专题：医学大数据
Medical Big Data

客座编辑

邹北骥（1961- ），男，博士，中南大学计算机学院教授、博士生导师，曾任中南大学信息科学与工程学院院长。中国计算机学会杰出会员，担任教育部计算机专业类教学指导委员会委员、湖南省普通高等学校计算机专业类教学指导委员会主任、湖南省高等教育学会计算机教育专业委员会理事长等职务。长期在计算机视觉、数字图像处理、计算机辅助设计与图形学和医疗大数据分析等领域从事研究工作，先后主持国家自然科学基金项目4项，国家863计划项目2项，国家973项目子课题1项，国家重大研究计划“人工智能2030”课题1项和企业委托项目20余项。在国内外权威期刊和学术会议上发表论文120余篇。获湖南省教学成果奖一等奖1项，湖南省自然科学奖二等奖1项和科技进步奖三等奖1项。

导读

医学是人类重点关注的领域之一。医学水平与人类健康息息相关，医学的进步是人类健康生活的重要保障。医学领域包括医疗、生物、药物等多个方面，每天产生的数据在EB级以上，医学数据是典型的大数据。采集、分析并挖掘医学大数据中的高价值信息对于利用信息技术开展医学研究、提升临床医疗诊断水平、发现新药物、开展基因分析与各类生物实验等具有重要的意义。《大数据》期刊专门策划了“医学大数据”专题，旨在阐述医学大数据领域的科学问题、研究方法，展示医学大数据领域的最新研究成果，开拓学者的研究视野。本期“医学大数据”专题共收集4篇学术论文。

陈恩红等人撰写的《一种基于深度神经网络的临床记录ICD自动编码方法》，针对国际疾病分类（international classification of diseases，ICD）自动编码问题，提出了一种基于多尺度残差图卷积网络的自动ICD编码方法，采用多尺度残差网络捕获临床文本的不同长度的文本模式，并基于图卷积神经网络抽取标签之间的层次关系，以加强自动编码能力。在真实医疗数据集MIMIC-III上的实验结果表明，该方法在所有指标上均优于现有的模型，显著提高了预测性能。

彭绍亮等人撰写的《基因组大数据变异检测算法的并行优化》，针对海量基因组大数据中的序列比对和变异测序分析问题，采用OpenMP、MPI等技术，对比对算法和测序算法进行了多级并行优化，在不同数据集和并行规模下的测试结果显示，新算法在保证精度的前提下获得了良好的并行性能和可扩展性，有效提高了基因组大数据变异检测能力。

孔桂兰等人撰写的《医疗大数据在学习型健康医疗系统中的应用》，总结了医疗大数据与学习型健康医疗系统（learning health system，LHS）的发展现状，给出了LHS的典型应用，阐述了医疗大数据在LHS中的应用方法和特点，对于推动个性化医疗与精准医学的发展具有重要的意义。

唐艳等人撰写的《基于生成对抗网络的医学数据域适应研究》，提出了一种基于生成对抗网络的方法，以解决在基于深度学习方法的医疗影像辅助诊断技术中因训练数据集样本少导致的预测模型精度低的问题，并针对男女脑影像的差异性研究开展了相关实验，证明了所提出的方法能在一定程度上提升预测模型的泛化能力，缓解由于某个域训练样本较少导致的预测模型在该域测试数据上表现不佳的问题。

由于篇幅有限，本专题不能涵盖医学大数据的方方面面。但仍然希望通过阐述医学大数据的重点研究方向，推动医学大数据的进一步发展。

一种基于深度神经网络的临床记录ICD自动编码方法

杜逸超[1]，徐童[1]，马建辉[1]，陈恩红[1]，郑毅[2]，刘同柱[3]，童贵显[3]
1. 中国科学技术大学计算机科学与技术学院，安徽 合肥 230027；
2. 华为技术有限公司，浙江 杭州 310007；3. 中国科学技术大学附属第一医院，安徽 合肥 230027

摘要
随着国际疾病分类（international classification of diseases，ICD）编码数量的增加，基于临床记录的人工编码难度和成本大大提高，自动ICD编码技术引起了广泛的关注。提出一种基于多尺度残差图卷积网络的自动ICD编码技术，该技术采用多尺度残差网络来捕获临床文本的不同长度的文本模式，并基于图卷积神经网络抽取标签之间的层次关系，以加强自动编码能力。在真实医疗数据集MIMIC-III上的实验结果表明，该方法的P@k和Micro-F1分别为72.2%和53.9%，显著提高了预测性能。

关键词
ICD编码；多尺度；残差网络；图卷积网络

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020040

An automatic ICD coding method for clinical records based on deep neural network

DU Yichao[1], XU Tong[1], MA Jianhui[1], CHEN Enhong[1], ZHENG Yi[2], LIU Tongzhu[3], TONG Guixian[3]
1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China
2. Huawei Technologies Co., Ltd., Hangzhou 310007, China
3. The First Affiliated Hospital of USTC, Hefei 230027, China

Abstract
With the increase in the number of the international classification of diseases (ICD) codes, the difficulty and cost of manual coding based on clinical records have greatly increased, and automatic ICD coding technology has attracted widespread attention. A multi-scale residual graph convolution network automatic ICD coding technology was proposed. This technology uses a multi-scale residual network to capture text patterns of different lengths of clinical text and extracts the hierarchical relationship between labels based on the graph convolutional neural network to enhance the ability of automatic coding. The experimental results on the real medical data set MIMIC-III show that the P@k and Micro-F1 of this method are 72.2% and 53.9%, respectively, which significantly improves the prediction performance.

Key words
ICD coding, multi-scale, residual network, graph convolutional network