基于异构网络表征的
微生物-药物关联预测研究

向 前 进

学　　院：计算机科学与技术学院　　专　　业：计算机科学与技术

学　　号：170110119　　指导教师：李君一

2021 年 6 月

哈爾濱工業大學

毕业设计（论文）

题　　目　基于异构网络表征的微生物-药物关联预测研究

专　　业　计算机科学与技术

学　　号　170110119

学　　生　向前进

指 导 教 师　李君一

答 辩 日 期　2021 年 6 月 8 日

摘　要

随着计算机技术的快速发展和生物数据的爆炸性增长，将计算机技术应用于生物数据之上来解决生物领域的相关问题的研究在国内外广泛开展。同时，网络表征学习方法的进步为复杂网络中的相关问题提供了有效的解决方案，被广泛应用于推荐系统、知识图谱、链接预测等领域。微生物-药物关联是生物领域中的重要信息之一，获取微生物和药物的关联关系是与人类健康有关的重要课题,有利于微生物治疗、药物开发等领域的发展。随着相关微生物-药物关联数据的公开，利用这些有价值的数据和网络表征技术来解决微生物-药物关联的预测问题十分必要。

本研究提出了一种微生物-药物关联预测方法 HNetMDA，使用异构网络表征学习算法 Metapath2vec，在微生物-药物异构网络中挖掘微生物节点和药物节点的低维稠密特征向量。本方法采用基于图神经网络的方法构造微生物-药物关联预测模型。实验结果表明，与现有的一些微生物-药物关联预测方法相比，本文提出的预测方法表现出了优良的预测性能，并且在时间和空间上有着很大的优势，能够以较高的准确率来预测未知的微生物-药物关联。

关键词：异构网络表征学习；复杂生物网络；微生物-药物关联预测；图神经网络

Abstract

With the rapid development of computer technology and the explosive growth of biological data, the application of computer technology in biological data to solve the related problems in the field of biology has been widely carried out at home and abroad.At the same time, the progress of network representation learning methods provides effective solutions for related problems in complex networks, which are widely used in recommendation systems, knowledge graphs, link prediction and other fields.Microbe-drug association is one of the most important information in the field of biology. Obtaining the association between microbes and drugs is conducive to the fields of microbial therapy and drug development, and is an important topic related to human health.With the publication of relevant microbe-drug association data, it is necessary to use these valuable data and network embedding technologies to solve the prediction problem of microbe-drug association.

This study presents a microbe-drug association prediction method named HNetMDA.It applies the representation learning algorithm *metapath2vec* to microbe-drug association heterogeneous network for uncovering the low-dimensional embeddings of microbe and drug.In this method, the graph neural network is used to construct the microbe-drug association prediction model.Experiments indicate that, compared with some existing microbe-drug association prediction methods, the method in this study has excellent prediction performance, and has great advantages in time and space.The result proves this model can predict the unknown microbe-drug association with high accuracy.

Keywords: Heterogeneous network representation learning,complex biological networks,microbe-drug association prediction,graph neural networks

目　录

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

随着生物科学领域方面技术的高速发展，生物信息学的数据资源也表现出极速增长之势。在海量的生物信息学数据中，出现了越来越多的关系型数据并成为现代生物信息学的重点研究对象之一。各类型热门的生物信息网络，例如基因共表达网络、细胞相互作用网络、蛋白质相互作用网络等都是研究生物系统的重要材料。复杂网络是用来建立信息结构、描绘数据关系的有效工具，作为描述关联属性数据的模型在近些年被广泛的使用，在复杂网络的研究领域中，链路预测本质上是挖掘网络产生连边的驱动力，是研究具有网络结构的数据演化和分析结构化数据特性不可或缺的重要工具[1]。在对具有高维度和较强稀疏性等特点的生物信息网络的处理方面，控制计算时间成本、存储空间成本对于目前大部分的网络分析方法都是一个挑战，并且这些方法在深入挖掘网络特征的能力方面也有所欠缺。令人惊喜的是，网络表征学习技术的快速发展为解决复杂网络中的分析问题提供了有效的解决方案，在知识图谱、推荐系统等领域得到了广泛应用。网络表征方法为生物复杂网络的分析带来了新的机遇。

微生物是单细胞或多细胞的微观活生物体的一类。积累的证据表明，微生物群落主要由细菌、病毒、原生动物和真菌组成，与人类宿主有密切联系。它们通常被认为是人类的“被遗忘的”器官，因为他们具有保护免受病原体攻击，提高代谢能力和增强免疫系统的功能。例如，微生物可提供保护，防止机会性病原体入侵；促进难消化多糖的代谢，通过合成必须的维生素来增强 T 细胞的反应。人类微生物在药物开发和精密医学中起着至关重要的作用。如今，如何系统地理解人类微生物与药物之间复杂的相互作用机制仍然是一个挑战。识别微生物-药物之间的联系不仅可以为了解作用机制提供重要见解，而且可以促进药物发现和已存在药物用途的扩展。考虑到生物实验的高成本和高风险，使用计算方法来解决是一种极佳的选择。但是，目前很少有计算方法能够解决这个任务。

基于上述内容，本课题提出了一种微生物-药物关联预测方法 HNetMDA，在使用已知的微生物-药物关联、微生物相似性和药物相似性信息构建出的异构网络中应用异构网络表征学习算法 Metapath2vec[2]，在微生物-药物异构网络中挖掘节点特征，而后使用深度学习的方法构建微生物-药物关联预测模型，将挖掘出的特

征作为模型的输入完成网络上的链接预测任务。本课题的研究补充了在生物网络中微生物（包括病毒）-药物关联预测方面的暂时还略有缺乏的内容，为高成本、高风险的生物药物试验提供更好的参考资料，挖掘出药物和微生物之间可能存在的关联关系，为基于微生物的治疗方式和药物发现提供有效帮助。

1.2 网络表征技术在生物信息网络中的应用发展概况

近几年来，网络表征学习方法在网络分析领域，特别是复杂生物网络分析中展现出了很高的学术价值和应用价值。网络表征学习方法可以更加深入地挖掘出传统的网络分析方法所得不到的网络的底部特征，为下游任务提供更好的服务，在许多的生物网络问题中得以应用，取得了非常先进的结果。

对于大规模网络（如社会网络、生物网络等），网络表征学习方法的目标是获得网络中节点或边的低维稠密表示，在将高维复杂网络转换为低维空间的同时，能够最大化地保留网络的结构特性，将网络中的节点或边表示为嵌入空间的特征向量，将得到的特征用于节点分类、节点聚类、边识别、链接预测等不同的任务[3]。网络表征学习方法在生物信息网络中表现出的有效性和潜力为生物信息研究领域带来了巨大的机遇，其在生物信息网络中主要用于解决链接预测和节点分类这两种下游任务。

1.2.1 链接预测任务

生物科学研究领域最重要的任务之一就是探索发现生物网络中的新链接关系，这对药物研发和致病基因发现等研究都有着巨大的帮助。已经有大量关于生物网络潜在链接预测问题的研究在国内外开展，例如 Zhu 等人开展的在异构网络上基于图正则化回归的药物-疾病关联预测[4]，Li 等人基于反向传播的图神经网络做的微生物-疾病关联预测[5]，Albert 等人进行的蛋白质相互作用预测[6]，Peng 等人基于神经网络开展的 microRNA-疾病关联预测[7]等。在复杂的生物网络处理方面，一般的网络分析方法面临着两个主要问题：第一，生物学特征难以提取并且成本很高；第二，难以保证提取出的特征的准确性和全面性。然而，网络表征学习方法的快速发展为苦恼受制于以上问题的研究者们带来了福音。例如药物-疾病的关联预测问题使用到了 SVD 算法，一种基于矩阵分解方法的网络表征方法；基于随机游走的方法的网络表征技术 RWR 被应用于微生物-药物关联预测等问题。网络表征学习方法为下游的链接预测等任务和原始网络数据之间搭建了一个桥梁，并且有着相当出色的表现。

1.2.2 节点分类任务

针对生物网络中的节点分类研究的主要目的在于挖掘节点在网络中的所扮演的角色以此来预测这些节点在具体的生命活动中的作用，这对生物实验的开展有着重要的指导意义。节点的分类主要根据节点在网络中的重要性，以及节点在网络中的分布情况这两种不同的角度来进行表征学习继而进行节点分类[8]。在生物网络的节点分类问题中，网络表征方法也得到了普遍的应用，例如基于随机游走的RWR[9]、Node2vec[10]被应用于蛋白质功能预测，基于深度神经网络的 SDNE[11]等方法被应用于基因功能分析，网络表征技术在生物信息网络的节点分类问题研究中也表现出了很优秀的效果。

1.3 基于网络表征的链接预测研究发展概况

链接预测问题可以说是一个图上的问题，它的目标是基于网络中的非线性信息以及节点的辅助信息，根据相似性、矩阵分解等内部逻辑进行节点关系预测，从而得到新的边，在社交网络中，它可用于用户商品推荐；在知识图谱中，它可用于实体关系学习；在基础研究中，它可进行图结构捕捉；在生物学领域，它可用于相互作用发现。

由于数据库资源的匮乏，早期时候的链接预测模型一般都被用于节点类型和边类型单一的同构网络上。有的研究方法从节点结构相似性的角度出发，认为与研究的节点在网络中有局部拓扑结构相似性的节点更倾向于产生新的联系。通过计算评定节点间的相似性，进而预测节点之间是否具有边的关系。这种基于相似性的计算可以依据网络结构中的一些观测指标或节点的外部特征等实现。这种研究方法比较直观、易于解释，但是与网络的耦合性高，在前期观察筛选衡量指标时要求比较高。另一种研究思路是将链接问题看作边是否存在的二分类问题，一般受限于网络稀疏性影响，在这种思路下，特征筛选和处理就显得极为重要。在后期任务中，模型一般基于二元关系的链接预测，这些研究继承了一部分单一网络预测的研究成果，例如，矩阵分解和相似性等算法被扩展应用到了二分网络中。

如今，主流的链接预测任务实施大致分为三步：第一，搜集相关数据，构建图网络；第二，使用一些方法，例如 Node2vec、RWR 等，进行网络的特征提取，挖掘网络信息；第三，搭建下游预测模型，利用获取的特征来完成下游的链接预测任务。这种模式下，特征提取和下游的链接预测任务得以解耦合，研究者可以更容易地尝试不同的方法来设计实现这两个模块，在这种模式下有着许多成功案例，这也是本课题实验开展的大致流程。在微生物-药物关联预测研究方面，也已

经有了许多优秀的预测方法，比如 KATZHMDA[12]、NTSHMDA[13]、WMGHMDA[14]、IMCMDA[15]、GCNMDA[16]等等，这为本课题研究提供了很好的参照对象。

1.4 本文的主要研究内容

本课题提出一个微生物-药物关联的预测方法 HNetMDA，以更加高效、更加准确的解决微生物-药物之间是否有关联的问题，为实际进行的相关生物实验提前指明方向，减小其实验成本，为了实现这一目标，本课题的主要研究内容包括以下几个方面：

（1）收集并且预处理已知的微生物-药物关联数据资源，构建微生物-药物异构网络。数据主要来源于 MDAD、aBiofilm、DrugVirus 三个数据库，其中包括了已知存在的一些微生物-药物关联关系，并进一步使用生物方面相关的计算方法得出微生物功能相似性网络和药物结构相似性网络，和微生物-药物关联网络一同构成微生物-药物异构网络。

（2）基于异构网络表征学习算法 Metapath2vec，使用不同类型，不同长度的元路径在构建出的微生物-药物异构网络中挖掘网络的结构特性，为药物节点和微生物节点构建出低维稠密的特征向量表示，之后在下游的链接预测任务中，利用链接预测结果的正确率和其他一些指标来评定生成的节点特征表示的质量，即特征是否具有较高的区分性和准确性。

（3）采用深度学习算法构建用于微生物-药物异构网络的链接预测模型。深度学习算法模型选用图卷积神经网络（GCN），许多研究已经开发了基于 GCN 的方法来处理生物信息学的相关任务。模型利用得到的节点特征表示，对异构网络中预设定的链接类型通过相关计算函数计算出节点之间存在链接可能性的得分，并且加入激活层函数，将得分进行归一化，投影到（0，1）区间，设定一定阈值来判定该链接是否存在。进一步的，在三个不同的数据库的数据下训练模型，并且进行交叉验证，获取模型的预测效果，并通过查阅文献的研究结果来对预测的结果进行比较验证。

第 2 章 微生物-药物关联数据

2.1 微生物-药物关联预测研究的相关数据库简介

随着近年来生物信息数据的迅猛增长，丰富的生物数据资源极大地推动了在生物信息网络方面的研究，一些数据库公开提供经实验验证的微生物-药物关联数据，这些数据的公开使得使用机器学习和深度学习技术来预测新的微生物-药物关联成为可能，相关数据资源主要包括：微生物-药物互作关系数据。

2.1.1 微生物-药物关联数据库

与人类相关的微生物群是多样化的和复杂的。它在人类健康和行为中起着重要的作用，与疾病的发生和发展密切相关。虽然微生物群落的多样性和分布已经得到了广泛的研究，但人们对人体微生物的功能和它们与药物之间相互作用的复杂机制知之甚少，这对药物的发现和设计很重要。

生物信息领域一直在微生物和药物之间关联的数据方面十分缺乏，主要是因为揭示微生物和药物之间是否真正存在关系需要进行常规湿实验室实验（例如基于培养的方法）来进行验证，这些方法不仅耗时、费力且昂贵，并且药物和微生物的种类繁多，要想探求它们之间每一对都是否有着关联关系，这显然是代价高昂且不切实际的想法。最近，一些经实验验证的微生物-药物关联的数据库公开可用，本课题所使用的微生物-药物关联数据库如表 2-1 所示。

表 2-1 微生物-药物关联数据库

数据库	网址
MDAD[17]	http://www.chengroup.cumt.edu.cn/MDAD/
aBiofilm[18]	http://bioinfo.imtech.res.in/manojk/abiofilm/
DrugVirus	https://drugvirus.info/tech_doc/

其中 MDAD 数据库是由中国深圳大学计算机科学与软件工程学院计算机科技系，中国矿业技术大学信息控制工程学院联合开发的针对微生物和药物的关联研究的数据库，是更深入理解微生物和药物相互作用的有用资源。该数据集包括介于 1388 种药物和 174 种微生物之间的 5505 种经过临床或实验验证的微生物-药物关联关系。在进行数据处理，删除了冗余信息后，最终获得了 1373 种药物和 173 种微生物之间的 2470 种关联。

aBiofilm 数据集记录了 1720 种独特的抗生物膜剂/药物，其中针对包括细菌和真菌在内的 140 多种生物和微生物，在微生物-药物关联研究中也被广泛使用。在过滤掉重复的数据后，最终得到了 2884 种微生物-药物关联，涉及 1720 种药物和 140 种微生物。

DrugVirus 数据集总结了 118 种化合物/药物的活动和发育状态，这些药物总共针对 83 种人类病毒，并且包括最近出现的新型冠状病毒 SARS-CoV-2。此外，还从药物数据库和相关出版物中手动搜集了 76 种药物与 12 种病毒之间的 57 种经过临床或实验验证的药物-病毒关联。最终得到了 933 种药物-病毒相互作用的关系，其中一共包括 175 种药物和 95 种病毒。

总的来说，以上三个微生物-药物关联数据集的统计数据如表 2-2 所示。

表 2-2 微生物-药物关联数据集的统计数据

数据集	微生物种类	药物种类	关联数量
MDAD	173	1373	2470
aBiofilm	140	1720	2884
DrugVirus	95	175	933

2.1.2 微生物-药物关联数据分析

一种微生物和一种药物的关联定义为一个元组(m, d)，其含义就是微生物 m 和药物 d 是存在互作关系的，即二者有关联，根据上一小节统计的微生物-药物关联数据可以看出，微生物的种类数量和药物的种类数量是极不平衡的，在 MDAD 数据集、aBiofilm 数据集和 DrugVirus 数据集中，药物的种类分别约为微生物种类的 7.94 倍、12.29 倍和 1.84 倍。而同时，每一种微生物关联的药物数量也有差别，例如在 MDAD 数据库中，针对全部微生物，每个微生物和药物关联数量的平均数约

为 14.28，中位数仅为 2，但是其中与药物关联最多的为序号为 121 的微生物，名称为绿脓杆菌（Pseudomonas aeruginosa），在自然界中分布极为普遍，是临床上较为常见的条件致病菌之一，会导致中耳炎、脑膜炎和呼吸道感染等疾病，它与 MDAD 记录的 1373 种药物中的 430 种都具有互作关系。微生物和药物的关联具体数量分布如图 2-1 所示，其中，有 69 种微生物只与一种药物具有关联关系，占所有微生物的 39.9%，而同时仅有 36 种微生物与药物的关联数量大于等于 10，占所有微生物数量的 20.8%。从该数据可以初步得出，针对微生物-药物的关联预测问题具有很强的不平衡性。

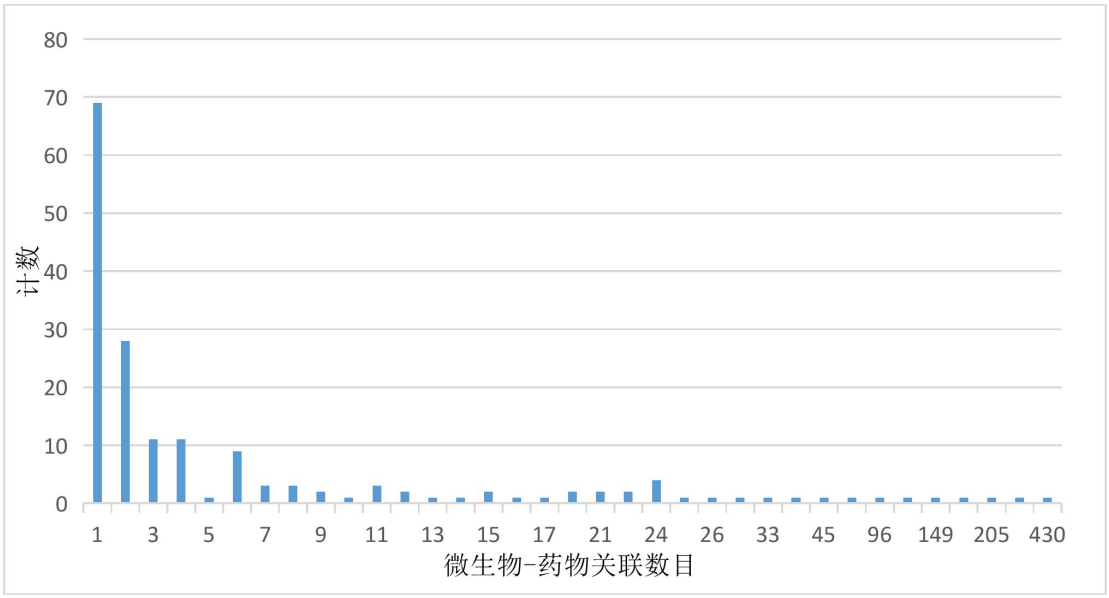

图 2-1 微生物-药物关联数量分布

2.2 本章小结

本章首先简要介绍了微生物-药物关联预测研究中常用的数据资源。进而介绍了从各个数据库中提取出来的微生物数目、药物数目以及它们的关联数目，并对这些数据进行了初步分析，简要论述了这类数据间存在的不平衡性，这也为本课题后续的异构网络表征方法选择奠定了基础，对数据的收集和处理为后续构建微生物-药物异构网络做好了数据准备。

第 3 章 微生物-药物异构网络构建

3.1 微生物-药物异构网络简介

在获取到微生物-药物关联的数据后，本课题对数据进行一定处理和定义，可以直接得到微生物-药物关联网络 I，而后使用相关计算方法和生物信息学上的一些计算工具，根据已有的微生物和药物信息来得到微生物功能相似性网络 FM 和药物结构相似性网络 DS，进而构建好整个微生物-药物异构网络，其概念图如图 3-1 所示。

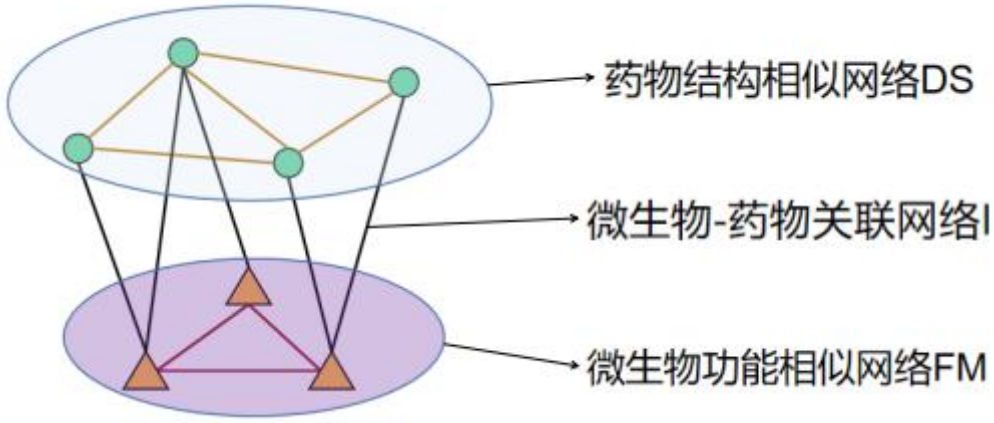

图 3-1 微生物-药物异构网络

3.2 微生物-药物关联关系网络

该关联网络的定义比较简单，基于第 2 章搜集到的微生物-药物关联数据来构建该关系网络 I，定义邻接矩阵 $I \in R^{nd \times nm}$ 来表示微生物-药物的关联关系，其中 nd 代表药物的种类数量，nm 表示微生物的种类数量。微生物-药物关联网络的邻接矩阵 I 的元素 I_{ij} 定义见公式 3-1。

$$I_{ij} = \begin{cases} 1 & d_i \text{和} m_j \text{有关联关系} \\ 0 & d_i \text{和} m_j \text{没有关联关系} \end{cases} \tag{3-1}$$

例如在 MDAD 数据集中有 173 种微生物和 1373 种药物，设某种药物编号为 d_i，某种微生物编号为 m_j，如果观察到药物 d_i 与微生物 m_j 之间的关联关系，那么令邻接矩阵中 I_{ij} 处的值等于 1，否则等于 0。容易观察到 MDAD 数据集中，微生物和

药物之间可能存在 237529 个链接，但实际上已知的仅有 2470 个链接，可以知道邻接矩阵 I 是一个稀疏矩阵，所以本课题在清洗完冗余数据之后选择以稀疏矩阵的形式来存储该网络，即将每一个有关系的节点对(m, d)存储在文件中，减少数据占用的空间。

3.3 微生物功能相似性网络 FM

在上述基础上，实验进一步构建微生物功能相似矩阵 FM 和药物结构相似矩阵 DS。其中，微生物功能相似矩阵 FM 使用 Kamneva 于 2017 年提出的根据微生物基因组和结构来计算微生物功能相似性的计算方法[19]计算得出，其流程如图 3-2 所示。

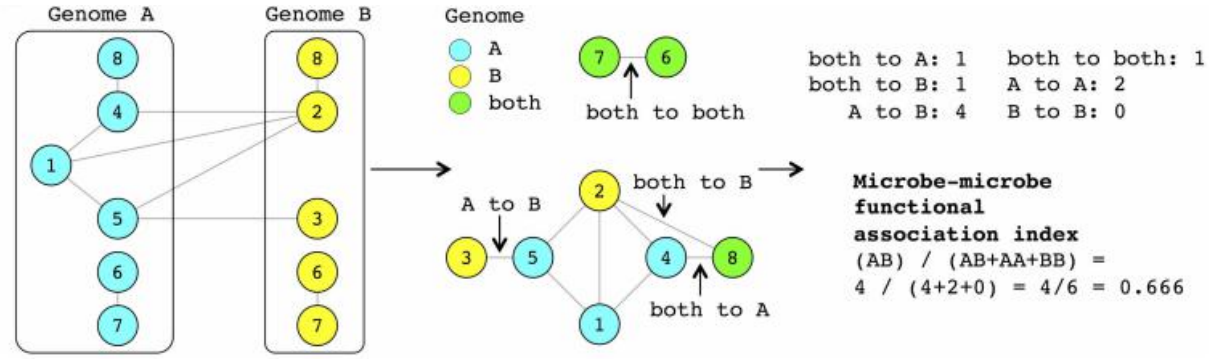

图 3-2 微生物功能相似性计算示意图[19]5

两个物种 A 和 B 的基因组编码的基因分别来自 6 个和 5 个基因家族。基因家族是一组具有相同重要特征的多个基因，这些基因在功能和结构上都具有相似性，编码相似的蛋白质产物，因为它们有着相似的 DNA 序列。同一家族基因紧密排列在一起，形成一个基因簇，但多数时候，它们是分散在同一染色体的不同位置上，或者存在于不同的染色体上的，各个基因有着独特的表达调控模式，这是生物体适应环境、提高工作效率的一种组织形式，便于调控和有效地开启和关闭。如图 3-2 中所示，有三个基因家族完全编码在基因组 A（1、4 和 5）中，有两个基因家族完全编码在基因组 B（2 和 3)中，还有三个（6、7 和 8）完全编码在两个基因组中，连接基因家族的边分为 6 类。通过不同种类边的计算，最终得到不同微生物功能相似性的结果。

以 MDAD 数据集的微生物为例，在以上计算步骤以后，再经过后续一些相似关系强化的处理，最终得到如表 3-1 所示的微生物功能相似性矩阵 FM。

表 3-1 微生物功能相似性矩阵 FM

微生物 ID	0	1	2	3	…	169	170	171	172
0	1.000	0.932	0.326	0.869	…	0.060	0.870	0.375	0.705
1	0.932	1.000	0.304	0.810	…	0.057	0.810	0.350	0.657
2	0.326	0.304	1.000	0.326	…	0.023	0.326	0.140	0.264
3	0.869	0.810	0.326	1.000	…	0.061	0.870	0.375	0.704
…	…	…	…	…	…	…	…	…	…
169	0.060	0.057	0.023	0.061	…	1.000	0.070	0.040	0.075
170	0.870	0.810	0.326	0.870	…	0.070	1.000	0.432	0.704
171	0.375	0.350	0.140	0.375	…	0.040	0.432	1.000	0.304
172	0.705	0.657	0.264	0.704	…	0.075	0.704	0.304	1.000

3.4 药物结构相似性网络 DS

药物结构相似网络 DS 使用 SIMCOMP2[20]方法测量得出，提出 SIMCOMP2 方法的作者将该方法做成了一个网站(www.genome.jp/tools/simcomp2/)以方便使用。SIMCOMP2 方法的数据库基于 KEGG[21]，即京都基因和基因组百科全书，KEGG 是一个数据库资源，是一个整合了基因组、化学和系统功能信息的数据库，旨在揭示生命现象的遗传与化学蓝图。它是由人工创建的一个知识库，是基于使用一种可计算的形式捕捉和组织实验得到的知识而形成的系统功能知识库，数据库中的记录用 C Number 唯一表示，每条分子都有对应的化学式，结构式，分子量等基本信息。SIMCOMP2 可以进行用户提供的两组指定化合物之间的化学结构相似性的全部计算。输入集可以是 KEGG Compound ID，或者是药物 ID，亦或是 MOL 文本文件。本课题通过药物名称查询到其在 KEGG 中的 ID，并以此作为 SIMCOMP2 方法的输入，最终获取到初步的药物结构相似性的信息。

而显然这还不够，很容易就知道，微生物功能相似性矩阵 FM 和药物结构相似性矩阵 DS 都是稀疏的，因为目前的数据库中仍然缺乏微生物的功能信息和药物的结构信息，许多微生物之间或者药物之间在 FM 和 DS 中并没有相似性信息，也就没有相似性评分。

为了发现更有价值，更加有利于研究展开的相似性信息，本课题根据文献[16]3-4 中提出的方法，选择利用高斯相互作用剖面核函数来计算微生物之间和药物之间的高斯核相似性，关键的思想是相似的微生物（药物）与相似的药物（微生物）相互作用，产生相似的相互作用特征，更具体的说，在第一步获得的微生物-药物关联关系矩阵 I 中，根据上述思想，将第 i 行 $\mathrm{I}(d_i)$ 和第 j 列 $\mathrm{I}(\mathrm{m_j})$ 定义为药物和微生物的相互作用关系，之后计算药物和微生物的高斯相互作用核相似度矩阵 GD 和 GM 的计算公式如式 3-2、式 3-3 所示：

$$\mathrm{GD}(d_i,d_j)=\exp(-\eta_{\mathrm{d}}\|I(d_i)-I(d_j)\|^2) \tag{3-2}$$

$$\mathrm{GD}(\mathrm{m}_i,m_j)=\exp(-\eta_{\mathrm{m}}\|I(m_i)-I(m_j)\|^2) \tag{3-3}$$

其中，η_d 和 η_{m} 分别表示药物和微生物的归一化核带宽，它们的定义如公式 3-4 和公式 3-5 所示：

$$\eta_d=\eta'_d/(\frac{1}{nd}\sum_{\mathrm{i}=1}^{nd}\|I(d_i)\|^2) \tag{3-4}$$

$$\eta_{\mathrm{m}}=\eta'_m/(\frac{1}{nm}\sum_{i=1}^{nm}\|I(m_i)\|^2) \tag{3-5}$$

其中 η'_d 和 η'_{m} 表示原始带宽，二者都设为 1。通过整合药物结构相似性和高斯核药物相似性，建立了最终的药物相似性矩阵 S_d，其定义如公式 3-6 所示：

$$S_d(d_i,d_j)=\begin{cases}\dfrac{GD(d_i,d_j)+DS(d_i,d_j)}{2} & DS(d_i,d_j)\neq 0\\ GD(d_i,d_j) & DS(d_i,d_j)=0\end{cases} \tag{3-6}$$

相似的，最终的微生物相似性矩阵 $\mathrm{S_m}$ 的定义如公式 3-7 所示：

$$S_{\mathrm{m}}(m_i,m_j)=\begin{cases}\dfrac{GM(m_i,m_j)+FM(m_i,m_j)}{2} & FM(m_i,m_j)\neq 0\\ GM(m_i,m_j) & FM(m_i,m_j)=0\end{cases} \tag{3-7}$$

最终得到的微生物-药物的异构网络 G 的邻接矩阵 A 表示如式 3-8 所示：

$$\mathrm{A}=\begin{bmatrix}S_d & I\\ I^T & S_m\end{bmatrix} \tag{3-8}$$

以 DrugVirus 数据集中的药物为例，最终的药物结构相似性矩阵 DS 的部分数据如表 3-2 所示：

表 3-2 药物结构相似性矩阵 DS

药物 ID	0	1	2	3	…	171	172	173	174
0	1.000	0.269	0.391	0.072	…	0.185	0.391	0.153	0.391
1	0.269	1.000	0.687	0.087	…	0.687	0.687	0.391	0.687
2	0.391	0.687	1.000	0.127	…	0.472	1.000	0.391	1.000
3	0.072	0.087	0.127	1.000	…	0.060	0.127	0.105	0.127
…	…	…	…	…	…	…	…	…	…
171	0.185	0.687	0.472	0.060	…	1.000	0.472	0.269	0.472
172	0.391	0.687	1.000	0.127	…	0.472	1.000	0.391	1.000
173	0.153	0.391	0.391	0.105	…	0.269	0.391	1.000	0.391
174	0.391	0.687	1.000	0.127	…	0.472	1.000	0.391	1.000

3.5 本章小结

本章首先简单介绍了组成微生物-药物异构网络的三个子网络以及它们的获取来源，之后分别对这三个子网络，微生物-药物关联网络 I、微生物功能相似性网络 FM、药物结构相似性网络 DS 进行了详细的介绍，尤其是它们之中元素值是如何计算得出的，I 很简单，如果微生物和药物之间有关联，那就把这个节点值定义为 1，否则就是 0；FM 通过美国加利福尼亚州斯坦福大学斯坦福大学生物学系的 Kamneva 提出的计算方法计算得出；DS 通过 SIMCOMP2 计算工具分析得出；将初步的 FM 和 DS 再用于进一步的相似性信息提取，利用高斯相互作用剖面核函数计算微生物之间和药物之间的高斯核相似性，最终得到相似性矩阵 S_d 和 S_m，构建出微生物-药物异构网络 G，为下一步特征提取奠定了良好基础。

第 4 章 基于异构网络表征的链接预测方法

4.1 异构网络简介

信息网络被定义为一个带有对象类型映射 $\varphi: V \to A$ 和链接类型映射 $\psi: E \to R$ 的有向图 G=(V, E)。每个对象 $v \in V$ 属于某一个特定对象类型 $\varphi(v) \in A$，且每个链接 $e \in E$ 属于关系类型集合 R：$\psi(e) \in R$。异构网络定义如下：一个信息网络中，对象的类型总数 $|A|>1$ 或者链接的类型总数 $|R|>1$，那么这样的网络就称为异构网络，否则就是同构网络[22]。在现实生活中，大量相互作用的、多类型的组件组成了大部分真实的系统，然而大多数的网络研究并没有区分网络中不同类型的对象和链接，仅仅将它们建模为同质信息网络。近年来，越来越多的研究者开始将这些互相联系的、具有多个类型的数据视为异构信息网络，并利用网络中对象和链接的结构类型的丰富语义来发展结构分析方法。相较于同构信息网络，异构信息网络包含着更加丰富的结构和语义信息，这给当代网络信息挖掘研究提供了很多机遇，也带来了很多挑战。

4.2 微生物-药物关联预测方法

破解微生物和药物之间的关联关系对于药物开发和微生物治疗技术的进步来说至关重要。本课题设计的微生物-药物关联预测方法流程如图 4-1 所示，图中左侧为构建微生物-药物关联预测模型的大致流程，右侧则是在每个阶段所需要完成的具体任务。本课题需要提取每一个微生物节点和药物节点的特征，特征提取方法采用针对异构网络的网络表征学习算法 Metapath2vec，特征来源于微生物-药物异构网络中，并对比不同长度、不同种类的元路径提取出特征在链接预测任务中表现出的效果。而后，构建基于 GCN 的链接预测模型，由在特征工程中提取出的节点特征计算出节点间存在链接的可能性得分，设置合适的激活层函数，在每条链接的得分激活层处理后，让链接的可能性得分分布在 0 到 1，最后根据一定阈值筛选出预测出的新链接，并且利用正负样本的得分来计算模型的性能参数 AUC 和 AUPR。其中，微生物-药物关联数据准备以及异构网络的构建等内容在第 2 章中已经介绍，本章主要介绍本课题的特征工程和模型构建部分。

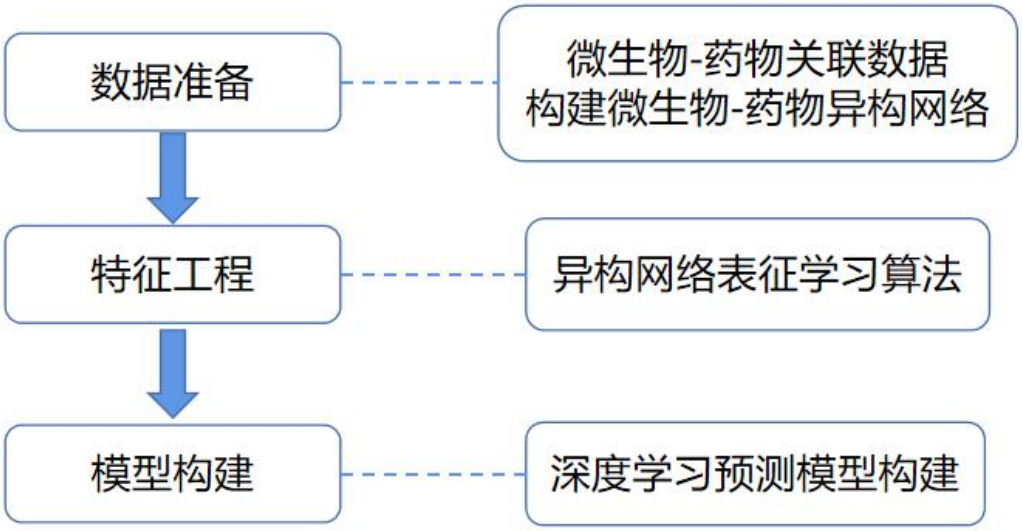

图 4-1 微生物-药物关联预测方法流程及框架图

4.2.1 学习微生物和药物节点的特征表示

网路表征学习被定义为：给定一个网络 G，学习一个 d 维的潜在表征 $X \in R^{|V|*d}, d << |V|$ 可以表征网络中顶点之间的结构信息和语义场景关系，首先介绍异构网络表征学习算法 Metapath2vec，这个算法为本课题研究的核心算法，课题研究中将其应用于微生物-药物异构网络中为微生物节点和药物节点学习特征表示，学习到的特征将被应用与后续的链接预测任务中，通过预测结果的正确率等参数来评定学习到的特征的优劣水平。

大量基于 word2vec 的网络表示学习框架，如 DeepWalk[23]、Node2vec[24]和 LINE[25]等被广泛应用于同构图的表示学习领域，对复杂网络的表示学习也具有很好的效果。这些表示学习方法可以自动发现有用的和有意义的（潜在的）特征，然而到目前为止，这些工作都集中在同构网络的学习上，仅仅代表单一类型的节点和节点间的联系，但是在现实社会中，大部分的信息网络本质上是异构的，涉及到节点类型的多样性和节点之间关系的多样性。这些异构网络带来的问题并不能被这些专为同构网络设计的表示学习方法所解决。本课题的关键之一就是研究异构网络中的表示学习问题，它的独特挑战来自于异构网络中多种类型的节点和链路的存在，这限制了传统网络嵌入技术的可行性。因此，本课题采用异构网络表征学习算法学习微生物-药物异构网络中的特征表示，具体使用的网络表征算法为 Metapath2vec，该算法的基本原理介绍如下。

在 Metapath2vec 中采用的方式和 DeepWalk 是类似的，采用 Skip-Gram 模型来学习图节点的表示，可以分为以下两步：1、利用元路径从图中获取游走序列。2、

利用异质 Skip-Gram 模型来学习节点的嵌入表示。Metapath2vec 不仅在各种异构网络挖掘任务（如节点分类、聚类等任务）中优于最先进的嵌入模型，并且能够识别不同网络对象之间的结构和语义相关性，这在大量的实验中都得到了验证。

首先介绍第一步：获取基于元路径的随机游走路径。Metapath2vec 的目标是最大化的保留一个异构网络的结构和语义信息的似然，首先使用基于 meta-path 的随机游走获取异构网络中每种不同类型顶点的异构领域，例如本实验中的一种 meta-path——“DMDMD”，即“药物-微生物-药物-微生物-药物”，然后利用扩展的 Skip-Gram 模型对之前得到的顶点邻域进行处理，最后学习得到各个类型节点的网络嵌入表示。基于元路径的随机游走可以定义成如下形式：

$$V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \ldots V_t \xrightarrow{R_t} V_{t+1} \ldots V_{l-1} \xrightarrow{R_{l-1}} V_l$$

在给定的一个异构网络图 G = (V, E, T)中，V 表示节点，E 表示图中的边，T 表示节点或者边的类型，且$|T_V|+|T_E|>2$，再有一个 meta-path 的模式，那么在随机游走的过程中，在第 i 步的转移概率 p 的定义如式 4-1 所示：

$$p(v^{i+1} \mid v_t; \rho) = \begin{cases} \dfrac{1}{N_{t+1}(v_t^j)} & (v^{i+1}, v_t^i) \in E, \varphi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t^i) \in E, \varphi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases} \tag{4-1}$$

其中$v_t^i \in V_t$，ρ 为预先设定的 meta-path，$N_{t+1}(v_t^i)$ 代表的是节点 v_t^i 的邻居中类型为 V_{t+1} 的节点集合，也就是说，在异构网络中的随机游走是在设定好的 meta-path ρ 的条件上进行的。在异构网络中的游走概率示意图如图 4-2 所示：

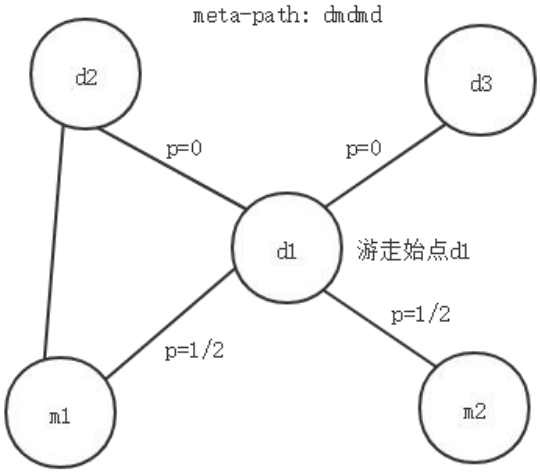

图 4-2 微生物-药物异构网络中随机游走概率示意图

一般来说，meta-path 具有对称的性质，即它的第一个节点 V_t 和最后一个节点 V_l 具有相同的类型，这个特性促使了它对随机游走的递归引导，如式 4-2 所示：

$$p(v^{i+1} \mid v_t) = p(v^{i+1} \mid v_l), if\,(t = l) \tag{4-2}$$

这种基于 meta-path 的随机游走策略很好地保证了在不同类型节点之间，语义关系可以正确的用于 Skip-Gram 模型中，通过训练得到节点的嵌入。

第二步，使用异质 Skip-Gram 模型获取节点的嵌入。对于异构图 G=(V,E,T)，目标是在给定节点 v 后，使其上下文内容存在的概率最大化，如式 4-3 所示：

$$\arg\max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} \log p(c_t \mid v; \theta) \tag{4-3}$$

其中，$N_t(v)$ 表示在节点 v 的邻接节点中类型为 t 的节点集合，概率函数 $p(c_t \mid v; \theta)$ 则为 softmax，如式 4-4 所示：

$$p(c_t \mid v; \theta) = \frac{e^{X_{c_t} \cdot X_v}}{\sum_{u \in V} e^{X_u \cdot X_v}} \tag{4-4}$$

这里的 X_v 就是从矩阵 X 中取出来的表示节点 v 的嵌入向量，可以注意到，softmax 中分母的累加需要遍历所有节点，计算量很大，引入负采样如式 4-5 所示：

$$O(X) = \log \sigma(X_{c_t} \cdot X_v) + \sum_{m=1}^{M} E_{u^m \sim P(u)} [\log \sigma(-X_{u^m} \cdot X_v)] \tag{4-5}$$

其中 σ 为 sigmoid 函数，P(u)是用于一共采样 M 次的节点负采样函数，采样时不区分节点类型，对所有节点都均匀采样。Skip-Gram 模型的流程图如图 4-3 所示：

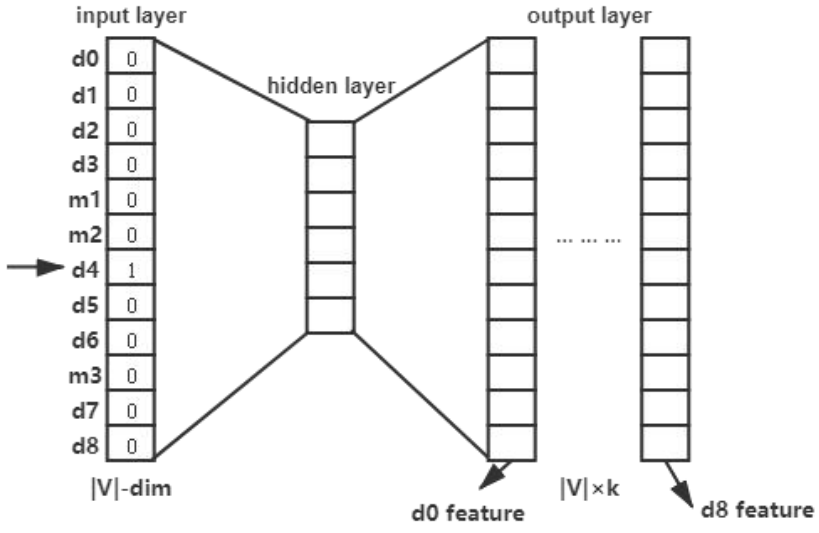

图 4-3 Skip-Gram 模型示意图

4.2.2 预测模型构建

针对微生物-药物异构网络中的链接预测问题，本课题选用的深度学习算法为图卷积神经网络(GCN)，具体模型为关系图卷积神经网络 RGCN[26]，RGCN 与 GCN 相似点在于它依旧基于消息传递，与 GCN 最大的不同点就在于 RGCN 会考虑边的类型和方向。在 GCN 中，每个节点的隐藏表示在 $(l+1)^{th}$ 层的计算方法如式 4-6 所示，$W^{(l)}$ 是 l 层中所有的边共享的，而 RGCN 中，不同类型的边使用不同的权重，只有相同关系 r 的边才能使用相同的映射权重 $W_r^{(l)}$，故在 RGCN 中，节点在 $(l+1)^{th}$ 层的隐藏表示计算方法如式 4-7 所示：

$$h_i^{l+1}=\sigma(\sum_{j\in N_i}\frac{1}{c_i}W^{(l)}h_j^{(l)}) \tag{4-6}$$

$$h_i^{l+1}=\sigma(W_0^{(l)}h_i^{(l)}+\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)}) \tag{4-7}$$

式 4-6 中，c_i 是正则化常数。式 4-7 中，N_i^r 表示在满足 $r\in R$ 的关系下，节点 i 的邻居节点的集合，$c_{i,r}$ 和 r 是正则化常数。

本课题中的关系图卷积神经网络由两个异构图卷积层堆叠而成，可以针对某一种或者多种类型的边进行图卷积运算，然后将所选类型的边上的消息根据聚合函数进行聚合。此处，激活层函数为 ReLU，聚合函数为 sum，卷积后将新的特征存入图上节点中，关系图卷积神经网络模型的示意图如图 4-4 所示。

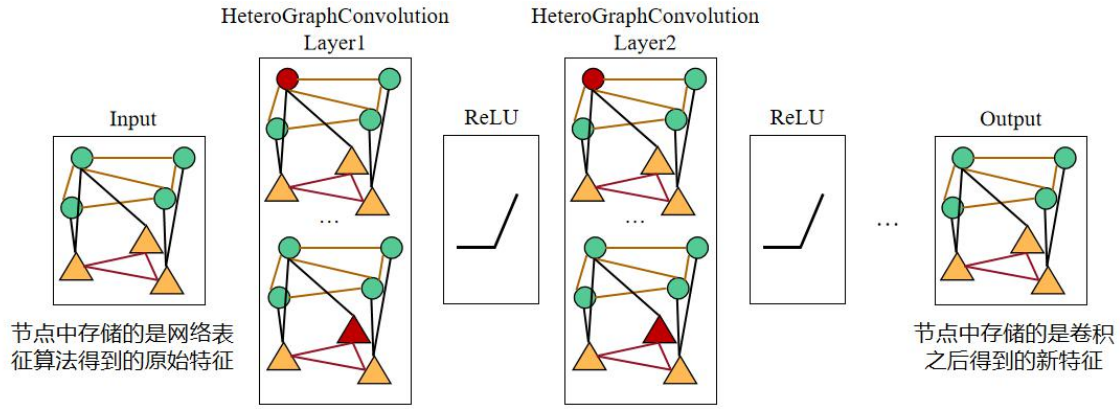

图 4-4 关系卷积网络示意图

基于图数据来完成图上的链接预测任务是非常常见且好用的一种链接预测方法，在本课题的研究中，将微生物-药物异构网络以图的形式实现。在图中不仅有节点的特征数据，还有图的结构，即节点之间是如何进行链接的。对于前者，我们很容易就可以获得每个节点的数据，而对于后者，想要提取出关于网络结构的

深层信息并非易事，将节点的特征与结构信息同时作为输入，然后让机器自己去决定要利用哪些信息是非常有效且方便的方法，这也就是使用网络表示学习算法的原因。GCN 是一种能够直接作用于图并且利用其结构信息的卷积神经网络。基于图卷积网络 GCN 的链接预测模型的基本思想是通过使用所需预测的节点对(u, v)的节点表示 $h_u^{(L)}$ 和 $h_v^{(L)}$，计算节点之间存在关联的可能性得分 $y_{u,v}$，其中的节点表示 $h_u^{(L)}$ 和 $h_v^{(L)}$ 在初始赋予特征工程中提取出的节点特征的基础上，再由多层 GCN 计算得出，如式 4-8 所示：

$$y_{u,v} = \phi(h_u^{(L)}, h_v^{(L)}) \tag{4-8}$$

训练一个链接预测模型涉及到对比两个存在链接的节点之间的得分与任意一对节点得分的差异。例如，对于一条给定的链接 u 和 v 的边，一个好的模型希望 u 和 v 之间的得分 $y_{u,v}$ 要高于 u 和从一个任意的噪音分布 $v' \sim Pn(v)$ 中所采样的节点 v' 之间的得分，这样的方法称为负采样。许多损失函数，例如交叉熵损失(见式 4-9)、贝叶斯个性化排序损失(见式 4-10)和间隔损失(见式 4-11，M 为常数项超参数)等都可以实现上述目标，本次模型构建选择的损失函数为间隔损失，常数项超参数 M 选取为 1。

$$L = -\log\sigma(\mathrm{y_{u,v}}) - \sum_{v_i \sim P_n(v), i=1,\ldots,k} \log[1-\sigma(\mathrm{y_{u,v_i}})] \tag{4-9}$$

$$L = \sum_{v_i \sim P_n(v), i=1,\ldots,k} -\log\sigma(y_{u,v} - y_{u,v_i}) \tag{4-10}$$

$$L = \sum_{v_i \sim P_n(v), i=1,\ldots,k} \max(0, M - y_{u,v} + y_{u,v_i}) \tag{4-11}$$

本课题研究中基于 GNN 的链接预测模型的整个流程为：将 Metapath2vec 算法得到的每个节点的特征向量赋给异构网络中的对应节点，将其作为每个节点进行卷积的初始特征信息，在训练模型时需要输入的信息有：异构网络、预测的边的类型、以字典形式保存的节点特征、节点特征向量的维度，用于计算节点间存在链接可能性的得分的函数选择有边类型选择的点积计算函数。训练过程中，进行针对需要链接预测的边类型的负采样，构造负采样图，训练时使用模型进行前向传播计算，得到原始图上需要预测边的得分和负采样图中负采样得到的边的得分，而后基于得到的链接预测得分 pos_score 和 neg_score 来进行损失函数计算，从而完成后向传播。最后，使用训练好的模型进行在测试集上的预测，由于任务为链接预测而非分类，在归一化时选用 sigmoid 函数，使链接预测的的得分分布到 0 和 1 之间，便于后续的 AUC 和 AUPR 等性能参数的计算。

4.3 本章小结

本章详细介绍了微生物-药物关联预测模型 HNetMDA 的流程，如图 4-5 所示。首先简要介绍了异构网络，也就是本课题研究的数据表示形式。接下来详细介绍了特征工程和预测模型构建部分，包括在微生物-药物异构网络上应用异构网络表征学习算法 Metapath2vec 得到微生物和药物节点的嵌入空间向量特征，链接预测模型构建部分介绍了图卷积神经网络 GCN 在进行链接预测任务时的基本思想、预测模型的整个运行流程和其中使用到的数据和函数。

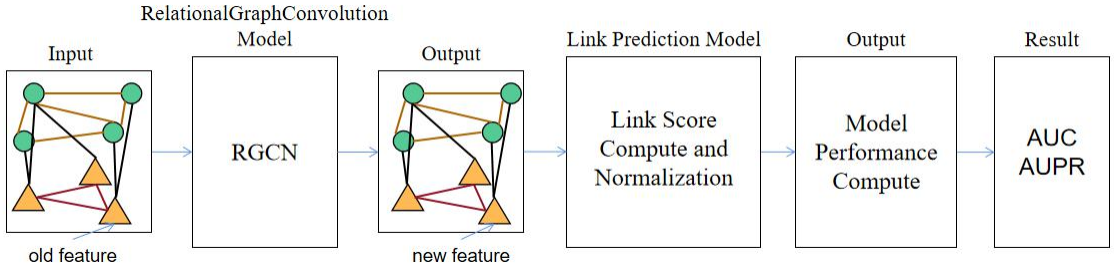

图 4-5 HNetMDA 模型示意图

第 5 章 微生物-药物关联预测模型验证

5.1 实验设置

实验采用的评价指标主要有两种：AUC 和 AUPR。AUC 为 ROC 曲线下面积的标准度量(Area Under the ROC Curve)，AUPR 为 PR 曲线下面积的标准度量(Area Under the Precision-Recall Curve)。AUC 是衡量学习器优劣的一种性能指标，一般的，ROC 曲线一直都处于 y=x 这条直线的上方，所以 AUC 的取值范围在 0.5 和 1 之间。AUC 越接近 1.0，那么这个方法/模型的真实性越高，等于 0.5 时最低，也就是说该检测方法没有应用价值。PR 曲线是由召回率和正确率组成的曲线图，AUPR 就是 PR 曲线下的面积，如果只有较高的召回率，只能说明模型或者方法可以预测出较多的数据，但是不能保证预测出的样本是正确的；如果只有较高的正确率，只能说明所预测的样本是正确的，但仅仅局限于一小部分数据集上。

实验验证方法选用 n-折交叉验证，n 取 2、5 和 10，相较于直接将数据分为训练集和测试集后仅训练一次就得出结果的常规方法，交叉验证可以对模型做出更加合理准确的评估，并且在一定程度上减小过拟合。用于分组的数据为异构网络中需要预测的种类的链接，将总的链接一共分为 n 份，在 n 次循环中，每一份数据依次作为该次循环的测试集，不参与模型训练，其余的 n-1 份数据作为训练集用于训练模型，对测试结果取平均值，由于需要计算 AUC 和 AUPR，不能只有正样本，从负采样图中随机抽取和测试集大小相同的测试负采样集，也就是本来不存在的链接，最终通过一定的计算方法得到预测结果，该预测结果也代表了 Metapath2vec 异构网络表征学习方法所提取出的特征的水平。同样的，针对 Metapath2vec 算法的两个主要参数特征维度 dim，元路径模式 P 进行参数测试调优。由于数据集的规模有些许差异，实验中针对三个数据集 MDAD、aBiofilm 和 DrugVirus 分别选取在下游链接预测任务中取得 AUC 数值最高的参数 dim 和 P，并尝试分析不同参数设置下提取出来的特征水平差异的原因。同样的，在链接预测模型中也有一个重要参数，负采样次数 k，实验过程中也将对这个参数进行调优，以达到模型最佳的性能。最后，将本课题研究得到的模型与目前已有的一些模型进行 AUC 和 AUPR 性能指标的对比，以得出微生物-药物链接预测模型的性能水平。

5.2 微生物-药物关联预测模型表现

本节将从模型在三个数据集上的 AUC 值、AUPR 值等性能参数方面展示模型 HNetMDA 的预测结果，分析不同参数设置对特征提取和预测模型性能的影响。

5.2.1 微生物-药物关联预测模型参数的影响

为了检验 Metapath2vec 算法两个主要参数表征维度 dim 和元路径模式 P 以及链接预测模型中负采样次数 k 对预测模型性能的影响。实验分别对上述三个参数进行调优，其中 dim 的取值范围设置为{4，8，16，32}，负采样次数 k 取值范围设置为{1，2，3，4，5，6，7，8}，元路径模式 P 从长度上进行范围设置，长度取值范围为{3，5，7，9}，训练次数设置为 100，采用五折交叉验证方法，比较的性能参数为 AUC 值。首先是模型中的负采样次数 k，在 Metapath2vec 算法提取特征部分统一设定 dim=16，P='dmdmdmd'，其 AUC 表现如图 5-1 和表 5-1 所示。

表 5-1 不同负采样次数 k 下微生物-药物关联预测模型的表现

数据集	AUC 最大值	负采样次数 k	AUC 标准差
MDAD	0.9119	4	0.0030
aBiofilm	0.9198	4	0.0025
DrugVirus	0.8081	4	0.0075

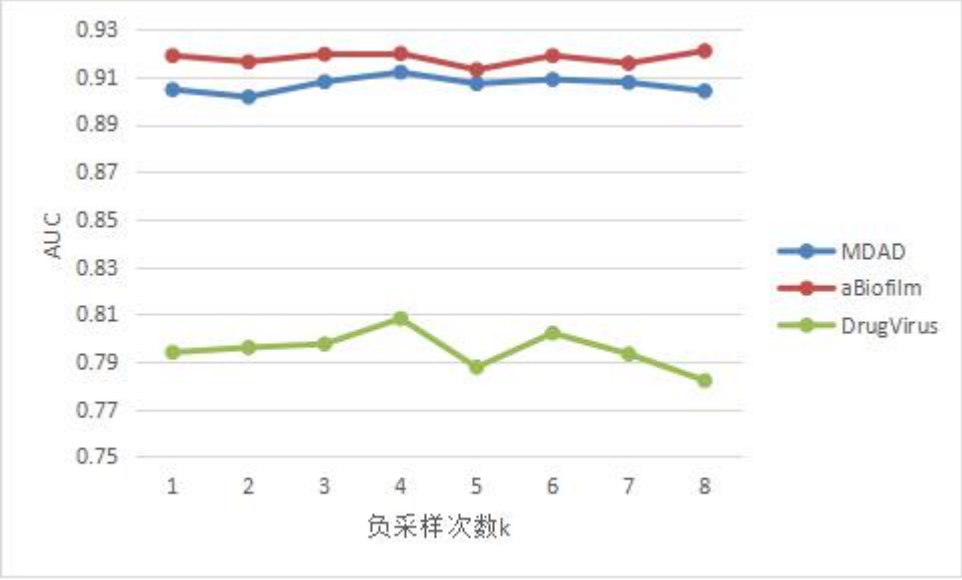

图 5-1 负采样次数 k 对预测模型性能的影响

由图 5-1 中折线走势和表 5-2 中的标准差可以看出，负采样次数 k 对模型性能影响并不大，且对于三个数据集，最大值都在 k=4 处取得，故在后续的实验中，负采样次数 k 都设置为 4。

在固定负采样次数后，实验对特征的维度进行了参数测试，测试时统一设定 k=4，P='dmdmdmd'，训练次数 epoch=100，采用五折交叉验证方法，比较的性能参数为 AUC 值。其结果如表 5-2 和图 5-2 所示，可以看到 AUC 值随着 dim 的变大而增大，但是通过标准差也可以看出，对于 AUC 的影响也不是特别强。也就是说 Metapath2vec 算法提取特征时，节点特征表示的维度在一定范围内越大，所能学习到的节点信息越多，用于链接预测时的效果就越好，本课题在参数测试后设置表征算法中的特征维度 dim 为 32。

表 5-2 不同特征维度 dim 下微生物-药物关联预测模型的表现

数据集	AUC 最大值	特征维度 dim	AUC 标准差
MDAD	0.9215	32	0.0056
aBiofilm	0.9437	32	0.0129
DrugVirus	0.8124	16	0.0105

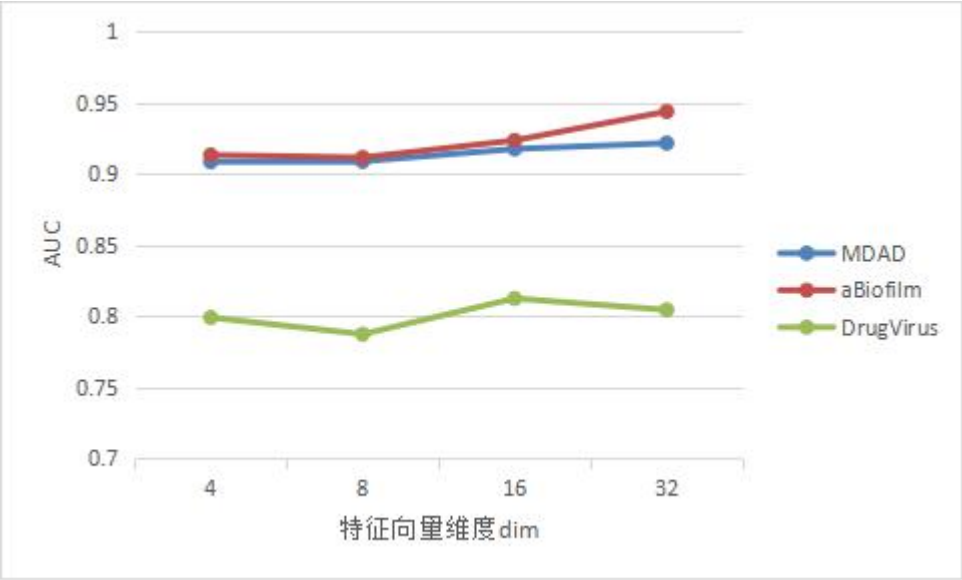

图 5-2 特征向量维度 dim 对预测模型性能的影响

实验已经确定了两个参数的选择：负采样次数 k 和 Metapath2vec 表征算法中的特征维度 dim，接下来需要对表征算法中的元路径模式 P 在 P 的长度设置方面进行测试，结果如表 5-3 和图 5-3 所示。

表 5-3 不同元路径模式 P 下微生物-药物关联预测模型的表现

数据集	AUC 最大值	元路径模式 P 长度	表征节点数（总数）
MDAD	0.9303	7	1531(1546)
aBiofilm	0.9465	9	1848(1860)
DrugVirus	0.8067	7	245(270)

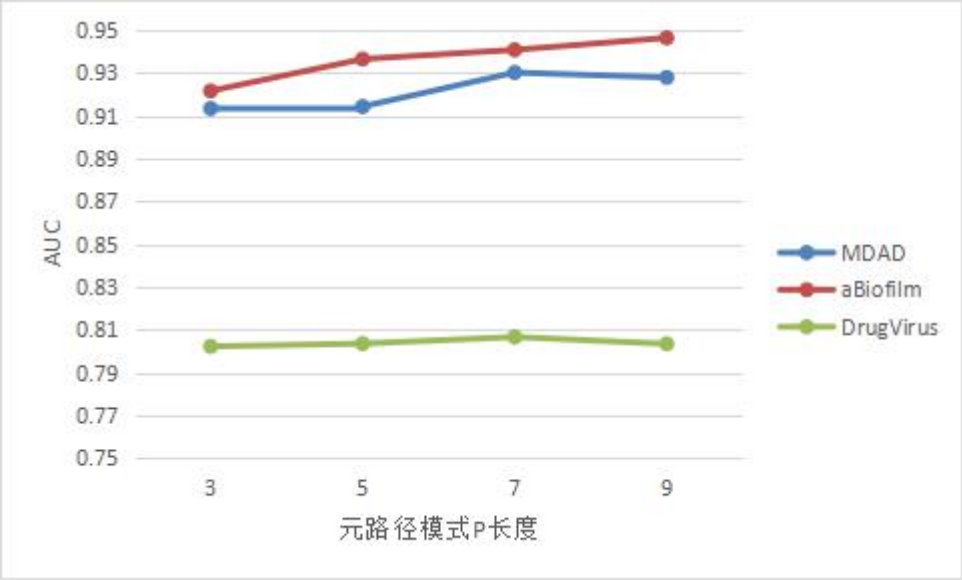

图 5-3 元路径模式 P 对预测模型性能的影响

从实验结果来看，元路径长度越长，所能表征到的节点越多，这也更加有利于模型的预测，而当元路径模式长度为 9 的时，虽然表征的节点数量有所增加，但不是所有数据集的预测结果都更好，因为数据集中的数据量不同，所以对不同的数据集要选择不同的元路径模式。

5.2.2 微生物-药物关联预测模型性能表现

针对三个不同的数据集，都进行二折、五折和十折交叉验证，在测试集上进行简单的预测正确率计算，参数设置均为实验测试出的最优参数：特征维度 dim=32，元路径模式 P 各自采用最优模式，负采样次数 k=4，训练次数 epoch=100，AUC 和 AUPR 的结果取 n 折交叉验证的平均值，保留小数点后四位，其结果如表 5-4、表 5-5 所示。

表 5-4 微生物-药物关联预测模型 AUC 值表现

数据集	2-fold CV	5-fold CV	10-fold CV
MDAD	0.9248±0.0138	0.9312±0.0061	0.9331±0.0141
aBiofilm	0.9368±0.0045	0.9472±0.0050	0.9516±0.0068
DrugVirus	0.7995±0.0037	0.8063±0.0158	0.8084±0.0284

表 5-5 微生物-药物关联预测模型 AUPR 值表现

数据集	2-fold CV	5-fold CV	10-fold CV
MDAD	0.9031±0.0022	0.9201±0.0089	0.9224±0.0188
aBiofilm	0.9234±0.0055	0.9340±0.0079	0.9403±0.0101
DrugVirus	0.7589±0.0138	0.7597±0.0210	0.7624±0.0443

在五倍交叉验证下，ROC 曲线和 PR 曲线如图 5-4、图 5-5、图 5-6 所示。

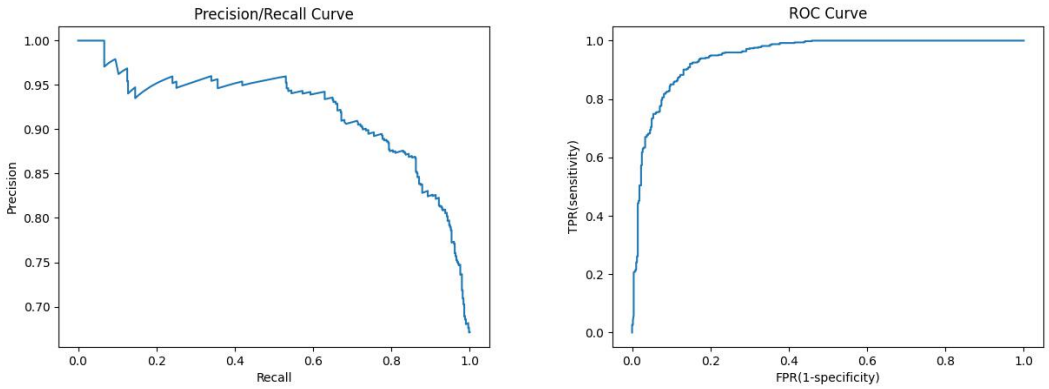

图 5-4 MDAD 数据集上的 PR 和 ROC 曲线

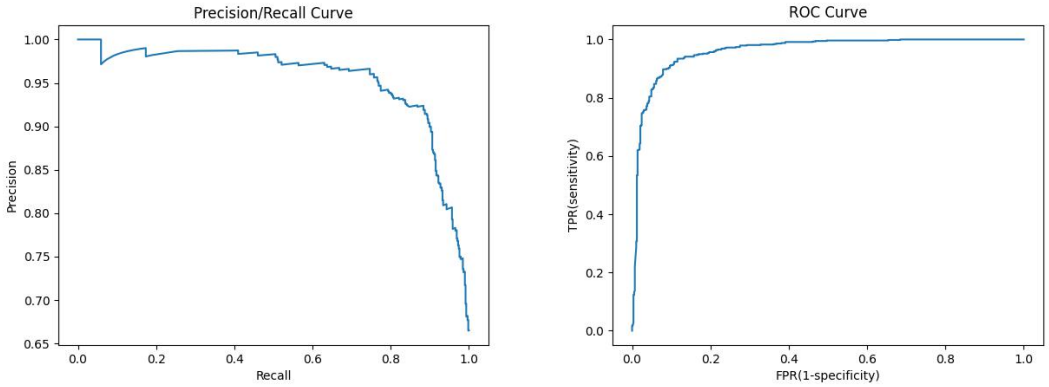

图 5-5 aBiofilm 数据集上的 PR 和 ROC 曲线

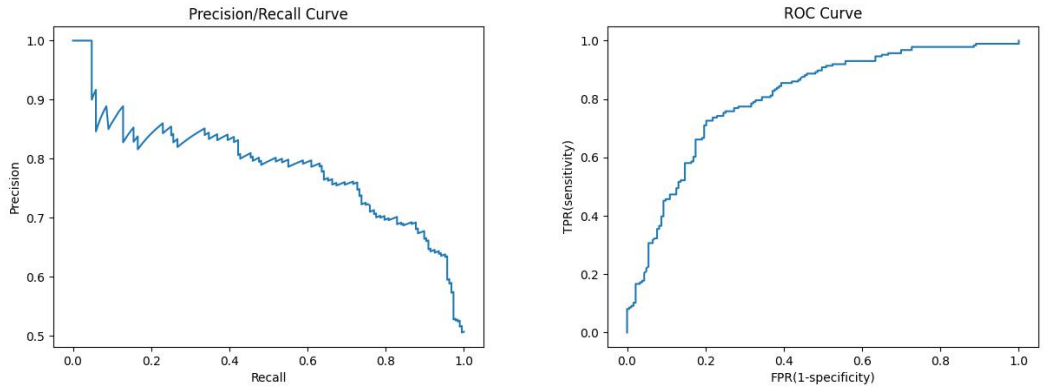

图 5-6 DrugVirus 数据集上的 PR 和 ROC 曲线

为了体现本课题做出的模型 HNetMDA 的性能优越性，在查阅相关文献后获得其他一些预测微生物-药物关联的方法在这三个数据集上的性能，以五倍交叉验证为例，对比实验中 epoch 设置为 200，AUC 和 AUPR 值对比如表 5-6 所示，其中最好的结果用加粗标记，第二好的结果用下划线标记。

表 5-6 HNetMDA 与其他预测方法性能比较

Methods	MDAD		aBiofilm		DrugVirus	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
KATZHMDA	0.8723	0.8384	0.9013	0.9020	0.7809	0.7554
NTSHMDA	0.8302	0.7924	0.8213	0.7639	0.7389	0.6973
WMGHMDA	0.8654	0.8381	0.8451	0.8903	0.7230	0.7687
IMCMDA	0.7466	0.7773	0.7750	0.8572	0.6235	0.6962
GCNMDA	0.9423	0.9376	0.9517	0.9488	**0.8986**	**0.9038**
HNetMDA	**0.9661**	**0.9629**	**0.9734**	**0.9666**	0.8076	0.7831

可以看到，本课题的研究结果在性能上已经超越了绝大部分已有的预测方法，并且有着运行时间快，数据处理简单的优势。其中的 GCNMDA 方法，该方法提取的特征由于注意力机制和条件随机场，所以用于预测时的效果非常好，尤其是对于 DrugVirus 这个数据平衡性很差的数据集，它也做到了很好的预测效果，本课题的未来优化方向也是在特征提取部分，使用一些强化初始数据的方法或对表征算法提取出的特征进行后处理等手段得到更好的表征结果，让模型性能更加优秀。

5.3 本章小结

本章通过实验来验证了本课题研究提出的微生物-药物关联预测模型的性能表现。详细展示了实验过程中的参数调优过程，并且对参数选取的理由进行了一定的说明和分析，绘制了在不同数据集上的 ROC 曲线和 PR 曲线图，并且最后与如今已有的微生物-药物关联预测方法在相同的数据集上进行了性能参数 AUC 和 AUPR 的对比，充分展现了本课题研究提出的预测模型的性能优越性，可以得出结论，本文所提出的微生物-药物关联预测方法 HNetMDA 能够有效地从异构网络中提取出优质的节点特征，并做出置信率较高的链接存在预测，能够为微生物和药物关联的相关生物研究提供有效的参考资料。

结　论

微生物与人类的关系是极为密切的，绝大部分微生物都是对人无害甚至是有益的，它们可以被看作人身上的功能“器官”，帮助人体免受病原体攻击，合成维生素。预测微生物-药物之间可能的关联关系有利于帮助生物领域的研究避开一些不必要的实验，将人力和物力都朝着最有可能的方向投入，从而减小实验的成本，有助于微生物治疗方法的开发和新药物的研究。随着科学技术的进步，复杂生物网络的数据也快速增多，利用这些数据提取出有价值的信息来解决预测微生物-药物关联问题是很有必要的，与此同时，网络表征技术为解决在复杂网络中的特征挖掘任务提供了有效范例。本课题为了更准确地预测微生物和药物之间潜在的关联关系，提出一种基于异构网络表征算法的微生物-药物关联预测方法 HNetMDA，取得的主要研究成果如下：

(1) 对公开的微生物-药物关联数据进行搜集和处理，得到了用于课题研究的微生物-药物关联数据，并利用一些计算方法和工具得出了微生物功能相似性和药物结构相似性，构建出微生物-药物异构网络。对数据集中微生物和药物的关联数量进行分析发现少数的微生物却占有了大部分的关联，数据存在很强的不平衡性。

(2) 首次将异构网络表征学习算法 Metapath2vec 应用于微生物-药物异构网络中，不同于一般的同构网络表征算法，Metapath2vec 可以有效地挖掘异构网络中特殊的结构信息，为微生物节点和药物节点构建出低维稠密的特征向量表示。经过链接预测任务实测，Metapath2vec 异构网络表征算法能够挖掘出准确的、具有区分性的微生物和药物节点特征，为下游任务提供较好的输入。

(3) 针对在异构网络中的链接预测问题，研究采用基于图神经网络的 RGCN 算法构建模型，其在异构网络上的链接预测、节点分类等任务中都有着极佳的性能表现。经实验验证，本课题提出的微生物-药物关联预测模型 HNetMDA 在三个数据集上有着较好的预测表现，相较于已有的大部分预测方法都有着更优秀的性能。

本文提出的基于异构网络表征的微生物-药物关联预测方法能够以较高的准确率预测微生物-药物关联，但仍存在不足和需要改进的地方：

(1) 进一步优化提取出的节点特征，针对未被游走到的节点进行后处理，减少其对预测模型性能的影响，同时也能更好解决数据不平衡的问题。

(2) 加入注意力机制，更好地利用微生物相似性和药物相似性数据，以提升微生物-药物关联预测模型的性能。

参考文献

[1] 司璀琪. 基于复杂网络结构表征学习下的链路预测[D].西安电子科技大学，2019.

[2] Dong Y， Chawla N V， Swami A . metapath2vec: Scalable Representation Learning for Heterogeneous Networks[C]// the 23rd ACM SIGKDD International Conference. ACM, 2017.

[3] 涂存超, 杨成, 刘知远,等. 网络表示学习综述[J]. 中国科学:信息科学, 2017, 047(008):P.980-996.

[4] Zhu Q , Luo J , Ding P , et al. GRTR: Drug-Disease Association Prediction Based on Graph Regularized Transductive Regression on Heterogeneous Network[J].*Lecture Notes in Computer Science*, 2018,vol 10847.Springer, Cham.

[5] Li H , Wang Y , Zhen Z Z , et al. Identifying Microbe-Disease Association Based on a Novel Back-Propagation Neural Network Model[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, PP(99):1-1.

[6] Albert István, Albert Réka. Conserved network motifs allow protein-protein interaction prediction[J]. *Bioinformatics.*2012,(18):3346-3352.

[7] Jiajie, Peng, Weiwei, et al. A learning-based framework for miRNA-disease association identification using neural networks.[J]. *Bioinformatics*, 2019,35(21):4364-4371.

[8] 林佳伟. 基于网络表征学习的生物网络节点分类[D]. 厦门大学，2019.

[9] Le D H， Pham V H . Random walk with restart: A powerful network propagation algorithm in Bioinformatics field[C]// The 5th International Conference on Computational Social Networks (CSoNet 2016). 2016..

[10] Grover A , Leskovec J . node2vec: Scalable Feature Learning for Networks[C]// the 22nd ACM SIGKDD International Conference. ACM, 2016:855-864.

[11] Wang D , Cui P , Zhu W . Structural Deep Network Embedding[C]//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2016:1225-1234.

[12] Xing, Chen, Yu-An, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases[J]. Bioinformatics, 2016, 33(5):733-739.

[13] J Luo, Long Y . NTSHMDA: Prediction of Human Microbe-Disease Association based on Random Walk by Integrating Network Topological Similarity[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2018:1-1.

[14] Long Y , Luo J . WMGHMDA: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network[J]. BMC Bioinformatics, 2019, 20.

[15] Chen X , Wang L , Jia Q , et al. Predicting miRNA-disease association based on inductive matrix completion[J]. Bioinformatics, 2018(24):4256-4265.

[16] Long Y , Wu M , Keong K C , et al. Predicting human microbe‐drug associations via graph convolutional network with conditional random field[J]. Bioinformatics, 2020(19):19.

[17] Ya-Zhou, Sun, De-Hong, et al. MDAD: A Special Resource for Microbe-Drug Associations.[J]. Frontiers in Cellular & Infection Microbiology, 2018.

[18] Akanksha R , Anamika T , Shivangi S , et al. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance[J]. Nucleic Acids Research, 2017(D1):D1.

[19] Kamneva O K . Genome composition and phylogeny of microbes predict their co-occurrence in the environment[J]. PLoS Computational Biology, 2017, 13(2):e1005366.

[20] Masahiro H , Nobuya T , Minoru K , et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses[J]. Nucleic Acids Research, 2010(suppl_2):W652.

[21] Ogata H , Goto S , Sato K , et al. KEGG: kyoto Encyclopedia of Genes and Genomes[J]. Nucleic Acids Research, 1999, 27(1):29-34.

[22] Shi C , Yu P S . Heterogeneous Information Network Analysis and Applications[J]. Springer International Publishing, 2017.

[23] Perozzi B , Al-Rfou R , Skiena S . DeepWalk: Online Learning of Social Representations[J]. ACM, 2014.

[24] Grover A , Leskovec J . node2vec: Scalable Feature Learning for Networks[C]// the 22nd ACM SIGKDD International Conference. ACM, 2016:855-864.

[25] Tang J， Qu M， Wang M , et al. LINE: Large-scale information network embedding[J]. International Conference on World Wide Web Www, 2015.

[26] Chen J， Hou H， Gao J , et al. RGCN: Recurrent Graph Convolutional Networks for Target-Dependent Sentiment Analysis[M]. 2019.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学（深圳）攻读学士学位期间，所提交的毕业设计（论文）《基于异构网络表征的微生物-药物关联预测研究》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：　　日期：　　年　　月　　日

致　谢

衷心感谢导师李君一教授在整个毕业设计的过程中对我的精心指导和耐心答疑，感谢她对我前期考研备考的支持和理解。她在论文选题，构思和论文最终定稿的过程中给了我巨大的帮助，她对自己学生的关心和帮助让我印象深刻，同时也倍感欣喜能在这样优秀的老师的带领下完成毕设任务，她的言传身教将使我终身受益。

感谢实验室的张晓帅学长，他为我的毕设具体实施给予了巨大的帮助，在我迷茫无措的时候给了我非常好的建议，让我的毕设最终有了比较令人满意的结果，他的学术水平同样令我敬佩，是我应该学习的榜样，同样也感谢实验室的其他学长学姐，感谢他们对我一个本科生的热情关心，让我很快融入实验室学习。

感谢我的舍友们，我们一起创造了和谐友好的寝室氛围，是他们让我拥有了充实快乐的大学生活，一起考研互相鼓励互相监督互相较劲的日子这辈子都难以忘却，虽然今后必定会各奔东西，但希望我们都能迈向光明的未来。

感谢计算机学院 17 级的同窗们，很荣幸也很幸运能和这么多优秀的学子在同一个屋檐下学习和竞争。

感谢哈尔滨工业大学给了我学习的环境，了解学术领域的机会，规格严格、功夫到家的校训必将铭刻在心，我以我是一名哈工大人为荣。

由衷地感谢我的父母，感谢他们从小到大对我的悉心培养、无私给予、无条件的支持和信任，感谢他们让我能够如此幸福地成长，感谢他们给了我在学业道路上走得更远的动力，生养之恩一生难报，我永远爱他们。

最后，感谢在百忙之中评阅论文和参与答辩的各位老师，您辛苦了。