卷积神经网络模型在儿科疾病预测中的应用

李小整[①] 王华珍[*] 熊英杰[①] 曾宇晨[①] 何霆[①] 吴谨准[②] 陈坚[③]

摘　要　目的：针对儿童看病需求量大导致的儿科诊疗服务效率和准确率偏低等问题，利用自然语言处理和深度学习技术，从儿科历史病历数据中自动“学习”专家医生诊断模式，形成智能辅助诊断模型，从而对新的儿科病历数据输出疾病诊断决策。结果：基于深度卷积神经网络的七分类疾病智能诊断模型的正确率为84.26%，F1-score为84.33%，基本达到可投入实际应用的级别。结论：智能诊断决策作为预诊信息提供给医生进行确诊参考，对提升医生诊断速度效果明显。

关键词　深度学习　中文电子病历　卷积神经网络　自然语言处理

Doi:10.3969/j.issn.1673-7571.2018.10.004

[中图分类号] R319;NO31　　**[文献标识码] A**

Application of Convolutional Neural Network in Pediatric Disease Prediction / LI Xiao-zheng, WANG Hua-zhen, XIONG Ying-jie, et al//China Digital Medicine.-2018 13(10): 11 to 13

Abstract　Objective: To solve the inferior efficiency and accuracy of pediatric diagnosis because of the large demand for children's medical treatment, by natural language processing and deep learning technology to automatically "learn" the expert doctor diagnosis mode based on pediatric historical medical record data, and then establishes an intelligent auxiliary diagnosis model. The model can provide disease diagnosis decisions for new pediatric medical record. Results: The accuracy of intelligent diagnosis model based on deep convolutional neural network with respect to seven-classification application is 84.26% and F1-score is 84.33%, which initially achieves the practical level. Conclusion: The intelligent diagnosis decision serves as pre-diagnosis and can offer to the doctor for reference, which can greatly improve the speed of clinical diagnosis.

Keywords　deep learning, Chinese electronic medical records, convolutional neural network, natural language processing

Fund project　National Natural Science Foundation of China under Grant(No. 71571056); The Natural Science Foundation of Fujian Province in China under Grant(No.2012J01274)

Corresponding author　College of Computer Science and Technology, Huaqiao University, Xiamen 361021, Fujian Province, P.R.C.

电子病历记录（Electronic Medical Records，EMRs）作为一种医疗信息化手段，储存了海量的医疗事实数据，是不可多得的医学资源。如何从海量的数据中提取有用的知识，影响了EHR研究的发展[1]。实践中医生通过病人的主诉、现病史、既往史、家族史、相关检查等病历信息对患者进行疾病确诊。但医生的诊断准确性往往取决于个体医学知识和临床经验，特别是低阶医生和偏远地区医生的诊断准确性较低。因此，若能够借助大量的专家电子病历记录进行机器学习，获得病历到疾病的映射关系，构建智能辅助诊断模型，以辅助普通医生的临床决策，提高医疗水平，显得十分重要和有现实意义。当前大部分学者都只针对某一特定疾病对电子病历数据进行数据挖掘，采用的方法也基本属于传统的机器学习的方法，如SVM、KNN、DT等，其应用价值和和临床效果有待提高[2-5]。

儿科诊疗对象是儿童，年龄周期跨度大，身体体质特征差异明显，同时儿科也是医疗垂直领域下的大综合平台，培养出一个合格的儿科医师周期长达十余年。儿童看病需求量大导致儿科诊疗服务效率和准确率偏低。针对儿科电子病历数据病种覆盖面广（类别数量多）、数据分布极端不平衡、诊断机理复杂等缺陷，采用深度卷积神经网络模仿儿科专家医生的

基金项目：国家自然科学基金面上项目（编号：71571056）；福建省自然科学基金面上项目（编号：2012J01274）
*通讯作者：华侨大学计算机学院，361021，福建省厦门市集美区集美大道668号
①华侨大学计算机学院，361021，福建省厦门市集美区集美大道668号
②厦门大学附属第一医院儿科，361003，福建省厦门市镇海路55号
③智业软件股份有限公司，361012，福建省厦门市软件园二期观日路20号

诊疗过程进行建模，构建儿科智能辅助诊断系统。研究内容包括基于儿科电子病历的EMRs分词、EMRs词向量表达、EMRs诊断模型训练，并与电子病历常用的智能诊断研究进行比较。

1 资料与方法

1.1 资料来源 研究的电子病历数据来源厦门大学附属第一医院儿科门诊电子病历系统。课题收集了145 712份有效病历样本，其中包含63种儿科疾病。每份电子病历文本包含年龄、性别、目前病情状态、主诉、现病史、既往史、家族史、体格检查和医生初步诊断（即疾病类型）。鉴于数据集中“急性上呼吸道感染”的样本数占总数50%以上，为了减轻数据集不平衡分布对模型预诊的影响，实验依次尝试选取样本量排名靠前且差异度较大的前8种疾病和样本量排名前32种疾病，以及去掉“急性上呼吸道感染”留下的7种疾病来构建实验样本集。因此，将对7类、8类、32类以及63种疾病共四种儿科EMRs分类问题分别进行研究，以探索CNNs模型对儿科门诊智能预诊的普适性。4种儿科EMRs分类问题实验数据集的分布特性如表1所示。

表1 四种儿科EMRs分类问题实验数据集的分布特性

疾病种类数	疾病名称	样本数
7种	变应性鼻炎、支气管炎、急性支气管炎、呼吸道疾病、支气管哮喘，非危重、腹泻、咳嗽变异性哮喘	49 333
8种	**急性上呼吸道感染**、变应性鼻炎、支气管炎、急性支气管炎、呼吸道疾病、支气管哮喘，非危重、腹泻、咳嗽变异性哮喘	93 428
32种	如附录所示	133 861
63种	如附录所示	145 712

1.2 分析方法 利用自然语言处理和深度学习技术从儿科历史病历数据中自动“学习”专家医生诊断模式。选用卷积神经网络（Convolutional Neural Networks，CNN）来构建智能辅助诊断模型，从而对新的儿科病历数据输出疾病诊断决策。模型构建包括分词、词向量表达、模型训练三个主要步骤，详细设计流程如图1所示。首先，从医院中文电子病历库抽取训练语料，然后基于此语料构建一个医学分词词典，作为分词的外部知识库。之后对语料进行词向量表达，并用10折交叉验证对构建的CNN模型进行评估。

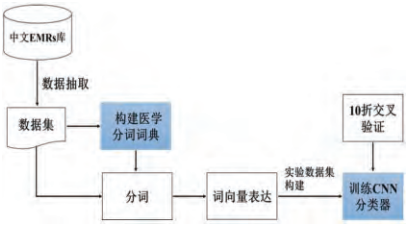

图1 实验架构设计流程图

1.2.1 语料预处理 中文医学语料需要通过分词把文本切分成最小的、独立表达语义的词。分词词典的完备性与准确性将在很大程度上决定了分词的性能[6-7]。利用已获取的电子病历构建了一份基于儿科的医学词典，并将其作为专家知识库在分词时加以调用，以提高分词准确性。

1.2.2 词向量表达 分词后的每个词语需要转化为机器学习能使用的词向量，该向量的相似度可以反映出词语语义上的相似度[8]。采用word2vec的CBOW模型进行词向量表达，以7分类为例，每个词嵌入到50维向量空间中。其中“咳嗽”词向量表示为 $[-3.982, -0.670, -1.754, \cdots, 3.048]_{50}$，为进一步说明word2vec词向量的有效性，其他词与“咳嗽”的向量距离计算如表2所示。

表2 词向量的语义相似度

单词	距离
反复咳嗽	0.6350
轻咳	0.6196
有痰	0.5434
呕吐	29.48
下午	23.41

1.2.3 模型设计 CNN算法框架是一个多层的神经网络，包括嵌入层、卷积层、池化层、全连接层四个部分[9]。CNN通过输入层的卷积计算，每块局部的输入区域连接一个输出神经元。运用不同卷积计算可以形成多通道输出。进而通过池化层采样，最后汇总输出结果。CNN能够进行样本特征的自动、多层次和多角度提取，具有良好的建模能力[10-11]。针对嵌入层的维度，卷积层和池化层的个数，卷积核的大小和数量等框架组成进行设计，同时对模型的参数，如池化方法，激活函数，优化方式，批样本量等，进行调优。文本以7分类问题采用网格搜索的方式，获得了最优性能的参数配置，并用于其他多分类数据集中。实验中采用的卷积神经网络结构主要由一个词嵌入层，两个卷积层和两个池化层构成，四个数据集的词向量维度依次设置为50，80，100，100，卷积核窗口的大小分别设为5×5和3×3，卷积核的个数为128和64，池化层采用的是 max-pooling 方法，dropout值为0.5，激活函数是relu，mini-batch值为95，并采用Adamax更新规则。所有实验都是使用Python 3.5的Python包进行的。

1.2.4 模型评价 基于十折交叉验证，对每一个数据集均采用三个指标，即准确率、精确率和F1-score，来衡量算法的性能。

2 结果

表3展示四种分类问题的模型性能指标。从表3可以看出：7分类CNN模型的准确率、精确率和F1-score都高于84%，达到一个较优的性能，这一结果表明CNN对中文EMRs分类的有效性；7分类CNN模型是四种分类应用表现最好的模型。

表3 基于最优的CNN模型4种分类问题的性能指标

疾病种类数	准确率(%)	精确率(%)	F1-score(%)
7种	84.26	84.58	84.33
8种	82.83	82.59	82.55
32种	73.57	73.23	72.65
63种	71.07	70.17	69.45

3 讨论

对7类、8类、32类以及63种疾病共四种儿科EMRs分类问题分别进行研究，以探索CNNs模型对儿科门诊智能预诊的普适性。7分类CNN模型在四种分类应用中表现最好，原因在于，其他三种数据集的样本表现出极端的类分布不平衡，有些类别没有足够的样本进行训练。

基于卷积神经网络的儿科智能诊断模型准确率最高达到84.26%，F1-score达到84.33%，进一步说明利用CNN实现中文电子病历的智能预诊具有现实意义，基本能够达到医学可用的水平。

参考文献

[1] 胡育.基于病历信息的智能诊断技术研究[D].成都:电子科技大学,2015.

[2] Ekong VE,Onibere EA,Imianvan A.A Fuzzy Cluster Me ans System for the Diagnosis of Liver Diseases[C]// International Journal of Computer Science & Technology,2011,2(3):35-44.

[3] 傅廷君,吴庆斌,钟军锐.基于微信平台的移动医疗应用[J].中国数字医学,2015(9):61-63.

[4] 许幸,张启蕊.基于KNN算法的医药信息文本分类系统的研究[J].计算机技术与发展,2009,19(4):206-209.

[5] 王亮,邹志鹏,姜虹.基于微信小程序的医患交流平台的设计与研究[J].中国数字医学,2017(11):71-73.

[6] 赵明,杜会芳,董翠翠,等.基于word2vec和LSTM的饮食健康文本分类研究[J].农业机械学报,2017,48(10):202-208.

[7] 郭瞳康.基于词典的中文分词技术研究[D].哈尔滨:哈尔滨理工大学,2010.

[8] 黄仁,张卫.基于word2vec的互联网商品评论情感倾向研究[J].计算机科学,2016,43(S1):387-389.

[9] Lecun Y,Boser B,Denker JS,et al.Backpropagation Applied to Handwritten Zip Code Recognition[J].Neural Computation,2014,1(4):541-551.

[10] 刘小明,张英,郑秋生.基于卷积神经网络模型的互联网短文本情感分类[J].计算机与现代化,2017(4):73-77.

[11] 余本功,张连彬.基于CP-CNN的中文短文本分类研究[J].计算机应用研究,2018(4):15-19.

【收稿日期：2018-08-27】
（责任编辑：肖婧婧）

（上接第07页）现信息录入，医生需要一定的培训和熟悉过程。

5.2 语音识别准确率的问题 作为西南地区顶尖的三甲医院，来院的病人多，医院的工作环境复杂，医生的口音多样等问题更加突出，如何排除外界的干扰，更准确地识别出专业的医学术语，也是这个项目面临的重要挑战之一。

5.3 移动医疗新环境的问题 目前移动医护、移动查房越来越普及，而移动端设备普遍存在屏幕小的问题，如何解决在移动端小屏幕上的文本输入、文本识别也是目前亟待解决的问题。

5.4 医疗信息化系统不统一 医院的信息系统种类繁多、涉及的厂商繁多，如何协调相关系统的对接也是目前存在的巨大问题。

5.5 医院数据亟待优化 由于医院原有数据的标签不统一及不规范，导致部分信息无法通过语音描述获得。如何优化医院数据质量，保证数据准确获得也是需要解决的难题。

参考文献

[1] 禹琳琳.语音识别技术及应用综述[J].现代电子技术,2013(13):43-45.

[2] 周英.关于语音识别技术发展趋势的分析[J].计算机光盘软件与应用,2012(19):141-142.

[3] 于俊婷,刘伍颖,易绵竹.国内语音识别研究综述[J].计算机光盘软件与应用,2014(10):76-78.

[4] 科大讯飞.探索语音识别技术的前世今生[J].科技导报,2016,36(9):76-77.

[5] 徐利军.基于DTW的孤立词语音识别研究[J].软件导刊,2012,11(2):137-139.

[6] 李光源.高效语音增强与端点检测技术研究[D].北京:清华大学,2011.

[7] 侯一民,周慧琼,王政一.深度学习在语音识别中的研究进展综述[J].计算机应用研究,2017,34(8):2241-2246.

[8] 张珍.智能机器人语音识别技术[J].现代电子技术,2011,34(12):57-60.

【收稿日期：2018-04-05】
【修回日期：2018-07-06】
（责任编辑：肖婧婧）