

西安交通大学

# 毕业设计（论文）

题 目 基于统计学习的骨关节炎风险预测模型构建

生命学院 理科试验班(化学生物 H) 专业 81 班

学生姓名 郭骐瑞

学 号 2186113661

指导教师 郭燕

设计所在单位 西安交通大学

2022 年 6 月



## 摘要

骨关节炎是一种由遗传因素与环境因素共同作用所产生的关节退行性病变，目前尚无有效的治疗方法。全球范围内骨关节患者群体庞大且规模逐渐扩大，使得骨关节炎成为导致残疾与疼痛的主要因素之一。研究显示通过风险预测模型的骨关节的早期预防与诊断能显著改善患者预后，然而现有的骨关节炎风险预测模型存在着性能差，可解释性弱的缺点，远不能达到临床使用需求。本文因此提出并设计了一种以图神经网络为核心的基于个体基因型与表型信息的骨关节炎风险预测模型。

本文主要完成以下三个方面工作：一、本文从 UKBiobank 数据库获取了 13706 名个体的基因型与表型数据，并根据现有的全基因组关联研究与相关指标对该数据进行质控与预处理，用以模型的训练与测试。二、本文设计了以切比雪夫图神经网络为核心的结合基因型与表型信息的风险预测模型，该模型能够通过图估计器将输入无结构数据转化为图数据并进行图神经网络处理，并同时具备预测结果解释模块。三、本文对该风险预测模型的功能与性能加以测试评估。本文选择了一系列指标对模型的预测性能加以评估，并同包括多基因风险评分模型、常见机器学习算法在内的传统疾病风险预测模型相比较。证明本模型（AUC 0.74）较传统疾病风险预测模型（AUC 0.5）而言有着十分明显的性能改善。本文还通过对图估计器工作过程分析展现了本模型优秀的可解释性。

在模型中，本文创新性使用了基于变分期望最大化的图估计器结合图神经网络来处理无结构的基因型数据。研究结果证实该方法不仅给出了较好的风险预测准确度，还能通过挖掘基因型数据中潜藏的信息对疾病病因、分型等因素加以推断。本文的研究成果一方面为无结构数据在图神经网络中的处理提供了新方法，另一方面也为疾病风险预测模型的构建提供了新思路，对图神经网络与疾病风险预测模型的广泛应用具有积极意义。

**关键词：**骨关节炎；风险预测模型；图神经网络

## ABSTRACT

Osteoarthritis is a degenerative joint disease for which there is no effective treatment. The large and expanding population of osteoarthritis patients worldwide makes osteoarthritis one of the leading causes of disability and pain. Studies have shown that the early prevention and diagnosis of osteoarthritis through risk prediction models can significantly improve the prognosis of patients. Nevertheless, the existing osteoarthritis risk prediction models are of poor performance and weak interpretability, which are far from meeting the needs of clinical use. Therefore, this paper proposes a GNN-based (Graph Neural Network) osteoarthritis risk prediction model based on individual genotype and phenotype information.

This paper mainly carries out three parts: First, this paper obtains the genotype and phenotype data of 13,706 individuals from the UKB database and conducts quality control and preprocessing on the data according to GWAS research and related indicators. Second, this paper designs a Chebyshev-GNN-based osteoarthritis risk prediction model, which can integrate phenotype and genotype information. Equipped with a graph estimator, this model is also capable of transforming plain data to a graph for GNN, the model interpreter qualifies the model to elaborate on the predicted result. Third, this paper introduces a series of indicators to evaluate the model's predictive performance and compares it with traditional disease risk prediction models. It is demonstrated that this model (AUC 0.74) has a very significant performance improvement compared with the traditional disease risk prediction model (AUC 0.5). This paper also demonstrates the model's excellent interpretability by analysing the graph estimator's working process.

This paper innovatively uses a graph estimator based on VEM combined with a graph neural network to process unstructured genotype data. The study's results confirmed that the model gives better risk prediction accuracy and can infer factors such as disease aetiology and classification by mining the hidden information in genotype data. The result of this study provides a new method for the processing of unstructured data in the graph neural network; it also provides a new idea for the construction of the disease risk prediction model.

**KEY WORDS:** Osteoarthritis; Risk Prediction Model; Graph Neural Network

## 目 录

1	绪论 .....	1
1.1	研究背景与意义 .....	1
1.2	现状研究综述 .....	1
1.2.1	骨关节炎与风险预测模型 .....	1
1.2.2	图神经网络 .....	3
1.3	本文研究内容 .....	3
2	数据预处理 .....	5
2.1	数据来源 .....	5
2.2	样本标注 .....	5
2.3	基因型位点质控 .....	6
2.4	特征筛选 .....	7
2.4.1	基于卡方的特征筛选 .....	8
2.4.2	基于支持向量机的特征筛选 .....	8
2.5	数据集描述 .....	9
2.6	本章小结 .....	9
3	骨关节炎风险预测模型构建 .....	11
3.1	图估计器 .....	11
3.1.1	图神经网络中无结构数据的处理方法 .....	11
3.1.2	问题描述 .....	11
3.1.3	基于变分期望最大化的图估计 .....	13
3.2	图神经网络 .....	22
3.2.1	基本定义 .....	22
3.2.2	切比雪夫层 .....	23
3.2.3	结构设计 .....	24
3.3	表型融合 .....	25
3.4	图解释器 .....	26
3.5	本章小结 .....	27
4	结果讨论与分析 .....	28
4.1	评估指标 .....	28

4.2 模型预测效果评价 .....	29
4.2.1 基准计算 .....	29
4.2.2 特征筛选方法对比 .....	29
4.2.3 图神经网络处理 .....	31
4.2.4 表型融合 .....	32
4.3 图估计效果 .....	32
4.3.1 估计器在训练过程中的效果 .....	32
4.3.2 图估计器工作过程分析 .....	36
4.4 图解释器结果案例分析 .....	45
4.5 本章小结 .....	45
5 总结与展望 .....	48
5.1 工作总结 .....	48
5.2 展望 .....	49
附录 A 外文文献原文 .....	50
参考文献 .....	50
附录 B 外文文献译文 .....	60
致 谢 .....	73

# 1 绪论

## 1.1 研究背景与意义

骨关节炎是一种由遗传因素与环境因素共同作用所产生的关节退行性病变，目前尚无有效的治疗方法。世界范围内至少三亿人罹患骨关节炎且患者群体逐年扩大。而其较差的预后也使得骨关节炎成为了全球范围内导致残疾与疼痛的主要因素之一。临床研究显示，早期诊断与介入对骨关节炎患者病程控制有着积极的影响。因此，构建骨关节炎患病风险预测模型有助于通过早期筛查与预警的方式帮助潜在患者控制病程发展。目前已有诸多关于骨关节炎基因型的全基因组关联研究（GWAS）鉴定了若干骨关节炎的易感 SNP 位点，也有相关研究基于这些位点建立了骨关节炎风险预测模型。但是这些模型的预测效果较差，远不能达到临床预测的要求。此外，现有的模型也无法处理输入基因型数据之间的复杂网络关系。同时这些模型只能根据输入的位点信息给出判断，无法解释输入位点之间的关联与特定位点在预测过程中的贡献值，模型可解释性方面存在较大不足。以上存在的问题与不足极大限制了骨关节炎风险预测模型的实际应用，如何改进构建模型的方法以提高模型的预测效果，也成为相关研究亟待解决的关键问题。

图作为一种数据结构能通过对节点与边的描述同时反映节点信息与节点之间关系，目前已被广泛应用于分子生物学代谢网络、动物行为学社交网络等生物学领域的研究中。鉴于基因型位点相互关联乃至成网的特征，同时考虑到图在提取和处理处理节点信息（特征）及节点间关系过程中的优势，本文使用图来描述基因型位点特征与位点-位点关系。但是，图数据作为一种非欧数据，传统的诸如卷积神经网络在内的机器学习方法并不适用于直接对图数据进行处理和分析。近年来随着图数据在各个领域的广泛应用，为了解决这一问题，适用于处理图数据的图神经网络应运而生。图神经网络能够基于图信息完成诸如图分类、节点分类等任务并且具有效果好、可解释性强等优点，适合用来处理基因型数据网络。因此，本文将基于图形式的基因型位点，通过图神经网络，结合患者表型信息来构建一种高效的、可解释的骨关节炎风险预测模型。该模型能够基于患者基因型及表型信息，在给出高效患病风险预测结果的同时对患者基因型网络进行解释，因此对骨关节炎的早期诊断与分型具有重要的临床意义。

## 1.2 现状研究综述

### 1.2.1 骨关节炎与风险预测模型

骨关节炎是最常见的关节退行性病变之一，目前报道的案例显示骨关节炎主要影响着人体膝、髋、手等若干关节。骨关节炎的症状主要包括关节疼痛，僵硬、柔韧性丧失，甚至导致关节失能。更严重的是，目前对骨关节炎的治疗以缓解患者痛苦为主，尚

无对骨关节的有效临床治愈手段。以上原因也使得骨关节炎成为了导致失能与残疾的主要因素之一，并且严重影响到了患者的生活质量。<sup>[?]</sup>同时，骨关节炎在世界范围内有着较庞大的患者群体：世界范围内至少有三亿人罹患骨关节炎<sup>[?]</sup>，仅在英国范围内的70岁以上群体中就有约40%的个体受骨关节炎影响<sup>[?]</sup>，而在全年龄个体中则至少有一千万患者，这导致了每年至少148亿英镑的直接医疗支出。<sup>[?]</sup>而研究显示，对骨关节炎的早期诊断与介入能很大程度上延缓关节异常生长进程，对骨关节炎患者的预后改善有着积极的效果<sup>[?]</sup>。因此许多研究也将注意力转到了基于风险预测模型的骨关节炎的早期诊断与预防上，目前已有的风险预测模型主要着眼于揭示与骨关节炎病程发展相关的影响因子，包括肥胖，关节错位，关节损伤，骨质增生与高强度的运动<sup>[? ? ?]</sup>。同时有研究发现骨关节炎也受遗传因素调控。<sup>[?]</sup>并试图通过全基因组关联分析研究骨关节炎的基因背景，以此鉴定并得到了丰富的具有统计学意义、能够作为疾病风险预测模型标志物的疾病易感单核苷酸多态性位点（Single Nucleotide Polymorphism, SNP）<sup>[? ?]</sup>。然而，基于这些疾病易感位点建立的风险预测模型性能却不尽如人意：例如<sup>[?]</sup>等人建立的基于PRS算法的风险预测模型，其效果与随机预测相当，远不能达到临床需求，还存在着极大的改进空间。

目前基于基因信息的疾病风险预测模型主要分为两种思路，一种思路通过统计学分析计算个体基因型位点对目标性状的贡献，并根据总体值对样本患病风险进行评估，该方法以PRS为代表<sup>[?]</sup>。该方法及其变体已被运用于精神分裂症、I型糖尿病与过敏性肠炎的诊断与筛查中<sup>[? ? ?]</sup>；另一思路则基于目前具有广泛应用机器学习算法，通过对样本信息进行学习，进而输出模型预测的样本患病风险。构建疾病风险预测模型所常用的机器学习方法主要分为基于回归的机器学习算法与基于树的机器学习算法。前者主要包括决策树与随机森林算法，该算法主要通过构建决策分类规则来完成输入输出数据的建模。有研究便通过随机森林法构建了II型糖尿病的疾病预测模型。<sup>[?]</sup>该研究采用的随机森林法相较于基于回归的支持向量机法有着较高的预测准确性。而后者主要有逻辑回归法、支持向量机、神经网络等算法。这类算法通过参数或非参数回归的方法构建损失函数并完成回归计算。这些算法已经被运用于癌症、老年痴呆症、心脏病以及糖尿病的风险预测<sup>[? ? ? ? ?]</sup>。而近年来随着神经网络的广泛应用，基于其发展来的深度学习疾病风险预测模型也受到越来越多的关注。一项研究肥胖预测模型的研究展现了其发掘样本信息的能力<sup>[?]</sup>。相较于传统机器学习算法，深度学习算法具有更好的预测准确性。

但是以上常见的风险预测模型构建方法仍存在着许多问题：首先，对于诸如基因型位点网络或代谢物网络等存在复杂结构的数据，上述方法都无法深层次挖掘数据的内在联系。目前的算法只将输入的位点作为独立的数据点处理，这便会导致输入阶段潜在信息的丧失。其次，这些基于机器学习或者深度学习的方法只能建立从输入数据到输出数据的映射关系，但无法结合输入数据对该映射关系的形成过程给出因果解释，即所谓的“黑箱化”。该问题使得基于以上算法构建的风险预测模型虽然能给出预测值，但是无法因此了解到使得模型做出该预测的决策过程，使得潜藏在输出数据内部的信

息被浪费。

综上，目前基于基因型信息的骨关节炎风险预测模型仍存在着预测准确率低，处理网络数据乏力，可解释性差等问题，相关风险预测模型建立的方法亟待改进。

### 1.2.2 图神经网络

图是由顶点与边组成的一种数据结构，该数据结构既描述了顶点的性质，也描述了顶点与顶点之间的相关关系。这种数据原理上同生物学领域的许多概念相契合，能够用来描述常见的例如代谢网络、SNP 网络的性质。本文主要以图的形式来整合并对输入的 SNP 数据进行处理。但是，同我们耳熟能详的图片、文本、序列等数据不同，图数据不满足平移不变性，不能投影到欧几里得空间中。而平移不变性又是目前常见的诸如卷积神经网络，递归神经网络等深度学习网络所依赖的关键假设。<sup>[?]</sup> 因此这些神经网络不能被直接用来处理图数据。但是随着图数据的广泛应用，能够处理图数据的神经网络也在逐渐发展。A. Sperduti and A. Starita<sup>[?]</sup> 率先提出了一种能够应用于有向图的神经网络。Gori<sup>[?]</sup> 等人则定义了图神经网络这一概念。而随着卷积方法在传统的成功，图卷积方法也成为了图神经网络中一项热门的研究方向，并产生了诸多已被应用到实际生产生活中用来解决诸如节点分类、图分类等问题的神经网络框架。

图卷积神经网络根据卷积核功能分为两种类：基于空间的图卷积与基于谱的图卷积。基于空间的卷积借助了信息传播<sup>[25]</sup> 的思想，认为图中的节点信息通过边进行扩散，卷积核作用于节点的空间邻域，继而通过该空间邻域计算节点信息。目前得到广泛应用的空间卷积算法主要有 GIN<sup>[?]</sup>，GAT<sup>[?]</sup>，DCNN<sup>[?]</sup> 等；而基于谱的图卷积则借助图的拉普拉斯量将图结构于傅里叶空间展开，有助于识别图结构中的潜藏结构。基于这一方法的谱卷积方法主要有 GCN<sup>[?]</sup>，ChebyNet<sup>[?]</sup>，AGCN<sup>[?]</sup> 等。目前图神经网络已被广泛应用于包括药物筛选<sup>[? ?]</sup>，计算机视觉<sup>[?]</sup> 等图数据的分析。Ghosal<sup>[?]</sup> 等人便将图神经网络用于阿尔兹海默症疾病风险的基因型预测图分类问题中，证明了使用图神经网络构建基于基因型的疾病风险预测模型的可行性。

此外，由于图的结构及其蕴含的信息特点，也有研究<sup>[?]</sup> 开发出了对图神经网络的解释器。该解释器通过分析已训练好的图神经网络与预测结果，给出与预测结果相关的子图。利用该类型解释器，我们可以在生成风险预测结果时同时获取与该预测结果相关的子图信息，并为致病位点或疾病分型相关工作提供便利。而这类型工作是传统疾病预测模型所无法实现的。

## 1.3 本文研究内容

综上所述，本文将基于谱图卷积网络，试图构建一个可解释的基于患者基因型信息与表型信息的高效骨关节炎风险预测模型。本文工作主要分为三个方面：首先，本文根据目前已发表的 GWAS 研究及公共数据库 UK BioBank 获取患者表型与基因型数据并进行数据预处理；之后构建了包括特征筛选、邻接矩阵估计、图卷积神经网络、表型

信息融合四个模块在内的骨关节炎风险预测模型；最后本文对该模型的性能以及预测结果进行了进一步的分析和解读，继而对模型的预测准确性，可解释性等指标进行评估。具体研究内容分为以下五个章节

第一章，绪论。该章首先阐述论文的研究背景与意义，并对骨关节炎、风险预测模型以及图神经网络领域加以综述。最后介绍了本文的主要研究内容。

第二章，数据预处理。该章介绍了根据已有研究从公共数据库获取患者基因型与表信息以及根据一定评判标准对该数据进行筛选与预处理的过程。

第三章，风险预测模型搭建。该章介绍了从特征筛选、邻接矩阵估计、图卷积神经网络、表型信息融合四个模块出发构建骨关节炎风险预测模型的过程与理论原理。

第四章，模型性能评价。该章分析了本文搭建模型的预测准确率并将其同传统风险预测模型准确率进行比较。同时还从模型可解释性角度出发分析了模型预测结果。一方面证明本文建立模型相较于传统模型具有较高程度的提升，另一方面也展现了模型对骨关节炎性状之外信息的揭示能力。

第五章，总结与展望。总结本文对骨关节风险预测模型的研究工作，并对未来本研究还需要解决的问题进行展望。

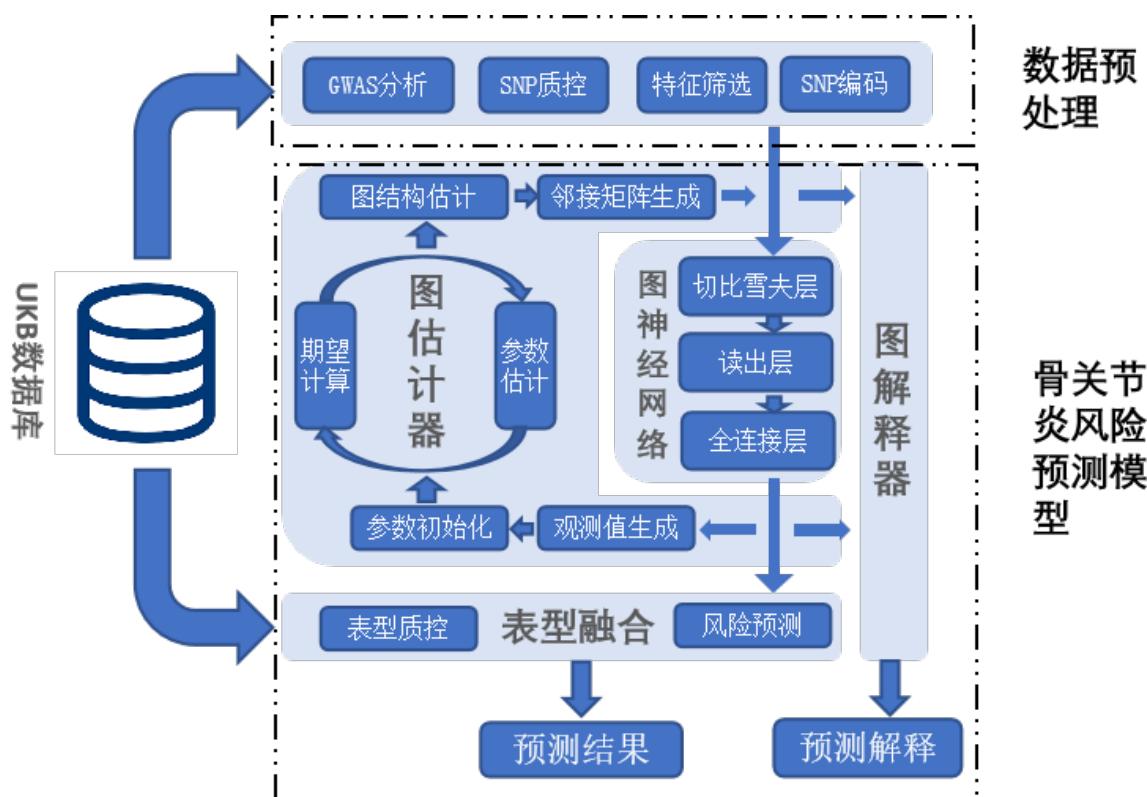


图 1-1 技术路线图

## 2 数据预处理

### 2.1 数据来源

本文所使用的基因型与表型数据主要来自于 UK Biobank[?] 数据库。该数据库收集并整合了 2006-2010 年间于英国招募的约 500000 名志愿者的生物样本、身体状况、功能评估、多种表型及遗传学数据。由于其所提供的丰富多样的遗传与表型信息，UK Biobank 已经成为了研究多种复杂表型和遗传病的重要数据来源。本文主要基于 UK Biobank 的第三批发布数据开展骨关节炎相关研究。

### 2.2 样本标注

我们从 UK Biobank 数据库中获取骨关节炎患者（阳性样本）与非骨关节炎患者（阴性样本），并使用个体疾病伤害分类标准编码来对 UK Biobank 所提供的样本信息进行区分。国际疾病伤害及死因分类标准（The International Statistical Classification of Diseases and Related Health Problems, ICD）是由世界卫生组织所制定的根据特定规则对人群可能出现疾病进行归类的一套编码系统。[?] UK Biobank 已经对数据库内样本进行了 ICD 标注，因此我们可通过这一方法筛选研究对象。

参考其他骨关节炎的相关研究，[?] 我们确定如下筛选规则。我们认定包含如表2-1所含 ICD 编码的个体为骨关节炎患者。

表 2-1 患者评判依据

ICD 编码	分类
M15-X	骨关节炎
M16-X	髋关节炎
M17-X	膝关节炎
M18-X	腕掌关节关节炎
M19-X	其他关节炎

在筛选非骨关节患者时，为防止其他关节症状的干扰，我们也根据文献报道对表2-2所含编码的个体予以排除。

依照以上规则，我们从 UK Biobank 数据库中筛选并获得 6706 名骨关节炎患者作为模型训练的阳性集。同时，为保证风险预测模型训练数据的均衡性，我们又挑选出了 7000 名正常个体作为阴性集。两部分数据共同构成本次研究所使用的数据集。

表 2-2 非患者评判依据

ICD 编码	分类
M11-X	软骨钙化症
M20-X、M21-X	获得性关节畸形
M22-X	髌骨功能异常
M23-X	膝关节异常
M24-X	其他关节异常
M25-X	关节疼痛
M42-X	脊柱软骨病

## 2.3 基因型位点质控

人类基因组具有复杂性与多样性，本研究中的每个样本个体基因型数据都由数十万乃至数百万个基因型位点组成。由于目前任何算法都无法短时间内对如此多的样本进行处理，因此需要从其中选取部分具有统计学意义的位点。本文根据目前已发表的骨关节炎的 GWAS 结果 [? ?] 对样本的基因型位点进行关联显著值（P-Value）与等位基因频率（Allele Frequency, AF）质控。

关联显著值是用来衡量 GWAS 研究中单核苷酸突变与给定性状关联显著性的一个统计量。在 GWAS 研究中，我们设定零假设为数据中没有 SNP 位点与特定性状相关联，而备择假设为至少有一个 SNP 位点同特定性状相关联。同时我们定义统计量  $p$  为当零假设为真时观察到该关联的概率。显然， $p$  值越小，有越高把握在观察到关联时认为零假设为假。因此我们可以设定一个阈值，当统计量  $p$  低于该阈值时拒绝原假设，认为该位点同目标性状相关。[?] 目前 GWAS 研究中对位点统计显著值的描述可通过 Manhattan 图进行，Manhattan 图的横坐标为 SNP 位点在基因组中的坐标，纵坐标为对应 SNP 位点在 GWAS 研究中的统计显著值。本研究所选区的 SNP 位点 Manhattan 图如图2-1所示。

但是，由于遗传漂变的影响，随机的基因突变也有可能具有类似相关关系。因此我们并不能认为  $p$  值较小的位点同目标位点直接相关，在指定阈值时需要将可能由遗传漂变所产生的位点筛去。因此我们还需要通过 QQ 图来完成阈值界定。QQ 图是一种根据分位数对两概率分布所作的图，该图来比较两分布差别的方法。[?] GWAS 中的 QQ 图横轴为遗传漂变所产生的分布，纵轴为实际分布，当图像偏离对角线时认为这类数据并非随机漂变。因此，根据 QQ 图2-2本文确定  $p$  阈值为  $10^{-5}$ 。

等位基因频率也是进行基因型数据质控时常见的评判指标。在 GWAS 研究中往往会出现一些出现频率很低的突变。受限于 GWAS 原理，大多数研究不能很好计算低等位频率位点与性状的相关性。因此在实际研究时，还需要通过等位基因频率对所得位点进行进一步质控。根据文献我们将该阈值设定为 0.05[? ]。

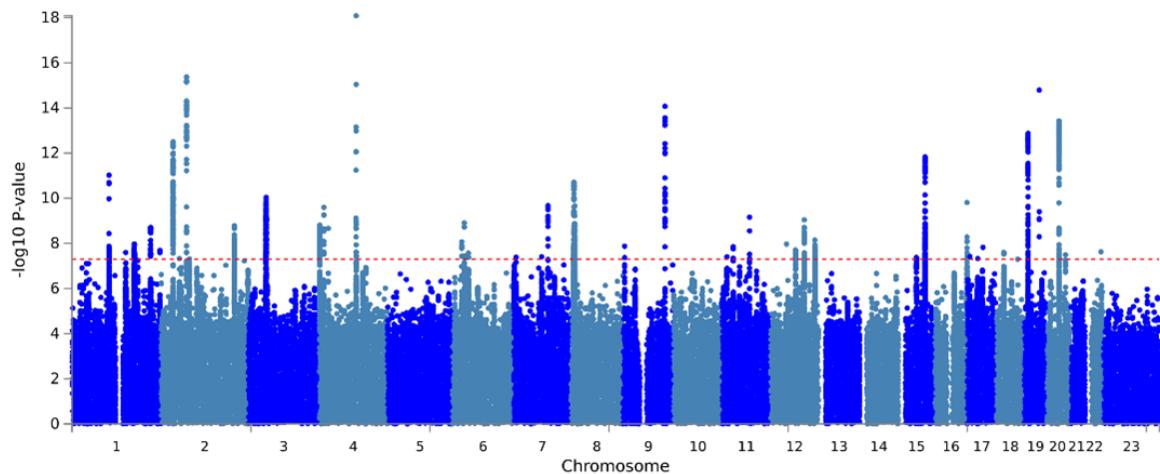


图 2-1 GWAS 研究的 Manhattan 图

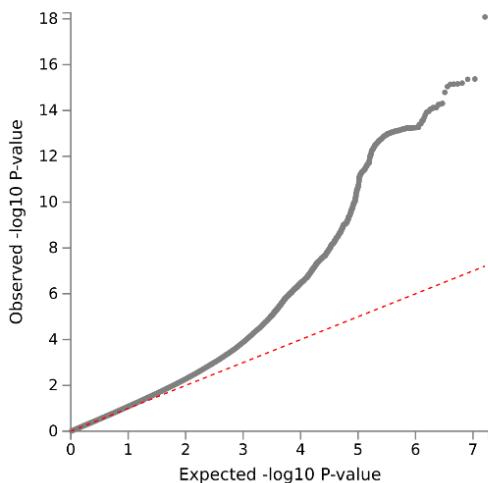


图 2-2 GWAS 研究的 Q-Q 图

综上，本文根据关联分析显著值与等位基因频率对所得 SNP 位点进行筛选，最终每个个体选取选取 8687 个 SNP 位点作为待分类基因型位点。同时，我们根据位点目标突变的出现频率对数据进行编码，将基因型信息转化为算法可直接读取的数值信息。

## 2.4 特征筛选

目前我们已经从 UKB 数据库中选取了 6706 位骨关节炎患者作为阳性样本，每个患者包括 8687 个基因型特征。考虑到特征数大于样本数，可能对预测模型的准确率产生不利影响。<sup>[?]</sup> 本文还使用卡方法与支持向量机法对现有数据进行了特征筛选。

### 2.4.1 基于卡方的特征筛选

卡方法使用统计学方法对输入特征与样本标注的关联进行计算，并根据计算结果筛选与样本标注相关程度较高的特征。其具体过程如下：

1. 确定零假设与备择假设：对于特征与样本标注，定义零假设  $H_0$  为该特征与样本标注无关；定义备择假设  $H_1$  为该特征与样本特征相关
2. 计算特征卡方值：根据公式计算特征与样本标注的卡方值

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2-1)$$

其中：

- $c$  为该卡方分布的自由度
  - $O$  为观测值
  - $E$  为期望值
3. 查卡方表，拒绝或接受假设：对于计算所得卡方值，查卡方表。对照计算所得卡方值与一定自由度与置信度下的标准卡方值。当计算卡方值大于标准卡方值时接受零假设，认为该特征与样本标注无关；当计算卡方值小于标准卡方值时拒绝零假设，认为该特征与样本标注相关。基于此选择该特征

### 2.4.2 基于支持向量机的特征筛选

支持向量机（Support Vector Machine, SVM）是一种著名的机器学习算法。其通过构建数据空间内的分类平面来实现数据分类算法。其算法核心为以下优化过程：

$$\min_{W,b} [C \sum_i \max(0, 1 - y_i(X_i^T W + b)) + l(W)] \quad (2-2)$$

其中：

- $X_i$  为样本数据
- $W, b$  为超平面参数
- $y_i$  为样本标注
- $C$  为惩罚系数，即错误分类情况下对优化函数的惩罚
- $l(W)$  为超平面拟合程度的评价函数

从上式可以看出，该优化目标主要分为两部分：一部分衡量超平面对数据点的分类能力；另一部分衡量超平面对数据的拟合能力。目前常用的 SVM 算法中评价函数通常具有如式??所示形式。

$$l(W) = \frac{1}{2} w^T w \quad (2-3)$$

此时该拟合函数评价分离两类数据点超平面在超空间上的距离。距离越大，证明该超平面对数据的拟合能力越好。但是，当该函数变为

$$l(W) = e^T |w| \quad (2-4)$$

此时该拟合函数又称  $w$  的  $l_1$  范数，又称 Lasso 惩罚 [? ]，该范数可以用来评价特征在整体分类效果中的贡献 [? ]。因此本文可通过该范数对特征进行筛选。

## 2.5 数据集描述

经过以上步骤，我们制得包括 13706 位个体、每个体选取 200 个基因型位点的基因型数据集。数据集大小 2.3GB。其中患病正样本 6706 例，正常负样本 7000 例。每个基因型位点我们通过 8 个字段描述，字段及其含义见表2-3。

表 2-3 基因型位点字段信息

字段	含义
RSID	SNP ID
Chromosome	SNP 所在染色体
BP	以 GRCh37 为参考的碱基对位置
Effect Allele	与表型相关的效应等位位点
Non-Effect Allele	与表型无关的等位位点
Beta	效应值
P	统计显著性
MAF	等位基因频率

同时为了计算与处理方便，我们对患者基因型位点 SNP 数据根据效应等位位点  $E$  与非效应等位位点  $N$  进行编码，编码方式见表2-4。

表 2-4 基因型位点编码方式

SNP 信息	编码
$EE$	(1,1)
$EN$	(1,0)
$NN$	(0,0)

## 2.6 本章小结

本章介绍了本文所使用数据集获取与预处理过程。首先，根据 ICD 编码我们从 UKB 数据库中选取了 6706 名患病个体（阳性集）与 7000 名非骨关节炎个体（阴性集）作为研究对象。其次，根据关联分析显著性及基因频率从已发表的骨关节炎 GWAS 中筛选出感兴趣的位点并获取研究样本相应的基因型数据，之后再根据位点信息对其进行编码。最后，为了进一步减少无关的样本特征，加快计算速度，本文通过卡方法与支持向量机法对数据进行特征筛选。最终完成了由每个样本包含 200 个基因型信息位点的 13706 样本所构成的骨关节炎风险预测模型基因型数据集的构建。

### 3 骨关节炎风险预测模型构建

第二章我们获取了骨关节炎患者的数据并对其进行了预处理与特征筛选。本章则主要对骨关节炎风险预测模型做以介绍。该模型首先根据基于变分期望最大化算法的图估计器构建了输入基因型数据的关联并通过图来表示该关联关系。之后构建了图神经网络对建立的图进行学习与处理并最终完成图的分类。该模型还通过融合患者表型进一步增强模型性能。最后本模型还构建了基于预测值与输入数据的模型解释器。

#### 3.1 图估计器

##### 3.1.1 图神经网络中无结构数据的处理方法

图神经网络只适用于图类型的数据，对常见无结构数据（例如文本，图像）的处理较为困难。本文目前获得的基因型数据也属于无结构数据，原理上不适用于图神经网络。但是，考虑到图神经网络优秀的性能与可解释性，我们需要基于该无结构基因型数据构建图，继而输入图神经网络。目前根据无结构数据建图的方法主要有两种方法：一种是通过分析数据构建静态图，神经网络训练过程中不对该图结构加以更新 [? ?]，例如在处理无结构的图像数据时，Monti[?] 提出了一种通过超像素将图像数据转为图的方法，使得图神经网络能够处理图像数据；第二种是通过学习方法构建动态图结构，并在网络训练过程中不断更新图结构 [?]。这类通过学习方法获取图结构特征的过程也被称为图结构学习（Graph Structure Learning, GSL）。目前的图结构学习过程主要分为三个步骤：首先通过已知数据通过  $k$  近邻法 [?] 或者阈值法 [?] 构建出一个初始图结构或观测图结构；然后使用诸如随机块模型（Stochastic Block Model, SBM）等方法 [?] 对该观测图结构进行建模；最后通过算法估计出数据可能具有的图结构。在图估计的过程中，基于统计学的算法受到了广泛关注：例如 Zhang[?] 等人提出了一种基于贝叶斯思想的图估计器。其认为图结构根据依赖一定参数的分布所产生，如果通过蒙特卡洛方法对该参数进行估计，就可因此得到该分布的具体信息，继而对图结构进行估计。Elinas[?] 同样基于贝叶斯思想提出了一种使用变分推断法的图估计方法。但是，以上研究仅适用于图神经网络中的节点分类问题，同本研究所属图分类问题不符，不能直接应用于本研究中。但考虑到这类贝叶斯方法思路简洁，原理清晰。因此，本文也将基于贝叶斯思想构建适用于本文图分类问题的图结构估计器。

##### 3.1.2 问题描述

贝叶斯思想认为：事件的观测值并不能反映事件的真实特性。贝叶斯方法因此主要解决由事件观测值向事件真实值的推断过程。但是在讨论推断过程前，我们首先需要对观测值与真实值加以定义。

### 3.1.2.1 观测值构建

对于本文研究对象基因型数据而言，未经图神经网络处理的原始数据所含信息量较少，较难依据初始数据构建观测值。但是有研究显示，经过谱图神经网络处理之后，具有较强关联的节点具有相似的值。<sup>[?]</sup> 我们因此认为在图神经网络的输出值中，对任一节点，如果有一其他节点与其有着相似的值，则两节点之间可能存在关联。这种关联可以通过 k 近邻算法<sup>[?]</sup> 计算从而得到 k 近邻网络，而该 k 近邻网络可以很好地描述节点的局部特征。在 Wang<sup>[?]</sup> 等人的研究中，也有通过根据图神经网络输出值构建 k 近邻网络的描述。我们因此根据每个样本经图神经网络处理后的输出值建立 k 近邻网络，再将所有网络相加求平均以获得单个观测值。同时为了防止信息丢失，我们将首次经图神经网络处理时输入的图初始化为全连接图，即认为每个节点都同其他节点存在关联。

### 3.1.2.2 真实值构建

本研究中所感兴趣的真实值为实际的未知图结构。但是对图结构的直接计算过于复杂，我们因此需要将图结构参数化，即使用少数几个参数来描述图结构。目前在图论领域常用的图参数化模型为随机块模型<sup>[?]</sup>，该模型认为图中的节点属于某几个簇，节点之间是否相连仅与两节点所在的簇相关且服从某一参数化分布。因此我们只需要得出支配该分布的参数就能对图结构加以复现，继而用少数参数描述庞大的网络。因此本文将该类参数作为推断的目标，而由其产生的图结构作为模型输出值。

### 3.1.2.3 推断过程

界定了观测值与真实值之后，我们便可对图估计器所解决的问题加以严谨描述。

图估计器首先根据式3-1有图神经网络的输出构建观测值  $O$ 。

$$O^{ij} = \frac{1}{N} \sum_N o_n^{ij} \quad (3-1)$$

其中  $O^{ij}$  描述了节点  $i, j$  之间关联的观测值， $N$  为样本数， $o_n^{ij}$  为由第  $n$  个样本神经网络输出值构建的 k 近邻网络中节点  $i, j$  之间的关联。通过该观测值我们构建图  $G = (V, O)$ ，其中  $V = \{v_1, \dots, v_n\}$  描述图中节点，即样本 SNP 位点。

为了通过 SBM 模型参数化图结构，我们假定图中共含有  $C$  个簇，每个节点  $i$  从属于簇  $m$  的概率由式3-2来描述，且  $z_i$  相互独立并满足由式3-3所描述的多项分布。

$$z_{im} \in \mathbf{Z}^{N \times C}, i \in \{1, \dots, n\}, m \in \{1, \dots, C\} \quad (3-2)$$

$$z_i \stackrel{iid}{\sim} M(1, \alpha) \quad (3-3)$$

基于此，我们使用  $z_i$  来描述节点之间关联。我们认为任意两节点之间的关联相互独立并满足如式3-4所描述二项分布。

$$o_{ij} | z_{im}, z_{jn} = 1 \sim B(1, \pi_{z_i z_j}) \quad (3-4)$$

可以看出，只要解出参数  $\pi$  与  $\alpha$ ，就可以此构建出如图3-1所示的 SBM 模型，继而对图结构加以推断；因此本文给出以下目标：

Representation of Stochastic Block Model

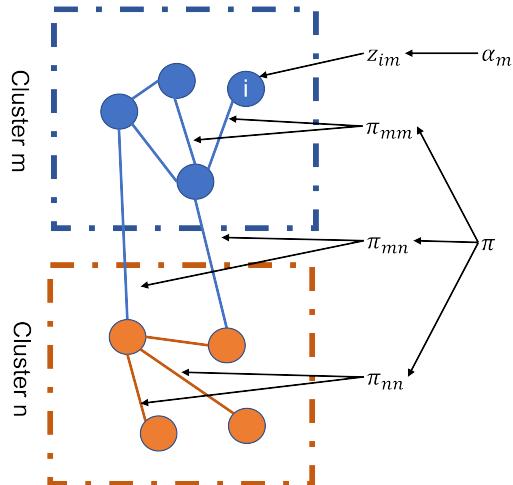


图 3-1 随机块模型及其参数化示意图

基于观察到的节点关联信息  $O$  以及隐变量  $Z$ ，通过式3-5估计参数  $\theta = \{\alpha, \pi\}$ ，并根据该参数给出推断图结构

$$\hat{\theta} = \arg \max l_\theta(O) \quad (3-5)$$

### 3.1.3 基于变分期望最大化的图估计

#### 3.1.3.1 推导与整理

在节点关联与节点簇信息均已知的情况下，我们首先由式3-6计算似然函数  $l_\theta$ ，根据边缘概率公式，有

$$l_\theta(O) = \sum_Z \log p_\theta(O, Z) = \sum_Z (\log p_\pi(O|Z) + \log p_\alpha(Z)) \quad (3-6)$$

其中

$$\sum_Z (\log p_\pi(O|Z) + \log p_\alpha(Z)) = \sum_Z \left( \sum_{i,j}^n \sum_{m,n}^C \log p_\pi(o_{ij}|z_{im}z_{jn}) + \sum_i^N \sum_m^C \log p_\alpha(z_{im}) \right) \quad (3-7)$$

给定参数时，有

$$p_\pi(o_{ij}|z_{im}z_{jn}; \pi) = \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} \quad (3-8)$$

$$p_\alpha(z_{im}) = \alpha_m^{z_{im}} \quad (3-9)$$

将式3-8与3-9代入式3-7可得式3-10。

$$\begin{aligned} l_\theta(O; \theta) &= \sum_Z \left( \sum_{i,j}^n \sum_{m,n}^C \log p_\pi(o_{ij}|z_{im}z_{jn}) + \sum_i^N \sum_m^C \log p_\alpha(z_{im}) \right) \\ &= \sum_Z \left( \sum_{i,j}^n \sum_{m,n}^C \log \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} + \sum_i^N \sum_m^C \log \alpha_m^{z_{im}} \right) \end{aligned} \quad (3-10)$$

### 3.1.3.2 期望最大化算法

考虑到似然函数中含隐变量  $Z$ ，通常我们采用期望最大化算法 [?] 计算使得似然函数取极大的参数。我们对似然函数式3-6做如下处理

$$\begin{aligned} l(O; \theta) &= \log p(O; \theta) \\ &= \log \int_Z p(O, Z; \theta) dz \\ &= \log \int_Z q(Z) \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \log \mathbb{E}\left[\frac{p(O, Z; \theta)}{q(Z)}\right] \end{aligned} \quad (3-11)$$

其中  $q(Z)$  为一辅助函数且满足  $\int q(Z) = 1$ 。根据 Jensen 不等式 [?]，对任意一凸函数，有

$$\mathbb{E}[f(x)] \leq f(\mathbb{E}[x]) \quad (3-12)$$

当且仅当  $x = \mathbb{E}[x]$  时等号成立。又因为  $\log x$  为一凸函数，因此将式3-12代入式3-11，有

$$\begin{aligned}
 l(O; \theta) &= \log p(O; \theta) \\
 &= \log \mathbb{E}\left[\frac{p(O, Z; \theta)}{q(Z)}\right] \\
 &\geq \mathbb{E}\left[\log \frac{p(O, Z; \theta)}{q(Z)}\right]
 \end{aligned} \tag{3-13}$$

当且仅当式3-13满足式3-14所示条件时等号成立。

$$\frac{p(O, Z; \theta)}{q(Z)} = c \tag{3-14}$$

对式3-14进行变化，有：

$$q(Z) = c \cdot p(O, Z; \theta) \tag{3-15}$$

又因为  $\int q(Z) dz = 1$ ，有：

$$\begin{aligned}
 \int_Z q(Z) dz &= \int_Z c \cdot p(O, Z; \theta) dz = 1 \\
 c &= \frac{1}{\int_Z p(O, Z; \theta) dz}
 \end{aligned} \tag{3-16}$$

因此：

$$q(Z) = \frac{p(O, Z; \theta)}{\int_Z p(O, Z; \theta) dz} = \frac{p(O, Z; \theta)}{p(O; \theta)} = p(Z|O; \theta) \tag{3-17}$$

将式3-17代回原不等式3-13，我们由式3-18构造辅助函数  $Q$ 。

$$l(O; \theta) \geq \mathbb{E}\left[\log \frac{p(O, Z; \theta)}{q(Z)}\right] = \int_Z p(Z|O; \theta) \log \frac{p(O, Z; \theta)}{p(Z|O; \theta)} dz = Q(\theta^{t+1}; \theta) \tag{3-18}$$

显然，在取等条件成立时， $Q$  为似然函数  $l(O; \theta)$  的下界。期望最大化算法因此首先计算给定预期参数  $\theta$  时下界  $Q$  的值，再通过优化参数  $\theta$  提高下界  $Q$ ，通过不断迭代计算使得似然函数极大的参数  $\hat{\theta}$ 。具体步骤描述如下：

- E 步骤：计算  $Q(\theta^{t+1}; \theta)$
- M 步骤：优化参数  $\theta$

但在下界  $Q$  的计算过程中我们由式3-18设  $q(Z) = p(Z|O; \theta)$ ，但是在本问题中， $p(Z|O; \theta)$  过于复杂，无法计算，因此也无法完成后续步骤。我们需要采取其他方法来估计参数  $\theta$ 。

### 3.1.3.3 变分期望最大化

**3.1.3.3.1 变分推理** 变分推理是贝叶斯统计领域遇到无法计算的分布时常见的处理方法 [? ]。它通过构建辅助函数来模拟无法计算的分布，并通过优化过程使得辅助函数逐渐逼近真实分布。[?] 该方法与本文所处理问题较为契合，我们因此结合变分推理来处理期望最大化过程中无法计算的分布  $p(Z|O; \theta)$ 。

**3.1.3.3.2 证据下限与 KL 散度** 在进一步解释之前我们需要先对可能使用到的概念加以说明。在上一节中我们已经证明，根据 Jensen 不等式，含隐变量的似然函数具有一下界。我们将该下界定义为似然函数的证据下界 (Evidence Lower Bound, ELBO)[? ]。其具有如式3-19所示形式。

$$\begin{aligned} l(O; \theta) &= \log \int_Z q(Z) \frac{p(O, Z; \theta)}{q(Z)} dz \\ &\geq \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= J(q(Z); \theta) \end{aligned} \tag{3-19}$$

我们同时根据式3-20定义似然函数与证据下界之间的差为分布  $q(Z)$  与分布  $p(Z|O)$  之间的 KL 散度 (Kullback-Leibler divergence)。[? ]

$$\begin{aligned} l(O; \theta) - \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz &= \int_Z q(Z) \log p(O; \theta) dz - \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \int_Z q(Z; \theta) \log \frac{p(O; \theta)q(Z; \theta)}{p(O, Z; \theta)} dz \\ &= - \int_Z q(Z; \theta) \log \frac{p(Z|O; \theta)}{q(Z; \theta)} dz \\ &= KL(q(Z)||p(Z|O)) \end{aligned} \tag{3-20}$$

可以看出，该散度衡量了两分布间的差异，可以用于后续辅助函数的构造。

**3.1.3.3.3 辅助函数构建** 基于变分推理思想，我们构建辅助分布  $q_\psi(Z)$  来模拟无法计算的  $p(Z|O)$ 。基于期望最大化思想，我们构建如式3-21所示似然函数。

$$l(O; \theta) = J(q_\psi(Z); \theta) + KL(q_\psi(Z)||p(Z|O)) \tag{3-21}$$

式3-21中首项描述了在使用辅助分布替换  $p(Z|O)$  时原始似然函数的下界，第二

项描述了辅助分布与原始分布之间的差值。值得注意的是，KL 散度具有非负性，即  $KL(\cdot||\cdot) \geq 0$ 。我们因此给出如式3-22所示不等式。

$$l(O; \theta) \geq J(q_\psi(Z); \theta) \quad (3-22)$$

借助期望最大化思想，我们如果能够通过迭代计算辅助函数与参数使得式3-22中下界  $J(q_\psi(Z); \theta)$  不断升高，就能使得该似然函数最终取最大。展开该下界，得式3-23。

$$\begin{aligned} J(q_\psi(Z); \theta) &= \int_Z q(Z) \log \frac{p(O, Z; \theta)}{q(Z)} dz \\ &= \int_Z q(Z) \log p(O, Z; \theta) dz - \int_Z q(Z) \log q(Z) dz \\ &= \int_Z q(Z) \left( \sum_{i,j}^n \sum_{m,n}^C \log \binom{1}{O_{ij}} [\pi_{mn}(1 - \pi_{mn})^{1-O_{ij}}]^{z_{im}z_{jn}} + \sum_i^N \sum_m^C \log \alpha_m^{z_{im}} \right) dz \\ &\quad - \int_Z q(Z) \log q(Z) dz \\ &= \mathbb{E}_{q_\psi(Z)} \left[ \sum_i^N \sum_m^C z_{im} \log \alpha_m \right] - \sum_i^N \sum_m^C \mathbb{E}_{q_\psi(Z)} [z_{im}] \log \mathbb{E}_{q_\psi(Z)} [z_{im}] \\ &\quad + \mathbb{E}_{q_\psi(Z)} \left[ \sum_{i < j}^N \sum_{m,n}^C z_{im} z_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \right] \end{aligned} \quad (3-23)$$

此时，我们构造基于服从参数  $\tau$  的多项分布的辅助函数，使得该辅助函数满足式3-24。

$$q_\psi(Z) = \prod_i^N q_\psi(Z_i) = \prod_i^N M(Z_i, \tau) \quad (3-24)$$

对于该变分参数  $\tau$ ，我们依据多项分布的性质给出如下限制条件  $\tau \in [0, 1]^C, \sum_m \tau_{im} = 1$ 。可以看出，该参数同原始隐变量  $Z$  的意义相同，都描述了节点属于簇的概率。因此我们给出如式3-25与3-26所示关系。

$$\tau_{im} = p(q_\psi(z_{im} = 1)) = \mathbb{E}_{q_\psi(Z)} [z_{im}] \quad (3-25)$$

$$\tau_{im}\tau_{jn} = p(q_\psi(z_{im} = 1, z_{jn} = 1)) = \mathbb{E}_{q_\psi(Z)} [z_{im}z_{jn}] \quad (3-26)$$

将式3-25与3-26代入式3-23，有式3-27所示形式。

$$\begin{aligned} J(O; \theta) = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-27)$$

此时我们可以发现，虽然该下界的形式较为复杂，但是在给定辅助函数与参数的情况下依然能够给出结果，我们采用变分法的目的也就达到了。因此我们设计如下变分期望最大化算法。

- VE 步骤：固定参数  $\theta$ ，根据  $J$  更新变分参数  $\tau$ ，并计算更新参数后下界  $J$  的值
- M 步骤：固定变分参数  $\tau$ ，根据  $J$  更新参数  $\theta$ ，直至收敛。

### 3.1.3.3.4 参数求解

**VE 步** 此步骤中主要基于式3-28根据式3-33求解变分参数  $\tau$ 。

$$\hat{\tau} = \arg \max_{\tau} J(O, \theta; \tau) \quad (3-28)$$

$$\begin{aligned} J(O, \theta; \tau) = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-29)$$

考虑到式3-33中的参数存在限制条件  $\sum_m \tau_{im} = 1$ ，我们采用拉格朗日乘数法 [? ] 对该式加以计算得3-30。

$$\begin{aligned} L(J, \lambda) = & J(O, \theta; \tau) + \lambda_i (\sum_m \tau_{im} - 1) \\ = & - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m + \lambda_i (\sum_m \tau_{im} - 1) \\ & + \sum_{i < j} \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-30)$$

将式3-30分别对  $\tau_{im}$  与  $\lambda$  求偏导，得式3-31与3-41。

$$\frac{\partial L}{\partial \tau_{im}} = \sum_j^N \sum_n^C \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) + \log \alpha_m - \log \tau_{im} + \lambda_i - 1 = 0 \quad (3-31)$$

$$\frac{\partial L}{\partial \lambda_i} = \tau_{im} - 1 = 0 \quad (3-32)$$

对式3-31加以变换，我们得到式3-33。

$$\log \tau_{im} = \sum_j^N \sum_n^C \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) + \log \alpha_m - \lambda_i + 1 \quad (3-33)$$

我们因此基于式3-33给出  $\tau_{im}$  的解析解

$$\begin{aligned} \tau_{im} &= e^{-\lambda_i+1} \alpha_m \left[ \prod_j^N \prod_n^C \left[ \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right]^{\tau_{jn}} \right] \\ &\quad \forall i \in \{1 \dots N\}; \forall m \in \{1 \dots C\} \end{aligned} \quad (3-34)$$

可以看出，每个变分参数的计算都依赖于其他变分参数，虽然可以通过迭代至收敛的方法计算，但这也对参数初始值的选取提出了要求，具体细节将于后文讨论。

**M步** 此步骤中主要基于式3-35求解目标参数  $\theta$ 。

$$\hat{\theta} = \arg \max_{\theta} J(O, \tau; \theta) \quad (3-35)$$

首先我们根据式3-36计算  $\theta$  中的  $\pi$ 。由于  $\pi$  在式中相对较为独立，我们直接对其求偏导得式3-37。

$$\begin{aligned} J(O, \tau, \alpha; \pi) &= - \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m \\ &\quad + \sum_{i < j} \sum_{m, n} \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) \end{aligned} \quad (3-36)$$

$$\frac{\partial J(O, \tau, \alpha; \pi)}{\partial \pi_{mn}} = \sum_{i < j} \tau_{im} \tau_{jn} \left( \frac{O_{ij}}{\pi_{mn}} - \frac{1 - O_{ij}}{1 - \pi_{mn}} \right) \quad (3-37)$$

求解以上方程，得如式3-38所示的  $\pi$  的解析解。

$$\hat{\pi}_{mn} = \frac{\sum_{i < j} \tau_{im} \tau_{jn} O_{ij}}{\sum_{i < j} \tau_{im} \tau_{jn}} \quad (3-38)$$

对于  $\theta$  中的  $\alpha$ ，由于其也存在限制条件  $\sum_m \alpha_m = 1$ ，我们依旧采用拉格朗日乘数法计算得式3-39。

$$\begin{aligned}
L(J, \lambda) &= J(O, \pi, \tau; \alpha) + \lambda_i(\sum_m \alpha_m - 1) \\
&= -\sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_i \sum_m \tau_{im} \log \alpha_m + \lambda_i(\sum_m \alpha_m - 1) \\
&\quad + \sum_{i < j}^N \sum_{m,n}^C \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right)
\end{aligned} \tag{3-39}$$

将该式分别对  $\alpha_m$  与  $\lambda$  求偏导，得式3-40与式3-41。

$$\frac{\partial L}{\partial \alpha_m} = \frac{\sum_i \tau_{im}}{\alpha_m} + \lambda_i = 0 \tag{3-40}$$

$$\frac{\partial L}{\partial \lambda_i} = \alpha_m - 1 = 0 \tag{3-41}$$

对式3-40加以处理，得：

$$\begin{aligned}
\frac{\sum_i \tau_{im}}{\alpha_m} + \lambda_i &= 0 \\
\sum_i \tau_{im} + \alpha_m \lambda_i &= 0 \\
\sum_m \sum_i \tau_{im} + \sum_m \alpha_m \lambda_i &= 0 \\
\lambda_i &= -\sum_m \sum_i \tau_{im}
\end{aligned} \tag{3-42}$$

代入式3-40，得如式3-43所描述的  $\alpha$  的解析解。

$$\begin{aligned}
\hat{\alpha}_m &= \frac{\sum_i \tau_{im}}{\sum_m \sum_i \tau_{im}} \\
&= \frac{\sum_i \tau_{im}}{N}
\end{aligned} \tag{3-43}$$

至此，我们已经完成了所有参数的估计。确定参数后只需将其代回原始模型即可得出估计的节点之间关联，继而完成图结构的构建。

### 3.1.3.4 起始与终止条件

之前在对变分参数  $\tau$  的计算中我们发现，任一参数的计算都依赖于其他参数。但是如果以零为初始值，参数在计算时只会考虑项  $e^{-\lambda_i+1} \alpha_m$ ，无法对参数  $\pi$  形成有效估计。因此我们需要通过一定方法给予该变分参数初值，并通过迭代的方法计算  $\tau$  直至收敛。目前的研究中常使用 k-Means 法 [?] 先对节点进行聚类，根据聚类结果对变分参数进行赋值。但是 k-Means 法需要预先确定图中节点簇数  $C$ ，且对于一般问题而言

该簇数未知，因此需要一种能够评估所选簇数对最终似然函数影响的方法。目前研究中，一种叫做贝叶斯信息指标（Bayesian Information Criterion, BIC）[?] 的方法可以用来完成类似工作，其定义为式3-44。

$$BIC(C) = \log P(O; \theta) - \frac{V_C}{2} \log N \quad (3-44)$$

其中  $\log P(O; \theta)$  描述对应模型下的似然函数， $V_C$  描述了选择对应簇数  $C$  时模型参数数量，但是该方法涉及到无法求解似然函数的计算，在本研究中无法实现。但是，基于类似的思想，Daudin 等人提出另一种评价指标：整合分类似然（Integrated Classification Likelihood, ICL）[?]，该指标并不直接计算原始似然函数，而是通过计算变分时使用的变分似然函数给出评判依据。该指标定义如式3-45。

$$ICL(C) = \sum_i \sum_m \tau_{im} \log \tau_{im} + \sum_{i < j} \sum_{m,n} \tau_{im} \tau_{jn} \log \left( \binom{1}{O_{ij}} \pi_{mn} (1 - \pi_{mn})^{1-O_{ij}} \right) - \frac{1}{2} \left( \frac{C(C+1)}{2} \log \frac{N(N-1)}{2} - (C-1) \log N \right) \quad (3-45)$$

基于以上信息，我们给出图3-2描述的基于变分期望最大化的图估计算法全流程。

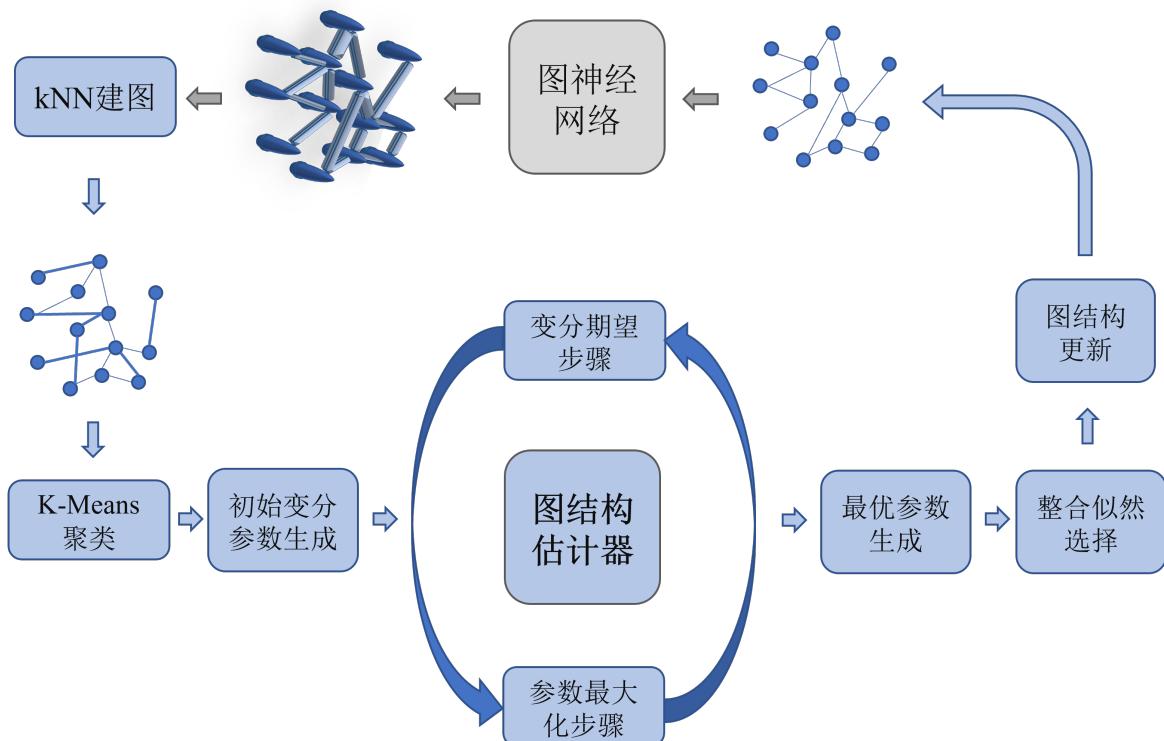


图 3-2 基于变分期望最大化的图估计算法流程

---

**Data:** 观测值  $O$  与聚类数目  $c$

**Result:** 模型参数  $\theta$

**Input:** 观测值, 确定簇数范围

1 使用 k-Means 法对输入值进行聚类, 给出初始变分参数  $\tau_0$

2 **while**  $\theta$  不收敛 **do**

3    VE 步

4    **while** 变分参数  $\tau$  不收敛 **do**

5     | 循环计算  $\hat{\tau} = \arg \max_{\tau} J(O, \theta; \tau)$

6    **end**

7    根据该变分参数计算  $J$

8    M 步

9    根据  $\hat{\theta} = \arg \max_{\theta} J(O, \tau; \theta)$  计算参数  $\theta$

10   计算最优似然下界  $\hat{J}$  与该簇数下的整合分类似然  $ICL(C)$

11 **end**

**Output:** 选择使得整个分类似然最小的簇数  $C$ , 选择该簇数计算出的参数  $\hat{\theta}$

---

算法 3-1 基于变分期望最大化的图估计算法

## 3.2 图神经网络

在完成了输入图结构的估计后, 我们便可开始设计模型中图处理的核心——图神经网络。基于绪论中对图神经网络的讨论与综述, 本文决定使用谱图神经网络中的切比雪夫层来处理生成的图数据。

### 3.2.1 基本定义

为了高效处理图数据, 我们将图估计器所生成的节点关系通过矩阵表示, 这类描述图节点之间关联的矩阵被称为邻接矩阵 (Adjacency Matrix), 用  $A$  来表示

$$A_{ij} = \{0, 1\}^{N \times N} \quad (3-46)$$

$A_{ij} = 1$  意味着节点  $i, j$  之间存在相关, 反之  $A_{ij} = 0$  意味着两节点之间无相关关系。我们同时定义无向图中节点的度 (Degree) 为与该节点相关节点的数量并定义图的度矩阵  $D$  为

$$D_{ii} = \sum_{i \neq j} A_{ij} \quad (3-47)$$

可以看出, 该度矩阵为一对角矩阵。同时为了后续矩阵的特征分解操作, 我们定义图的拉普拉斯矩阵  $L$  为度矩阵与邻接矩阵之差, 即

$$L = D - A \quad (3-48)$$

### 3.2.2 切比雪夫层

切比雪夫网络是一种谱图神经网络，其通过卷积方式来从数据中学习复杂信息 [?]。但是同常见的卷积神经网络不同，谱图神经网络通过傅里叶变换的方式来实现图这种非欧数据上的卷积。根据卷积定理 [?]，两个函数的卷积为他们傅里叶变换后的结果的点积的傅里叶逆变换，即

$$x * y = F^{-1}\{F\{x\} \cdot F\{y\}\} \quad (3-49)$$

因此如果我们能够在图上定义一傅里叶变换，我们就能以此定义图数据上的卷积。通常来说，函数的傅里叶变换是函数在拉普拉斯算子特征函数这一组标准正交基上的投影。类似地，我们根据图的拉普拉斯矩阵定义该变换。

$$Lu = \lambda u \quad (3-50)$$

其中  $L$  为图的拉普拉斯量， $u$  为该正交基， $\lambda$  为特征值。我们同时定义矩阵  $U = [u_1, \dots, u_N]$ ，使得

$$L = U^T \Lambda U \quad (3-51)$$

因此我们定义图  $\phi$  上的傅里叶变换为

$$\begin{aligned} \hat{\phi} &= U^T \phi \\ &= \text{diag}(\hat{\phi}(\lambda_l)) \\ &= \text{diag}\left(\sum_i^N \phi_i u_{li}\right) \end{aligned} \quad (3-52)$$

因此根据卷积定理，对于滤波器  $g$  与图信号  $x$  而言，在图  $G$  上定义的卷积为

$$\begin{aligned} (g * x)_G &= U(U^T g \cdot U^T x) \\ &= U g_\theta(\Lambda) U^T x \\ &= g_\theta(L)x \end{aligned} \quad (3-53)$$

其中  $g_\theta$  为卷积核在图中的傅里叶变换， $\theta$  为该卷积核的参数。因此我们可以定义谱图神经网络中的一层为

$$y = \sigma(g_\theta(L)x) \quad (3-54)$$

但是，此时  $g_\theta$  较难确定。Deffend[65] 等人依靠切比雪夫多项式对  $g_\theta$  如式3-55进行了近似。

$$g_\theta(\Lambda) = \sum_{k=0}^K \theta_k T_k(\tilde{\Lambda}), \quad \tilde{\Lambda} = 2\Lambda_n/\lambda_{max} - I_n \quad (3-55)$$

其中  $\theta$  代表需要在训练中学习的参数， $T_k$  代表  $k$  阶切比雪夫行列式， $\tilde{\Lambda}$  代表特征矩阵。其中切比雪夫行列式计算方式如式3-56。

$$\begin{aligned}
 T^{(0)} &= X \\
 T^{(1)} &= \tilde{L}X \\
 T^{(k \geq 2)} &= 2 \cdot \tilde{L}T^{(k-1)} - T^{(k-2)},
 \end{aligned} \tag{3-56}$$

因此切比雪夫网络中的一层表示为式3-57所示形式。

$$Y = \sigma \left( \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})x \right) \tag{3-57}$$

### 3.2.3 结构设计

基于切比雪夫层的特点，我们构建如表3-1的图神经网络。

表 3-1 图神经网络结构

层	功能	超参数	输出形式
切比雪夫层	图卷积	$k=1, Channel=64, activation='ReLU'$	$(200, 16)$
切比雪夫层	图卷积	$k=1, Channel=64, activation='ReLU'$	$(200, 16)$
读出层	将图卷积结果转为矩阵	-	$(3200)$
全连接层	学习图数据	$N=512, activation='ReLU'$	$(256)$
Drop-Out 层	抑制网络过拟合	Dropout rate = 0.5	
全连接层	学习图数据	$N=256, activation='ReLU'$	$(64)$
全连接层	学习图数据	$N=64,$ $activation='Sigmoid'$	$(1)$

对于切比雪夫层，我们首先选择参数  $k = 1$ ，这意味着该层只取图切比雪夫矩阵的前两项。由式3-56与式3-57我们可知，在该参数条件下，每个节点输出值至于该节点及其相邻节点相关。这与传统卷积神经网络中选取的  $(3 \times 3)$  卷积核功能类似。同时为了充分学习图中潜藏的信息，我们设定该层的通道数为 16，这意味着同时有 16 个卷积核在一图中学习，增强了模型对潜藏信息的挖掘能力。同时根据文献报道，我们选择如式3-58所示  $ReLU$  函数作为该层的激活函数。[? ]

$$ReLU(x) = \max\{0, x\} \tag{3-58}$$

在对图信息充分学习与处理后，我们将图数据再次通过读出层转化为数据矩阵，准备后续通过全连接层对疾病风险进行预测。考虑到读出层共产生 3200 个特征，我们选择  $512 - 256 - 64$  的全连接层组合以求最大限度提高模型性能。同时为了防止该规模下模型对数据的过拟合，我们加入 Dropout 层以在训练中随机屏蔽某些神经元。输出层的最后，本模型通过如式3-59所示  $Sigmoid$  激活函数来给出根据患者基因型信息的骨关节炎患病风险。

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (3-59)$$

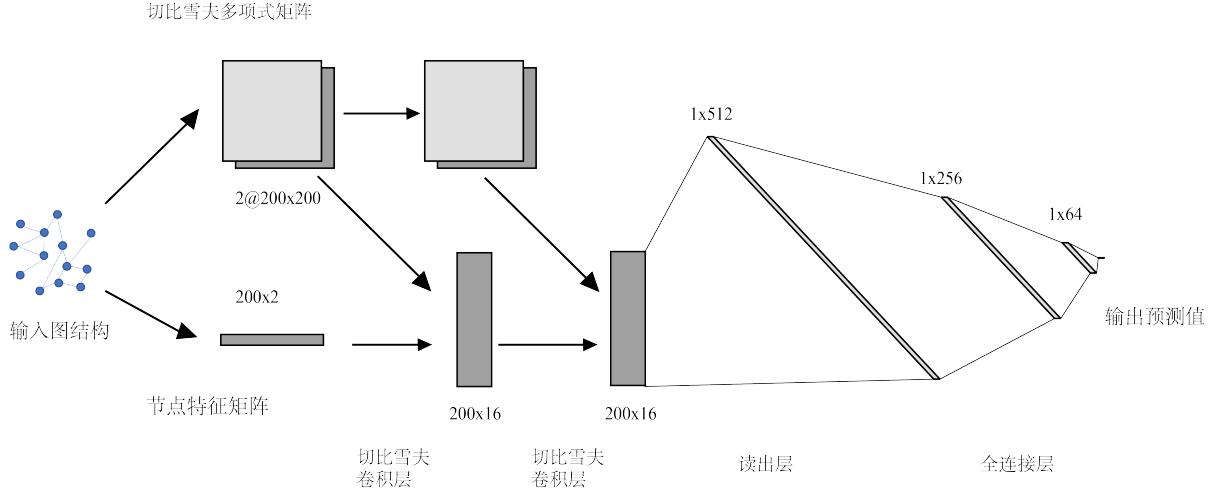


图 3-3 图神经网络架构

在完成了工作过程如图3-3的图神经网络构建后，我们将其通过算法2与已经搭建好的图估计器相融合。使其具备根据输入基因型信息给出风险预测的能力。

**Data:** 基因型数据  $X$

**Result:** 患病风险预测值  $y$ , 最优图结构  $G$ , 神经网络参数  $\Theta$

**Input:** 基因型数据  $X$

- 1 依据全连接法构建初始图结构  $G$
- 2 **while** 未到达设定迭代数 **do**
- 3     神经网络训练
- 4      $\Theta = \arg \min_{\Theta} l(y, G)$
- 5     图估计
- 6     根据图神经网络输出估计可能图结构并更新  $G$
- 7 **end**

**Output:** 输出最优图结构及图神经网络参数, 根据该参数给出风险预测

算法 3-2 图估计切比雪夫图神经网络

### 3.3 表型融合

考虑到骨关节炎作为一种复杂多基因疾病，单纯以基因型为输入可能导致风险预测效果不佳。因此有必要同时融合表型数据。根据 Boer[?] 等人的报道，我们选择同骨关节炎密切相关的表型做为协变量并从 UKB 数据库中获取相关表型数据。具体表型如表3-2。同时我们选择多模态数据融合中的中间融合法 [?] 将基因型数据与表型数据相

表 3-2 选取协变量表型

UKBID	表型
32883-31	性别
32883-48	髌围
32883-49	腕围
32883-2463	是否发生过骨折
32883-21001	BMI
32883-21002	体重
32883-21003	年龄

融合。我们将已构建的图神经网络的中间输出层与所获取的表型数据相融合，并输入新的分类器进行训练，最终给出融合表型后的预测结果。

### 3.4 图解释器

图是一种包含了节点信息与节点之间关系的数据结构，但是，本文目前为止建立的模型只能直接给出对患病风险的预测，与传统机器学习方法无异。如果能够充分利用图中的关联信息并对产生该预测结果的原因加以解释，不仅能够增强预测结果可信性，还能根据其解释内容发掘数据中潜藏的其他信息。目前对于图神经网络解释器的研究已有颇多进展，本文基于其中备受关注的 GNNExplainer [?] 来构建模型解释器。

GNNExplainer 通过输入的已训练图神经网络模型及其预测结果从输入数据中选择对预测结果影响最大的子图。其数学描述如下

$$\max_{G_s} MI(Y, G_s) = H(Y) - H(Y|G = G_s) \quad (3-60)$$

其中

- $MI$  表示两变量之间的互信息（Mutual Information），其代表两变量之间相互依赖的程度。
- $Y$  表示模型给出的预测结果
- $G_s$  表示同预测结果相关的子图
- $H$  表示变量的信息熵

该文献描述：对于给定的 GNN 模型， $H(Y)$  恒定，此时只需要最小化  $H(Y|G = G_s)$  就可达到解释目的。该文献同时描述了一种变分近似方法来实现该优化过程，最终生成了与预测结果相关的最小子图。

### 3.5 本章小结

本章描述了骨关节炎风险预测模型的构建过程与基本模块。本文首先针对现存基因型数据结构缺失的问题基于变分期望最大化算法构建了一个图结构估计器。该估计器能根据神经网络的输出动态更新与预测图结构，同时能够对图中的节点聚类分析。其次，本文根据谱图神经网络中的切比雪夫层构建了包裹卷积层、读出层、全连接层在内的图神经网络用于图数据的处理以及患病风险的预测。考虑到单纯通过基因型数据预测患病风险结果可能较差，本文还基于文献描述的同骨关节炎密切相关的表型构建了表型融合风险预测模型。最后，为了尽最大可能挖掘出图数据中的潜藏信息，本文还根据已有的图神经网络解释器构建了预测结果解释器，在给出风险预测结果的同时获取同预测结果相关的子图。以上模块共同构成了本文预期的骨关节炎风险预测模型。

## 4 结果讨论与分析

本章将对搭建的骨关节炎风险预测模型的图估计效果、患病风险预测能力进行分析与评估。同时将结合具体案例分析模型解释器所产生的解释结果。

### 4.1 评估指标

为了对模型的风险预测能力进行科学的评估，本文依照真实样本标签与预测结果构建如图4-1混淆矩阵（Confusion Matrix）并根据该矩阵计算如表4-1所示常用评价指标[? ]。

		True Class	
		Positive	Negative
Predicted Class	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

图 4-1 混淆矩阵构成

表 4-1 混淆矩阵相关指标

指标	计算公式	含义
准确率 Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	模型整体判断准确程度
召回率 Recall	$\frac{TP}{TP+FN}$	正确预测阳性样本占总阳性样本的比例
特异度 Specificity	$\frac{TN}{TN+FP}$	正确预测阴性样本占总阴性样本的比例
精确度 Precision	$\frac{TP}{TP+FP}$	正确预测阳性样本占预测结果为真的样本比例
F1 分数	$\frac{2*Precision*Recall}{Precision+recall}$	综合召回率与精确度评估模型预测准确程度

此外，在评估模型预测效果时还有另一种不依赖混淆矩阵的指标，本文使用了其中

的两种——接收器操作特性曲线下面积 (Area Under Receiver Operating Curve, AUROC) 以及柯尔莫哥洛夫-斯米尔诺夫指标 (Kolmogorov-Smirnov Statistic, K-S Statistic)。接收器操作特性曲线 (ROC) 是用来评价二元分类器分类能力的一种曲线，其以真阳性率为纵轴假阳性率为横轴绘制模型不同阈值下真假阳性率的坐标。其曲线与横轴所围成面积被称为曲线下面积 (AUC)，对于一分类模型而言，AUC 越高表示其做出正确判断的能力越强。<sup>[?]</sup> 并且同准确率相比，AUC 评价时不受训练集正负样本比例所影响。<sup>[?]</sup> 因此本文使用该指标衡量模型预测能力。具有相似功能的指标还有 K-S<sup>[?]</sup>，它是一种用来检验两经验分布之间是否相同的统计量，K-S 量越大表明模型的区分能力越强。

## 4.2 模型预测效果评价

### 4.2.1 基准计算

在对本模型效果进行分析前，我们首先需要计算使用传统 PRS 方法时骨关节炎的风险预测效果。PRS 方法，又称多基因风险评分 (Polygenic Risk Score) 方法<sup>[?]</sup>，是一种评估 SNP 位点对某一特定性状的累计影响的量化指标。它由已有 GWAS 研究结果中 SNP 对表型影响的效应值以及统计显著性构建。其计算公式为如式4-1

$$S = \sum_i^N X_i \beta_i \quad (4-1)$$

其中  $S$  为个体的风险评分， $N$  为总潜在 SNP 的数量， $X_i$  为第  $i$  个 SNP 中潜在效应等位基因的数量， $\beta_i$  为第  $i$  个 SNP 对形状的效应值。我们根据该公式对现有基因型数据进行计算得到个体评分，再通过逻辑回归法构建基于 PRS 的风险预测模型并得到如图4-2-a 所示 ROC 曲线。可以发现，传统 PRS 模型的风险预测效果较差，AUC 仅为 0.51，与随机预测相当。

同时我们也使用了包括决策树算法在内的若干传统机器学习模型对基因型数据进行处理与预测，所得结果如图4-2-b 与基于 PRS 的模型无显著差异。我们还通过混淆矩阵相关指标对两种模型进行评估，得到表4-2。从表中我们可以看出两模型预测性能均较差，这意味着使用 PRS 模型与传统机器学习模型对骨关节炎的预测几乎没有任何实际意义。而这也与 Boer<sup>[5]</sup> 研究中构建的 PRS 模型结果相符。因此本文以该 PRS 模型为基础讨论本文提出模型对预测效果的改善。

### 4.2.2 特征筛选方法对比

之前的讨论中我们提到，特征筛选对于特征数与样本数之比较高的数据集有着很好的提高模型性能的效果。本文也因此提到了两种特征筛选方法——基于卡方的样本筛选与基于支持向量机的样本筛选。对于使用同样超参数的决策树模型，我们输入使用不同特征筛选方式处理后的数据集并比较两种特征筛选方法对模型效果的影响。我

表 4-2 混淆矩阵相关指标

指标	基于 PRS 的模型	决策树模型
auc	0.51	0.50
f1	0.45	0.29
accuracy	0.52	0.51
precision	0.50	0.50
recall	0.41	0.21

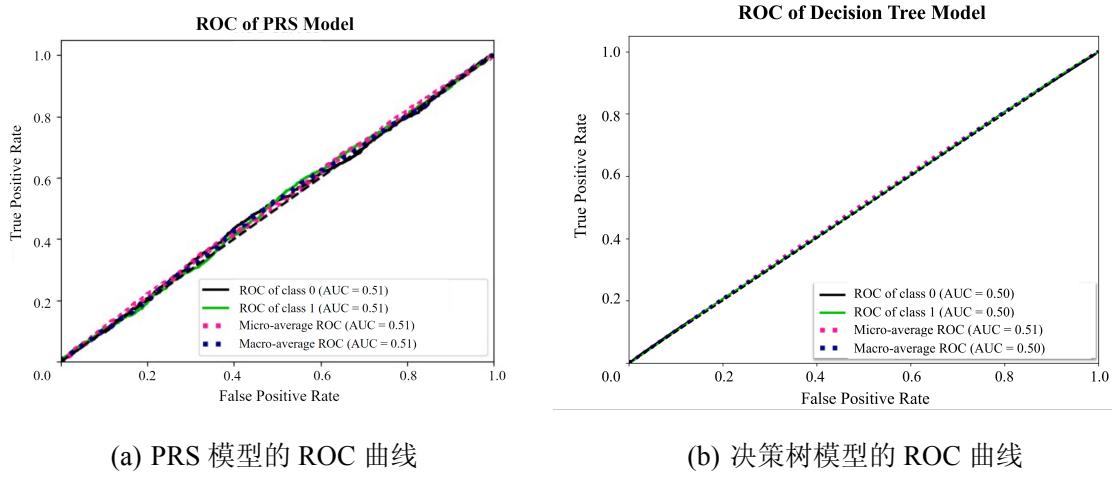


图 4-2 基准模型 ROC 曲线

们首先通过混淆矩阵相关指标对该结果进行评估，结果如表4-3。

表 4-3 特征筛选方法对比

指标	经卡方法筛选特征	经支持向量机法筛选特征
f1	0.38	0.43
accuracy	0.52	0.54
precision	0.51	0.55
recall	0.29	0.35

可以看出，使用两者特征筛选方法所得结果虽然均同 PRS 法没有明显提高，但是相比较而言支持向量机法筛选特征在决策树模型中的效果更好。我们又对两种方法所得数据在决策树模型中的性能通过非混淆矩阵指标进行评估，得到图4-3。图中我们可以看到，支持向量机法筛选所得数据训练出的模型在 AUC、KS 指标方面均优于通过卡方法筛选所得数据，本文因此使用支持向量机法筛选特征。

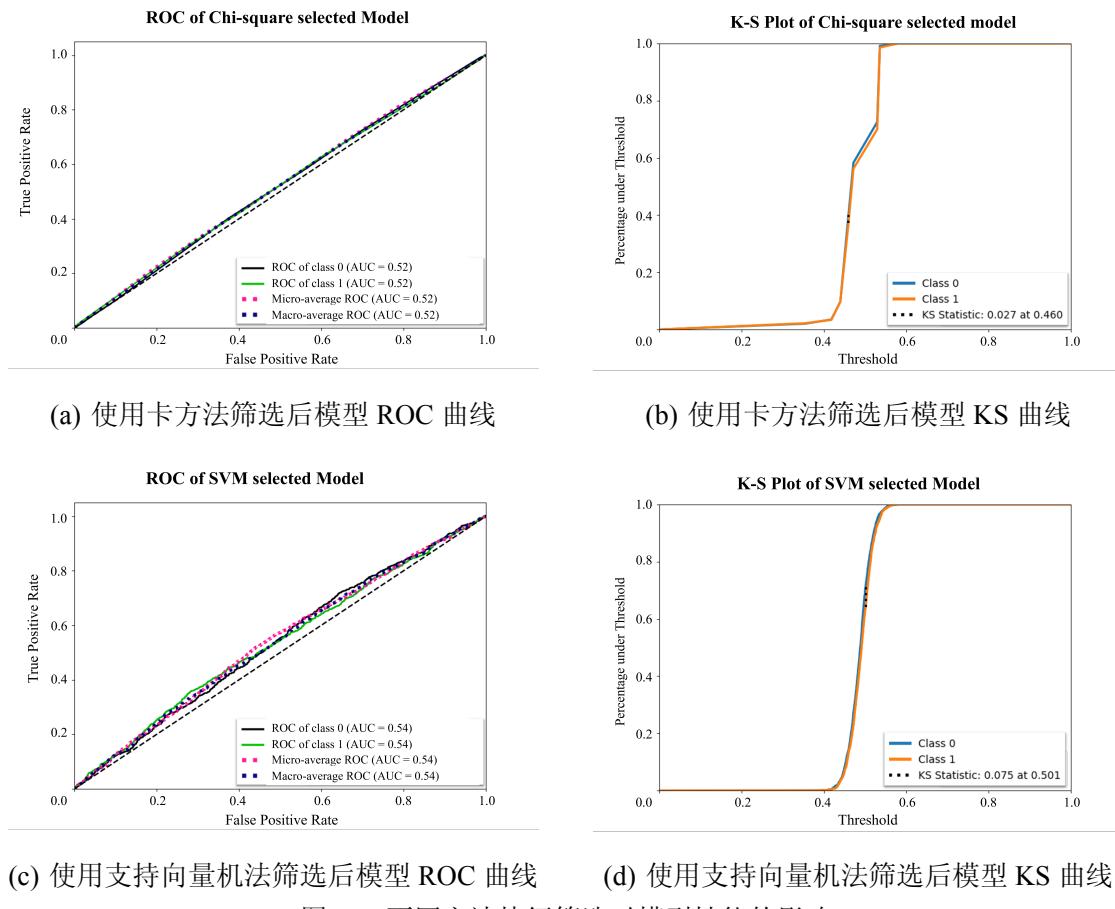


图 4-3 不同方法特征筛选对模型性能的影响

### 4.2.3 图神经网络处理

将经过支持向量机法筛选所得特征输入本文提出的风险预测模型中进行图估计与图神经网络处理，选择最优的图估计结构，并完成参数调优后，本文训练出根据个体基因型数据完成骨关节炎风险预测的图神经网络模型。我们首先通过混淆矩阵相关指标对该结果进行评估，结果如表4-4。

表 4-4 图神经网络处理

指标	图神经网络预测结果	PRS 模型
f1	0.56	0.45
accuracy	0.60	0.52
precision	0.60	0.50
recall	0.53	0.41

根据该结果我们发现，训练所得图神经网络在预测准确性、f1 分数、精准度、灵敏度上均优于传统 PRS 模型。我们还通过非混淆矩阵指标对图神经网络性能加以评估，

得到图4-4。

可以看出，本文建立的图神经网络较 PRS 模型在 AUC 方面有了明显改善。同时模型在模型 Precision-Recall 曲线中我们也发现随着召回率的提升，模型准确率下降速度较慢这也预示着模型分类性能较好。在模型 Lift 曲线中，随着深度增加，模型 lift 值在深度较大时下降较快，意味着模型良好的分类性能。综上，本文使用的单纯依靠基因型数据的图神经网络模型相较于传统模型而言有了明显的性能进步，然而，该模型绝对性能依然不如人意。

#### 4.2.4 表型融合

为进一步增强模型性能，本文引入同骨关节炎密切相关的表型参与风险预测，该模型其给出如表4-5混淆矩阵相关指标结果。

表 4-5 融合表型信息对模型性能的影响

指标	融合表型	单纯图神经网络	PRS 模型
f1	0.66	0.57	0.45
accuracy	0.67	0.58	0.52
precision	0.66	0.57	0.50
recall	0.69	0.58	0.41

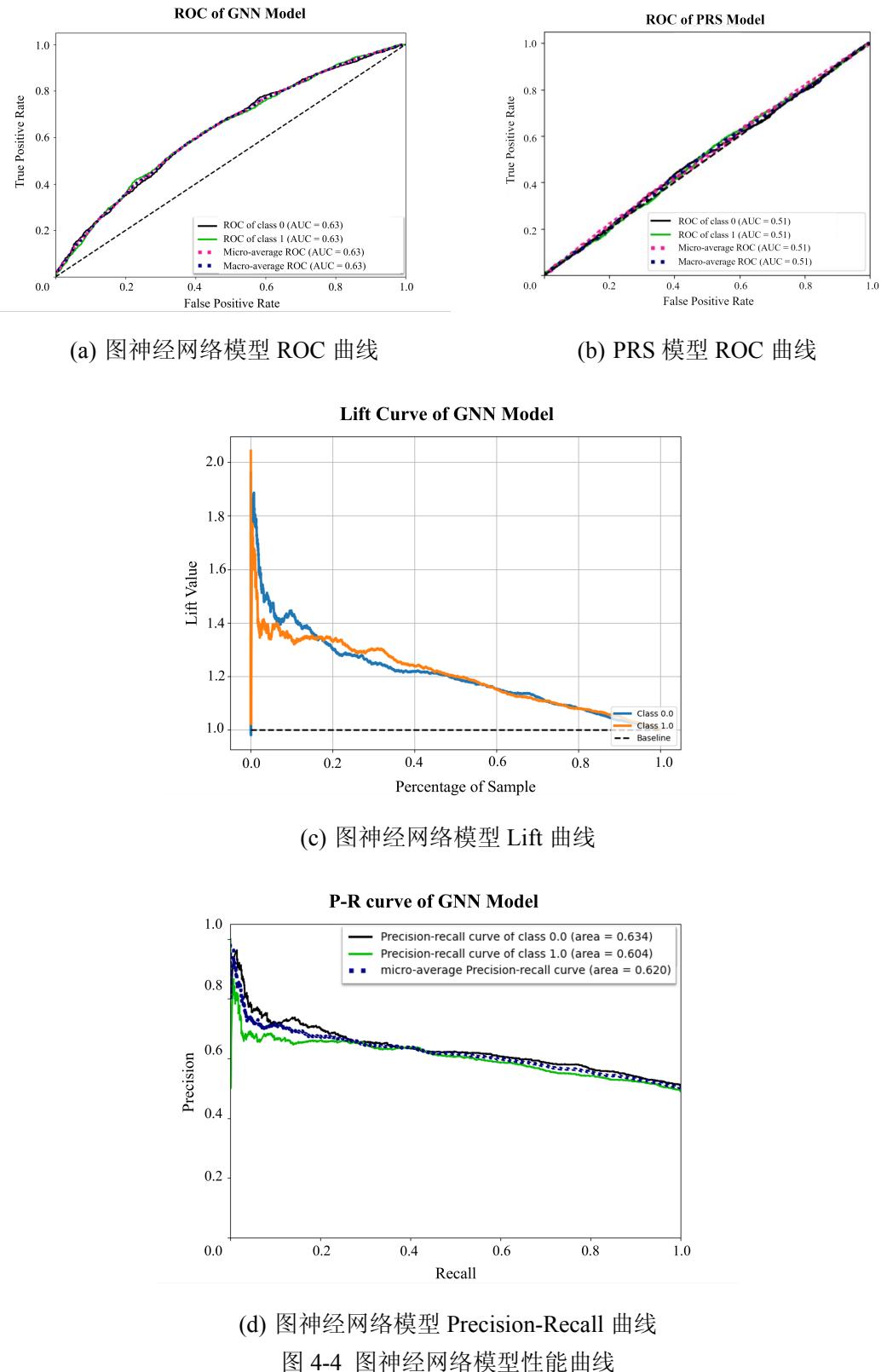
可以看出融合表型之后模型性能相较于单纯使用基因型数据的图神经网络模型有了进一步提升。我们还对该模型性能的非混淆矩阵指标进行了计算并同图神经网络模型加以比较。在融合表型模型中，预测 AUC 已经达到 0.74，相较于单纯利用基因型数据的图神经网络预测准确能力显著增强。同时模型的 KS 值（0.356）相较于单纯利用基因型数据的图神经网络 KS 值（0.202）也有明显提高。Precision-Recall 曲线方面，融合表型模型的 PR 曲线下降较单纯使用基因型的图神经网络模型更慢，意味着模型良好的分类性能。而在 Lift 曲线中，融合表型模型 Lift 曲线较图神经网络模型曲线更为陡峭，也提醒模型预测性能的增强。综上，通过表型信息的融合，本文提出的骨关节炎风险预测模型较传统模型相比取得了可观的性能提升，已经具有较好的风险预测能力，可以应用于实际的风险预测之中。

### 4.3 图估计效果

#### 4.3.1 估计器在训练过程中的效果

##### 4.3.1.1 可行性验证

为了验证本文设计图估计器对无结构数据的处理能力，在使用该估计器估计本研究所使用骨关节炎患者基因型数据结构之前，本文先在 MNIST 数据集上进行测试。MNIST



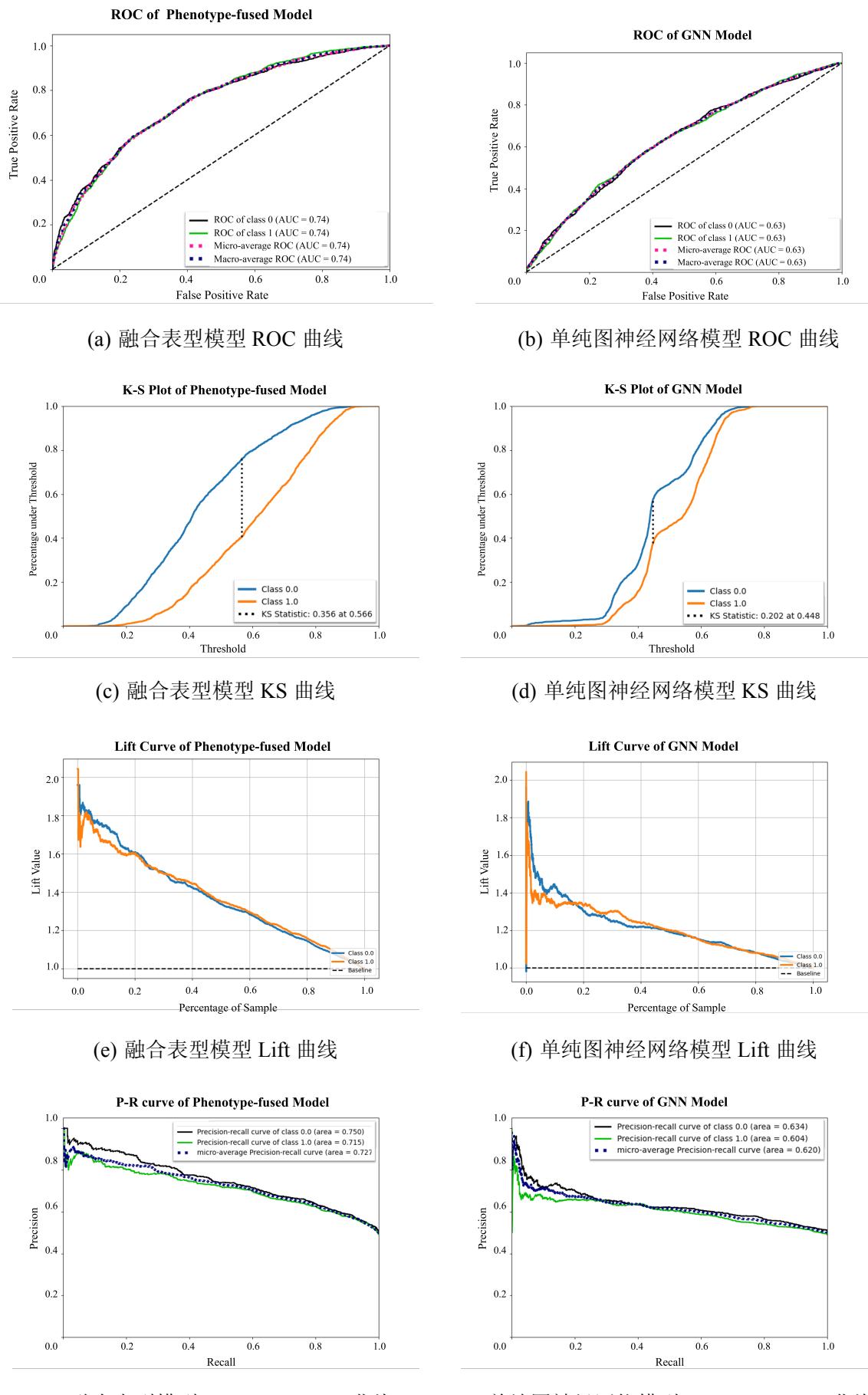
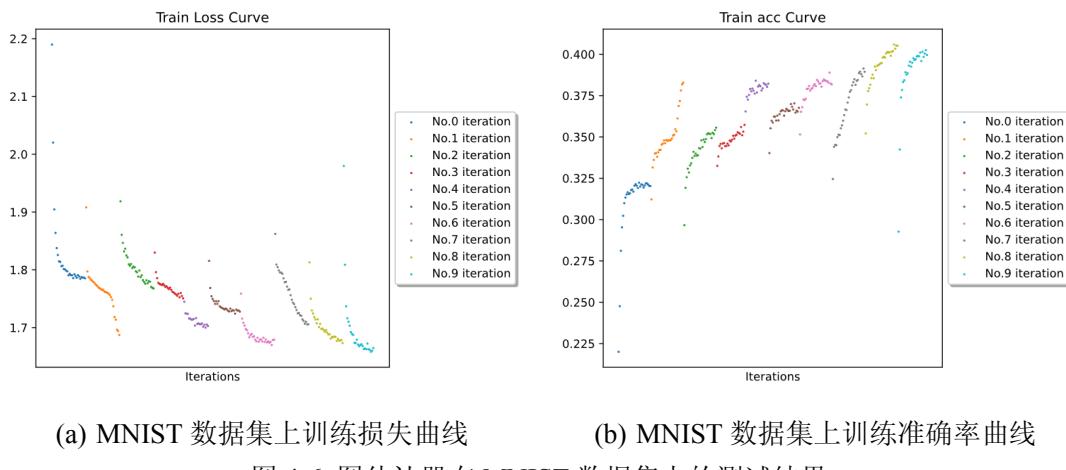


图 4-5 融合表型信息对模型性能的影响

数据集 [?] 由若干已标注手写数字图片组成，也是一种无结构数据。我们因此在该数据集上应用我们的模型并将结果加以记录如图4-6。可以看到，在每个迭代的图估计器工作之后，模型损失有着明显下降，同时分类准确率也有显著提升。这证明了本文提出图估计器对图神经网络在无结构数据上的性能有着明显改善，为后续对骨关节炎患者基因型数据的处理工作提供可行性基础。



#### 4.3.1.2 对照组

图估计器的工作过程涉及到图神经网络参数与图结构的循环更新。为了排除多次重启训练对模型性能的潜在影响，我们设计对照组，即使用全连接法构建图矩阵并且每次图神经网络训练之后不更新图结构。并对对照组的模型训练情况做以记录如图4-7。可以看到，在不更新图结构时，每次迭代中模型损失与准确率的变化趋势基本一致，同时不同迭代末期模型 AUC 变化不大。因此我们认为模型效果与无图结构更新过程的迭代次数无关。

#### 4.3.1.3 实际数据

该组试验中我们正式将图估计器与图神经网络运用于骨关节炎患者基因型数据的处理过程中。我们首先使用全连接法构建初始图，在每个迭代中我们训练神经网络并给出数据在该网络下的输出，再将输出作为图估计器的观察值并基于此更新图结构，再将该图结构作为数据的结构重新输入图神经网络中训练。如此迭代若干次，记录模型性能如图4-8。

可以看出，首次迭代中由于使用全连接网络，模型性能变化同对照组中模型相仿。但是在首次迭代结束生成新图结构后并以此进行第二次迭代的图神经网络训练时我们可以发现，训练末期模型损失明显下降，模型准确率明显上升。模型预测能力 AUC 也有明显提高。最终我们选择预测能力最好的迭代作为最优模型，参与到骨关节炎风险预测过程之中。

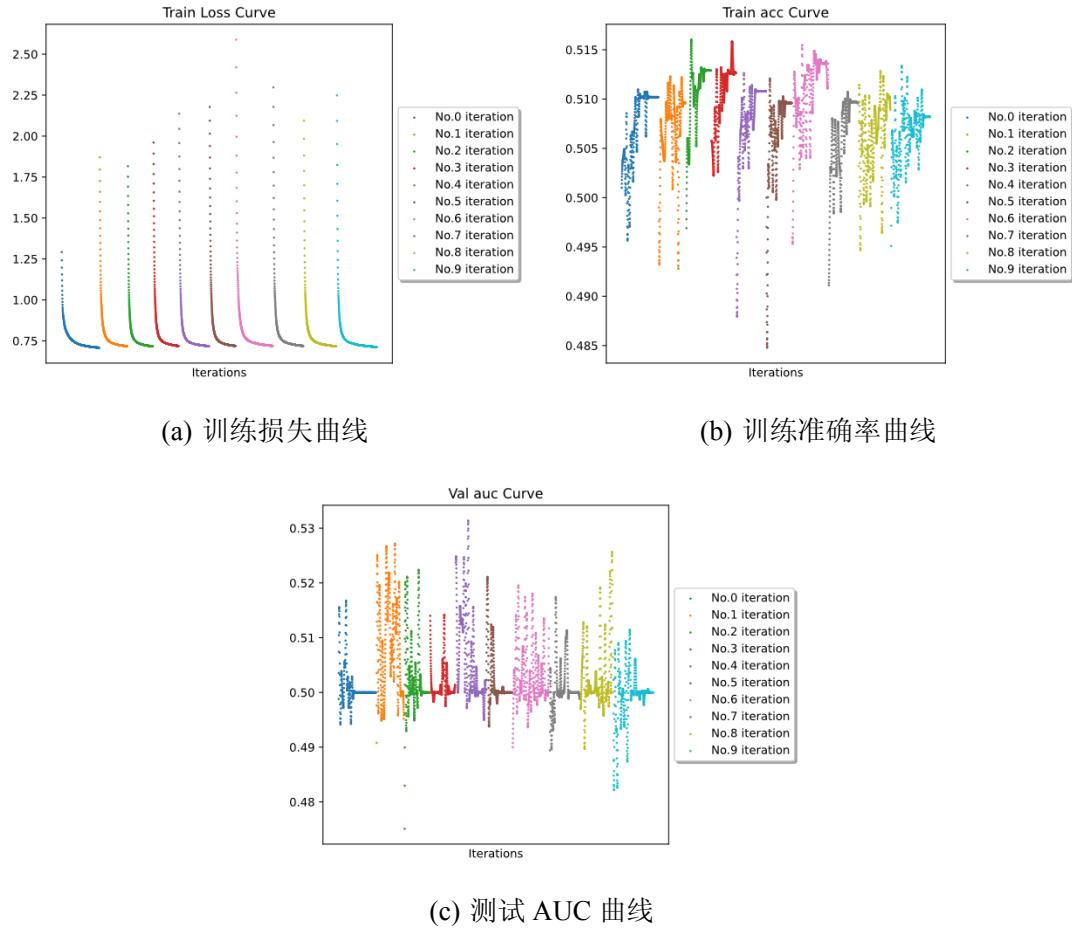


图 4-7 对照组训练结果

### 4.3.2 图估计器工作过程分析

#### 4.3.2.1 聚类参数选择

在讨论图估计器时我们提到，图估计器在输入观察值的同时还需要人工选择聚类参数  $k$  以初始化变分参数，且该聚类参数会对估计结果产生直接影响，该影响可通过整合分类似然评估。为了选择最优聚类参数，我们计算了使用一定范围内的聚类参数时的模型整合分类似然，结果如图4-9。

可以发现， $k$  取 6 时模型的整合分类似然最低，性能最好。同时我们考虑了在使用该范围内参数时变分期望最大化算法所给出的证据下界 ELBO，训练中 ELBO 的变化如图4-10。我们发现虽然  $k$  取 6 的组证据下界并非最高，但是同其他组相比也处于高位。因此我们将  $k$  定为 6 并估计该参数下可能的图结构。

#### 4.3.2.2 生成图结构分析

根据前文对变分期望最大化算法的讨论我们可以发现， $k$  取 6 意味着最终生成的图结构可以分为 6 簇。为了探究该网络结构以及其潜藏的信息，我们将估计图结构中的

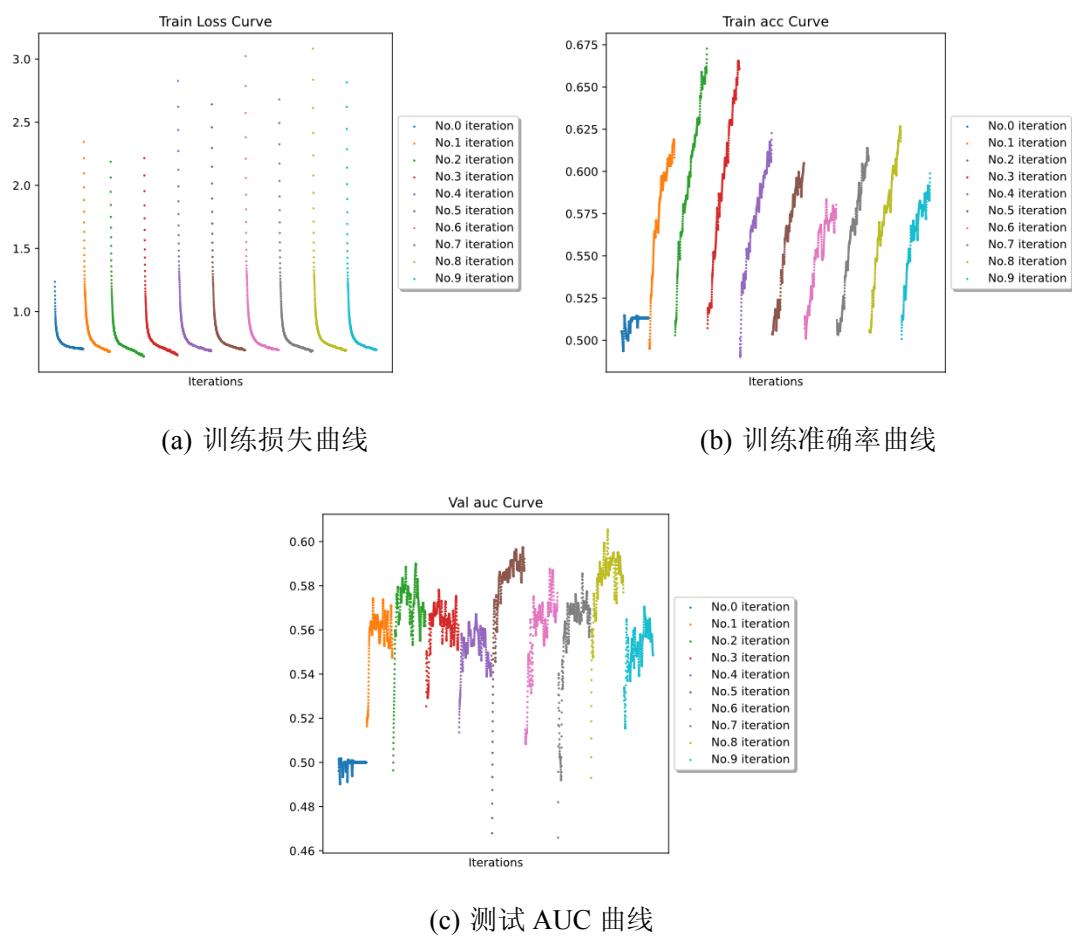


图 4-8 实际数据训练结果

ICL curve on different initial clusters

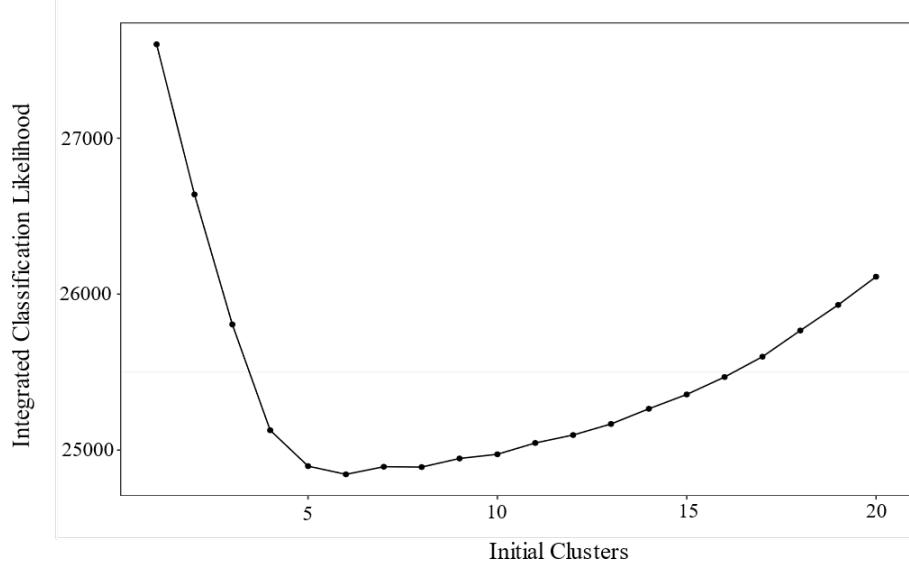


图 4-9 整合分类似然变化曲线

节点依照其所属的簇加以归类并表示如图4-11。

我们同时将计算所得的簇与簇之间的关系表示如图4-12。可以发现，簇3与簇4、

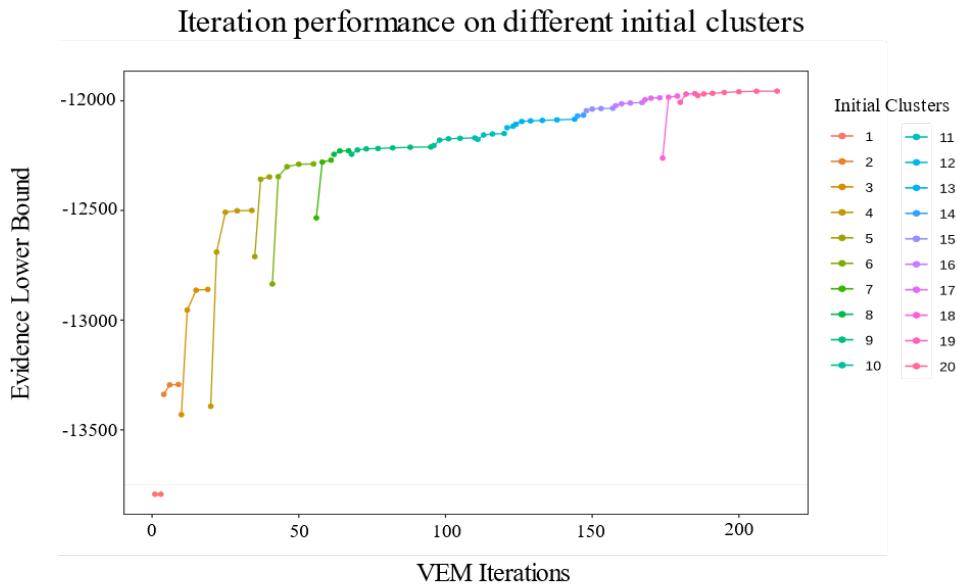


图 4-10 证据下界曲线

### Estimated Graph Structure

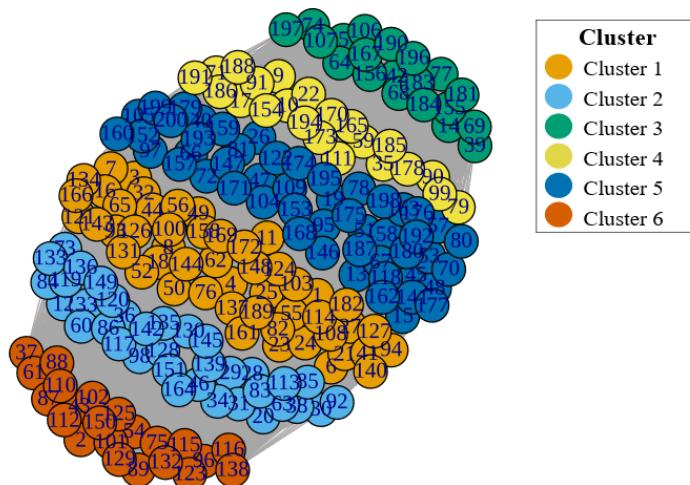


图 4-11 节点簇信息

簇 2 与簇 6 之间存在着很强的相关信号。且簇 6 与除簇 1, 2 之外的其它簇关联都不强。我们同时发现簇 3、5 之间也存在一相关信号。除此以外数学分析无法从该图中获取更多信息，我们因此需要结合图中节点的生物学意义对该图结构与簇关系加以分析。

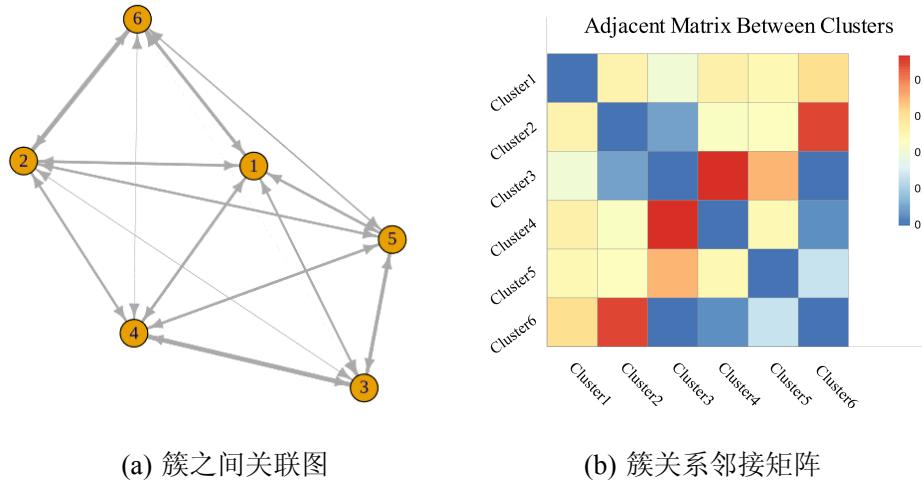


图 4-12 簇之间关系

#### 4.3.2.3 结合表型分析

为了进一步探究生成图结构的生物学意义，我们需要了解其中节点，也就是 SNP 位点对人的不同表型的贡献。同 GWAS 研究类似的，全表型关联研究（Phenome-wide Association Study, PheWAS）主要研究某一 SNP 位点与已知性状的关联 [? ]，因此我们对图中每个节点进行 PheWAS 研究。我们对节点的研究主要在 GWASATLAS 数据库内进行 [? ]，该数据库包含了在 UKB 数据上进行的对 3302 个表型的约 4756 个 GWAS 研究，可以较为全面地展现 SNP 位点对表型的影响。对于每个 SNP，我们编写爬虫由数据库获取与其相关的表型以及该相关的统计学显著性，经过统计学显著性筛选 ( $p < 10^{-5}$ ) 后将与每个 SNP 相关的表型按照领域分为 15 类。对于每个簇，我们统计簇内 SNP 位点对每类性状的关联并依照式4-2进行簇内标准化，得出如图4-13所示热图。

$$x = \frac{x - \mu}{\sigma / \sqrt{n}} \quad (4-2)$$

可以看出，每个簇内与簇内 SNP 相关联程度最高的表型均集中在代谢与免疫领域，这也与骨关节炎本身炎症性质相关。为了能进一步分析簇间差异，我们还依照式4-2对每个领域内关联进行簇间标准化，得到如图4-13所示热图。

相较于簇内标准化，我们可以从此图中发现簇间的明显差异，尤其是簇一、簇五在环境、内分泌、免疫等领域表型关联同其他簇存在明显差异。我们对六簇在这三个领域中所观察到的关联事件绘制提琴图，得到图4-15。提琴图中的簇间差别也印证了在热图中观察到的簇一、簇五的明显差别，因此本文对该两簇进行深入研究。

我们首先对簇一内布关联性状分布、数量、统计显著性进行分析并得出图4-16。通过该组图可以看到同其他簇相同，与簇一内 SNP 位点相关表型主要集中在代谢与免疫领域。但是我们从热图中发现簇一有着较强的环境相关信号。我们对该环境相关信号加以统计，结果如表4-6所示。出乎我们意料的是，与簇一内 SNP 相关的环境相关表型

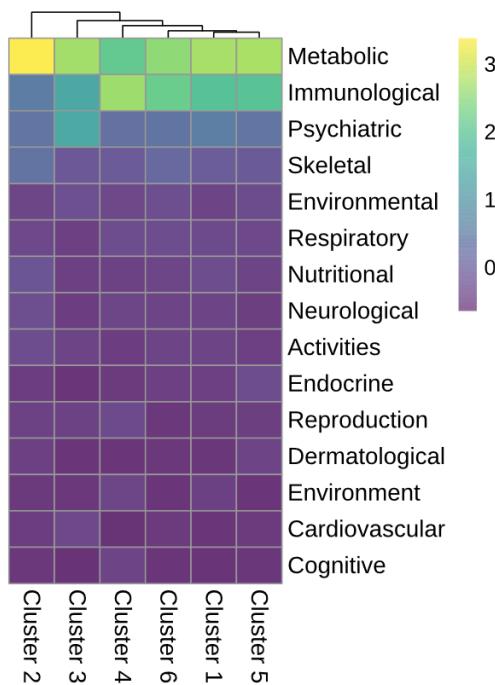


图 4-13 簇与性状关联簇内标准化热图

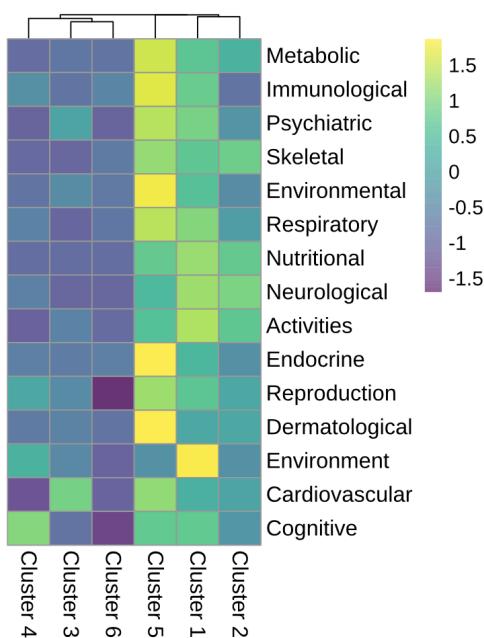
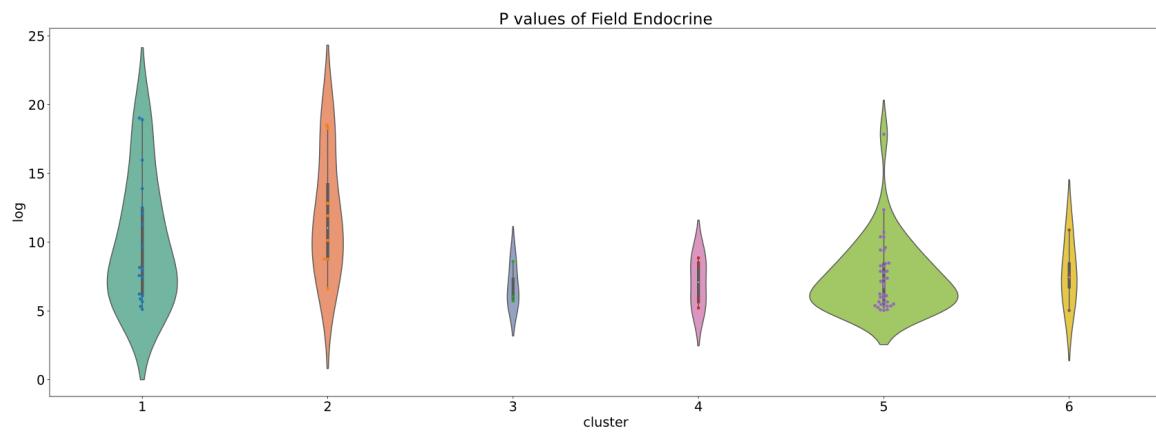
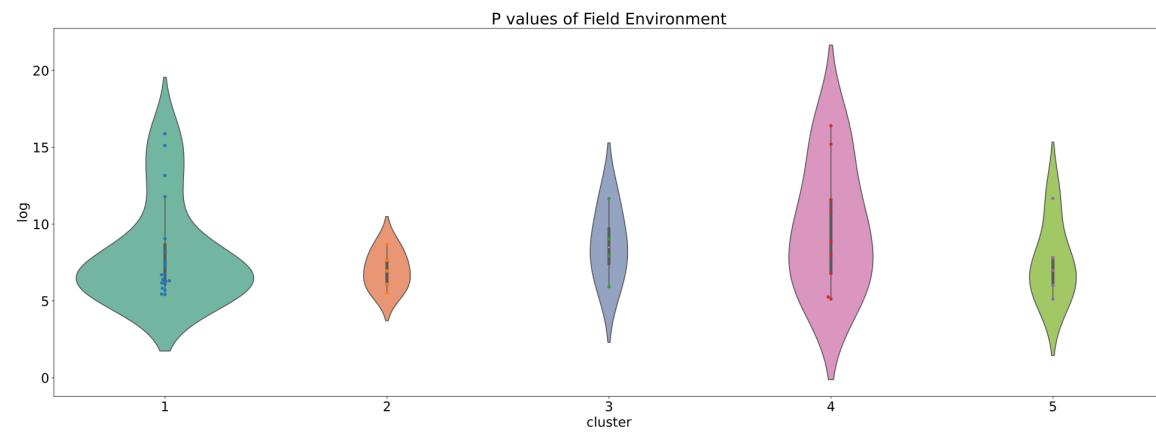


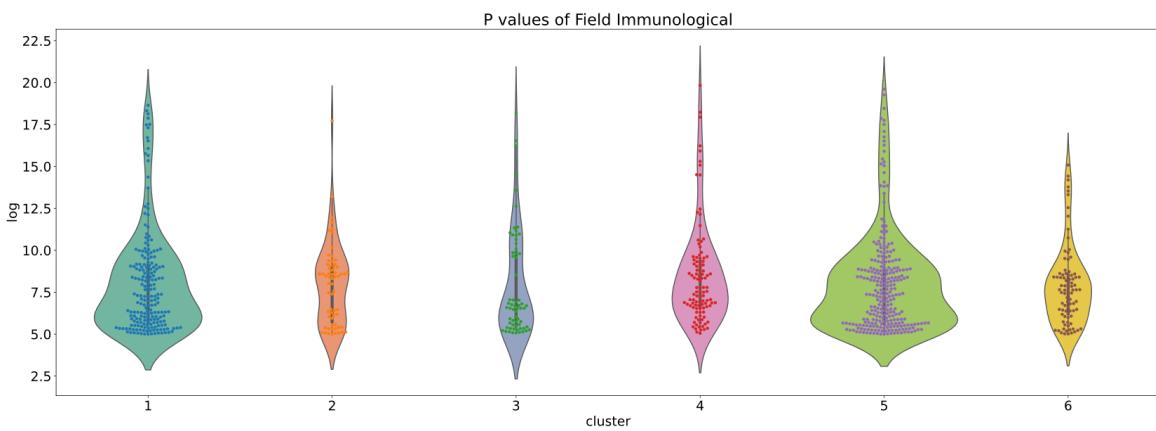
图 4-14 簇与性状关联簇间标准化热图



(a) 内分泌领域表型关联分布提琴图

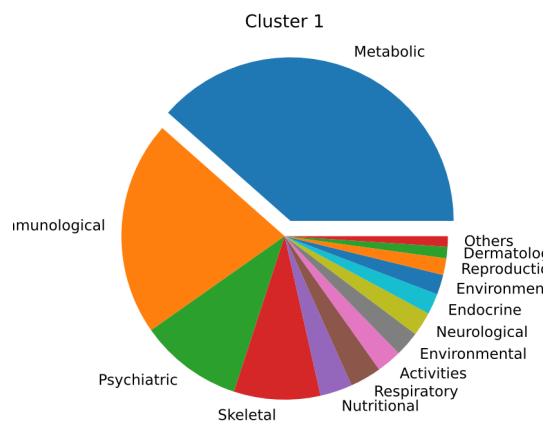


(b) 环境领域表型关联分布提琴图

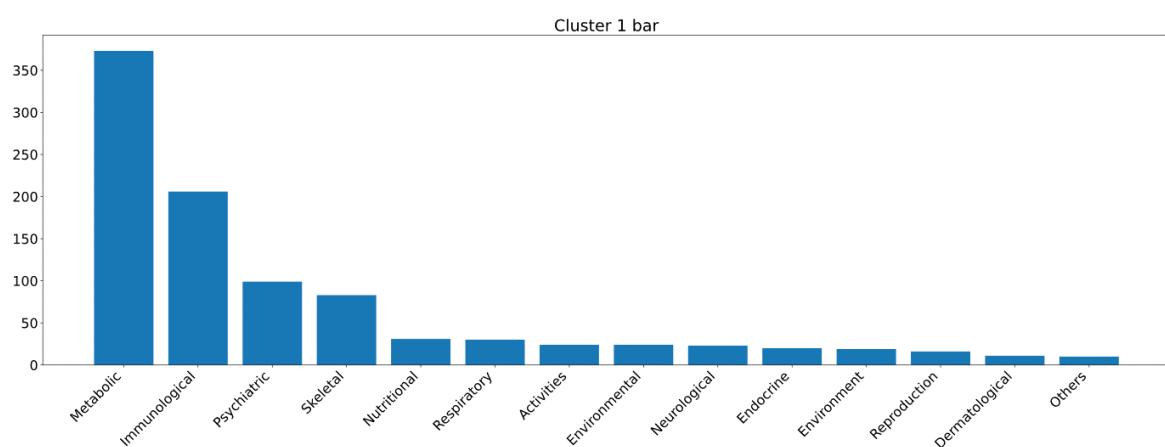


(c) 免疫领域表型关联分布提琴图

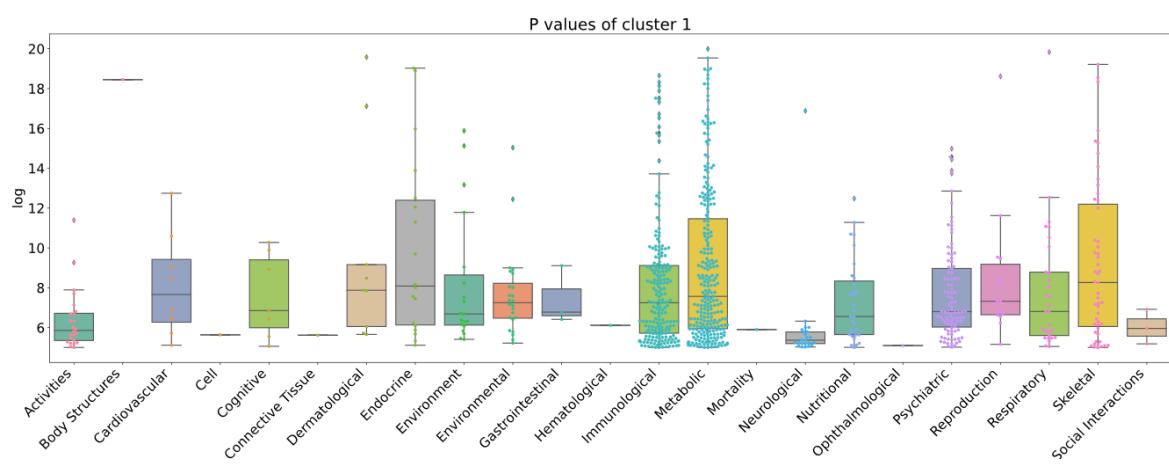
图 4-15 不同领域性状关联分布提琴图



(a) 簇相关性状分布饼图



(b) 簇相关性状分布柱图



(c) 簇相关性状 p 值图

图 4-16 簇一相关性状分领域图

出现了受教育程度、工作是否涉及重体力劳动等看起来与骨关节炎毫不相干的性状。但是，在查阅文献 [?] 后我们了解到，骨关节炎作为一种退化性疾病有可能与关节过度磨损相关。因此我们做出猜测，受教育程度较低的个体趋向于从事体力型劳动，导致关节加速磨损，最终导致骨关节炎的发生。而该簇内的 SNP 可能在受到这种环境影响时更易导致骨关节炎的发生。我们猜测该簇内 SNP 可能与环境因素导致关节磨损继而导致的骨关节炎相关。

表 4-6 簇一相关表型

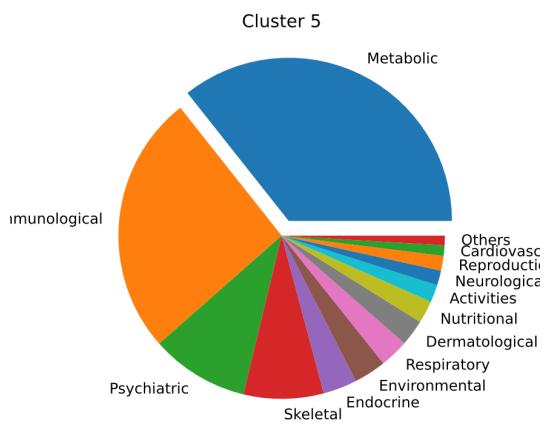
领域	表型	计数
Environment	Educational attainment	10
Environment	Education - Qualifications	3
Environment	Attendance/disability/mobility allowance: Blue badge	1
Environment	Job involves heavy manual or physical work	1
Environment	Maternal smoking around birth	1

我们也对簇五进行了类似分析，得到图4-17. 相较于其他簇，簇五在免疫、内分泌、皮肤病领域有着很强相关。我们对这些领域内的典型性状加以统计得到表4-7。可以看出，簇五内的 SNP 与二型糖尿病、巨噬细胞、白细胞与粒细胞计数相关表型都存在着相关。二型糖尿病是一种由于机体胰岛素抗性而生成的糖尿病，其主要由超重乃至肥胖而导致。[?] 而肥胖导致的高血脂又会诱发体内免疫系统包括巨噬细胞与粒细胞的激活。[?] 这也与簇内的观察一致，因此我们猜测簇五内的 SNP 位点主要与个体自身肥胖导致的糖尿病与超重继而诱发的自身型骨关节炎相关。

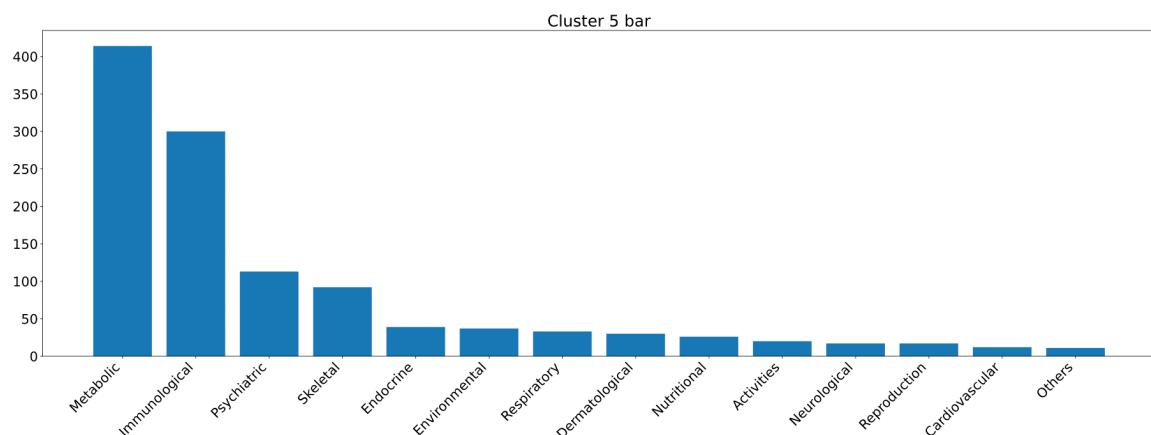
表 4-7 簇五相关表型

领域	表型	计数
Endocrine	Type 2 Diabetes	25
Immunological	Myeloid white cell count (three-way meta)	14
Immunological	White blood cell count (three-way meta)	14
Immunological	Granulocyte count (three-way meta)	13
Dermatological	Male pattern baldness	10

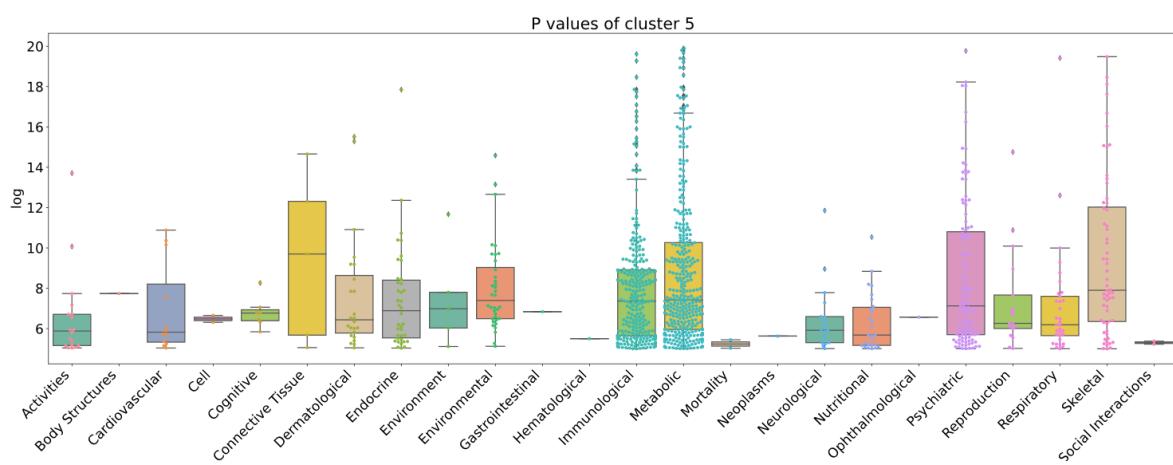
以上我们通过对图估计生成图结构的分析将输入的致病 SNP 分为六簇并对其中两个典型簇进行具体生物学意义分析。同时根据两簇内 SNP 的相关性状提出了骨关节炎



(a) 簇相关性状分布饼图



(b) 簇相关性状分布柱图



(c) 簇相关性状 p 值图

图 4-17 簇五相关性状分领域图

的两种诱因，即由环境因素导致关节过度磨损诱发的骨关节炎以及由于超重诱发的自身型骨关节炎。而这类分析是传统机器学习模型所无法实现的，这也进一步印证了图神经网络在学习数据深度信息时的出众能力。

#### 4.4 图解释器结果案例分析

我们之前提到本风险预测模型中的图解释器能对产生预测结果的原因，即输入基因型图中同预测结果相关的最大子图，加以分析与计算。本节通过模型对 UKB 数据库中一个案的分析结果展示该能力。本文从 UKB 数据库中选取样本 (UID=5125713)，并将其基因型数据根据图估计器给出的最佳图结构转为图数据。利用训练好的图神经网络模型对该图进行处理，预测结果为该样本有很高概率患骨关节炎。我们将该预测结果与模型输入图解释器中，最终图解释器给出如图4-18相关子图。

Subgraph from GNN Explainer

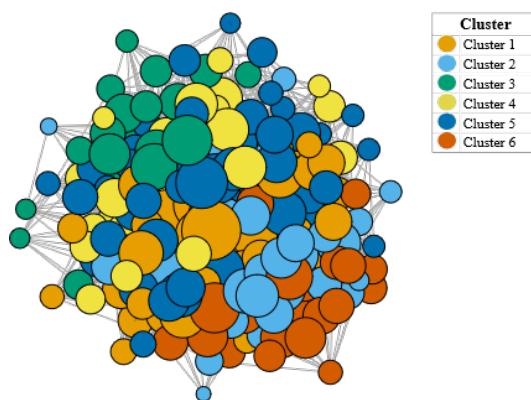


图 4-18 图解释器生成与预测结果相关子图

在该子图中，边的权重为该关联对预测值的贡献，权重越高代表该边对预测值的生成越“重要”。我们将该图按照边的权重进行稀疏化，得到如图4-19系列子图。可以看到，随着权重阈值的不断提升，子图的骨干逐渐显现，而该主干则主要由簇一、五中的节点构成。上一节的分析中我们提到簇五中所含位点可能与二型糖尿病相关，我们因此推测该个体可能由肥胖造成的关节过度磨损诱发骨关节炎，同时该患者可能合并二型糖尿病。我们从 UKB 数据库中提取患者其他表型得到表4-8。

患者实际 BMI 为 35.2，体脂率为 46%，已为肥胖体型，并且合并确诊糖尿病。可以看出，患者实际情况同我们根据图结构以及图解释器做出的解释一致。体现了本文提出模型除高效预测骨关节炎风险之外对疾病分型以辅助治疗的能力。

#### 4.5 本章小结

本章对本文搭建的骨关节炎风险预测模型的预测准确性通过若干指标加以评估并同现有的常见疾病风险预测模型算法进行比较，发现本文搭建模型较传统 PRS 方法与机器学习算法而言在准确性上有着显著提高，已可作为临床辅助工具参与骨关节炎的

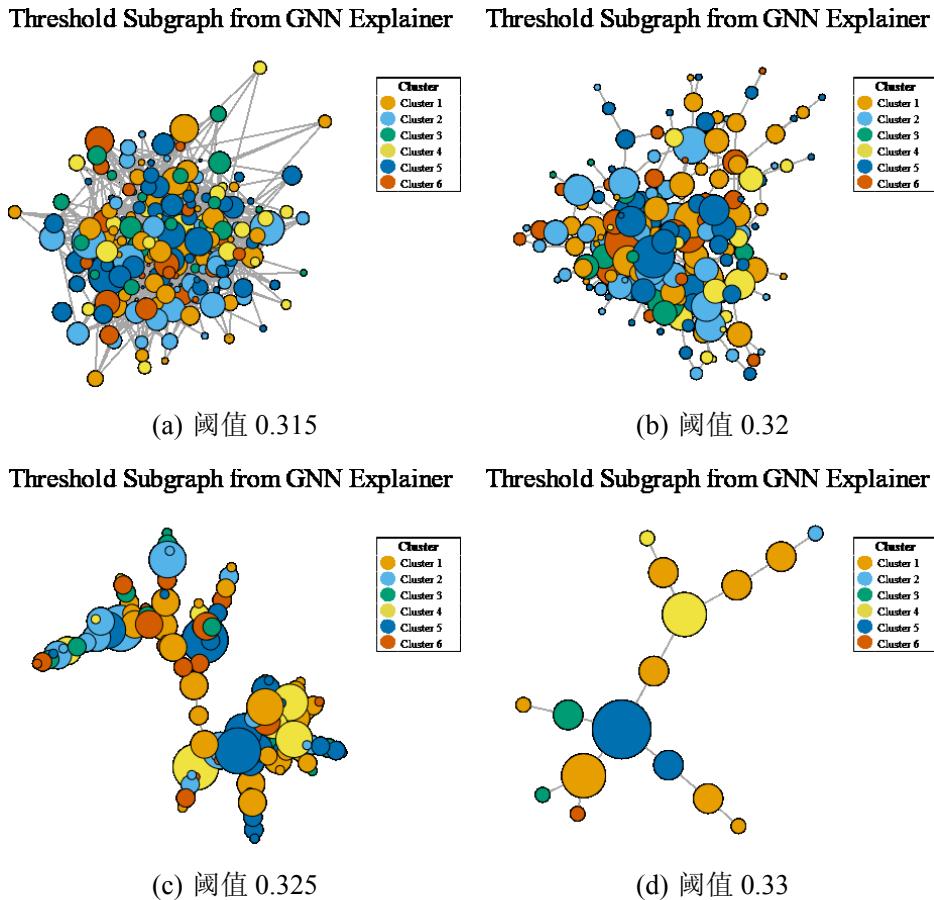


图 4-19 子图阈值化结果

表 4-8 该个体其他表型

UKB-ID	表型	值
31	性别	男
48	腰围	113
49	胯围	121
2443	确诊糖尿病	是
21001	BMI	35.2
23098	体重	105.5
23099	体脂率	46.1

早期诊断。本章同时对图估计器的工作过程加以记录与展示，一方面验证了算法的可行性；另一方面也证实了图估计器可以显著改善无结构数据在图神经网络中的表现。为了进一步挖掘潜藏在估计器所得图结构中的信息，本章还结合图估计器输出数学模型与图中节点的 PhEWAS 研究结果对图中的簇聚节点之间关系加以分析并且识别出了两个典型簇，并且根据典型簇内节点信息提出了关节炎的两种诱因。最后本章利用模型解释器进行案例分析，在给出预测结果的同时还利用解释器的结果与识别出的典型簇

预测了疾病的分型。以上内容完整展现了本模型作为高效准确可解释骨关节炎风险预测模型的功能与性能。

## 5 总结与展望

### 5.1 工作总结

构建高效可解释骨关节炎对骨关节炎的早期诊断与防治有着重要的意义。本文首先根据目前已发表的 GWAS 研究及公共数据库 UK BioBank 获取患者表型与基因型数据并进行数据预处理；之后构建了包括图结构估计器、图卷积神经网络、表型信息融合与图解释器四个模块在内的骨关节炎风险预测模型；最后本文对该模型的性能以及预测结果进行了进一步的分析和解读，继而对模型的预测准确性，可解释性等指标进行评估。本文对主要完成的工作总结如下。

1. 构建了适用于图分类问题的图结构估计器：目前提出的用来处理无结构数据的图结构估计器多适用于图节点分类问题，对适用于图分类问题的图结构估计器研究较少。本文提出了一种基于变分期望最大化算法的图结构估计器，该估计器通过对图神经网络输出的处理通过参数估计的方法预测图节点之间关联可能性与图整体结构，继而在图神经网络训练中动态更新图结构。经测试，该估计器能明显改善图分类问题中无结构数据在图神经网络上的表现。同时，参数估计方法使得估计器生成的图结构具有明显的局部结构，能够用于图深层信息的挖掘。该估计器弥补了图分类问题图估计器的空白，对无结构数据的图神经网络处理具有积极意义。

2. 构建了具有实用价值的骨关节炎风险预测模型：现有的骨关节炎风险预测模型效果差，不能满足骨关节炎患者早期诊断与筛查的需求。本文所构建的风险预测模型通过结合表型信息与基因型信息输入图神经网络进行处理，最终实现了较好的预测准确率，具备实用风险预测价值。同时本文提出模型通过结合图解释器对风险预测结果结合图神经网络给出预测值解释，使得模型在输出预测患病风险的同时能够对潜在的病因进行预测，对疾病分型与精准治疗具有积极意义。

3. 通过估计器所得图结构界定出骨关节炎两可能诱因：通过对图估计器所得图结构并结合 PheWAS 表型信息研究，我们发现了图结构中的两典型簇。其中一簇主要与受教育程度、是否从事中体力劳动等环境因素相关，本文推测该簇内 SNP 可能与环境因素导致关节磨损继而导致的骨关节炎相关；另一簇主要与免疫、内分泌等自身因素相关，本文推测该簇主要与个体自身肥胖导致的糖尿病与超重继而诱发的自身型骨关节炎相关。本文因此认为骨关节炎存在包括重体力劳作在内的环境诱因与包括肥胖、二型糖尿病在内的自身诱因，与目前对骨关节炎研究的观点一致。以上分析均是在单纯依靠图结构的基础上进行的，可以推广到其他疾病风险预测模型的研究之中，对复杂多基因疾病的风险预测与分型诊断有着积极意义。

## 5.2 展望

虽然本文所提出风险预测模型在给出可信预测结果的同时兼具可解释性，但是本文中还存在若干问题需要进一步研究

1. 模型超参数与结构还存在调整空间：本模型中四个模块都有着数量较多的决定模型性能的超参数，由于时间限制本文未能对这些超参数进行细致调整，可能对最终模型的性能产生不利影响。后续研究中仍要对其中一些参数加以分析调整以求继续提高模型性能。同时随着图神经网络研究的快速发展，本研究进行之时不断有效果更好的图卷积方式出现，后续研究中还需基于这些研究对图卷积层进行优化。

2. 图估计器仍需继续研究：为了简化问题，本模型图估计器所生成的图结构中各边的权重一致。但是现实生活中的图数据中各边权重通常不一致，并且有着具体意义，忽略该权重意味着潜藏信息的损失。如何通过参数描述各边的权重并通过算法估计该参数仍需后续研究。

3. 图估计产生图结构仍需进一步分析：本研究中图估计器生成的图结构共分为六簇，但是由于 SNP 位点生物学意义的复杂性，本文只对其中两簇进行了浅显的分析，并且没有对簇内节点之间与簇与簇之间的关联加以深入解释。后续研究中仍要对产生该结构的具体原因与意义加以具体分析