DOI:10.7544/issn1000-1239.2021.20200402
58(3):513-527,2021

电子病历文本挖掘研究综述

吴宗友[1] 白昆龙[2,3,4] 杨林蕊[3,4,5] 王仪琦[2,3,4] 田英杰[1]
[1](中国科学院大学经济与管理学院 北京 100049)
[2](中国科学院大学计算机与科学技术学院 北京 100049)
[3](中国科学院虚拟经济与数据科学研究中心(中国科学院大学) 北京 100190)
[4](中国科学院大数据挖掘与知识管理重点实验室(中国科学院大学) 北京 100190)
[5](中国科学院大学中丹学院 北京 100049)
(bossbit@126.com)

Review on Text Mining of Electronic Medical Record

Wu Zongyou[1], Bai Kunlong[2,3,4], Yang Linrui[3,4,5], Wang Yiqi[2,3,4], and Tian Yingjie[1]
[1](*School of Economics and Management, University of Chinese Academy of Sciences, Beijing* 100049)
[2](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 100049)
[3](*Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences (University of Chinese Academy of Sciences), Beijing* 100190)
[4](*Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences (University of Chinese Academy of Sciences), Beijing* 100190)
[5](*Sino-Danish College, University of Chinese Academy of Sciences, Beijing* 100049)

Abstract Electronic medical records (EMR), produced with the development of hospital informationization and contained rich medical information and clinical knowledge, play important roles in guiding and assisting clinical decision-making and drug mining. Therefore, how to efficiently mine important information in a large amount of electronic medical records is an essential research topic. In recent years, with the vigorous development of computer technology, especially machine learning and deep learning, data mining in the special field of electronic medical records have been raised to a new height. This review aims to guide future development in the field of electronic medical record text mining by analyzing the current status of electronic medical record research. Specifically, this paper begins with an introduction to the characteristics of electronic medical record data and introduces how to preprocess electronic medical record data; then four typical tasks around electronic medical record data mining (medical named entity recognition, relationship extraction, text classification and smart interview) introduce popular model methods; finally, from the perspective of the application of electronic medical record data mining in characteristic diseases, two specific diseases of diabetes and cardio-cerebrovascular diseases are combined and a brief introduction to the existing application scenarios of electronic medical records is given.

Key words electronic medical records; natural language processing; data mining; machine learning; deep learning

收稿日期:2020-06-08;**修回日期**:2020-08-20
基金项目:国家自然科学基金项目(71731009,61472390);中国科学院科技服务网络计划项目(KFJ-STS-ZDTP-060)
This work was supported by the National Natural Science Foundation of China (71731009, 61472390) and the Science and Technology Service Network Program of Chinese Academy of Sciences (KFJ-STS-ZDTP-060).
通信作者:王仪琦(wangyiqi18@mails.ucas.edu.cn)

摘 要 电子病历是医院信息化发展的产物,其中包含了丰富的医疗信息和临床知识,是辅助临床决策和药物挖掘等的重要资源.因此,如何高效地挖掘大量电子病历数据中的信息是一个重要的研究课题.近些年来,随着计算机技术尤其是机器学习以及深度学习的蓬勃发展,对电子病历这一特殊领域数据的挖掘有了更高的要求.电子病历综述旨在通过对电子病历研究现状的分析来指导未来电子病历文本挖掘领域的发展.具体而言,综述首先介绍了电子病历数据的特点和电子病历的数据预处理的常用方法;然后总结了电子病历数据挖掘的4个典型任务(医学命名实体识别、关系抽取、文本分类和智能问诊),并且围绕典型任务介绍了常用的基本模型以及研究人员在任务上的部分探索;最后结合糖尿病和心脑血管疾病2类特定疾病,对电子病历的现有应用场景做了简单介绍.

关键词 电子病历;自然语言处理;数据挖掘;机器学习;深度学习

中图法分类号 TP391

在过去的几十年里,互联网信息技术蓬勃发展,使得数据的管理与传输变得更为高效,同时医疗机构也构建了大量的电子病历信息库.在众多的现代化医疗数据中,电子病历数据是最重要的医疗数据资源之一.每天都有海量的电子病历数据在各级医院产生,形成了医疗大数据的重要组成部分.电子病历以患者为主体,比较完整地记录患者长期的医疗信息,并且经过计算机技术的整理、储存、共享和分析,是纸质病历的升级版本.电子病历的特性使得相关信息库中同时包含着结构化数据和非结构化数据,结构化数据如表格数据,非结构化数据如文本数据和医学图像等.

对电子病历进行数据挖掘有巨大的应用前景,然而电子病历数据量大、非结构化数据多的特点使得数据挖掘十分困难.虽然早期部分研究人员利用一些简单的电子病历数据挖掘方法如决策树[1]等辅助诊断,但是由于技术和方法的限制,对医疗数据的信息挖掘效果十分有限.而近年来,机器学习和深度学习在其他领域展现的巨大潜力给电子病历的挖掘带来了新的希望[2].本文主要针对电子病历中的文本数据部分,结合国内外的研究现状,分析电子病历数据挖掘面临的挑战以及现有的解决方案.

1 电子病历

电子病历往往包含了患者就诊过程中产生的多种数据格式,如文本、图像、表格等.本文主要针对的是电子病历中的文本部分,包括患者基本信息、患者病史、患者的症状、医生的诊断说明等.电子病历中的文本数据同样分为结构化信息和非结构化信息.结构化信息如身高、体重等往往通过简单的数据清洗就可以作为各种机器学习算法的输入;非结构化的信息如诊断信息、用药信息、检查信息、临床记录等,这些数据则需要较为复杂的预处理和自然语言处理(natural language processing, NLP).对电子病历信息进行分析与利用需要大量有专业知识的人工付出大量的时间,成本昂贵.而NLP的发展为自动处理电子病历文本数据提供了基础.对电子病历的数据挖掘有助于节省人力物力,提高医生的诊断效率,实现智慧医疗.

电子病历的数据挖掘始于数据收集和预处理,在获得海量电子病历数据后需要针对数据的特性和挖掘分析的目标对数据进行预处理.电子病历数据预处理常用的方法主要包括脱敏处理、数据清洗、数据集成、数据选择和数据规约.

1) 脱敏处理.电子病历的一个特性是隐私性,这意味着相关数据的收集只能以大型的医院信息部门或者一些专业的医疗机构作为数据来源,并且公开数据集时要对数据进行处理,隐藏掉部分信息以保护患者的隐私权.

2) 数据清洗.数据清洗其主要目的是将错误或杂乱无章的数据处理成干净、标准的数据以供后续数据统计和挖掘使用.数据清洗主要包括补齐、去重和降维等方法,需要针对不同数据选择具体的方法.对电子病历文本挖掘前的数据清洗需要考虑电子病历的特点,电子病历的建立通常需要医护人员和患者协作,在记录过程中可能会有数据缺失、噪声数据、错误语义甚至是自相矛盾的数据或文本存在.对于缺失的数据,通常可以只删除带有缺失数据的样本或者采用均值将缺失数据填充,再者,可以采用类似于回归、贝叶斯、决策树等机器学习方法来确定填充数据的最佳值.而噪声数据通常指的是数据出现了明显的不正常数值,如血压数值高于常人数倍,面对这样的数据,通常采用平滑处理或异常值分析的

方法.平滑处理通常用数据周围点的均值进行处理,而异常值分析方法通常通过聚类方法来构建类别从而处理数据.另外,面对一些语法错误和语义错误,只能通过人为再编辑或者 NLP 技术进行修正.

3) 数据集成.经过数据清洗过后干净的数据需要经过数据集成操作,数据集成通常是指将不同源头的电子病历数据集成到同一个数据库中,可以扩大数据规模,方便模型的训练和后续算法的研究.但是数据集成也存在着问题,从不同源集成的数据之间可能存在结构的不一致或数据存在冗余的现象,那么整理之后的数据要重新进行数据清洗.在电子病历数据中,同一个患者的数据可能来自不同医院的不同科室,这些数据很容易出现异构或重复的现象.

4) 数据选择.针对不同的研究目标,对整理出的电子病历数据进行内容筛选处理也十分重要.根据研究目的的不同,选择出电子病历中不同的病历记录.经过数据选择确定研究数据,一方面可以剔除掉无关属性和噪声对研究目标的影响,另一方也起到了对数据的降维作用.高质量的数据选择是数据预处理中的关键步骤,影响着研究目标的实现结果.

5) 数据规约.数据规约是对电子病历文本数据进行规范化调整,将数据调整为适合进行数据挖掘的形式,包括对输入数据的归一化处理(最小最大归一化、零均值归一化和分数归一化等)、遗漏数据的处理以及错误信息的纠正.数据规约是为了使数据更加规范化,使得数据在接近原始数据的基础上更加易于处理.

2 电子病历数据挖掘任务和方法

电子病历中包含着大量非结构化文本信息,要从非结构化的文本信息中挖掘出潜在的规律需要识别出大量专业词汇和如疾病-症状等特殊鲜明的实体关系,要对这些文本信息进行数据挖掘,关键的 2 个基本任务分别为命名实体识别和关系抽取.近年来,随着 NLP 技术的发展,对电子病历数据的分类任务和问答任务也有了一定的突破.下文将简单介绍电子病历文本挖掘中常见的 4 种任务:命名实体识别、关系抽取、文本分类和问答系统,并介绍任务常用的数据分析方法.

2.1 医学命名实体识别

命名实体识别(named entity recognition, NER)也称为概念抽取,即从指定的自由文本中抽取出相关的具有特定意义的词语,它在医学文本研究中被称为生物医学命名实体识别(biomedical named entity recognition, BioNER).电子病历命名实体识别是 BioNER 的子领域,其主要任务是识别出患者的电子病历中具有特定意义的实体,并对它们进行标注,这些实体根据研究目的不同而有所区别.通常中文电子病历中的实体类型包括疾病、病因、临床表现、检查方法、药品名称、手术、身体部位等.电子病历在被标注实体之后可以提高医生查看病历的工作效率.同时,标注的结果也将辅助后续的如关系抽取和知识图谱构建等研究.随着相关技术的进步和研究的进展,发展出了很多电子病历命名实体识别的方法,最开始的方法是基于词典与规则的方法和基于统计学的机器学习方法.而在过去几年中,基于深度学习的方法在该任务中效果显著并得到广泛使用,如 2018 年出现的 BERT(bidir-ectional encoder representation from transformers)等深度学习框架进一步改善了生物医学命名实体识别的性能.目前主流的中文电子病历命名实体识别的方法仍然是条件随机场和双向长短时记忆网络[3].下面将对 3 类不同方法进行详细梳理.

1) 基于词典和规则的方法

基于词典的方法在识别过程中通常是依靠术语词典,然后采取匹配算法进行命名实体识别.因此,对于电子病历这种专业性较强的文本,标注语料即词典的规模和质量起到了相当关键的作用.在医疗领域,中文电子病历的标注规范也在不断的探索当中,并形成语料库,如曲春燕等人[4]在 2015 年参照 i2b2 2010 的标注规范制定了中文电子病历的标注规范,并在 2 名临床医生的帮助下对标准语料进行了检验,后来在他们的标注语料基础上,一些研究人员也进行了改进[5-6],这些都使得中文电子病历标注语料规模和质量变得更加可靠.虽然完全基于词典的命名实体识别准确率有一定的保障,但是电子病历的标注语料库的构建需要医疗专业知识,通常需要医学方面的专业人员共同协作,且随着时间的推移语料库的维护也耗时耗力.由于词典规模有限且需要及时更新等原因,仅使用词典往往并不能取得特别好的效果,因此后来词典常作为特征帮助以提升自然语言处理的效果.

与基于词典的方法不同,基于规则的方法主要是通过对整个文本进行分析来构建规则模板,利用规则模板,通过匹配的方式实现命名实体的识别.基于规则的方法更加直观且方便维护,但规则的构建也需要相关领域专家的人力且耗时较大,而且在没有

明显规则时基于规则模板的识别将较困难.同时,不同领域文本与实体大不相同,无法从某个医学领域直接扩展到其他医学领域.因此,规则与词典相似,后来也通常被用于辅助命名实体识别模型.例如利用规则优化词典等的特征,再结合条件随机场对中文电子病历进行命名实体识别,比单纯使用条件随机场的效果要好[7].

2) 基于统计学习的方法

随着机器学习的发展和流行,针对词典和规则的方法存在的缺点,基于统计机器学习进行命名实体识别的方法被提出并得到深入的研究和应用,词典和规则则作为一种辅助手段用于提高机器学习实体识别的效果.机器学习方法需要的专业人工相对较少,成本也较低,所以近年来应用相对广泛.传统机器学习方法可以分为有监督学习、半监督学习和无监督学习 3 类.其中有监督学习方法在命名实体任务中占了主流,它通常需要大规模带标签的训练集,将命名实体识别任务转换成分类问题,训练集用于模型的训练,生成目标模型后才可以对未标注语料中的实体进行识别.常用的序列标注模型包括隐马尔可夫模型、最大熵模型、条件随机场模型和支持向量机等.

① 隐马尔可夫模型.隐马尔可夫模型(hidden Markov model, HMM)最初由 Bikel 等人[8]提出并发表在统计学的系列论文中,该模型在后续研究中被证实在语言识别、自然语言处理以及生物信息学等多个领域的应用都体现了很大价值[9].

在序列标注中使用 HMM 时,目的在于给定观测序列是 $\boldsymbol{X}$ 的条件下,求解使条件概率 $P(\boldsymbol{X}|\boldsymbol{Y})$ 最大的标记序列 $\boldsymbol{Y}^*$.根据贝叶斯公式推导可知,HMM 的实质为求解联合概率 $P(\boldsymbol{X},\boldsymbol{Y})$.在获得模型参数后,命名实体识别问题的解码(常用 Viterbi 算法)过程目标为得到相对于观测序列的最优命名实体标记序列,解码序列.HMM 虽然是序列标注的最常用且有效的方法之一,然而 HMM 是以独立性假设为前提的,即观测元素为独立于观测序列中的其他元素的单元.事实上元素之间一般并非独立,且可能具有长距离依赖关系,如文本语句远距离上下文之间的语义联系,严格的独立性假设不能够真实地描述数据序列所包含的信息,这是 HMM 的主要缺陷.

② 最大熵模型.熵(entropy)[10]表示能量在空间中分布的均匀程度.香农在描述信息量时用了这个概念,提出了信息熵的概念,来表示系统的平均信息量.最大熵模型(maximum entropy, ME)是在最大熵原理[11]的基础上实现的,主要思想是在已知部分知识的前提下选择熵最大的概率分布,即在满足约束条件的情况下选择不确定性最大,信息量最大的模型.

最大熵模型在特征选择时相对灵活可以引入特征提高模型的准确率,且不需要 HMM 必须的独立性假设.但是其迭代过程计算量巨大,计算的时间复杂度较高.

③ 条件随机场模型.条件随机场(conditional random fields, CRFs)[12-13]是一种用于序列标记任务的概率统计模型.CRF 是最大熵 HMM 模型在标注问题上的改进.假设 $\boldsymbol{X}$,$\boldsymbol{Y}$ 分别表示为需要标记的观测序列和相对应的标记序列的联合分布的随机变量,那么 CRF 就是一个以观测序列 $\boldsymbol{X}$ 为作为全局条件的无向图模型.在命名实体识别任务中,$\boldsymbol{X}$ 可能是一句话,而 $\boldsymbol{Y}$ 则是相对应的类别标记序列.在对标记序列进行建模时,最简单也是最常用的图形结构就是:观测节点与标记序列中的节点构成简单的一阶链形式,此时图中的标记序列形成了一条马尔可夫链.CRF 克服了 HMM 的独立性假设条件,考虑了整个 $\boldsymbol{X}$ 即上下文的信息,虽然也具有时间复杂度大导致的训练难度高等问题,但是仍然被广泛使用,对比其他传统机器学习方法,是最受欢迎的用于命名实体识别的机器学习方法.在电子病历的命名实体识别任务中也是如此,如燕杨等人[14]针对中文电子病历的命名实体识别问题,提出使用层叠条件随机场且在第 2 层中使用包含实体和词性等特征的特征集,对疾病名称和临床症状 2 类命名实体进行识别,该模型相比于传统层叠 CRF 模型和单层 CRF 模型总体性能有显著提高.

④ 支持向量机.支持向量机(support vector machine, SVM)是较为经典的模式识别方法,其在解决小样本、线性不可分及高维度等模式识别问题中发挥了重大作用,在多个领域成功应用,其中包括电子病历文本挖掘.其主要思想是利用高维特征空间转化使其变为线性可分问题处理,再基于结构风险最小理论构建最优分割超平面,目标是使得学习器得到全局最优化.支持向量机在电子病历文本挖掘中除了文本分类任务也可以被用来完成命名实体识别,例如 Tang 等人[15]研究了结构化支持向量机(structed support vector machine, SSVM)用于临床命名实体识别的方法,该算法结合了 CRFs, SVMs 以及词表征.评价结果表明,当使用相同的特征时,基于 SSVMs 的 NER 系统在临床实体识别

方面的性能优于单纯基于 CRFs 的系统.将 2 种不同类型的单词表征与 SSVMs 相结合,最终系统精度表现最高达到 85.82%.

基于统计学习的方法学习过程不需要太多的人工干预,便于在不同领域之间进行模型的移植,因此大受欢迎,且有不少学者尝试使用多个统计模型来提高医学命名实体识别任务的效果.但是在广泛使用的有监督学习模型实施中,前期大规模标注语料的构建成本高,如何获取高质量、可靠的语料也是主要挑战之一.所以也有不少研究利用半监督学习方法对电子病历进行命名实体识别[16-17].无监督学习方法最典型的就是聚类,在命名实体识别中的主要目标就是通过相似的上下文将内容或格式相似的实体聚在一起.

3) 基于深度学习的方法

近年来,随着深度学习的兴起,为降低人工消耗和训练代价,研究者们也开始将神经网络应用于自然语言处理领域,获得不少成果.在自然语言处理任务中,常用的深度学习模型包括卷积神经网络、循环神经网络、长短期记忆网络、Word2Vec 模型和 2018 年出现的 BERT 模型等.

① 卷积神经网络.卷积神经网络(convolution neural network, CNN)是由卷积层、池化层和全连接层组成,卷积层利用不同的卷积核提取不同的输入特征,池化层是为了降维提取主要特征,全连接层为了结合最后损失函数进行分类.

CNN 不仅在图像处理领域有很好的效果,在 NLP 的诸多任务也可以实现特征抽取等目标,从而提升最终的性能.利用 CNN 对词向量输入进行特征抽取,是 CNN 在 NLP 的一大应用.每一个词向量可以视为一个 1 维的输入,而对于一个由词语构成的序列,它可以作为 2 维的数据(和 2 维图像一样),作为 CNN 的输入.为了保证卷积操作的可解释性,通常过滤器的某个维度会设置成和词向量的维度一样,而在另一个维度上的设置则是考虑上下文语境的长度,并在该维度上进行移动与卷积操作.CNN 中的过滤器具有一定的感受野,考虑了前后语境的影响,这个过程也是 n 元语言模型的一种体现.CNN 在计算上还有一个巨大的优势:它支持并行计算,无论是单个过滤器在不同位置的卷积操作,还是不同过滤器之间,都互不影响,这也意味着在并行计算中 CNN 具有极高的自由度.但是通常来讲,单个的卷积层只能够捕捉到局部短距离的依赖关系(如三元语言模型),想要建立更长距离的语言特征、依赖关系,需要多层的卷积层,但深层网络的参数优化也会相应的更加困难.另外一个 CNN 在 NLP 任务应用的缺点是池化层在一定程度上丢弃了卷积层保留的相对位置关系,在 NLP 中有时候相对位置关系尤为重要,这也导致了一定程度的信息丢失.

CNN 在生物医学命名实体识别任务研究中有大量应用.Gehrmann 等人[18]将卷积神经网络与传统的基于规则的实体提取系统进行了对比和测试.结果显示 CNN 优于其他算法,基于 NLP 的深度学习方法提高了患者表型的性能.如 Wu 等人[19]将 CNN 应用在中文临床记录文本的命名实体识别任务中,他们使用 CNN 对文本进行词向量的预训练,以此提高基准模型的准确率.Crichton 等人[20]将每个单词标记及其周围的上下文单词作为输入,设计了有监督的多任务 CNN 模型,结果表明多任务学习的引入带来了更好的效果,且对小型数据集很有用.Luo 等人[21]同时应用 CNN 和 RNN 对来自 i2b2-VA 挑战数据集的出院摘要中的医学概念之间的语义关系进行分类,并表明具备单词嵌入特征的 CNN 和 RNN 可以在挑战中获得与具有大量特征的系统相似性能.

② 循环神经网络.在 NLP 领域,最常使用的深度学习算法是基于循环神经网络(recurrent neural network, RNN)的深层结构.传统的神经网络无法处理像自然语言这种具有时间序列特性的连续输入,而 RNN 则通过添加指向自身的回路,使得网络能够利用输入的序列特征,因而在处理概念抽取、词性标注等时间序列标注任务时有着先天的优势.RNN 的改进之处在于添加了指向自身的回路,每个神经元的输出除了沿层间连接向上传递之外,还直接传输给了下一个序列.理论上,RNN 可以处理任意变长的序列,然而,随着时间序列的不断累积,梯度会出现指数级衰减的现象,这导致 RNN 难以记录距离较远的历史信息,其性能也因此而受到制约.为了解决这个问题,1997 年 Hochreiter 等人[22]第一次提出长短期记忆网络(long short-term memory, LSTM)概念,并从理论上证明了这种结构能够很好地解决梯度消失和爆炸问题.

③ LSTM. LSTM 也是一种时间递归神经网络.在 LSTM 算法中加入了判断信息有用与否的输入门、遗忘门和输出门,LSTM 是解决长距离依赖问题的有效技术.

另外,在有些 NLP 任务中,某个时间的输出不仅和过去信息有关,也取决于它的未来信息,例如在命名实体识别任务中,一个词语是否为命名实体,由其上下文共同决定.因此,为了同时考虑过去和后续的信息对当前时刻的影响,我们可以在原有的 LSTM 中增加一个反向的信息流,来传递后续时刻的信息.Schuster 等人[23]基于 LSTM 提出了双向长短期记忆网络(bi-directional LSTM, Bi-LSTM)概念,这种方法不仅从前到后对序列建模,而且从后到前也对序列建模,所以每一个时刻的状态不仅包含前面的信息,而且囊括了后面的信息.

在过去几年中,使用 LSTM 和 CRF 结合的模型,BioNER 的性能得到了很大改善.LSTM 是解决传统 RNN 中梯度消失问题的一种方法,而双层循环神经网络 Bi-LSTM 改进了 LSTM,使得在做命名实体识别时可以既利用正向序列信息同时利用反向序列信息.之后又在 Bi-LSTM 后加入了 CRF 层,Bi-LSTM 可以充分利用字词信息和位置信息得到特征,将其隐层输出输入到 CRF 层中来做标签的预测.Habibi 等人[24]将通过学习一个实体注释的金标准语料库(gold standard corpora, GSC),结合预先学习词嵌入(word embedding)的大型语料库(大量来自 PubMed 的摘要)得到特征,并作为 BiLSTM-CRF 模型的输入.在包括 5 种不同的实体类型的不同标准语料库进行了准确率评估.平均而言,它比基于词典的 NER 工具提升 5%,比单独使用 CRF 方法提升 3%.Wang 等人[25]为 BioNER 提出了一个使用字符级的多任务神经网络的学习框架.该框架考虑了具有依赖于上下文的 Bi-LSTM 层的 BiLSTM-CRF 模型.通过重用相应 Bi-LSTM 单元中的参数,来自不同数据集的输入可以有效地共享字符和单词级表征.文献[25]的作者将提出的多任务模型与多个 BioNER 系统和基线神经网络模型在 15 个基准 BioNER 数据集上进行比较,并观察到更好的性能.Gorinski 等人[26]对比了基于规则、运用深度学习和迁移学习 3 种不同的系统在对脑卒中患者的脑成像报告中的命名实体识别任务的应用效果.实验表明基于规则的系统因为有相关领域专家提供的规则效果最精确,而运用 BiLSTM-CRF 模型的系统减少了设计新规则时对专家知识的需求,学习效率更高,迁移学习虽然仍然需要大量人工操作但是表现很好有代替基于规则的系统的可能.但是无论是 RNN 还是 CNN,在处理 NLP 任务时都有缺陷.CNN 是其卷积操作不适用于序列化的文本,RNN 的缺点则是其不能实现数据处理并行化,这会导致对内存的要求过高.在中文电子病历的命名实体识别研究中,张聪品等人[27]构建了 LSTM-CRF 模型对电子病历进行命名实体识别,准确率达到了 96.29%.

在深度学习中不得不提到的是 Word2Vec 模型[28]和 Bert 模型,它们采用分布式表示方法将自然文本转换到词向量,之前提到的特征都是基于词向量挖掘而被具体的神经网络学习最终服务于各种任务,包括命名实体识别、关系抽取任务等.

④ Word2Vec 模型.传统的自然语言词语处理方法将词语看作一个符号,被称作 one-hot 表示,这种方法导致词与词之间的关系被独立开,当词表过大时,向量维度也随着变大,Word2Vec 的提出解决了这一问题.Google 在 2013 年提出了一种新的用于计算词向量的方法 Word2Vec[28],Word2Vec 模型是一种快速训练词向量模型的方法.使用 Word2Vec 模型的目的在于从大量的文档医学文本数据中训练出高质量的词向量,Word2Vec 被用来解决大规模语料的词向量表达问题,在 Word2Vec 出现之前,很多的 NLP 技术都是采用 one-hot 结构,这样构建出的词向量忽略了词语之间的相似性和关联性,Word2Vec 则提出了词向量的分布式表示方法,利用浅层神经网络,在大规模无标注语料库上训练低维稠密的词向量.Word2Vec 框架提出基于分布式词表示的思想,即要理解一个词语的意思只需要通过了解词语出现的上下文即可.由此提出了 2 种用于训练的模型,一种为连续词袋模型(continuous bag of words, CBOW)模型,其基本思想是给定上下文来预测其中心词;另一种为连续跳跃模型(continuous skip-gram model, Skip-Gram),是利用给定的中心词来预测上下文进行训练.在生物医学文本挖掘任务中,由于生物医学语料库与一般领域语料库在词汇和表达方面存在较大差异,在应用于生物医学数据时需要在医学文本语料上进行训练.如 Zhu 等人[29]在包含临床报告和临床领域相关的 Wikipedia 页面的语料库上训练上下文单词嵌入模型,然后训练 BiLSTM-CRF 模型.

⑤ BERT 模型.以 Word2Vec 为代表的词向量模型有一个明显的缺陷,即训练好的词向量是固定的,单词不会因为处在不同的语境而改变,这对于一词多义的情况是十分不利的.因此在 Word2Vec 出现之后,涌现了许多利用语言模型建立基于上下文的词向量方法,如 ELmo[30],OpenAI[31],BERT[32],其中的模型 BERT 是最具影响力的方法.BERT 模型

是基于 Transformer 提取特征,并采用双向语言模型.其训练方式区别于传统的从左到右的训练方式,而采用 2 种新的方法进行大规模无监督训练,2 种方法分别是 Masked LM 和 Next Sentence Prediction. Masked LM 方法是在给定一句话并随机抹去这句话中的 1 个或几个词语,然后利用 BERT 模型去预测这几个被抹去的词语.Next Sentence Prediction 是给定一句话,判断下一句话是否与给定的句子属于同一个上下文中.其在文本分类任务、语句对分类任务和 NER 中均取得 SOTA(state-of-the-art)的效果.BERT 采用多层双向 Transformer 编码器,可以学习生成考虑语境的语言模型,并可以在调整后针对各种任务(如命名实体识别和关系抽取).Mao 等人[33]在系统 Hadoken 中利用 BERT 模型对训练数据进行预训练,然后将临床病例的表征提供给 CRF 输出层以进行分类,并且发现其适用于多语言命名实体识别任务.

深度学习的模型通常需要大量的标注训练数据,然而在电子病历文本挖掘任务中构建大型训练集需要专业人士的知识,成本非常高.因此用于生物医学文本挖掘任务的训练数据较少,大多数生物医学文本挖掘模型无法充分利用深度学习的能力.为了解决训练数据的缺乏,近年来有研究集中在训练多任务模型或是借助迁移学习的力量.如 Yoon 等人[34]提出使用多个 NER 模型(在文中指的是一组 BiLSTM-CRF 模型)组合的 CollaboNet.在 CollaboNet 中,在不同数据集上训练的模型彼此连接,成功地减少了错误分类实体的数量并提高了性能.另一方面,在深度学习中应用迁移学习的思想,用在其他数据集训练好的模型初始化部分甚至所有的神经网络中的参数,为用目标数据集训练做准备.Lee 等人介绍了针对生物医学文本挖掘任务的语境化语言表示模型 BioBERT[35],其在原结构 BERT 的基础上重新训练.他们使用迁移学习来解决缺乏训练数据的问题,即使用一般语料库和医学领域语料库对 BioBERT 进行预训练.实验证明,BioBERT 能有效地将大量其他语料库文本的知识迁移到医学文本挖掘任务中,只需要针对特定任务的体系结构进行少量修改.BioBERT 在 3 个具有代表性的生物医学文本挖掘任务、生物医学命名实体识别、生物医学关系提取和生物医学问答系统上明显优于其他模型.

2.2 关系抽取

关系抽取(relation extraction, RE)通常基于命名实体识别的结果之上,也是自然语言处理中一个重要的子任务,但是由于深度学习的发展,也有不少深度神经网络将命名实体识别和实体关系抽取看作一个完整的任务.理论上,关系抽取任务分为 2 步,首先判断一个实体对是否存在关系,若有关系,则进一步判断属于哪种关系.在实际模型设计中,通常把无关系当作一种特殊的关系,直接将关系抽取看作是多类别分类任务.关系抽取是医疗健康知识库建立维护的基础.在医学领域,不同实体间的关系有不同的定义标准,根据 I2B2 2010 评估会议[36],EMR 中的实体关系可以分为 3 类,包括疾病之间的关系、疾病与医学检查的关系以及疾病与治疗之间的关系.在医学领域,常常采用基于共生[37]、传统机器学习和深度学习方法来进行关系抽取.基于共生的方法是基于 2 个实体同时出现的频率越高关系越强这一假设的统计方法.最广泛使用的方法还是传统机器学习方法和深度学习方法,例如 Bhasuran 等人[38]采用有监督的机器学习方法,即使用深度集成支持向量机来训练,利用语法和语义属性的特征集并结合词嵌入,从 4 个标准语料库中提取基因-疾病关系.实验显示在 EUADR,GAD,CoMAGC,PolySearch 四个语料库处理结果的 F-measure 分别达到 85.34%,83.93%,87.39%,85.57%.

同时,将机器学习与深度学习相结合,也可以提高电子病历中关系抽取性能.如张玉坤等人[39]将 CNN,SVM,CRF 三者结合,然后通过联合学习方式来对医疗文本进行关系抽取,取得了不错的效果.自注意力机制也是医疗关系抽取中常用的方法,宁尚明等人[40]针对文本特征的每个通道都计算注意力权重,实现了电子病历实体关系的抽取,在 2010 i2b2 和 SemEval 2013 DDI 中 $F1$ 值分别达到 69.72% 和 72.32%.

时序性是电子病历数据的一大特点,因为电子病历通常不仅包括患者当前的治疗状况和指标,还包括患者过去所经历的临床事件.为了自动构建这些事件之间的时间线,需要抽取临床记录中事件和时间的关系.自动检测并抽取患者记录中的时间和事件之间的关系能帮助医务人员了解疾病进展,如 Tian 等人[41]提出了一种基于深度学习的汉语电子病历时间信息提取框架.他们运用 LSTM-CRF 模型提取临床实体如疾病、治疗等以及时间等相关实体,之后用 CNN 进行时序关系分类.

另外,从临床记录中提取药物相关实体与药物之间的关系,可以帮助患者避免药物引发的不利效果,比如药物的不良事件(adverse drug events, ADE)

中的实体关系能反映某些药物引起的副作用,以及副作用对患者产生的影响.电子病历是挖掘 ADE 的丰富信息来源,学者们已经广泛应用 NLP 技术以使用关系抽取方法挖掘在电子病历数据中的 ADE 相关信息.Christopoulou 等人[42]提出了一种集成学习方法,其运用神经网络模型中的 Bi-LSTM 注意力机制和 Transformer 神经网络来提取药物和药物相关实体之间的关系.

2.3 文本分类

文本分类是文档级的自然语言处理任务,目标是给文档打上预定义的文档级标签.电子病历中的医学报告是一种具有丰富信息的资源,特别是主要用自然语言写的自由文本部分,这部分包括医生的临床推理信息及其思维过程,其能通过提供详细的病人情况来帮助解决不同的临床问题,而且它们通常不能被其他结构化数据所代替.文档分类有助于处理和提取这类数据,辅助后续的分类预测任务,并提高叙述性临床笔记的利用效率,因此文档分类成为临床预测分析的一个重要研究领域.早期的文献中,自动分类研究需要利用手工实现且由专家制定规则的知识工程,这非常耗时耗力,且对人力资源有较高的要求.最近,机器学习算法和自然语言处理技术已经被用来处理电子病历以支持临床决策,目前效果较好的文档分类方法之一是基于 Word2Vec 的神经网络模型.深度神经网络可以通过算法本身学习复杂的数据特征,而不需要手工设计的特征提取规则.深度学习在医疗文本分类中通常使用的方法是先利用网络将医疗文本以字为单位应映射到向量空间,然后再利用 SVM 等模型对向量空间进行分类[43],例如 Hughes 等人[44]将基于单词分布式表示的 CNN 应用于句子级的医学文本分类任务中,取得了很好的性能.

文本分类可以被用于分类临床记录,以此来辅助识别患者所患疾病,例如 Turner 等人[45]评估了多种传统分类器(包括神经网络、随机森林、朴素贝叶斯、支持向量机等)在系统性红斑狼疮患者识别中的性能,其中具有统一医学语言系统(unified medical language system, UMLS)概念唯一标识符(CUIs)的浅神经网络和同时具有 CUIs 和词袋模型(bag-of-words, BoW)的随机森林表现最优.Topaz 等人[46]针对从临床记录中识别糖尿病患者这一任务建立了一个基于 NLP 的分类系统 Nimble-Miner,运用了包括 SVM,RNN 等多种机器学习方法.Doing-Harris 等人[47]使用聚类算法,基于词汇和词性的数据表征,通过无监督学习区分不同文档类型和信息来源,并取得了良好的性能.Kocbek 等人[48]使用基于 bag-of-phrases 的 SVM 检测各种疾病类别,对潜在疾病的人入院进行分类.

也有学者从临床记录中的句嵌入方法入手临床文本分类问题,与词嵌入相似,句嵌入是一种高维向量,它可以表示单词序列的特征.句子嵌入的使用通常是考虑到在训练数据较少的情况下,依赖许多词向量的 RNN 捕获诊断信息来分类文档比较困难,而少量的句子向量就可以获得丰富的语义信息.如 Ormerod 等人[49]通过 LSTM 构建电子病历分类模型,并同时显示文档中哪些句子对患者的病情诊断最有帮助.另外,中文病历的文本分类与英文电子病历的文本分类有所不同,因为中文语言有其特殊性:断句方式不同、需要新的分词工具等,且中文电子病历既有短文本又有长文本,因为在借鉴国外优秀成果时需要结合中文特点作出调整.杜宝琛[50]在设计基于电子病历的辅助治诊断系统时同时考虑长短文本,采用了双通道下不同神经网络同时学习.吕愿愿等人[51]在对电子病历进行自然语言处理后,利用 TF-IDF 和潜在语义分析(latent semantic analysis, LSA)方法提取特征,在挖掘出依存关系后对病历短文本进行分类.

2.4 智能问诊

问答系统(question answering system, QA)是自然语言处理中的传统任务,相比于日常的机器问答,由于在医学领域有更多的专业名词,实现问答则更为困难,传统的问答系统通常需要特征工程、语言工具或外部资源的帮助.通常是使用语言模型提取语义特征,利用决策树等模型[52-53]来识别问题的答案.尽管这些方法具有一定的有效性,但它们需要额外的资源和特征工程并使用语言工具,模型的复杂性较高.

目前,问答研究主要集中在利用深度学习技术自动提取句子特征,且多在通用数据集上进行测试.例如 Wang 等人[54]使用 LSTM 框架将答案选择任务转化为分类和排序问题.Xiong 等人[55]使用动态协作注意网络将斯坦福问答数据集上的 $F1$ 值增加到 80.4%.此外,Tan 等人[56]提出了一个基于注意力机制的 RNN 模型,将问题注意引入到答案表征中,建立了基于 BiLSTM 模型的匹配问答对,并利用余弦相似度计算了它们的接近度.Dong 等人[57]提出了一种改进的多列卷积神经网络,从响应路径、上下文和答案类型 3 个方面学习问题和答案的分布

式表示.Santos 等人[58]提出了一种基于特征加权的双向注意力机制,通过特征工程、注意机制等提高问答匹配的准确性.

使用大型带注释的数据集构建的几个开放领域的机器理解系统使自动问答取得了长足的进步.然而在临床领域,自动问答仍然在探索阶段.由于缺乏大规模的临床标注数据集,目前还没有一个通用的系统来回答医生在病人的电子病历上提出的自然语言问题.医生们通常希望根据电子病历中找到有关医疗实体和关系的问题答案,这需要计算机对临床笔记有更深的理解.电子病历中数据的特性包括大量的非结构化数据、大量的专业术语、多个疾病之间有时序性和拼写错误等,而这些都是机器理解电子病历时的难点,现有的 NLP 工具难以应付这种复杂情况.此外,在挖掘答案时也有难点,因为答案可能是隐式的,而且可能需要多个临床领域的知识和推理.由于这些挑战的存在,为患者构建可信的 QA 系统变得十分困难,同时 QA 系统一般需要大规模的问答注释.然而构建数据集涉及到个人健康信息的隐私问题以及大量人力资源,手工构建大型注释数据集不切实际[59].

QA 数据集主要分为两大类:使用非结构化文档的机器理解(machine comprehension, MC)数据集和使用知识库的问题-答案对数据集.MC 系统旨在回答任何针对参考文本提出的问题.最近在云资源和搜索引擎方面的进展导致了 MC 数据集的爆炸式增长,但其中有价值的数据比例却较少.另一方面,特定领域 MC 数据集如 MCTest,BioASQ,InsuranceQA 等对专家注释的需求高,同时也涉及隐私问题,这使其在规模上受到了限制(500～10 000).Pampari 等人[60]利用 i2b2 数据集中针对各种 NLP 任务的临床笔记上的现有专家注释,为电子医疗记录生成大型问题-答案对数据集.得到的语料库有 100 万个问题形式和 40 多万个问题-答案对,在问题-答案的关系抽取时用了带注意力层的端对端模型.Roberts 等人[61]通过在 468 个电子病历问题上手工注释标签,生成了语料库.随着医学 QA 系统的发展,学者们也结合了传统的方法和深度神经网络方法来构建混合模型.这些模型结合了神经网络模型的精确性和传统方法中符号表示的可解释性.

总而言之,医疗问答系统方面还处于探索和研究阶段,没有能够切实有效的落地应用,但其未来的潜力巨大,是一个非常有前景的研究方向.

3 电子病历文本挖掘在糖尿病和心脑血管疾病中的应用

除了常见的命名实体识别、关系抽取、文本分类和医疗问答等基本任务外,对于电子病历文本挖掘的应用广泛存在于不同的领域,且在不同的疾病中往往有不同的表现形式.

3.1 糖尿病

国际糖尿病联合会 2017 年修订的第 8 版本数据显示,全球有 4.25 亿糖尿病患者.这意味着每 11 个成年人里就有 1 个糖尿病患者,而中国有超过 1 亿人患有糖尿病,所以不管是对个体患者还是在整个人类范围内,管理糖尿病都是非常重要的[62].糖尿病是一种慢性疾病,健康的胰腺分别通过 α-细胞和 β 细胞动态控制胰岛素和胰高血糖素激素的释放,以维持正常血糖[63],而糖尿病特征是患者体内不存在葡萄糖稳态.糖尿病可以分为多种,当身体的免疫系统攻击产生胰岛素的细胞并完全停止产生胰岛素时,就会导致Ⅰ型糖尿病;当身体不能产生足够的胰岛素或细胞产生胰岛素抵抗时,会导致Ⅱ型糖尿病,Ⅱ型糖尿病可能是遗传、饮食不良、缺乏运动或肥胖的结果;另外还有妊娠期糖尿病(妊娠中期或晚期确诊糖尿病且在妊娠前没有糖尿病症状)和由于其他原因引起的特定类型的糖尿病,例如单基因糖尿病综合征、外分泌胰腺疾病和药物或化学诱导的糖尿病[64].糖尿病护理在很大程度上取决于患者的日常自我管理,包括吃什么以及何时运动,以及确定部分患者需要的胰岛素剂量和时间.在这种情况下,每个患者每天都产生与糖尿病相关的大量数据,这些数据来源包括电子病历、胰岛素泵、传感器、血糖仪和其他可穿戴设备,还包括实验中糖尿病相关的基因组学、蛋白质组学、代谢组学和微生物学数据[65].所以在糖尿病相关的研究中有许多数据挖掘的应用.最早在 2002 年 Breault 等人[66]应用 CART 分析方法对糖尿病数据库进行了分析,虽然准确率仅为 59%,但是首次验证了数据挖掘技术在糖尿病问题领域的应用前景.随着标准化的电子病历系统在中国的兴起,糖尿病相关的诊断和风险管理等也接受到电子病历文本挖掘的辅助,本节将从糖尿病的诊断角度阐述机器学习在糖尿病上的应用.

糖尿病的诊断需要通过包括 α-糖酸盐血红蛋白(A1C)实验、随机血糖实验、空腹糖试验或口服葡萄糖耐量实验在内的几项实验.无论是Ⅰ型还是Ⅱ型

糖尿病,早期诊断和预测对于延缓疾病发展,有针对性地选择药物,延长患者预期寿命,减轻症状和相关并发症的发作都至关重要.生物标志物(例如生物分子)是代表健康和疾病状态的特定病症的可测量指标,通常在体液(血液、唾液或尿液)中测量.在研究糖尿病的情况下,生物标志物可以反映患者是否存在高血糖及其严重程度,或是否存在糖尿病相关并发症及其严重程度.而机器学习方法中的特征选择可以帮助挖掘出新的生物标志物,辅助糖尿病的确诊,且在特征选择步骤之后,分类算法可以被用来评估所选特征的预测准确度.例如 Jelinek 等人[67]研究了在糖化血红蛋白(glycated hemoglobin, HbA1c)水平低于或等于6.5%的情况下,找到的 2 种生物标志物与 HbA1c 一起参与检测,提高了糖尿病的诊断准确性.也有学者利用特征提取的算法选取预测糖尿病的特征,如 Bagherzadeh-Khiabani 等人[68]使用了 803 名有 55 个特征的糖尿病前期女性的临床数据集,比较了 19 种常用的特征选择算法来预测糖尿病.Sideris 等人[69]提出了一种基于聚类的特征提取框架,使用疾病诊断信息产生的特征群,并用作预测患者病情严重程度和患者再入院风险.

许多利用电子病历的机器学习方法和框架被运用到Ⅱ型糖尿病的早期诊断上[70-71].集成学习方法和关联规则学习也被大量运用到糖尿病的诊断中.如 Tapak 等人[72]比较了 5 种机器学习模型 ANN, SVM, FCM(fuzzy k-means)、随机森林(random forst, RF), LDA(linear discriminant analysis)来分类是否患有糖尿病的个体.集成学习也被逐渐应用到诊断糖尿病的分类系统中[73].Han 等人[74]提出了一种基于 SVM 和 RF 的规则提取集成学习方法.另外,通过挖掘一些与糖尿病相关的属性也可以预测糖尿病风险,提醒体检的人注意某些习惯预防糖尿病的发生.

深度学习方法也对糖尿病的诊断研究作出了贡献,尤其是在对电子病历非结构化数据的处理分析中,例如可以有效地识别病历中未明确指出的糖尿病病例,从而显著改善糖尿病病例发现现状.EMR 的非结构化数据存在于临床记录、手术记录、出院记录、放射学报告和病理报告中.其中临床记录包含的信息有患者的病史(疾病和治疗措施等)、疾病家族史、环境和生活方式等,因此提供了很多可供研究的细节信息[75].Zheng 等人[76]针对已有算法无法大量识别糖尿病电子病历案例中非结构化数据的问题,使用 RF 的方法实现更完整的糖尿病诊断.Pham 等人[77]针对个性化医疗中的预测患者疾病和护理过程建模问题,考虑了包括时序性等几项特性,提出了一种端到端的深层动态神经网络.其基于 LSTM,引入了处理不规则且有时序性事件的方法,还模拟医疗干预措施改变病程,根据历史和当前健康状态来估计未来结果.最近,Liu 等人[78]提出了一个多任务学习框架来预测包括糖尿病在内的慢性疾病的发病,并比较了不同深度学习架构(包括 CNN 和 LSTM)的性能.

糖尿病作为影响人类健康的常见疾病,长久以来一直损害着社会的经济,用自动化、低成本的方式来管理糖尿病的患者,辅助医疗,将创造巨大的社会效益.

3.2 心脑血管疾病

心脑血管疾病是心脏血管和脑血管疾病的统称,泛指由于高脂血症、血液黏稠、动脉粥样硬化、高血压等所导致的心脏、大脑及全身组织发生的缺血性或出血性疾病.其中心血管疾病(cardiovascular disease, CVD)是全球众多致死疾病之一,因其死亡人数占全球死亡人数的 1/3[79].2 种疾病都严重威胁人类,特别是 50 岁以上中老年人的健康,且有幸存者生活不能完全自理的可能性,或者有严重的并发症,例如心力衰竭(heart failure, HF).但是电子病历文本挖掘可以在多个角度辅助患者的治疗和风险管理等.例如通过电子病历计算患者 HF 存活风险评分,识别高风险患者并应用个体化治疗和健康生活指导将降低其死亡风险[80],且可以在出院时确定再入院风险的患者.另外,Li Bin 等人[81]在心血管疾病的许多严重的预后疾病如急性心肌梗死、肺栓塞、严重的脑神经系统疾病等研究中发现风险预警模型可以探讨其风险因素,筛选出与危重疾病预后相关的严重疾病(中风、心力衰竭、肾功能衰竭).本节主要就心脑血管疾病的预测讨论数据挖掘技术辅助心血管疾病治疗的作用.

对于心脑血管疾病的预测,在医疗领域,建立可预测患者疾病的模型可以提高医院的治疗效果和效率.传统的对于心血管的治疗预测等都是通过评分来辅助决策,例如美国心脏病学会(American College of Cardiology, ACC)/美国心脏协会(American Heart Association, AHA)提出基于风险因素的组合 Framingham 风险评分,包括高血压、糖尿病、胆固醇和吸烟状况等这些常规因素的预测模型预测心血管疾病.然而,随着电子病历系统的迅速普及,患者的大部分数据都以电子格式存储.确定疾病所需

的风险因素数据存在于电子病历中,包括结构化数据如心电图、血管造影等和临床记录等非结构化数据.然而,通常为了利用结构化数据,需要大量的人力物力资源来对数据进行筛选和清洗,同时从非结构化电子病历数据中手动提取所需成本也十分昂贵[82].且心血管疾病本质上是复杂的,由多种遗传、环境(例如空气污染)和行为因素(例如饮食)引起的,需要更有效的工具来准确地预测结果,而不是依靠简单的评分系统.在数据挖掘领域,人工智能技术(如机器学习)正在彻底改变医生制定临床决策和诊断的方式,并提高心脑血管疾病风险自动化预测的水平.将医学信息技术与机器学习技术相结合,使用疾病相关数据生成的预测模型,可以提高预测准确性.其中,有监督学习算法已成功应用于心脑血管疾病的预测.Kim 等人[83]使用与心血管疾病相关的健康数据进行统计分析,找出与心血管疾病相关的变量,并建立了基于深度信念网络(deep belief nets, DBN)的心血管风险预测模型.但是有监督学习也有一定的缺点,首先其需要大型数据集来训练模型并通过其他数据集进行验证.通常还需要手动标记训练数据集,比如苏嘉等人[84]针对中文电子病历特点构建的心血管疾病风险因素的标注语料库,以预测死亡率和再入院率等.此外,即使模型能在给定的训练数据集和测试集上表现良好,但是它可能由于训练数据与真实数据的差异和过拟合情况而导致偏差.针对这些问题,也有不少无监督学习算法应用到心血管疾病预测模型中,在最近的趋势下无监督深度学习在这一领域表现较好.其次,深度学习可用于分类来自异质 CVD 的新基因型和表型,例如肺动脉高压和心肌病等.另外,深度学习预测模型可以通过高血压、肾功能异常、肝功能异常、年龄、药物治疗和酒精摄入等因子之间的加权来预测出血和中风的风险评分,以确定患者的最佳剂量和抗凝治疗持续时间[85].最后,通过深度学习,可以从心电图模式或超声心动图预测冠状动脉钙化评分.事实证明,深度学习比其他机器学习技术(如 SVM)更好.但是深度学习也有缺点,比如其通常是非线性分析,有很多参数和多层,因此可能导致过度拟合而预测性能不佳.而且,深度学习还需要大量的训练数据集,这需要各机构之间的协作,对计算机硬件的要求也较高.

脑血管疾病主要的表现就是脑卒中,也叫中风,主要分为缺血性脑卒中和出血性脑卒中.中风的预测从简单的到复杂的模型各不相同.脑卒中的风险因素是复杂的,可以从直接和间接 2 方面找到不同程度的因素.Leira 等人[86]采用逐步回归法对数据库中选择的 1 266 例患有缺血性脑卒中患者和复发脑卒中患者的医疗记录进行分析并选择 20 个临床变量进行评估.Goyal[87]利用 ICD-10 编码(包含疾病特征和分类)和脑卒中患者的电子病历数据进行分析,最终利用 LSTM 建立脑卒中的预测模型.除此之外大多数的数据挖掘模型都结合电子病历中的医学图像辅助脑血管疾病的预测.

简而言之,对患有心脑血管疾病患者的电子病历进行数据挖掘,可以从病前、病中、病后 3 个阶段进行有效的预测,从而来配合医生和患者做出更好的决策.

4 总结与展望

在医疗领域中,文本电子病历是医疗单位对患者临床诊疗的数字化相关信息载体.电子病历数据中的知识对于临床决策和医药研发等都有很强的指导意义,其非结构化特征导致很难利用计算机直接进行批量分析.故将人工智能技术和大数据数据挖掘的手段应用在电子病历中是大势所趋,但是由于电子病历数据的特性,机器学习方法的应用也有特定的挑战和难点,其吸引了国内外广大学者的研究.本综述针对电子病历数据挖掘,尤其是其中的非结构数据挖掘的主要分析流程和方法进行了梳理,简要介绍了传统机器学习和深度学习常见网络结构,综述其在电子病历等方面的最新研究进展,并且探讨了在糖尿病和心脑血管疾病这样特定疾病中的应用现状和前景,为后续文本数据挖掘的研究应用提供参考.

参 考 文 献

[1] Mao Lifeng, Qu Haibin. A new computer-aided method for diagnosis of breast cancer based on decision tree [J]. Journal of Southern Yangtze University, 2004, 3(3): 227-229 (in Chinese)
(毛利锋, 瞿海斌. 一种基于决策树的乳腺癌计算机辅助诊断新方法[J]. 江南大学学报, 2004, 3(3): 227-229)

[2] Belle A, Thiagarajan R, Soroushmehr S, et al. Big data analytics in healthcare [J]. BioMed Research International, 2015, 2015: No.370194

[3] Yang Feihong, Zhang Yu, Qin Lu, et al. A research progress in named entity recognition of Chinese EMR [J]. China Digital Medicine, 2020, 15(2): 9-12 (in Chinese)

(杨飞洪, 张宇, 覃露, 等. 中文电子病历的命名实体识别研究进展[J]. 中国数字医学, 2020, 15(2): 9-12)

[4] Qu Chunyan, Guan Yi, Yang Jinfeng, et al. Construction of Chinese EMR named entity annotation corpus [J]. Chinese High Technology Letters, 2015, 25(2): 143-150 (in Chinese)
(曲春燕, 关毅, 杨锦锋, 等. 中文电子病历命名实体标注语料库构建[J]. 高技术通讯, 2015, 25(2): 143-150)

[5] Yang Jinfeng, Guan Yi, He Bin, et al. Corpus construction for named entities and entity relationships on Chinese electronic medical records [J]. Journal of Software, 2016, 27(11): 2725-2746 (in Chinese)
(杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报, 2016, 27(11): 2725-2746)

[6] Banda J M, Halpern Y, Sontag D, et al. Electronic phenotyping with APHRODITE and the observational health sciences and informatics (OHDSI) data network [J]. AMIA Summits on Translational Science Proc, 2017, 2017: 48-57

[7] Zhai Juye, Chen Chunyan, Zhang Yu, et al. A study on named entity recognition of Chinese electronic medical records based on the combination of CRF and rules [J]. Journal of Baotou Medical College, 2017, 33(11): 124-125, 130 (in Chinese)
(翟菊叶, 陈春燕, 张钰, 等. 基于CRF与规则相结合的中文电子病历命名实体识别研究[J]. 包头医学院学报, 2017, 33(11): 124-125, 130)

[8] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name [J]. Machine Learning, 1999, 34(1/2/3): 211-231

[9] Kupiec J. Robust part-of-speech tagging using a hidden Markov model [J]. Computer Speech & Language, 1992, 6(3): 225-242

[10] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 39-71

[11] Jaynes E T. Information theory and statistical mechanics [J]. Physical Review, 1957, 106(4): 620-630

[12] Lafferty J D, Mccallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] //Proc of the 18th Int Conf on Machine Learning. New York: ACM, 2001: 282-289

[13] Wallach H M. Conditional random fields: An introduction [J]. Technical Reports, 2004, 53(2): 267-272

[14] Yan Yang, Wen Dunwei, Wang Yunji, et al. Chinese medical record named entity recognition based on stacked conditional random fields [J]. Journal of Jilin University: Engineering Edition, 2014, 44(6): 1843-1848 (in Chinese)
(燕杨, 文敦伟, 王云吉, 等. 基于层叠条件随机场的中文病历命名实体识别[J]. 吉林大学学报:工学版, 2014, 44(6): 1843-1848)

[15] Tang Buzhou, Cao Hongxin, Wu Yonghui, et al. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features [J]. BMC Medical Informatics and Decision Making, 2013, 13(Suppl 1): S1

[16] Alicante A, Corazza A, Isgrò F, et al. Unsupervised entity and relation extraction from clinical records in Italian [J]. Computers in Biology and Medicine, 2016, 72: 263-275

[17] Zhang Shaodian, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts [J]. Journal of Biomedical Informatics, 2013, 46(6): 1088-1098

[18] Gehrmann S, Dernoncourt F, Li Yeran, et al. Comparing rule-based and deep learning models for patient phenotyping [J]. arXiv preprint, arXiv:1703.08705, 2017

[19] Wu Yonghui, Jiang Min, Lei Jianbo, et al. Named entity recognition in Chinese clinical text using deep neural network [J]. Studies in Health Technology and Informatics, 2015, 216: 624-628

[20] Crichton G, Pyysalo S, Chiu B, et al. A neural network multi-task learning approach to biomedical named entity recognition [J]. BMC Bioinformatics, 2017, 18(1): 368-368

[21] Luo Ling, Li Nan, Li Shuaichi, et al. DUTIR at the CCKS-2018 Task1: A neural network ensemble approach for Chinese clinical named entity recognition [C] //Proc of the CCKS Tasks. Berlin: Springer, 2018: 7-12

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

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

[24] Habibi M, Weber L, Neves M, et al. Deep learning with word embeddings improves biomedical named entity recognition [J]. Bioinformatics, 2017, 33(14): i37-i48

[25] Wang Xuan, Zhang Yu, Ren Xiang, et al. Cross-type biomedical named entity recognition with deep multi-task learning [J]. Bioinformatics, 2019, 35(10): 1745-1752

[26] Gorinski P J, Wu Honghan, Grover C, et al. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches [J]. arXiv preprint, arXiv:1903.03985, 2019

[27] Zhang Congpin, Fang Tao, Liu Yuliang. Research and application of named entity recognition technology based on LSTM-CRF [J]. Computer Technology and Development, 2019, 29(2): 106-108, 142 (in Chinese)
(张聪品, 方滔, 刘昱良. 基于LSTM-CRF命名实体识别技术的研究与应用[J]. 计算机技术与发展, 2019, 29(2): 106-108, 142)

[28] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint, arXiv:1301.3781, 2013

[29] Zhu Henghui, Paschalidis I C, Tahmasebi A. Clinical concept extraction with contextual word embedding [J]. arXiv preprint, arXiv:1810.10566, 2018

[30] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [J]. arXiv preprint, arXiv:1802.05365, 2018

[31] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [EB/OL]. 2018 [2020-05-11]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[32] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv:1810.04805, 2018

[33] Mao Jihang, Liu Wanli. Hadoken: A BERT-CRF model for medical document anonymization [C] //Proc of the Iberian Languages Evaluation Forum co-located with 35th Conf of the Spanish Society for Natural Language Processing. Aachen: CEUR Workshop Proceedings, 2019: 720-726

[34] Yoon W, So C H, Lee J, et al. CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition [J]. BMC Bioinformatics, 2019, 20(10): 249-249

[35] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240

[36] Uzuner Ö, South B R, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556

[37] Jelier R, Jenster G, Dorssers L C, et al. Co-occurrence based meta-analysis of scientific texts: Retrieving biological relationships between genes [J]. Bioinformatics, 2005, 21(9): 2049-2058

[38] Bhasuran B, Natarajan J. Automatic extraction of gene-disease associations from literature using joint ensemble learning [J]. PloS One, 2018, 13(7): No.e0200699

[39] Zhang Yukun, Liu Maofu, Hu Huijun. Chinese medical entity classification and relationship extraction based on joint neural network model [J]. Computer Engineering and Science, 2019, 41(6): 1110-1118 (in Chinese)
(张玉坤, 刘茂福, 胡慧君. 基于联合神经网络模型的中文医疗实体分类与关系抽取[J]. 计算机工程与科学, 2019, 41(6): 1110-1118)

[40] Ning Shangming, Teng Fei, Li Tianrui. Entity relationship extraction of electronic medical records based on multi-channel self-attention mechanism [J]. Chinese Journal of Computers, 2020, 43(5): 916-929 (in Chinese)
(宁尚明, 滕飞, 李天瑞. 基于多通道自注意力机制的电子病历实体关系抽取[J]. 计算机学报, 2020, 43(5): 916-929)

[41] Tian Bing, Xing Chunxiao. Deep learning based temporal information extraction framework on Chinese electronic health records [C] //Proc of Int Conf on Web Information Systems and Applications. Berlin: Springer, 2018: 203-214

[42] Christopoulou F, Tran T T, Sahu S K, et al. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods [J]. Journal of the American Medical Informatics Association, 2020, 27(1): 39-46

[43] Zhou Yang. Research on medical text analysis and mining technology based on machine learning [D]. Beijing: Beijing Jiaotong University, 2019 (in Chinese)
(周阳. 基于机器学习的医疗文本分析挖掘技术研究[D]. 北京: 北京交通大学, 2019)

[44] Hughes M, Li I, Kotoulas S, et al. Medical text classification using convolutional neural networks [J]. Studies in Health Technology & Informatics, 2017, 235: 246-250

[45] Turner C A, Jacobs A D, Marques C K, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus [J]. BMC Medical Informatics and Decision Making, 2017, 17(1): 126-126

[46] Topaz M, Murga L, Bar-Bachar O, et al. NimbleMiner: An open-source nursing-sensitive natural language processing system based on word embedding [J]. Computers, Informatics, Nursing, 2019, 37(11): 583-590

[47] Doing-Harris K, Patterson O, Igo S, et al. Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts [C] //Proc of the 7th Int Workshop on Data and Text Mining in Biomedical Informatics. New York: ACM, 2013: 9-12

[48] Kocbek S, Cavedon L, Martinez D, et al. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources [J]. Journal of Biomedical Informatics, 2016, 64: 158-167

[49] Ormerod M, Martínez-Del-Rincón J, Robertson N, et al. Analysing representations of memory impairment in a clinical notes classification model [C] //Proc of the 18th BioNLP Workshop and Shared Task. Stroudsburg: ACL, 2019: 48-57

[50] Du Baochen. Research and implementation of text mining based on medical record data [D]. Beijing: Beijing University of Posts and Telecommunications, 2019 (in Chinese)
(杜宝琛. 基于病历数据的文本挖掘研究与实现[D]. 北京: 北京邮电大学, 2019)

[51] Lu Yuanyuan, Deng Yongli, Liu Mingliang, et al. Classification of short medical records using the structural features of entity and dependency syntax [J]. Chinese Journal of Medical Instrumentation, 2016, 40(4): 245-249 (in Chinese)
(吕愿愿, 邓永莉, 刘明亮, 等. 利用实体与依存句法结构特征的病历短文本分类方法[J]. 中国医疗器械杂志, 2016, 40(4): 245-249)

[52] Heilman M, Smith N A. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions [C] //Proc of Human Language Technologies: The 2010 Annual Conf of the North American Chapter of the ACL. Stroudsburg: ACL, 2010: 1011-1019

[53] Wang Mengqiu, Manning C D. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering [C] //Proc of the 23rd Int Conf on Computational Linguistics. Beijing: Coling 2010 Organizing Committee, 2010: 1164-1172

[54] Wang Di, Nyberg E. A long short-term memory model for answer sentence selection in question answering [C] //Proc of the 53rd Annual Meeting of the ACL and the 7th Int Joint Conf on Natural Language Processing (Volume 2: Short Papers). Stroudsburg: ACL, 2015: 707-712

[55] Xiong Caiming, Zhong V, Socher R. Dynamic coattention networks for question answering [J]. arXiv preprint, arXiv: 1611.01604, 2016

[56] Tan M, Santos C D, Xiang B, et al. LSTM-based deep learning models for non-factoid answer selection [J]. arXiv preprint, arXiv:1511.04108, 2015

[57] Dong Li, Wei Furu, Zhou Ming, et al. Question answering over freebase with multi-column convolutional neural networks [C] //Proc of the 53rd Annual Meeting of the ACL and the 7th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: ACL, 2015: 260-269

[58] Santos C D, Tang Ming, Xiang Bing, et al. Attentive pooling networks [J]. arXiv preprint, arXiv:1602.03609, 2016

[59] Lee C, Luo Zhaojing, Ngiam K Y, et al. Big Healthcare Data Analytics: Challenges and Applications [M]. Berlin: Springer, 2017: 11-41

[60] Pampari A, Raghavan P, Liang J, et al. emrQA: A large corpus for question answering on electronic medical records [J]. arXiv preprint, arXiv:1809.00732, 2018

[61] Roberts K, Demner-Fushman D. Annotating logical forms for EHR questions [C] //Proc of Int Conf on Language Resources and Evaluation. Paris: ELRA, 2016: 3772-3778

[62] Gan D. Diabetes Atlas [M]. Belgium: International Diabetes Federation, 2015

[63] Group N D D. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance [J]. Diabetes, 1979, 28(12): 1039-1057

[64] Association A D. Standards of medical care in diabetes—2010 [J]. Diabetes Care, 2010, 33(S1): S11-S61

[65] Zheng Tao, Xie Wei, Xu Liling, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records [J]. International Journal of Medical Informatics, 2017, 97: 120-127

[66] Breault J L, Goodall C R, Fos P J. Data mining a diabetic data warehouse [J]. Artificial Intelligence in Medicine, 2002, 26(1-2): 37-54

[67] Jelinek H F, Stranieri A, Yatsko A, et al. Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis [J]. Computers in Biology and Medicine, 2016, 75: 90-97

[68] Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, et al. A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results [J]. Journal of Clinical Epidemiology, 2016, 71: 76-85

[69] Sideris C, Pourhomayoun M, Kalantarian H, et al. A flexible data-driven comorbidity feature extraction framework [J]. Computers in Biology and Medicine, 2016, 73: 165-172

[70] Ramezankhani A, Pournik O, Shahrabi J, et al. An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database [J]. International Journal of Endocrinology and Metabolism, 2015, 13(2): e25389-e25389

[71] Bernardini M, Romeo L, Misericordia P, et al. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine [J]. IEEE Journal of Biomedical and Health Informatics, 2019, 24(1): 235-246

[72] Tapak L, Mahjub H, Hamidi O, et al. Real-data comparison of data mining methods in prediction of diabetes in Iran [J]. Healthcare Informatics Research, 2013, 19(3): 177-185

[73] Anderson J P, Parikh J R, Shenfeld D K, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: An application of machine learning using electronic health records [J]. Journal of Diabetes Science and Technology, 2016, 10(1): 6-18

[74] Han Longfei, Luo Senlin, Yu Jianmin, et al. Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes [J]. IEEE Journal of Biomedical and Health Informatics, 2014, 19(2): 728-734

[75] Helgheim B I, Maia R, Ferreira J C, et al. Merging data diversity of clinical medical records to improve effectiveness [J]. International Journal of Environmental Research and Public Health, 2019, 16(5): 769-769

[76] Zheng Le, Wang Yue, Hao Shiying, et al. Web-based real-time case finding for the population health management of Patients with Diabetes Mellitus: A prospective validation of the natural language processing-based algorithm with statewide electronic medical records [J]. JMIR Medical Informatics, 2016, 4(4): No.e37

[77] Pham T, Tran T, Phung D, et al. Predicting healthcare trajectories from medical records: A deep learning approach [J]. Journal of Biomedical Informatics, 2017, 69: 218-229

[78] Liu Jingshu, Zhang Z, Razavian N. Deep ehr: Chronic disease prediction using medical notes [J]. arXiv preprint, arXiv:1808.04928, 2018

[79] Davidson J A, Warren-Gash C. Cardiovascular complications of acute respiratory infections: Current research and future directions [J]. Expert Review of Anti-Infective Therapy, 2019, 17(12): 939-942

[80] Panahiazar M, Taslimitehrani V, Pereira N, et al. Using EHRs and machine learning for heart failure survival analysis [J]. Studies in Health Technology and Informatics, 2015, 216: 40-44

[81] Li Bin, Ding Shuai, Song Guolei, et al. Computer-aided diagnosis and clinical trials of cardiovascular diseases based on artificial intelligence technologies for risk-early warning model [J]. Journal of Medical Systems, 2019, 43(7): 228

[82] Zhao Juan, Feng Qiping, Wu P, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction [J]. Scientific Reports, 2019, 9(1): 1-10

[83] Kim J, Kang U, Lee Y. Statistics and deep belief network-based cardiovascular risk prediction [J]. Healthcare Informatics Research, 2017, 23(3): 169-175

[84] Su Jia, He Bin, Wu Hao, et al. Cardiovascular disease risk factor annotation system and corpus construction based on Chinese electronic medical records [J]. Acta Automatica Sinica, 2019, 45(2): 420-426 (in Chinese)
(苏嘉, 何彬, 吴昊, 等. 基于中文电子病历的心血管疾病风险因素标注体系及语料库构建[J]. 自动化学报, 2019, 45(2): 420-426)

[85] Krittanawong C, Zhang Hongju, Wang Zhen, et al. Artificial intelligence in precision cardiovascular medicine [J]. Journal of the American College of Cardiology, 2017, 69(21): 2657-2664

[86] Leira E C, Chang K C, Davis P H, et al. Can we predict early recurrence in acute stroke? [J]. Cerebrovascular Diseases, 2004, 18(2): 139-144

[87] Goyal M. Long short-term memory recurrent neural network for stroke prediction [C] //Proc of Int Conf on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer, 2018: 312-323

Wu Zongyou, born in 1983. PhD candidate. His main research interests include big data, machine learning and management science.
吴宗友,1983 年生.博士研究生.主要研究方向为大数据、机器学习和管理科学.

Bai Kunlong, born in 1994. Master candidate. His main research interests include data mining and machine learning.
白昆龙,1994 年生.硕士研究生.主要研究方向为数据挖掘和机器学习.

Yang Linrui, born in 1996. Master candidate. Her main research interests include natural processing language and deep learning.
杨林蕊,1996 年生.硕士研究生.主要研究方向为自然语言处理和深度学习.

Wang Yiqi, born in 1996. PhD candidate. His main research interests include natural processing language, object detection and deep learning.
王仪琦,1996 年生.博士研究生.主要研究方向为自然语言处理、目标检测和深度学习.

Tian Yingjie, born in 1973. Professor, PhD supervisor. His main research interests include machine learning, deep learning, data science and intelligent knowledge management.
田英杰,1973 年生.教授,博士生导师.主要研究方向为机器学习、深度学习、数据科学和智能知识管理.