河南工學院
HENAN INSTITUTE OF TECHNOLOGY

毕 业 设 计

药物与药物相互作用预测网络设计

Design of Drug-Drug Interaction Prediction Network

学院名称：计算机科学与技术学院

专　　业：物联网工程

班　　级：物联网 172

学生姓名：陈帅帅

学　　号：1710331243

指导教师姓名：王鲜芳

指导教师职称：教授

2021 年 05 月 28 日

毕业设计（论文）原创性声明和使用授权说明

原创性声明

本人郑重承诺：所呈交的毕业设计（论文），是我个人在指导教师的指导下进行的研究工作及取得的成果。尽我所知，除文中特别加以标注和致谢的地方外，不包含其他人或组织已经发表或公布过的研究成果，也不包含我为获得河南工学院及其它教育机构的学位或学历而使用过的材料。对本研究提供过帮助和做出过贡献的个人或集体，均已在文中作了明确的说明并表示了谢意。

作　者 签 名: _____日　期: _____

指导教师签名: _____ 日　期: _____

使用授权说明

本人完全了解河南工学院关于收集、保存、使用毕业设计（论文）的规定，即：按照学校要求提交毕业设计（论文）的印刷本和电子版本；学校有权保存毕业设计（论文）的印刷本和电子版，并提供目录检索与阅览服务；学校可以采用影印、缩印、数字化或其它复制手段保存论文；在不以赢利为目的前提下，学校可以公布论文的部分或全部内容。

作者签名: _____ 日　期: _____

药物与药物相互作用预测网络设计

摘要：药物在疾病治疗过程中起着非常重要的作用，对于同一疾病，医生往往依据经验会使用多种药物，然而这些药物具有不同的生化结构，不同药物之间往往会有相辅或者相克两种作用，相辅作用有利于疾病的治疗，而相克会对病人病情造成雪上加霜的结果，因此对药物和药物的相互作用研究尤为重要。传统的方法是依据药理生化实验，但这种方法往往需要大量的时间以及事件作为验证，耗时费力。随着计算机技术和人工智能技术的发展，利用数据挖掘技术研究药物与药物之间相互作用关系，能够有效缩短药物筛选时间，对医生药物选择、药物研发、保障人们生命健康具有十分重要的意义。

本设计来源是国家自然科学基金项目：“基于多源异构数据的人类冠状病毒药物协同重定位关键技术研究”（项目编号：62072157），主要基于知识图谱的方法对药物-药物相互作用进行分类预测。从 DrugBank 上下载原始数据，利用 Cytoscape 软件制作成药物-靶标关联网络的知识图谱，作为 GraphSAGE 模型的输入；使用Agg^{sum}加和算子汇聚邻域深度分别为 1、2、3 层药物邻居节点的特征，将其拼接得出药物对的特征向量；将药物对的特征向量作为多层感知机模型的输入，经过四层网络的学习计算出药物对之间相互作用得分来判断药物-药物之间是否存在相互作用；经过训练建立药物-药物互作用模型，并通过测试验证模型的有效性。实验结果表明，模型选取邻域深度为 1 层时表现最佳，模型的可靠性达到 0.94，准确度指标为 0.88，模型扩展性指标为 0.92，实现了较为精准的药物间相互作用判断，达到预期设计目标。

关键词：药物-药物；药物-靶标；知识图谱；深度学习

Design of Drug-Drug Interaction Prediction Network

Abstract： Drugs in the treatment of disease plays a very important role. For the same disease, doctors often will use a variety of drugs on the basis of experience, but these drugs have different structure that different drugs tend to have matched or mismatched between two kinds of function that supplemented will get good effect for the treatment of diseases and phase grams will cause worse results to patients' condition. Therefore, the study of drug and drug interaction is particularly important. The traditional method which is based on pharmacological and biochemical experiments often needs lots of time and events to verify and make the verification time-consuming and laborious. With the development of computer technology and artificial intelligence technology, using data mining technology to study the interaction relationship between drugs can effectively shorten the time of drug screening and also is helpful in doctors' drug selection, drug research and the protection of people's life and health.

This design was funded by the National Natural Science Foundation of China: "Research on Key Technologies for Collaborative Drug Relocation of Human Coronavirus Based on Multi-source Hexogeneous Data" (Project No. : 62072157), which mainly used the method of knowledge graph to classification and prediction of drug-drug interactions. The knowledge map of drug-target association network was made by downloading the original data from DrugBank and using Cytoscape software as the input of GraphSage model. Agg^sum operator was used to aggregate the features of drug neighbor nodes with neighborhood depth of 1, 2 and 3 layers, respectively, and the eigenvectors of drug pairs were obtained by splicing them together. The eigenvector of drug pair was used as the input of multi-layer perceptron model, and the score of drug pair interaction was calculated by learning four-layer network to judge whether there was interaction between drug and drug. The drug - drug interaction model was established by training, and the validity of the model was verified by testing. The experimental results show that the model performs best when the depth of the neighborhood is 1 layer. The reliability, accuracy and scalability of the model are 0.94, 0.88 and 0.92 respectively, which achieve a more accurate judgment of drug interactions and achieve the expected design goal.

Keywords: Drugs-Drugs; Drug-Target; Mapping Knowledge Domain; Deep Learning

目 录

第 1 章 [1]绪论

1.1 研究背景

当下，世界气候变化剧烈，尘封于极地并随后出现了剧烈变异的远古病毒，横空出世的新型病毒等在不同地区都有出现不同的变异现象，这种情况对于人类造成了相当的威胁，同时，在针对各种疾病的药物作用实验中往往会使用多种药物作用于病灶，由于各种药物之间化学成分的不同，经常会出现多种药物之间反应出现药物活性增强、减弱，甚至出现毒害病体的情况。针对这种情况，药物与药物相互作用(Drug-Drug-Interaction, DDI)的重要性愈发凸显，建立一个药物与药物相互作用预测网络具有一定的理论和实践意义。

1.2 研究意义

当前研究在国内主要是对于两种药物之间的相互作用关系以及体外药物作用，对于多种药物之间的网络图预测研究较少。所以本设计对此展开研究，通过对基于药物网络图的药物相互作用进行设计，首先，对相关医疗专家筛选最优药物组合具有借鉴意义；其次，可以给予医生用药建议，降低患者因药物反应受到二次伤害的可能性。最后，可以加速针对未知症状的药物的研发与投产，减少不良反应的发生。

1.3 国内研究现状

国家药品监督管理局药品评审中心于 2020 年 9 月 11 日发布了 Drug-Drug-Interaction(DDI)评审征求意见稿[9]，依据于 2021 年 1 月 25 日发布试行版的药物相互作用指导原则可以发现国内对于 DDI 研究主要包括体外药物作用研究和药物临床试验 DDI 预测模型两大类。医学领域主要涉及包括代谢酶介导药物相互作用，转运体介导药物相互作用研究，代谢产物之间的相互作用三个方面的体外药物作用研究大类。

在代谢酶介导药物相互作用研究上，主要研究代谢酶负责药物的代谢清除之后的体内药物浓度方面，多涉及 CYP450 酶系，该酶系具有可抑制和可诱导性，底物代谢速率

[1] 本设计来源于国家自然科学基金项目：“基于多源异构数据的人类冠状病毒药物协同重定位关键技术研究”（项目编号：62072157）,2021.1-2024.12

易被外界第三方影响导致代谢速率的变化从而引起底物在体内的药动学变化，和化学成分类似酶抑制剂或诱导剂的药物产生 DDI 反应。该项研究从 1958 年开始至今一直有新成员的出现以及新的生化特性发现，仍具有高度的研究意义。

在转运体介导药物相互作用研究方面，由于转运体参与药物的吸收，分布与排出，在长期的研究过程中基本认为药物体内过程和药效反应的产生大多数和药物的透膜有着密不可分的联系，而药物透膜方式中就包含了载体介导转运，并由此开始了对转运体的分类、功能、底物上开展了多种多样的研究，当下研究最多的是 P-gp 转运体，P-gp 转运体即 P 糖蛋白，在正常人体组织内广泛分布，不同组织内的 P-gp 转运体功能特性存在差异，如在胃肠道、肝脏组织内有降低底物吸收、降低生物利用度的生化特性，而在肾脏、肾上腺组织内则增加肾清除，同时底物也极易影响其生化特性，故主要研究人体摄入有着诱导剂性质的药物作用时发生药物相互作用。

代谢产物之间的相互作用，此项研究类别较多，常见研究为微生物与代谢产物之间相互作用，主要研究思想认为机体将代谢和免疫作为生存基本需求，两者长期协同存在已经构建成为机体稳态调节的核心机制。此研究美国于 2019 年发布了指导原则，已有较为成熟的研究结果，国内尚处于大方面的指导，缺乏具体指导原则，仍处于尚不成熟阶段。

随着 2016 年谷歌研发的围棋机器人阿尔法狗击败韩国李世石，在世界范围内引发巨大反响，机器学习开始广泛开展到各个方面，计算机领域也于此时开始迈向医学领域，随着几年的发展，逐渐发展出成熟的药物临床试验 DDI 预测模型，即基于生理药物代谢动力学(physiologically based pharmacokinetic，PBPK)模型的 DDI 预测模型。

基于生理药物代谢动力学(physiologically based pharmacokinetic，PBPK)模型的 DDI 预测。PBPK 即生理药代动力学，该模型给出药物机体内的浓度与时间关系，输入参数后动态反应药物在机体内吸收、分布、代谢及排泄的影响，可以依据外部用药浓度变化来预测各个不同组织内的药物浓度，给出相关专家用药浓度的参考建议，但不足点在于忽略了药物进入机体内之后对于机体的影响，对于不同药物之间是否存在生化反应没有给出判别结果。

实现药物与药物相互作用的预测网络，首先需要了解药物的作用机理，经过网上查询以及对相关医学专业同学的咨询后得到一个初步结论：传统西医药理学将药物-靶标-疾病作为判别一个药物有效性的理论基础，本设计也是基于所得初步结论开展。其次，需要了解药物的属性特征，通过对 DrugBank 等相关网站的数据检索发现，药物一般包括靶标，酶，化学式，人血清蛋白，基因五种关键属性，酶由于存在人体内生化反应，

化学式存在异构，手性因素，人血清蛋白以及基因涉及的相关医学知识已经超出可以快速了解范围，故采用靶标作为 key 值进行药物间相互作用的判别依据。然后，将数据整理成为网络图，并初步处理出相应的药物间相互作用数据矩阵作为预测模型的训练集。最后，用训练后的模型进行正确率判断以验证模型的可靠性。

1.4 本设计结构安排

全文内容分为五章，具体安排如下：

第一章主要介绍了本课题的研究背景，针对目前药物之间相互作用预测存在的问题及国内外研究现状提出本设计的主要研究内容。

第二章阐述了对于研究药物之间相互作用有关的基础理论研究。

第三章对原始数据集进行处理。将原始数据集整理出知识图谱，让人类知识和机器学习产生更高的结合。

第四章构建模型。

第五章对模型进行测试。

第六章进行总结。

第 2 章 药物-药物相互作用预测相关理论

2.1 相关软件及开发工具

2.1.1 Python 和扩展工具

Python 是典型的程序性语言，语法定义相对较少，对逻辑关联要求较低，可以更关注于执行内容，在数据挖掘和信息处理上应用较多。本设计集成开发环境使用的是 PyCharm；Python 版本为 3.6，使用 NumPy 对数据进行科学计算，使用 CUDA 进行 GPU 运算，使用 Keras 和 TensorFlow 进行深度学习模型的构建。主要工具包对应版本信息如表 2.1 所示：

表 2.1 工具包版本

工具名称	使用版本
CUDA	11.1
Keras	2.3.1
NumPy	1.2.3
TensorFlow	1.5.0
Python	3.6

2.1.2 Cytoscape

Cytoscape 是一个基于 Java 的开源项目，Cytoscape 官网图如图 2.1 所示。该软件主要功能是进行网络的可视化构建，支持多种网格描述形式，同时支持大量的数据分析扩展工具，极大简化了本设计关联网络知识图谱的构建。

图 2.1 Cytoscape 官网图

2.2 随机游走算法

随机游走[7]，就是粒子的随机运动，将粒子每一时刻的位置变化刻画为随机变量，依据该思想可以给定粒子的运动模型规则，粒子从一个顶点开始遍历一张图，随着粒子的不断运动扩散，会得到一个由该粒子运动轨迹形成的概率分布，这个概率分布描绘了每一个顶点被访问的概率，将此概率分布作为下一次游走的输入并经过多次训练后会得到一个比较稳定的概率分布。

设$f(x)$是一个含有 n 个变量的多元函数, $x = (x1, x2, \ldots, xn)$为 n 维向量。给定初始迭代点x, 控制精度 ϵ, 初次行走步长λ，手动设置迭代控制次数N，k为当前迭代次数，置$k = 1$。当$k < N$时，随机生成一个$(-1,1)$之间的 n 维向量$\mu = (\mu_1, \mu_2, \cdots, \mu_n), (-1 < \mu_n < 1, i = 1,2, \cdots, n)$, 并将其标准化得到

$$\mu' = \frac{\mu}{\sqrt{\sum_{i=1}^{n} \cdot \mu_i^2}} \tag{2-1}$$

令$x1 = x + \lambda u'$, 进行第一步游走。当$f(x1) < f(x)$, 表明找到了一个比较合适的点，此时将k重置为 1，进行再次迭代，否则$k = k + 1$, 继续重复进行游走。当连续N次都找不到更好的点，那么此时将认定最优解位于以当前最优解的点为中心，当前步长为半径的多维空间内。此时，如果$\lambda < \epsilon$，则结束算法；否则令$\lambda = \lambda/2$，开始新的一轮游走。

改进的随机游走算法的将原来产生一个随机向量μ进行优化，将固定值随机向量μ替换为产生 n 个随机向量$u_1, u_1, \cdots, u_n$，并将随机向量进行标准化得到$u_1', u_2', \cdots, u_n'$，令

$$xi = x + \lambda \mu_i', \tag{2-2}$$

使$min\{x1, x2, \cdots, xn\}$替换原来的$x1 = x + \lambda u'$，改进之后将降低对于初始训练参数的依赖并提高寻优能力降低寻优时间。

2.3 GraphaSAGE

GraphaSAGE[7]由图卷积网络（Graph Convolutional Network，GCN）发展而来，对 GCN 模型进行优化。

GCN，顾名思义是作用于图结构的神经网络，多运用于给定多个原始结构的训练集后预测一个具有参考意义的结构图，例如预测当下众人非常关心的治疗新冠肺炎的有效药物的化学结构，或者运用于给定多个有给定结论并具有多项特征的实体预测出一个已知特征的实体具有何种结论，例如预测某人是否是罪犯的情景，最后就是预测实体间的关系，本设计即属于该思想下的应用。

当下存在多种图神经网络的变种，基本思想都是由基本卷积网络发展而来，即给定一个卷积核对一个向量化的数据做卷积，对选中的区域向量做内积计算出当前卷积层的特征，然后移动卷积核重复上述操作得到下一层的特征。进行多次的重复训练之后可以训练出一个较为理想的模型。GCN 模型中核心在于层间的传播，即

$$F^{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}F^{l}\omega^{l}\right) \tag{2-3}$$

公式中F^{l}是节点特征，$\tilde{A} = A + I$，A是节点的邻接矩阵，I是单位阵，这种方式解决了提取特征时不包括节点自身特征的问题。$\tilde{D}$是对选定节点的邻居深度，即度矩阵，ω^{l}是权重矩阵，可经过模型训练获取，σ是激活函数。

通过 GCN 模型可以发现节点特征只和上一层的邻居节点特征有关，GCN 模型中需要对整张图进行训练，往往会出现超级节点导致计算的时间成本指数级增加，GraphSAGE 可以限定邻居深度，通过对范围内邻居节点的采样使得训练数据保持相对低纬度的状态下，有效提升了计算效率并节省了大量的空间。采样邻居图如图 2.2 所示：

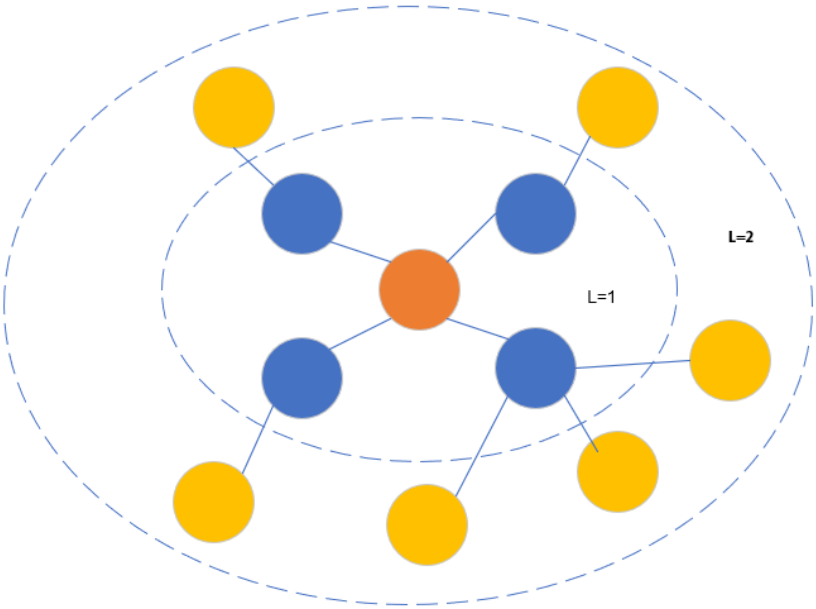

图 2.2 采样邻居图

GraphSAGE 的关键在于对邻居的聚合操作，需要保证在聚合数量非固定，可以随着节点变化而变化，同时需要可导便于对模型进行优化，常规采用加和算子，本设计也是采用该思想进行聚合。

$$Agg^{sum} = \sigma\ (\sum \omega H_j + b) \tag{2-4}$$

H_j是第 j 层的节点特征，b 为偏移量。

对图 2.3 运算示例图里的s^i节点做卷积以提取其特征

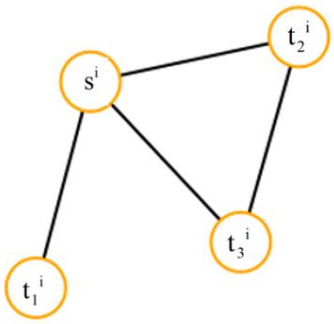

图 2.3 运算示例图

将彼此间没有精密联系的稀疏向量进行降维处理，使其成为只有语义联系的类似索引信息的低维数据，将每个具有高维含义的信息用一个向量表示，此时我们也得到了第一层隐藏层的节点特征。当我们想要得到第二层隐藏层特征时，我们只需要聚合第一层隐藏层中s^i节点的邻居节点和第一代图模型中给定的s^i节点信息即可，即

$$s^2 = \omega\sigma(t_1^1 + t_2^1 + t_3^1 + s^0) \tag{2-5}$$

其中s^2为第二层隐藏层中s^i节点特征，t_1^1, t_2^1, t_3^1分别为第一层隐藏层中t_1^i，t_2^i，t_3^i节点特征，s^0为最初的s^i节点数据，i代表第几层隐藏层，ω为权重，给定后需要训练优化以得到最优模型。

2.4 多层感知机

多层感知机（Multi-Layer Perceptron , MLP）[7]由神经元模型发展而来，基本神经元模型图如图 2.4 所示：

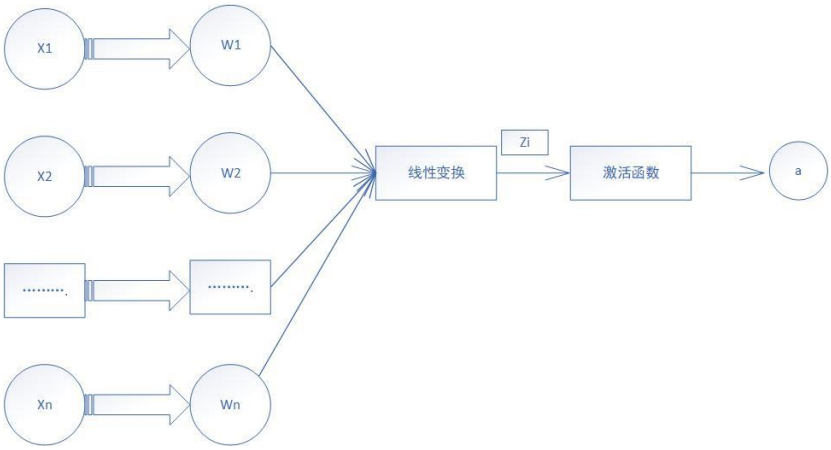

图 2.4 神经元模型图

有输入层、线性组合和非线性激活函数组成，上层向下层输出信息时都会经过一个激活函数，有函数判断是否该输出次类型信号，即

$$a = \begin{cases} 1 & if\ wx + b > 0 \\ -1 & if\ wx + b \leq 0 \end{cases} \quad (2\text{-}6)$$

x 是输入向量，w 是输入向量的权重矩阵，b 是偏执向量。

神经元处理能力有限，无法应对复杂的学习，也没有办法输出更多的结果，仅能完成二分类任务。基于人类大脑多神经元的思想，多层感知机应运而生，多层感知机模型图如图 2.5 所示。

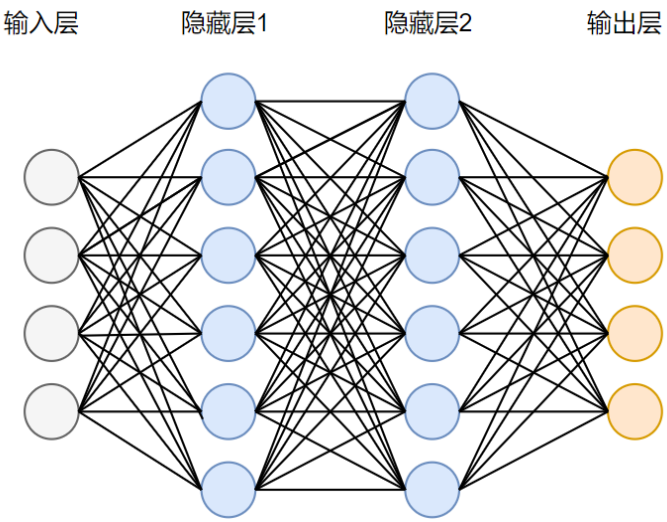

图 2.5 多层感知机模型图

本层的神经元与下层的全部神经元都建立连接，这类连接可称为全连接，这样设计的好处是，每一层都可以用上一层的输出进行线性组合，对输出结果不断降维优化，并由激活函数和权重的积表示。常用的激活函数有 ReLU 函数和 Sigmoid 函数。

ReLU 函数定义当$x \geq 0$时，函数输出值不变，当$x < 0$时，函数输出值为 0，即

$$f(x) = max(0, x) \quad (2\text{-}7)$$

这种方式可以有效的缓解稀疏矩阵带来的梯度消失问题，同时也导致了输出值为负的神经元功能性死亡，不存在任何价值。

Sigmoid 函数即

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2\text{-}8)$$

将输入都转化为 0 到 1 之间的值，可以实现对结果概率预测。

第 3 章 药物-靶标关联网络构建

3.1 获取药物属性数据

目前主流与药物相关的医药相关数据库有 pharmmapper、STITCH、TCMSP、DrugBank 这四种数据库，其中 DrugBank 数据库整合了生物信息学和化学信息学资源，并提供详细的药物数据与药物-靶标信息及其机制的全面分子信息，包括药物化学、药理学、药代动力学、ADME 及其相互作用信息。目前 DrugBank 包含了 10971 种药物和 4900 种蛋白靶标的信息。这些药物包括 2391 种 FDA 批准的小分子药物，934 种批准的生物技术药物，109 种营养药物和 5090 多种实验药物，最重要的是该数据库中的所有药物靶点信息都得到过实验验证，具有相对最高的数据可靠性，于是本设计从 DrugBank(https://go.drugbank.com/)上随机下载了 105 种包括 target、pathway、name、ID，hsa、enzyme 属性的药物信息，由这 105 种药物得出 511 种药物-靶标信息对作为原始输入集。

3.2 建立药物-靶标关联网络

药物与药物之间产生相互作用有多种因素导致，从本设计选定的角度出发，选择将药物中的靶标属性作为各种药物之间相互作用的关键值。使用下载的 105 种药物的相关数据构建一个简易的药物-靶标关联网络作为预测模型的输入。所采用的部分数据集如表 3.1 部分药物-靶标作用关系表所示：

表 3.1 部分药物-靶标作用关系表

Drug	靶标 ID	Interact
Glucosamine	P14780	1
Glucosamine	Q00653	1
Caffeine	P30542	1
Caffeine	P29274	1
Methotrimeprazine	P14416	1
Fluticasonepropionate	P06401	1
Fluticasonepropionate	P47712	1
Fluticasonepropionate	P08235	1

将药物编号、名称及靶标数据导入 Cytoscape 软件，将药物编号作为源数据，靶标作为目标数据建立一个原始的网络结构，可以看到原始数据集相对杂乱且不具备进行建模分析训练的理论基础。原始药物-靶标关联网络图如图 3.1 所示：

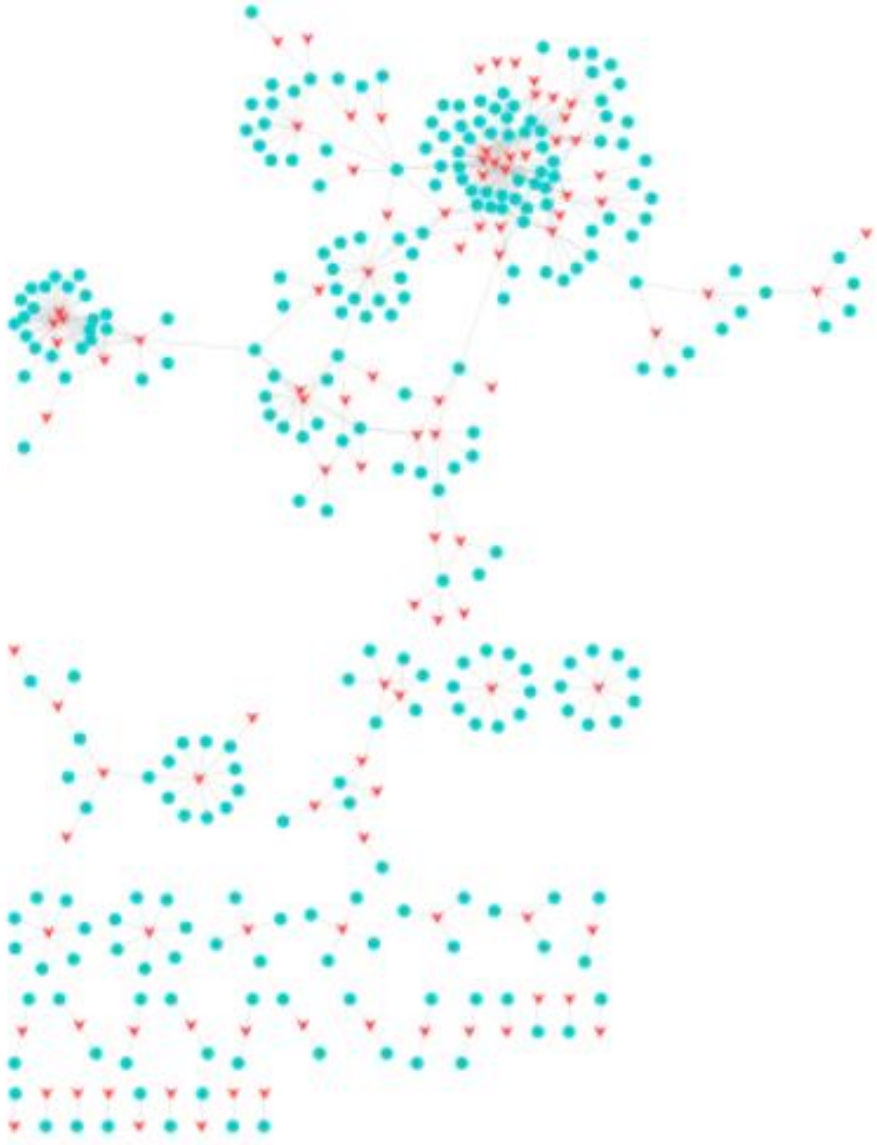

图 3.1：原始药物-靶标关联网络图

（红色倒三角代表药物，浅绿色为靶标数据）

由图 3.1 可以粗略的看到下载的原始数据构建的网络图存在没有关联的药物，以及存在大量对于药物间关联无用的靶标数据。对数据进行深度分析发现包括靶标在内大部分节点的邻居深度都不超过 3，如图 3.2：原始药物深度所示具有比较直观的相互作用效果，邻居节点也多分布在深度在 3 以下的范围内，如图 3.3：邻居节点-深度关系图所示，因此可以认为在深度范围在 3 以内的药物之间存在相互作用

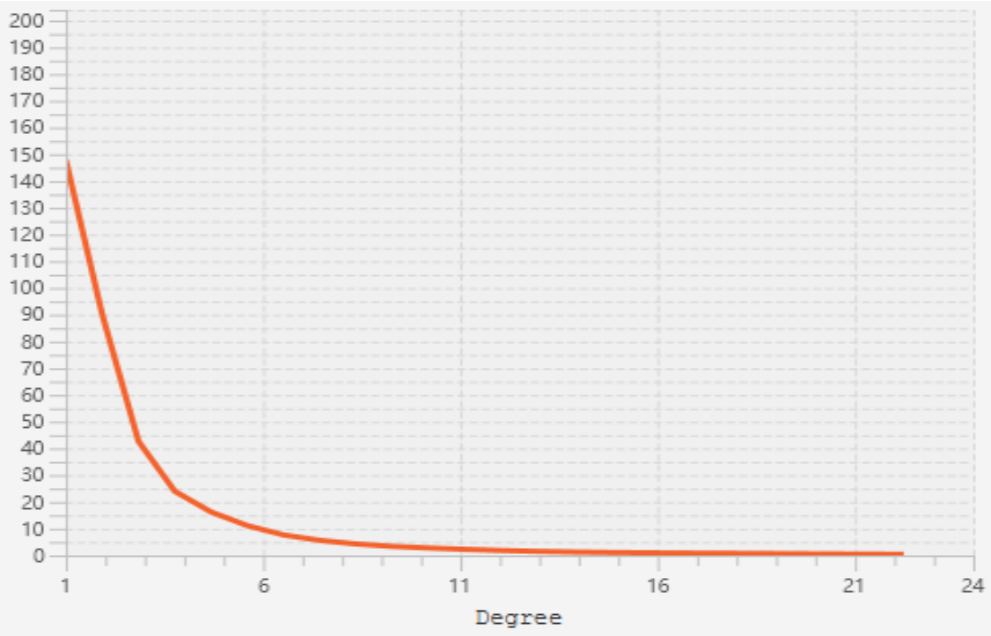

图 3.2：原始药物深度

（药物节点深度的分布，X 轴代表与药物有关联的邻居节点深度，Y 轴代表药物-靶标在不同深度的节点数量）

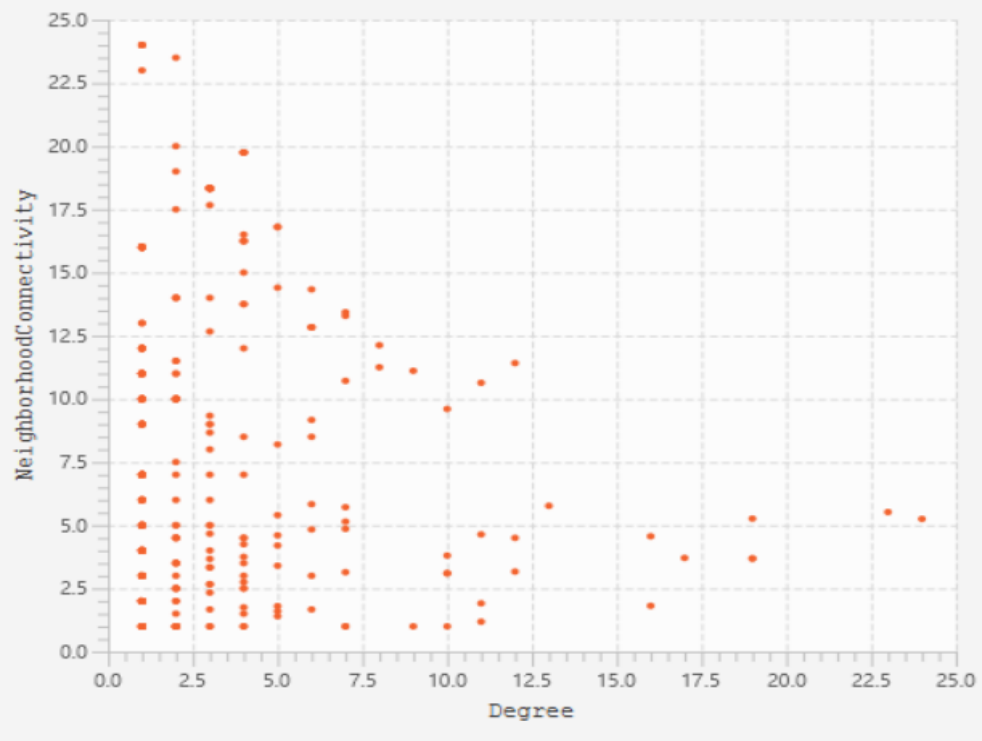

图 3.3：邻居节点-深度关系图

（X 轴所有节点包括药物和靶标互相之间的深度关系，Y 轴代表位于不同深度的各节点的邻居关系数量）

经过分析处理后的部分药物-靶标网络药物节点数据如表 3.2 部分药物节点相关数据表所示

表 3.2 部分药物节点相关数据表

ID	Name	degree	NeighborhoodConnectivity	SelfLoops
DB01296	Glucosamine	5	1.6	0
DB01195	Flecainide	3	3.333333	0
DB00201	Caffeine	11	1.181818	0
DB01403	Methotrimeprazine	19	5.263158	0
DB00588	Fluticasonepropionate	4	2.75	0
DB00915	Amantadine	3	5	0
DB00461	Nabumetone	2	3.5	0
DB01223	Aminophylline	4	1.75	0
DB00648	Mitotane	5	3.4	0
DB00783	Estradiol	10	3.1	0
DB00933	Mesoridazine	2	10	0

在原始数据网络图中加入过滤器，筛选出邻居节点深度在 1 到 3 的药物节点，将筛选后的药物之间视为具有相互作用，并提取出包含三个连通片的子网，如图 3.4 药物-靶标子网 1，图 3.5 药物-靶标子网 2，图 3.6 药物-靶标子网 3 所示：

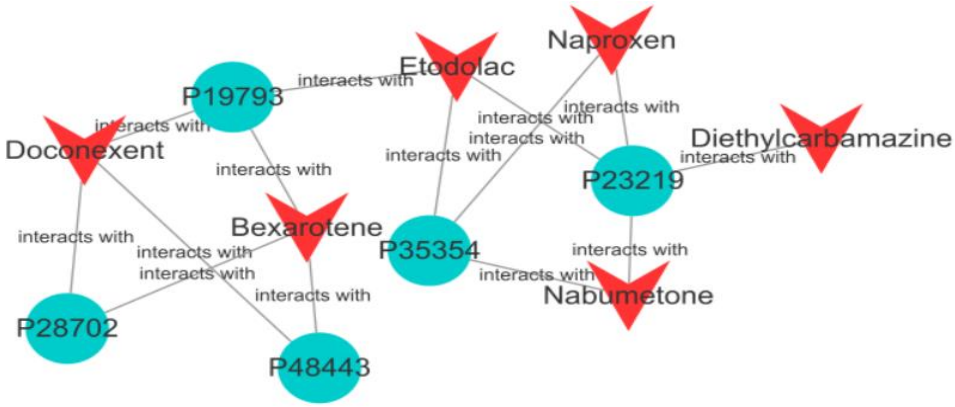

图 3.4 药物-靶标子网 1

（图中红色为药物数据，浅绿色为靶标 ID 数据）

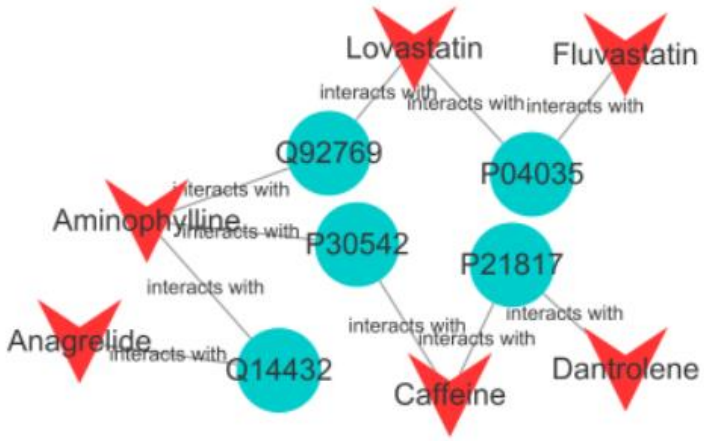

图 3.5 药物−靶标子网 2

（图中红色为药物数据，浅绿色为靶标 ID 数据）

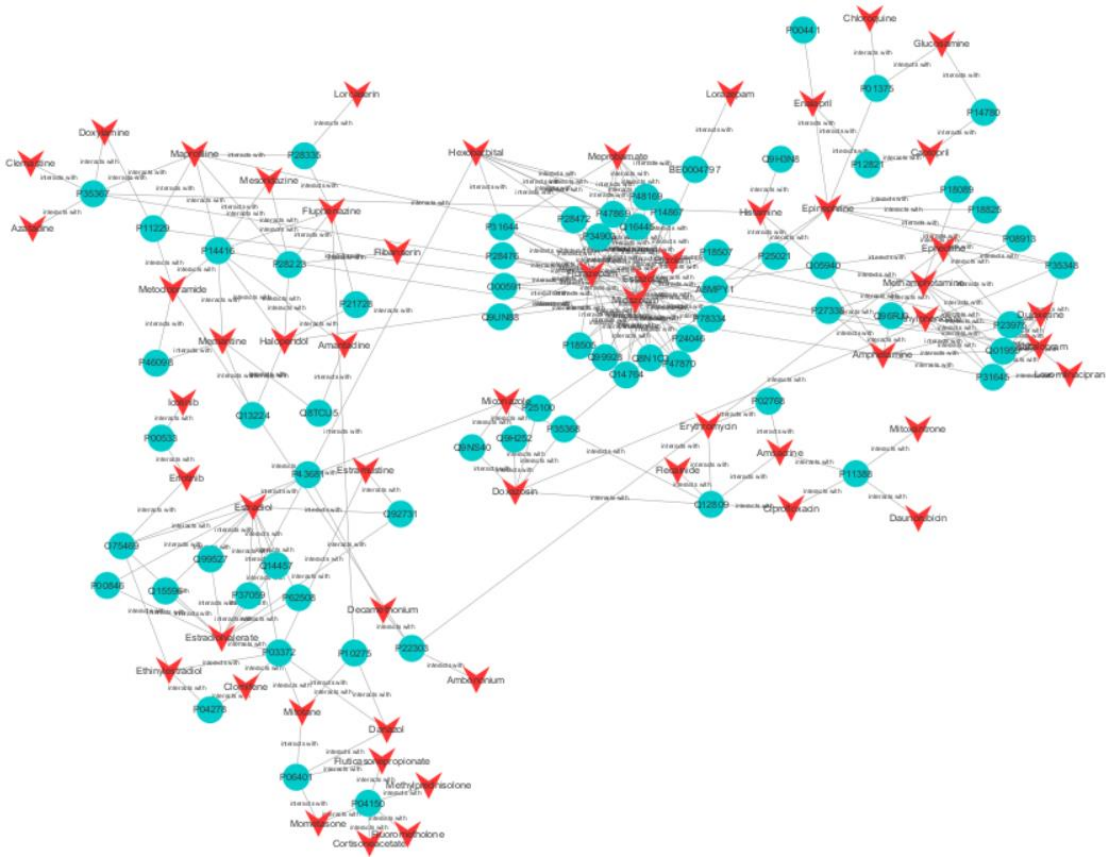

图 3.6 药物−靶标子网 3

（图中红色为药物数据，浅绿色为靶标 ID 数据）

3.3 小结

本章通过对原始数据进行处理，得到一张代表药物-靶标之间关系的知识图谱和数据集，这些结果是后面章节模型的输入，可以去除数据冗余，降低机器学习的时间成本，提高机器学习的准确率。

第 4 章 基于图的药物相互作用预测模型

4.1 模型介绍

计算机科学与医学领域的大规模交叉融合都可以追溯到深度学习阶段，当下对于医学相关内容的研究从深度学习上入手是相对科学成熟的做法。从深度学习发展史上看，传统深度学习算法曾广泛应用于欧氏空间数据的特征提取，而本设计中的药物-靶标关联图具有显著的邻居节点数量不固定的非欧氏空间的数据特征，目前随着图论的兴起以及图在当下数据预测中越发重要，用于处理图这种非欧氏空间数据的图神经网络得到了众多人士的研究探索并取得长足进步，已经成熟的处理方式有把非欧空间的图转换成欧式空间和找出一种可处理变长邻居结点的卷积核在图上抽取特征俩种，显然对于具有不定多样邻居节点的药物-靶标网络而言，第二种方法极为适合，使用其中发展成熟的 GraphSAGE 模型将邻居节点信息与药物节点自身特征聚合得出新的药物节点的特征向量，并将药物节点特征向量拼接成药物对特征向量。

采用拼接后的药物对特征向量作为输入，由于本设计的药物属性关系为二维关系，可以采用多层感知机模型进行预测，保证了个人设备的可执行性的基础上最大程度提升结果的可靠性，最后由得出的药物之间关联分数做一个分类器来判断药物之间是否存在相互作用，模型结构图如图 4.1 所示：

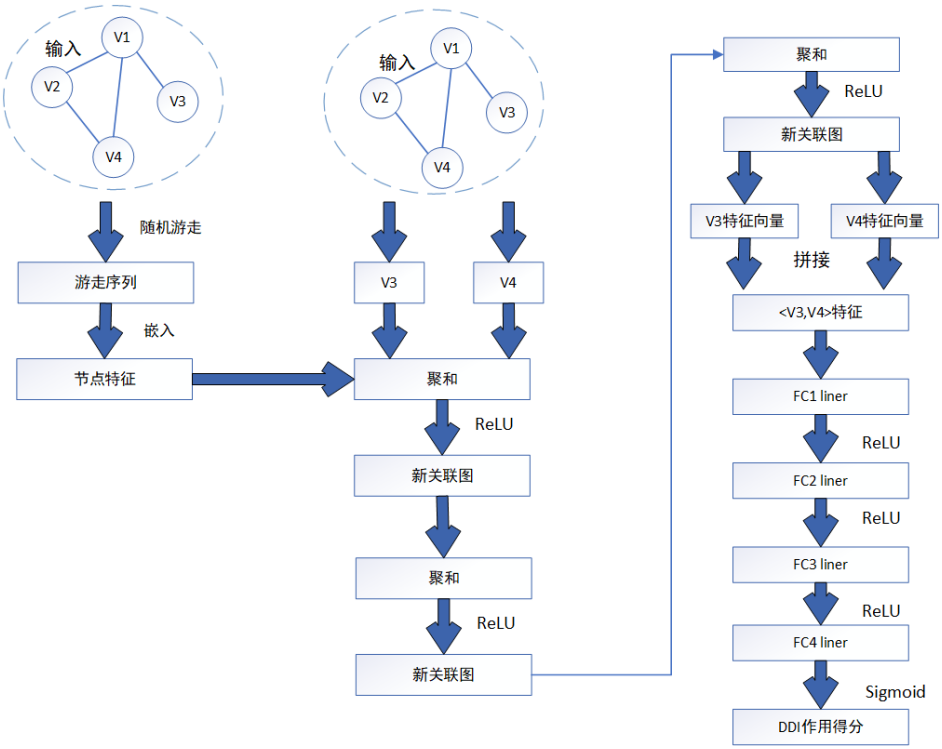

图 4.1 模型结构图

4.2 基于 GraphSAGE 的药物特征提取

GraphSAGE 是一图神经网络架构下的一个变种，对于本设计中的药物关联图数据来说，在第一次进行训练时节点特征是未知的，所以采用随机游走算法得到每个节点的游走序列和节点间的深度关系，Python 内置定义了 Embedding layer， 将相邻采样大小定义为 3，在原始数据集中提取出关键样本，将取样空间维度定义为 8，保证机器学习能够尽快高效计算出空间特征，将初始化器设置为正态分布（glorot_normal）模式，设置输出层正则函数值，随后调用其即可完成 embedding 化，计算出每个节点的特征值作为第一层的初始化特征。初始化特征获取流程图如图 4.2 所示:

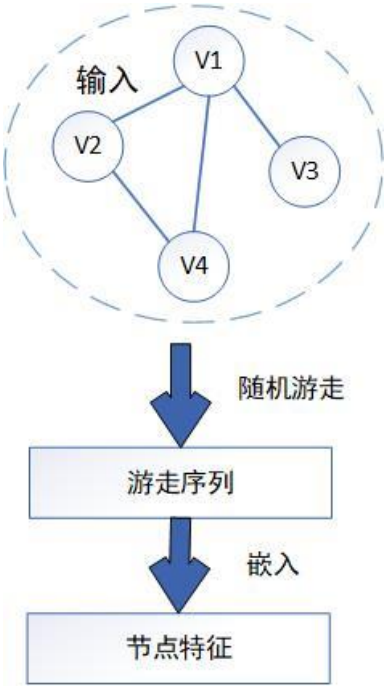

图 4.2 初始化特征获取流程图

在获取了第一层即初始化层的节点特征之后，对药物关联图进行三次聚类，选用加和算子Agg^{sum}作为聚类器，即可获取邻居节点深度为 3 的所有邻居节点特征和自身初始化特征作为最终药物节点的特征向量，每一层的权重 ω 由上一层的药物特征值与上一层的邻居节点特征值的和确定，每一层的偏移量 b 都由上一层的药物节点特征值确定。层间使用 $ReLU$ 函数作为激活函数。最后将获取之后的两种药物特征向量进行拼接，即将 v3 和 v4 的特征向量组合成[v3 v4]的药物对特征向量，邻居采样提取药物特征图如图 4.3 所示：

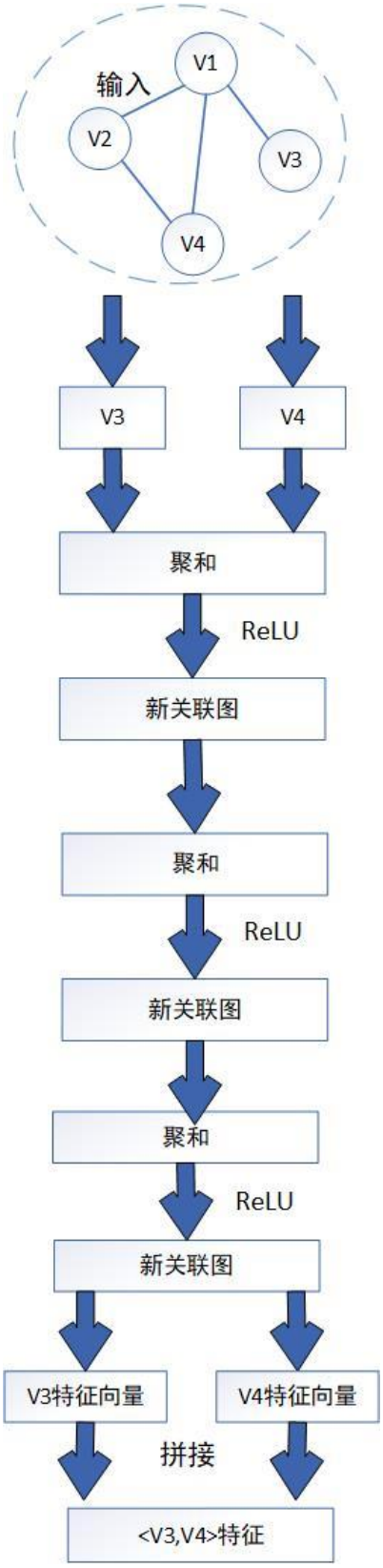

图 4.3 邻居采样提取药物特征图

4.3 基于多层感知机的药物-药物相互作用预测

多层感知机预测模型本质上是一个具有预测相互作用关系的二元分类模型。将拼接的药物-药物对特征向量作为预测模型的输入，采用四层多层感知机预测药物和药物之间的相互作用值，每层神经元的数量依次为 64，32，16，2。前三层使用 ReLU 函数作为激活函数，最后一层采用 Sigmoid 函数作为激活函数，输出药物-药物之间具有相互作用的概率值，多层感知预测流程图如图 4.4 所示：

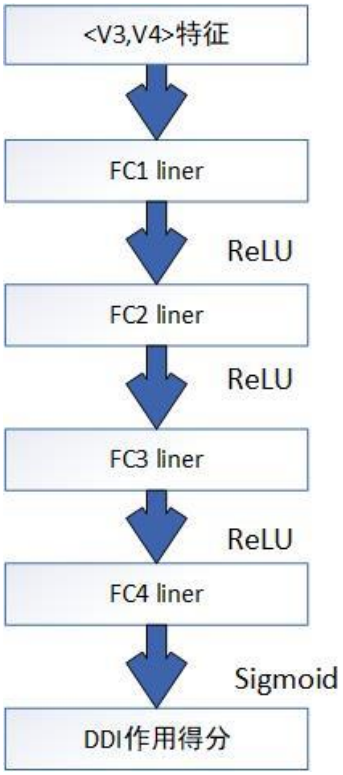

图 4.4 多层感知预测流程图

第 5 章 模型评估

5.1 药物对关系数据集构建

将多层感知机预测模型计算出的药物对之间的得分关系作为药物-药物关联网络的边属性，将数据整理为[药物 A，药物 B，相互作用]的三元结构 DDI 事件的数据集，药物 A 与药物 B 之间存在相互作用即将相互作用置为 1，没有相互作用即将相互作用置为 0。部分药物-药物作用关系表如表 5.1 所示：

表 5.1 部分药物-药物作用关系表

Drug1	Drug2	Interact
Abiraterone	Dexchlorpheniraminemaleate	1
Abiraterone	Dexfenfluramine	1
Abiraterone	Dexlansoprazole	1
Acebutolol	Amiloride	1
Acebutolol	Aminophylline	1
Acebutolol	Amiodarone	1
Abemaciclib	Aprepitant	1
Abemaciclib	Atomoxetine	1
Abemaciclib	Bortezomib	1

5.2 建立药物关联网络

将药物-药物之间的关系矩阵建立成知识图谱，便于直观分析药物-药物间的相互作用关系，使药物-药物作用关系表中的 Drug1 作为源数据，Drug2 作为目的数据，Interact 作为边属性，值为 1 则源数据和目的数据存在药物相互作用关系，值为 0 则为没有相互作用关系，部分药物关联网络图如图 5.1 所示：

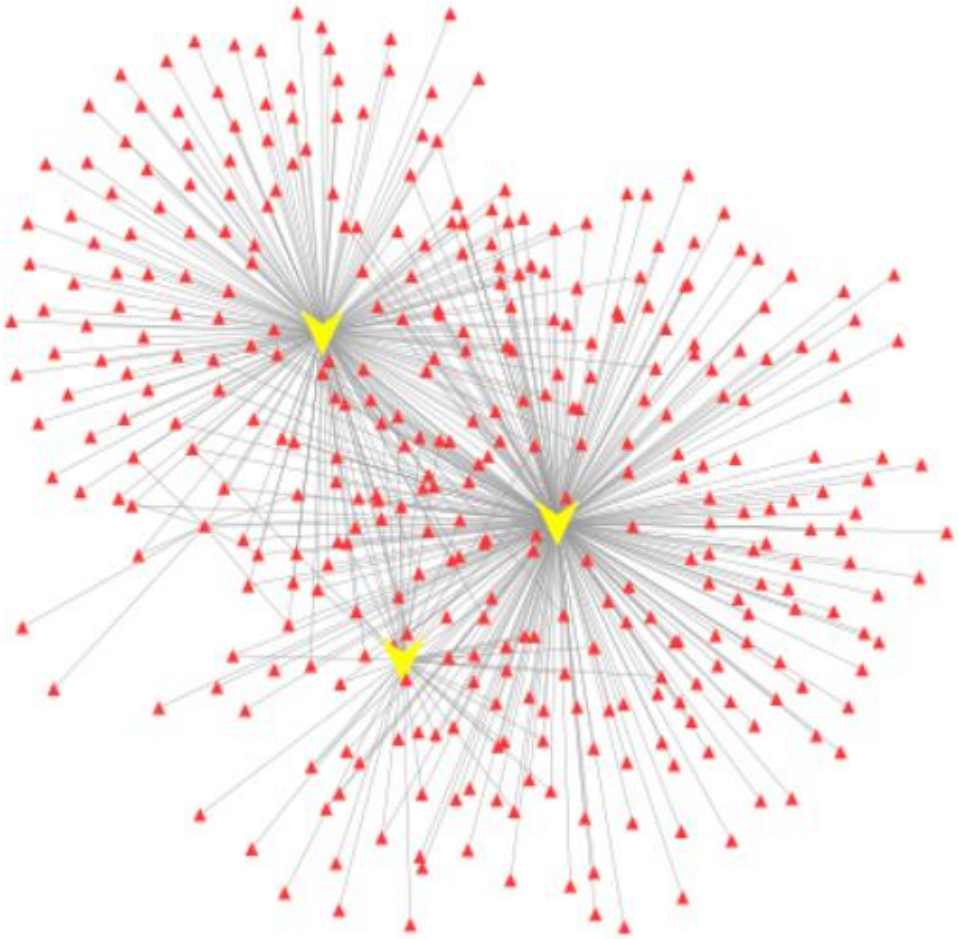

图 5.1 部分药物关联网络图

（其中黄色 V 型节点为 Drug1 类，红色三角形节点为 Drug2 类，边表示 Drug1 类与 Drug2 类存在相互作用关系）

图 5.1 中黄色 V 型节点自上而下代表 Acebutolol（醋丁洛尔）、Abiraterone（坦度酮罗）、Abemaciclib（玻玛西林）；Acebutolol（醋丁洛尔）与 Furosemide（呋塞米）、Amobarbital（异戊巴比妥）、Fentanyl（芬太尼）等药物具有相互作用；Abiraterone（坦度酮罗）与 Epoprostenol（依前列醇）、Dopamine（多巴胺）、Azelastine（盐酸氮卓斯汀）、Clevidipine（氯维地平）等药物具有相互作用；Abemaciclib（玻玛西林）与 Dabrafenib（达拉菲尼）、Dasatinib（达沙替尼）、Lorpiprazole（洛吡呱唑）、Midostaurin（米哚妥林）等药物具有相互作用。

5.3 评价指标

本设计采用准确率（accuracy，ACC），ROC-AUC Score，召回率和正确率的比值（AUPR）作为模型评价指标。

准确率是分类正确的样本数与样本总数之比

$$ACC = \frac{TP + TN}{ALL} \tag{5-1}$$

ROC-AUC Score 描述了正确预测样本在全部预测样本中的概率，即

$$ROC - AUC\ Score = \frac{TP(FP + TN)}{TP(FP + TN) + FP(TP + FN)} \tag{5-2}$$

召回率和正确率的比值（AUPR）用于评价模型在更大的数据集上的可执行能力，即

$$AUPR = \frac{TP}{TP + FN} \cdot \frac{ALL}{TP + TN} \tag{5-3}$$

TP 是预测存在关系成功的样本数量；*TN* 是预测不存在关系成功的样本数量；*FP* 是预测误判存在关系的样本数量；*FN* 是预测误判不存在关系的样本数量；*ALL*是全部样本数量。

5.4 测试结果

对模型进行 50 次训练之后得到一个最优化的评价模型，对测试集进行对比，采用 AUC 评价指标对模型优劣进行评估，得出正确的预测在错误的预测前面的概率，采用 ACC 评价指标得出正确的预测概率，采用 AUPR 指标，得出召回率和正确率的比值评估出模型是否具有评估大规模数据的价值。对邻居节点深度分别为 1、2、3 层下模型的测试，模型评估测试图如图 5.2 所示：

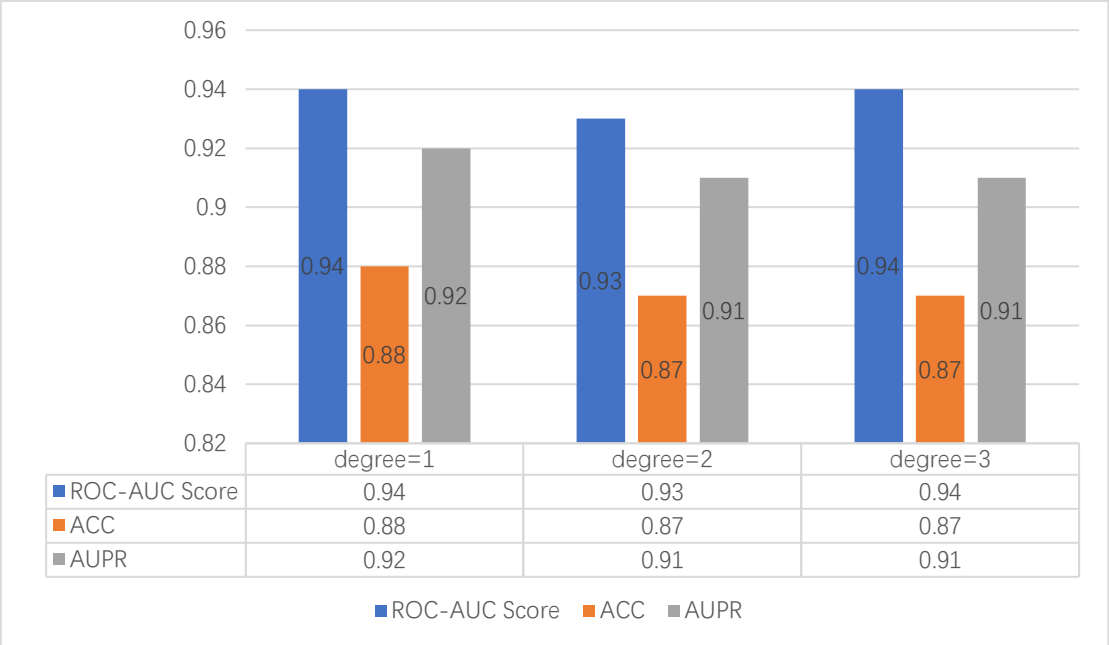

图 5.2 模型评估测试图

从测试结果中可以看到，该模型选取邻域深度为 1 层时表现最佳，预测的正确率为 0.88，模型可靠性指标为 0.94，大规模数据扩展指标为 0.92，模型整体性能一般，基本可以完成对药物-药物间相互作用的预测，在进行大规模数据预测时模型的可靠性也有一定的保证。

第 6 章 结 论

准确的预测出药物与药物之间是否存在相互作用是药物实验中的一个关键点。对于这个关键点进行更一步的研究可以构建出更多的药物与药物之间相互作用的关系网络，可以为缺乏医疗知识的普通大众提供用药指导、为新药的开发投产提供一个数据基础，尽量减少因用药不当对机体产生二次伤害的概率。本设计首先将看似毫无联系的各种药物数据整理成一个清晰明了的药物-靶标相互作用知识图谱和药物与靶标之间的关系矩阵，然后采用图神经网络对知识图谱中的药物进行特征提取，最后将提取的药物-药物拼接矩阵作为多层感知机的输入层数据，对药物与药物之间相互作用进行评估，模型进行预测之后与测试集数据对比后正确率在 0.85 左右，模型基本完成目的需求。

在进行知识图谱的构建中只选取了药物的靶标属性，忽略了药物通路等其他属性，导致药物的邻居节点特征值不具有显著代表性，同时使用靶标数据作为药物-药物相互作用的评判标准使得药物-药物之间只存在二维的距离关系，当加入更多属性作为评判标准时，更高维的空间结构就出现了，之间的结构关系是人类无法想象的，只能使用计算机进行模拟，本设计模型很遗憾未能实现高维度的模拟，与药物相互作用关系还存在一些差异，模拟出高维度中药物相互作用关系是下一步的主要研究方向。

参考文献

[1] 刘宁宁，琚生根，熊熙, 等. 基于胶囊网络的药物相互作用关系抽取方法[J]. 中文信息学报，2020, 034(001):80-86,96.

[2] 杨旭华，俞佳，& 金林波. (2017). 一种基于二阶局部群落和大度节点有利的预测网络连边的方法. CN106603313A.

[3] 叶林虎, 王 森, 徐永寿,等. 抗新型冠状病毒肺炎(COVID-19)药物潜在相互作用研究进展[J]. 现代药物与临床, 2020, 035(004):607-613.

[4] 刘文斌, 陈杰, 方刚,等. 基于药物互作网络的协同与拮抗预测研究[J]. 电子与信息学报, 2020, v.42(06):121-128.

[5] 陈杰. 基于复杂网络的药物互作预测研究[D].温州大学,2020.

[6] 马龙. 基于深度神经网络的药物关系挖掘方法研究[D].西北大学,2019.

[7] Ethen Alpaydin，《机器学习导论》[M]，机械工业出版社，2016

[8] Ali M A , Rizvi S , Syed B A . Trends in the market for antihypertensive drugs[J]. Nature Reviews Drug Discovery, 2017, 16(5):309-310.

[9] www.cde.org.cn Drug-Drug-Interaction(DDI)评审征求意见稿，2020.

[10]Ghosal A . Evaluation of the clearance mechanism of non-CYP-mediated drug metabolism and DDI as a victim drug - ScienceDirect[J]. Identification and Quantification of Drugs, Metabolites, Drug Metabolizing Enzymes, and Transporters (Second Edition), 2020:237-271.

[11]Liao M , Jaw-Tsai S , Beltman J , et al. Evaluation of in vitro absorption, distribution, metabolism, and excretion and assessment of drug-drug interaction of rucaparib, an orally potent poly(ADP-ribose) polymerase inhibitor[J]. Xenobiotica, 2020:1-34.

致　谢

经过了几个月的努力，我最后完成了论文的写作。从开始的论文选题到系统的实现，再到论文的完成，每都一步对我来说都是新的尝试和挑战，这也是我在大学期间独立完成的最大项目。这段时间我学到了很多新的知识，也有很多感受，从一无所知到独立思考的学习，查看相关的文献资料和书籍，心中模糊的概念也逐渐清晰，正是这些进步才成就了今天的我。

在本次的毕设设计过程中，首先感谢我的学校，给了我学习的机会和优质的平台，一个成功的大学生活离不开良好的学习环境。其次就是感谢我的毕业设计指导导师王鲜芳老师，老师从论文整体框架到细致知识点给予了分析，培养学生独立解决问题的思维，提出了很多宝贵的意见与推荐，提供了许多优质的论文和参考资料。正是老师高度的敬业精神和不厌其烦的指导，给我论文的写作带来了很大的帮助，这篇论文正是在老师的精心指导和大力支持下完成的。最后感谢所有授我以业的老师们和热心帮助我的同学们，没有这些年的积淀，我没有那么大的动力和信心完成这篇论文。感恩之余，诚恳请各位老师对我的论文多加批评指正，使我及时完善论文的不足之处。