医疗领域的跨网络数据表示研究

李懿

2018 年 1 月

中图分类号： TQ028.1

UDC分类号： 540

医疗领域的跨网络数据表示研究

作 者 姓 名 李懿

学 院 名 称 计算机学院

指 导 教 师 礼欣副教授

答辩委员会主席 刘玉树教授

申 请 学 位 工学硕士

学 科 专 业 计算机科学与技术

学 位 授 予 单 位 北京理工大学

论 文 答 辩 日 期 2018 年 1 月

Network Representation Learning for Clinical Prediction

Candidate Name:	**Yi Li**
School or Department:	**School of Computer Science**
Faculty Mentor:	**Associate Prof. Xin Li**
Chair, Thesis Committee:	**Prof. Yushu Liu**
Degree Applied:	**Master of Computer Science**
Major:	**Computer Science and Technology**
Degree by:	**Beijing Institute of Technology**
The Date of Defence:	**January, 2018**

医疗领域的跨网络数据表示研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

作者签名：_____　　签字日期：_____

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：① 学校有权保管、并向有关部门送交学位论文的原件与复印件；② 学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③ 学校可允许学位论文被查阅或借阅；④ 学校可以学术交流为目的, 复制赠送和交换学位论文；⑤ 学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

作者签名：_____　　导师签名：_____

签字日期：_____　　签字日期：_____

摘要

随着医疗信息化的发展和电子设备的升级，越来越多的医疗领域数据以电子化的形式被记录和存储，这些蕴含丰富信息的电子医疗记录数据为医疗服务的精准化和个性化提供了基础的数据支持。虽然目前电子医疗记录数据包含非常多的数据，例如病人信息、实验室检验结果、药品使用记录、医生诊断结果和医疗影像等，可以为医生提供高质量的辅助决策并提高医疗个性化水平。但如何将数据转变为信息，仍然是急需解决的问题。这其中就包括如何从这些海量数据，特别是跨网络的数据中，自动地提取医疗信息、发掘疾病间潜在关联、减轻数据计算负担等。本文针对医疗领域跨网络数据表示问题，提出一种结合医疗实体关系网络向量化表示算法和深度神经网络模型的疾病预测算法。与传统的疾病预测算法不同，本算法将医疗领域的跨网络数据表示学习分为三个步骤。首先，从未经处理的电子医疗记录中构建医疗实体关系网络，使用基于节点相似性的网络向量化表示算法，对医疗实体关系网络中的疾病节点学习特征向量表示；接着，使用降噪自编码器栈模型和深度神经网络模型，以病人的实验室检验结果为输入，学习病人的向量化体征表示；最后，使用第一步学习的疾病特征向量构建疾病相似矩阵，作为医学辅助知识对神经网络的参数学习进行约束，让预测模型生成的潜在疾病预测结果时，既满足过往数据经验又符合医学常识。为了验证本文所提出的医疗领域跨网络数据表示学习算法的有效性，在真实的重症监护室数据集MIMIC-III 上设计了对比实验，分别进行了疾病预测准确度实验和对比算法实验。实验结果表明，本文提出的医疗领域跨网络数据表示算法，不仅合理地使用网络向量化表示算法学习医学辅助信息，还在预测潜在疾病任务上有较好的表现。

关键词： 跨网络数据；表示学习；医疗；神经网络

Abstract

With the development of medical informatization and the upgrading of clinical equipment, great amount of medical data were recorded and stored with electronic form. These electronic medical record (EHR) contains rich information about patient and diseases, like patient's demographic information, test results in laboratory, drug usage records, physician diagnoses records and medical imaging. Although EHR already contains various type of data, it's still struggling to transform such data into high-quality supportive decisions information and accelerate the progress of personalized medicine. To solve this challenge, the algorithm should be capable to extract medical information automatically with network data, to find out the potential relationship between diseases and to accelerate the data computing. In this paper, a disease prediction algorithm was proposed. The aim is to solve disease prediction mission by utilizing deep neural network model and the prior knowledge learned with network embedding. The proposed algorithm can be devided into three sections. Firstly, it extract the network of diseased entities from unprocessed electronic health records. Using network embedding algorithm based on the similarity of the neighborhood to learn the eigenvectors of the diseased nodes in the network, and then calculate a disease similarity matrix by using the disease feature vector. Secondly, it constructs a denosing autoencoder model to reduce the dimension of data which takes the patient's laboratory test results as input. Finally, it build a deep neural network to predict the potential disease, in which the prediction was constraint by the prior clinical knowledge learned by network embedding in the first section.In order to validate the effectiveness of the proposed algorithm, a comparative experiment was designed on a real intensive care unit dataset MIMIC-III. By comparing with some traditional disease prediction algorithm, the experimental results show that the proposed algorithm could learn useful representation of entity in network and provide accurate disease prediction.

Key Words: network embedding; representation learning; medical; neural network

目录

第 1 章　绪论

1.1　研究背景

随着计算机技术的快速更新、医疗领域信息化的推进以及传感器硬件技术的发展，医学检查从单纯地依靠专家人工经验进行诊断，逐步走向电子设备辅助监护配合人工检查的方式。医疗电子化的演变也带来了更加丰富的信息，从各个监护仪和检查仪器获取的结果，将病人的生理指标量化为直观的数据，让医生能够全面掌握病情的发展，及时对突发的生理变化做出响应，提高病人的存活率并有效降低误诊率。近年来，医疗领域逐步从简单的文本病历记录发展到多维度、大数据量的新型医疗诊断系统。

除了伴随医疗信息化产生的大量数据外，基于医疗历史信息和医生诊断记录提取的医疗知识网络也引发了极大的关注。依托跨网络医疗数据表示学习得到的知识，可以快速改善落后地区的医疗水平。不仅如此，对医疗领域的跨网络数据进行表示学习，还催生了精准医疗的落地。医生根据每个病人的实时体征数据，能够及时调整用药方案和诊断结果。以 IBM、微软为代表的传统软件公司，以及以谷歌和阿里巴巴为代表的互联网巨头，都投入大量的人力和财力进行医疗数据挖掘相关的项目，希望能够使用模型和算法提高医疗知识的精确度并促成医疗知识的自动化提取。

由于现在的大型医院基本都实现了信息化，与病人相关的各种数据都使用对应的电子标准格式进行存储，这些数据被统称为电子医疗数据（EHR）[1]。在后续使用的过程中，医疗人员只要依据相应的数据格式，就能进行解析并顺利读取信息。从 2001 年开始，每年都有大量的电子医疗记录被保存和使用，随着电子化普及率的提高以及医疗器械的发展，电子医疗数据日益呈现爆发式的增长[2]。例如，一个急性心脏病人，光生命体征监护设备一项，每天就能产生 5M 到 20M 的数据。这样多的数据量对于每天需要进行大量诊疗工作的医生而言，是非常难进行处理的，急需新的技术手段自动化地从数据中学习和提取有效信息。同时，系统还需要结合病人自身人口统计学数据，以及专业的医疗背景知识，才能顺利完成后续的预测任务。所以，如何对复杂的跨网络医疗信息进行表示学习，是成功构建智能医学辅助诊断系统的基础。

现实世界的数据并不总是孤立的，从各个信息源采集的数据经过有效的对齐和表示学习后，往往可以挖掘出实体的潜在关系。传统的社交网络通常关注用户之间的好

友关系，但随着移动设备的普及，由 GPS 模块产生的地理网络信息将网络上的虚拟用户绑定到每一个真实世界中的实体。通过将社交网络信息和地理位置信息融合，可以从这样的跨网络数据表示中提取用户的兴趣地点进行精准的推荐。医疗领域的数据同样也存在许多跨网络的结构特性，例如病人的实时体征数据和医生给出的诊断信息就是在不同数据网络中存储的。我们使用适当的模型算法，就可以从历史数据中挖掘出病人的潜在向量化表示，在此基础上使用可视化技术或者对应的分类算法，可以进一步辅助医生进行快速诊断。

因此，跨网络数据表示已经成为数字化医疗的研究热点之一。跨网络数据表示旨在从海量的复杂网络中，通过算法自动寻找节点之间的相关性与潜在联系，并将网络结构数据中的节点映射到对应的医疗场景中，实现医疗问题的自动特征提取，减少医生的工作负担，提高疾病的预测准确率[3]。在整个流程中，疾病预测系统首先会根据跨网络的电子医疗记录学习病人的向量化表示，然后根据结点向量之间的相似度关系预测潜在的疾病。一般情况下，疾病预测系统需要使用三方面的数据：病人的人口统计学数据、医生的历史诊断结果和医学疾病知识。

1.2 研究意义

随着计算机软件和硬件技术的发展，越来越多的医疗健康数据被采集和存储。包括病人的电子医疗记录、研究机构的药品实验数据、可穿戴设备采集的实时体征、甚至包括社交网络上的数据在内，都变得更加容易被获取。对这些医疗数据进行有效的探索已经成为热门的研究领域，数据驱动的健康管理也成为提高生活质量的趋势。

随着医疗电子化的快速推广，智能终端可以自动采集病人的各项数据和诊疗结果，为医疗领域带来巨大的机遇和挑战。并且，随着大众对于医疗科技的要求越来越高，传统的纯人工诊疗机制已经无法满足个性化医疗的需求。由于医疗数据的增加速度已经远超人工处理的能力上限，大量蕴含宝贵知识的数据没有办法被利用，无疑限制了传统医疗行业的现代化转型。除此之外，医疗是公认的复杂学科，不但涉及到某项生理指标的变化，还受各个关联指标的综合影响。所以在医疗领域采集到的数据通常伴随严重的噪声值、缺失值，同时由于病人实际病情与检查项目之间的巨大差异，数据的稀疏性也不能忽视。因此，如何从这样一个复杂的多来源跨网络数据中，学习病人的向量化表征，辅助医生进行诊断，成为一个非常紧迫但又具有很强现实意义的学科问题。

医疗数据至今难以被大规模使用的一个重要原因，在于数据本身的高维度、异质性、时间依赖性、稀疏性和不规则性，数据处理的巨大难度给后续的使用带来了很大困难。例如，同一种医学概念在不同的数据源中，会使用不同的命名规则，包括了医学领床术语系统化命名法（SNOMED-CT）[4]，统一医学语言系统（UMLS）[5]，第 9 版国际疾病分类标准（ICD-9）[6] 等。在电子医疗记录中，2 型糖尿病和 ICD-9 编码 250.00 就是同一个概念。在人工建立医疗专家系统的过程中，这些数据上的复杂问题，造成只能在某个小范围疾病上建立数据库和专家知识，难以进行扩展。

1.3 研究现状

1.3.1 跨网络数据表示学习的发展

人类在解决现实世界的问题时，通常会将问题的背景和目的抽象化为概念之间的逻辑关系，再使用逻辑推导和实际生活中的知识得到问题的解决方案。同样，计算机程序在解决问题时，也需要将数据转换为程序可以处理的形式。这种把实际问题数据化并学习对应表示的过程，通常称为数据的表示学习。目前机器学习方法的性能主要受到两方面影响，一方面是模型自身结构和算法优化，另一方面是根据不同的应用场景选择合适的数据表示方法[7]。例如针对语音识别和信号处理领域的语义分析工具和时序数据处理算法，针对图像和视频领域的边缘检测与实体识别算法等[8]。高效的数据表示方法，不仅可以降低数据本身的数据量，还可以有针对性地提取数据中显式和隐式的数据特征，将高维信息和知识提取出来。

现实世界中的各个实体之间并不是完全独立的，而是以复杂的实体网络形式存在，并相互关联和影响，所以网络是表示实体关系的最重要的形式。针对实体间网络关系的学习，被称为跨网络数据的表示学习。网络数据表示学习的研究对象通常是一个由顶点和边组成的网络 $G = (V, E)$，可以使用邻接矩阵 $A \in R^{|V|\times|V|}$ 记录节点之间的关系，其中如果 $A_{ij} = 1$ 则表示节点 v_i 和 v_j 之间有一条边 $(v_i, v_j) \in E$[9]。这种方式虽然直观地将所有顶点和对应的边都保存为计算机可以处理的格式，但由于现实中的网络一般都是稀疏网络，如果用邻接矩阵的方式表示网络会包含大量冗余信息，既不方便存储也给程序处理增加不小的负担。所以在跨网络数据表示学习的研究中，会首先将一个或多个相关网络使用恰当的表示学习算法，转换为以向量形式存储的网络向量化表示，再将这种网络特征向量应用在实际问题中，整体的流程如图1.1所示。

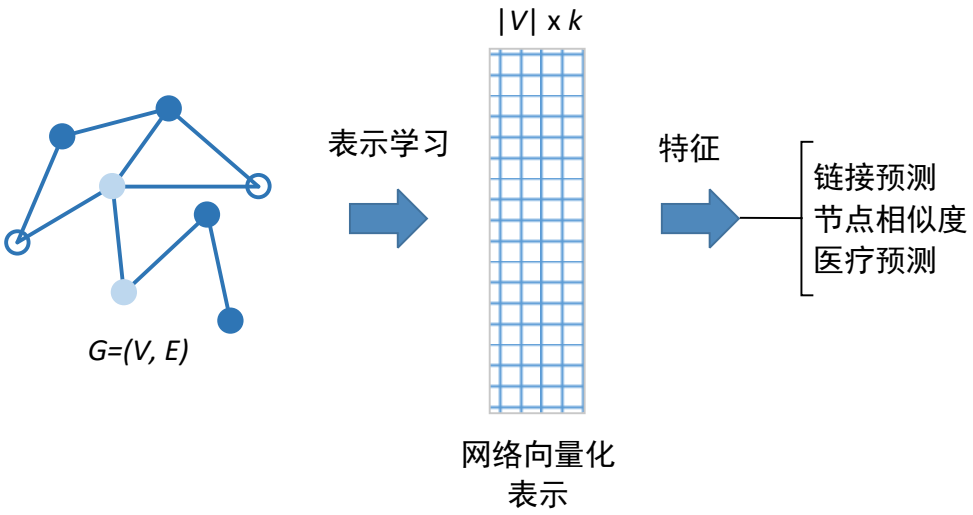

图 1.1　跨网络表示学习流程图

后续的跨网络表示研究从网络结构出发，以实现更高密度的表示学习方法。其中 Roweis 等人在 2000 年提出基于局部线性表示的方法学习跨网络数据的表示[10]，后续 Chen 等人通过有向图表示方法扩展了 Laplace 特征表示方法[11]。这些基于矩阵特征向量计算的算法虽然一定程度上解决了稀疏性的问题，但是由于矩阵计算需要大量的计算时间和资源，在大规模数据上难以扩展。对于关系矩阵的分解方法在另一个角度完成了矩阵的降维，相比特征向量计算节省大量的时间，典型算法包括通过 SVD 分解的 GraRep 算法[12] 等。

随着神经网络的发展，对大规模网络的处理出现了 DeepWalk 算法[13] 和使用生成神经网络的 SDNE 算法[14]。除了网络本身的信息之外，还有很多学者将外部的文本信息、标签信息等引入模型中，结合网络本身的结构信息和外部信息一起学习。Yang 等人就在 2015 年基于矩阵分解的框架，在跨网络表示学习中引入节点的文本信息[15]；Grover 等人也使用随机游走的模型改变序列层次方式，通过半监督的方式刻画出网络中结构的局部相似性[16]。

不论使用何种跨网络数据表示的算法，最终得到的都是网络节点的向量化表示或网络结构本身的向量化表示，这些向量化表示都作为网络的特征在后续的应用场景中被使用。

1.3.2 医疗领域数据表示学习的研究

从复杂、高维度的医疗数据中获取知识和有价值的标记数据是一项非常困难的挑战，传统的机器学习方法基本遵循特征抽取、模型构建和拟合学习的模式，高度依赖数据处理和特征学习。但面对庞杂的医疗数据，人工特征抽取或者传统机器学习方法由于缺乏必要的医疗领域专家知识，很难得到满足要求的特征。而基于神经网络的深度学习方法，可以实现多角度多层次的数据表示。同时得益于神经网络的特性，可以学习特征数据之间的非线性关系，进行更为抽象的高维度表示学习。并且，深度学习不需要人工对数据进行特征选择，可以实现大规模数据集的自动特征抽取，在文本、语音和图像这样的高维度任务中发挥巨大优势[17][18]。

在医疗领域的跨网络数据分析方面，目前多数任务集中在医疗影像识别和电子健康记录分析两个方面。医疗数据分析涉及到结构化的数值数据和非机构化的文本手写数据，重点关注挖掘数据时间特性和从海量数据中自动提取有效信息的算法。在模型方面，研究学者将卷积神经网络（CNN）[19] 迁移到医疗图像问题上，实现病变组织的自动标注。同时利用卷积神经网络模型自动抽取相邻区域内信息，挖掘医疗文本和数值数据的潜在关系。另一种经常被使用的模型是长短期记忆机制（LSTM）[20]，这个模型改进循环神经网络（RNN）[21]，增强了历史信息的记忆和存储能力，能够对医疗领域常见的时序数据进行分析处理。同时，随着深度神经网络在各个领域的成功，基于深层神经网络的模型也被使用在医疗数据处理上，极大提高系统对海量数据的处理能力。目前基于神经网络的医疗数据表示已有很多成功的案例，例如 IBM 公司的 Watson 系统就在 2015 年成功预测一名女性患者可能患有白血病并提出治疗方案[22]、谷歌公司的 DeepMind 团队使用糖尿病病人的眼底影像建立视网膜病变预测系统以及斯坦福大学的皮肤病预测算法等[23]。

1.3.3 跨网络数据表示学习的应用场景

关于跨网络数据表示学习方法的应用场景，根据要解决的是网络内部问题还是网络外部问题，大致可以分为两大类。一类是涉及到网络内部结构的链接补全问题和节点分类问题，另一类是用于解决网络外部问题的社群发现问题。

节点关系链接的补全问题，主要解决现实网络的复杂性，对网络关系进行采集的过程中难免会遗漏某些节点之间的关系，需要有效的方式对这些潜在关系进行预测[24]。这种链接预测的应用场景很广，例如在社交网络中，对用户推荐另一个关注相

同兴趣领域但还未互动的用户，可以促进整个社交网络的活跃度；在医疗领域，通过分析药品成分和化学反应，发现已有药物在原有领域之外的其他治疗作用或者发现药品之间的不良反应，能够减少昂贵且耗时的临床试验，降低药品研发成本。

节点分类问题也是跨网络数据表示学习的一个应用场景，它涉及对已有节点的属性划分和新加入节点的属性预测。举例来说，在基于地点的商家推荐系统中，对于还没有正式划分属性的新商户，除了人工给出标签的方式之外，还可以基于顾客评价和周边商户的类别预测新商户的类型[25]。这种使用稀疏数据或少量已标注数据进行分类的应用场景，同样也可以实现对未知疾病的发病原理探索和新型药物功效的发现，在医疗领域具有非常重要的实际作用。

社群发现在社交网络分析中也非常重要，人群根据兴趣、工作、居住环境，天然的被划分为相互独立的社群。由于社交网络的匿名性，导致研究者无法直接对用户进行社群划分。需要算法通过分析用户的行为，找到隐藏在用户中的社群[26]，作为好友推荐和运营活动的基础。同样的应用在医疗领域也存在，比如根据蛋白质性质，将蛋白质网络进行分类处理等。

除了上述三种应用场景，跨网络数据表示学习本身还同时具有实用性和可扩展性，在医疗和其他领域发挥重要作用。

1.4 研究内容

通过对研究背景以及研究现状的深入分析，针对传统医疗数据表示模型中缺少跨网络数据融合场景的问题，本研究提出一种新型医疗领域跨网络数据表示方法。主要从解决数据稀疏性和融合跨网络数据角度进行考虑，不仅能够处理大规模电子医疗记录数据，还利用深度神经网络模型建立具有实际使用价值的疾病预测模型。有别于以往数据表示算法，本方法分两个阶段实现数据的表示和跨网络数据融合任务，首先使用降噪自编码器栈对原始的高维医疗数据进行特征学习，再使用基于网络结构的节点向量化表示方法学习病人特征，最后建立多标签预测任务进行潜在疾病的预测及评估。下面将从三个阶段分别进行具体阐述。

1.4.1 稀疏医疗数据的表示学习

由于现实世界的复杂性，大部分对实体进行数据化得到的结果都是具有很高维度的数据，例如对一个人的量化可以从生理指标也可以从人口统计学角度，这些维度综

合在一起可能有成千上万维，对于目前的计算机硬件水平，还很难进行解析和处理，而且这些稀疏的数据还会占用大量存储空间，不利于快速检索数据，所以需要对这样的稀疏数据进行降维，提取对目标任务有价值的特征向量表示。

在医疗领域，稀疏数据是由数据采集方式决定的。对一个急性心脏病发作的病人，需要监护仪实时监控病人的心率、呼吸率、血压和体温等指标，但这些指标通常会稳定在一个较小的区间内波动，只有在出现严重生理反应时才会出现异常值。在记录数据时，普通情况下的正常值虽然很稠密，但是对于数据分析没有很高的价值，反而是稀疏的异常值能够反映病人出现的特殊生理状况。在电子医疗记录（EHR）中，病人的生理数据和检查数据大部分使用数值型保存，每个病人 ID 对应一个时刻的检查结果，造成数据稀疏。考虑到数据获取难度，一般采用类似主成分分析法的非监督方法进行稀疏数据的降维处理，受到[27] 的启发，本文采用降噪自编码器栈的形式，对医疗电子记录数据进行非监督的表示学习。因为自动编码器模型不依赖先验条件或其他知识，能够实现任务无关的数据表示学习，可以得到病人生理状态的向量化表示，输入到后续的预测模型中。

1.4.2 跨网络数据的融合

在本文使用的医疗数据库 MIMIC-III 中，涵盖了病人的生理数据、诊断记录和医生开具的药品处方信息[28]。这些数据不仅来源不同，数据结构也完全不同，无法直接使用数据，所以需要对不同数据网络学习到的医疗数据表示进一步融合，实现跨网络的数据表示学习。

病人的诊断信息在原始数据库中是直接使用 ICD-9 代码进行存储的，每个病人的一次就诊记录保存组成一段信息，为了从医生的诊断结果中学习不同疾病之间的关联，本文将诊断记录信息转换为由诊断代码为实体的网络结构。在这个网络中每个节点代表一种疾病的诊断代码，节点之间的边代表不同疾病在同一个病人身上同时出现的可能性，两种疾病同时出现的次数越多则两个节点之间的边强度越强。最后可以将所有重诊断数据转为一整个网络。本文采用基于网络中节点关系的结构信息，进行网络节点的向量化表示学习。这种方法根据不同节点周围的邻居分布关系，衡量两个节点的相似性，如果一个节点与另一个节点共享很大部分的直接邻居，那么这两个节点可以视为相似节点，这两个节点所对应的向量化表示在特征空间上也应该是比较接近的。我们根据这种相似关系可以得到不同疾病之间的相似性度量矩阵，提供给之前数

据表示阶段学习得到的向量使用，一起融合之后传递给最后的疾病预测模型使用。

1.4.3 潜在疾病的预测

疾病预测任务是根据病人当前的身体状况，预测病人潜在疾病的发病概率。由于医疗水平的限制，医生只有病人基础数据时，可能会出现诊断上的遗漏，造成病情治疗的延误。潜在疾病的预测任务可以作为一种辅助手段，通过模型自动地对一些平时比较少考虑到的疾病进行发病概率的预测，当某种疾病的发病概率超过阈值时，就可以及时提醒医生针对这种疾病进行进一步的检测。

由于疾病之间不是互斥关系，同一个病人可能同时发生多种疾病，所以不能使用传统的多分类预测模型。在本方法中，我们使用的是多层感知机的神经网络进行多标签预测任务，这个模型的输入是自动编码器学习的病人体征向量，输出针对重症监护室内发生频率较高且对病人危害大的几种疾病的发病概率。

1.5 论文组织结构

第一章概述了研究的背景、意义和研究现状，通过对研究问题和现状的分析，阐述本文提出的解决方法。

第二章具体介绍了实现医疗领域跨网络数据表示需要使用到的模型算法，包括文本向量化算法、网络表示算法和神经网络预测模型。

第三章详细介绍了医疗实体关系网络中的表示学习模型。

第四章详细介绍了跨网络数据融合与疾病预测模型构建。

第五章介绍了本算法与其他方案的横向对比结果，说明本文提出的方案的高效性与可行性。

第 2 章　跨网络数据表示中的算法模型

2.1　数据表示模型

医疗领域的一个常用研究手段是先让医学专家明确目前已知的病症特征（Phenotypes），然后根据这些特征规则进行学习。但是有监督的特征学习通常会导致特征空间过于狭窄，并且无法发现新的潜在数据关联。与之相反，使用算法模型进行数据挖掘和知识抽取的工作，可以自动发现新关联。经过了数十年的更迭，从简单的数据挖掘算法应用，逐步发展为结合医疗专家知识的系统工程。当数据规模较小时，使用简单的关联规则分析和时间序列分析算法就足以提取基本的医学规则；随后数据规模不断增大，知识图谱被引入医疗数据分析中，用于构建程序可理解的医学规则；现在，医疗任务逐渐从基础的识别任务转向数据驱动的精准医疗，各种神经网络模型成为处理大规模数据和分析数据的首选算法。

2.1.1　基于关联规则分析的医疗数据表示

关联规则是数据挖掘领域的经典方法，它通过在数据集中对各个项集的发生频率进行计算，结合支持度和置信度分析不同项集之间的潜在相关性。针对交易数据库的分析任务，可以定义 $I=\{i_1,i_2,...,i_m\}$ 为交易数据集 T 的项集集合，$D=\{t_1,t_2,...,t_n\}$ 为事务集合。事务集合中的任意一项 t_i 表示一条交易记录，是项集集合 I 的一个真子集。整个挖掘的过程大致可以分成两个阶段，第一阶段通过预设的最小支持度在原始的数据库中找出所有的频繁项集，第二阶段通过预设的最小置信度生成满足条件的规则。由于生成频繁项集的过程需要花费大量的时间，目前一般采用 Agrawal 等[29]在 1993 年提出的快速生成频繁项集 Apriori 算法。通过合并两个只有一项不同的频繁 $(k-1)-$ 项集，生成频繁 $k-$ 项集。根据递推规则，首先产生频繁 $1-$ 项集 L_1，再产生频繁 $2-$ 项集 L_2，直到无法生成新的频繁项集为止。这种算法解决了遍历整个频繁项集空间时，最大复杂度为 $O(n^2)$ 的情况，极大减少了程序运行时间。

在医疗领域进行数据的关联规则分析，本质上是将病历数据和疾病数据电子化，并针对症状和疾病之间的关系挖掘对应的关联信息，建立起疾病症状和潜在疾病之间的联系。为了完成这个任务，王华等人[30]在 2006 年提出了临床上数据挖掘的四个步骤。第一步，将病人的病历档案数据电子化并转换为可供数据挖掘算法分析的格式，

建立症状对照表；第二步，建立医疗代码和病历数据之间的关联数据库；第三步，使用关联规则分析算法，产生频繁项集；最后，根据置信度建立关联规则。使用关联规则建立医疗数据表示的意义在于，由于病人缺乏专业的医学知识背景，在向医生表述自身症状时可能会使用非常模糊的词语，在进一步的检查之前医生只能根据自身的经验做出判断。但即使受过专业训练的医生也无法将病人的表述和具体疾病之间潜在的所有可能性考虑在内，造成诊断病情的延误，给病人带来不必要的痛苦乃至引起医疗纠纷。关联规则算法简单直观，生成的结果都是人工可以处理的逻辑推理结果，在经过专业人士的筛选后，完全可以作为一种医疗辅助手段，帮助医生快速识别病人的真实病情。

2.1.2 基于知识图谱的医疗数据表示

随着网络和电子设备的发展，互联网上出现了越来越多散落的元知识，在这个背景下诞生的知识管理和知识表示方法就是知识图谱，用以强化语义检索能力。在人工智能发展浪潮下，知识图谱覆盖的知识表示、抽取、推理、融合和问答等关键问题也取得了一定程度的突破。知识图谱的前身是万维网之父 Tim Berners-Lee 提出的语义网，通过结合本体和知识组织及管理方面的概念，让电子化的知识在不同计算机以及不同人之间能够更容易地流通和加工。一个知识图谱包含概念实体和概念之间的语义关系边。目前大规模使用的通用知识图谱有谷歌公司的“Knowledge Graph”、搜狗公司的“知立方”等，在医学领域也涌现了很多垂直领域的知识图谱应用，如 wishart 提出的药品知识图谱 drugbank[31]、医疗本体知识库 SNOMED-CT[4]、贾李蓉在 2015 提出的中医药知识图谱[32] 等。构建一个医疗知识图谱主要涉及五个方面的内容：医疗知识表示、医疗知识抽取、医疗知识融合、医疗知识推理和质量评估五个部分。

知识表示是通过约定的符号对数据知识进行模式化的过程，将数据存储为计算机可用格式。由于医学数据来源复杂、种类繁多，不同来源的电子病历间通常不能通用，导致建立医疗知识图谱表示有很大困难。早期主要使用传统的谓词逻辑表示法和语义网表示法，例如 SNOMED-CT 知识图谱就是使用这些方法生成的。但是由于这些基于语法的表示方式能力有限，目前主要作为辅助构建的手段作为补充。而以三元组（实体 1，关系，实体 2）来表示知识图谱中节点关系的本体表示法，逐渐被认可。使用本体表示法可以提高医疗术语之间的整合表示能力，建立强大的医疗信息系统并实现不同数据的聚合，而且本体表示法中三元组的关系可以是广义上的关系，恰好能够表达

医学上一些晦涩的关系。目前，医疗知识图谱的规模日益扩大，网络中的节点数量增加到传统方法都无法处理的程度，通过引入机器学习方法将知识图谱中的语义信息表示为低维空间上的稠密向量，有效解决了数据稀疏的问题，提升知识融合和推理的性能。

医疗知识抽取的目的是从非结构化的数据中，自动抽取实体、属性和关系。主要分为人工抽取和自动抽取两种，但是因为人工抽取需要大量医学专家进行工作，代价太大，使自动抽取技术成为主流。实体抽取可以标准化地识别医学概念，有基于医学词典及规则的方法、基于医疗数据统计学的方法和基于机器学习及深度学习的方法等。

医疗知识融合包括实体对齐和知识库融合两方面，实体对齐主要解决知识图谱中复杂来源带来的知识重复和知识质量问题，知识库融合解决不同垂直领域的知识库合并来解决知识图谱中数据的多样性和异构性问题。

医疗知识推理是从已经构建好的知识库中挖掘有价值的隐含信息，通过合理的推理方法选择减少人工参与。知识推理可以帮助医生高效地完成患者医护数据整理和收集、控制误诊率、预测潜在疾病等任务。这部分必须具备处理大规模重复数据和矛盾数据的能力，是医疗诊断自动化工程中不可缺少的部分。主要使用传统的描述推理、规则推理、案例推理等，但也存在学习能力不足和错误率较高等问题。近年来将知识推理和机器学习及深度学习结合的趋势下，数据维度和查询效率明显提高。

质量评估是保障数据正确性和可用性的关键环节，主要用于量化数据质量和筛选置信度高的数据。与其他领域不同，医疗关系到患者的健康和性命安全，更加要求质量评估在严谨性和可靠性上的水平。

医疗图谱在人工智能的飞速发展下，正受到国内外企业和学界的广泛关注，有望给大众提供更精准、高效、廉价的医疗技术。目前医疗问答系统、医疗信息检索系统、医疗决策支持系统都是重点开发的方向。

2.2 自动编码器模型

真实世界的数据通常反映了事物之间的复杂关系，所以数据必须具有很高的维度才能完整表示这些复杂的现象。但是当数据维度超过目前计算机的处理能力时，就会引发“维度灾难”，造成很难处理高维数据并有选择性地提取出对应的有效特征。同时，高维度也存在“维度福音”，这是指现实世界的客观现象和信息包含在这些极高

的数据维度内，蕴含着解决问题所必须的信息。自动编码器使用无监督的学习方法，通过重构原始数据的方式实现对数据特征的学习，能够在降低数据维度的同时，将最能体现数据特征的向量化表示学习出来。

2.2.1 自编码器结构

自动编码器由编码器、隐含层和解码器三个部分组成，如图2.1所示。给定 n 个输入数据 $X = x^{(1)}, x^{(2)}, ..., x^{(n)}$，对于每个 p 维数据向量 $x = x_1, x_2, ..., x_p$，由输入层到隐含层的映射函数转为 m 维的隐含层向量 $y = y_1, y_2, ..., y_m$，映射函数为

$$y = f_\theta(x) = s(Wx + b) \tag{2.1}$$

公式中的 $s(x)$ 是非线性激活函数，一般情况下采用 Sigmoid 函数

$$sigmoid(z) = \frac{1}{1 + z^{-1}} \tag{2.2}$$

下一步将求出的隐含层表示 y 映射为输出向量 z，要求 x 与 z 尽可能相似，即复现输入数据在降低维度的同时最大限度地保留对数据中主要的特征要素。解码器将隐含层映射为输出数据的函数为：

$$z = g_{\theta'}(y) = s(W'y + b') \tag{2.3}$$

公式中的 s 是解码器的激活函数，也使用 Sigmoid 函数。在将输入向量 $x^{(i)}$ 映射为隐含层向量 $y^{(i)}$，再映射为重构输出向量 $z^{(i)}$ 的过程中，自编码器模型的参数最优解应该满足重构误差最小化的条件，也就是寻找满足最小化 x 和 z 之间的重构误差时的参数 θ 和 θ'。满足以下表达式：

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) = \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} Lx^{(i)}, g_{\theta'}[f_\theta(x^{(i)})] \tag{2.4}$$

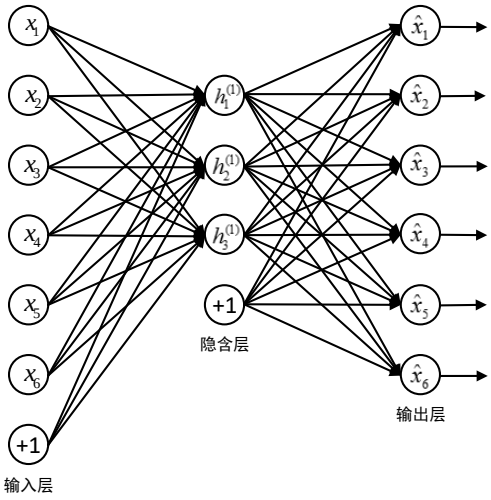

图 2.1　自动编码器结构

2.2.2　降噪自编码器及深度自编码器栈模型

原始的自动编码器只考虑了最简单情况下的数据还原情况，没有办法适应现实中的数据在采集的过程中就可能出现各种缺失情况。降噪自编码器（Denoising Auto Encoder, DAE）是通过引入人工噪音，来提高原始自动编码器的鲁棒性的一种改进模型，改造过后的降噪自编码器原理如图2.2所示。首先对于原始输入 x 进行一次随机映射，得到 $\tilde{x} \sim p(\tilde{x} \mid x)$，即按照预先设定的比例随机地对数据进行置零操作，人为地毁坏数据。随后按照原始自动编码器的结构，进行隐藏层映射和输出层映射，得到输出层结果 z，使 z 和未破坏的原始数据 x 尽可能接近。这种处理方式通过程序的方法模拟了人在处理缺损或低质量数据时的模式，当一个人在看到一幅部分被遮挡的图片时，还可以通过剩余部分的特征进行判断，得到图片的主题。

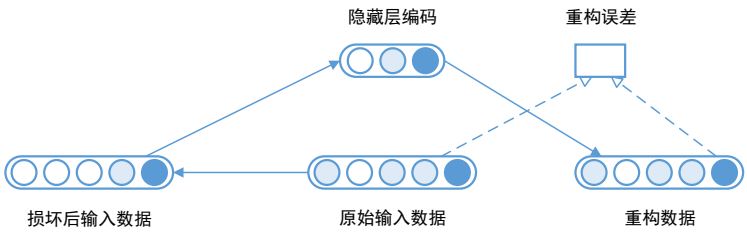

图 2.2　降噪自动编码器原理

随着神经网络的发展，深度自编码器栈模型通过将自编码器结构扩展到深度模型

上，实现了一种使用无监督逐层贪心预训练的非线性多层网络结构，可以在原始高维数据中学习分层特征的神经网络结构。自动编码器最早由 Rumelhart 在 1985 年提出[33]，并由 Hinton 在 2006 年提出逐层预训练的改造方案，得到降噪自编码器结构[34]。深度自编码器栈在原始自编码器结构上，增加了隐藏层的层数。训练时，首先以无监督的方法对第一层神经网络训练，最小化重构损失；然后以上一层隐藏层的输出作为输入，训练下一层；重复逐层训练的思路直到深度自编码器栈的所有参赛训练完成，将最后一层的隐含层输出作为最终的降维特征，提供给后续有监督学习方法使用。

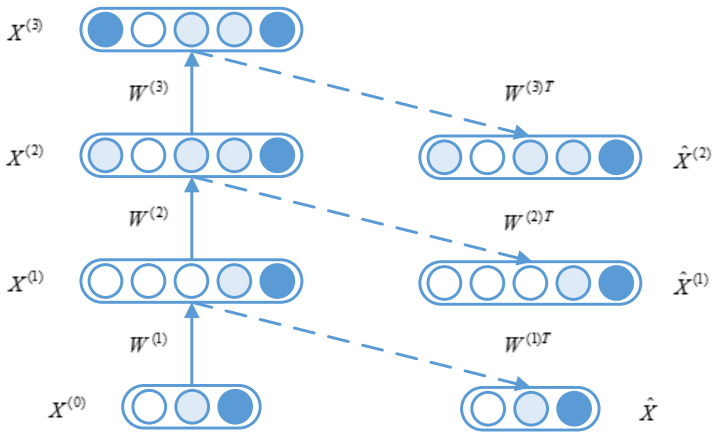

图 2.3　深度自动编码器栈原理

将降噪自编码器用在医疗领域的应用也非常多，lasko 等在 2013 年使用降噪自编码器对稀疏的医疗数据提取疾病表征[35]，悉尼大学的刘博士团队在 2014 年实现了阿尔兹海默症进行早期预测[36]，台湾大学的郑博士使用无标签的医学影像对乳房硬化进行诊断[27]。

由于降噪自编码器简单明了的结构，研究者针对不同应用场景进行改造，可以快速实现数据的降维和特征抽取任务。但是由于神经网络仍然设计大量隐藏层节点和层数的设定，在数据规模大的情况下，现有的硬件条件比较难满足计算要求，需要根据具体任务调整神经网络的结构。

2.3　网络向量化表示算法

向量化表示是计算机领域的一种常用数据表达，对于一个实体或者抽象概念，使用若干维度组成的向量形式对该实体概念进行表示，每一个维度对应的数值代表这个

实体在该维度上对应的强度大小。例如在对一个人进行向量化表示时，我们可以选用性别、年龄、身高和体重四个维度，当我们用 0 代表女性，1 代表男性后，一个人可以被一个长度为 4 的向量表示。

向量化表示的目标是学习一个映射函数，将一个实体或者抽象概念映射到同一个特征空间中，让这些不同类型和来源的对象可以在特征空间中直接进行计算和比较。对于一个数据网络而言，实体构成了网络的节点，实体间关系构成了节点之间的边，在使用网络向量化表示算法后，实体间边上的关系被映射到特征空间上的某种抽象特征中，使得我们可以很方便地对没有直接连接关系的节点进行比较。网络向量化表示算法是来源于自然语言处理领域的词向量化表示算法，随后研究者通过结合随机游走机制和其他模型，逐步发展出针对网络结构数据的向量化表示算法。

2.3.1 词向量化表示

自然语言处理（Nature Language Proceccing, NLP）中，经常遇到需要预测在已知一段文本序列的情况下，接下来出现哪段文本的任务。这样的预测任务对于机器翻译、问答系统和文本补全都有非常重要的意义，如果我们能够清楚地预测目标结果的概率，就可以挑选概率最高的候选词作为回应。将这个任务用统计语言模型进行表示，一段文本 $S = w_1, w_2, ..., w_T$ 出现的概率是每个词出现的条件概率的乘积：

$$P(S) = P(w_1, w_2, ..., w_T) = \prod_{t=1}^{T} p(w_t \mid w_1, w_2, ..., w_{t-1}) \tag{2.5}$$

这样的联合概率因为参数空间太过于巨大难以计算，所以提出了 N-gram 模型，在一个较小的范围内进行概率统计的计算。N-gram 模型将每个词使用一个 one-hot 向量进行表示，这个向量的长度是整个词表的长度，也就是词向量的某一位为 1 即代表该词语。这种词向量表示方法在实际应用时，面对上万词汇的词典很容易导致维度灾难问题。需要配合矩阵分解等降维方法，将稀疏矩阵映射为稠密的词语离散化表示使用。但是由于 N-gram 模型本质上依然是将每个词作为独立的个体进行计算，忽略了段落本身具有的上下文环境和语境，导致对结果的预测有效性不高。针对这些不足，Bengio 等人在 2003 年提出了基于神经网络的统计语言模型处理框架[37]，通过将词表中的每一个单词用一个连续的向量进行表示，进而使用 softmax 方法计算候选词出现的概率。但是由于计算量仍然很大，并且只能处理定长的序列，所以 Mikolov 等人在

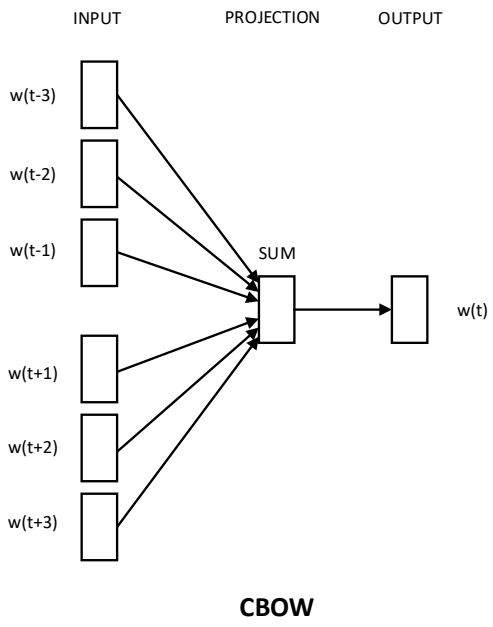

图 2.4　word2vec 算法的 CBoW 模型

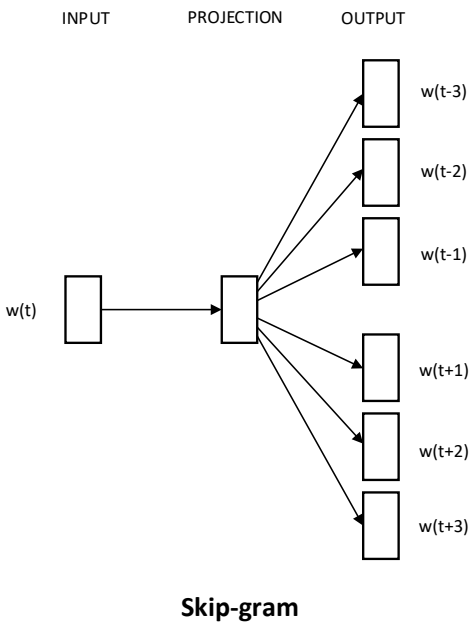

图 2.5　word2vec 算法的 Skip-gram 模型

2013 年提出了新的词向量化模型 Word2vec[38]。

word2vec 算法包含两个模型，分别是 CBoW 模型（Continuous Bag-of-Word Model）和 Skip-gram 模型。其中 CBoW 模型是根据词袋模型，将预测词语附件一定窗口内包含的上下文词语忽略次序直接进行融合后，使用映射函数得到输出的词向量化表示，结构如图2.4所示。这种取消神经网络的隐藏层，直接使用 softmax 层连接融合后的词向量与输出的方式，大大降低了算法的计算复杂度，同时也充分考虑了上下文环境信息。CBoW 模型学习到的是根据上下文预测目标词语学习到的词向量表达，也可以反过来，通过目标词语预测上下文，也就是 Skip-gram 模型如图2.5。Skip-gram 模型从本质上来看，是计算输入词语的向量化表示与目标词语向量化表示之间的余弦相似度，同时使用 softmax 函数进行归一化处理。其数学形式是，当 V_i 是输入向量，U_j 是输出向量时：

$$p(w_o \mid w_i) = \frac{e^{U_o \cdot V_i}}{\sum_j e^{U_j \cdot V_i}} \tag{2.6}$$

对于 Skip-gram 模型和 CBOW 模型的计算，Mikolov 给出了两种计算框架，分别是层次化 Softmax（Hierarchical Softmax）和负采样（Negative Sampling）。

层次化的 Softmax 是 Bengio 在 2005 年引入语言模型中的一种方法，它把原先复

杂的归一化计算公式分解为一系列层次结构的概率连乘形式，具体计算如公式2.7所示。其中每个概率对应一个二分类的概率问题，将复杂度降低到 $logV$ 个词语的出现概率计算。在实现层次化 softmax 时，首先根据词语在文档中出现的频率构建一颗哈夫曼树，树上的每一个叶节点为一个词。则可以根据这颗哈夫曼树对词语出现的概率进行计算。

$$p(v \mid context) = \prod_{i=1}^{m} p(b_i(v) \mid b_1(v), ..., b_{i-1}(v), context) \tag{2.7}$$

哈夫曼树结构的采用虽然减少了计算复杂度，但在实际使用时仍然需要耗费大量的计算时间。所以负采样方法使用更加简单的随机负采样方法用来提升性能。在日常生活中，如果我们需要考虑一个预测系统的好坏，不仅会衡量它对正确样本的分类正确率，还会同时考虑它能否将一个错误样本识别出来。在语言信息处理任务中，就体现为负采样算法需要从两个角度考虑损失函数，一个角度是对于目标单词与上下文单词的得分，另一个角度是目标单词与非上下文单词也就是噪音单词之间的得分。这种方法通过对 k 个样本进行负采样就实现了类似 softmax 的效果。公式如下：

$$J(\theta) = \frac{1}{T} \sum_{t=1}^{T} J_t(\theta) \tag{2.8}$$

$$J_t(\theta) = log\sigma(u_o^T v_c) + \sum_{i=1}^{k} E_j P(w)[log\sigma(-u_j^T v_c)] \tag{2.9}$$

2.3.2 段落和篇章向量化表示

在词向量化的基础上，还针对长度更长的段落和篇章进行向量化的模型。在段落向量化（paragraph2vector）算法中，作者仍然实现了从上下文进行目标词预测和从目标词进行上下文预测两种模型，但与原始的词向量化方法不同，段落向量化算法定义了一个新的向量形式，也就是段落向量。

第一种情况被称为分布式记忆模型 PV-DM（Paragraph Vector: Distributed Memory model），在初始情况下的段落向量是随机生成的一个定长向量，这个向量被同一个段落中生成的上下文窗口共享，结构如图2.6。可以这种情况下的段落向量当做一个特殊的词语，它通过学习上下文中词语出现的概率，存储这个段落的高维抽象主题，作

为一个虚拟的记忆单元。接下来的词向量和段落向量和 word2vec 算法相同，都是通过随机梯度下降训练并使用反向传播算法求解梯度。这种模型的优点是继承了词向量模型中”词语语义“的特性，保留了”powerful“和”strong“这样的单词非常相似的含义；另一个优点是它保留了词语的语序。

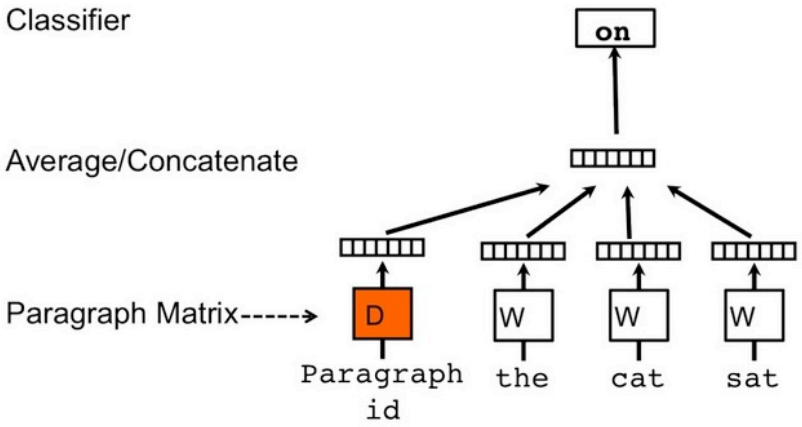

图 2.6　paragraph2vec 算法的 PV-DM 模型

第二种情况中，需要从一个段落向量推算出段落中应该出现的词语，这种模型称为针对段落向量的分布式版本的词袋模型 PV-DBOW（Distributed Bag of Words version of Paragraph Vector）。在随机梯度下降的每一个迭代中，首先抽取一个滑动窗口内的文本，再随机抽取一个词语来构建分类器的目标，结构如图2.7所示。根据 PV-DM 模型和 PV-DBOW 模型学习得到的段落向量可以组合在一起使用，以便更好地表示段落中的语义概念。

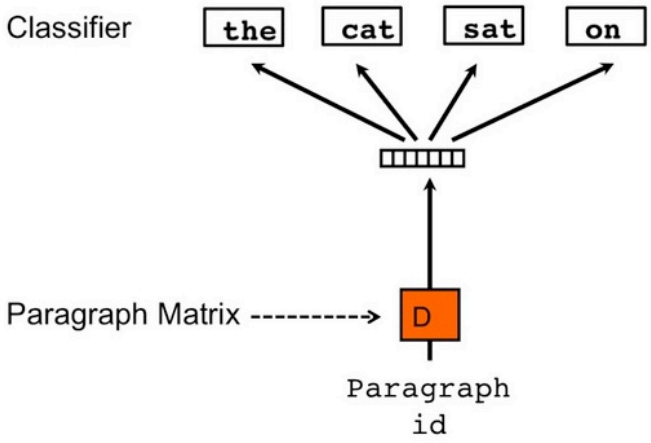

图 2.7　paragraph2vec 算法的 PV-DBOW 模型

2.3.3 网络节点向量化表示

对一个网络中的节点进行向量化表示学习的过程，其实就是最大限度保留网络邻居关系的过程，当我们使用合适的算法对节点向量化后，通过节点向量的计算我们可以还原网络中的邻居关系。例如当计算两个节点向量的余弦相似度时，直接相连的节点相似度值大于不连接的，拥有共同邻居的节点相似度高于没有共同邻居的节点。所以，网络节点向量化的目标就是将节点映射到一个低维的特征空间中，并最大程度保留节点之间的关系。

在 Grover 等人于 2016 年发表的网络特征学习算法 node2vec 中，在深度游走模型[13] 的基础上提出了一种通过设计带偏置的随机游走方法实现网络结构的保留[16]。当一个节点在网络上进行运动的时候，会产生一段轨迹，这段轨迹记录了所经过的所有节点。这样一条轨迹和文本处理中的一段上下文非常相似，也可以通过 Skip-gram 模型，将每个节点当做一个词语进行处理，学习出对应的向量化表示。这种方法的主要难点在于对上下文，也就是轨迹的采集。传统算法一般使用深度优先搜索 DFS（Depth First Search）和广度优先搜索 BFS（Breadth First Search），模拟一个点在网络上的运动轨迹，如图2.8所示。

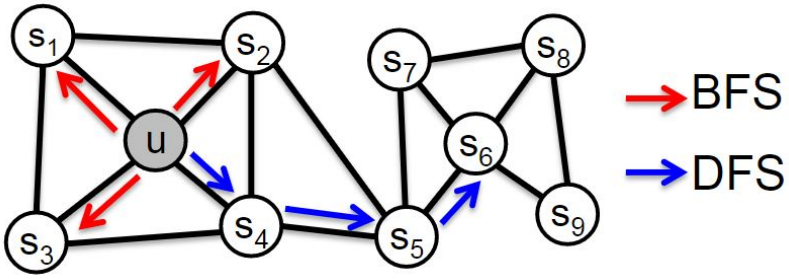

图 2.8　传统的图搜索算法

虽然这两种搜索方法分别从宏观上和微观上进行网络结构的探索，但是不管哪种方法都存在局限性，无法完全展示网络结构，所以作者提出了基于随机游走概念的图搜索算法。首先定义随机游走时节点之间的转移概率为：

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & \text{if}(v,x) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

其中的 π_{vx} 表示未经正则化的节点 v 和节点 x 之间的转移概率，Z 是正则化常数。

表示如果节点之间存在边，则按照转移概率进行下一跳节点的选择；如果节点间不存在边，则不可能发生转移。与传统的随机游走算法不同，node2vec 算法的作者认为现实世界中是深度优先搜索和广度优先搜索的混合，要表示这种复杂的混合情况可以通过设定搜索时的偏置 α，实现对现实情况的最大程度模拟。当出现如图2.9的情况时，点在网络上刚刚经过了 (t,v) 边来到节点 v 所在的位置。在 node2vec 算法中，设置未经正则化处理的转移概率为 $\pi_{vx}=\alpha_{pq}(t,x)\cdot w_{vx}$。$\alpha$ 具体为：

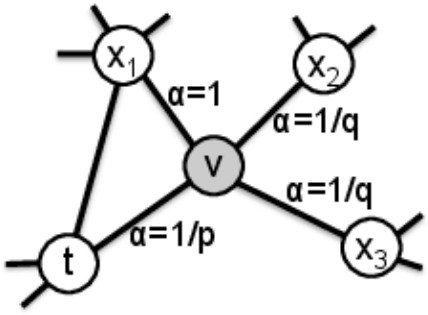

图 2.9　节点向量化算法

$$\alpha_{pq}(t,x)=\begin{cases}\frac{1}{p} & \text{if} d_{tx}=0,\\ 1 & \text{if} d_{tx}=1,\\ \frac{1}{q} & \text{if} d_{tx}=2.\end{cases}$$

其中 d_{tx} 表示待计算的下一跳节点 x 与 t 之间的距离，当下一步跳回节点 t 时距离为 0，当下一步跳到节点 x_1 时距离为 1，当下一步跳到节点 x_2 或 x_3 时距离为 2。根据实际情况就可以计算出每个节点对应的转移概率。直观上看，参数 p 和 q 控制点在网络上游走时，离开目标节点 u 的邻居范围，进行探索的速度。所以，实际上就可以通过设置参数控制深度搜索和广度搜索之间的切换。参数 p 是返回参数，表示点重新返回之前走过的路径的概率，当参数大于 1 时倾向于向外探索，当参数小于 1 时倾向返回。参数 q 是控制点在运动时侧重深入还是扩展的参数，在图2.9中则体现为当参数值大于 1 时倾向进行广度优先搜索；当参数小于 1 是倾向原理 t 节点的深度优先搜索。

和单纯的深度优先搜索或者广度优先搜索相比，采取随机游走的方式不论在时间还是空间上都更加高效。设定好每段轨迹的长度和 α 参数后，就能在网络轨迹采样时获得最能体现真实情况的数据。接下来使用词向量化算法中的 Skip-gram 模型，就能够得到网络中每个节点对应的向量化表示。

2.4 基于神经网络的数据表示模型

不管是传统的关联规则分析、知识图谱还是新兴的神经网络模型，本质上都是在挖掘数据本身的规律，研究病人状态的向量化表示和生理状态与疾病之间的关系，但是在处理复杂来源的医疗领域跨网络数据时，还是难以直接应用。因此，在解决跨网络数据表示、稀疏数据特征抽取方面的研究依然面临着不小的挑战。

2.4.1 多层感知机

多层感知机模型（Multi-Layer Percetron，MLP）是基于前向反馈型的一种多层神经网络模型，主要用来解决非线性不可分时的计算任务。组成结构也有输入层、隐含层和输出层构成，如图2.10所示。

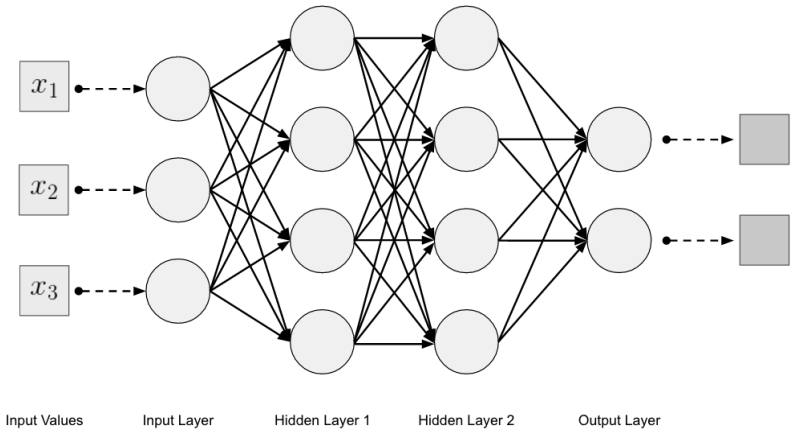

图 2.10　多层感知机结构

2.4.2 基于循环神经网络的数据表示算法

与图像数据不同，由于医疗服务是一个持续的过程，在医疗领域经常会采用时序数据的方式保存病人在每次检查时的生理数据以及对应的医生医嘱。在病人住院情况可以分为两种情况，第一种属于普通疾病或者慢性疾病，在确诊后一般以 12 小时为周期由专业的医护人员对病人的状况进行检查，持续几天到几周时间，会产生稀疏但时间跨度较大的数据；第二种属于急性疾病或威胁生命的情况，会在病人身上佩戴实时记录生理体征的监护仪，持续几小时到若干天，会产生稠密但时间跨度小的数据。对于稀疏数据，一般可以使用传统的机器学习方法进行特征点的标记和提取。但是对

于短时间内的大量稠密数据，医护人员很难有时间对所有数据进行有效的分析，这就需要针对时序数据进行数据挖掘和表示的算法模型，以便自动提取病人特征。

在电子医疗记录（EHR）上进行时间序列建模的最近一个模型，是由 Choi 等人在 2016 年提出的 Doctor AI 模型[39]。面对 26 万个心脏病病人的数据，研究者经过筛选后得到人均 54 条访问记录（visit），每次记录对应若干个包含时间戳的诊断代码或者药物代码。这些经过筛选和处理的医疗时序数据，首先作为高维原始记录使用传统的嵌入式表示方法，得到低维表示，然后在使用两层结构的循环神经网络进行学习，得到病人的向量化状态表示，最后使用全连接神经网络，预测病人对应的诊断代码和下次进行诊疗的间隔时间。Lipton 等人在 2016 年通过分析儿科急诊室类的医疗电子记录，在小时粒度上完成了对单独病人在数个月之内数据进行学习，并使用 LSTM 学习病人的潜在死亡率[40]。循环神经网络在处理这样的医疗数据时，能够充分考虑病人之前的身体状态和未来可能发生的疾病，充分学习两者的关联。

传统的感知机结构在处理时序数据时，只能按照顺序依次读入并单独计算输出，这种方法没有考虑到时序数据前后的关联性。循环神经网络（RNN）的基本结构中既有原始感知机的前馈通路，又有反馈通路。新增加的反馈通路可以将过去时刻的计算结果输送到当前时刻，与此时的原始输入一起作为神经元计算的输入，在实际使用过程中，为了降低模型的复杂度，通常只使用上一时刻的状态。经典的循环神经网络结构如图所示。研究学者还针对原始模型不同方面的缺点，设计了很多循环神经网络的改造模型。例如，Elman 等人在 1990 年提出的 Simple RNN 来解决传统标准的多层感知机无法处理序列数据的问题[41]，Schuster 等人在 1997 年提出的双向循环神经网络（BiRNN）从前向和后向两个角度解决时序问题[42]，还有 Hochreiter 等人提出的长短期记忆机制（LSTM）解决长时间序列的依赖问题[43]。

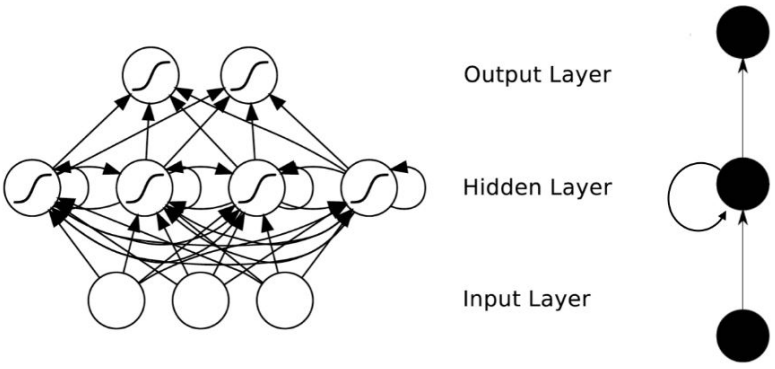

图 2.11　循环神经网络结构

2.4.3 基于卷积神经网络的数据表示算法

卷积神经网络有非常悠久的历史，最早由 Hubel 等人在上世纪 60 年代根据视觉解剖学的研究发现视觉的局部感受野（Receptive Field）机制[44]，基于局部感受野概念 Fukushima 等人在 1980 年提出了的自组织多层神经网络模型[45]，通过引入反向传播算法 Lecun 等人在 1998 年提出了适用于手写数字识别的 LeNet-5 模型将卷积神经网络概念转为实际可用的模型[46]，随着深度学习的发展 Krizhevsky 等人在 2012 年的 ImageNet 上使用卷积神经网络取得巨大成功[47]，标志着卷积神经网络模型进入快速扩展期。

随着医疗领域的检验技术发展，使用仪器对人体进行断层扫描和三维影像采集成为一种趋势，在为医生提供清晰直观的病人数据的同时，也产生大量隐含着病人体征信息的影像数据。通过计算机程序从海量医疗图像数据中学习病人的体征表示，成为具有现实意义的研究方向。作为受到生物学实验启发的神经网络模型，卷积神经网络在处理图像数据时具有天然的可解释性，通过学习图像中的纹理特征并进行非线性组合，能够自动地学习到计算机可理解的数据特征。得益于卷积神经网络的发展，医疗领域跨网络的数据虽然表示形式复杂，但都可以使用卷积神经网络根据局部的特征提取方式，得到有效的数据表示。Cheng 等在 2015 年对电子医疗记录进行疾病发病率预测[48]，Gulshan 等在 2016 年就使用卷积神经网络对糖尿病病人的眼底照片进行分析，建立起眼底病变的实际预测模型[49]；Esteva 等在 2017 年也成功将卷积神经网络应用到皮肤癌的检测和分类上，并且实现了专业皮肤病专家的水平，其成果刊登在 Nature 杂志上[50]。

卷积神经网络的典型结构由输入层、卷积层、下采样层（池化层）、全连接层和输出层组成。卷积的概念是指一个卷积核在一幅原始图像 X 上，依照一定顺序进行对图像的局部进行卷积操作，最后通过非线性的激活函数得到卷积层的输出，公式表示为 $H_i = f(H_{i-1} \otimes W_i + b_i)$。因为经过卷积操作的输出结果仍然有很高的维度，所以引入下采样层（池化层）进行降维的池化操作，在保证图像特征尺度不变的前提下尽量压缩数据维度，通常采用最大函数作池化。最后将图片在一维角度上展开，形成一个向量与全连接神经网络相连，进行进一步的非线性特征组合，最后使用激活函数得到应用相关的输出层结果。

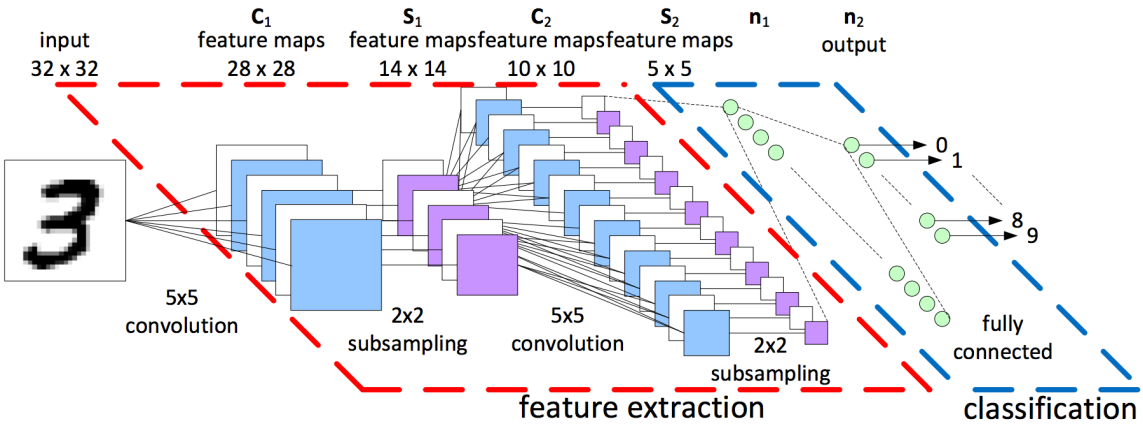

图 2.12　卷积神经网络结构

2.4.4　深度神经网络模型

深度神经网络模型通常是指在神经网络中，隐含层的数量超过一定数量后，导致梯度计算出现困难而需要采用新的训练方法的模型。

在医疗领域，通常涉及到诊断、药品、诊疗过程等诸多涉及健康分析的概念。但是，对于病人在医院内的一次就诊记录而言，生成的电子医疗记录包含了其中多种概念，可以分别反映出序列中的次序关系和一个序列中概念的相关性情况。在 Choi 等人在 2016 年提出的基于多层感知机的医疗概念学习模型中[51]，作者提出的模型不仅可以学习医疗概念分布式表示，还能有很强的可解释性并由医疗专家验证。算法不使用额外的专家知识，就能通过简单且具有鲁棒性方法来学习医疗表示,，并且在大规模的数据集上有良好表现。

第 3 章　医疗实体关系网络中的表示学习

电子医疗记录（Electronic Health Record，EHR）是当前医疗机构在病人住院期间，将所有医疗过程数据化记录的结果，其中包含了病人的生理监控数据、医生给出的诊断结果、病人接受的治疗过程等等。合理使用这些信息丰富的电子医疗记录，结合医疗辅助系统帮助医生进行诊断，是实现医疗研究信息化和诊疗个性化的基础。但是，面对海量原始数据和不统一的数据格式，对于数据聚合和表示学习任务确实是巨大的挑战，阻碍电子医疗记录在辅助诊疗的过程中发挥更大的作用。

为了使用这些电子医疗记录数据进行疾病预测，本文提出了一种基于跨网络向量化表示和深层神经网络的疾病预测模型。可以通过对网络结构的表示学习，无监督地自动抽取数据降维方法，同时结合疾病网络的先验知识进行预测结果约束。模型分为三个部分，第一部分使用网络结构化信息对医疗实体关系网络中的疾病向量化表示进行学习，同时得到疾病之间的相似性矩阵；第二部分使用降噪自编码器栈实现电子医疗记录数据的表示学习，降低数据维度；第三部分使用多层神经网络模型，将前两部分的结果相结合进行多标签的疾病预测。本章节将介绍第一部分，关于医疗实体关系网络中的表示学习解决方案及实现，具体流程如图3.1所示。

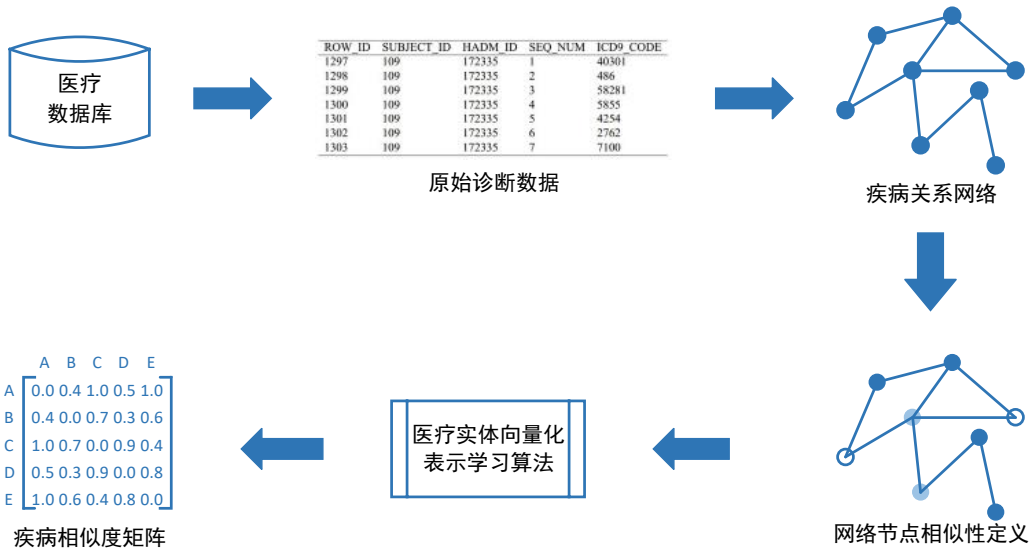

图 3.1　医疗实体关系网络中的表示学习算法流程

表 3.1 原始诊断数据

ROW_ID	SUBJECT_ID	HADM_ID	SEQ_NUM	ICD9_CODE
1297	109	172335	1	40301
1298	109	172335	2	486
1299	109	172335	3	58281
1300	109	172335	4	5855
1301	109	172335	5	4254
1302	109	172335	6	2762
1303	109	172335	7	7100

3.1 医疗实体关系网构建

一个医疗实体关系网络定义为$G = (V, E)$的形式，其中V代表网络中疾病实体顶点的集合，E代表实体之间产生关系连接的集合。对于顶点u和v，边为$e = (u, v) \in E$，连边上的权重为w_{uv}。

有了定义好的医疗实体关系网络，就需要从实际的医疗数据中抽取对应的数据来构建网络。一个病人在医疗数据库 MIMIC-III 中都对应一个唯一的身份编号$subject_id$和每次住院时分配的$hadm_id$，根据这两个唯一编号就能在数据库中检索出该病人在这次住院期间的相关数据，其中就包括病人的疾病诊断数据。当病人住院并经过检查和处理后，医生就会给每个病人做出一系列诊断，这些诊断结果在数据库中以诊断代码的形式保存。诊断代码的形式是一个最长五位数的国际疾病与相关健康问题分类标准编码（ICD-9），小数点前的数字代表疾病的大类别，小数点后的数字代表该类别下的具体疾病。例如 288.9 代表白血病，401.9 代表高血压等。从医疗数据库中检索得到的原始数据如表3.1所示。

根据数据库中获取的医疗诊断结果表，每条诊断结果包含若干疾病的 ICD-9 代码，表示病人被诊断出了这些疾病。当两种疾病同时出现在一个病人的诊断结果中，说明这两个疾病之间可能存在潜在的关联，体现在疾病网络中也就是两个疾病节点之间因为一次共同出现有了一条边。但是，为了使用病人的疾病诊断数据组成医疗实体关系网络，还需要进行适当的数据转换。

首先，由于 ICD-9 编码数量巨大，在实际使用的过程中，需要先将数据进行归约操作。根据编码之间的层级关系，聚合为数量较少的类别编码。归约操作后，将得到一个诊断代码频次统计矩阵，每一行表示一个病人在某次住院时的诊断，其中某列的数值为 1 表示病人被诊断出该种疾病。

接着，需要根据统计矩阵构造医疗实体的关系。算法中定义两个疾病实体共同出现在同一个病人的诊断记录中，就表示这两个实体有一条连边。两种疾病之间可能存在若干次共同出现的情况，就将共同出现的次数作为两个节点之间连边的强度。最后，可以得到疾病网络的文本表示，即 $< u, v, w_{uv} >$ 这样的一个三元组代表节点之间的连边关系。

最后，对于某些权重过大的边，要进行适当的降权，避免权重对后续的网络实体向量化表示学习造成影响。

3.2 医疗实体网络相似性定义

构建好医疗实体网络后，还需要根据医疗领域的特定关系，对医疗实体在网络上的相似性做出定义。受到社交网络中用户相似性分析启发，在分析医疗实体网络中节点之间关系时，提出了两种医疗实体相似性关系。

3.2.1 显式相似性

在社交网络中，如果一个用户节点与另一个用户节点存在连边，则代表这对用户有相互关注的操作，潜在地体现出一对用户之间存在显式的兴趣关系。将这种显式关系带来的相关性引申到疾病网络中，就表现为两个疾病节点因为同时在病人身上被诊断出来，说明这两种疾病之间存在的强相关性。例如，当病人由于外力受伤时通常会有肌肉损伤和流血等症状，糖尿病和眼底病变也可能同时出现在病人身上等。这样的强相关性可以形式化地记录为显式相似性，并使用节点连边强度 w_{uv} 代表这种显式关系的强弱。

3.2.2 隐式相似性

但是，在将现实世界中的关系保存为网络结构的形式时，通常只能智能保存下很少量的显式关系，大部分的实体相似性都是通过隐式关系体现出来的。例如，在社交网络中，即使你和同事之间并没有在虚拟的社交网络中由互相关注的操作，现实世界中对应的工作关系依然存在。同理，在医疗实体关系网中，没有在同一个病人身上同时发生的疾病，也并不是完全没有关联的。如果仅依据显式相似性评价医疗网络，由于没有连边，两个潜在相关的医疗实体就无法体现相关性。所以，还必须从其他角度

考察网络中的两个节点之间的相似性关系。

虽然很多实体之间不存在直接关系，但是由于实体间的隐式相似性必然导致两个实体之间共享一部分邻居节点。例如根据社交网络的兴趣属性，自然就能想到从一个人的朋友圈关系，去分析他的兴趣爱好。当一个用户关注了很多体育明星时，即使不存在显式的连边，他与另一个关注了相同体育明星的用户之间，仍然还存在很多共同兴趣。所以就可以通过观察一个节点的邻居信息，来定义节点对之间的隐式相似性。

隐式相似性定义为节点对 (u,v) 之间所属的邻居网络结构的重合程度。对于节点 u，它与整个网络中的节点之间的显式相似性用向量 $\vec{p_u} = (w_{u,1}, ..., w_{u,|V|})$ 表示，对于节点 v 同样也可以得到这样一个显式相似性向量 $\vec{p_v}$ 表示。隐式相似性就是对于两个显式相似性向量表示 p_u 和 p_v 之间的相似度比较，可以选择余弦相似度计算公式或者其他计算公式，计算得到的数值越高，说明这两个节点之间的隐式相似度越高。

3.3 医疗实体向量化表示学习

对特定领域内的概念学习一个对应的表示，已经被证实是机器学习任务中的一个非常重要的基础。在医疗领域，可以从两个角度理解数据。一个角度是考虑病人在同一次住院期间显式的疾病并发症关系；另一个角度是根据病人每次在医院的来访记录，探索疾病之间的隐式关系。为了综合这两种关系学习得到对于医疗实体的向量化表示，就需要分别从显式角度和隐式角度学习，再将结果结合到一起，这种学习过程如伪代码1所示。接下来的部分将详细说明显式相似性学习和隐式相似性学习的过程。

3.3.1 基于显式相似性学习向量化表示

首先对网络节点的显式相似性进行建模。当存在一条边 (i,j) 时，节点 v_i 和 v_j 之间的显式相似性定义为：

$$p_1(v_i, v_j) = \frac{1}{1 + exp(-\vec{u_i}^T \cdot \vec{u_j})} \tag{3.1}$$

其中 $\vec{u_i} \in R^d$ 是顶点 v_i 在低维特征空间中的向量化表示，也是最终需要学习到的形式。根据上述公式可以将这个网络对应的 $|v|$ 个顶点之间的显式相似性都计算出来，得到 $p_1(\cdot,\cdot)$。而根据网络结构本身的信息，显式相似性关系代表的节点间直接关系强弱可以用边的强度表示，此时节点 v_i 和 v_j 之间的理想显式相似性是 $\hat{p}_1(v_i, v_j) = \frac{w_{ij}}{W}$，

Algorithm 1 医疗实体向量化表示算法伪代码

Input: 医疗实体关系元组 $< v_i, v_j, w_{ij} >$
Output: 医疗实体相似度矩阵 **sim**

```
根据实体关系元组，建立医疗实体图 G = (V, E)
初始化显式相似性对应向量化表示矩阵
for step < step_max do
    随机从边集合 E 中抽取一条边 e
    计算 e 对应的节点显式相似度，同时更新显式向量表示矩阵
end for
初始化隐式相似性对应向量化表示矩阵
for step < step_max do
    随机从顶点集合 V 中抽取一个顶点 v_i
    for v_j ∈ V and v_j ≠ v_i do
        计算 v_i 与 v_j 的隐式相似度
    end for
    根据真实节点相似性计算 v_i 对应的损失函数，同时更新隐式向量表示矩阵
end for
将显式和隐式相似性向量表示结合，计算医疗实体相似度矩阵 sim
```

$W = \sum_{(i,j)\in E} w_{ij}$ 是所有边上强度的累加和作为归一化因子。模型的训练目标就是让根据学习到的低维向量化表示得到的显式相似性与真实的显式相似性越接近越好，用数学公式描述最小化差距就是：

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{3.2}$$

$$\begin{aligned} O_1 &= d(\hat{p}_1(\cdot,\cdot), p_1(\cdot,\cdot)) \\ &= \sum_{(i,j)\in E} \hat{p}_1(i,j) log \frac{\hat{p}_1(i,j)}{p_1(i,j)} \\ &= -\sum_{(i,j)\in E} \frac{w_{ij}}{W} log\left(p_1(i,j)\frac{W}{w_{ij}}\right) \\ &= -\frac{1}{W}\sum_{(i,j)\in E} w_{ij} log(p_1(i,j)) - \frac{1}{W}\sum_{(i,j)\in E}\frac{W}{w_{ij}} \end{aligned} \tag{3.3}$$

其中 d 函数表示计算两种概率分布之间距离的函数，当使用 KL 散度公式3.2 作为计算函数时，省略公式3.3中的常数项，得到公式3.4：

$$O_1 = -\sum_{(i,j)\in E} w_{ij} log p_1(v_i, v_j) \tag{3.4}$$

3.3.2 基于隐式相似性学习向量化表示

在衡量隐式相似性时，假设一个节点和另一个节点共享很多邻居节点，代表了隐式相似性高的情况，此时一个顶点既可能是一个待比较的原始点，也是作为其他节点邻居时的上下文节点。为了体现节点在扮演这两种不同角色时的不同特性，需要使用两种向量化表示来表示。对于顶点 v_i，向量 $\vec{u_i}$ 是作为顶点时的向量化表示，向量 $\vec{u_i}'$ 是作为上下文时的向量化表示。那么，对于一条有向边 (i,j) 来说，上下文节点 v_j 有顶点 v_i 生成的概率为：

$$p_2(v_j \mid v_i) = \frac{exp(\vec{u_j}'^T \cdot \vec{u_i})}{\sum_{k=1}^{|V|} exp(\vec{u_k}'^T \cdot \vec{u_i})} \tag{3.5}$$

其中 $|V|$ 表示顶点 v_i 的所有邻居节点个数。根据上述公式，能够得到表示由向量化表示学习计算得到的顶点 v_i 的隐式相似性概率分布 $p_2(\cdot \mid v_i)$，学习的目标函数就是最小化隐式相似性概率分布和真实分布 $\hat{p}_2(\cdot \mid v_i)$ 之间的差距，表示为：

$$\begin{aligned} O_2 &= \sum_{i \in V} \lambda_i d(\hat{p}_2(\cdot \mid v_i), p_2(\cdot \mid v_i)) \\ &= \sum_{(i,j) \in E} \hat{p}_2(v_j \mid v_i) log \frac{\hat{p}_2(v_j \mid v_i)}{p_2(v_j \mid v_i)} \\ &= -\sum_{(i,j) \in E} \frac{w_{ij}}{d_i} log \left(p_2(v_j \mid v_i) \frac{d_i}{w_{ij}} \right) \\ &= -\frac{1}{d_i} \sum_{(i,j) \in E} w_{ij} log p_2(v_j \mid v_i) - \frac{1}{d_i} \sum_{(i,j) \in E} \frac{d_i}{w_{ij}} \end{aligned} \tag{3.6}$$

其中，真实条件概率分布 $\hat{p}_2(v_j \mid v_i) = \frac{w_{ij}}{d_i}$ 可以根据节点 v_j 和节点 v_i 之间的边强度，占从节点 v_i 出发的所有边强度总和的比例表示，其中 $d_i = \sum_{k \in N(i)} w_{ik}$。当同样适用 KL 散度作为评估概率分布相似性的方式时，隐式相似性的学习目标方程3.6省略公式中的常数项，得到如下结果：

$$O_2 = -\sum_{(i,j) \in E} w_{ij} log p_2(v_j \mid v_i) \tag{3.7}$$

通过学习两种向量化表示 $\{\vec{u}_i\}_{i=1...|V|}$ 和 $\{\vec{u}'_i\}_{i=1...|V|}$，就能够学习到每个顶点对应的低维向量化表示 $\vec{u}_i$。

对于之前学习到的基于显式相似度的向量化表示和基于隐式相似度的向量化表

示，可以将两个特征向量拼接到一起，作为整体表示医疗网络中顶点的向量化表示。

3.4 疾病相似矩阵建立

通过上述的方法，得到了两种网络中节点的向量化表示。在医疗疾病网络中，可以使用余弦相似度来度量这些低维特征空间中的向量表示。顶点 v_i 和 v_j 的余弦相似度定义为：

$$sim_{ij} = \frac{\sum_{k=1}^{d} u_{ik}u_{jk}}{\sqrt{\sum_{k=1}^{d} u_{ik}^2} * \sqrt{\sum_{k=1}^{d} u_{jk}^2}} \tag{3.8}$$

计算疾病网络中所有疾病节点的相似度，能够得到疾病相似矩阵，其中的每个值都是显式和隐式相似度向量拼接后的向量表示计算得到的。这个相似矩阵在后续的模型中将作为一种先验知识对模型优化函数进行约束。

第 4 章　基于表示的跨网络数据融合与模型验证

面对海量医疗领域跨网络数据的表示问题，本文提出的算法主要从两方面入手进行解决。一方面是使用降噪自编码器模型降低医疗数据维度，另一方面是将算法自动学习的专业医疗领域知识与疾病预测任务融合，本章将详细说明上述两个创新点的具体实现步骤。

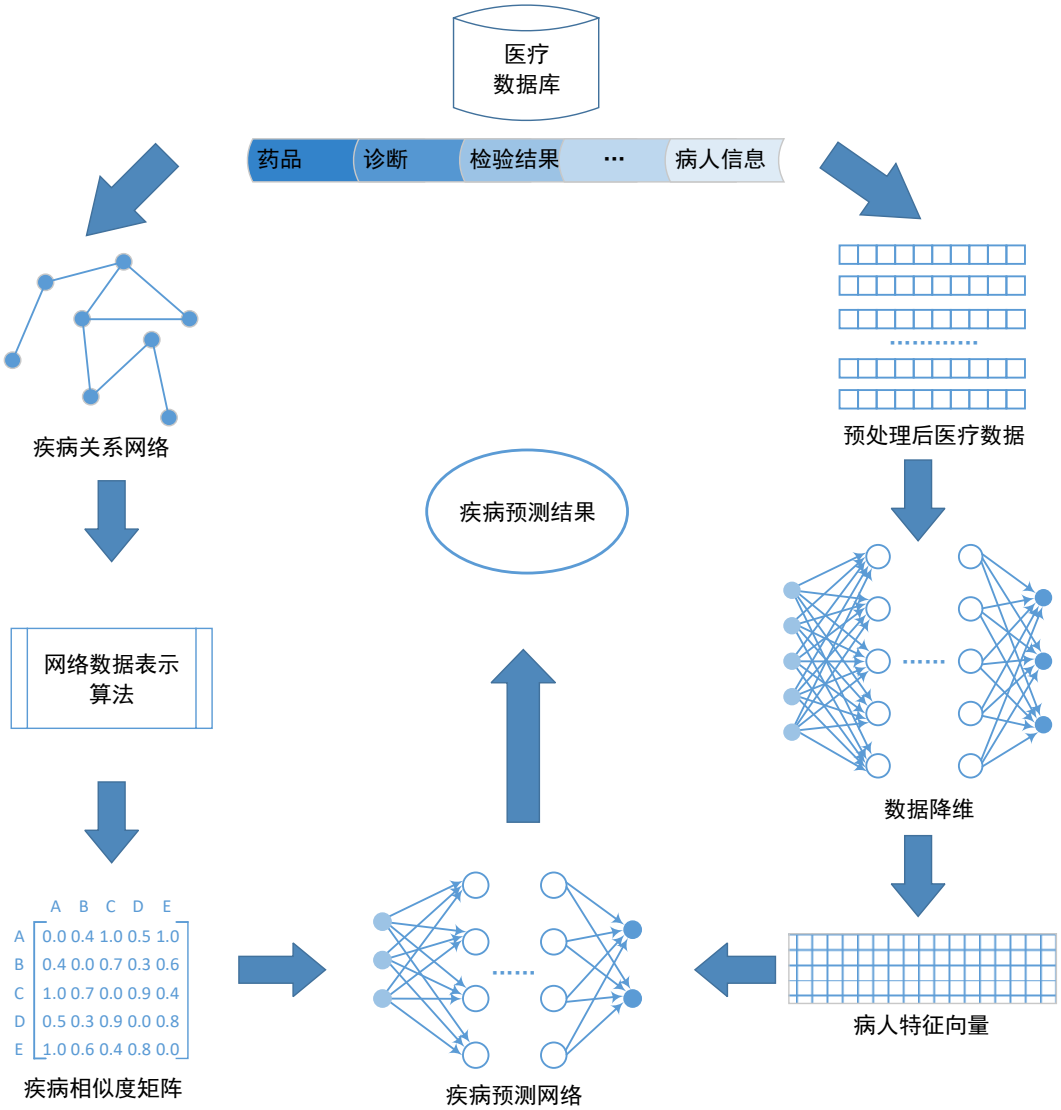

图 4.1　跨网络数据表示学习算法流程

本研究方法的具体思想是，模拟医生进行诊断的过程中，参考实验室检验结果进

行病情分析这一步骤。使用含有丰富信息的原始检验结果数据进行病人特征学习，再通过多层神经网络强大的非线性拟合能力进行预测。在实现疾病预测任务时，算法使用病人在医院期间的实验室检验结果，作为病人身体状况的数值化表现。针对这些数值数据，算法使用降噪自编码器作为特征提取方法，并结合多层神经网络模型进行多标签预测任务，最后与上一章节学习到的疾病相似度矩阵想结合，完成对潜在疾病的预测。整个任务的流程如图4.1所示。

算法模型主要可以分为三部分。第一部分如图左侧所示，对于从原始医疗数据集中抽取出的疾病关系网络，使用上一章节介绍跨网络数据表示算法，学习得到疾病相似度矩阵；第二部分如图右侧所示，使用降噪自编码器模型对抽取出的人口统计学信息和实验室实验室检验结果，进行数据降维后得到病人的特征向量；第三部分如图中下部所示，建立多层神经网络模型，利用降维后的特征向量进行疾病的多标签预测，同时根据疾病相似度矩阵计算损失函数，最终得到病人的潜在疾病预测结果。

4.1 医疗数据特征向量提取

4.1.1 原始数据处理

为了从原始的医疗数据库 MIMIC-III 中抽取病人的特征向量，首先要根据病人编号和入院时的住院编号，将与病人相关的实验室检验结果数据检索出来。由于检索结果中采集时间、数值单位等信息和后续处理无关，首先会去除这些数据，最终检索得到的原始数据如表5.1所示。表中最重要的是代表检查项目的 ITEMID 和表示数值结果的 VALUENUM，此外还有代表检查项目是否超标的标记位 FLAG。此外，因为一个病人在住院期间可能会进行多次检查，还需要对多次结果做均值处理，作为对应身体指标的结果。统计检索得到的原始数据，得知数据总共包含 753 个检查项目。仿照文本处理中词向量的表示方法进行数据处理，使用一个 753 位的向量存储检验结果，其中向量的每一位都对应一种检查项目和检查后的数值。最终，可以得到所有病人住院期间的实验室检验结果向量 V_{lab}。

除了实验室检验结果数据，要建立病人特征向量需要使用到的其他数据来源，就包括病人本身的人口统计学信息。人口统计学信息属于病人主动填写或者由门诊人员主动收集的，属于高可信度的信息。使用人口统计学信息，不仅可以更全面地了解患者的基本信息、增加对病症的描述，还能帮助算法根据这些信息进行特征筛选。例如，

男性病人不可能患有女性独有的卵巢癌，幼儿和青少年患有风湿病的比例较低，婚姻状况不同的女性患有某些生殖系统疾病的几率不同，等等。所以，需要根据病人编号从 MIMIC-III 数据库中检索对应人口统计学信息，包括病人的性别、年龄、婚姻状况、地区和保险情况信息等。检索得到这部分信息后，就可以与之前获得的实验室检验结果向量 V_{lab} 拼接在一起，构成病人的原始特征向量 V。

根据上述步骤得到的病人特征向量没有经过预处理，向量中的数值对应的数据范围不同，并且有缺失值等问题。如果直接使用这些原始数据会影响算法的计算，降低预测准确度，所以还需要进行一次归一化操作。归一化计算首先需要根据 $max(\vec{v_i})$ 和 $max(\vec{v_i})$ 函数得到对应特征维度上的最大值和最小值，再根据转换函数将数值线性映射到 $[0-1]$ 的区间上。最终得到的经过归一化后的病人特征向量将提供给后续计算使用。

$$x^* = \frac{x - min}{max - min} \tag{4.1}$$

4.1.2 医疗数据特征学习

病人在实际住院的过程中，医生会根据初步诊断结果，有针对性地进行检查筛选。由于这种有目的性的筛查通常集中在某一类检验项目上，数量上并不多，所以在上一步抽取出病人特征向量后，这时候的特征向量数据还是非常稀疏的。医疗数据的这种稀疏性，不仅会给存储系统增加很多不必要的负担，还会影响后续预测模型的训练。当输入数据稀疏性很高时，模型必须对每一位特征都设定对应的参数，在训练过程中，如此庞大的参数很难进行更新操作，而且会影响预测的准确率。所以必须使用有效的降维方法，在原始特征向量的基础上，进一步降低维度，得到更加稠密的特征向量。

本文提出的医疗领域跨网络数据表示学习算法中，使用的是降噪自编码器栈模型，目标是将特征向量的维度从原始的 758 维降低到可接受的程度。图4.1右侧部分，从宏观的角度展示了如何从原始的电子医疗记录中学习病人的向量化表示。数据降维模型首先从数据库中抽取与病人相关的医疗数据并进行预处理，然后使用降噪自动编码器栈模型，逐层地对特征向量进行学习。

降噪自编码器栈（stack of denoising autoencoders，SDA）模型，可以实现无监督地对原始数据进行降维和特征抽取。其中模型的输入是病人的原始特征向量 $\vec{x_0} \in [0,1]^{d_0}$，模型首先对向量做变换并加入噪声。特征向量加入噪音后变成 $\hat{\vec{x_0}} \sim p(\hat{\vec{x_0}} \mid \vec{x_0})$，则第

一层降噪自编码器的映射函数为：

$$y_1 = f_\theta(\hat{x_0}) = s(W_0\hat{x_0} + b_0) \tag{4.2}$$

得到第一层隐含层表达 $y_1 \in [0,1]^{d_1}$，其中 d_1 是第一层隐含层的神经元数量。对应的解码器映射函数为：

$$z_1 = g_{\theta'}(y_1) = s(W_0'y_1 + b_0') \tag{4.3}$$

得到第一层重构向量表达 $z_1 \in [0,1]^{d_0}$。要学习降噪自编码器对应的最优参数就是寻找满足一下最优化函数时的参数解：

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} L(x_0, z_1) = \arg\min_{\theta,\theta'} \frac{1}{N}\sum_{i=1}^{N} L\left(x_0^{(i)}, z_1^{(i)}\right), \tag{4.4}$$

其中的 $L(\cdot)$ 是损失函数，N 是训练集中病人的数量。在本模型中使用重构交叉熵函数作为损失函数：

$$L_H(x_0^{(i)}, z_1^{(i)}) = -\sum_{k=1}^{d_0}[x_{0k}^{(i)} log z_{1k}^{(i)} + (1 - x_{0k}^{(i)}) log(1 - z_{1k}^{(i)})] \tag{4.5}$$

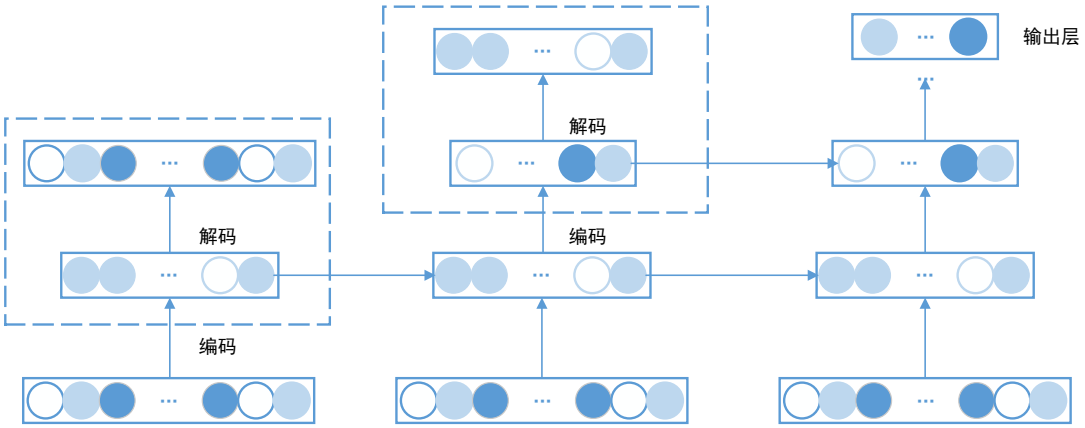

图 4.2　降噪自编码器逐层学习过程

因为这里使用的是由多层降噪自编码器组成的栈式结构，在学习参数的过程中，需要按照上述方法对每一层逐层训练。如图4.2所示，需要先将降噪自编码器栈的第一

层训练完毕后，将参数固定下来，再训练第二层的参数，以此类推，直到所有层的参数都学习完毕。这种逐层训练的方法在简化计算复杂度的同时，还模仿了人脑处理信息的过程，也就是先将粗糙的外界知识提炼成较为抽象的类似语言和文字的表示，然后在根据语言和文字的抽象表示学习更为抽象的信息，最后才根据高层抽象信息进行推理判断任务。

4.2 多标签疾病预测模型

上一小节中完成了对病人特征向量的抽取和降维，得到一个长度为 200 维的低维特征空间中的稠密向量，现在就要使用多层神经网络模型建立对患者潜在疾病的预测模型。在实现疾病预测模型时，使用了深度学习框架 TensorFlow 进行编程。整个多标签疾病预测模型的伪代码如算法2所示：

Algorithm 2 多标签疾病预测模型训练过程伪代码

Input: 降维后的病人特征向量 $\vec{x_i}$，病人发生疾病的标记向量 $\vec{y_i}$
Output: 多标签疾病预测模型

```
读取数据，将 tensor 存入对应的 placeholder
初始化神经网络中所有的权重矩阵 weight 和偏置 bias
for layer_i < N_dae do
    while epoch < pre_train_epochs do
        对降噪自编码器栈的 layer_i 层进行预训练
    end while
end for
for step < step_max do
    从训练集中随机读取一批数据
    使用模型得到预测结果，结合相似度矩阵计算损失值
    与真实值比较，得到函数综合损失值
    使用随机梯度下降算法更新参数
end for
```

与一般情况下的只预测单疾病的发病情况不同，医疗领域的疾病本身就带有并发症的特性，例如糖尿病可能诱发病人眼底发生病变、白血病会造成免疫力下降导致免疫疾病等等。因为不同疾病之间不是互斥的关系，所以在设计预测模型时不能够使用传统的基于 Softmax 的概率模型，而是需要对每个待预测的疾病都建立一个对应的预测神经元。结构如图4.3所示。

隐含层和输出层之间使用非线性的 ReLU 函数作为激活函数，因为采用 sigmoid 函数时需要进行指数运算，而且在反向传播求解误差梯度的时候计算量大，而采用

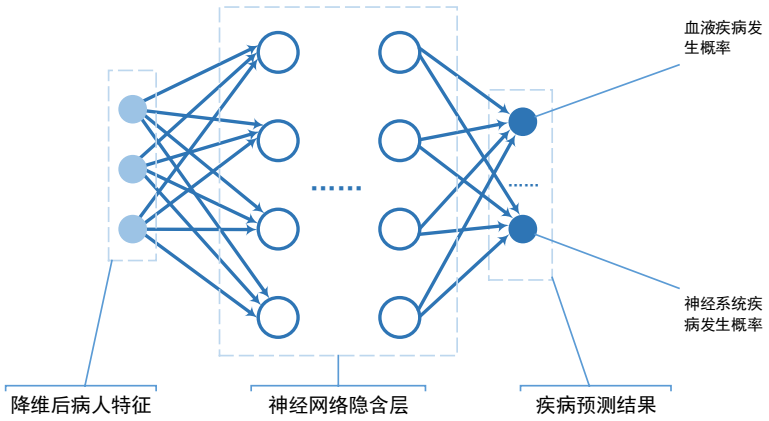

图 4.3　多标签疾病预测模型结构

Relu 激活函数就可以节省很大部分的计算量。

为了实现对潜在疾病的多标签预测，本研究中采用的是多层全连接神经网络模型，在神经网络使用上一层的输出作为输入，在结合权重以及偏置进行计算后，根据激活函数输出值传递到下一层。每一层神经网络使用的公式如下：

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{4.6}$$

其中 $\vec{w}$ 表示权重向量，$\vec{x}$ 表示输入向量，b 表示偏置。$f(\cdot)$ 是神经网络中使用的激活函数，本模型中使用 Sigmoid 函数：

$$f(z) = \frac{1}{(1+e^{-z})} \tag{4.7}$$

4.3　基于医疗背景知识的参数优化算法

医生在诊断时需要依据专业的医疗背景知识，判断疾病之间的发病关联。在本文提出的模型中，使用上一章中学习到的疾病相似度矩阵，作为预测结果的概率分布先验条件，就可以模拟医疗知识库对预测结果进行约束的情况。这种改进方法使得跨网络的数据表示学习算法，既能够得到体现病人实时身体状况的预测结果，又能融合专业医疗知识给出专业诊断，实现对病人的潜在疾病精准预测。

首先，需要根据疾病预测结果和真实结果的差距，计算损失函数。特征向量经过多层神经网络的计算后得到的是一个疾病发病概率向量，每个神经元代表的是一种疾

病和其可能的发病概率。此时，整个模型的学习目标是最小化预测结果和真实结果之间的差距，所以选择交叉熵函数作为损失函数的计算公式如下：

$$Loss_1 = -\sum_{k=1}^{q}[y_k log(y'_k) + (1-y_k) log(1-y'_k)] \tag{4.8}$$

其中 q 是待预测的疾病数量，y_k 代表病人在第 k 个疾病上真实的发病情况，y'_k 是由多层神经网络计算得到的疾病发病概率，整个损失函数从预测准确率的角度对模型的预测情况进行评价。

接着，需要根据上一章中得到了疾病之间的相似矩阵，计算预测结果与这部分先验知识之间差距对应的损失值。因为对于医学领域而言，有很强的因果相关性，如果学习的模型给出完全无关的两个疾病同时发病的预测，需要进行适当的惩罚，对于模型根据先验知识给出有内在关联的疾病，同时发生的概率很高的情况，需要进行鼓励。使用这种方法可以让学习到的模型具备现实中专业医生那样的判断逻辑，能够通过疾病的潜在相似性给出更具有专业性的预测结果。假设对于 q 个待预测疾病，模型得到的预测概率向量是 $\vec{y'} = (y'_1, y'_2, ..., y'_q), \text{for} y'_k \in [0,1]$，之前学习到的疾病相似矩阵为 $\mathbf{D}, \text{for each} d_{ij} \in [0,1]$，可以得到根据先验知识计算的损失函数为：

$$Loss_2 = -P_0(\vec{y'}) \propto -exp(\vec{y'}^T \mathbf{D} \vec{y'}) \tag{4.9}$$

其中 $P_0(\cdot)$ 表示基于先验知识的概率分布，它和疾病预测向量与相似矩阵的运算结果成正比，在模型中使用后者代替，表示预测结果向量和先验知识的匹配程度。这部分损失值越大，说明预测结果向量和先验知识越不匹配。

最后，还需要从模型参数复杂度的角度考虑，增加针对复杂度的惩罚项，让模型可以学习出尽量精简的参数组合。

$$Loss_3 = || \ w \ ||_k^2 \tag{4.10}$$

模型将以上三个部分的损失值累加到一起，通过设定合适的权值，就得到了最终的损失函数为：

$$
\begin{aligned}
J(w,b) &= Loss_1 + Loss_2 + Loss_3 \\
&= -\sum_{k=1}^{q}[y_k log(y'_k) + (1-y_k) log(1-y'_k)] - \gamma_1 log P_0(\vec{y'}) + \gamma_2 ||w||_k^2
\end{aligned}
\tag{4.11}
$$

根据计算得到的损失函数，可以进一步使用随机梯度下降方法和方向传播算法，更新多层神经网络中的参数。

$$
\begin{aligned}
W_{ij}^{(l)} &= W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) \\
b_i^{(l)} &= b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W,b)
\end{aligned}
\tag{4.12}
$$

神经网络的最后一层输出的是隐含层经过 sigmoid 函数之后的激活值，属于在区间 $[0,1]$ 上的连续概率值。表示如果概率值大于阈值则说明有可能出现该疾病，反之则不出现。

4.4 模型总结

本章节着重介绍了如何根据病人的实验室检验结果信息和其他辅助信息，使用降噪自编码器栈模型和多层神经网络模型计算病人患有某种潜在疾病的概率。同时使用根据网络结构得到的疾病相似矩阵作为医疗专业知识，采用基于先验知识的参数优化方法，形成最终的疾病预测结果。

第 5 章　测试与评估

本章节中，为了验证本文提出的医疗领域跨网络表示学习算法的有效性，首先针对病人的潜在疾病进行预测，结合多种传统的多标签分类方法设计对比实验。同时，还针对跨网络数据表示方法中的多种向量化学习算法，设计不同参数对应的不同学习方法的优劣对比，探究跨网络数据表示的改进思路。

5.1　数据集

关于数据集，本文使用的是记录了急诊室医疗信息的大型数据库 MIMIC-III[28]，该数据库包含了 2001 年至 2012 年间 Beth Israel Deaconess 医疗中心接诊过的超过四万名病人。为了保护病人隐私权，所有数据都去除了涉及个人隐私的数据，例如对病人记录中对应的日期进行了人为偏移，在保证相对日期正确的情况下，无法通过反推的方式确定病人就诊实际日期以及病人的出生日期。数据内容方面涵盖了病人的人口统计学信息、生命体征测量记录、实验室化验结果、医疗服务记录、药品记录、医生诊断结果等各个方面的数据，基本覆盖了医疗领域的常见电子记录数据，给实验提供了丰富的数据。对于本文中研究的医疗领域跨网络数据的表示研究问题，主要使用了数据集中的人口统计学数据、医生诊断结果和实验室检测结果三方面的数据。这三方面数据组成的数据集，在实验前被随机分为两部分，训练集包含 80% 的数据样本，测试集包含 20% 的数据样本，分别用于模型训练阶段和模型测试阶段。

5.1.1　人口统计学数据

病人入院时会被要求填写基本的个人信息，其中包括了性别、出生日期、婚姻状况等。在处理这部分信息时，为了后续程序处理的方便起见，会将性别和婚姻状况这样的类别信息转换为 0 或 1 的数值型数据，同时将本身就是数值型的年龄数据进行归一化处理，转换为 0 到 1 区间上的连续值。在人口统计学数据中，通常蕴含了疾病类型和病人身体状况的最直接关联，例如很多慢性病是老年人的高发疾病，病人性别有直接排除某些生殖相关的疾病等。使用该部分数据的目的是增加算法的分析数据的维度，避免不合理情况的发生。

统计数据集内病人的年龄分布情况，结果如图5.1所示。从图中可以看出，在 0 至

100 的年龄分布区间内，35 岁以下的病人数量较少，低于 35 岁的病人大约占 7%；从 35 岁开始逐渐增多，35 岁至 60 岁的病人大约占 33%；直至 55 岁达到顶峰期，60 岁以上的病人占 60%。

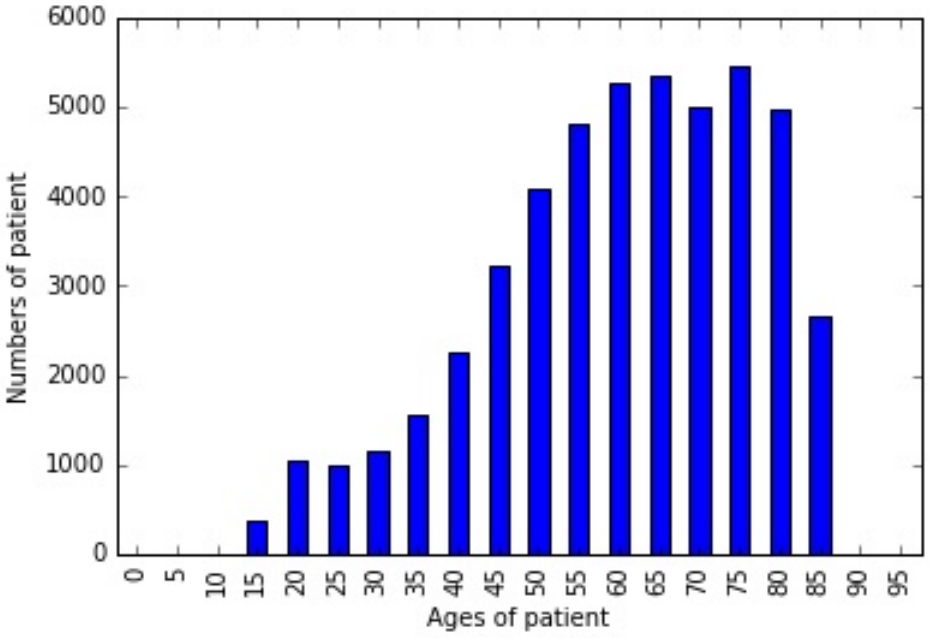

图 5.1　数据集中年龄分布

5.1.2　实验室检验结果数据

实验结果主要包含了 753 种不同检测项目的结果，原始数据如表5.1所示，包含了病人编号、住院编号、检验项目编号、结果数值和异常结果标记，其中的异常结果标记表示该条记录超过了正常范围，属于异常结果，可能代表病人身体中对应部位发生了病变。为了在后续的模型中使用这些病人的原始特征数据，要进行数据清洗和整理的操作。首先，需要根据病人编码和住院编码抽取出发生异常的检验数据向量 $< x_1, x_2, ..., x_753 >$，每个病人的一次住院记录对应一个数据向量，每一项记录了发生异常的检验项目所对应的结果数值；接着，对超出正常值范围很多的结果剔除，进行数据清洗操作；最后，根据每种检验指标的区间范围，使用最大最小化归一函数，将所有 $x_i \in R$ 转为 0 到 1 的区间上 $x_i \in [0, 1]$，便于后续程序读取。

表 5.1　原始实验室化验结果数据

ROW_ID	SUBJECT_ID	HADM_ID	ITEMID	VALUENUM	FLAG
283	3	103219	50802	-1	
284	3	103219	50804	22	
285	3	103219	50808	0.93	abnormal
287	3	103219	50813	1.8	
288	3	103219	50818	33	
289	3	103219	50820	7.42	

5.1.3　诊断结果数据

数据集中的关于诊断结果的数据是以 ICD-9 编码为标准存储的 < 唯一编码，病人编码，住院编码，序列号，ICD-9 编码 > 五元组，其中“唯一编码”用来唯一标识此条诊断结果，“病人编码”和“住院编码”一起组成抽取数据使用的索引，“序列号”用来标识医生的一次诊断结果中若干条疾病诊断的重要性顺序，“ICD-9 编码”标识疾病诊断结果。对于最后的实验部分，只使用 < 病人编码，住院编码，ICD-9 编码 > 的三元部分信息。接着将同一个病人的一次住院记录用向量形式保存，形式为 $< c_1, c_2, ..., c_n >$，向量中的每一个值为一个疾病编码。

由于 ICD-9 编码的数量众多，在实验中无法对所有编码都进行预测，所以根据 ICD-9 编码的层级特性，将同属于一个大类的疾病编码规约为同一编码。例如将以 280-289 开头的凝血功能障碍（286.0）、白细胞数量减少（288.5）等归约为血液和血液相关疾病编码，将以 520-579 开头的胆道疾病（576.9）、胃炎（535.1）、十二指肠炎（535.6）等归约为消化道疾病编码。通过类似的归约操作，将所有具体的疾病 ICD-9 编码转换为数据集中常见的 21 种大类疾病对应的 $[0-20]$ 编码，如表5.2所示。最终得到每个病人住院记录对应的一个长度为 21 的疾病向量，向量上的每一位表示病人时候被诊断出该大类疾病。

5.2　评价指标

本文中提出的医疗领域跨网络数据表示算法，主要是从医疗实体网络的结构表示角度出发，结合神经网络模型对病人的潜在疾病做出预测。模型在进行病人潜在疾病预测的过程中，需要对所有可能出现的疾病都进行逐一预测，最后给出每种疾病的发病概率。疾病预测模型输出的结果是一个多标签分类任务的结果向量 $y =<$

表 5.2　疾病种类对照表

大类编号	IDC-9 编码范围	大类疾病名称
0	001-139	传染病和寄生虫病
1	140-239	肿瘤
2	240-279	内分泌、营养和代谢疾病
3	280-289	血液和造血器官疾病
4	290-319	精神障碍
5	320-359	神经系统疾病
6	360-389	感觉器官疾病
7	390-459	循环系统的疾病
8	460-519	呼吸系统的疾病
9	520-579	消化系统的疾病
10	580-629	泌尿生殖系统的疾病
11	630-679	妊娠并发症、分娩和产褥期
12	680-709	皮肤和皮下组织疾病
13	710-739	肌肉骨骼系统和结缔组织疾病
14	740-759	先天性异常
15	760-779	怀孕期间导致的疾病
16	780-799	症状体征不明确的疾病
17	800-999	损伤和中毒
18	E	其他外力损伤
19	V	补充疾病分类
20	其他	其他情况

$p1, p2, ..., p21 >$，其中每一项的值对应该大类疾病的发病概率，数值越高表示越可能在病人身上出现。

为了模拟多标签预测任务在实际使用情况下的表现，需要从医生使用算法结果的方式上进行分析。当医生收到疾病预测结果的时候，如果直接使用预测结果作为参考指标，会将预测值和真实结果从预测的准确率和结果召回率两方面进行评估。当病人出现的疾病数量比较少的时候，会侧重考察算法模型预测结果是否和病人真是发病情况一致，如果可以成功预测病人发生的疾病就表示一次成功的预测。当病人同时出现多种疾病的时候，会比较算法给出的疾病预测结果是否包含了所有病人的发病，一个优秀的算法需要充分覆盖病人可能出现的所有情况，也就是从召回率的角度进行评估。

5.2.1 准确率与召回率

准确率是用来计算分类器再预测数据集上，正确进行分类的比例。在多标签预测任务中，需要统计每种预测目标的准确率，进行综合判断。首先定义 TP 为正确被推荐的结果个数，FP 为错误被推荐的结果个数，TN 为正确的未被推荐结果个数，FN 为错误的未被推荐的结果个数。则准确率（precision）可以用下列公式计算：

$$P = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{TP_u}{TP_u + FP_u} \tag{5.1}$$

召回率（recall）评价指标是用来评价在一个预测结果集合中，预测准确的疾病数量占病人真实发病情况中疾病数量的比例，这个比例越高说明预测模型得到的结果越接近实际情况。与上述准确率类似，当测试集中有 $|U|$ 个病人时，召回率为：

$$R = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{TP_u}{TP_u + FN_u} \tag{5.2}$$

5.2.2 K 长度准确率

使用准确率作为评价标准时，如果预测结果中包含了病人真实出现的疾病，就代表一次成功的预测，进行预测准确率计算。考虑到医生是根据预测概率，按照从高到低的顺序使用预测结果，所以越靠前的结果被医生关注到的可能性更大。在计算准确率时，如果要将这种顺序信息考虑在内，就需要使用不同长度的预测结果进行计算。对于长度为 N 的结果，准确率计算公式如下：

$$P@N = \frac{1}{|U|}\sum_{u=1}^{|U|}P_u@N = \frac{1}{|U|}\sum\frac{S_u}{N_u} \tag{5.3}$$

其中，$|U|$ 表示测试集中病人的数量，S_u 和 N_u 分别表示针对病人 u 预测正确的次数和预测的总次数。

5.2.3 F1 值

通过上述公式计算得到准确率和召回率后，还可以使用信息检索领域的一种常用评价指标 F1 值（F1-Measure），综合评估预测算法的准确率。假设准确率为 P，召回

率为 R，则 F1 值计算公式为：

$$F1 = \frac{2P \star R}{(P + R)} \tag{5.4}$$

5.3 对比算法

由于在传统的跨网络表示研究中，只涉及网络结构的向量化表示部分，没有针对医疗领域的数据集进行专门的处理和分析。所以在对比实验部分主要针对潜在疾病的多标签预测任务，选择随机森林算法和逻辑回归算法进行实验，能够公平地比较跨网络的数据表示给疾病预测任务带来的预测准确度提升。

5.3.1 逻辑回归算法

逻辑回归是一种基础的分类算法，主要用来解决二分类问题。对于一组包含 N 个有标签的训练数据 $D = (x^1, y^1), (x^2, y^2), ..., (x^N, y^N)$ 的训练集，逻辑回归的目标是使用最大似然法，找到一组合适的参数，能够将样本属于对应标签的概率最大化，概率计算公式如下：

$$p(y = 1 \mid \mathbf{x}; \theta) = \sigma(\theta^T \mathbf{x}) = \frac{1}{1 + exp(-\theta^T \mathbf{x})} \tag{5.5}$$

对于某种疾病的逻辑回归判别模型，当概率大于阈值 0.5 时就判定为会发生这种疾病，概率小于阈值 0.5 时就判定不会发生这种疾病。

5.3.2 随机森林算法

随机森林算法是集成学习的一种模型，具体使用决策树作为基本决策单元，再将多颗决策树的结果通过投票或其他方式综合得到结论。其中的决策树（Decision Tree）模型具有树形的结构，在构建过程中，根据输入特征向量的每一个维度所具有的信息增益，选择信息增益最高的特征作为分裂节点，将所有样本通过这个特征的不同值，划分到不同子树中。如果划分后的样本所对应的标签仍然是不一样的，就需要继续在剩余的特征中选择信息增益最高的特征进行分裂。建立决策树的过程就是不断寻找合适的分裂特征，最终将所有样本划归为一种标签。在预测时，就根据这些特征将样本

表 5.3　不同疾病结果对比

	precision			recall			f1-score		
disease	ours	lr	rf	ours	lr	rf	ours	lr	rf
2	**0.76**	0.75	**0.76**	0.88	**0.89**	0.73	**0.82**	0.81	0.74
4	**0.64**	0.62	0.4	0.21	0.17	**0.4**	0.31	0.27	**0.4**
8	0.7	**0.72**	0.62	**0.71**	0.63	0.61	**0.71**	0.67	0.62
15	**0.85**	0.83	0.79	0.75	**0.76**	0.75	**0.8**	0.79	0.77
19	**0.73**	0.69	0.68	0.74	**0.78**	0.69	**0.73**	**0.73**	0.68

划分到对应的标签中。随机森林就是将若干颗独自建立的决策树集成到一起，通过少数服从多数的投票方式，形成一个强分类器。对于某种疾病的随机森林模型，是否发生疾病服从多数基本判别器的结果。

5.4　实验结果

5.4.1　实验平台及环境

本文提出医疗领域跨网络表示学习算法使用 Python 语言（2.7 版本）和深度学习框架 TensorFlow（r1.2 版本）实现，实验中使用一台装有 NVIDIA Titan 图形处理器（GPU）、Intel Core i7 CPU（3.2GHz, 8core）和 8GB 内存的电脑进行程序运行。

5.4.2　疾病预测结果评估

为了比较本文提出的算法对于潜在疾病的预测效果，评估环节使用了包含 1 万多名病人数据的测试集，与其他多标签预测算法进行了对比实验。由于是多标签预测任务，所有模型针对可能出现的 21 种大类疾病，都分别给出了发病概率值。由于篇幅限制，选择了比较有区分度的 5 种疾病的预测结果如下表5.3所示。表中每行分别对应疾病大类编号为 2、4、8、15、19 的统计值，在准确率（precision）、召回率（recall）和 F1 值（f1-score）三种指标上，分别给出了本文提出的算法（ours）、逻辑回归（lr）和随机森林（rf）算法对应的计算结果，其中加粗数值为针对当前疾病的最高值。

从表中可以看出，在准确率方面，本文提出的算法在 4 种疾病上都有最好的表现，即使是表现较差的呼吸系统的疾病（8 号）也仅落后 2%，说明算法的准确率是对比算法中最好的。在召回率指标上，本文提出的算法比逻辑回归算法差，但仍然由于随机森林算法，说明针对多种疾病同时发生的情况，算法在结果覆盖率方面还有提高的

空间。结合准确率和召回率，从 F1 值的结果上进行综合比较，本文的模型仍旧占有很大的优势，说明总体预测结果优于对比算法。

5.4.3 疾病预测准确度评估

在预测准确度对比实验部分，使用本文提出的医疗领域跨网络数据表示算法和其他多标签分类算法，对病人的潜在疾病进行预测，对于总长度为 21 的结果，针对医生在实际处理预测结果时使用的场景，对于按照概率从高到低排序的结果，分别比较长度从 1 到 21 的情况下，三种模型的准确度如图5.2所示。图中红色圆点的曲线代表的是本文提出的模型，紫色曲线方块点对应的是逻辑回归模型的结果，绿色菱形点的曲线对应的是随机森林模型的预测结果。每种模型都在不同预测结果长度的情况下，计算了疾病预测的准确率。其中横轴表示的是从结果长度 1 到结果长度 21，对应的预测情况。

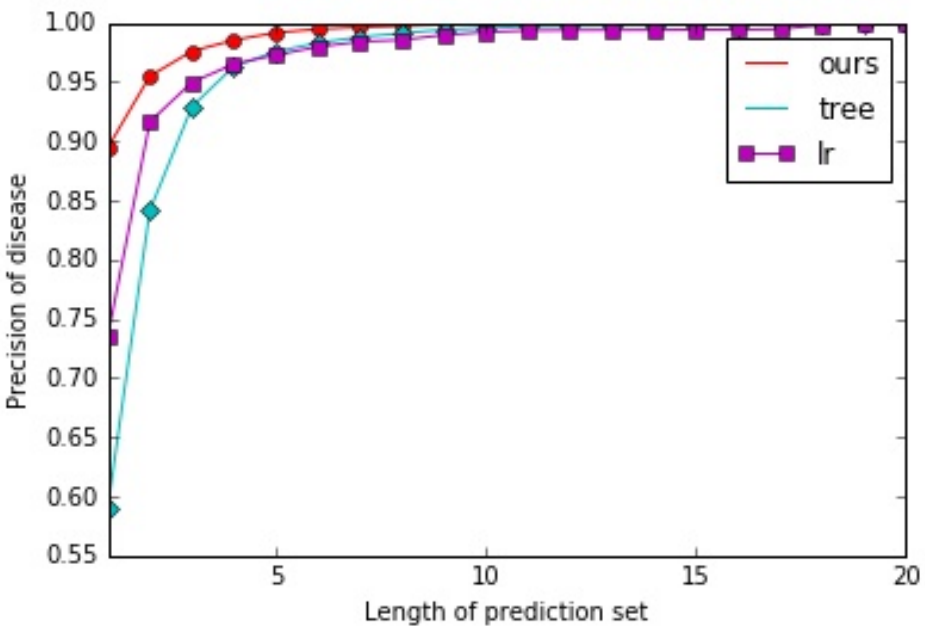

图 5.2 疾病预测准确度对比

从结果图上，可以看出本文提出的模型在任意长度的预测结果中，准确率都超过了对比模型。结果还反映出另一个值得注意的现象，就是当预测结果长度大于 10 的时候，所有模型对应的准确率基本都不再提升。在医院里就诊的真实病人中，大部分都只患有几种疾病，较少出现大量疾病同时发生的情况，这个结果和实际的医疗情况是十分贴合的。其次，预测结果的前 5 项基本上已经覆盖了 95% 的病人患病情况，这

种结果帮助在后续的实际应用中，可以只显示概率最高的前 5 项结果作为疾病预测值。这种设定不仅让呈现出的预测结果更简洁，而且也能够让医生更专注于某些大类疾病诊断。

5.4.4 算法性能指标评估

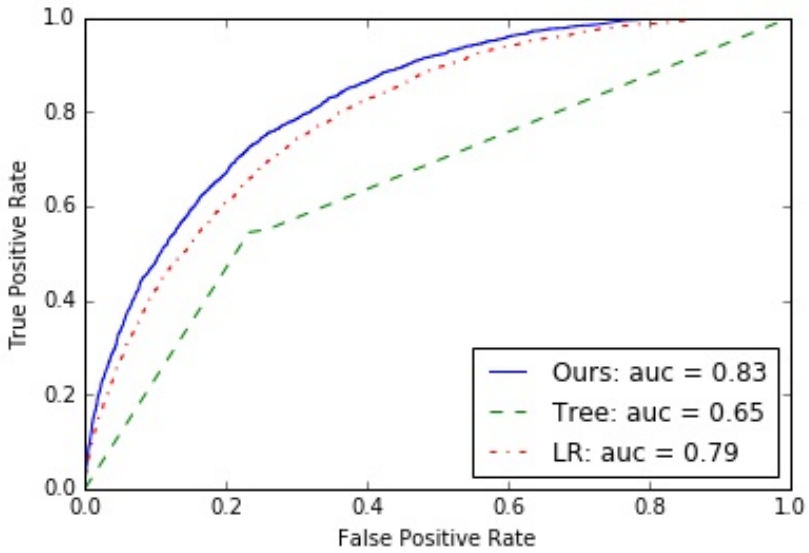

图 5.3 血液疾病预测的 ROC 曲线

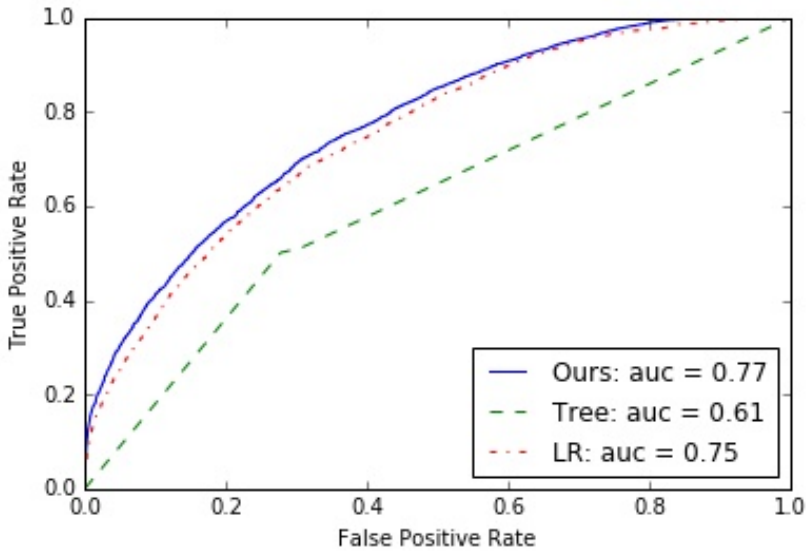

图 5.4 消化系统疾病预测的 ROC 曲线

针对单种疾病的预测结果，使用受试者工作特征曲线（ROC，receiver operation characteristic curve）进行可视化的对比展示。ROC 曲线通过针对目标变量设定不同的

临界值，可以考察算法的敏感性和特异性。曲线下方的面积（AUC，area under curve）越大，说明预测的准确性越高，图中左上角的点为敏感性和特异性的最高值。对比实验选择了血液疾病和消化系统疾病，这两个与日常生化关系最紧密也是最容易发生的疾病作为考察情况，结果如图5.3与5.4所示。从图中可以看出本文提出的算法对应的ROC 曲线更靠近（0，1）点，具有更好的敏感性和特异性。

结论

本文首先充分调研了医疗领域表示学习的研究现状和跨网络数据表示的研究进展，阐明了跨网络数据表示在医疗应用中的重要性和现实意义。通过合理使用电子医疗记录数据（EHR），不仅可以自动生成医疗信息以辅助医疗诊断，还促进了医疗精准化和个性化的发展。对于大众健康保障水平的提高，以及促进社会发展方面都有巨大的价值。

在本文的研究过程中，提出了一种融合医疗实体关系网络向量化表示算法和深层神经网络模型的疾病预测算法，将医疗领域的跨网络数据表示学习分为三个步骤。首先，从未经处理的电子医疗记录中构建疾病实体网络，使用基于节点相似性的网络向量化表示算法，对医疗实体关系网络中的疾病节点学习向量化特征表示；接着，使用降噪自编码器栈模型和深度神经网络模型，以病人的实验室检验结果为输入，学习病人的向量化体征表示；最后，使用第一步学习的疾病特征向量构建疾病相似矩阵，作为医学辅助知识对神经网络的参数学习进行约束，让预测模型生成的潜在疾病预测结果，既满足过往数据经验又符合医学常识。

针对现有疾病预测算法的不足，本文提出的模型分别从医疗跨网络数据表示和医疗数据降维方面提供了解决思路。使用这种算法可以自动地从海量数据中学习医疗领域的专家知识，用来约束后续预测模型的结果，大大降低了人工提取专家知识的时间成本和人力成本。其次，使用降噪自编码器结合深度神经网络模型的疾病预测方法，能够辅助医生进行初步的疾病筛查，在减少医生工作量的同时避免医生因为误判造成的损失。

由于研究时间问题，本研究中还有诸多方面需要完善。未来需要进一步研究跨网络表示学习中，医疗实体向量化的其他算法，使网络节点的向量化表示可以和后续的疾病预测模型结合。此外，在疾病预测模型中，还需要研究不同参数对模型准确率的影响，进一步提高预测准确率，达到更好的医疗辅助水平。

参考文献

[1] Mantas J. Electronic health record. [J]. Studies in health technology and informatics, 2002, 65: 250–257.

[2] 彭友霖, 叶和杨, 王志坚. 医用电子医疗器械市场分析 [J]. 商场现代化, 2008 (23): 176–176.

[3] Yu M, Yi Y, Rexford J, et al. Rethinking virtual network embedding: substrate support for path splitting and migration [J]. ACM SIGCOMM Computer Communication Review, 2008, 38 (2): 17–29.

[4] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth [J]. Studies in health technology and informatics, 2006, 121: 279.

[5] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology [J]. Nucleic acids research, 2004, 32 (suppl_1): D267–D270.

[6] Dubberke E R, Reske K A, McDonald L C, et al. ICD-9 codes and surveillance for Clostridium difficile–associated disease [J]. Emerging infectious diseases, 2006, 12 (10): 1576.

[7] 张长水, 杨强. 机器学习及其应用 [M]. 清华大学出版社, 2013.

[8] 罗希平, 田捷, 诸葛婴, et al. 图像分割方法综述 [J]. 模式识别与人工智能, 1999 (3): 300–312.

[9] Fischer A, Botero J F, Beck M T, et al. Virtual network embedding: A survey [J]. IEEE Communications Surveys & Tutorials, 2013, 15 (4): 1888–1906.

[10] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding [J]. science, 2000, 290 (5500): 2323–2326.

[11] Chen M, Yang Q, Tang X. Directed Graph Embedding. [C]. In IJCAI, 2007: 2707–2712.

[12] Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information [C]. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015: 891–900.

[13] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C]. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014: 701–710.

[14] Wang D, Cui P, Zhu W. Structural deep network embedding [C]. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016: 1225–1234.

[15] Yang C, Liu Z, Zhao D, et al. Network Representation Learning with Rich Text Information. [C]. In IJCAI, 2015: 2111–2117.

[16] Grover A, Leskovec J. node2vec: Scalable feature learning for networks [C]. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016: 855–864.

[17] Abdel-Hamid O, Mohamed A-r, Jiang H, et al. Convolutional neural networks for speech recognition [J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22 (10): 1533–1545.

[18] Cho K, Courville A, Bengio Y. Describing multimedia content using attention-based encoder-decoder networks [J]. IEEE Transactions on Multimedia, 2015, 17 (11): 1875–1886.

[19] Chua L O, Yang L. Cellular neural networks: Applications [J]. IEEE Transactions on circuits and systems, 1988, 35 (10): 1273–1290.

[20] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM [J], 1999.

[21] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. [C]. In Interspeech, 2010: 3.

[22] Aggarwal C C, Al-Garawi F, Yu P S. Intelligent crawling on the World Wide Web with arbitrary predicates [C]. In Proceedings of the 10th international conference on World Wide Web, 2001: 96–105.

[23] Gibney E, et al. DeepMind algorithm beats people at classic video games [J]. Nature, 2015, 518 (7540): 465–466.

[24] 东昱晓, 柯庆, 吴斌. 基于节点相似性的链接预测 [J]. 计算机科学, 2011, 38 (7): 162–164.

[25] Bi W, Kwok J T. Mandatory leaf node prediction in hierarchical multilabel classification [C]. In Advances in Neural Information Processing Systems, 2012: 153–161.

[26] 胡云, 王崇骏, 吴骏, et al. 微博网络上的重叠社群发现与全局表示 [J]. 软件学报, 2014 (12).

[27] Cheng J-Z, Ni D, Chou Y-H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans [J]. Scientific reports, 2016, 6: 24454.

[28] Johnson A E, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database [J]. Scientific data, 2016, 3.

[29] Agrawal R, Srikant R, et al. Fast algorithms for mining association rules [C]. In Proc. 20th int. conf. very large data bases, VLDB, 1994: 487–499.

[30] 王华, 胡学钢. 基于关联规则的数据挖掘在临床上的应用 [J]. 安徽大学学报: 自然科学版, 2006, 30 (2): 21–25.

[31] Wishart D S, Knox C, Guo A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets [J]. Nucleic acids research, 2007, 36 (suppl_1): D901–D906.

[32] 贾李蓉, 刘静, 于彤, et al. 中医药知识图谱构建 [J]. 医学信息学杂志, 2015 (8): 51–53.

[33] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation [R]. 1985.

[34] Hinton G E, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18 (7): 1527–1554.

[35] Lasko T A, Denny J C, Levy M A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data [J]. PloS one, 2013, 8 (6): e66341.

[36] Liu S, Liu S, Cai W, et al. Early diagnosis of Alzheimer's disease with deep learning [C]. In Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, 2014: 1015–1018.

[37] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3 (Feb): 1137–1155.

[38] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]. In Advances in neural information processing systems, 2013: 3111–3119.

[39] Choi E, Bahadori M T, Schuetz A, et al. Doctor ai: Predicting clinical events via recurrent neural networks [C]. In Machine Learning for Healthcare Conference, 2016: 301–318.

[40] Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks [J]. arXiv preprint arXiv:1511.03677, 2015.

[41] Elman J L. Finding structure in time [J]. Cognitive science, 1990, 14 (2): 179–211.

[42] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45 (11): 2673–2681.

[43] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735–1780.

[44] Huang G B, Lee H, Learned-Miller E. Learning hierarchical representations for face verification with convolutional deep belief networks [C]. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012: 2518–2525.

[45] Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition [M] // Fukushima K, Miyake S. Competition and cooperation in neural nets. Springer, 1982: 1982: 267–285.

[46] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278–2324.

[47] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]. In Advances in neural information processing systems, 2012: 1097–1105.

[48] Cheng Y, Wang F, Zhang P, et al. Risk prediction with electronic health records: A deep learning approach [C]. In Proceedings of the 2016 SIAM International Conference on Data Mining, 2016: 432–440.

[49] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs [J]. Jama, 2016, 316 (22): 2402–2410.

[50] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks [J]. Nature, 2017, 542 (7639): 115–118.

[51] Choi E, Bahadori M T, Searles E, et al. Multi-layer representation learning for medical concepts [C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1495–1504.

攻读学位期间发表论文与研究成果清单

[1] 礼欣，李懿，翟艳梅. 一种基于用户话题权威性的微博重排序方法 [P]. 中国专利：CN104317881B, 2017-11-24.

[2] Li X, Yang L, Li Y, et al. Deep trajectory: a deep learning approach for mobile advertising in vehicular networks[J]. Neural Computing and Applications, 1-13.

致谢

研究生生活很快就要迎来它的终点，感谢一路帮助过我的师长、同辈、家人，谢谢你们和我一起走过这段两年半的人生旅程，让所有弯路与收获都在我的生命里留下刻痕。

礼欣老师，我的本科学业指导老师，同时也是研究生期间的导师，7 年时间言传身教让我成为一个独立的社会个体。经常说为人师表，礼欣老师就是一位用自己的实际行动给学生做榜样的老师。她对学术认真严谨的态度，对研究领域新知识的不断学习，都起到了不可替代的表率作用。每当看到礼欣老师在办公室加班或者读论文，就会督促自己加倍努力地去学习。她同时也是一个无私地对同学好的导师，亲自组织学术讨论和娱乐活动，亦师亦友的关系让我由衷敬佩。

1029 实验室的众多小伙伴们也是研究生期间收获的至宝。刚进实验室的时候看到师兄认真研究的态度，让我对学术有了敬畏和憧憬，鞭策我不断前进。除了和师兄师姐的交流，同样记忆犹新的还有与同级的亮哥、飞哥、刘琳和祺旺一起上课、读论文、做实验的日子，大家一起朝着目标奋斗、准备实习、交流论文经验的日子是研究生活最宝贵的财富。此外，新入学的师弟师妹们也用他们的年轻和冲劲，带动我前行。

同样要感谢的还有我优秀的室友们，让我在科研之外的生活中不断丰富和反思自己。还要感谢北京理工大学和计算机学院提供科研条件和优秀的老师指导我的学习。

最后要感谢的是我的家人，谢谢你们对我无条件的信任和支持，让我顺利完成又一次人生挑战。