

# 机器学习算法在中医诊疗中的研究综述

张晓航<sup>1,2</sup> 石清磊<sup>4</sup> 王斌<sup>5</sup> 王炳蔚<sup>1</sup> 王永吉<sup>1,3</sup> 陈力<sup>1,2</sup> 吴敬征<sup>1</sup>

(中国科学院软件研究所协同创新中心 北京 100190)<sup>1</sup> (中国科学院大学 北京 100049)<sup>2</sup>

(中国科学院软件研究所计算机科学国家重点实验室 北京 100190)<sup>3</sup>

(西门子医疗系统有限公司北京分公司临床科研部 北京 100102)<sup>4</sup>

(中国中医科学院中医临床基础医学研究所 北京 100700)<sup>5</sup>

**摘要** 机器学习算法包括传统机器学习算法和深度学习算法。传统机器学习算法在中医诊疗领域中的应用研究较多,为探究中医辨证规律提供了参考,也为中医诊疗过程的客观化提供了依据。与此同时,随着其在多个领域不断取得成功,深度学习算法在中医诊疗中的价值越来越多地得到业界的重视。通过对中医诊疗领域中使用到的传统机器学习算法与深度学习算法进行述评,总结了两类算法在中医领域中的研究与应用现状,分析了两类算法的特点以及对中医的应用价值,以期机器学习算法在中医诊疗领域的进一步研究提供参考。

**关键词** 机器学习,深度学习,中医

**中图分类号** TP181 **文献标识码** A

## Review of Machine Learning Algorithms in Traditional Chinese Medicine

ZHANG Xiao-hang<sup>1,2</sup> SHI Qing-lei<sup>4</sup> WANG Bin<sup>5</sup> WANG Bing-wei<sup>1</sup> WANG Yong-ji<sup>1,3</sup> CHEN Li<sup>1,2</sup> WU Jing-zheng<sup>1</sup>

(X-Lab, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)<sup>1</sup>

(University of Chinese Academy of Sciences, Beijing 100049, China)<sup>2</sup>

(State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)<sup>3</sup>

(Diagnostic Imaging Scientific Research Department, Siemens Healthcare Limited Company Branch of Beijing, Beijing 100102, China)<sup>4</sup>

(Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China)<sup>5</sup>

**Abstract** Machine learning algorithms include traditional machine learning algorithms and deep learning algorithms. There exist more reports for traditional machine learning algorithms in the field of traditional Chinese medicine(TCM) diagnosis and treatment, which provides reference used for exploring the dialectical laws of TCM and provides the basis for the objectification of TCM diagnosis and treatment. At the same time, the latest advances in deep learning technologies provide new effective paradigms in obtaining end-to-end learning models from complex data. Deep learning algorithms have gained great success and become increasingly popular in more and more areas. The value of deep learning algorithms in TCM diagnosis and treatment has been paid more and more attention to by the industry. In this paper, the review of traditional machine learning algorithms and deep learning algorithms used in the advance of the TCM domain overe given. Firstly, the research and application status of the two algorithms in the TCM domain was summarized. Then in view of the analyzed work, different characteristics and limitations were found between traditional machine learning algorithms and deep learning algorithms. Finally, these characteristics and limitations were discussed and the existing problems and recommendations were put forward, so as to provide a reference for the further study of machine learning algorithm in the field of TCM.

**Keywords** Machine learning, Deep learning, Traditional Chinese medicine

## 1 引言

机器学习作为人工智能最基本的形式,包括传统的机器学习算法和深度学习算法<sup>[1]</sup>,已被广泛应用于医疗领域,在疾

病的诊疗中发挥着重要的作用。长期以来,由于传统的机器学习算法通过选取合适的特征值,在小数据集上也可以取得较准确的效果,因此基于传统机器学习算法的研究在中医诊疗领域出现了较多的报道,这些研究在探寻基于中医的分类

本文受国家重点研发计划项目(2017YFB1002300,2017YFC1703505),国家自然科学基金(61772507)资助。

张晓航(1995-),男,硕士生,主要研究方向为图像处理、深度学习;石清磊(1982-),男,硕士,主要研究方向为MR及AI新技术在临床中的应用;王斌(1976-),男,博士,副研究员,主要研究方向为医疗信息方向的数据采集与利用;王炳蔚(1991-),男,硕士,主要研究方向为机器学习;王永吉(1962-),男,博士,教授,CCF高级会员,主要研究方向为虚拟化技术、隐蔽信道、实时系统、人工智能,E-mail: ywang@itechs.iscas.ac.cn(通信作者);陈力(1989-),男,博士生,主要研究方向为实时系统、优化算法、可满足性模理论;吴敬征(1982-),男,博士,副研究员,CCF专业会员,主要研究方向为隐蔽信道分析、网络信息安全、安全操作系统。

分型规则、促进中医的标准化和客观化方面发挥了重要的作用。同时,基于传统机器学习算法的中医智能辅助诊疗系统也已经投入使用<sup>[2]</sup>。

与传统机器学习算法通常需要专家在原始数据上进行特征选择不同,深度学习算法作为多层次的表示学习算法,可以从原始数据中自动提取特征,并且逐层将低层次的表示抽象为更高层次的表示。在深度学习算法广泛应用于语音识别、图像识别以及自然语言处理等领域的同时,基于丰富的医疗数据,深度学习算法也逐渐开始在医疗领域扮演着越来越重要的角色<sup>[3]</sup>。深度学习在辅助医疗、医学影像、药物挖掘以及健康管理等方面都有着巨大的应用潜力<sup>[4]</sup>。而随着中医药信息化工作的开展,中医领域建立了共享的数据中心和共享网络,并且大多数中医医院已建立了医院信息系统与电子病历系统,中医领域基于深度学习的研究也开始起步。

本文主要对传统机器学习算法与基于神经网络的深度学习算法在中医诊疗中的研究与应用现状进行了总结,进而讨论了两类算法的特点和未来的应用前景,指出了基于机器学习算法在中医领域开展研究所面临的挑战,以期能为机器学习算法在中医诊疗领域的进一步研究与应用提供参考和依据。

## 2 机器学习算法

### 2.1 传统机器学习算法

机器学习算法是人工智能的一个分支,目标是不通过复杂的编程能够赋予计算机学习的能力,使得计算机能够利用现有数据对内部机制不明确或不能有效定义的问题进行有效的预测。从学习方式的角度来看,机器学习算法可以分为监督式学习、无监督式学习、半监督式学习以及强化学习。在监督式学习算法中,每组训练数据都有一个明确的标识;在非监督式学习中,数据并不被特别标识,学习目标是推断出数据的内在结构;在半监督式学习中,部分数据被标记,部分数据没有标记;强化学习是智能系统从环境到行为映射的学习,通过将环境提供的强化信号作为动作信号好坏的评价标准来改进行动方案以适应环境,从而达到学习的目的。在医学诊疗中,传统机器学习算法中的聚类算法、分类算法、回归算法以及关联规则算法均得到了应用,其中聚类算法与关联规则算法属于无监督式学习,而分类算法与回归算法则属于监督式算法。传统的机器学习算法通常需要对研究问题的领域知识有深入的了解,进而完成特征的设计与提取,最后才可以对模型进行构造和训练。中医医疗领域主要使用到的传统机器学习算法如表 1 所列。

表 1 应用在中医领域的传统机器学习算法

算法名称	算法简介
k-means	k-means 算法是典型的基于距离的聚类算法,采用距离作为相似性的评价指标,把得到紧凑且独立的 $k$ 个簇作为最终目标。
Fuzzy c-means	Fuzzy c-means 是一种允许同一数据属于多个不同类的聚类算法。它通过优化模糊目标函数得到每个样本点对类中心的隶属度,从而决定样本点的归属。
决策树 (Decision Tree, DT)	决策树算法通常是一个递归地选择最优特征,并根据该特征对训练数据进行分割,使得对各个子数据集有一个最好分类的过程。
支持向量机 (Support Vector Machine, SVM)	支持向量机在高维或无限维空间中构造超平面或超平面集合,其可以用于分类、回归或其他任务。
关联规则 (Association rule)	关联规则是从事务集中挖掘事务之间关联关系的无监督机器学习算法。

### 2.2 深度学习算法

深度学习算法的概念最早在 1986 年被引入机器学习领域,被认为是受生物大脑的工作模式启发而设计出来的算法系统,其在发展过程中大量借鉴了统计学与应用数学等领域的知识。2006 年,在发现基于无监督标准的贪心逐层训练过程能够为多层联合训练过程找到一个好的初始值并成功训练全连接的结构之后<sup>[5]</sup>,深度学习开始被研究者们广泛关注。与传统的机器学习算法不同,深度学习不需要对领域知识进行深入的理解掌握就能完成特征的构造与提取。同时,相较于传统的三层人工神经网络,深度学习模型在输入层和输出层之间拥有多个隐含层。深度学习模型的多层结构将复杂映射分解为一系列嵌套的简单映射,通过逐层抽象实现从局部特征到整体特征的提取。图 1 给出了深度学习系统如何通过组合简单的概念(例如转角和轮廓)逐层抽象而最终完成图像分类的操作<sup>[6]</sup>。

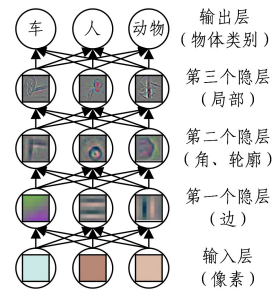


图 1 深度学习算法示意图

近年来,得益于计算机硬件的进步、规模较大数据集的完备、更有效的训练技术和各种深度学习模型框架的发展,深度学习算法的普及性和实用性都得到了极大的发展。深度学习算法主要包含自动编码器、卷积神经网络、循环神经网络和深度置信网络 4 类,表 2 对这些算法进行了简单的介绍。在医疗领域,基于深度学习算法的研究中,卷积神经网络(Convolution Neural Networks, CNN)主要应用于医学影像数据的处理,循环神经网络(Recurrent Neural Networks, RNN)主要应用于电子病历数据的处理,自动编码器(Auto Encoders, AE)和深度置信网络(Deep Brief Networks, DBN)在医学影像数据和电子病历数据中都有所应用。

表 2 深度学习算法的分类

算法类别	算法简介
自动编码器	自动编码器是输入和输出相同的深度学习模型。
深度置信网络	深度置信网络是由一系列受限玻尔兹曼机组成的深度学习模型。
卷积神经网络	卷积神经网络是受启发于猫的视觉神经单元,引入卷积操作,层与层之间局部连接且连接权值共享的深度学习模型。
循环神经网络	循环神经网络是引入了定向循环,隐含层内部神经元存在连接的深度学习模型。

## 3 机器学习算法在中医领域中的应用

### 3.1 传统机器学习算法在中医领域中的应用

#### 3.1.1 聚类算法在中医领域中的应用

聚类算法是一种无监督学习算法,该算法基于数据的内部结构对样本进行分类,使同一类别内的个体具有尽可能高的同质性,而类别之间则具有尽可能高的异质性。裴卫等<sup>[7]</sup>以 589 首中医治疗肝硬化的处方为研究对象,运用 k-means

方法对方剂中涉及到的 257 味中药进行聚类分析,结果显示中医治疗肝硬化的中药主要可以归结为 5 类,体现了药物的 5 类特征,经临床确认基本符合中医临床的用药规律。Dai 等<sup>[8]</sup>针对中医对人体体质类别划分的问题设计了 60 个特征指标,并针对这些特征指标制作了调查量表,运用 k-means 算法在调查量表统计结果的基础上进行聚类分析,该方案对当前 9 种人体体质类别进行了进一步划分并给出了划分依据,k-means 算法使得研究者对人体体质类别划分问题有了进一步的认识。在脉象研究中,由于不同医生对脉象的体会不同,因此不同的医生对同一脉象的区分标注可能会有所不同。冯冰等<sup>[9]</sup>针对中医脉象标注问题,对仪器采集到的脉诊信号数据首先进行双树复小波变换,然后以梅尔倒谱系数进行特征提取,最后应用 Fuzzy c-means 聚类算法进行聚类。该方法对脉象的识别准确率达到 78.2%,一定程度上避免了医生主观体会对脉象标注的影响。以 k-means 为代表的聚类算法作为无监督学习算法,以不可知的分类目标对样本进行归类,在很大程度上避免了人为分类产生的主观影响,推动了中医的客观化。同时,在分类后进行总结发现,聚类算法能够帮助研究者对中医中的药物性味、功能主治以及症状特征等问题进行进一步的认识和归纳。

### 3.1.2 分类算法在中医领域中的应用

分类方法是一种对离散型随机变量建模或预测的监督学习算法。为了探索鼻鼾的中医辨证规律,黄嘉韵等<sup>[10]</sup>制定了中医鼻鼾辨证分型诊断标准,并对原始数据中每个用于分型的变量指标进行了量化。通过对 560 例鼻鼾病例构建决策树模型,筛选出 5 条判断规则,对鼻鼾症型分类获得了 91.5% 的准确度。同时,通过将筛选出的 5 条判断规则结合在一起,研究者进一步明确了判断鼻鼾的辨证分型分类规则。针对慢性胃炎症状、体征的选择和实证症候分类识别的问题,徐玮斐等<sup>[11]</sup>组织临床专家成立研究小组,制定了包含 113 个特征变量的调查量表。通过决策树算法中的随机森林算法对慢性胃炎症候进行建模,在 113 个特征变量中挑选出了 15 个特征症状,对慢性胃炎症候分类的准确率达到 83%。该方案为证候诊断的客观化提供了直接的参考和依据,也为建立慢性胃炎诊断标准提供了借鉴。为了分析中医临床记录中症状与症候类别之间的关系,叶培等<sup>[12]</sup>针对包含 2000 多种症状以及 269 种症型的 2597 例病例开展了研究,该研究首先对症状和病例进行了规范化,然后建立了规范的症型症状表并进行了特征选取,最后通过最大熵模型,根据症状对症型分类。该方法的准确率达到 85.32%,召回率达到了 93.08%,在中医临床诊断中具有较大的应用价值。对于舌诊图像的质量评价问题,王亚真等<sup>[13]</sup>参照中医诊断学中望舌的标准,提取基于自然场景统计的特征、纹理、颜色和几何特征作为输入,并通过支持向量机算法将舌诊图像分为合格与不合格两类,在 616 例舌诊图片的分类实验中,该方案的分类准确度达到 83.12%,从而证明该方法在中医舌诊图像的质量评价中具有较大的应用价值。中医智能诊断本质上是中医病例样本的分类问题,将专家筛选出的特征作为输入并进行模型构造和训练的分类算法模型在辅助医生诊断上具有较大的应用潜力。

### 3.1.3 回归分析算法在中医领域中的应用

回归分析算法是分析两个及两个以上变量间相互依赖关

系的一种统计分析方法,常用于预测分析以及变量之间因果关系的分析。徐琳等<sup>[14]</sup>以 1474 例慢性乙型肝炎患者为研究对象,运用二元 Logistic 回归分析方法探讨慢性乙型肝炎症候与症状特征间的关系。该方案对湿热内阻症筛选出 32 项症状指标,对肝郁脾虚症筛选出 31 项特征指标,对肝肾阴虚症筛选出 39 项特征指标,以预测概率 0.5 为判别分界点,正确分类能力分别达到了 85.8%,86.1%以及 89.8%。戴方瑜等<sup>[15]</sup>以 128 例自发性幕上脑出血患者为研究对象,运用二元 Logistic 回归分析方法从 32 个特征指标中筛选出年龄、中医辨证分型以及脑室出血 3 个特征作为预测因素,对自发性幕上脑出血预后情况的预测特异度达到 94.6%,预测率达到 87.1%。该方案能够为自发性幕上脑出血急性期最佳治疗方案提供参考依据,也能够为指导临床预后估计提供依据。王秀凤等<sup>[16]</sup>以卵巢早衰相关的 12 个生物因子为研究对象,运用支持向量机回归算法分别建立 14 个预测模型来探讨各个生物分子之间在定经汤作用下的相互调控关系,结果显示 14 个预测模型的复相关系数均大于 0.78,对生物分子含量变化的预测比较可靠,能够很好地帮助研究者对定经汤调节卵巢早衰机制的进一步认识。回归算法在探讨中医症状与症候之间的关系上得到了一定的应用。

### 3.1.4 关联规则算法在中医领域中的应用

关联规则算法是从大量的数据中寻找数据之间关联规则的机器学习算法。通常以蕴含式的形式表示关联规则,以支持度衡量规则的普遍性,并以置信度衡量规则的可靠度。季涛等<sup>[17]</sup>以中医古方方剂数据库中 557 首治疗消渴症的方剂为研究对象,运用关联规则算法对中医治疗消渴症方剂的高频药物、药对、药组进行分析。该方案总结出高频药物 93 种、药对 187 组、3 味中药药组 117 组以及 4 味中药药组 27 组,显示出对于消渴症不同症型遣方用药的不同侧重。陶方方等<sup>[18]</sup>以沈敏鹤治疗卵巢癌 778 份医案为研究对象,运用关联分析方法研究沈敏鹤治疗卵巢癌的用药规律。该方案总结出 21 味高频药物、18 对核心药对、42 条关联规则、3 个核心方剂以及沈敏鹤治疗卵巢癌的 3 个创新药对,较好地体现了沈敏鹤中医师治疗卵巢癌的临床思维路径以及主要治疗方法。郭宏等<sup>[19]</sup>以中医师治疗鼻鼾的 560 例门诊病例为研究对象,运用关联规则算法探讨鼻鼾药物配伍关系。该方案除了得出郭宏中医师治疗鼻鼾方剂配伍使用频次较高且关联性较强的配伍药对和组药外,还得出鼻鼾“寒热错杂、虚实并见”症型中的特殊关联药组。关联规则分析算法的优点是简单,容易被解释和理解,因此其在中医临床资料研究尤其是用药规律的研究中被广泛运用。关联规则分析在研究用药与症状、证型、疾病分期,乃至节气等之间的联系问题上表现出了较大的应用潜力。

### 3.2 深度学习算法在中医领域中的应用

由于中医诊断的各项特征极其复杂,且症状之间存在大量的多重共线性关系和协同关系,浅层的神经网络模型很难准确地模拟症状和用药效果之间的复杂关系,更难以描述中医药诊疗的真实情况。面对这种困境,深度学习作为一种新的思路被应用到了中医药相关问题的研究中。

中草药是中医预防、治疗疾病所使用的独特药物,对中草药知识掌握的不足可能引起药物的误用。Weng 等<sup>[20]</sup>首先使用变形、上下移位以及亮度调节等多种数据增强方法对中草药

药图片进行预处理,然后使用包含 5 个卷积层和 3 个全连接层的卷积神经网络模型 CafeNet<sup>[21]</sup>开发中医药识别 APP。在涵盖 11 种中草药的 2200 张图片的交叉验证实验中,该应用对中草药的识别准确率达到了 96%,可以满足实际应用对精确度的要求。

在中医中,病患的体质是诊疗过程中需要考虑的重要因素。但是,在大多数医学研究中,研究者往往忽略了对病患体质的考虑,而只考虑了疾病本身。文献[22-23]使用卷积神经网络模型对 Wang<sup>[24]</sup>划分出的 9 种不同体质进行分类。Huan 等<sup>[22]</sup>分别从 3 个中医门诊医院采集了 5330 张面部图像,使用由 7 个卷积层、3 个池化层、1 个起始层以及 2 个全连接层组成的卷积神经网络模型对病患体质进行分类,达到了 65.29% 的分类准确度。Li 等<sup>[23]</sup>从 1661 名参与者身上的 6 个部位分别采集脉搏信号数据,在这些数据中不同体质的数量分别是 867,79,83,205,234,76,43,33,43。使用包含输入、输出层在内的 12 层卷积神经网络模型在该数据集上针对人体体质分类问题进行训练测试,结果该模型在人体体质分类问题上达到了 95.33% 的准确率。

舌诊是中医诊疗的重要手段。Huo 等<sup>[25]</sup>针对裂纹舌、芒刺舌以及齿痕舌的不同特征,首先使用 Gabor 滤波器算法和轮廓检测算法对舌形图像分别进行了预处理,然后通过卷积神经网络模型 AlexNet<sup>[26]</sup>的每个卷积层后添加批归一化层构造二值分类模型来对舌形图像进行分类。该分类模型在裂纹舌、芒刺舌以及齿痕舌上的分类准确率分别达到了 98%,90% 和 81.2%,相对于不经过预处理的图像数据,分类准确率都有所提升。同时借助于批归一化层,卷积神经网络的训练时间缩短了 14.2%,效率得到了提升。

表 3 和表 4 分别对深度学习算法在中医和西医中的应用情况进行了总结。

表 3 深度学习算法在中医领域中的应用总结

算法类别	算法应用	参考文献
自动编码器	无	
卷积神经网络	通过中草药影像数据,利用卷积神经网络对中草药种类进行识别	[20]
	通过面部影像数据,利用卷积神经网络对体质进行分类	[22]
	通过脉搏信号数据,利用卷积神经网络对体质进行分类	[23]
	通过舌头影像数据,利用卷积神经网络对舌形进行分类	[25]
循环神经网络	无	
深度置信网络	无	

表 4 深度学习算法在西医领域中的应用总结

算法类别	算法应用	参考文献
自动编码器	通过电子病例数据,利用降噪自动编码器对个体罹患各类疾病的风险进行评估	[27]
卷积神经网络	通过电子病例数据,利用卷积神经网络对个体在 6 个月内罹患各类疾病的风险进行评估	[28]
	通过视网膜影像数据,利用卷积神经网络对糖尿病视网膜病变和视网膜黄斑水肿进行检测	[29]
	通过皮肤图像数据,利用卷积神经网络对皮肤癌进行检测	[30]
	通过彩色眼底照片,利用卷积神经网络对出血情况进行检测	[31]
循环神经网络	通过电子病例数据,利用循环神经网络进行药物推荐和疾病复发时间的推测	[32]
深度置信网络	通过 MRI 影像数据,利用深度置信网络对多发性硬化进行检测	[33]

通过分析对比可知,深度学习算法在中医领域的应用尚处于起步状态,主要是通过卷积神经网络在图像数据的基础上进行研究。而西医领域借助于丰富的公共数据集,深度学习中的自动编码器、卷积神经网络以及循环神经网络等各种模型在文本、医学影像以及可穿戴设备数据上均得到了一定程度的应用。同时,以 DeepMind Health 公司开发的 Stream<sup>[34]</sup>为代表,医疗软件在辅助医护人员工作的任务中也已经起到了积极的作用。对于将深度学习算法应用到实际中医诊疗过程中的有关情况,作者未见文献报道,中医领域基于深度学习的研究亟待进一步的深入。

## 4 希望与挑战

### 4.1 中医内在规律的总结发现与中医客观化

由于传统中医诊断学在理论及操作技能中存在着大量的模糊性描述,缺乏客观的评价指标,因此中医的普及、推广和发展受到了严重的制约。传统机器学习算法则为中医的客观化以及中医内在规律的总结和发现提供了新手段。首先,聚类算法作为一种无监督的机器学习方法,可以避免传统分类方法的人为主观性。一方面,对于用药规律的研究,聚类算法从药物客观属性出发进行聚类;另一方面,对于辨证规律的研究,无监督的聚类在帮助发现症状、症候之间的共性,推动症状、症候辨别的客观化与标准化方面起到了重要作用。其次,分类算法和回归算法在优化分类或回归结果的过程中,对于发现影响症候诊断的主要症状有着明显的效果,分类算法和回归算法对于完善规范化的症候诊断具有重要作用。最后,通过药物与药物之间、处方与症状之间以及药物与症状之间的关联规则分析,药物配伍规律以及不同医师的用药侧重点等中医内部规律等都能够得到进一步的发掘与总结。传统机器学习算法在中医内在规律的总结发现以及推动中医客观化、标准化的工作上仍有重要的研究价值。

### 4.2 传统机器学习算法与深度学习算法的比较

从数据集规模的角度来看,深度学习算法模型通常需要海量数据来支持模型的训练和优化,而传统的机器学习算法则在小数据集上也可以取得很好的效果。从特征提取的角度来看,深度学习算法可以从原始数据中自动提取特征,不需要研究者对领域知识有十分深入的了解;传统的机器学习算法则通常需要领域专家在原始数据上进行特征的设计和提取,然后才可以用于算法模型的训练和优化。从可解释性的角度来看,深度学习模型在大多数情况下依旧被看作一个黑盒,中间过程难以解释;传统机器学习算法中的某些算法,如决策树算法等,则可以对输出结果做出清晰的说明,具有良好的可解释性,这在实际应用中对智能诊断软件的可信性和可靠性具有重要的影响。因此,对于特征提取容易并且定义和量化完备的问题,传统机器学习算法依旧可以发挥重要的作用。而对于中医领域中的其他问题,如患病风险评估、用药推荐等,深度学习算法凭借其无需人工构造特征的特性,具有更为重要的研究价值。

### 4.3 中医数据

机器学习算法的成功均离不开数据的支持,尤其是深度学习算法更是需要大量的数据。中医数据主要可以分为临床医疗数据、中医药文献数据、临床研究数据和实验研究数据 4 类。临床医疗数据主要来源于临床信息系统的电子病历等,中医药文献数据主要包括报纸、期刊文献以及论文等,临床研究数据主要包括临床试验数据等,实验研究数据主要包括动

物实验、临床前药理学数据等。同时,随着可穿戴技术在医疗领域的广泛应用,可穿戴设备的数据也是重要的数据来源。然而,一方面,中医领域不同于其他研究领域,数据种类繁多、结构多样并且缺乏统一的标准,如何保证数据集的质量面临着巨大的挑战;另一方面,相对于西医领域的各种机构组织构造了丰富的公共数据集来支持深度学习的研究,中医领域则缺乏相应的工作,中医数据集获取困难并且规模十分有限。为了支持和促进机器学习算法在中医领域研究的开展,针对中医诊疗中的各项问题制定统一的标准并收集数据,是在中医领域基于机器学习开展研究所面临的首要挑战。

## 参考文献

- [1] YU D, DENG L. Deep learning and its applications to signal and information processing[J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154.
- [2] DAOSH. Hypertension TCM Auxiliary Syndrome [EB/OL]. [2017-12-08]. <http://www.daosh.com/en/product/detail.aspx?id=12>.
- [3] MIOTTO R, WANG F, WANG S, et al. Deep learning for healthcare: review, opportunities and challenges[J]. Briefings in Bioinformatics, 2017, 18(3): 1-11.
- [4] MIN S, LEE B, YOON S. Deep learning in bioinformatics[J]. Briefings in Bioinformatics, 2017, 18(5): 851-869.
- [5] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[M]. Cambridge, Mass: MIT Press, 2006.
- [6] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision. Cham, Springer, 2014: 818-833.
- [7] 裴卫, 吴辉坤, 李晓东, 等. 基于机器学习的中医治疗肝硬化组方规律研究[J]. 世界科学技术-中医药现代化, 2017, 19(6): 950-956.
- [8] DAI Y, LONG Y, XING X, et al. Exploring Body Constitution in Traditional Chinese Medicine with K-Means Clustering[C]// Advances in Services Computing: 10th Asia-Pacific Services Computing Conference. Zhangjiajie: Springer International Publishing, 2016: 52-64.
- [9] 冯冰, 李绍滋. 中医脉诊信号的无监督聚类分析研究[J]. 智能系统学报, 2017, 13(4): 564-570.
- [10] 黄嘉韵, 郭宏, 邴艳萍. 基于决策树算法的鼻鼾辨证规律初步研究[J]. 中华中医药杂志, 2016, 31(11): 4770-4773.
- [11] 徐玮斐, 顾颀杰, 刘国萍, 等. 基于随机森林和多标记学习算法的慢性胃炎实证特征选择和证候分类识别研究[J]. 中国中医药信息杂志, 2016, 23(8): 18-23.
- [12] 叶培. 最大熵在中医智能辨证的研究及其应用[J]. 计算机时代, 2015(3): 50-52.
- [13] 王亚真, 张新峰, 胡广芹, 等. 基于支持向量机的中医舌图像质量评价研究[J]. 北京生物工程, 2015, 34(6): 551-557.
- [14] 徐琳, 赵瑜, 彭景华, 等. 慢性乙型肝炎常见证候特征的二元 Logistic 回归分析[J]. 中华中医药杂志, 2015, 30(5): 1780-1783.
- [15] 戴方瑜, 王珏, 於孝龙, 等. 脑出血中医分型与良好预后影响因素的 Logistic 回归分析[J]. 浙江中医药大学学报, 2015(2): 115-119.
- [16] 王秀凤, 张磊, 伍庆华, 等. 基于支持向量回归机的定经汤调节卵巢早衰作用机制研究[J]. 中国中医基础医学杂志, 2015(12): 1516-1518.
- [17] 季涛, 宿树兰, 尚尔鑫, 等. 基于关联规则的中医药治疗消渴症的用药规律与特点探析[J]. 中华中医药杂志, 2016, 31(12): 4982-4986.
- [18] 陶方方, 沈敏鹤, 孔丽娅, 等. 基于关联规则和相关系数的沈敏鹤主任医师卵巢癌用药规律研究[J]. 中华中医药杂志, 2015, 30(1): 235-237.
- [19] 郭宏, 黄嘉韵, 邴艳萍. 基于数据挖掘技术的郭宏治疗鼻鼾用药规律初步研究[J]. 中华中医药杂志, 2017, 32(3): 1332-1335.
- [20] WENG J C, HU M C, LAN K C. Recognition of Easily-confused TCM Herbs Using Deep Learning[C]// Proceedings of the 8th ACM on Multimedia Systems Conference. Taipei: ACM, 2017: 233-234.
- [21] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando: ACM, 2014: 675-678.
- [22] HUAN E Y, WEN G H, ZHANG S J, et al. Deep Convolutional Neural Networks for Classifying Body Constitution based on face image[J]. Computational and Mathematical Methods in Medicine, 2017, 2017(2): 9846707.
- [23] LI H, XU B, WANG N, et al. Deep Convolutional Neural Networks for Classifying Body Constitution [C]// International Conference on Artificial Neural Networks. Cham: Springer, 2016: 128-135.
- [24] WANG Q. Classification and diagnosis basis of nine basic constitutions in Chinese medicine[J]. Journal of Beijing University of Traditional Chinese Medicine, 2005(4): 1-8.
- [25] HUO C M, ZHENG H, SU H Y, et al. Tongue shape classification integrating image preprocessing and Convolution Neural Network[C]// 2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS). Wuhan: IEEE, 2017: 42-46.
- [26] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012: 1097-1105.
- [27] MIOTTO R, LI L, KIDD B A, et al. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records[J]. Scientific Reports, 2016(6): 26094.
- [28] NGUYEN P, TRAN T, WICKRAMASINGHE N, et al. A Convolutional Net for Medical Records [J]. IEEE Journal of Biomedical & Health Informatics, 2017, 21(1): 22-30.
- [29] 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.
- [30] ESTEVA A, KUPREL B, NOVOA R A, et al. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.
- [31] VAN G M, VAN G B, HOYNG C, et al. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images [J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1273-1284.
- [32] CHOI E, BAHADORI M T, SCHUETZ A, et al. Doctor ai: Predicting clinical events via recurrent neural networks[C]// Machine Learning for Healthcare Conference. Boston: PMLR, 2016: 301-318.
- [33] YOO Y, BROSCHE T, TRABOULSEE A, et al. Deep Learning of Image Features from Unlabeled Data for Multiple Sclerosis Lesion Segmentation[C]// International Workshop on Machine Learning in Medical Imaging. Cham: Springer, 2014: 117-124.
- [34] DeepMind Health. Streams in NHS hospitals[A/OL]. [2017-09-18]. <https://deepmind.com/applied/deepmind-health/working-nhs/how-were-helping-today>.