

浙 江 大 学

本 科 生 毕 业 论 文



题目 基于深度学习的疾病预测模型

姓 名 彭靖田

学 号 3120000013

指导教师 郑小林 副教授

专 业 计算机科学与技术

学 院 计算机学院

A Dissertation Submitted to Zhejiang
University for the Degree of Bachelor of
Engineering



TITLE: From clinic notes to semantic
representation: A deep learning-based
model for predicting diagnoses

Author: Jingtian Peng

Supervisor: Assi.Prof. Xiaolin Zheng

Major: Computer Science and Technology

College: Computer Science and Technology

Submitted Date: 2016-5-30

浙江大学本科生毕业论文（设计）诚信承诺书

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。
2. 本人在毕业论文（设计）中引用他人的观点和参考资料均加以注释和说明。
3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。
4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

毕业论文（设计）作者签名：

_____年_____月_____日

摘要

大数据+医疗是如今的热门话题，在医疗大数据分析蓬勃发展的同时，也相应产生了一些问题。其中，如何减少医学人员的数据标注成本，合理有效的分配医疗资源至关重要。医疗大数据分析，本质上是利用人工智能和算法辅助医生诊疗。而研究如何利用当下的深度学习技术，构建一种无监督学习的疾病预测模型，为医生提供诊断建议，对于缓解当下医疗资源不足的矛盾是有意义的。

鉴于医疗领域对数据可靠性的要求，本文使用了 MIT 开发的真实电子病历系统数据集 MIMIC，通过文本挖掘进行特征抽取，构建住院记录序列和特征矩阵，通过 Word2Vec 训练语言模型，并利用词向量的组合计算，提出一种基于时序性的无监督学习疾病预测模型。最后，通过 PCA 降维分析词向量的潜在语义表达，以及影响预测模型的因素。

通过实验及分析，我们发现空间维度、上下文距离和时间衰减系数会影响预测模型，而深度神经网络能够提升模型准确率。

关键词 文本挖掘，Word2Vec，语义分析，无监督学习

Abstract

Big Data + Healthcare is a hot topic nowadays. In the past years, it has achieved high-speed development., but it has also made some problems at the same time. So, the most essential things are how to reduce the cost of data annotation by medical personal and how to allocate the medical resources in a reasonable and effective way. In essence, Big Data + Healthcare is to use the AI and algorithms to assist doctors and clinics. It's meaningful to find a way to apply the Deep Learning Technology to an unsupervised disease prediction model which is used to provide some diagnostic advice for doctors and clinics, and it can alleviate the current storage of medical resources contradiction as well.

In view of the requirements of the reliability of the data in medical field, I used the electronic medical records system MIMIC data set developed by MIT lab. I extracted the feature through text mining, building the hospital records sequence and feature matrix, and trained the language model by Word2Vec. Then, I present an unsupervised learning model based on the time-ordered by computing the distributed representation and its combination. Finally, through PCA dimensional reduction, I analyzed the latent semantic of the distributed representation and the factor of the prediction model.

Through experiments and analysis, I found that the spatial dimension, the distance of context and time attenuation coefficient will affect the prediction model. And the deep neural network can improve the accuracy of the model.

Keywords Text Mining, Word2Vec, Semantic Analysis, Unsupervised Learning

目录

摘要..... I

Abstract..... II

第 1 章 绪论..... 1

 1.1 课题背景..... 1

 1.1.1 大数据+医疗 1

 1.1.2 大数据医疗现阶段问题 2

 1.1.3 工业界发展现状 2

 1.2 本文研究目标和内容..... 3

 1.3 本文结构安排..... 3

第 2 章 文献综述..... 5

 2.1 医疗大数据..... 5

 2.1.1 医疗大数据的采集和存储 5

 2.1.2 医疗大数据的特点 6

 2.2 处理医疗大数据的算法..... 7

 2.2.1 分类 7

 2.2.2 回归分析 7

 2.2.3 聚类 8

 2.2.4 关联规则 8

 2.2.5 知识表达 8

 2.3 医疗大数据应用..... 9

 2.3.1 临床决策支持系统 9

 2.3.2 电子档案分析与公共健康 9

 2.3.3 疾病模式分析与个性化医疗 10

 2.4 本章小结..... 10

第 3 章 研究方案..... 12

 3.1 概述..... 12

 3.2 文本挖掘..... 13

3.2.1 数据集 MIMIC III.....	13
3.2.2 自然语言处理	15
3.3 Word2Vec	17
3.3.1 潜在语义提取	17
3.3.2 CBOW	17
3.3.3 Skip-gram	19
3.3.4 Hierarchical Softmax	20
3.3.5 Negative Sampling	22
3.4 无监督疾病预测模型.....	23
3.4.1 住院记录矩阵 R	23
3.4.2 特征矩阵 M	24
3.4.3 预测模型 Y	24
3.4.4 时序预测模型	25
3.5 本章小结.....	27
第 4 章 实验结果.....	28
4.1 数据描述.....	28
4.1.1 MIMIC III 数据.....	28
4.1.2 潜在语义提取	30
4.2 Word2Vec 训练参数.....	31
4.2.1 确定最佳向量空间维度 size	32
4.2.2 确定最佳上下文窗口 window	33
4.3 疾病预测模型对比.....	34
4.4 本章小结.....	36
第 5 章 分析与讨论.....	37
5.1 Word2Vec 潜在语义提取分析.....	37
5.2 影响疾病预测模型的因素分析.....	37
5.2.1 向量空间维度对疾病预测的影响	37
5.2.2 上下文窗口对疾病预测的影响	38
5.2.3 时序性对疾病预测的影响	38
5.3 本章小结.....	38

第 6 章 本文总结.....	39
6.1 论文主要工作.....	39
6.2 将来的工作.....	39
参考文献.....	41
致谢.....	46

第1章 绪论

1.1 课题背景

1.1.1 大数据+医疗

医疗大数据分析是如今国内外的热门话题，不论是在学术界，还是工业界，各国都投入了大量的资源。在 IDF2013 英特尔信息技术峰会上，英特尔全球健康解决方案架构师吴闻新的《大数据在医疗行业的应用》报告中提到医疗服务产生的数据总量正在急剧增长。到2020年，医疗数据将达到35ZB(1ZB=2⁴⁰GB)，相当于 2009 年数据量的 44 倍，而其中最有可能首先得到利用的便是电子健康病例（EHR）。

电子健康病历（Electronic Health Records, EHR），是人们在健康相关活动中直接形成的具有保存备查价值的电子化历史记录。它是存储于计算机系统之中、面向个人提供服务、具有安全保密性能的终身个人健康档案。EHR 是以居民个人健康为核心，贯穿整个生命过程，涵盖各种健康相关因素、实现多渠道信息动态收集，满足居民自我保健、健康管理和健康决策需要的信息资源。

此外，各种健身，健康可穿戴设备的出现，使得血压、心率、体重，血糖，心电图（EKG）等的监测都变为现实和可能，信息的获取和分析的速度已经从原来的按“天”计算，发展到了按“小时”，按“秒”计算。比如，一家名为 Blue Spark 的科技公司已经生产出能 24 小时实时监测体温的新型温度计贴片 temptraq。

这种数据的扩展速度和覆盖范围是前所未有的，数据的格式也五花八门，可能是无格式文件（flat file），CSV，关系表，ASCII/ 纯文本文件等等。同时，数据的来源也纷繁复杂，可能来自不同的地区，不同的医疗机构，不同的软件应用。

但相对于各种可穿戴设备采集的医疗数据，EHR 经过多年的积累，已经拥有了一套相对完善的规范和标准。在时间的维度上，可以描述出一个更加完整的病人画像，有助于我们的疾病预防和医学研究。

有效的整合和利用数字化的医疗大数据对个体医生，康宝中心，大型医院，和医疗研究机构都有着显著的好处。潜在的利益包括（W.Raghupathi &

Raghupathi, 2014[1]):

- 更多更准确的数据使得疾病能在早期被监测，从而使治疗更容易和有效。
- 通过对特定个体或人群的健康管理，快速有效地监测保健诈骗。
- 基于大量的历史数据，预测和估计特定疾病或人群的某些未来趋势，比如：预测特定病人的住院时间，哪些病人会选择非急需性手术，哪些病人不会从手术治疗中受益，哪些病人会更容易出现并发症，等等。麦肯锡估计，单单就美国而言，医疗大数据的利用可以为医疗开支节省出 3 千亿美元一年。

1.1.2 大数据医疗现阶段问题

根据一份针对美国和加拿大 333 家医疗机构及 10 家其他机构的调查 (IHIT, 2013), 2013 年, 医疗机构累积的数据量比 2011 年多出了 85%, 但 77% 的医疗健康行政人员对自己机构在数据管理方面的能力评价为“C”。此外, 仅有 34% 报告他们能从电子健康记录 (EHR) 中获取数据用来帮助病人, 而有 43% 报告他们不能收集到足够多的数据来帮助病人。由此可见, 在北美的医疗系统中, 医疗大数据的管理使用准备工作还有一大段路要走。中国也是处在起步阶段, 国内种类繁多的医疗信息系统导致产生大量的信息孤岛, 如果无法整合各家医院的 EHR 数据, 也就无法完整的描述出一个病人的病例画像。

国家卫生计生委在《2016 年卫生计生工作要点》的第(八)点加快卫生计生信息化建设中, 明确指出“开展健康医疗大数据应用发展试点示范工作, 积极实施‘互联网+健康医疗’服务, 通过信息化手段, 放大群众的获得感。”

因此, 如何有效的利用 EHR 数据成为当今的热门话题。现阶段对于结构化的 EHR 数据已经有了许多突破性的进展, 然而对于住院病历这类医生手写的自然语言数据, 却仍然没有得到充分的利用。

1.1.3 工业界发展现状

我曾经在 2015 年 5 月—2015 年 11 月在《半个医生》团队实习半年, 主要工作内容是研究疾病预测的半监督机器学习算法。目前同类型的大数据医疗公司如《春雨医生》、《快速问医生》和《丁香医生》等, 都存在以下几个比较明显的特点:

- 主要使用来源单一的结构化 EHR 数据;

- 需要医学专业人士的支持；
- 需要大量人工标注数据；

基于以上背景，我希望构造一种计算简单，无需医学人员标注的无监督学习模型。我将通过大量病人报告构建语言模型，通过抽取住院报告数据中的疾病诊断、体格检查和治疗药物等 3 类特征数据，构建特征矩阵和住院记录矩阵，将时序性纳入模型作为一个权重因子进行考量。

1.2 本文研究目标和内容

鉴于国内暂时没有面向学生个人研究的医疗数据库，所以我选择了由 MIT 实验室开发的公共医疗数据集 MIMIC[2]（包含了超过 40000 去识别化的病人电子记录）。通过 CITI Program 和 Protecting Human Research Participants 的考试，拿到相关资质证书后，由我在 UCSD 医学院访学期间的导师 Prof.Jiang 作为担保，我准备对以下三个方向做出研究：

- **对比国内大数据医疗公司的疾病预测模型**

国内目前有多家公司已经上线了大数据医疗的疾病预测服务。通过对比已有的疾病预测模型不仅有助于我理解基于特定需求的算法和模型，还提供了—个学习和改进现有算法的机会。

- **基于深度学习的无监督疾病预测模型**

将大量真实的医学病历作为语料库，使用 Word2Vec 训练得到语言模型，最终利用词向量的组合计算，以及医疗事件和疾病之间的相关性构建—个无监督的疾病预测模型。

- **分析时序性对疾病预测模型的影响**

在基于医学知识和文献阅读的基础上，考虑时序性对于疾病预测的影响，我认为距离当前时间越远的医疗事件影响应当越小。因此，我希望通过自己的研究在这方面获得—些新的发现和进展。

1.3 本文结构安排

本文共分六个章节，每个章节的研究内容和主要贡献如下：

第 1 章介绍了本文的研究背景、研究目的和意义、研究内容和主要贡献。

第 2 章主要介绍了大数据医疗的相关背景和应用，与大数据医疗相关的文

献，分析了文献中疾病预测的常见模型。在相关技术小节阐述了本文自然语言处理中用到的一些信息抽取的方法和原理。

第 3 章首先提出了本文的数据来源和研究方案，结合目前国内外 EHR 数据普遍为信息孤岛的情况下，提出一种计算简单的无监督疾病预测模型。该模型通过 Word2Vec 训练大规模真实 EHR 数据，在自然语言处理的前提下，不再需要专业的医学人士进行人工标注，省去了大量人力成本和时间成本。然后，考虑时序性对于预测结果的影响，将时间权重纳入模型。

第 4 章主要通过 MIMIC III 数据集来验证第 3 章中提出的预测模型。与经典的 TF-IDF 进行效果对比，同时考察时序性对于预测结果的影响。

第 5 章主要通过第 4 章的实验结果，分析影响预测模型的因素，为无监督疾病预测模型的改进提出个人建议。

第 6 章对本文的研究内容进行总结，并对未来的工作进行展望。

第2章 文献综述

2.1 医疗大数据

大数据是最近几年非常热门的话题，如今随着移动互联网、物联网、社交网络的蓬勃发展与应用，全球的数据容量正以空前的速度增长。2011 年《Science》推出关于数据处理的专刊《Dealing with data》，讨论了数据洪流所带来的挑战[3]；IEEE 在 2013 召开 Big Data 国际会议并成立首届学术工作组，对大数据的理论、管理、信息安全、基础结构在互联网技术、医疗健康、环境科学等多个方面的应用和发展进行了深入探讨[4-5]。

本章将详细介绍医疗大数据的采集、存储及数据特点等，以及国内外现阶段用大数据服务于医疗健康的算法和应用。

2.1.1 医疗大数据的采集和存储

医院信息系统(Hospital Information System, HIS),最早由 Collen 提出：利用电子计算机和通讯设备，为医院所属各个部门提供病人诊疗信息和行政管理信息的采集、存储、处理、提取和数据交换的能力并满足授权用户的功能需求的平台[6]。HIS 以计算机为基础简化管理医院医疗信息，是医疗大数据发展的来源和应用领域。随着数据传输和存储技术的发展，HIS 得到快速发展，根据 2005 年 CHIMA 医院信息化调查资料数据，在我国许多医院临床信息系统已经得到快速的应用和发展，LIS 系统占 39.14%，住院医生工作站系统占 35.04%，门诊医生工作站系统占 32.99%，在沿海经济发达地区的信息化突出的医院，电子病历、全院 PACS、移动、无线、PDA、Tablet PC、RFID、万兆网络、服务器集群等先进的系统和先进的 IT 技术已经开始应用[7]。

电子病历系统，与上述系统不同，电子病历以病人为中心，将病人诊断过程中产生的诊疗数据和检查数据集合为具有统一形式的记录，是最具有价值的数据来源[8]，也是真正可以与长期健康监测整合为统一数据样本的医疗数据信息。以计算机化的病历系统或者基于计算机的病人记录用以集成患者医疗信息，可以为教学、科研和决策提供资料来源。电子病历的发展目标主要是加速患者医疗信息流通，使患者信息在医疗系统内随时随地可以得到,提供纸张病历无法

提供的服务，在信息化的角度来讲，方便了大数据挖掘中的信息提取，标准化的电子病历有效的提高决策效率和医疗服务质量。

本文实验使用的病历数据就来自于北美某医疗机构的电子病历系统。

2.1.2 医疗大数据的特点

传统医疗行业中，医院信息系统完成了医院内部的流程控制、数据积累等工作。医疗行业早就遇到了海量数据和非结构化数据的挑战，而近年来很多国家都在积极推进医疗信息化发展，这使得很多医疗机构有资金来做大数据分析[9]。医疗数据是医疗人员对病人诊疗过程中产生的数据，包括病人的基本情况、行为数据、诊疗数据、管理数据、检查数据、电子病历等。现代医院中将上述数据存储于医院的各个信息系统之中，是医疗大数据分析的基础。

按照大数据的概念来看，医疗数据的 Volume、Velocity、Variety、Value 四个特征都是显而易见的，除此之外，医疗大数据具有多态性、不完整性、时效性、冗余性、隐私性等特点[10]。

多态性：医疗数据的表达格式包括文本型、数字型和图像型。文本型数据包括人口特征、医嘱、药物使用、临床症状描述等数据；数字型数据包括检验科的生理数据、生化数据、生命体征数据等；图像型数据包括医院中的各种影像学检查如 B 超、CT、MRI、X 光等图像资料。在文本型数据中，数据的表达很难标准化，对病例状态的描述具有主观性，没有统一的标准和要求，甚至对临床数据的解释都是使用非结构化的语言。多态性是医学数据区别于其他领域数据的最根本和最显著的特性。这种特性也在一定程度上加大了医疗数据的分析难度和速度。

不完整性：医疗数据的搜集和处理过程存在脱节，医疗数据库对疾病信息的反映有限。同时，人工记录的数据会存在数据的偏差与残缺，数据的表达、记录有主观上的不确定性。同一种疾病并不可能全面由医学数据反映出来，因此疾病的临床治疗方案并不能通过对数据的分析和挖掘而得出。另外，从长期来看，随着治疗手段和技术手段的发展，新类型的医疗数据被创造出来，数据挖掘的对象维度是在不停的增长的[10]。

时效性：病人的就诊、疾病的发病过程在时间上有一个进度，医学检测的波形信号（比如说心电、脑电）和图像信号（MRI、CT 等）属于时间函数，具有时效性。例如心电信号检测中，短时的心电无法检出某些阵发性信号，而只

能通过长期监测的方式实现心脏状态的监测[11]。

冗余性：医疗数据中存在大量的相同或类似信息被记录下来。比如常见疾病的描述信息，与病理特征无关的检查信息。

隐私性：在对医疗数据的数据挖掘中，不可避免的会涉及到患者的隐私信息，这些隐私信息的泄露会对患者的生活造成不良的影响。特别是在移动健康和医疗服务的体系中，将医疗数据和移动健康监测甚至一些网络行为、社交信息整合到一起的时候，医疗数据的隐私泄露带来的危害将更加严重。大数据分析中隐私保护要注意两个方面：其一，用户身份、姓名、地址和疾病等敏感信息的保密；其二，经分析后所得的私人信息的保密[12]。

2.2 处理医疗大数据的算法

医疗数据是持续、高增长的复杂数据，蕴含的信息价值也是丰富多样的，对医疗数据的有效存储、处理、查询和分析，挖掘其潜在价值，发现医学知识，将深切影响人类健康水平和治疗手段。在传统的医学统计方法的基础上，新的模型与技术的出现，为从数据中获取新知识提供了新的思路。

医疗数据挖掘和分析常用的算法包括分类、回归分析、聚类、关联规则、知识表达。针对不同的类型的病人对不同类别的生理数据进行推理判断，大数据分析技术实现了服务临床治疗、预测疾病发病情况、跟踪病人病情等目的。下面将介绍在医疗健康领域中上述大数据算法的应用现状。

2.2.1 分类

分类是找出数据库中一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到某个给定的类别[13]。其实例可以应用到用户的分类、用户的属性和特征分析、用户治疗效果评价等。利用病人的人口学信息、对情绪图片库的生物反馈信息、自主神经的生理学特性信息作为输入特征，使用决策树算法用以对生物唤起讯号分类[14]。对比基于规则、基于决策树、基于人工神经网络的糖尿病人的健康数据的模式分类方法[15]。

2.2.2 回归分析

回归分析方法反映的是事务数据库中属性值在时间上的特征，产生一个将

数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特征、数据序列的预测以及数据间的相关关系等[16]。例如对医院信息系统的医疗风险因素的回归分析，分析各个影响因素与医疗风险之间的联系及引起风险的概率变化，用以指导医院的风险管理[17]。临床心理治疗中通过对实验结果的回归分析，研究人员可以高效地区分有效预测和无效预测，并且发现预测变量之间的联系，并对临床的治疗决策提供预测模型[18]。使用回归模型对父母调查报告、孩子调查报告、生物标记、行为与儿童焦虑抑郁建模，并将模型应用到临床指导[19]。

2.2.3 聚类

聚类分析是把一组数据按照相似性和差异性分为几个类别，其目的是使得属于同一类别的数据间的相似性尽可能大，不同类别中的数据间的相似性尽可能小[20]。通过对青少年和成年人的酗酒成瘾状况及心理测试的结果聚类分析来完成酗酒人格模式的知识发现[21]。通过对疼痛反应结果的聚类分析，完成了对热性疼痛、压力性疼痛、缺血性疼痛的诱因分析[22]。

聚类方法的一般思路是通过观测数据样本的亲疏关系的统计量，根据某种准则使同类型的数据差别较小，类与类之间差别较大，以此完成个体或者变量的区分。在医疗健康记录的关键词分类[23]、生理信号分析[24-25]中发挥了重要作用。

2.2.4 关联规则

关联规则是描述数据库中数据项之间所存在的关系的规则，即根据一个事务中某些项的出现可导出另一些项在同一事务中也出现，即隐藏在数据间的关联或相互关系。在电子健康档案中，大量用户的个人信息、健康信息、临床诊疗信息等，可以应用到疾病的检测和预测中。通过对比房颤病人的医疗数据研究房颤与脑梗塞的关联规则[26]。对门诊病人的医疗档案和异常的筛查结果进行了关联规则分析[27]。提出了一种基于医院信息系统和医疗记录的临床推荐系统[28]。医疗记录中的关联规则发现，有助于新的医学知识的发现和发病风险分析[29]。

2.2.5 知识表达

Fei Wang 等人在 TPAMI 上发表过一篇在医疗健康数据上挖掘事件序列特征

的结构和其应用方法的论文[30]，详细阐述了一种全新的知识表达，即一种能够在单一和多样事件序列中表达、抽样和挖掘事件隐含关系的结构，并且能够在大规模长周期多样化的医疗事件数据中提取特征。这种结构将多样化的医疗数据整合，并且给出了不同事件间的关系，对于之后的研究打下了坚实的基础。

2.3 医疗大数据应用

医疗行业的传统数据应用具有重要的参考价值。在对用户的诊疗数据、健康监测数据的采集和分析的基础之上，可以实现用户身体状况的预测、监控，甚至可以确定用户是哪一类的疾病的易感人群。提高用户的健康状况水平，降低用户的患病风险。精准分析包括病人体征数据、费用数据和疗效数据在内的大型数据集，可以帮助医生确定临床上最有效和最具有成本效益的治疗方法。医疗护理系统将有可能减少过度治疗，比如避免副作用大于疗效的治疗方式。

2.3.1 临床决策支持系统

临床决策支持，是指医生在诊疗过程中，能对医生的实时诊疗决策制定做出帮助的各种资源。常见的有科研文献、在线期刊、专家会诊意见、循证医学证据、临床决策支持系统（CDSS）等。临床决策支持系统，是通过数据、模型等，以人机交互辅助临床工作人员决策的计算机应用系统。

得益于对非结构化数据的分析能力的日益加强，临床决策支持系统在大数据分析技术的帮助下变得更加智能。比如可以使用图像分析和识别技术，识别医疗影像数据，或者挖掘医疗文献数据建立医疗专家数据库，从而为医生提出诊疗建议。文献[50]介绍一种基于生理数据的云计算用药决策支持系统，使用了基于生理数据和药物剂量及临床表现的历史数据用以指导早产儿药物剂量。利用临床心脏影像大数据支持的人工智能和先进计算，用以实现个性化的治疗[31]。利用机器学习对临床数据建模,用以实现疾病的预测、康复和临床决策支持，为医生的治疗提供了新的思路[32-34]。

2.3.2 电子档案分析与公共健康

在病人档案方面应用高级分析可以确定哪些人是某类疾病的易感人群，进行药物使用的安全性分析。通过对相关病人的电子病历以及药品代理商的药物资料进行数据分析，用以完成药品安全监测，防止药品滥用事件的发生[35]。通

通过对基因和遗传等数据进行建模来推断疾病的潜伏期、易感群体、传染性的方法[36]。通过电子医疗记录进行公共卫生监测和传染病控制的几种方法[37]，以心衰为例子综述了使用电子病历进行疾病预测建模的挑战与策略，并对几种智能算法进行了分析[38]。电子病历的数据挖掘有利于新的疾病分级策略的建立和对未知疾病的相关临床症状进行分析，结合基因数据，可以实现对基因表达的生物机制研究[39]。

2.3.3 疾病模式分析与个性化医疗

美国罗彻斯特大学医学院精神病学和内科教授恩格尔（G. L. Engel）在1977年科学杂志上发表的题为“需要新的医学模式，对生物医学的挑战”的文章[40]中指出，现有的占统治地位的疾病模式是生物医学模式，以分子生物学的可测量的生物学变量来分析疾病，没有将病患的社会、心理和行为方面纳入到医学模式之中。通过对病人生理参数的长期监测，挖掘病人电子档案，实现疾病的预测、疾病的建模已经广泛应用在医疗领域[41-43]。大数据背景下，把患者的健康数据包括锻炼习惯、生活习惯、社交媒体信息等等纳入到疾病模式的分析和建模中来，可以更有针对性的针对个体实现个性化的治疗，也是生物心理社会医学的一个发展方向，例如可以通过对社交媒体数据进行文字关键字分析来分析青少年心理压力[44]。

2.4 本章小结

本章从大数据的基本特点和在医疗健康领域的发展出发，分析了医疗大数据的数据来源和数据特点，综述了基于医疗大数据的应用，最后对医疗大数据应用中使用的算法模型进行了概述。

首先，医疗领域的海量数据的积累，并不是完全的新的概念，而是在医疗机构的临床治疗、实验中一直存在的，随着信息化程度的加深，越来越多的诊疗数据以可分析的方式逐渐积累。如何在海量的数据基础上获得有价值的信息以及如何使用新的方法来完善现有的诊疗知识库成为了当下的挑战。分类、回归、聚类等数据挖掘方法在医疗大数据分析中获得的结果在临床决策支持、电子档案分析与公共健康方面、疾病模式分析与个性化医疗方面都得到了广泛的应用。

医疗健康与人类的生活息息相关，随着技术的发展，如何更好的利用技术

服务人类，促进人类的发展，在大数据时代背景下变得更加迫切。

第3章 研究方案

3.1 概述

本文的研究是基于 MIMIC III 公共医疗数据集，通过 Word2Vec 训练语言模型，对其中非结构化数据进行文本挖掘，构造特征矩阵和医疗事件序列，建立一个无监督的疾病预测模型。然后将时序性影响加入模型，分析时间因素对预测模型的影响。最后对比 TF-IDF 与该模型的预测结果。

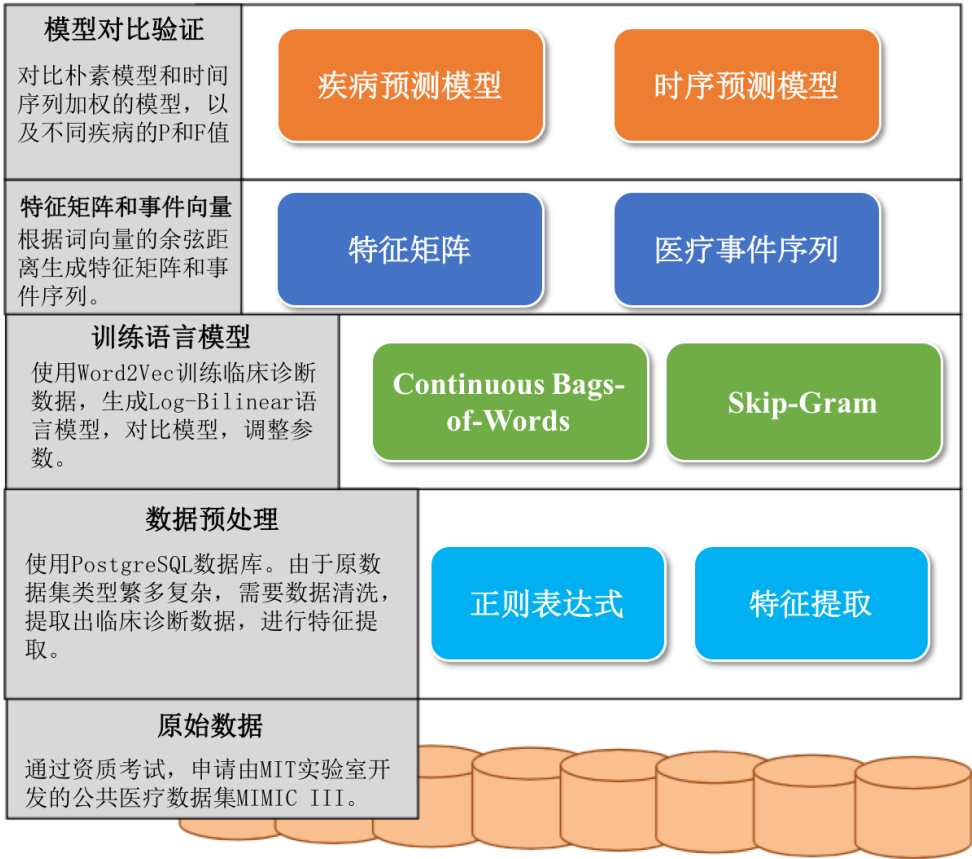


图 3-1 模型框架图

在具体的研究方法方面，首先需要查阅相关文献和熟悉美国隐私保护法案，了解目前该领域研究现状和医疗数据利用情况。对比工业界和学术界对医疗大数据的不同使用情况，最终决定将 EHR 非结构化数据，即医生手写的住院记录

作为原始数据。用正则表达式和关键词技术提取出医疗事件和疾病诊断作为特征数据。通过训练得到 Log-Bilinear 语言模型，生成医疗事件序列和特征矩阵。最终对比不同模型和在各种疾病上的预测效果，采用 P 和 F 值作为评价标准。

3.2 文本挖掘

3.2.1 数据集 MIMIC III

MIMIC III(Medical Information Mart for Intensive Care III)[2]，由 MIT 实验室开发，是一个大规模、免费的公共医疗数据集。拥有超过 40000 个去识别化病人的完整电子病历，这些病人均为 2001 年—2002 年在 Beth Israel Deaconess 医疗中心重症监护室的患者。

MIMIC III 中的数据包括患者的生命特征（如脉搏、呼吸、体温、血压）、服用药物、体格检查、实验结果、护理记录和住院记录等。MIMIC 广泛支持流行病学研究分析、临床决策支持和电子工具开发。该数据集有 3 个显著特点：

- 1) 面向全球的科研工作者免费开放（需要相关证书和研究申请）；
- 2) 拥有高质量和多样化的 ICU(Intensive Care Unit)患者完整病历；
- 3) 包含高频率和细颗粒度的实验结果、电子文档、病床监护仪趋势和波形图。

MIMIC III 本身是以关系数据库的形式保存的表格，每张表格都是关于病人的相关信息，其中每列都是不变的患者信息，每行为一个病人实例。同时提供以“D_”开头的索引表，用以唯一确定患者和其相关信息。

ADMISSIONS	Every unique hospitalization for each patient in the database (defines HADM_ID)
CALLOUT	Information regarding when a patient was cleared for ICU discharge and when the patient was actually discharged
ICUSTAYS	Every unique ICU stay in the database (defines ICUSTAY_ID)
PATIENTS	Every unique patient in the database (defines SUBJECT_ID)
SERVICES	The clinical service under which a patient is registered
TRANSFERS	Patient movement from bed to bed within the hospital, including ICU admission and discharge

表 3-1 定义患者和跟踪患者轨迹表

CAREGIVERS	Every caregiver who has recorded data in the database (defines CGID)
CHARTEVENTS	All charted observations for patients
DATETIMEEVENTS	All recorded observations which are dates, for example time of dialysis or insertion of lines.
INPUTEVENTS_CV	Intake for patients monitored using the Philips CareVue system while in the ICU
INPUTEVENTS_MV	Intake for patients monitored using the iMDSoft Metavision system while in the ICU
NOTEVENTS	De-identified notes, including nursing and physician notes, ECG reports, imaging reports, and discharge summaries.
OUTPUTEVENTS	Output information for patients while in the ICU
PROCEDUREEVENTS_MV	Patient procedures for the subset of patients who were monitored in the ICU using the iMDSoft MetaVision system.

表 3-2 急诊治疗记录表

CPTEVENTS	Procedures recorded as Current Procedural Terminology (CPT) codes
DIAGNOSES_ICD	Hospital assigned diagnoses, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system
DRGCODES	Diagnosis Related Groups (DRG), which are used by the hospital for billing purposes.
LABEVENTS	Laboratory measurements for patients both within the hospital and in out patient clinics
MICROBIOLOGYEVENTS	Microbiology measurements and sensitivities from the hospital database
PRESCRIPTIONS	Medications ordered, and not necessarily administered, for a given patient
PROCEDURES_ICD	Patient procedures, coded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system

表 3-3 住院记录表

D_CPT	High-level dictionary of Current Procedural Terminology (CPT) codes
D_ICD_DIAGNOSES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD) codes relating to diagnoses
D_ICD_PROCEDURES	Dictionary of International Statistical Classification of Diseases and Related Health Problems (ICD) codes relating to procedures
D_ITEMS	Dictionary of ITEMIDs appearing in the MIMIC database, except those that relate to laboratory tests
D_LABITEMS	Dictionary of ITEMIDs in the laboratory database that relate to laboratory tests

表 3-4 字典索引表

除以上四种基本表以外，MIMIC 中还有许多的派生表，提供了患者在 ICU 期间的附加信息和情况总结。这些派生表大多是由一些社区或是研究人员自行编写的程序和 SQL 脚本产生的，其中的大部分都可以在官方的 github 项目下获取。

3.2.2 自然语言处理

该实验的一大难点在 NLP，即如何高效、准确的从非结构化的住院记录中抽取特征数据。我主要采用了正则表达式和模式匹配的方式进行处理，参考了一部分 python NLTK 的 API，再针对住院记录的关键字实现了特征识别。最终抽取 3 类特征数据：

- 1) 疾病诊断(Diagnoses)
- 2) 服用药物(Medications)
- 3) 体格检查(Procedures)

其中，疾病诊断和服用药物在住院记录中通常会有关键字标示，如“Discharge Diagnoses:”、“Discharge Medications:”，之后便会依次列出疾病和药物。

```
7013  DISCHARGE DIAGNOSES:
7014  1. Community-acquired pneumonia complicated by loculated
7015  effusion and probable empyema.
7016  2. Parkinson disease with associated [**Last Name (un) 305**] body dementia.
7017  3. Hypertension.
7018  4. Depression.
7019  6. Tobacco abuse.
7020  7. Respiratory distress; resolved.
7021
```

图 3-2 住院记录中的疾病诊断

```
8316  Discharge Medications:
8317  1. Insulin Regular Human 100 unit/mL Solution Sig: One (1)
8318  Injection ASDIR (AS DIRECTED).
8319  2. Albuterol Sulfate 0.083 % Solution Sig: One (1) Inhalation
8320  Q6H (every 6 hours) as needed.
8321  3. Fluoxetine 20 mg Capsule Sig: One (1) Capsule PO DAILY
8322  (Daily).
8323  4. Heparin (Porcine) 5,000 unit/mL Solution Sig: One (1)
8324  Injection TID (3 times a day).
8325  5. Acetaminophen 325 mg Tablet Sig: One (1) Tablet PO Q4-6H
8326  (every 4 to 6 hours) as needed.
8327  6. Albuterol 90 mcg/Actuation Aerosol Sig: 1-2 Puffs Inhalation
8328  Q4H (every 4 hours) as needed.
8329  7. Erythromycin 5 mg/g Ointment Sig: One (1) Ophthalmic QID (4
8330  times a day).
8331  8. Polyvinyl Alcohol-Povidone 1.4-0.6 % Dropperette Sig: [**2-7**]
8332  Drops Ophthalmic Q2H (every 2 hours).
8333  9. Levofloxacin 500 mg Tablet Sig: One (1) Tablet PO Q24H (every
```

图 3-2 住院记录中的出院药物

由于体格检查数据没有标志词，因此就需要更加复杂的正则表达式来进行匹配。

```
Pattern = "^(^[^,.)+ ]\\*{2}([\\d]{1,2})-([\\d]{1,2})\\*{2}\\]:.*"
```

```

161 Radiology:
162 CXR [**9-28**]: Diffusely increased opacities at the lung fields
163 bilaterally. In an immunocompromised patient, this is concerning
164 for PCP [**Name Initial (PRE) 2**]. Radiographically, the differential
165 pulmonary edema. Additionally, there is a faint opacity at the
166 right lung base, which may represent atelectasis or focal
167 pneumonic process.
168 .
169 CT-Head [**9-28**]: Focus of low attenuation within the subcortical
170 white matter of the right medial frontal lobe. This may
171 represent a subacute infarction; however, an underlying mass
172 lesion cannot be completely excluded. An MRI examination with
173 gadolinium and diffusion-weighted imaging is recommended for
174 further evaluation. No intracranial hemorrhage noted.
175 .
176 MR-head-w&w/o gadolinium [**9-30**]:
177 Signal abnormality in the medial right frontal lobe involving

```

图 3-3 住院记录中的体格检查

```

1 Total Discharge Diagnoses: 38215
2 (u'hypertension', 4884)
3 (u'primary:', 3381)
4 (u'', 2914)
5 (u'primary diagnosis:', 2141)
6 (u'hyperlipidemia', 1930)
7 (u'coronary artery disease', 1663)
8 (u'secondary:', 1626)
9 (u'atrial fibrillation', 1581)
10 (u'htn', 1428)
11 (u'pneumonia', 1369)
12 (u'acute renal failure', 1221)
13 (u'anemia', 1166)
14 (u'depression', 1005)
15 (u'urinary tract infection', 920)

```

图 3-4 抽取出疾病诊断数据的统计结果

最终通过设置阈值和人工筛选出特征数据，由于 Word2Vec 的训练单位是单词（word），因此将非单词的特征词用“_”连结作为一个整体词，如“acute_renal_failure”。然后在语料库中整体替换，用作下一步 Word2Vec 的输入。

3.3 Word2Vec

3.3.1 潜在语义提取

自从 Google 发布 Word2Vec 以来，深度学习迎来了学术界和工业界的一次新的高潮。在 Tomas Mikolov 等人发表的《Distributed Representation of Words and Phrases and their Compositionality》[45]中提出了一种新的 Log-Bilinear 语言模型。相对于传统的统计语言模型，如 N-gram[46]、N-pos[47]通常只考虑 2-3 个上下文单词（由于计算量随着 N 呈指数级增长，因此普遍使用 bigram 或者 trigram）。Word2Vec 使用的 CBOW 和 Skip-gram 能够将上下文窗口提升至 200 仍可用。

由于使用了词向量表示单词（Distributed Representation）[48]，不仅大大减小了时间复杂度和空间复杂度，同时也使得模型拥有了更强的潜在语义提取的能力。谷歌官方给出了一个经典的词向量推理例子，即：

$$\text{vec}(\text{"Paris"}) \approx \text{vec}(\text{"Madrid"}) - \text{vec}(\text{"Spain"}) + \text{vec}(\text{"France"})$$

这说明 Word2Vec 确实学到了语言背后的语义，并且这种学习能力是跨语言的，人类在发明语言的时候，从向量空间的角度来看，设计者们应当是“串通一气”的。因为不同语言在表达相同意思时，他们的词向量出奇一致。

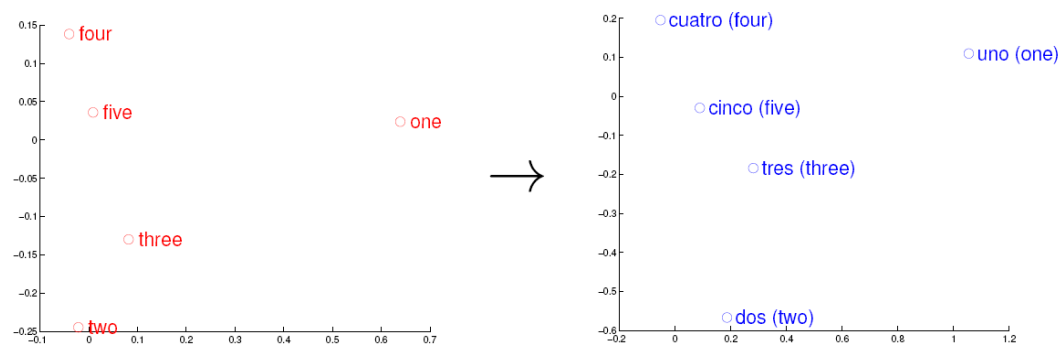


图 3-5 英语与西班牙语的同义词向量经 PCA 降维后向量也几乎相等

这给了我启示，如果利用 word2vec 训练后的词向量的推理能力研究英文病例，如果模型效果不错，那么作用在中文语料库下也应当具有等同的疾病预测。

3.3.2 CBOW

CBOW[49] (Continuous Bags-of-Words) 模型是一种与前向 NNLM 类似的模型,不同点在于 CBOW 去掉了最耗时的非线性隐层且所有词共享隐层。CBOW 是通过上下文预测当前词 $P(w_t | w_{t-k}, w_{t-(k-1)} \dots, w_{t-1}, w_{t+1} \dots, w_{t+k})$ 。

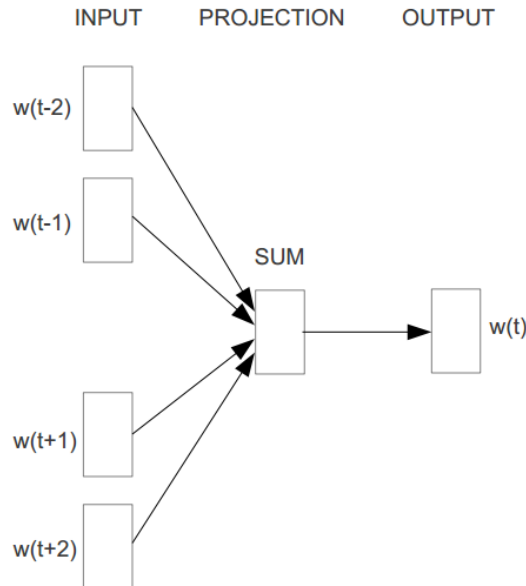


图 3-6 CBOW 模型

从输入层到隐层所进行的操作实际就是上下文向量的加和,具体的代码见图 3-7。其中 *sentence_position* 为当前 word 在句子中的下标。以一个具体的句子“W X Y Z”为例,第一次进入到下面代码时当前 word 为 W, *sentence_position* 为 0。b 是一个随机生成的 0 到 window-1 的词,整个窗口的大小为 $(2 * window + 1 - 2 * b)$,相当于左右各看 window - b 个词。可以看出随着窗口的从左往右滑动,其大小也是随机的 $3(b = window - 1)$ 至 $2 * window + 1(b=0)$ 之间随机变动,即随机值 b 的大小决定了当前窗口的大小。代码中的 neu1 即为隐层向量,也就是上下文(窗口内除自己之外的词)对应 vector 之和。

```
// in -> hidden
for (a = b; a < window * 2 + 1 - b; a++) if (a != window) {
    c = sentence_position - window + a;
    if (c < 0) continue;
    if (c >= sentence_length) continue;
    last_word = sen[c];
    if (last_word == -1) continue;
    for (c = 0; c < layer1_size; c++) neu1[c] += syn0[c + last_word * layer1_size];
}
```

图 3-7 CBOW 模型输入层到隐层操作代码片段

3.3.3 Skip-gram

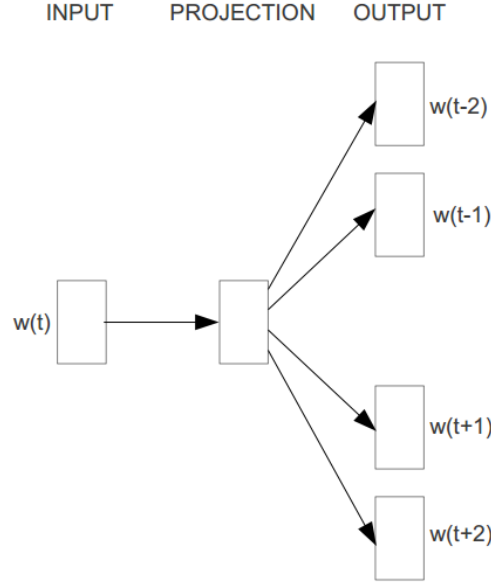


图 3-8 Skip-gram 模型

Skip-gram[49]与CBOW正好方向相反,是通过当前词预测上下文 $p(w_i|w_t)$,其中 $t-c \leq i \leq t+c$ 且 $i \neq t$, c 是上下文窗口大小的常数。 c 越大则需要考虑的词对(pair)越多,一般能够带来更精确的结果,但是训练时间也会增加。假设存在一个 $w_1, w_2, w_3, \dots, w_T$ 的词组序列, Skip-gram 的目标是最大化:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t)$$

基本的 Skip-gram 模型定义为 $p(w_o|w_i)$ 为:

$$p(w_o|w_i) = \frac{\exp(v'_{w_o} v_{w_i})}{\sum_{w=1}^W \exp(v'_w v_{w_i})}$$

从公式不难看出, Skip-gram 是一个对称的模型,如果 w_t 为中心词时 w_k 在其窗口内,则 w_k 也必然在以 w_k 为中心词的同样大小窗口内,即:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_t|w_{t+j})$$

同时, Skip-gram 中的每个词向量表征了上下文的分布。Skip-gram 中的skip是指在一定窗口内的词两两都会计算概率,就算他们之间隔着一些词,这样的好处是“白色汽车”和“白色的汽车”很容易被识别为相同的短语。

3.3.4 Hierarchical Softmax

在 CBOW 和 Skip-gram 模型中, 都会用到 Hierarchical Softmax 算法或 Negative Sampling 算法来优化求解过程。前面提到 Skip-gram 中的条件概率为:

$$p(w_o|w_l) = \frac{\exp(v'_{w_o} v_{w_l})}{\sum_{w=1}^W \exp(v'_w v_{w_l})}$$

这其实是一个多分类的 Logistic Regression, 即 Softmax 模型, 对应的 Label 是 One-hot Representation[50], 有且仅有当前词对应的位置为 1, 其他为 0。

普通的方法是 $p(w_o|w_l)$ 的分母要对所有词汇表里的单词求和, 这使得计算梯度很耗时。

另外一种方法[51]是只更新当前 w_o, w_l 两个词的向量而不更新其他词对应的向量, 也就是不管归一化项, 这种方法也会使得优化收敛的很慢。

Hierarchical Softmax 则是介于两者之间的一种方法, 使用的办法其实是借助了分类的概念。假设我们是把所有的词都作为输出, 那么“桔子”、“汽车”都是混在一起。而 Hierarchical Softmax 则是把这些词按照类别进行区分的。对于二叉树来说, 则是使用二分类近似原来的多分类。例如给定 w_l , 先让模型判断 w_o 是不是名词, 再判断是不是食物名, 再判断是不是水果, 再判断是不是“桔子”。虽然 word2vec 论文[45]里, Mikolov 是使用哈夫曼编码构造的一连串两分类。但是在训练过程中, 模型会赋予这些抽象的中间结点一个合适的向量, 这个向量代表了它对应的所有子结点。因为真正的单词公用了这些抽象结点的向量, 所以 Hierarchical Softmax 方法和原始问题并不是等价的, 但是这种近似并不会显著带来性能上的损失同时又使得模型的求解规模显著上升。

结合了哈夫曼编码的 Hierarchical Softmax 算法, 每个词 w 都可以从树的根结点沿着唯一的一条路径被访问到。这种哈夫曼编码用于神经网络语言模型并不是 word2vec 首创, 作者在他之前的论文中就有提到。假设 $n(w, j)$ 为这条路径上的第 j 个结点, 且 $L(w)$ 为这条路径的长度, 注意 j 从 1 开始编码, 即 $n(w, 1) = \text{root}$, $n(w, L(w)) = w$ 。对于第 j 个结点, Hierarchical Softmax 定义的 Label 为 $1 - \text{code}[j]$, 这里其实也可以把 Label 定义为 $\text{code}[j]$, 得到的向量也差不多。

1) 在 CBOW 模型中, 输出 f 为:

$$f = \sigma(\text{neu1}^T \cdot \text{syn1})$$

Loss 为负的 Log 似然, 即:

$$\text{Loss} = -\text{Likelihood} = -(1 - \text{code}[j]) \log f - \text{code}[j] \log(1 - f)$$

梯度为:

$$\begin{aligned}
 \text{Gradient}_{\text{neu1}} &= \frac{\partial \text{Loss}}{\partial \text{neu1}} \\
 &= -(1 - \text{code}[j]) \cdot (1 - f) \cdot \text{syn1} + \text{code}[j] \cdot f \cdot \text{syn1} \\
 &= -(1 - \text{code}[j] - f) \cdot \text{syn1} \\
 \text{Gradient}_{\text{syn1}} &= \frac{\partial \text{Loss}}{\partial \text{syn1}} \\
 &= -(1 - \text{code}[j]) \cdot (1 - f) \cdot \text{neu1} + \text{code}[j] \cdot f \cdot \text{neu1} \\
 &= -(1 - \text{code}[j] - f) \cdot \text{neu1}
 \end{aligned}$$

2) 在 Skip-gram 模型中, 输出 f 为

$$f = \sigma(v'_{n(w,j)}{}^T v_l)$$

条件概率 $p(w|w_l)$ 为:

$$p(w|w_l) = \prod_{j=1}^{L(w)-1} \sigma(\llbracket n(w, j+1) = \text{ch}(n(w, j)) \rrbracket \cdot v'_{n(w,j)}{}^T v_l)$$

其中:

$$\llbracket x \rrbracket = \begin{cases} 1, & \text{if } x \text{ is true} \\ -1, & \text{else} \end{cases}$$

$\text{ch}(n(w, j))$ 既可以是 $n(w, j)$ 的左儿子, 也可以是右儿子结点。

$\text{Loss}_{\text{pair}}$ 为负的 Log 似然 (因采用随机梯度下降, 这里只看一个 pair), 即:

$$\begin{aligned}
 \text{Loss}_{\text{pair}} &= -\log \text{Likelihood}_{\text{pair}} \\
 &= -\log(p(w|w_l)) \\
 &= -\sum_{j=1}^{L(w)-1} \log(\sigma(\llbracket n(w, j+1) = \text{ch}(n(w, j)) \rrbracket \cdot v'_{n(w,j)}{}^T v_l))
 \end{aligned}$$

与 CBOW 相比, 这里的 Loss 公式把 $\sigma(\text{neu1}^T \cdot \text{syn1})$ 变成 $\sigma(v'_{n(w,j)}{}^T v_l)$, 则对应的第 j 层的 loss 为:

$$\text{Loss} = -\log \text{Likelihood} = -\log(\sigma(\llbracket n(w, j+1) = \text{ch}(n(w, j)) \rrbracket \cdot v'_{n(w,j)}{}^T v_l))$$

A) 如果 $\llbracket n(w, j+1) = \text{ch}(n(w, j)) \rrbracket$ 为 true, 即当前结点为左儿子, 则:

$$\text{Loss} = -\log(\sigma(v'_{n(w,j)}{}^T v_l))$$

梯度为:

$$\text{Gradient}_{v_{n(w,j)'}} = \frac{\partial \text{Loss}}{\partial v'_{n(w,j)}} = -(1 - \sigma(v'_{n(w,j)}{}^T v_l)) \cdot v_l$$

$$\text{Gradient}_{v_l} = \frac{\partial \text{Loss}}{\partial v_l} = -(1 - \sigma(v'_{n(w,j)}{}^T v_l)) \cdot v'_{n(w,j)}$$

B) 如果 $\llbracket n(w, j+1) = ch(n(w, j)) \rrbracket$ 为 false, 即当前结点为右儿子, 则:

$$\text{Loss} = -\log(\sigma(-v'_{n(w,j)}{}^T v_l)) = -\log(1 - \sigma(v'_{n(w,j)}{}^T v_l))$$

梯度为:

$$\text{Gradient}_{v_{n(w,j)'}} = \frac{\partial \text{Loss}}{\partial v'_{n(w,j)}} = \sigma(v'_{n(w,j)}{}^T v_l) \cdot v_l$$

$$\text{Gradient}_{v_l} = \frac{\partial \text{Loss}}{\partial v_l} = \sigma(v'_{n(w,j)}{}^T v_l) \cdot v'_{n(w,j)}$$

合并 A)和 B)得

$$\text{Gradient}_{v_{n(w,j)'}} = \frac{\partial \text{Loss}}{\partial v'_{n(w,j)}} = -(1 - \text{code}[j] - \sigma(v'_{n(w,j)}{}^T v_l)) \cdot v_l$$

$$\text{Gradient}_{v_l} = \frac{\partial \text{Loss}}{\partial v_l} = -(1 - \text{code}[j] - \sigma(v'_{n(w,j)}{}^T v_l)) \cdot v'_{n(w,j)}$$

3.3.5 Negative Sampling

Negative Sampling 的原理在我翻译的文献中介绍的很详细, Mikolov 使用了噪声对比估计(Noise Contrastive Estimation, NCE)理论进行方法论证。简单来说, 负抽样就是随机抽取一些样例作为负样本, 被抽中的词 Label 为 0, 其余的为 1。

在 CBOW 模型中, 输出 f 不变, 仍为:

$$f = \sigma(\text{neu1}^T \cdot \text{syn1})$$

Loss 为负 log 似然为:

$$\text{Loss} = -\text{Likelihood} = -\text{label} \cdot \log f - (1 - \text{label}) \cdot \log(1 - f)$$

梯度为:

$$\text{Gradient}_{\text{neu1}} = \frac{\partial \text{Loss}}{\partial \text{neu1}} = -\text{label} \cdot (1 - f) \cdot \text{syn1} + (1 - \text{label}) \cdot f \cdot \text{syn1}$$

$$\begin{aligned}
&= -(\text{label} - f) \cdot \text{syn1} \\
\text{Gradient}_{\text{syn1}} &= \frac{\partial \text{Loss}}{\partial \text{syn1}} = -\text{label} \cdot (1 - f) \cdot \text{neu1} + (1 - \text{label}) \cdot f \cdot \text{neu1} \\
&= -(\text{label} - f) \cdot \text{neu1}
\end{aligned}$$

在 Skip-gram 模型中, Negative Sampling 和隐层往输入层传播梯度部分与 CBOW 类似, 不再赘述。

3.4 无监督疾病预测模型

3.4.1 住院记录矩阵 R

将 3.2.2 中抽取到的 3 类特征数据 Diagnoses、Medications 和 Procedures 统一定义为医疗事件 Events, 数量为 N, 则每一个 event 都拥有一个独立的 id。用 d_id、m_id 和 p_id 表示 3 类特征, 根据病人的住院记录构建医疗事件序列 Seq:

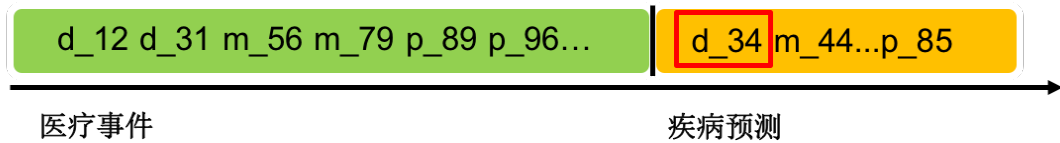


图 3-9 医疗事件序列 Seq

基于此, 我可以用 One-hot Representation 的方式根据病人的历史住院记录, 构造出一个 $1 \times N$ 的住院记录矩阵 $R = [r_{ij}]_{1 \times N}$, 即:

$$r(1, i) = \begin{cases} 1 \\ 0 \end{cases}, i \in N$$

其中, $r(1, i)=1$ 表示 $Events[i]$ 出现在该病人的住院记录中; $r(1, i)=0$ 表示 $Events[i]$ 未发生在该病人的住院记录中。

如表 3-5 所示, 我将样本中的 Diagnoses、Medications 和 Procedures 分类依次加入 R 中, 通过 R 我们就可以构造出一个病人的历史住院记录, 描绘出一个完整的病人画像。

d_0	...	m_32	...	p_78	...
0	...	1	...	1	...

表 3-5 住院记录矩阵 R

3.4.2 特征矩阵 M

首先定义 $D = \sum cnt(d_id)$ 为样本中 Diagnoses 的数量，接着构造一个 $N \times D$ 的特征矩阵 $M = [m_{ij}]_{N \times D}$ ，即：

$$m(i, j) = model.similar(event[i], event[j]) \quad , i \in N, j \in D$$

其中，*model* 为 Word2Vec 训练的 Log-Bilinear 语言模型，训练中可以调节的参数有 *size*（向量空间维度），*window*（上下文窗口大小），*sg*（0:CBOW，1:skip-gram），*similar(event[i], event[j])* 为向量空间中两个词的余弦相似度。

	d_0	...	d_31
d_0	1	...	0.4568
...
m_32	0.7613	...	0.2548
...
p_78	0.2549	...	0.8451
...

表 3-6 特征矩阵 M

其中每个值其实表示的是两个医疗事件之间的相关性，在图 3-10 可以看出疾病和对应的治疗药物和体格检查在降维后的向量空间中是极其接近的，这也就表示他们的 *cos_similar* 值会非常的高，这有助于我们作进一步的推理和预测。

例如，与高血压病最接近的词为高血脂、抑郁症、糖尿病等。

可以在 word2vec 训练完成后，用 PCA 降维观察 3 类特征数据的分布情况。

3.4.3 预测模型 Y

在定义了 R 和 M 后，我大胆假设病人住院记录中医疗事件之间的余弦相似度累积值越大，则病人患有该疾病的可能性也就越高。其数学形式如下：


$$y = \operatorname{argmax} \left(\sum_{i \in N, j \in D} r(1, i) * m(i, j) \right)$$

其中, i, j 均为 $event_id$,

$$patient = [\sum_{i \in N, j \in D} r(1, i) * m(i, j), \dots]$$

$patient$ 为一个 $1 \times D$ 的疾病序列, 其中每一项都是病人住院记录中医疗事件与疾病相似度的加和。如表

d_0	d_1	d_2	...
0.6841	0.8943	0.7648	...



Diagnoses	Precisions
Renal failure	0.6841
Asthma	0.8943
Hypertension	0.7648
...	...

图 3-10 根据疾病编号转为疾病名称

因此, 基于之前的假设, 我们不难得出 $event[y]$ 即为最大可能性患的疾病, 这个模型计算非常简单, 只需要将两个小矩阵 R 和 M 相乘得到 Y , 最后取 Y 中最大值的下标作为 $event$ 的索引值, 即可得出可能性最高的疾病。

该模型有一个很明显的缺点, 就是认为病人之前住院记录中的所有医疗事件权重相等。然而, 我们可以很直观的想到, 距离当前时间越近的医疗事件应该具有较大权重; 相反, 年代久远的医疗事件则应该影响较小。因此, 下一节将时序性加入预测模型, 作为一个小的改进, 与模型 Y 进行一个横向对比。

3.4.4 时序预测模型

考虑到时间因素对于医疗事件的影响, 在阅读大量文献后, 我认为可以用指数衰减函数来拟合时序性对于疾病预测的影响。

某变量下降速度和它的值成比例时, 称其服从指数衰减。用符号可以表示为下面的微分方程:

$$\frac{dN}{dt} = -\lambda N$$

其中 N 是变量值, λ 是指数衰减常数, 方程的解为:

$$N(t) = N_0 \cdot \exp(-\lambda t)$$

其中 $N(t)$ 是在 t 时刻的变量值, $N_0 = N(0)$, 为变量初值, 即在 0 时刻的变量值。

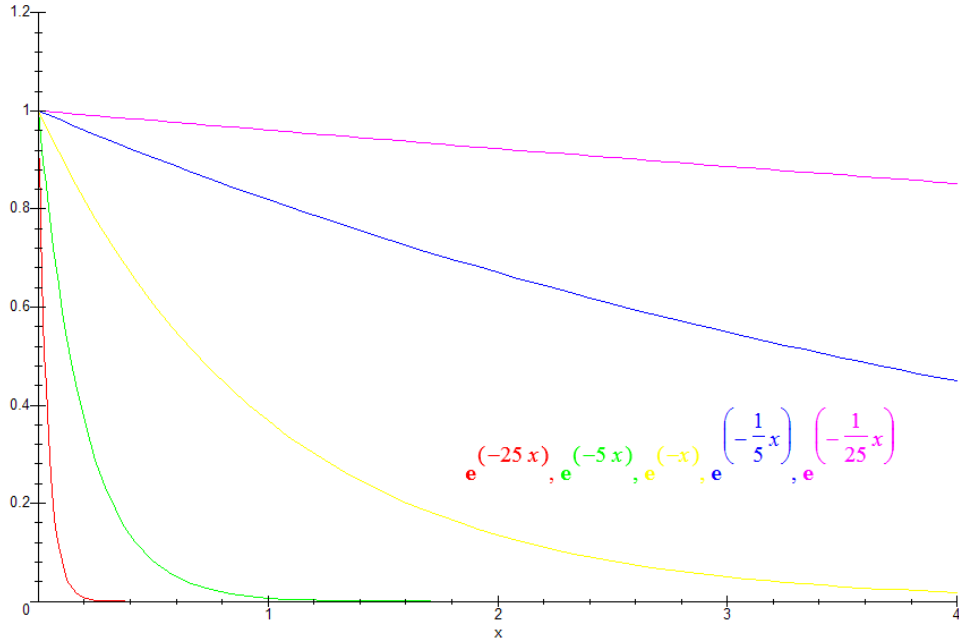


图 3-11 指数衰减函数图

定义时间权重 W :

$$W_i = \exp^{-\lambda * (t_{event} - t_0)}, \quad t_0 = t(0)$$

加入指数衰减函数后, 我可以得到全新的疾病序列:

$$patient' = [\sum_{i \in N, j \in D} \exp^{-\lambda * (t_{event}[i] - t_0)} * r(1, i) * m(i, j), \dots]$$

时序预测模型 $Y' = WRM$, 即:

$$y' = \operatorname{argmax} \left(\sum_{i \in N, j \in D} \exp^{-\lambda * (t_{event}[i] - t_0)} * r(1, i) * m(i, j) \right)$$

由于增加了新的参数 λ , 所以模型主要参数为 size、window 和 λ , 前两个为 word2vec 神经网络中的超参数。最终将 TF-IDF、预测模型 Y 和时序性预测模型 Y' 进行横向对比, 分析不同参数对于模型的影响。

3.5 本章小结

本章共分 4 个小节。

第 1 节主要概述了该实验的整体架构和研究方案。

第 2 节介绍 MIMIC III 数据集的来源和相关背景；通过自然语言处理，主要是依靠正则表达式和模式匹配进行特征数据的抽取。

第 3 节展示了 Word2Vec 在潜在语义提取、跨语言知识表达和逻辑推理上的优势，详细介绍了实现 Log-Bilinear 语言模型的两种算法 CBOW & Skip-gram，及其优化算法 Hierarchical Softmax 和 Negative Sampling。

第 4 节将 Log-Bilinear 语言模型与医疗事件相结合，构造了住院记录矩阵 R 、特征矩阵 M 和时序衰减因子 W ，提出了无监督疾病预测模型 Y ，以及时序预测模型 Y' 。

第4章 实验结果

4.1 数据描述

虽然本文采用的是全英文的 MIMIC III 数据集，但是在第 3 章中已经论证过 Word2Vec 在跨语言知识表达上的强大实力。因此，完全有理由相信当使用中文语料库时，预测模型的效果应当会相对稳定。同时，在现阶段没有可靠中文医疗数据集的情况下，使用英文数据集做实验，也不失为一种方法和策略。

4.1.1 MIMIC III 数据

MIMIC 官方使用 PostgreSQL 数据库搭建数据集，该数据库在北美学术界比较流行，完全开源，并且支持云服务和 Linux Ubuntu 平台。

4.1.1.1 数据结构

在 3.2 已经介绍了 MIMIC III 的各种表分类，我实验中主要处理非结构化的文本数据，即 NOTEEVENTS 这张病历记录表，共 4GB 大小。

其中 SUBJECT_ID 和 HAMD_ID 用来唯一标识病人和其在医院的记录；CATEGORY 为记录类型，如出院报告、护理报告、检查报告等；TEXT 为报告内容，医生或护理人员输入的自然语言。

在本文中，我将通过该表建立 7 张新表，下面简单介绍这些表的用途：

- DischargeReports: 所有超过 1 次住院记录的病人都可以被拿来作为测试样本，该表用来存储符合该条件病人的所有出院报告。
- SelDiagnoses: 通过文本挖掘将 DischargeReports 中的高频 Diagnoses 抽取出来，生成该表。
- SelMedications: 通过文本挖掘将 DischargeReports 中的高频 Medications 抽取出来，生成该表。
- SelProcedures: 通过文本挖掘将 DischargeReports 中的高频 Procedures 抽取出来，生成该表。

- SelEvents:将 selDiagnoses、selMedications 和 selProcedures 合并生成该表，统一编码，为每个医疗事件生成唯一的 id。
- TrainData:将最后 1 次出院前的报告当作训练数据，用来生成住院记录矩阵 R ，通过 id_test 与 TestData 关联。
- TestData: 由于无标注数据，将病人的最后 1 次出院报告作为测试集，测试无监督预测模型的效果。

4.1.1.2 数据处理

NoteEvents: 7,714 Discharge Reports;

TrainData: 3,256 Discharge Reports;

TestData: 2,117 Discharge Reports;

SelDiagnoses: 32 条高频疾病;

SelMedications: 32 条高频药物;

SelProcedures: 34 条高频体格检查;

SelEvents: 98 条高频医疗事件。

最终，经过数据处理的数据集特征如下描述：

- id_train: 训练集数据编号，TrainData 主键;
- id_test: 测试集数据编号，TestData 主键，TrainData 外键;
- id_events: 医疗事件编号，SelEvents 主键，SelDiagnoses 等表的外键;
- diagnosis: 疾病在 SelEvents 中的 id_events;
- medication: 药物在 SelEvents 中的 id_events;
- procedure: 体格检查在 SelEvents 中的 id_events;
- original_name: 医疗事件在原文中的名字
- connected_name: 医疗事件连接后的名字。
- event: SelDiagnoses 等表中 original_name 的整合

4.1.2 潜在语义提取

NoteEvents 的所有报告均用作 Word2Vec 训练的语料库，单词总数为 2,614,679,835 words(2,312,394,523 effective words)，词汇量为 612,837。

正如 Mikolov 在论文中举的例子一样，在经过大量医学报告训练后的 Log-Bilinear 语言模型也拥有强大的医学推理能力，如：

$$\text{vec}(\text{"sildenafil"}) \approx \text{vec}(\text{"torsemide"}) - \text{vec}(\text{"renal_failure"}) + \text{vec}(\text{"hypertension"})$$

其中，torsemide（拖拉塞米）是治疗 renal failure（肾功能衰竭）的袢利尿药，而 sildenafil（西地那非）是治疗 hypertension（高血压）的口服药。在没有任何人工参与的情况下，仅仅凭借神经网络训练就能达到如此效果。

并且，Word2Vec 还有一定的语义聚类效果，意思越相近的词在向量空间中的余弦相似度就越大。

Word	Cos_similar
hypertension	1.0000
htn	0.6692
hypotension	0.5383
depression	0.5027
siadenitis	0.4994
atrial_fibrillation	0.4909
hypertensive	0.4876
insufficinecy	0.4852
neurogenic	0.4817
diabetes_mellitus	0.4808
insufficiency	0.4783

表 4-1 与高血压余弦距离最近的单词

将 3 类特征数据用 PCA 降维后，可以非常明显的看到聚类的效果。尤其是 Diagnoses，聚类效果最为明显，几乎只有一个中心点，形成了一个类。而 Procedures 相对来说就要发散一些，有一个点甚至靠近 Medications 的中心点。

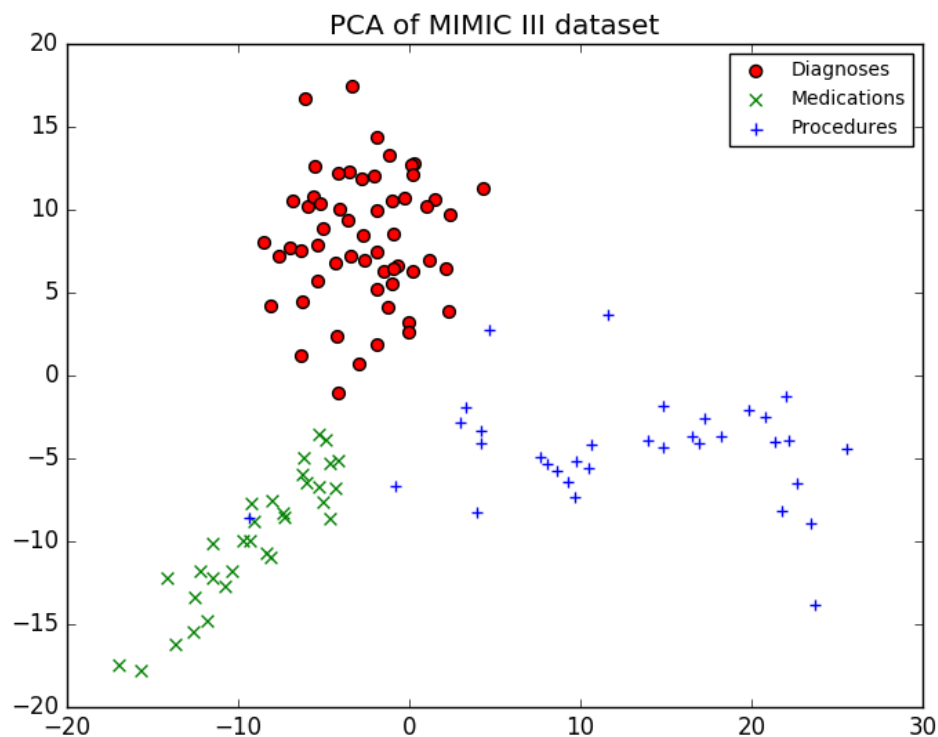


图 4-1 PCA 降维后的 3 类特征数据，聚类效果显著

4.2 Word2Vec 训练参数

根据 Google 训练 Wikipedia 的参数和其他文献中的经验值，我希望能够找出适合本文语料库的 size 和 window 最优值。

	MIMIC III	Wikipedia
File Size	3.7GB	12GB
Total Words	2,614,679,835	1,994,415,728
Total Vocabulary	612,837	1,969,354
Size	100-400	400
Window	50-200	5

表 4-2 MIMIC III 与 Wikipedia 数据规模和数据类型比较

4.2.1 确定最佳向量空间维度 size

由于 Wikipedia 词汇量达到了 MIMIC III 的 3 倍多，因此将 size 分为 100/200/300/400 训练，经过对比发现仍然是 size=400 时准确率最高。

parameters	precision			
	window=200 size=100	window=200 size=200	window=200 size=300	window=200 size=400
hypertension	0.458	0.572	0.821	0.891
sepsis	0.667	0.697	0.735	0.794
diabetes_mellitus	0.491	0.521	0.615	0.632
hyperthyroidism	0.491	0.521	0.814	0.853
aspiration_pneumonia	0.454	0.742	0.784	0.769
dyslipidemia	0.333	0.714	0.754	0.743
arthritis	0.560	0.667	0.707	0.750
asthma	0.354	0.584	0.624	0.667
respiratory_failure	0.541	0.561	0.481	0.524
anxiety	0.484	0.504	0.504	0.547
obstructive_pulmonary	0.197	0.217	0.312	0.355

表 4-3 对比不同 size 对疾病预测准确率的影响

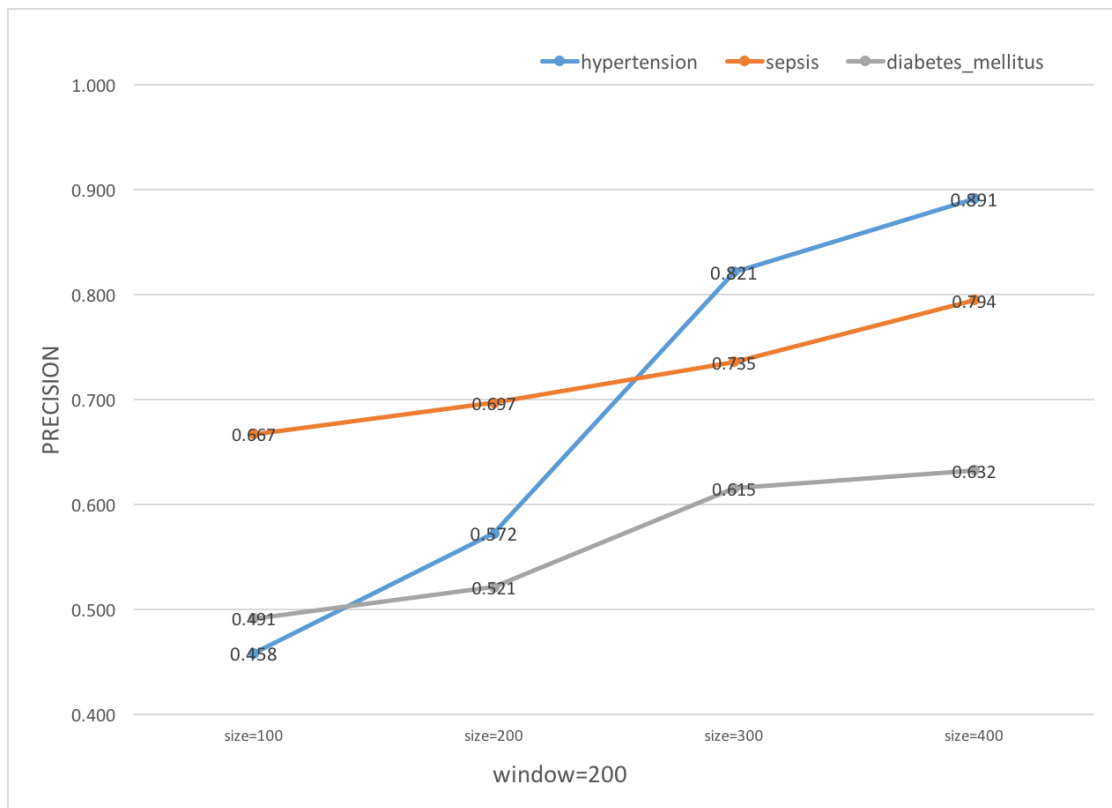


图 4-2 对比不同 size 对高血压、败血症和糖尿病的影响

4.2.2 确定最佳上下文窗口 window

由于临床报告中含有大量噪声，因此选择较大的上下文窗口，实验结果也证明 window 在 200 时效果最好，相对于 size，window 的影响并不算大。

parameters	precision			
	size=400 window=50	size=400 window=100	size=400 window=150	size=400 window=200
hypertension	0.833	0.856	0.871	0.891
sepsis	0.736	0.759	0.774	0.794
diabetes_mellitus	0.568	0.593	0.609	0.632
hyperthyroidism	0.766	0.789	0.804	0.853
aspiration_pneumonia	0.752	0.775	0.790	0.769
dyslipidemia	0.705	0.717	0.760	0.743
arthritis	0.661	0.673	0.716	0.750
asthma	0.578	0.590	0.633	0.667
respiratory_failure	0.435	0.447	0.490	0.524
anxiety	0.458	0.470	0.513	0.547
obstructive_pulmonary	0.266	0.278	0.321	0.355

表 4-4 对比不同 window 对疾病预测准确率的影响

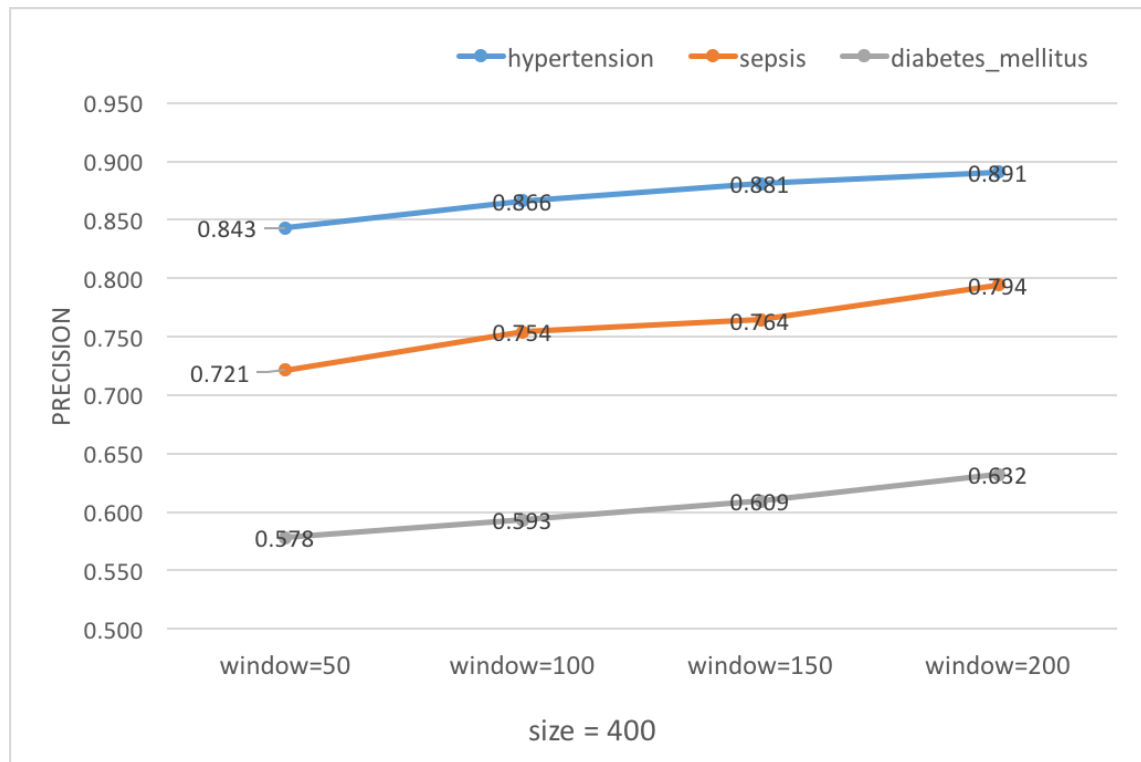


图 4-3 对比不同 window 对高血压、败血病和糖尿病的影响

4.3 疾病预测模型对比

首先，考察时间因素对预测结果的影响，找到 λ 的最优值，验证时序性预测模型是否能够提升预测的准确率。

parameters	precision			
	window=200	window=200	window=200	window=200
	size=100	size=400	size=400	size=400
	lambda=5	lambda=1	lambda=0.2	lambda=0.04
hypertension	0.781	0.874	0.915	0.911
sepsis	0.659	0.847	0.893	0.889
diabetes_mellitus	0.708	0.833	0.884	0.878
hyperthyroidism	0.688	0.831	0.872	0.868
aspiration_pneumonia	0.668	0.816	0.852	0.848
dyslipidemia	0.626	0.788	0.815	0.806
arthritis	0.870	0.747	0.784	0.780
asthma	0.807	0.650	0.715	0.711
respiratory_failure	0.666	0.571	0.594	0.590
anxiety	0.625	0.539	0.555	0.551
obstructive_pulmonary	0.492	0.395	0.412	0.408

表 4-5 对比不同 λ 对准确率的的影响

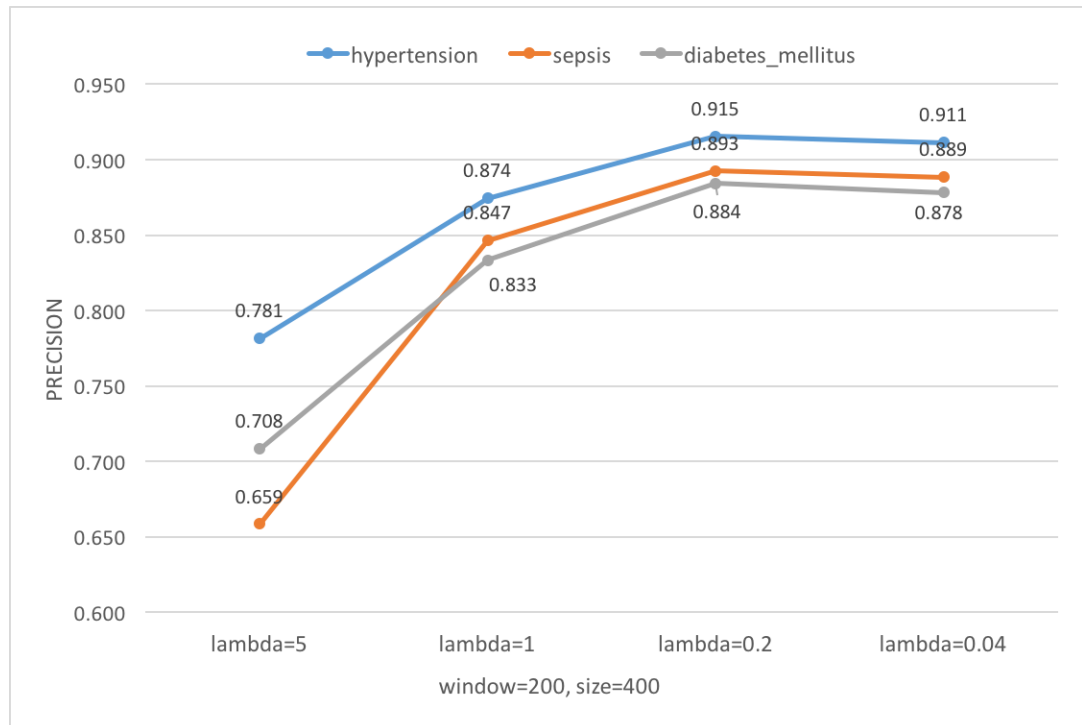


图 4-4 对比不同 λ 对高血压、败血症和糖尿病准确率的影响

由图 4-4 明显可以看出当 $\lambda=0.2$ 时，预测准确率在高频的几个疾病中达到最高。此时，我们已经得到模型的所有最优参数，即 $\text{size}=400$ 、 $\text{window}=200$ 和 $\lambda=0.2$ 。将其与不考虑时序性的朴素预测模型进行对比。

parameters	precision		
	Skip-gram $\lambda=0$	CBOW $\lambda=0$	CBOW $\lambda=0.2$
hypertension	0.884	0.891	0.915
sepsis	0.839	0.794	0.893
diabetes_mellitus	0.698	0.632	0.884
hyperthyroidism	0.824	0.853	0.872
aspiration_pneumonia	0.755	0.769	0.852
dyslipidemia	0.767	0.743	0.815
arthritis	0.739	0.750	0.784
asthma	0.667	0.667	0.715
respiratory_failure	0.509	0.524	0.594
anxiety	0.513	0.547	0.555
obstructive_pulmonary	0.368	0.355	0.412

表 4-6 对比朴素预测模型和时序预测模型

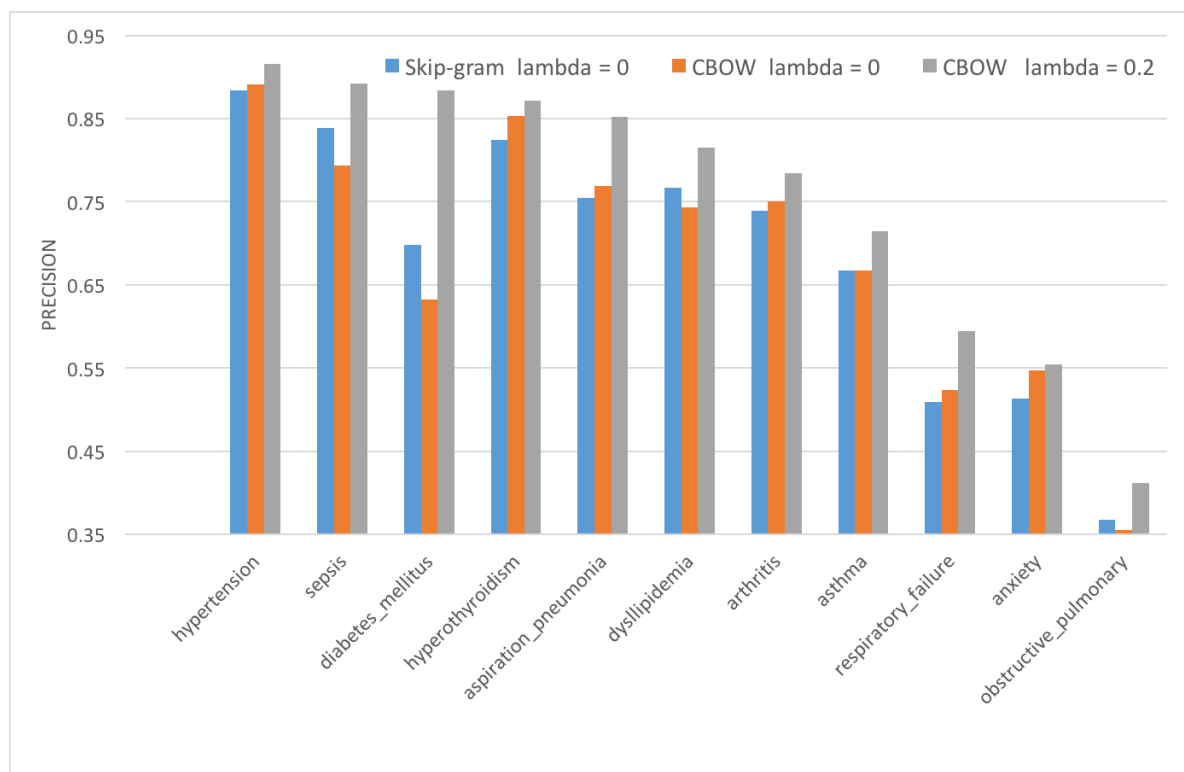


图 4-5 朴素预测模型和时序预测模型的对比

4.4 本章小结

本章主要通过 MIMIC III 数据集来验证第 3 章中提出的自然语言处理方法、word2vec 的潜在语义提取和无监督学习的疾病预测模型，以及时序疾病预测模型。分析了 Log-Bilinear 语言模型的训练结果，对特征数据使用 PCA 降维聚类，同时在医学知识上再次印证了词向量的逻辑推理能力，并通过与 Wikipedia 数据集的对比确定了 word2vec 的最佳参数。最后，简单分析了不同参数和时序性对疾病预测结果的影响。

第5章 分析与讨论

5.1 Word2Vec 潜在语义提取分析

Word2Vec 通过训练 Log-Bilinear 语言模型，把文本内容的处理简化为 size 维向量空间中的向量运算，通过向量空间的余弦距离来表示文本语义上的相似度，将潜在的语义用词向量的方式表示出来，本身就具备了一定的聚类效果。经过 PCA 降维后，聚类效果更加明显，可以清楚的看到 Diagnoses、Medications 和 Procedures 汇聚为 3 个大类。

同时，经过十亿级单词的训练后，以及高 window 值的训练，强化了模型的知识表达和逻辑推理能力。在没有任何人工标注的情况下，模型仍然能够完成肾衰竭和高血压及其治疗药物的类比推理，说明无监督学习的疾病预测模型还是具有继续深入发展的可能性。

5.2 影响疾病预测模型的因素分析

5.2.1 向量空间维度对疾病预测的影响

向量空间维度决定了 Word2Vec 能够映射的类别数 C 大小，CBOW 是一种与前向 NNLM 类似的模型[49]，即将词 w_t 经映射 $C(w_t)$ 为一个词向量。

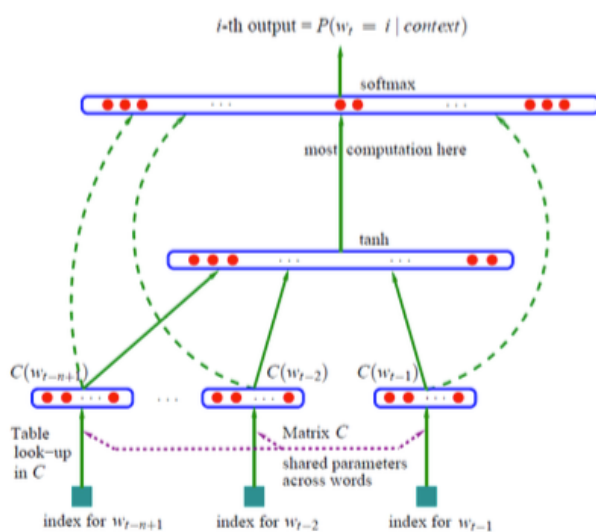


图 5-1 NNLM 模型

因此，当 C 的取值较小时，模型无法区分不同类别的单词，导致在向量计算时产生大量“语义”误差。根据 Google 训练 Wikipedia 的参数设置，对比 MIMIC III 的数据规模和词汇量，尽可能的对比不同参数对准确率的影响，

5.2.2 上下文窗口对疾病预测的影响

上下文窗口决定了模型学习逻辑的能力，当上下文环境越复杂、越丰富时，模型能够建立和验证更多的概率分布。因此，Word2Vec 相对于传统的 n -gram 和 n -pos 统计语言模型效果好了不少。

但是随着上下文窗口的增大，计算量也成倍增长。而实际的文本，通常也是逐句分段表意的，所以一味增加 window 并不能保证效果一直提升。同时，本文相对 wikipedia 来说，噪音较大，且有不少是对疾病诊断无意义的时间、地点和副词。

因此，在实际的参数选择时，应当兼顾语料库特点和计算成本两方面的考虑，再根据实际的模型效果，综合作出决策。

5.2.3 时序性对疾病预测的影响

时序性对疾病预测的影响是直观的，由于人的身体在治疗后是会康复痊愈的。因此，并不是所有的疾病都会“终身”影响人的健康状况。所有，减小时间久远的医疗事件影响，增大最近发生的事件权重是能够提升模型效果的。

在实际的医生诊疗过程中，医生通常也只参考近期的病历记录，太过久远的信息则会选择性忽略。因此，应当可以构建更加复杂和带反馈机制的时序性模型，将医疗事件的影响性分类后，再做时间衰减，这样便符合专业医生的诊疗过程，应该能够进一步提升模型的效果。

5.3 本章小结

本章通过分析和总结第 4 章的实验结果，分析了 Word2Vec 潜在语义提取的结果，解释了词向量在知识表达上的优势缘由；然后分析了向量空间维度、上下文窗口大小、时序性对疾病预测的影响，为更细颗粒度的时序性模型提出了个人的建议。

第6章 本文总结

6.1 论文主要工作

本文从大数据在医疗健康服务领域的发展出发，调查了大量国内外医疗大数据分析的应用现状，概述了相关的算法和技术。并基于 MIT 开发的 MIMIC III 数据集，针对非结构化文本和自然语言进行电子病历的信息抽取。

利用十亿级单词医学报告，通过 Word2Vec 训练 Log-Bilinear 语言模型，将医疗事件特征数据经过 PCA 降维后，印证了该模型在逻辑推理和知识表达上的优势，同类型特征数据聚类效果明显。

然后，构建住院记录序列和特征矩阵，提出了一种无监督学习的疾病预测模型，并将时序性作为影响因子加入预测模型。

最终，通过对比 Wikipedia 数据集，确定了最佳参数 size 和 window，并分析了向量空间维度、上下文窗口大小及时间衰减系数对疾病预测的影响。

6.2 将来的工作

在未来的 10 到 20 年，医疗大数据分析的发展仍将面对一个问题：各医疗机构信息孤岛的连结短期内无法实现，这将导致结构化 EHR 数据格式不一致。在这种情况下，在医学人员的报告记录和病人的住院病历上进行文本挖掘，并进一步过滤、抽取特征，构造辅助医生决策的疾病预测或病情预警模型，承担一部分简单、可预测的病患需求，使医学资源能够达到更加合理的分配。

由于时间关系，本文并未能做到尽善尽美，仍然有以下几个问题可以继续进一步研究：

- 文本挖掘：虽然 MIMIC III 数据集中的纪录能够通过正则表达式和模式匹配进行特征抽取，但是扩展到实际应用场景中，如果仍然是针对每个数据集设计一套模式，代价非常大。如何结合深度学习和文本挖掘，做到高效、准确的电子病历信息抽取将会是一个值得研究和探讨的方向。
- 参数训练：由于本文的输出为多种疾病，类似一个多分类问题。但是由于每个病人可能同时患有多种疾病，则又不是一个单纯的多分类。如果能够设计一个目标函数和最优化函数，直接在 CBOW 或者 Skip-gram 的输出层上再嵌套一层应用层，通过训练和反馈直接取得最有参数 window、

size 和 λ ，将会使结果更有说服力。

- 时序性影响：本文采用了指数衰减函数来拟合时序性对疾病预测的影响，但是实际情况肯定会更加复杂。如果能够根据疾病类型，分类处理、标记，用更加复杂的函数来拟合，也许效果会更好。
- 对比模型：由于数据集规模太大，所以没有采用主题模型如 LDA 进行模型效果对比。而其他的预测或分类模型则大都需要数据标注，这一点也是不容易实现的，所以实验只对比了基于 CBOW/Skip-gram 的朴素模型和加入时序性的预测模型。

参考文献

- [1] Wullianallur Raghupathi, Viju Raghupathi, Big data analytics in healthcare: promise and potential, Health Information Science and Systems. 2014, 2:3
- [2] MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).
- [3] OVERPECK J T, MEEHL G A., BONY S. Climate data challenges in the 21st century [J]. Science, 2011, 331 (6018): 700-702.
- [4] ZHENG Z, ZHU J, LYU M R. Service-generated Big Data and Big Data-as-a-Service: An Overview. IEEE International Congress on Big Data, 2013, 403-410.
- [5] 黄哲学,曹付元,李俊杰. 面向大数据的海云数据系统关键技术研究[J]. 网络新媒体技术, 2012, 1(6):20-26.
- [6] COLLEN M F. A brief historical overview of hospital information system evolution in the United States [J]. International journal of bio-medical computing, 1991, 29(3):169-189.
- [7] 中国医院协会信息管理专业委员会.中国医院信息化发展研究报告(白皮书) [J].中国数字医学,2008, (6): 11-19.
- [8] AMARASINGHAM, RUBEN, BILLY J. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care, 2010, 48(11): 981-988.
- [9] MURDOCH, TRAVIS B, ALLAN S. The inevitable application of big data to health care. JAMA, 2013, 309(13): 1351-1352.
- [10] 陈功, 范晓薇, 蒋萌. 数据挖掘与医学数据资源开发利用[J]. 北京生物医学工程, 2010, 29(3): 323-328.
- [11] PARTA A, FAES L, MASE M. An integrated approach based on uniform quantization for the evaluation of complexity of short-term heart period variability: application to 24h Holter recordings in healthy and heart failure humans [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2007,17(1): 015117.

- [12] 胡新平. 医疗数据挖掘中的隐私保护 [J]. 医学信息学杂志, 2009, 8: 1-4
- [13] SCHAFFER, CULLEN. Selecting a classification method by cross-validation. *Machine Learning*, 1993, 13(1): 135-143
- [14] FRANTZIDIS, CHRISTOS A, CHARALAMPOS B. On the classification of emotional bio signals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, 2010, 14(2): 309-318.
- [15] KAUR H, WASAN S K. Empirical Study on Applications of Data Mining Techniques in Healthcare [J]. *Journal of Computer Science*, 2006, 2: 194-200.
- [16] MOSTELLER, FREDERICK, JOHN W T. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*. 1977.
- [17] 樊震林,黎爱军,吴宏.医疗风险影响因素的有序多分类 Logistic 回归分析 [J].*中国卫生质量管理*,2009, 16(4): 11-13.
- [18] KING, MATTHEW W, RESICK P A. Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees. *Journal of Consulting and Clinical Psychology*, 2014.
- [19] HANRAHAN, KIRSTEN. From Research Results to Prediction and Translation: A Decision Support System for Children, Parents, and Distraction During Healthcare Procedures. In *Sigma Theta Tau International's 23rd International Nursing Research Congress*. STTI. 2012.
- [20] ANDERBERG, MICHAEL R. Cluster analysis for applications (No. OAS-TR-73-9). OFFICE OF THE ASSISTANT FOR STUDY SUPPORT KIRTLAND AFB N MEX. 1973.
- [21] ANDREA, LIVIA M, GARY L. Cluster analysis of adult children of alcoholics. *Substance Use & Misuse*, 1994, 29(5): 565-582
- [22] HASTIE, BARBARA A, JOSEPH L. Cluster analysis of multiple experimental pain modalities. *Pain*, 2005, 116(3): 227-237.
- [23] BOTSIS, TAXIARCHIS, MICHAEL D. Text mining for the Vaccine

Adverse Event Reporting System: medical text classification using informative feature selection. Journal of the American Medical Informatics Association, amiajnl-2010

[24] BORTOLAN G, DEGANI R, WILLEMS J L ECG classification with neural networks and cluster analysis. In Computers in Cardiology 1991, Proceedings. (pp. 177-180). IEEE.

[25] YEH, YUN C, CHE W. Analyzing ECG for cardiac arrhythmia using cluster analysis. Expert Systems with Applications, 2012, 39(1): 1000-1010.

[26] JUNG, SUN J, CHANG S. Association Rules to Identify Complications of Cerebral Infarction in Patients with Atrial Fibrillation. Healthcare informatics research, 2013, 19(1): 25-32

[27] HUANG, YI C. Mining association rules between abnormal health examination results and outpatient medical records. Health Information Management Journal, 2013, 42(2): 23.

[28] DUAN, LIAN, NICK W. Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterprise Information Systems, 2011, 5(2): 169-181.

[29] KHALILIA, MOHAMMED, SOUNAK C. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 2011, 11(1): 51.

[30] Fei Wang, Noah Lee, Jianying HU, Jimeng Sun, Shahram Ebadollahi, and Andrew F.Laine. A Framework for Mining Signatures from Event Sequences and Its Applications in Healthcare Data. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 35, No.2, Feb 2013.

[31] STEVEN E, DILSIZIAN, ELIOT. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. Current Cardiology Reports, 2013, 16:441

[32] SHOUVAL R, BONDI O, MISHAN H. Application of machine learning algorithms for clinical predictive modelling: a data-mining approach in SCT. Bone marrow transplantation, 2013, 49(3): 332-337.

[33] GREEN K, STILPHEN ,VILENSKY S. Computerized Clinical Decision Support System For Early Identification Of Patients Appropriate For Rehabilitation Services Improves Functional Status In Survivors Of Critical Illness. *Am J Respir Crit Care Med*, 2013, 187: A3621

[34] GULTEPE E, GREEN J P, NGUYEN H From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*, 2014, 21(2), 315-325.

[35] TRIFIRO G, PARIENTE A, COLOMA P M. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? [J]. *Pharmacoepidemiology and drug safety*, 2009, 18(12): 1176-1184.

[36] KAO, ROWLAND R, DANIEL T. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends in microbiology*, 2014, 22(5): 282-291.

[37] BROSSETTE, STEPHEN E, ALAN P. Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American medical informatics association*, 1998, 5(4): 373-381.

[38] Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 1648(6), S106-S113.

[39] JENSEN, PETER B, LARS J. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 2012, 13(6): 395-405.

[40] ENGEL, GEORGE L. The need for a new medical model: a challenge for biomedicine. *Science*, 1977, 196(4286): 129-136.

[41] EVANS, STEVEN, STEPHEN J. Automated detection of hereditary syndromes using data mining. *Computers and biomedical research*, 1997, 30(5): 337-348.

[42] HUANG, QI R, ZHENXING Q. Clinical patterns of obstructive sleep apnea and its comorbid conditions: a data mining approach. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*

2008, 4(6): 543.

[43] MELZER, TRACY R, RICHARD W. Arterial spin labelling reveals an abnormal cerebral perfusion pattern in Parkinson's disease. *Brain*, awq377, 2011.

[44] YUANGYUANG X, QI L, LI J. Detecting adolescent psychological pressures form Micro-Blog. *Health Information Science*, 2014, LNCS8423: 83-94

[45] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*, 2013

[46] Broder, Andrei Z.; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey (1997). "Syntactic clustering of the web".*Computer Networks and ISDN Systems* **29** (8): 1157–1166.

[47] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?", *Proceedings of the IEEE*, 88(8), 1270-1288, 2000.

[48] Hinton, Geoffrey E. "Learning distributed representations of concepts." *Proceedings of the eighth annual conference of the cognitive science society*. 1986.

[49] Efficient Estimation of Word Representations in Vector Space. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. *ICML*, 2013

[50] Harris, David and Harris, Sarah. *Digital design and computer architecture* (2nd ed.). San Francisco, Calif.: Morgan Kaufmann. p. 129.

[51] Ferguson, Thomas S. (1982). "An inconsistent maximum likelihood estimate". *Journal of the American Statistical Association* **77** (380): 831–834.

致谢

首先，我衷心感谢我的导师郑小林副教授在学习和生活中给予我的谆谆教诲和悉心关怀。本论文从选题、构思、修改到成文，每个环节都凝结了实验室老师大量的心血，充斥了实验室学长学姐的耐心指导。郑小林教授专业知识渊博，治学态度严谨，在学术上精益求精、积极进取，在工作中求真务实，在生活中平易近人，给我留下了深刻的印象。在这短短一年的相处时间里，我深深感受到了郑老师出色的人格魅力，这在以后的求学道路上无疑给我树立了旗帜鲜明的方向标，是我一生取之不尽的宝贵财富。

同时，我要感谢 UCSD 生物医疗信息学系的蒋晓谦教授，在我出国访学期间，对我实验数据获取的支持，以及在论文构思和修改上的建议和帮助。

此外，我要感谢浙江大学电子服务研究中心的所有师兄师姐们，他们是陈超超博士、扈中凯博士、朱梦莹博士、苏赞文、洪福兴、方崇豪、马国芳、王磊。在学习和工作中，我总能从他们身上学到很多东西，我的每一点进步都离不开他们对我的支持和帮助。

感谢计算机学院的各位老师，谢谢你们在我本科生阶段，带领我进入丰富多彩的计算机科学与技术的世界。

感谢对我本科论文进行评审的各位专家教授，感谢你们对论文的指导和宝贵的意见。

我要特别感谢我的父母，他们在我的成长中起到了不可替代的影响，是我人生中最重要依靠。

最后，感谢所有给予我指导、帮助、关心和支持的老师、亲人和朋友们。

彭靖田

2016 年 5 月

本科生毕业论文（设计）任务书

一、题目：基于深度学习的疾病预测模型

二、指导教师对毕业论文（设计）的进度安排及任务要求：

该毕业论文要求在对深度学习和疾病预测模型相关国内外研究现状进行深入分析的基础上，重点围绕无监督疾病预测模型展开研究。实验数据要求采用 MIT 实验室开发的 MIMIC III 电子病历系统真实数据，实验结果要清晰可信。

进度安排如下：

3.01-3.15 研究方案确定

3.16-4.15 模型的建立与算法实现

4.15-5.15 实验与分析

5.16-5.30 论文撰写

起讫日期 200 年 月 日至 200 年 月 日

指导教师（签名）_____ 职称 _____

三、系或研究所审核意见：

负责人（签名）_____

年 月 日

毕 业 论 文（设计） 考 核

一、指导教师对毕业论文（设计）的评语：

指导教师(签名) _____
年 月 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述 占（10%）	开题报告 占（20%）	外文翻译 占（10%）	毕业论文（设计） 质量及答辩 占（60%）	总评 成绩
分值					

答辩小组负责人（签名） _____
年 月 日