似性得分。无向图$G_g = (g, E_g)$,其中

$g=\{g_1,g_2,...,g_g\}$表示G_g中的基因的集合，$E_g \in G*G$为上文求得的基因之间的相似性得分。无向二分图$G_{m-g}=(m,g,E_{m-g})$，其中 m 和 g 分别代表 miRNA 集合和基因集合，$E_{m-g} \in M*G$表示 miRNA-基因对距离的集合。

首先，求出 miRNA-基因对的相似性得分，计算公式如下：

$$\mathrm{R(g,m)}=\sum_{g'\in G(m)} e^{-dis(g,g')} \tag{3-1}$$

Z 其中，$g' \in G(\mathrm{m})$表示对于任意 miRNA,与该 miRNA m 关联的所基因的集合；$-dis(g,g')$表示基因g与基因g'之间的相似度得分，具体计算方法为二者之间距离的平方。

步骤二：

求出 miRNA m_i与m_j之间的相似性得分。尽管在上文我们已经求出 miRNA 对之间的相似性得分，但是该相似性得分中没有考虑基因信息，因此我们用下面的公式重新计算 miRNA 对之间的相似性得分，该计算方法中加入了对基因相关的信息的考虑。

$$\mathrm{Sim}\left(m_i,m_j\right)=C_{m_i}+\sum_{g_i\in G(\mathrm{m}_i)}\beta_{\mathrm{m}_i g_i}R\left(g_i,\mathrm{m}_j\right) \tag{3-2}$$

其中，$\mathrm{Sim}(m_i,m_j)$表示 miRNA m_i,m_j之间的相似性得分，其中$\mathrm{g'} \in G(m)$表示对于任意 miRNA m，与该 miRNA 关联的所基因的集合，$R(g_i,m_j)$为步骤一求出的 miRNA-基因相似性得分，对于任意 miRNA，C_{m_i}是一个常数，表示某一种 miRNA 与其他 miRNA 之间的基本相似度，$\beta_{m_i g_i}$表示了基因g_i对 miRNA 相似度的贡献程度。具体计算方法是采用多元线性回归模型计算，公式如下，对于 miRNA m_i而言：

$$y=C_{m_i}+\beta_{m_1g_1}x_1\ldots+\beta_{m_ig_i}x_i+\xi \tag{3-3}$$

用矩阵表示为：

$$y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad x=\begin{bmatrix}1 & x_{m_1g_11} & \cdots & x_{m_ig_i1}\\1 & x_{m_1g_12} & \cdots & x_{m_ig_i2}\\\vdots & \vdots & & \vdots\\1 & x_{m_1g_1n} & \cdots & x_{m_ig_in}\end{bmatrix},\quad \beta=\begin{bmatrix}C_{m_i}\\\beta_{m_1g_1}\\\vdots\\\beta_{m_ig_i}\end{bmatrix},\quad \xi=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix} \tag{3-4}$$

其中，Y 为 miRNA m_i与其余 miRNA 之间的相似性得分，输入 x 为E_{m-g}中 miRNA-基因对的距离，g 表示与m_i相关的所有基因。这么做的目的是我们已知 m_i与 m_j之间的关系，但这一关系中不包含基因特征，因此用已知的 m_i与 m_j之间的关系通过多元线性回归可以拟合得到 miRNA m_i与基因的关系，再还原出 m_i与 m_j之间的关系，此时该得分就可以在基因层面反映 m_i与 m_j的关系。

步骤三：

求出基因与 miRNA 对的皮尔逊相关系数，公式如下：

$$\mathrm{Cor}(m,g)=\frac{\mathrm{cov}\left(S_m,R_g\right)}{\delta\left(S_m\right)\delta\left(R_g\right)} \tag{3-5}$$

其中$S_m=[Sim(m,m_1),Sim(m,m_2),\ldots,Sim(m,m_n)]$是 miRNA m 和其余 miRNA G_m的相似性得分组成的特征向量，$R_g=[R(g,m_1),R(g,m_2),\ldots,R(g,m_n)]$是基因 g 和 miRNA G_m的距离得分组成的特征向量，其中 cov 表示 s 和 r 的协方差，δ 表示标准差。

3.2.2 构建 miRNA-基因、疾病-基因特征

通过上面的步骤可以求得疾病-基因以及 miRNA-基因的关联性得分，通过以上得分我们构建出构建 miRNA-基因、疾病-基因特征，对于任意一个疾病，我们可以表示用向量 X_d表示：

$$X_d=\left[\mathrm{Cor}\left(d,g_1\right),\mathrm{Cor}\left(d,g_2\right),\ldots,\mathrm{Cor}\left(d,g_n\right)\right] \tag{3-6}$$

同理，对于任意一个 miRNA，我们可以用向量 X_m表示：

$$X_m=\left[\mathrm{Cor}\left(m,g_1\right),\mathrm{Cor}\left(m,g_2\right),\ldots,\mathrm{Cor}\left(m,g_n\right)\right] \tag{3-7}$$

为了减少一部分极值（如离群点）的影响，最后我们对数据进行了归一化处理，这里使用了 softmax 归一化的方法，softmax 的形式为：

$$P(y=i)=\frac{\exp\left(\sum_d w_{id}x_d\right)}{\sum_j \exp\left(\sum_d w_{jd}x_d\right)} \tag{3-8}$$

最终我们构建的基因-miRNA 以及疾病-基因特征计算公式如下：

$$X_m^{'}=\left[\frac{\exp\left(\mathrm{Cor}\left(m,g_1\right)\right)}{\sum_{i=1}^{n}\exp\left(\mathrm{Cor}\left(m,g_i\right)\right)},\ldots,\frac{\exp\left(\mathrm{Cor}\left(m,g_n\right)\right)}{\sum_{i=1}^{n}\exp\left(\mathrm{Cor}\left(m,g_i\right)\right)}\right] \tag{3-10}$$

$$X_d^{'}=\left[\frac{\exp\left(\mathrm{Cor}\left(d,g_1\right)\right)}{\sum_{i=1}^{n}\exp\left(\mathrm{Cor}\left(d,g_i\right)\right)},\ldots,\frac{\exp\left(\mathrm{Cor}\left(d,g_n\right)\right)}{\sum_{i=1}^{n}\exp\left(\mathrm{Cor}\left(d,g_i\right)\right)}\right] \tag{3-11}$$

其中，X'_d和X'_m分别表示某一疾病和某一 miRNA 与基因的关系的特征向量表示。

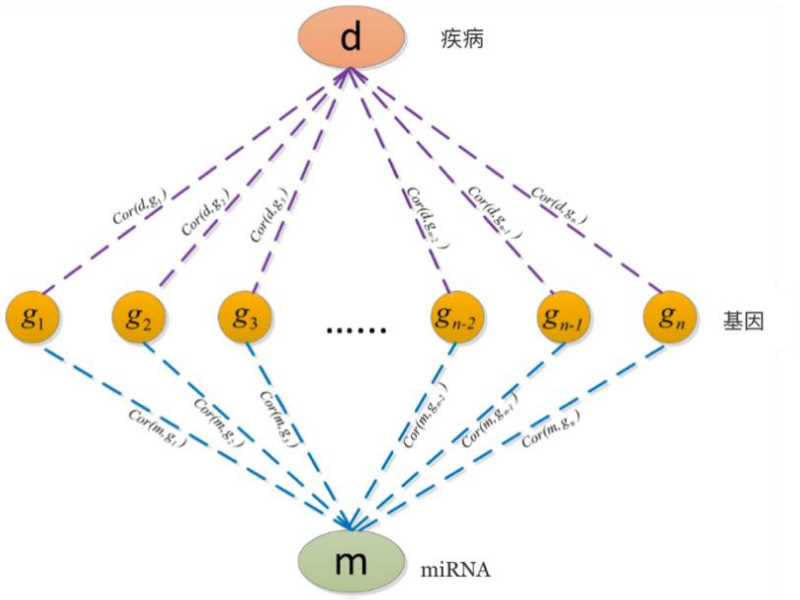

图 3.3 疾病-miRNA 的特征向量表示

3.3 基于自编码器的特征选择

3.3.1 数据降维的方法

数据挖掘的本质是从大量数据中获得有效、新颖、潜在有用、并且最终易于理解的模型的重要过程。很多时候我们得到的高维数据中存在着大量没有用的冗余信息，这些冗余信息一方面会影响计算的效率，另一方面对计算的准确率也会造成影响，数据的本质维数要比目前所得到的数据维数小很多。因此我们要通过降维方法保留数据中包含信息量较大的维度而忽略对数据描述不重要的维度。这

样一方面可以减少数据量从而增加计算效率，同时包含信息量最少的维度的特征向量有可能是数据的噪声信息，因此数据降维可以起到去噪的作用。数据降维的方法目前主要分为线性降维法和非线性降维法两种方法。

线性降维方法指降维后的数据在低维空间依然保持数据之间的线性关系，如局部保留投影(LPP)、线性判别分析(LDA)、主成分分析技术(PCA) 等[37]。但是当数据集在高维空间呈现高度扭曲时，这些方法线性方法难以发现嵌入在数据集中的非线性结构也很难恢复数据的内在结构[38]。

非线性降维方法包括局部线性嵌入(LLE)、核主成分分析(KPCA)、自编码器(Autoencoders)等[39]，本文使用自编码器来进行特征选择，从而起到去除疾病-基因-miRNA 三层网络中冗余信息和噪声的目的。

3.3.2 自编码器的结构及工作原理

自编码器是一种能够通过无监督学习学到输入数据高效表示的人工神经网络[40]。自编码器具有编码和解码两个部主要部分，编码部分(encoder)和解码(decoder)部分。编码部分是通过训练神经网络，将高维数据转化到具有一定维度的低维嵌套上；解码部分也称为重建部分，该部分可以看作编码的逆过程，是自编码器从编码过程得到的低维嵌套中还原与输入高维数据尽可能相似的数据的过程。在编码部分和解码部分之间还存在着一个内部隐藏层，称为“码字层(code)”，用于反应输入数据的本质规律。

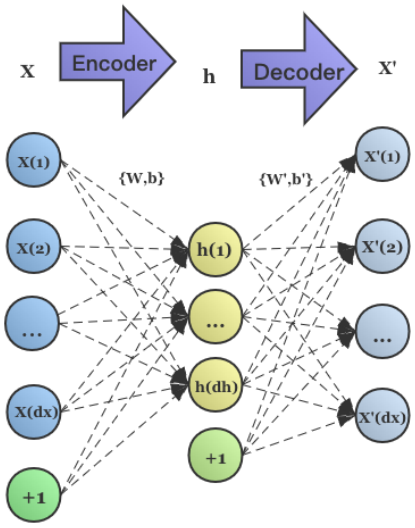

图 3.4 自编码器示意图

基本的自编码器如图所示，我们假设自编码器的输入是$x = [x_1, x_2, \ldots, x_{d_x}]^T \in R^{d_x}$，其中$d_x$是输入的维度，编码器通过映射函数 f 将 x 从输入层投影到隐含层$h = [h_1, h_2, \ldots, h_{d_h}]^T$，其中$d_h$是隐含层变量向量的维度。其中$f(x)$函数表示为：

$$h = f(x) = s_f(Wx + b) \tag{3-12}$$

其中W是$d_h \times d_x$权重矩阵，$b \in R_{d_x}$是偏差向量。编码器的激活函数s_f可以是 sigmoid 函数、tanh 函数或者 ReLu 函数。

在编码器中，通过映射函数 f将隐含层表示的 h 映射到输出层的 $x \in R_{d_x}$，其中 f表示如下：

$$\overline{x} = \overline{f}(h) = s_{\overline{f}}(\overline{W}h + \overline{b}) \tag{3-13}$$

W 是 $d_x \times d_h$权重矩阵，$b \in R_{d_x}$是输出层的偏差向量。同样 s_f的激活函数可以是 sigmoid 函数、tanh 函数或者 ReLu 函数。因此，自编码器的参数集是 $\theta = \{W, \overline{W}, b, \overline{b}\}$。

自编码器通过对网络施加限制来重新构建输出$Y(i, j) = 0$使其尽可能的与输入 x相似。每一个训练样本 x_i都被投影到隐含表示 h_i，然后被映射到重构数据 $\overline{x}_i$。通常通过计算均方重构误差最小化来重构损失函数来获得模型参数：

$$J(W, \overline{W}, b, \overline{b}) = \sum_{i=1}^{N} \left\| \overline{x}_i - x_i \right\|^2 / 2N = \sum_{i=1}^{N} \left\| g_\theta(x_i) - x_i \right\|^2 / 2N \tag{3-14}$$

本文使用堆栈自编码器[41]，相较于基本自编码器，堆栈自编码器具有更强大的表达能力及深度神经网络的优点。堆栈自编码器是一种具有分层结构的神经网络，由多个自编码器层逐层连接组成。堆栈自编码器的模型如下图：

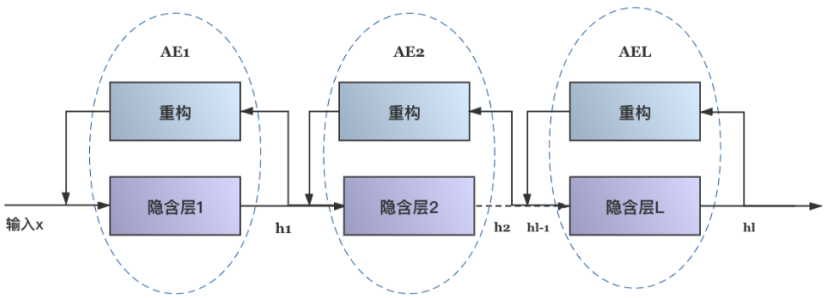

图 3.5 堆栈自编码器示意图

在预训练的步骤中，第一层自编码器通过最小化重构误差将原始输入数据映射到第一个隐含特征层。训练了第一个自编码器之后，第一个隐含层的输出被用作第二个隐含层的输入。然后训练第二层自编码器得到参数 $\{W_2, b_2\}$。通过这种方式，对整个堆栈自编码器进行逐层预训练直到得到最后一个自编码器层。在无监督预训练之后，将输出层加到堆栈自编码器的顶部用来微调权重和偏差。预训练的权重被用作每个隐含层的权重的初始化，可以随机输出层的参数 $\{W_0, b_0\}$。最后通过反向传播对整个网络进行微调，通过最小化目标变量的预测误差来获得改进的权重$\{W_l, b_l\}, l = 1,2,3,\ldots,L$。反向传播函数为:

$$J_O = \sum_{j=1}^{N_l} \left\| y_j - \hat{y}_j \right\|^2 / 2N_l \tag{3-15}$$

3.3.3 基因特征选择实现方法

本文使用了包含两个隐含层的堆栈自编码网络，首先用原始输入训练第一个自编码网络，这里我们的输入数据为维度为 3578 维，如图所示，在输入数据经过第一个自编码网络后，该自编码网络能够学习得到原始输入的一阶特征$h_k^{(1)}$，这里将 k 设置为 2048，表示在原始数据经过第一个自编码网络后，它的维度变为 2048 维：

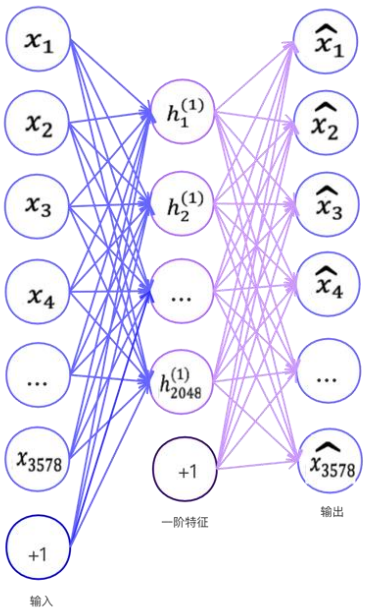

图 3.6 自编码器一阶特征学习过程

接着，再用刚刚训练好的一阶特征 $h_k^{(1)}$作为第二个自编码器的输入，此时的输入数据维度为 2048，该自编码器可以学习到原始数据的二阶特征 $h_k^{(2)}$，这里将 k设置为 1024，表示在原始数据经过第二个自编码网络后，它的维度变为 1024 维，如图所示：

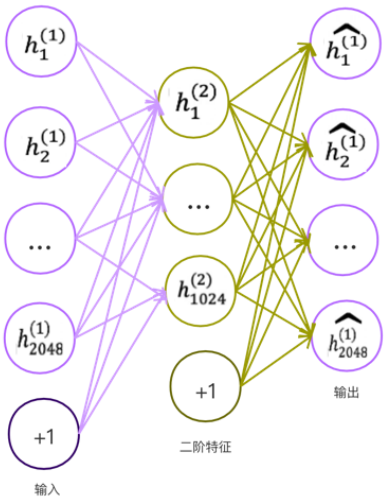

图 3.7 自编码器二阶特征学习过程

通过以上方法，就可以实现堆栈自编码降维的过程，我们将数据维度的从开始的 3578 维降至 1024 维，具体如图所示。同时，在自编码器初始化时我们向数据中加入了高斯噪声，自编码器从噪声中学习出数据的特征，将无规律的噪声略去，这样我们通过自编码器实现了降维和去噪的功能，最后，将通过自编码器处理的数据输入卷积神经网络进行分类。

3.4 基于卷积神经网络的结果预测

卷积神经网络是一种深度学习模型或类似于人工神经网络的多层感知器，常用来分析视觉图像，卷积神经网络常用来处理环境信息较为复杂，背景知识不太清楚，推理规则较为不明确的一些问题[42]。卷积神经网络的结构大致上分为，输入层、隐含层和输出层，隐含层又包括卷积层、池化层、全联接层。本文使用的卷积神经网络结构如下：

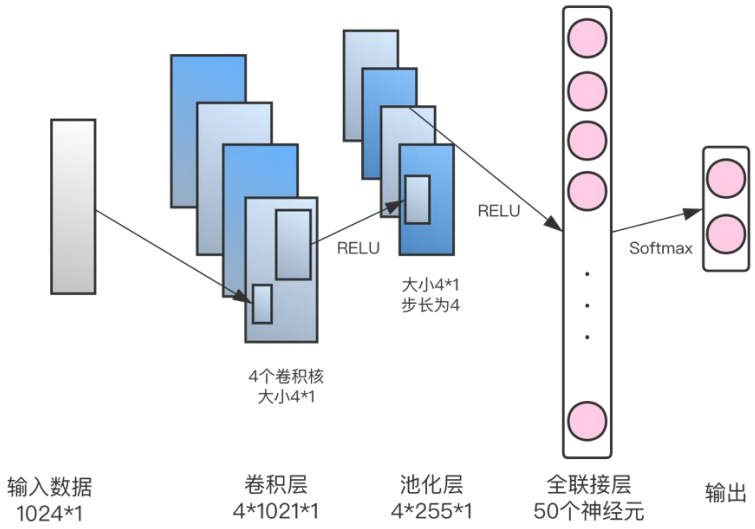

图 3.8 本文使用的卷积神经网络结构示意图

输入层：已知一维卷积神经网络的输入数据可以是一维或二维数组，二维卷积神经网络的输入数据可以是二维或者三位数组，本文的卷积神经网络的输入是一维数组，长度用 L 表示，这里的 L 为 1024。

卷积层：卷积层的作用是对输入数据进行特征提取，可以起到保留特征并且简化参数的作用。本文使用了 4 个大小为 4*1 的卷积核扫描二维数组，对于每一个卷积核而言，可以得到一个特征图，经过 4 个卷积核对二维数组的卷积，最终得到四张特征图。对于一个卷积核而言卷积过程如下图所示：

图 3.9 卷积过程示意图

本文设置步长为 1，采用的填充方式是 valid，因此最终得到的神经元长度为 $(L-4)+1$，有四个卷积核，最终得到的数据大小为$4*[(L-4)+1]$。采用 RELU 作为激活函数，RELU 计算方法为 $f(x)=\max(0,x)$，本文使用 RELU 作为激活函数的原因是 RELU 函数可以加快速度并且克服梯度消失问题。

池化层：池化是降采样的一种形式，我们常用非线性池化层来实现池化这一过程，非线性池化层有很多不同形式，本文使用最大池化层。其实现方法是将输入的数据划分为很多个矩形子区域，最大池化层输出每个子区的最大值作为这个位置的值。最大池化层的降维原理是我们只关注特征本身而忽略了特征的位置信息，因为特征的位置信息不如其他特征关系重要。池化层可以不断缩小数据空间的大小，因此它还减少了参数和计算的数量，也相应解决了过拟合的问题。本文选用的池化大小为4，步长为 4，最终输出的神经元大小为$[(L-4)+1]/4$ ，这里用到的激活函数为 RELU。

全联接层：全连接层可以将经过卷积和池化后神经网络的高层特征连接起来，减少特征位置带来的影响。它位于卷积神经网络最后的部分，特征图将在全联接层失去拓扑结构。本文的全联接层神经元个数为 50。

输出层：输出层的目的是输出分类结果，卷积神经网络的输出层的神经元通常不具有激活函数，因为它们本身就具有线性激活功能。本文输出层使用归一化指数函数(softmax function)输出分类标签，因为 softmax 分类器的分类结果在 0 到 1 之间，因此我们认为当输出结果大与 0.5 时表示该 miRNA 和疾病又关联，当小于 0.5 时表示该疾病与该 miRNA 无关联。

本文使用的卷积神经网路模型表示如图 3.8 所示，输入数据(1024*1)进入第一层卷积层，卷积层的激活函数为 RELU，该层包含 4 个卷积核并且每个卷积核的大小为4*1，步长为 1，经过这一步我们得到的神经元个数为1021*4；接着经过最大池化层，池化层大小为 4 并且步长也为 4，池化层的激活函数为 RELU，输出的数据大小为4*255；我们将4*255的数据输入模型的下一层全联接层，我们将参数设置为 50，这样输出结果为 50 个神经元；最后数据到达输出层，采用 softmax 分类器我们可以得到在 0 到 1 之间的分类结果。

3.5 本章小结

这一章主要介绍了本文使用的模型的构建方法，通过该模型可以预测 miRNA 与疾病的关联，结合 miRNA 和疾病的生物学原理以及 miRNA 在基因表达方面的重要作用，本文从疾病-基因-miRNA 三层网络中提取出 miRNA-基因以及疾病-基

因特征，这样做可以提取出基因层面的信息并且将基因信息融入模型中，接着对提取到的特征进行 softmax 归一化处理来消除离群点的存在；再将处理过的特征用堆栈自编码器进行降维处理，同时添加了高斯噪声的自编码器可以起到去噪作用；最后运用卷积神经网络进行分类。

第四章 实验结果与分析

4.1 实验结果

采用十折交叉验证的方式，对本文提出的模型进行评估，上文已经介绍模型的评估标准，包括精确度、准确度、召回率、F1 得分、ROC 曲线、PR 曲线等。首先介绍一下训练集的产生方式，我是通过上文提到的 HMDDv2.0 数据库中的数据得到的数据集，具体方法如下：

步骤一：通过 HMDDv2.0 数据库中的 miRNA 与疾病关联关系信息构建只包含 0 和 1 的邻接矩阵 Y，当 miRNA i 和疾病 j 有关联时，我们将 $Y(i，j)$设为 1，并记录 miRNA 和疾病的关联组合，共 5430 个符合关联条件的数据。

步骤二：若 $Y(i，j)$等于 0，则表示该位置的 miRNA 和疾病是否有关联这一信息尚未被实验验证，我们从中随机产生与正例相等数目的数据作为未知关联的反例数据。

步骤三：上面步骤所产生的数据集中所包含的 miRNA 和疾病有一部分无法在我们所建立的模型中得到，因此我们删去这一部分数据，最终得到的数据集大小为 3000，其中正例和反例各 1500。

本文使用十折交叉验证法，十折交叉验证法是将数据集等比例划分成十份，以其中的一份作为测试数据，其他的九份数据作为训练数据。交叉验证是把实验重复做 10次，每次实验都是从 10个部分中任意选取一份不同的数据作为测试数据(这里要保证 10个部分的数据都分别做过测试数据)，剩下的 9 个作为训练数据来训练模型，最后将得到的 10次实验的结果进行平分。

本文使用了精确度、准确度、召回率、F1 得分、ROC 曲线、PR 曲线来衡量模型性能，具体结果如表所示：

表 4.1 实验结果

精确度	准确度	召回率	F1 得分	AUROC	AUPR
0.881	0.871	0.904	0.893	0.925	0.948

通过在不同阈值之下的假正率和真正率绘制 ROC 曲线，具体 ROC 曲线如下图所示，ROC 曲线与坐标轴围城的面积称为 AUC，AUC 的值在 0-1 之间，AUC 的值越大，代表模型的准确性越高。

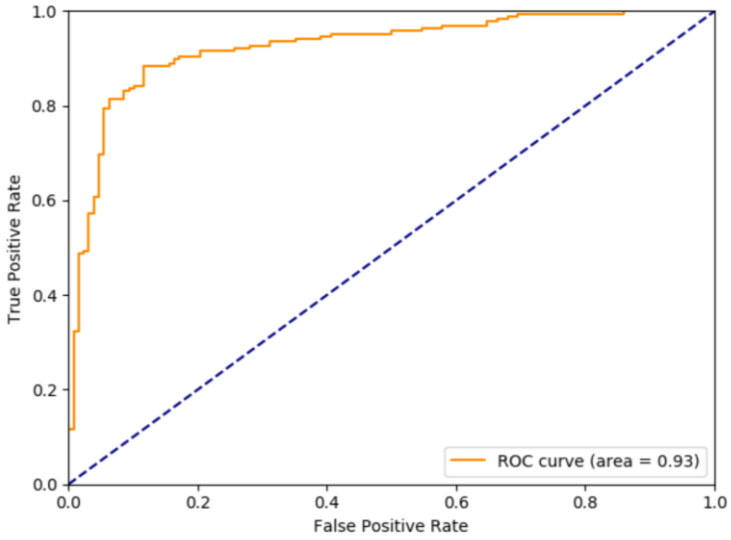

图 4.1 实验结果 ROC 曲线

通过在不同阈值之下的精确率和召回率绘制 PR 曲线，当 PR 曲线越接近(1,1)点时，代表模型的性能越好，PR 曲线与坐标轴围城的面积称为 AUC，AUC 的值越大，代表模型的性能越好。

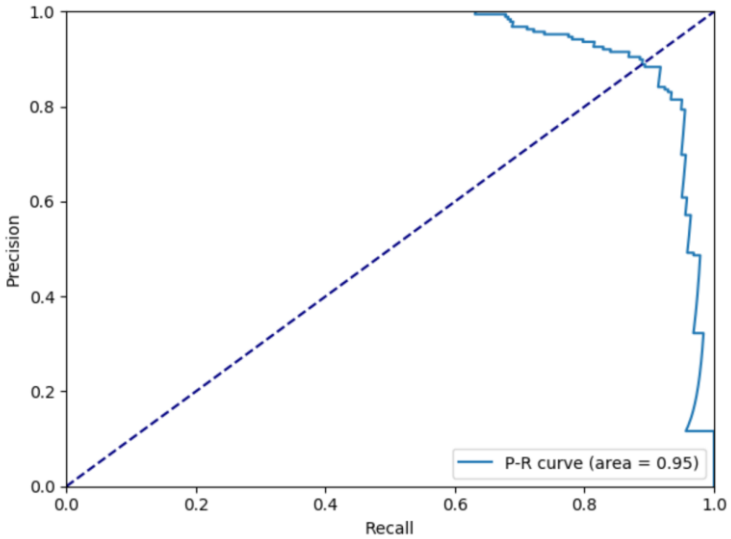

图 4.2 实验结果 P-R 曲线

4.2 结果对比

这里将本文提出的模型和其他四种模型进行比较，第一种是使用本文的特征提取与特征选择方法，分类器采用支持向量机(SVM)，第二种是 You[43]于 2017 年提出的模型 PBMDA，该模型构建了一个由三个相互关联的子图组成的异构图，并进一步采用了深度优先搜索算法来推断潜在的 miRNA-疾病关联。第三种是 Jiang[44]于 2010 年提出一种方法，该方法基于 miRNA 功能相似性网络和疾病相似性网络的超几何分布的方法来预测 miRNA 和疾病的潜在关联。第四种是 Shi[45]于 2013 年提出的模型 WPSMDA，该模型注重于蛋白质网络中 miRNA 靶标和疾病基因的关系，通过随机游走预测 miRNA 和疾病的关联。

表 4.2 实验结果对比

	精确度	**准确度**	**召回率**	**F1 得分**	**AUROC**	**AUPR**
本文模型	**0.881**	**0.871**	0.904	**0.893**	**0.925**	**0.948**
SVM 作为分类器	0.833	0.820	0.875	0.864	0.882	0.893
PBMDA	0.521	0.537	0.903	0.661	0.632	0.614
Jiang	0.589	0.631	**0.984**	0.741	0.790	0.834
WPSMDA	0.542	0.523	0.936	0.687	0.641	0.594

由实验结果对比可以看出，本文实现的模型在精确度、准确度、F1 得分以及 AUROC 和 AUPR 这些衡量指标方面表现的最好，我们可以看到在精确度方面，本文提出的模型比 PBMDA、Jiang 提出的模型以及 WPSMDA 精确度高出 30%以上，对于 F1 得分，本文提出的模型分别比 PBMDA、Jiang 提出的模型以及 WPSMDA 的得分高出 23.2%、15.2%和 20.6%，对于 AUROC 的值，本文模型的结果比 PBMDA、Jiang 的模型以及 WPSMDA 的结果中最好的值 Jiang 的模型高出 13.5%，对于 AUPR 的值，本文提出的模型的 AUPR 的值比三者中最好的结果 Jiang 的模型高出 11.4%。同时，本文采用的卷积神经网络作为分类器，与支持向量机(SVM)作为分类器的结果相比，本文提出的模型在衡量性能的各个指标上都是远超支持向量机的作为分类器的模型。

第五章 总结

从 1993 年人类首次发现 miRNA 到现在，人们逐渐发现 miRNA 在细胞生命活动的过程中起到了十分重要的调节作用。miRNA 可以通过与靶向基因特异性结合从而调控蛋白质的表达，进而影响功能和表型。因此 miRNA 与人类疾病发生息息相关，探索 miRNA 与疾病的关系一方面可以预测人类疾病的发生，在疾病的治疗方面也起到了至关重要的作用。计算机具有计算速度快，逻辑能力强及存储容量大等特点，因此在预测 miRNA 和疾病的关联过程中，通过计算机建立模型并进行预测变得也来越常见，人们在这领域也取得了十分重要的成果。本文也是通过生物信息学方法建立模型并对 miRNA 与疾病的关联进行预测，下面是对本文所做工作的总结：

本文提出一种通过有监督的机器学习方法对 miRNA 与疾病的关联进行预测，在模型构建方面，本文使用基因作为介导，构建 miRNA-基因-疾病三层网络并从这三层网络中提取了 miRNA-基因、疾病-基因特征，特征提取的方法参考了 Wang 在 2009 年发表的论文，主要思路是运用一种创新的回归算法求得不同疾病(MiRNA)之间的相关性的得分，基于上面的结果再通过皮尔逊相关系数求得疾病(MiRNA)与基因之间的关联得分。为了减少离群点和极值的影响，本文对提取到的特征进行了 softmax 归一化处理。接着本文对提取到的特征进行特征选择，这一步的目的一方面是降维可以提高模型的效率，同时向数据中加入高斯白噪声，这样也可以起到去噪的作用，本文采取堆栈自编码器进行上述处理。最后通过卷积神经网络对上述处理的到的数据进行分类，在数据集的产生方面，本文使用了 HMDD v2.0 数据库中的信息，正例为 HMDD v2.0 数据库中论文记载的经实验验证的 miRNA 和疾病的关联对，反例为随机产生与正例数目相同的 miRNA 与疾病对。在对模型进行评估方面，本文采用了十折交叉验证法对模型进行评估，采用的评估标准包括精确度、准确度、召回率、F1 得分、ROC 曲线、PR 曲线、AUC 得分等。同时，我们将本文提出的模型(卷积神经网络作为分类器)与支持向量机作为分类器的模型进行对比，可以看出以卷积神经网络作为分类器的分类效果明显好于用支持向量机作为分类器的模型，同时，我们将本文提出的模型和其他论

文结果进行对比，可以看出本文的模型明显优于其他模型。但针对预测 miRNA 与疾病关联这一问题，本文模型也存在一些问题，这些问题可以作为后续的研究方向：

(1) 在本文使用的数据量方面，因为本文使用的训练集和测试集都选自 HMDD v2.0 数据库，再经过筛选后符合本文条件的 miRNA-疾病关联条目共 3000 个，数据量过少可能导致模型的准确性降低，随着人们逐渐意识到 miRNA 和疾病的发生有关，越来越多的数据库可以为我们提供 miRNA 和疾病的关系数据，包括 miR2Disease、dbDEMC 等，运用更多的数据来预测 miRNA 和疾病的关联，会得到更好的效果。

(2) 在模型参数的选择方面，因为本文用到了堆栈自编码器和卷积神经网络，这两个机器学习框架都需要参数的支持，为了训练出最佳的模型，堆栈自编码器的参数和卷积神经网络的参数都要达到最佳，因此我们反复训练模型找到合适的参数能是模型达到更好的效果，但因为训练条件有限，不能保证本文模型使用的参数就是最佳参数。

(3) 对于模型的整体设计方面，本文的整体设思路是构建三层网络模型之后进行特征提取、特征选择、降维处理最后用卷积神经网络进行分类，取得了不错的结果，但是是否可以简化模型设计，让神经网络自己进行特征提取并分类是后续主要的研究方向，具体可以重点关注图卷积神经网络等较新的神经网络框架。

致 谢

四年的大学生活即将结束，今年是十分特殊的一年，从未想到大四的下半学期的大部分时间是在家度过的，也从未想过我的毕设要在家里完成，这对我而言也算是一个巨大的挑战。今年，是我大学生活的收场，曾经畅想着在这一学期与朋友们享受大学最后的时光，最终只能在家度过，有时会让我有些遗憾，但在今年如此特殊的情况下，我希望大家都可以永远健康、永远平安，这就足够了。

在论文即将完成之际，我想感谢我的论文指导老师鱼亮老师，如果没有她为我划定论文选题并对我进行细心指导，我很难写出这篇论文，她的指导使我终生受益。同时，感谢学长巨秉熠对我的帮助，他总是不厌其烦的回答我的每一个问题，为我论文的撰写提出了很多宝贵的意见，谢谢她(他)们。

今年对我来说真的很难，本来打算出国并且已经拿到 offer 的我因为疫情原因不得不改变自己的计划，时间紧急，在完成毕业设计的这段时间我也在努力找工作，每天都在写论文、复习面试知识、做笔试、答面试中度过。感谢 shopee 公司给我一个工作的机会，让我能静下心来完成论文。

还想感谢我的朋友们，尤其是 SCU 录取群中的朋友，可以说命运让我们这群有着相似经历的人走到一起，面对今年这样的情况，我们也曾绝望、也曾悲观、也曾抱怨，但是幸好我们大家都没有放弃，我们相互鼓励，相互分享找工作的资源，可以说，在面临着失学失业的双重心理压力下，你们让我认识到很多人都和我一样，我没有资格放弃。也正是因为相似，彼此之间的鼓励成为最有效的安慰剂，最终，我们群里的各位朋友们都取得了很好的结果，可以说，是我们相互成就了彼此，幸好我们没有放弃。

最后要感谢我的朋友与我的父母，是父母在我最绝望时让我重新振作起来，他们告诉我没关系，让我急躁的心情得以平静，向着看不清的未来而奋斗。他们为我带来了极大的安慰，让我可以安心找工作、写论文。因为我知道，父母永远在我身后默默支持我。

最后，由于我的学术水平有限，所写论文难免有思考不全面、不成熟的地方，恳请大家批评、指正。

参考文献

[1] Victor Ambros. The functions of animal microRNAs[J]. Nature, 2004, 431(7006): 350-355.

[2] Slack FJ, Basson M, Liu Z, 等. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor[J]. Molecular cell, 2000, 5(4):659.

[3] Thomas Tuschl, Phillip A. Sharp, David P. Bartel. A ribozyme selected from variants of U6 snRNA promotes 2 ', 5 '-branch formation[J]. Rna, 2001, 7(1):29-43.

[4] 于水澜, 于英君, 王桂云, et al. miRNA 与疾病相关性研究进展[J]. 牡丹江医学院学报, 2015, 036(001):94-96.

[5] 曹雪雁, 张晓东, 樊春海, et al. 聚合酶链式反应(PCR)技术研究新进展[J]. 自然科学进展, 2007, 17(5):580-585.

[6] 张庆峰, 高志贤, 王升启. 蛋白微阵列技术及其应用[J]. 中国生物工程杂志, 2004, 24(2):66-69.

[7] Jiang Q, Hao Y, Wang G, et al. Prioritization of disease microRNAs through a human phenome-microRNAome network[J]. BMC Syst Biol. 2010, Suppl 1(Suppl 1):1752-1759

[8] Mørk Søren, Pletscher-Frankild Sune, Palleja Caro Albert, et al. Protein-driven inference of miRNA-disease associations[J]. Bioinformatics, 2013(3):3.

[9] Xu C, Ping Y, Li X, et al. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles[J]. Molecular BioSystems, 2014, 10(11):2800-2809.

[10] Chen, X, Yan. et al. WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. Sci Rep 6, 2016, 21106 (2016):1356-1363.

[11] Chen X, Liu M X, Yan G Y. RWRMDA: predicting novel human microRNA-disease associations[J]. Molecular Biosystems, 2012, 8(10):2792-2798.

[12] Xuan P, Han K, Guo Y, et al. Prediction of potential disease-associated microRNAs based on random walk[J]. Bioinformatics, 2012, 10(11):11.

[13] Xing C, Clarence Y C, Xu Z, et al. HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction[J]. Oncotarget, 2016, 7(40):12-20.

[14] Xiangxiang Z, Li L, Lü Linyuan, et al. Prediction of potential disease-associated microRNAs using structural perturbation method[J]. Bioinformatics, 2014, 10(14):14.

[15] Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13):1644-1650.

[16] Xuan P, Han K, Guo M, et al. Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors[J]. Plos One, 2013, 8(8):70204.

[17] Chen X, Clarence Yan C, Zhang X, et al. RBMMMDA: predicting multiple types of disease-microRNA associations[J]. Scientific Reports, 2015, 5(1):13877.

[18] Pasquier C, Gardès, Julien. Prediction of miRNA-disease associations with a vector space model[J]. Scientific Reports, 2016, 6(1):27036.

[19] Chen X, Wu Q F, Yan G Y. RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction[J]. RNA Biology, 2017, 7(1):1-11.

[20] Xing C, Li H, Edwin W. LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction[J]. Plos Computational Biology, 2017, 13(12):e1005912.

[21] Fanyin Meng, Roger Henson, Hania Wehbe-Janek, et al. MicroRNA-21 Regulates Expression of the PTEN Tumor Suppressor Gene in Human Hepatocellular Cancer[J]. gastroenterology, 2007, 133(2):647-658.

[22] Huang Z, Shi J, Gao Y, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations[J]. *Nucleic Acids Res*, 2019, 47(D1):D1013–D1017.

[23] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants[J]. *Nucleic Acids Research*, 2017, 1(D1):D833–D839.

[24] Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans[J]. *Genome Res*, 2003, 13(10):2363-2371. 3

[25] Victor Ambros. The functions of animal microRNAs[J]. Nature, 2004, 431(7006):350-355.

[26] Huang Z, Shi J, Gao Y, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations[J]. *Nucleic Acids Res*. 2019;47(D1):D1013-D1017.

[27] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, et al.DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants[J]. *Nucleic Acids Research*, 2017, 1(D1):D833-D839.

[28] Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans[J]. *Genome Res*, 2003, 13(10):2363-2371.

[29] Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions[J]. Nature Methods, 2015, 12(8):697-697.

[30] Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13):1644-1650.

[31] You Z, Huang Z A, Zhu Z X, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction[J]. PLoS Computational Biology, 2017, 13(3):e1005455.

[32] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, et al..DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants[J]. *Nucleic Acids Research*, 2017, 1(D1):D833-D839.

[33] Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions[J]. Nature Methods, 2015, 12(8):697-697.

[34] Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans[J]. *Genome Res*, 2003, 13(10):2363-2371.

[35] Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13):1644-1650.

[36] Wu X, Jiang R, Zhang M Q, et al. Network-based global inference of human disease genes. Mol Syst Biol 4:189[J]. Molecular Systems Biology, 2008, 4(189):189.

[37] 孙平安, 王备战. 机器学习中的 PCA 降维方法研究及其应用[J].湖南工业大学学报, 2019, 33(1):73-78.

[38] 胡昭华, 宋耀良. 基于 Autoencoder 网络的数据降维和重构%Dimensionality Reduction and Reconstruction of Data Based on Autoencoder Network[J]. 电子与信息学报,2009, 031(005):1189-1192.

[39] 孙平安, 王备战. 机器学习中的 PCA 降维方法研究及其应用[J].湖南工业大学学报,2019,33(1):73-78.

[40] Kramer M A. Nonlinear principal component analysis using autoassociative neural networks[J]. AIChE Journal, 1991, 37(2):123-135.

[41] X. Yuan, B. Huang, Y. Wang, et al. Deep Learning-Based Feature Representation and Its Application for Soft Sensor Modeling With Variable-Wise Weighted SAE[J]. IEEE Transactions on Industrial Informatics, 2018, 14(7): 3235-3243.

[42] Alex Krizhevsky, I Sutskever, G Hinton. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2):33-38.

[43] You, ZH, Huang, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction[J]. PLOS COMPUTATIONAL BIOLOGY, 2017, 13(3):453-467.

[44]Jiang Q, Hao Y, Wang G, et al. Prioritization of disease microRNAs through a human phenome-microRNAome network[J]. BMC Systems Biology, 2010, 4 Suppl 1(Suppl 1):324.

[45] Shi H, Xu J, Zhang G, et al. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes[J]. BMC Systems Biology, 2013, 7(1):101.