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

1 引言

国际疾病分类(international classification of diseases, ICD)编码是在医院等医疗机构使用的统一的编码方法。它根据疾病的病因、病理、临床表现和解剖位置等特性将疾病分门别类,同时也包含手术、诊断和治疗程序的统一代码。ICD代码使用字母数字组合的形式表示具体的疾病或诊断,如E860.0(酒精饮料意外中毒)。ICD代码有多种用途,如报告疾病和健康状况、协助医疗报销决策、收集发病率和死亡率统计数据等。

临床记录包含了患者在医院就诊期间的人口统计学信息、床边的生命体征测量值、实验室测试结果、诊疗程序、药物使用情况、成像报告、死亡率和出院小结等信息。在医疗机构中,编码员通过查看医生的诊断说明和临床记录中的信息手动分配适当的ICD代码,这样的人工编码费时费力且容易出错。人工编码往往会出现以下几个难题:ICD代码的层次结构导致相同层次的疾病往往难以区分;医生在撰写诊断说明时,经常使用缩写词和同义词,极易与ICD编码的描述产生歧义[1];在很多情况下,密切相关的多个诊断描述应该被映射到某一特定ICD编码上,而没有经验的编码人员可能会分别对每种疾病进行编码。

为了降低人工编码的难度,一些工作开始尝试使用机器自动完成ICD编码任务。早期工作通常使用有监督的机器学习方法进行ICD编码,这种方法的效率相对较低。近期研究者采用卷积神经网络(convolutional neural network, CNN)和注意力机制(attention mechanism)结合的方式,大大提高了编码的效率和准确度[2]。虽然之前的方案有所成效,但是自动ICD编码依然存在一些挑战:一是临床记录往往拥有非常长的字符序列,但是其中仅有少部分关键文本片段与某一特定的ICD编码相关;二是ICD编码的标签空间非常庞大,在ICD-9-CM中有超过22 000个编码,而在新版的ICD-10-CM中有超过170 000个编码,庞大的标签空间意味着标签分布存在不平衡的问题。如**图1**所示,在被广泛用于自动ICD编码的重症加强护理病房(intensive care unit, ICU)医疗记录公开数据集MIMIC-III(Medical Information Mart for Intensive Care III)[3]中,共包含8 922个ICD编码,而在所有病历中出现次数小于5次的ICD代码共有4 344个,ICD代码的长尾分布意味着自动编码是一个非常大的挑战。

针对上述问题,笔者基于先前的方法提出了一种多过滤器残差图卷积网络的ICD自动编码技术,可以充分利用临床记录的非结构化数据实现较好的自动ICD编码水平。与之前的工作相比,本文的工作有以下3点贡献。

- 针对冗长、低质量的临床记录文本,之前的工作使用单卷积核进行特征抽取,难以适应每种ICD代码关注的文本片段长度。本文采用多过滤器卷积层抽取不同跨度的文本片段,并使用残差网络扩大接受域,提取长度种类更多的文本片段模式,以适应不同ICD代码关注的文本片段长度。
- 针对层次结构,使用图卷积神经网络(graph convolutional neural network, GCN)抽取标签之间的依赖关系,缓解了标签分布不平衡的现象,并加强了模型的泛化性能。
- 本文的模型提高了在真实的ICU医疗记录数据集MIMIC-III上的自动ICD编码水平。

2 相关工作

2.1 自动ICD编码

针对医疗记录的自动ICD编码一直是医学信息领域的热点问题。20世纪90年代，Larkey L S等人[4]集成了3种分类器：*K*-近邻（*K*-nearest neighbors）、关联反馈（relevance feedback）和贝叶斯独立分类器（bayesian independence classifier），并结合患者的医疗记录进行自动ICD编码，但是他们的方法仅为每个医疗记录分配一个ICD代码。Franz P等人[5]在非结构化的德语文本上针对医疗记录采用了一种诊断记录和ICD代码一对一映射的方式进行编码，显然这种方法与临床实践不符。Perotte A等人[6]使用“平面”和“分层”支持向量机结合MIMIC-II[3]数据集中的出院小结为患者自动分配ICD编码，前者针对代码单独进行预测，而后者仅在存在父亲ICD代码的情况下训练子代码。Kavuluru R等人[7-8]针对肯塔基大学医学中心的71 463条医疗记录中的非结构化文本，提出了一种无监督的集成方法和一种基于临床记录的特征抽取和选择方法，并结合排序算法实现多标签ICD自动编码。Koopman B等人[9]使用一种级联的支持向量机，根据死亡报告识别与癌症相关的死亡原因，模型的第一级根据ICD-10分类系统确定癌症是否存在，第二级为患者自动分配具体的癌症ICD代码。Scheurwegs E等人[10]基于覆盖度的特征选择方法和随机森林，并结合医疗记录中的结构化和非结构化文本信息实现了ICD-9和ICD-10的自动编码。早期工作通常使用有监督的机器学习方法来进行ICD编码，忽略了文本的上下文依赖关系以及关键词语的贡献，这样的方式难以对高噪声、高冗余的现代医疗记录进行自动ICD编码。

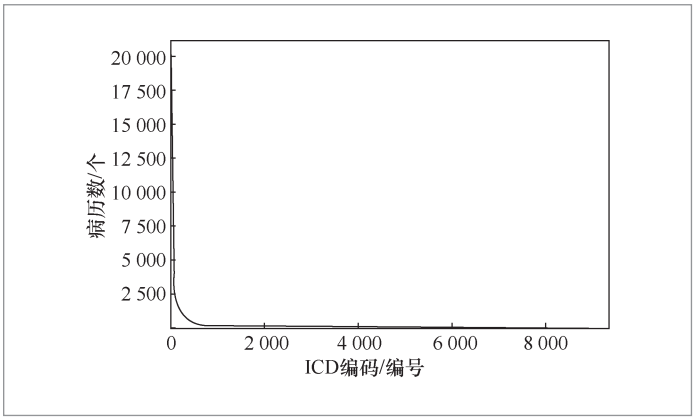


图1 MIMIC-III 数据集中 ICD 编码的分布情况

随着深度学习的发展，近期的许多方法将神经网络的架构应用到自动ICD编码中[11]。Lipton Z C等人[12]利用长短期记忆网络（LSTM）根据临床测量的时间序列预测诊断代码。Xu K等人[13]采用多种模态数据（包括非结构化文本、半结构化文本和结构化表格数据）构建了一个包含卷积神经网络、长短期记忆网络和决策树的混合系统来分配代码。Shi H等人[14]利用字符感知的长短期记忆网络生成书面诊断描述和ICD代码的隐层向量表示，并设计了一种注意力机制来解决诊断描述与相应代码之间不匹配的问题。Xie P等人[15]引入序列树长短期记忆网络（Tree-LSTM）来表示ICD代码的层次结构，并采用对抗网络学习不同医生的诊断记录风格的差异，最终将自动编码转换为语义匹配问题。Duarte F等人[16]利用门控循环单元（gated recurrent unit，GRU）和注意力机制，实现了对癌症病人的死亡证明的自动ICD-10编码。Prakash A等人[17]将维基百科作为知识来源，学习一种压缩记忆神经网络，以保留特征的层次结构，从而预测出现频繁的前50个和前100个ICD代码。Baumel T等人[18]使用具有标签依赖注意力机制的分层

GRU模型对ICD代码进行分类，同时提供了可解释的决策过程。Zeng M等人[19]使用在不同的医疗数据集中进行迁移学习的方式，并引入多尺度卷积神经网络，实现较好的自动ICD编码能力。Mullenbach J等人[2]仅使用MIMIC-III数据集的非结构化文本将卷积神经网络与标签注意力机制结合在一起，实现了自动ICD编码的最佳性能。

2.2 图卷积神经网络

图卷积神经网络主题最近受到越来越多的关注。许多研究者将成熟的神经网络模型（如适用于规则网格结构的CNN）推广到图结构中，以处理更复杂的结构和保存全局信息[20-24]。在这些工作中，Kipf T N等人[24]提出了一种简化的图神经网络模型，即GCN，该模型在许多基准图数据集上达到了先进的水平。近期，图卷积神经网络还被用于文本分类任务中，Yao L等人[25]提出了一种文本图卷积网络（Text-GCN），使用单词和文档的on-hot向量进行初始化，并联合学习单词和文档的表征，以提高文本分类的效果。Peng H等人[26]提出了一种递归正则化的图卷积网络，在单词共现图上进行大规模的文本分类。Rois A等人[27]提出了一种利用GCN学习标签结构化信息的方法，提高了在少样本、零样本情况下的自动ICD编码的性能。Wang W等人[28]将GCN和变分自编码器结合在一起，从而以统一的方式嵌入ICD代码，同时引入多任务学习方法，提高了ICD编码的预测能力。

3 基于多尺度残差图卷积网络的自动编码技术

在本节中，针对冗长且低质量的临床记录和标签空间极其庞大且类别不平衡的ICD代码，笔者提出了一种基于多尺度残差图卷积网络（multi-scale residual graph convolution network，MSResGCN）的方法进行自动ICD编码。

3.1 概述

与Mullenbach J等人[2]提出的方法类似，自动ICD编码可以被视作基于临床记录的多标签文本分类问题。针对临床记录实例i的编码可以被表示成将标签空间中的所有标签$l \in L$映射到$y_{i,l} \in \{0,1\}$（$y_{i,l}$=1表示将标签l分配给实例i）中。**图2**展示了模型的架构，模型包含5个主要组件：词向量查找层、特征抽取层、标签感知的注意力层、标签结构抽取层和输出层。首先通过词向量查找层为临床记录和标签描述生成向量表示；其次，使用含有多个尺度的卷积模块捕获不同长度的文本模式，并通过残差网络扩大接受域；接着，使用标签感知的注意力机制捕获与每个ICD代码最相关的n个连续出现的词语（n-gram），以克服临床记录冗长的问题；最后，通过$|L|$个二元分类器为临床记录分配ICD代码。

3.2 词向量查找层

与先前的工作类似，使用gensim工具包在整个MIMIC-III数据集上预训练word2vec[29]词嵌入向量$\boldsymbol{E}=[e_1,e_2,\cdots,e_N] \in \mathbb{R}^{N \times d_e}$，其中$N$是词表大小，$d_e$是预训练词向量的维度。本文的模型将临床记录序列$\{w_1,w_2,\cdots,w_n\}$作为输入，并通过词向量查找层为临床记录生成文档嵌入向量矩阵$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\} \in \mathbb{R}^{n \times d_e}$，其中$n$是临床记录的序列长度。类似地，依据标签的文本描述序列$\{w_{l1},w_{l2},\cdots,w_{ln}\}(l \in L)$为每个标签生成一个特征向量，以避免学习标签特定的参数，从而缓解标签空间不平

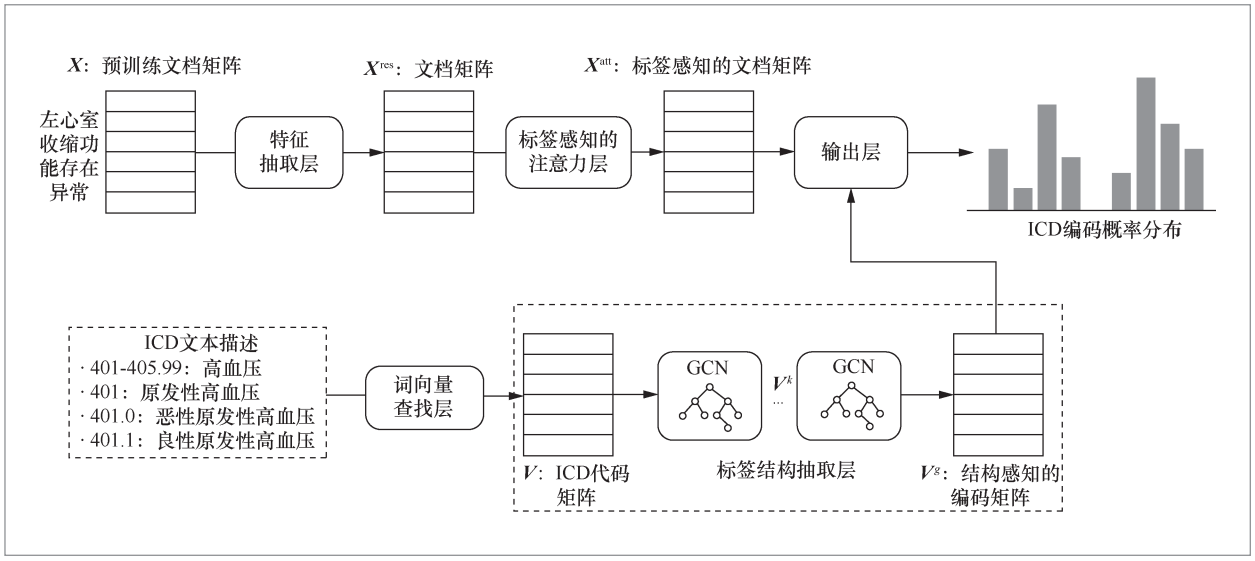


图 2 MSResGCN 整体架构

衡的问题。

$$v_i = \frac{1}{|N_i|}\sum_{j\in N_i} e_j, \quad i=1,\cdots,|M|$$
$$\boldsymbol{V}=[v_1,v_2,\cdots,v_{|M|}]\in\mathbb{R}^{|M|\times d_e} \qquad (1)$$

其中，$\boldsymbol{V}$表示所有标签的表征，v_i表示第i个标签的特征，N_i表示第i个标签的文本描述索引集合，M表示所有ICD代码集合，$|M|$表示M的势。

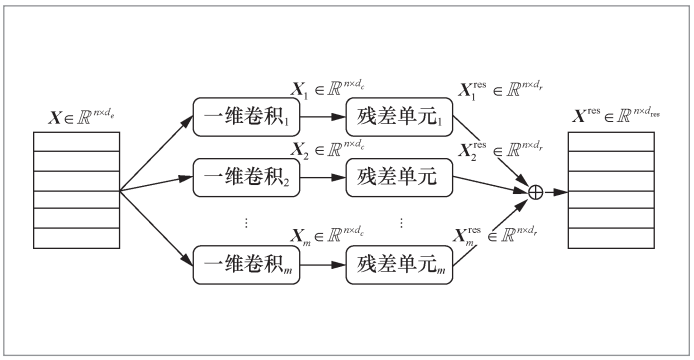


图 3 特征抽取层

3.3 特征抽取层

本文在特征抽取层中设置了两个组件：多尺度卷积层和残差卷积层。由于每个编码对应的临床记录的n-gram的长度会随着标签的改变而变化，多尺度卷积层可以使用多个不同尺度的一维卷积模块捕获多种长度的文本模式。接着通过残差卷积层扩大接受域，以捕获更长的文本模式。**图3**展示了特征抽取层的整体架构。

3.3.1 多尺度卷积层

多尺度卷积层包含多个并行的不同尺度的一维卷积单元。假设拥有m个不同尺度的卷积核，它们对应的尺寸分别为$\mathbb{R}^{s_1\times d_e},\mathbb{R}^{s_2\times d_e},\cdots,\mathbb{R}^{s_m\times d_e}$。对于给定的临床记录输入矩阵$\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\mathbb{R}^{n\times d_e}$，多尺度卷积操作可以被形式化地定义为：

$$\boldsymbol{X}_1=\mathcal{F}_1(\boldsymbol{X},\boldsymbol{W}_1)=$$
$$\tanh\left[\boldsymbol{W}_1^{\top}\boldsymbol{X}^{1:s_1},\cdots,\boldsymbol{W}_1^{\top}\boldsymbol{X}^{j:j+s_1-1},\cdots,\boldsymbol{W}_1^{\top}\boldsymbol{X}^{n:n+s_1-1}\right]$$
$$\cdots$$
$$\boldsymbol{X}_m=\mathcal{F}_m(\boldsymbol{X},\boldsymbol{W}_m)=$$
$$\tanh\left[\boldsymbol{W}_m^{\top}\boldsymbol{X}^{1:s_m},\cdots,\boldsymbol{W}_m^{\top}\boldsymbol{X}^{j:j+s_m-1},\cdots,\boldsymbol{W}_m^{\top}\boldsymbol{X}^{n:n+s_m-1}\right] \qquad (2)$$

其中，$\mathcal{F}(\boldsymbol{X},\boldsymbol{W}_1)$表示对矩阵$\boldsymbol{X}$进行卷积操

作，$\boldsymbol{W}_1\in\mathbb{R}^{(s_1\times d_e)\times d_c}$和$\boldsymbol{W}_m\in\mathbb{R}^{(s_m\times d_e)\times d_c}$表示对应的权重矩阵，$d_c$表示每个卷积层的特征映射维度，$s_1,s_2,\cdots,s_m$表示$m$种不同的卷积尺度，$\boldsymbol{X}^{j:j+s_1-1}\in\mathbb{R}^{s_1\times d_e}$和$\boldsymbol{X}^{j:j+s_m-1}\in\mathbb{R}^{s_m\times d_e}$为输入矩阵$\boldsymbol{X}$的子矩阵，分别表示临床记录文本的第$j$个到第$j+s_1-1$个字符和第$j$个到第$j+s_m-1$个字符的输入矩阵。为了表达简洁，在本文所有的计算式中忽略偏差。因为笔者希望输出矩阵可以保持输入矩阵的行数，所以对输入矩阵进行大小为$(s_m/2)$的填充，并使用步幅为1的一维卷积操作$\mathcal{F}_i,i=1,2,\cdots,m$，最终的输出为$m$个特征矩阵$\boldsymbol{X}_i\in\mathbb{R}^{n\times d_c},i=1,2,\cdots,m$。

3.3.2 残差卷积层

残差卷积层包含多个并行的残差单元，将m个并行的残差单元与多尺度卷积层中对应的一维卷积单元相连，每个残差单元的卷积核大小与对应的一维卷积单元保持一致，即$\mathbb{R}^{s_1\times d_e},\mathbb{R}^{s_2\times d_e},\cdots,\mathbb{R}^{s_m\times d_e}$。如**图4**所示，每个残差单元包含3个一维卷积单元，该单元可以通过扩大接受域来捕获更长的文本特征，并使用短路连接保证网络性能不会下降。

接下来，以第k个尺度的卷积单元的输出矩阵$\boldsymbol{X}_k$为第k个残差单元的输入为例，将残差单元形式化地定义为：

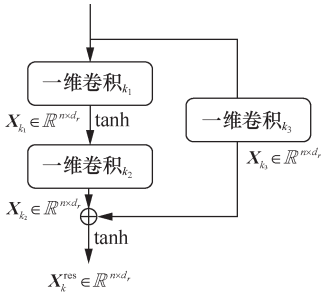


图4 残差卷积单元

$$
\begin{aligned}
&\boldsymbol{X}_{k_1}=\mathcal{F}_{k_1}(\boldsymbol{X}_k,\boldsymbol{W}_{k_1})=\\
&\tanh[\boldsymbol{W}_{k_1}^{\top}\boldsymbol{X}_k^{1:s_k},\cdots,\boldsymbol{W}_{k_1}^{\top}\boldsymbol{X}_k^{j:j+s_k-1},\cdots,\boldsymbol{W}_{k_1}^{\top}\boldsymbol{X}_k^{n:n+s_k-1}]\\
&\boldsymbol{X}_{k_2}=\mathcal{F}_{k_2}(\boldsymbol{X}_{k_1},\boldsymbol{W}_{k_2})=\\
&[\boldsymbol{W}_{k_2}^{\top}\boldsymbol{X}_{k_1}^{1:s_k},\cdots,\boldsymbol{W}_{k_2}^{\top}\boldsymbol{X}_{k_1}^{j:j+s_k-1},\cdots,\boldsymbol{W}_{k_2}^{\top}\boldsymbol{X}_{k_1}^{n:n+s_k-1}]\\
&\boldsymbol{X}_{k_3}=\mathcal{F}_{k_3}(\boldsymbol{X}_k,\boldsymbol{W}_{k_3})=\\
&[\boldsymbol{W}_{k_3}^{\top}\boldsymbol{X}_k^{1:1},\cdots,\boldsymbol{W}_{k_3}^{\top}\boldsymbol{X}_k^{j:j},\cdots,\boldsymbol{W}_{k_3}^{\top}X_k^{n:n}]\\
&\boldsymbol{X}_k^{\text{res}}=\tanh(\boldsymbol{X}_{k_2}+\boldsymbol{W}_{k_3}),k=1,2,\cdots,m
\end{aligned}
\tag{3}
$$

其中，$\boldsymbol{W}_{k_i}$为残差单元中第k_i个卷积单元的权重矩阵，具体的$\boldsymbol{W}_{k_1}\in\mathbb{R}^{(s_k\times d_c)\times d_r}$，$\boldsymbol{W}_{k_2}\in\mathbb{R}^{(s_k\times d_r)\times d_r}$，$\boldsymbol{W}_{k_3}\in\mathbb{R}^{(s_k\times d_c)\times d_r}$。每个残差单元的输出为$\boldsymbol{X}_k^{\text{res}}\in\mathbb{R}^{n\times d_r},k=1,2,\cdots,m$，其中$d_r$表示每个残差卷积层的特征映射维度。与多尺度卷积类似，采用相同的方式对输入矩阵进行填充，以保证输出矩阵和临床记录矩阵的序列长度一致。残差卷积层最终的输出为所有残差单元的拼接：

$$
\begin{aligned}
&\boldsymbol{X}^{\text{res}}=\text{concat}(\boldsymbol{X}_1^{\text{res}},\cdots,\boldsymbol{X}_m^{\text{res}})\in\mathbb{R}^{n\times d_{\text{res}}}\\
&d_{\text{res}}=(m\times d_r)
\end{aligned}
\tag{4}
$$

残差单元可以通过扩大接受域来捕获更长的文本特征，并使用短路连接保证网络性能不会下降。假设第k个单元的卷积核的宽度为s_k=3，多核卷积单元的输出$\boldsymbol{X}_k$的接受域为3，即可以捕获tri-gram的特征，残差卷积单元第一层输出$\boldsymbol{W}_{k_1}$可以捕获5-gram的特征，第二层输出$\boldsymbol{W}_{k_2}$可以捕获7-gram的特征，短路操作可以保持原有特征，从而防止网络退化。

3.4 标签感知的注意力层

与Mullenbach J等人[2]提出的工作类似，本文采用一种标签感知的注意力机制来克服临床记录中关键信息分散的问题。本文为每个ICD代码都分配了一个注意力向量，以确保能够捕捉到临床记录中所有与该ICD代码相关的关键信息。与Mullenbach J等人[2]提出的工作不同的是，

为了缓解标签不平衡的问题，本文将ICD代码的表征作为权重矩阵，而不是学习一个特定注意力参数矩阵。

首先，将特征提取层的输出矩阵$\boldsymbol{X}^{\text{res}}$通过简单的单层神经网络改变矩阵维度，以保证矩阵的第二维与标签向量的第二维一致：

$$\boldsymbol{X}^{r}=\tanh(\boldsymbol{X}^{\text{res}}\boldsymbol{W}_{\text{att}}) \quad (5)$$

其中，$\boldsymbol{X}^{r}\in\mathbb{R}^{n\times d_e}$为改变维度之后的矩阵，$\boldsymbol{W}_{\text{att}}\in\mathbb{R}^{d_{\text{res}}\times d_e}$为权重矩阵。接着，为每一个标签$l$生成注意力向量，并为每个编码生成注意力得分：

$$\begin{aligned}&\boldsymbol{a}_l=\text{softmax}(\boldsymbol{X}^{r}\boldsymbol{v}_l)=\left[a_l^1,a_l^2,\cdots,a_l^n\right],l\in L\\&x_l^{\text{att}}=\boldsymbol{a}_l^{\top}\boldsymbol{X}^{r},l\in L\end{aligned} \quad (6)$$

其中，$\boldsymbol{v}_l\in\mathbb{R}^{d_e}$为标签$l$的向量表示，softmax为归一化指数函数，$a_l^i$为在标签为$l$的前提下文档表示矩阵中第$i$行的注意力得分，$x_l^{\text{att}}\in\mathbb{R}^{d_e}$为文档表示矩阵$\boldsymbol{X}^{r}$与标签$l$有关的行的加权平均值。

3.5 标签结构抽取层

由于ICD编码拥有天然的树状层次结构关系，可以通过GCN捕获标签之间的依赖关系，以进一步缓解标签不平衡的问题。针对标签l的向量表示$\boldsymbol{v}_l$，可以通过结合它的父标签和子标签的向量来更新，第k次更新$\boldsymbol{v}_l$如下：

$$\boldsymbol{v}_l^{k}=f\left(\boldsymbol{W}^{k}\boldsymbol{v}_l^{k-1}+\sum_{j\in P}\frac{\boldsymbol{W}_p^{k}\boldsymbol{v}_j^{k-1}}{|P|}+\sum_{j\in C}\frac{\boldsymbol{W}_c^{k}\boldsymbol{v}_j^{k-1}}{|C|}\right) \quad (7)$$

其中，令$\boldsymbol{v}_l^{0}=\boldsymbol{v}_l$，$f$是激活函数，$\boldsymbol{W}^{k}\in\mathbb{R}^{|M|\times|M|}$，$\boldsymbol{W}_p^{k}\in\mathbb{R}^{|M|\times|M|}$，$\boldsymbol{W}_c^{k}\in\mathbb{R}^{|M|\times|M|}$是权重矩阵，$P$和$C$分别是标签$l$的父标签集合和子标签集合。需要说明的是，在进行标签结构抽取时，本文使用的是整个ICD-9-CM的编码，其中包含了在测试的数据集中没有的编码。选取图卷积神经网络输出的最后一层所形成的矩阵$\boldsymbol{V}^{g}\in\mathbb{R}^{|M|\times d_e}$的子集$\boldsymbol{V}^{f}\in\mathbb{R}^{|L|\times d_e}$作为最终的标签矩阵。

3.6 输出层

根据标签感知的注意力层输出的“ICD-文档”注意力矩阵和标签结构抽取层输出的标签矩阵为临床记录分配类别，定义如下：

$$\hat{y}_l=\text{sigmoid}\left(\boldsymbol{v}_l^{f}x_l^{\text{att}}\right),\quad l\in L \quad (8)$$

其中，$\boldsymbol{v}_l^{f}\in\boldsymbol{V}^{f}$为标签$l$的分类向量，$\hat{y}_l$为预测结果，表示是否将该标签分配给病人。

最后，通过最小化真实值y_l与预测值$\hat{y}_l$的二元交叉熵损失函数来训练本文的模型：

$$\text{loss}=\sum_{l=1}^{L}\left[-y_l\log\left(\hat{y}_l\right)-\left(1-y_l\right)\log\left(1-\hat{y}_l\right)\right] \quad (9)$$

4 实验与分析

4.1 数据集

下面在公开数据集MIMIC-III上对模型进行验证。该数据集包含2001年至2012年在贝斯以色列女执事医疗中心就诊的49 583位患者的58 976次入院记录。每条入院记录都有出院总结，包括病史、诊断结果、手术步骤、出院说明等，编码员根据重要性和相关性从高到低的顺序，为患者在住院期间发生的诊断和程序进行编码。根据患者ID分割数据集，以防止同一名患者同时出现在训练集和测试集中。**表1**是MIMIC-III数据切割与统计数据，共有46 157条出院小结用于训练，3 280条和3 285条数据分别用于验证与测试。该数据集中一共包含8 922个ICD编码，包括6 919个诊断编码和2 003个程序编码，其中训练集中包含8 579种不同的ICD代码。

对于数据的预处理，本文将所有字符

表1 MIMIC-III数据切割与统计数据

对比项	样本数/条	平均字符数	平均标签个数/个	标签种类/种
训练集	46 157	1 445	15.68	8 579
验证集	3 280	1 742	17.72	4 051
测试集	3 285	1 727	17.31	4 042

转换为小写并删除纯数字和符号，但不删除类似于“50 mg”的字符，并将出现次数少于3次的字符替换为“UNK”标记。遵循参考文献[2]的设定，使用gensim工具以连续词袋（CBOW）模型对训练集中的所有文本进行word2vec[27]词向量预训练，向量维度设置为100，窗口大小设置为5。同时由于医疗记录过于冗长，本文将字符长度大于2 500的文本截断，以保证训练速度。

4.2 评价指标

为了与之前的工作进行比较，本文使用多种不同的评价指标对模型进行评价，重点使用微平均值Micro-F1、宏平均值Marco-F1和ROC曲线下的面积（AUC）。Micro-F1是将每个“临床记录-ICD编码”对作为单独的预测来计算的，Marco-F1通过对每个类别计算的指标取平均值而得到。本文还计算了在基准值（ground truth）中出现的得分最高的前k个标签的比例，即P@k，在实验中k分别取8和15。

4.3 基准方法

为了证明本文提出的模型的有效性，将提出的MSResGCN与目前最先进的自动ICD编码方法进行了比较，包含传统的机器学习方法逻辑回归（LR）和3种深度学习方法Text-CNN[30]、CAML[2]和DR-CAML[2]。

- Text-CNN[30]：该方法包含一个单层卷积神经网络，没有标签依赖的注意力机制，仅使用最大池化的方法提取所有ICD编码的表示向量。
- CAML[2]和DR-CAML[2]：这两个方法在MIMIC-III数据集上取得了最优的分类效果。CAML使用Text-CNN进行文档表示学习。为了克服文档过长的情况，Mullenbach J等人[2]提出了标签依赖的注意力机制，以学习每种特定代码与临床记录最相关的n-gram。DR-CAML将标签表征作为损失函数的正则化项来增强CAML。他们假设ICD代码的描述在语义上与输入的文本片段相似，这些文本片段可以通过标签注意力机制来捕获，DR-CAML通过Text-CNN提取标签描述表示形式，然后使用均方损失在ICD编码向量表示和最终分类的权重之间进行正则化。

4.4 实验设置

本文所有实验均在一台处理器为Intel(R) Xeon(R) Gold 5218 CPU@2.30 GHz、内存大小为251 GB、GPU型号为Tesla V100-SXM2、显存大小为32 GB的Centos7服务器上进行。因为模型的超参数较多，所以遵循Mullenbach J等人[2]的工作对一些超参数进行设置，或者根据经验选择一些超参数。预训练词向量的维度d_e为100，多尺度卷积层中每个卷积核输出通道的尺寸d_c为100，学习率为0.000 1，批大小（batch-size）为16，随机失活率（dropout）为0.2，5个卷积核的大小$s_1,s_2,\cdots,s_5$分别为3、5、10、15、20，图卷积神经网络的隐层大小为300，图卷积层数为2。

4.5 实验结果

本文在MIMIC-III数据集上对提出

的模型MSResGCN与部分现有的自动ICD编码方法进行了比较。**表2**给出了所有模型在MIMIC-III数据集上的性能表现。从**表2**可以看出，本文提出的模型MSResGCN在所有指标上都优于之前的分类结果。与之前分类效果最好的模型CAML相比，MSResGCN在Micro-F1上提升了1.1%，在Marco-F1上提高了0.4%；在P@8上提高了0.8%，在P@15上提高了1.2%；同时在Micro-AUC上提高了0.4%，在Marco-AUC上提高了1.3%。从**表2**可以看出，LR在所有的指标上都低于深度学习方法，这是因为前者使用的是传统的人工特征。除此之外，还可以看出，简单的深度学习模型Bi-GRU和Text-CNN有相似的性能水平；CAML和DR-CAML相较于除MSResGCN外的其他3种方法性能较好，CAML将Text-CNN与标签感知的注意力机制结合在一起，提高了抽取临床记录中关键信息的能力。MSResGCN使用多尺度的残差卷积网络来捕获临床记录中不同长度的关键文本片段，同时对标签的层次结构的融合学习使得MSResGCN在标签不平衡的情况下具有高于其他模型的性能。

4.6 消融实验

下面通过设计消融实验来验证MSResGCN的每个组件的有效性。MSResGCN的主要贡献是在CAML上扩展了两个重要的组件，分别是多尺度残差卷积模块和标签结构抽取模块。实验的具体设置如下。

- w/o-MSRes：使用单个尺度的卷积层代替多尺度残差卷积模块，即除了多尺度残差卷积模块，其他组件都与CAML保持一致，卷积核的长度被设置为最优值10。
- w/o-GCN：删除标签结构抽取模块，保持其他组件与CAML一致，即使用特定的参数矩阵代替标签向量计算标签注意力得分和最后的分类。

表3展示了消融实验的结果，在删除多尺度残差卷积模块之后，多个指标都有小幅度下降；在删除标签结构抽取模块后，Marco-F1大幅下降，并且比CAML的Marco-F1低0.3%，同时AUC下降也较为明显。由此可以看出，标签结构抽取模块有助于改善标签不平衡的问题，多尺度残差卷积模块可以捕获更加丰富的关键文本片段。

表2 实验结果

模型	P@k		F1		AUC	
	8	15	Micro	Marco	Micro	Marco
LR	54.5%	41.7%	29.3%	1.2%	92.5%	57.1%
Bi-GRU	63.5%	46.2%	43.3%	4.2%	97.2%	82.4%
Text-CNN	60.1%	45.5%	43.9%	4.5%	96.9%	81.7%
DR-CAML	69.0%	55.3%	52.5%	7.2%	97.9%	87.2%
CAML	71.4%	55.6%	52.8%	7.2%	97.9%	86.4%
MSResGCN	72.2%	56.8%	53.9%	7.6%	98.3%	87.7%

表 3 消融实验结果

模型	P@*k*	F1		AUC	
	8	Micro	Macro	Micro	Macro
w/o-MSRes	71.7%	53.2%	7.5%	98.1%	87.5%
w/o-GCN	72.0%	53.8%	6.9%	98.0%	87.4%
MSResGCN	72.2%	53.9%	7.6%	98.3%	87.7%

4.7 可扩展性

考虑到数据的尺度对模型训练时间的影响，将训练集大小缩小为当前数据集大小的20%、40%、60%、80%进行训练。分别统计出训练时间为275 s/轮、541 s/轮、748 s/轮、1 135 s/轮。由此可以看出，随着数据的增加，MSResGCN的训练时间呈线性增长趋势，具有较好的扩展性。

5 结束语

本文提出了一种用于自动ICD编码的多尺度残差卷积神经网络模型，使用多尺度残差卷积网络来适应不同标签依赖的文本片段的长度，同时使用图卷积神经网络改善了标签不平衡的问题。笔者在MIMIC-III数据集上验证了本文方法的有效性。在接下来的工作中，笔者将考虑将更多的文本数据和模态信息进行融合，以进一步提升自动ICD编码的质量。

参考文献:

[1] SCHNAKERS C, VANHAUDENHUYSE A, GIACINO J, et al. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment[J]. BMC Neurology, 2009, 9(1).

[2] MULLENBACH J, WIEGREFFE S, DUKE J, et al. Explainable prediction of medical codes from clinical text[C]// The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.:s.n.], 2018.

[3] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III: a freely accessible critical care database[J]. Scientific Data, 2016(3).

[4] LARKEY L S, CROFT W B. Combining classifiers in text categorization[C]// The 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1996: 289-297.

[5] FRANZ P, ZAISS A, SCHULZ S, et al. Automated coding of diagnoses-three methods compared[C]// The AMIA Symposium. [S.l.:s.n], 2000.

[6] PEROTTE A, PIVOVAROV R, NATARAJAN K, et al. Diagnosis code assignment: models and evaluation metrics[J]. Journal of the American Medical Informatics Association, 2014, 21(2): 231-237.

[7] KAVULURU R, HAN S, HARRIS D. Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques[C]// 2013 Canadian Conference on Artificial Intelligence. Heidelberg: Springer-Verlag, 2013: 77-88.

[8] KAVULURU R, RIOS A, LU Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records[J]. Artificial Intelligence in Medicine, 2015, 65(2): 155-166.

[9] KOOPMAN B, ZUCCON G, NGUYEN A, et al. Automatic ICD-10 classification of cancers from free-text death certificates[J]. International Journal of Medical Informatics, 2015, 84(11): 956-965.

[10] SCHEURWEGS E, CULE B, LUYCKX K, et al. Selecting relevant features from the electronic health record for clinical code prediction[J]. Journal of Biomedical Informatics, 2017, 77: 92-103.

[11] SHICKEL B, TIGHE P J, BIHORAC A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis[J]. IEEE Journal of Biomedical and Health Informatics, 2017, 22(5): 1589-1604.

[12] LIPTON Z C, KALE D C, ELKAN C, et al. Learning to diagnose with LSTM recurrent neural networks[J]. arXiv preprint, 2015, arXiv: 1511.03677v7.

[13] XU K, LAM M, PANG J, et al. Multimodal machine learning for automated ICD coding[J]. arXiv preprint, 2018, arXiv: 1810.13348v1.

[14] SHI H, XIE P, HU Z, et al. Towards automated ICD coding using deep learning[J]. arXiv preprint, 2017, arXiv: 1711.04075v3.

[15] XIE P, XING E. A neural architecture for automated ICD coding[C]// The 56th Annual Meeting of the Association for Computational Linguistics. [S.l.:s.n.], 2018: 1066-1077.

[16] DUARTE F, MARTINS B, PINTO C S, et al. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text[J]. Journal of Biomedical Informatics, 2018, 80: 64-77.

[17] PRAKASH A, ZHAO S, HASAN S A, et al. Condensed memory networks for clinical diagnostic inferencing[C]// The 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 3274-3280.

[18] BAUMEL T, NASSOUR-KASSIS J, COHEN R, et al. Multi-label classification of patient notes: case study on ICD code assignment[C]// The 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018.

[19] ZENG M, LI M, FEI Z, et al. Automatic ICD-9 coding via deep transfer learning[J]. Neurocomputing, 2019.

[20] CAI H, ZHENG V W, CHANG K C C. A comprehensive survey of graph embedding: problems, techniques, and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.

[21] BATTAGLIA P W, HAMRICK J B, BAPST V, et al. Relational inductive biases, deep learning, and graph networks[J]. arXiv preprint, 2018, arXiv: 1806.01261v3.

[22] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint, 2013, arXiv: 1312.6203v2.

[23] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Advances in Neural Information Processing Systems. [S.l.:s.n.], 2016.

[24] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint, 2016, arXiv: 1609.02907v4.

[25] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]// The 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019.

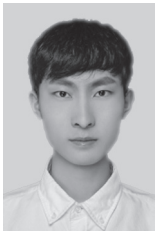
[26] PENG H, LI J, HE Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-CNN[C]// The 2018 World Wide Web Conference. [S.l.:s.n.], 2018: 1063-1072.

[27] RIOS A, KAVULURU R. Few-shot and zero-shot multi-label learning for structured label spaces[C]// The 2018 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2018: 3132-3142.

[28] WANG W, XU H, GAN Z, et al. Graph-driven generative models for heterogeneous multi-task learning[C]// The 34th AAAI Conference on Artificial Intelligence. Palo

Alto: AAAI Press, 2020.
[29] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv: 1301.3781v3.
[30] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint, 2014, arXiv: 1408.5882v2.

作者简介

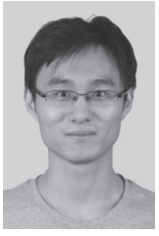


杜逸超（1997- ），男，中国科学技术大学计算机科学与技术学院硕士生，主要研究方向为数据挖掘、知识图谱。

徐童（1988- ），男，博士，中国科学技术大学计算机科学与技术学院副教授，主要研究方向为数据挖掘。

马建辉（1975- ），男，中国科学技术大学计算机科学与技术学院讲师，主要研究方向为数据挖掘。

陈恩红（1968- ），男，博士，中国科学技术大学计算机科学与技术学院教授，主要研究方向为数据挖掘和机器学习。



郑毅（1987- ），男，博士，华为技术有限公司自然语言处理技术专家，主要研究方向为自然语言处理和机器学习。

刘同柱（1967- ），男，博士，中国科学技术大学附属第一医院副研究员，主要研究方向为健康大数据和医院管理。

童贵显（1991- ），男，中国科学技术大学附属第一医院初级经济师，主要研究方向为健康大数据和医院管理。

收稿日期： 2020-07-06

基金项目： 国家自然科学基金资助项目（No.U1605251，No.61703386）；中央高校基本科研业务费专项资金项目（No.WK911000014）；安徽省重点研发计划项目（No.1804b06020377）

Foundation Items: The National Natural Science Foundation of China(No.U1605251, No.61703386), The Fundamental Research Funds for the Central Universities(No.WK911000014), The Key Research and Development Program of Anhui Province (No.1804b06020377)

基因组大数据变异检测算法的并行优化

崔英博[1]，黄春[1]，唐滔[1]，杨灿群[1]，廖湘科[1]，彭绍亮[2,3]
1. 国防科技大学计算机学院，湖南 长沙 410073；2. 湖南大学信息科学与工程学院，湖南 长沙 410082；
3. 国家超级计算长沙中心，湖南 长沙 410082

摘要
序列比对和变异检测是基因组数据分析的基础步骤，是后续各种功能性分析的前提，也是基因组数据分析中最耗时的环节。为有效处理高通量测序技术产生的海量基因组大数据，采用OpenMP、MPI等技术，对序列比对算法和SNP检测算法进行了多级并行优化，并对相关算法进行了改进。在不同数据集和并行规模下的测试中，核心算法加速比达到9倍以上，大规模测试中算法的并行效率保持在60%以上，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的能力。

关键词
序列比对；SNP；OpenMP；MPI

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020041

Parallel optimization of variation detection algorithms for large-scale genome data

CUI Yingbo[1], HUANG Chun[1], TANG Tao[1], YANG Canqun[1], LIAO Xiangke[1], PENG Shaoliang[2]
1. College of Computer, National University of Defense Technology, Changsha 410073, China
2. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
3. National Supercomputer Center in Changsha, Changsha 410082, China

Abstract
Sequence alignment and mutation detection are the basic steps of genomic data analysis. They are the premise of subsequent functional analysis, and the most time-consuming steps. In order to effectively deal with the massive genomic big data brought by high-throughput sequencing technology, MPI, OpenMP and other technologies to perform multi-level parallel optimization of sequence alignment algorithm and SNP detection algorithm were used. By testing on different data sets and parallel scales, the core algorithm reached more than 9x speedup, and the parallel efficiency remained above 60% in large-scale test. The improved algorithms obtain good parallel performance and scalability, that effectively improves the ability of genomic big data mutation detection.

Key words
sequence alignment, SNP, OpenMP, MPI