◆机器学习模型◆

基于多任务学习的国际疾病分类自动编码模型*

张　艺,滕　飞**,胡　节
(西南交通大学计算机与人工智能学院,四川成都　611756)

摘要:国际疾病分类(International Classification of Diseases,ICD)编码任务是将疾病编码分配给电子病历,每份电子病历分配一个或多个 ICD 编码。现有的方法大多考虑临床文本中症状与诊断之间的关系,而对诊断与诊断间关系以及症状与症状间关系缺乏考量。针对这一现状,对于诊断与诊断间关系,构造编码共现任务,采用多任务的形式使得预测结果不依赖于标签之间的顺序关系,且不会进行错误预测的传播;对于症状与症状间关系,使用对比学习获取有意义的表征,学习同一临床文本中的症状一致性。通过以上任务的组合,构建基于多任务学习的 ICD 自动编码模型框架。在 MIMIC-Ⅲ数据集上的实验表明,所提出的方法相较于优异模型在 Micro-f1 指标上提高了 1.0%,在 Micro-auc 指标上提高了 0.3%,在 P@5 指标上提高了 0.7%。

关键词:ICD 编码;多任务学习;编码共现;对比学习;自然语言处理

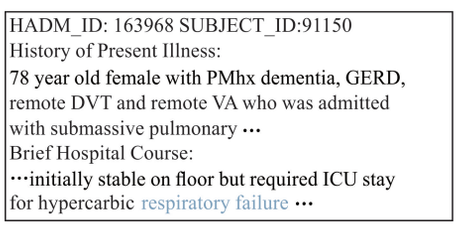
中图分类号:TP389.1　　**文献标识码**:A　　**文章编号**:1005-9164(2023)01-0114-07
DOI:10.13656/j.cnki.gxkx.20230308.013

国际疾病分类(International Classification of Diseases,ICD)提供疾病描述的诊断编码,其编码包含了关于流行病学、健康损害和治疗条件的关键信息。ICD 编码任务是将一个或多个疾病编码分配给电子病历。ICD 编码可用于临床研究和医疗保健等,如流行病学研究和服务计费[1]。ICD 编码通常依赖临床编码员进行基于经验的判断和人工检查,这往往费时并且容易出错。因此,近年来研究者开始使用深度学习技术进行 ICD 自动编码。而现有的一些研究[2-4]大多考虑临床文本中症状与诊断之间的关系,对诊断与诊断间关系以及症状与症状间关系缺乏考量。

在 ICD 自动编码任务中,每个给定的临床文本都与一组给定的 ICD 编码相关联,这些编码通常在统计和语义上相关,利用编码之间的共生关系可以构建具有较强泛化性能的多标签分类模型。与此同时,现实生活中一些疾病通常是并发的或相互之间存在因果关系[3],它们的编码在临床文本中通常同时出现,这种共现关系能够捕获编码间的相关性。对于一些仅用临床文本难以预测的编码,可以利用编码共现

收稿日期:2022-09-30　　修回日期:2022-12-05
*四川省国际科技创新合作项目(2022YFH0020)和四川省重点研发项目(2021YFG0136)资助。
【第一作者简介】
张　艺(1998-),女,在读硕士研究生,主要从事医学信息学研究。
【**通信作者】
滕　飞(1984-),女,副教授,主要从事医学信息学研究,E-mail:fteng@swjtu.edu.cn。
【引用本文】
张艺,滕飞,胡节.基于多任务学习的国际疾病分类自动编码模型[J].广西科学,2023,30(1):114-120.
ZHANG Y,TENG F,HU J. Auto-encoding Model Based on Multi-Task Learning for International Classification of Diseases [J]. Guangxi Sciences,2023,30(1):114-120.

对其进行正确预测。图1给出了一份MIMIC-Ⅲ数据集中的编码示例。图左侧为电子病历，右侧为其对应的ICD-9编码。示例中编码“518.81急性呼吸衰竭”能够从医疗记录文本中捕捉到显式语义信息，但是编码“496气道阻塞”却无法从文本中获取直接的相关语义。因为“气道阻塞”是引起“急性呼吸衰竭”的常见病因，所以可以通过编码的共现关系对那些难以从临床文本中获取直接语义的编码进行正确预测。



ICD-9 codes	Descriptions
276.2	Acidosis
496	Chronic airway obstruction, not elsewhere classified
518.81	Acute respiratory failure
530.81	Esophageal reflux

图1　MIMIC-Ⅲ数据集编码示例

Fig.1　Coding example of MIMIC-Ⅲ dataset

编码共现是对诊断与诊断间的相关性进行学习。Hatanaka等[5]使用基于seq2seq的方式捕捉编码之间的关系。然而基于seq2seq的方式依赖于预定义的标签顺序以及先前不正确的预测结果，并可能对频繁出现的标签序列过度拟合[6]，而ICD编码任务与标签顺序无关。基于此，本研究提出基于多任务学习的方法，该方法不依赖于标签之间的顺序关系也不会进行错误预测的传播。编码之间的共生关系能够反映出标签的相关性，通过编码共现任务来辅助ICD自动编码。编码共现任务的构建直接使用原始训练数据，通过二元分类器判断成对标签是否同时出现在相关标签集中。

对比学习是用于构建有意义表征的一种有效方法[7]，其目的是使相似的样本对彼此接近，不同的样本对彼此拉远。考虑到同一临床文本中症状与症状间存在一致性，使用对比学习来学习这种特性。在对比学习中构建高质量的样本对至关重要。由于临床文本通常篇幅长且存在冗余信息，一些先进的方法提出标签式注意力机制，使用编码描述作为查询来提取与编码相关的表示[2]。对临床文本中的各个词而言，其对最终编码预测结果的重要性是不同的。标签式注意力机制能够计算出每个词对每个标签的注意力权值。保留注意力权值大于阈值的词，对其余词进行遮蔽，将得到对应病历文本的主序列表示。输入序列和它们对应的主序列构成正样本，与同批次其他示例构成负样本。依据对比学习的思想，原始序列及其对应的主序列在表示空间中应当彼此接近，而与其他序列应该相距更远。主序列的多标签分类损失将用于主序列生成指导。

本研究拟通过构建编码共现任务，预测难以从医疗文本获取直接语义的编码，通过对比学习获取医疗文本有意义的表征，最终以多任务学习思想整合医疗文本中症状与症状间、诊断与诊断间的相关关系，提高ICD编码精度。

1　相关工作

ICD自动编码是医学自然语言处理领域一个活跃的研究课题，已有超过20年的研究历史[8]。早期一些传统的机器学习算法[9-11]被应用于ICD自动编码。Perotte等[9]使用词频-逆文档频率(Term Frequency-Inverse Document Frequency，TF-IDF)得到的关键词作为支持向量机(Support Vector Machine，SVM)分类器特征，并将ICD-9的层次结构考虑在内。与之相似的，Koopman等[10]研究了在死亡证书中与癌症相关的编码，并使用考虑层次结构的SVM进行编码分配。Ruch等[11]使用K最近邻(K-Nearest Neighbor，KNN)算法作为分类器，并证实使用既往病史和处方能够在统计学上有显著的改善。然而，这些方法往往需要手动进行数据处理及验证等工作。

随着深度神经网络的发展，出现了许多新方法。Mullenbach等[2]使用卷积神经网络(Convolutional Neural Network，CNN)提取文档语义，并使用注意力机制为每个可能的编码选择最相关的段。Li等[12]提出一种结合了多滤波器卷积层和残差卷积层的模型来改进卷积注意力模型。Teng等[13]将知识图谱和注意力机制扩展到ICD编码预测中，以此来提高预测结果的可解释性。Vu等[14]利用编码间的层次关系扩展了标签式注意力模型并使用联合学习机制提高罕见编码的性能。Yuan等[15]提出多个同义词匹配网络，利用同义词对编码进行更好的表示学习。杜逸超等[16]采用多尺度残差网络捕获不同长度的文

本模式并使用图卷积抽取标签层次关系。王天罡等[17]使用预训练语言模型提升了电子病历文本的特征表达能力。这些方法考虑的是症状与最终诊断之间的关系。

一些研究将诊断与诊断间的关系考虑在内。Cao 等[3]设计了一个图来捕捉编码之间的共生相关性。图的构建规则为临床文本中同时出现的两个编码之间有一条边,边的权重越高,表明编码间的依赖性越强。Schäfer 等[4]使用 apriori 算法查找关联规则并标记尸检报告中的共现。Duarte 等[18]应用非负矩阵分解产生的分解分量来利用具有树状结构的类别之间的共现。与以前的工作不同,本研究通过从原始训练数据中构建标签共现相关的二分类任务,学习诊断与诊断间的关系,用于辅助 ICD 自动编码任务。

2 方法

图 2 展示了多任务自动编码(Multi-Task Automatic Encoding,MTAE)模型的基本框架,该模型用于给患者的临床记录分配 ICD 编码。ICD 自动编码被视为一个多标签文本分类问题[19],$L=\{l_1,l_2,\cdots,l_n\}$ 为所有的 ICD 编码的集合,其中 n 表示 ICD 编码的数量。最终目标是训练 n 个二元分类器,对于输入文本,每个分类器输出预测结果 $y_i\in\{0,1\}$,其中 y_i 代表 L 中的第 i 个标签。

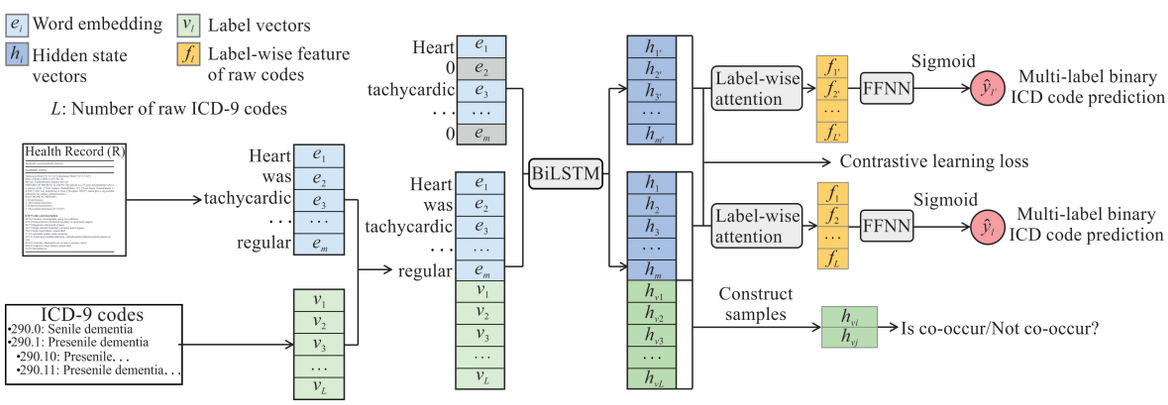


图 2 MTAE 模型框架

Fig. 2 Framework of MTAE model

2.1 原始多标签分类任务

2.1.1 嵌入层

给定一个包含 m 个单词的临床输入文本 R,使用矩阵 $R=[x_1,x_2,\cdots,x_m]$ 来表示。对于文本中每个单词 x_i,使用预训练词嵌入的方法得到维度为 d_e 的词嵌入向量 e_i。最后得到输入文本的词嵌入向量表示 $D=[e_1,e_2,\cdots,e_m]$。

2.1.2 文本编码器

对于给定的输入数据,使用双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)从词嵌入中学习文本中的语义信息。第 i 个词的隐藏状态计算公式为

$$\overrightarrow{h_i}=\overrightarrow{\mathrm{LSTM}}(e_1:e_i)\ , \tag{1}$$

$$\overleftarrow{h_i}=\overleftarrow{\mathrm{LSTM}}(e_i:e_m)\ 。 \tag{2}$$

最终的潜在向量 h_i 由 $\overrightarrow{h_i}$ 和 $\overleftarrow{h_i}$ 连接而成:

$$h_i=\overrightarrow{h_i}\oplus\overleftarrow{h_i}\ 。 \tag{3}$$

长短期记忆网络(Long Short-Term Memory,LSTM)的隐藏状态维度被设置为 u,因此最终潜在向量 h_i 的维度是 $2u$。文本 R 中所有的隐藏状态向量连接起来形成矩阵 $H=[h_1,h_2,\cdots,h_m]\in\mathbb{R}^{2u\times m}$。

2.1.3 标签向量

每个 ICD 编码 l 都有一个编码描述 l_d 与之对应,例如:ICD-9 中编码 401.9 对应的编码描述是“未特指的原发性高血压”。为了表示编码描述 l_d,需要对每个编码描述进行预处理。首先将所有单词转换为小写并删除停用词,然后通过平均剩余单词的词嵌入向量来形成标签向量 $v_l\in\mathbb{R}^{d_e}$。将标签向量 v_l 与输入文本的词嵌入向量 D 拼接起来经过编码器 BiLSTM 得到标签最终表示 h_{v_l}。

2.1.4 注意力机制

由于临床文本通常对应多个 ICD 编码,不同的 ICD 编码对应的关键短语不一致。针对这个问题,采用标签式注意力机制,使得模型可以关注文本的不同部分。标签式注意力机制由以下公式计算:

$$s_l=\mathrm{softmax}(\tanh(HW_a^{\mathrm{T}}+b_a)v_l)\ , \tag{4}$$

其中,s_l 是 H 中所有行的注意力分数,H 表示文本特征,W_a 和 b_a 为训练过程中的参数。

$$a_l = s_l^{\mathrm{T}} H, \tag{5}$$

其中,a_l 是 H 中与编码 l 最相关的信息。对于每个 ICD 编码总共有 L 个注意力向量。通过下式得到文本表示:

$$f_l = \mathrm{ReLU}(w_l a_l + b_l), \tag{6}$$

其中,w_l 表示编码 l 的权重,b_l 表示编码 l 的偏置。

2.1.5 输出层

多标签文本分类器为每一个标签生成二进制预测结果:

$$\hat{y}_l = \mathrm{Sigmoid}(w f_l + b), \tag{7}$$

其中,w 表示权重,b 表示偏置。

使用二元交叉熵损失函数 L_c 来训练多标签文本分类模型:

$$L_c(y,\hat{y}) = -\sum_{l=1}^{L}[y_l \log(\hat{y}_l) + (1-y_l)\log(1-\hat{y}_l)], \tag{8}$$

其中,L 表示标签总数,y_l 表示真实标签,$\hat{y}_l$ 表示预测标签。

2.2 编码共现任务

每份临床文本都包含相关的标签集 $L+$ 和不相关的标签集 $L-$。标签共现任务数据可以从 MIMIC-Ⅲ训练数据中直接构建。其中,共现标签对的采样规则为仅从临床文本相关的标签集 $L+$ 中采样;不共现标签对的采样规则为一部分从临床文本相关的标签集 $L+$ 中采样,另一部分从临床文本不相关的标签集 $L-$ 中采样。将采样得到的两个标签分别对应的表示向量 h_{v_i} 和 h_{v_j} 一起嵌入作为输入特征,添加二元分类器用于预测两个标签的状态为“共现”或者“不共现”。下式为二元分类的损失函数:

$$L_{\mathrm{co}} = -[q_{ij}\ln p_{ij} + (1-q_{ij})\ln(1-p_{ij})], \tag{9}$$

其中,$p_{ij} = p(l_j \mid R, l_i)$ 表示标签对共现的输出概率,q 是真实标签,$q_{ij}=1$ 表示共现,$q_{ij}=0$ 表示不共现。

2.3 对比学习任务

使用输入序列和它们对应的主序列来构造对比学习的正样本。主序列的生成规则:计算文本中各词嵌入对每个 ICD 编码的注意力权值,将注意力权值大于阈值 λ 的词保留下来,小于阈值 λ 的词使用 0 填充。主序列构造策略为

$$\hat{e} = \{e_i \text{ if } P_i > \lambda \text{ else } 0\}, \tag{10}$$

其中,P_i 表示文本中某词的注意力权值,λ 为设定的阈值。

主序列 $\hat{e}$ 也被送入相同的文本编码器 BiLSTM,得到 $\hat{H} = [\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_m] \in \mathbb{R}^{2u\times m}$。考虑到同一临床文本中疾病的一致性,输入序列与其对应的主序列经过文本编码器之后得到的表示应该是相似的。其他来自不同对的例子在表示空间中应该相距较远。

对于一个正对$(H, \hat{H})$,在其上添加非线性层:

$$g_i = W_2 \mathrm{ReLU}(W_1 H), \tag{11}$$

$$\hat{g}_i = W_2 \mathrm{ReLU}(W_1 \hat{H}), \tag{12}$$

其中,W_1 和 W_2 为训练过程中的参数。

对一批次中 $2N$ 个例子,每一个序列有 $2(N-1)$个负对,示例集合表示如下:

$$Z = \{z \in \{g_i\} \cup \{\hat{g}_i\}\}。 \tag{13}$$

对于其中某一示例 z_i,计算其 NT-Xent 损失[20]:

$$L_i^{\mathrm{con}} = -\log \frac{\exp(\mathrm{sim}(z_i, u(z_i))/t)}{\sum_{j,j\neq i}^{2N}\exp(\mathrm{sim}(z_i, z_j)/t)}, \tag{14}$$

其中,sim 表示余弦相似函数,t 是温度系数,u 函数为匹配函数,将其表示如下:

$$u(z_i) = \begin{cases} g_i, \text{if } z_i = \hat{g}_i \\ \hat{g}_i, \text{if } z_i = g_i \end{cases}。 \tag{15}$$

总的对比损失是所有例子计算得到的平均损失:

$$L_{\mathrm{con}} = \frac{1}{N}\sum_{i=1}^{2N} L_i^{\mathrm{con}}。 \tag{16}$$

使用主序列的多标签分类损失用于主序列生成指导,让生成的主序列与原始序列对应相同的预测标签。

2.4 目标函数

最终的损失函数由原序列的分类损失 L_c、对应主序列的分类损失 $L_{c'}$、编码共现模块对应的二分类损失 L_{co} 及对比损失 L_{con} 组合而得:

$$L = L_c + L_{c'} + L_{\mathrm{co}} + L_{\mathrm{con}}。 \tag{16}$$

3 实验与结果分析

3.1 数据集

MIMIC-Ⅲ是一个可开放获取的包含医院 ICU 文本和结构化记录的数据集。依据患者 id 拆分出院小结,并使用 top-50 编码进行实验。为便于结果对比,本研究按照 Mullenbach 等[2]的方法进行数据集划分,得到 8 066 个出院小结用于训练,1 573 个用于验证,1 729 个用于测试。对文本的预处理方式如下:进行分词处理,将所有标记转化为小写,并将文本最大截断长度设置为 4 000。

3.2 基线模型

CAML[2]:使用 CNN 对文本进行编码,并使用

标签式注意力机制增强模型的可解释性。

MultiResCNN[12]:使用多滤波器残差卷积神经网络对文本进行编码,获取深层且多样化的文本表征。

DCAN[21]:一种膨胀卷积注意力网络,其增强了感受野。

MSATT-KG[22]:采用多尺度语义特征及编码间的关系进行 ICD 自动编码。

G_Coder[13]:将知识图谱和对抗性学习扩展到医学编码预测中。

JointLAAT[14]:使用分层联合学习机制来提高低频 ICD 编码的性能。

3.3 实验设置

本实验超参数的设置如下:词嵌入的维度(d_e)为 100;LSTM 隐藏层状态维数(u)为 256;学习率(l_r)设置为 0.001 2;批尺寸(batch_size)设置为 16;对比学习模块中阈值(λ)在{0.06,0.13,0.26,0.5}中选择,为 0.26;温度系数(t)设置为 1,这是由于初始实验设置为该值时取得了不错的结果。

3.4 结果与分析

表 1 展示了在 MIMIC-Ⅲ top-50 测试集上的主要结果。在 Micro-f1、Micro-auc、P@5 指标上与优异模型 JointLAAT 相比,MTAE 模型在 Micro-f1 指标上提高了 1.0%,在 Micro-auc 指标上提高了 0.3%,在 P@5 指标上提高了 0.7%。实验结果证实了多任务学习模块的有效性。编码共现任务以及主序列生成策略对 ICD 编码的最终预测产生了积极的作用。编码共现通过捕获编码间的相关性来预测那些难以预测的编码。主序列生成策略维持同一病历的疾病一致性,使得同类数据经过编码器的编码结果尽可能相似。

表 1 在 MIMIC-Ⅲ top-50 测试集上的结果

Table 1 Results on MIMIC-Ⅲ top-50 test set

模型 Model	Micro-f1 (%)	Micro-auc (%)	P@5 (%)
CAML[2]	61.4	90.9	60.9
MultiResCNN[12]	67.0	92.8	64.1
DCAN[21]	67.1	93.1	64.2
MSATT-KG[22]	68.4	93.6	64.4
G_Coder[13]	69.2	93.3	65.3
JointLAAT[14]	71.6	94.4	67.3
MTAE	**72.6**	**94.7**	**68.0**

Note:bold indicates the optimal value of each indicator

表 2 给出了消融实验的结果。在不改变其他模块的情况下从完整模型中删除一个模块,并在表格中使用 no X 表示这样的基线。为了评估它们,对以下两种配置进行对比。①无编码共现任务模块:将编码共现指导的二分类任务从模型中移除。②无对比学习模块:将主序列生成及与之相关的多标签分类任务从模型中移除。可以观察到去除两个子任务对 MIMIC-Ⅲ数据集上的主要指标有一定程度的影响。编码共现模块用于学习诊断与诊断间的关系,通过学习诊断间的相关性,模型可对一些难以预测的疾病进行正确的预测,提高了模型的泛化能力。与完整模型相比,去除编码共现模块后,模型在 Micro-f1 指标上降低 0.2%,在 P@5 指标上降低 0.5%。对比学习模块用于学习症状与症状间关系,学习同类病历之间的共同特征,区分非同类病历之间的不同之处。通过维持同一病历原始序列与主序列文本的症状一致性,构建有意义的文本表征。与完整模型相比,去除对比学习模块后,模型在 Micro-f1 指标上降低 0.5%,在 Micro-auc 指标上降低 0.1%,在 P@5 指标上降低 0.7%。

表 2 消融实验结果

Table 2 Results of ablation experiment

方法 Method	Micro-f1 (%)	Micro-auc (%)	P@5 (%)
MTAE	**72.6**	94.7	**68.0**
no co-occurrence	72.4	94.7	67.5
no contrastive learning	72.1	94.6	67.3

Note:bold indicates the optimal value of each indicator

4 结论

本研究提出了一种基于多任务学习的 MTAE 模型用于 ICD 自动编码。除症状与诊断之间的关系之外,多任务学习框架还考虑了诊断与诊断间关系以及症状与症状间关系。首先通过构建编码共现任务,利用标签之间的相关性学习诊断与诊断间关系,辅助 ICD 自动编码,提高多标签分类模型的泛化能力。然后使用对比学习思想,在学习症状与症状间一致性的同时,获得主序列的表征。最后采用多任务学习思想整合症状与诊断、诊断与诊断以及症状与症状间的相关关系,通过各任务知识组合进行 ICD 自动编码。在 MIMIC-Ⅲ数据集上的实验表明,MTAE 模型的表现具有优越性。

参考文献

[1] NGUYEN A N,TRURAN D,KEMP M,et al. Computer-assisted diagnostic coding:effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings [C/OL]//AMIA Annual Symposium Proceedings. Online:AMIA,2018,2018:807-816 [2022-09-10]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371260.

[2] MULLENBACH J,WIEGREFFE S,DUKE J,et al. Explainable prediction of medical codes from clinical text [Z/OL]. (2018-02-15)[2022-09-10]. https://arxiv.org/abs/1802.05695.

[3] CAO P F,CHEN Y B,LIU K,et al. HyperCore:hyperbolic and co-graph representation for automatic ICD coding [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online:Association for Computational Linguistics,2020:3105-3114 [2022-09-10]. https://aclanthology.org/2020.acl-main.282.

[4] SCHÄFER H,FRIEDRICH C M. Multilingual ICD-10 code assignment with transformer architectures using MIMIC-Ⅲ discharge summaries [C/OL]//CAPPELLATO L,EICKHOFF C,FERRO N,et al. Working Notes of CLEF 2020-Conference and Labs of the Evaluation Forum. Online:CLEF,2020,2696:212 [2022-09-10]. https://ceur-ws.org/Vol-2696/paper_212.pdf.

[5] HATANAKA H,AKIBA T. Using a sequence-to-sequence model for large-scale automated ICD coding [C]//2019 International Conference of Advanced Informatics:Concepts,Theory and Applications (ICAICTA). Yogyakarta,Indonesia:IEEE,2019:1-6. DOI:10.1109/ICAICTA.2019.8904397.

[6] ZHANG X M,ZHANG Q W,YAN Z,et al. Enhancing label correlation feedback in multi-label text classification via multi-task learning [Z/OL]. (2021-07-06)[2022-09-10]. https://doi.org/10.48550/arXiv.2106.03103.

[7] KIM T,YOO K M,LEE S. Self-guided contrastive learning for BERT sentence representations [C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:Long Papers). Online:Association for Computational Linguistics,2021:2528-2540 [2022-09-10]. https://aclanthology.org/2021.acl-long.

[8] LARKEY L S,CROFT W B. Combining classifiers in text categorization [C]//Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York,USA:Association for Computing Machinery,1996:289-297.

[9] 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.

[10] 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.

[11] RUCH P,GOBEILL J,TBAHRITI I,et al. From episodes of care to diagnosis codes:automatic text categorization for medico-economic encoding [C/OL]//AMIA Annual Symposium Proceedings. Online:AMIA,2008,2008:636-640 [2022-09-10]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655971.

[12] LI F,YU H. ICD coding from clinical text using multi-filter residual convolutional neural network [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto,California,USA:AAAI Press,2020,34(5):8180-8187.

[13] TENG F,YANG W,CHEN L,et al. Explainable prediction of medical codes with knowledge graphs [J]. Frontiers in Bioengineering and Biotechnology,2020,8:867-878.

[14] VU T,NGUYEN D Q,NGUYEN A. A label attention model for icd coding from clinical text [Z/OL]. (2020-07-13)[2022-09-10]. https://doi.org/10.48550/arXiv.2007.06351.

[15] YUAN Z,TAN C,HUANG S. Code synonyms do matter:multiple synonyms matching network for automatic ICD coding [Z/OL]. (2022-03-03)[2022-09-10]. https://arxiv.org/abs/2203.01515.

[16] 杜逸超,徐童,马建辉,等.一种基于深度神经网络的临床记录ICD自动编码方法[J].大数据,2020(5):1-15.

[17] 王天罡,李晓亮,张晓滨,等.基于预训练表征模型的自动ICD编码[J].中国数字医学,2020,15(7):53-56.

[18] 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.

[19] MCCALLUM A K. Multi-label text classification with a mixture model trained by EM [C]//Working Notes of the AAAI'99 Workshop on Text Learning. California,USA:AAAI Press,1999:201-208.

[20] CHEN T,KORNBLITH S,NOROUZI M,et al. A simple framework for contrastive learning of visual representations [Z/OL]. (2020-02-13)[2022-09-10]. https://doi.org/10.48550/arXiv.2002.05709.

[21] JI S X,CAMBRIA E,MARTTINEN P. Dilated convolutional attention network for medical code assignment from clinical text [Z/OL]. (2020-09-30)[2022-09-10]. https://doi.org/10.48550/arXiv.2009.14578.

[22] XIE X C,XIONG Y,YU P S,et al. EHR coding with multi-scale feature attention and structured knowledge graph propagation [C]//CKIM'19:Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York,USA:Association for Computing Machinery,2019:649-658.

Auto-encoding Model Based on Multi-Task Learning for International Classification of Diseases (ICD)

ZHANG Yi,TENG Fei**,HU Jie

(School of Computer and Artificial Intelligence,Southwest Jiaotong University,Chengdu,Sichuan,611756,China)

Abstract:The task of International Classification of Diseases(ICD) automatic coding is to assign disease codes to electronic medical records,and each electronic medical record is assigned one or more ICD codes. Most of the existing methods consider the relationship between symptoms and diagnosis in clinical texts,while the relationship between diagnosis and diagnosis and the relationship between symptoms and symptoms are not considered. In view of this situation,for the relationship between diagnosis and diagnosis,the coding co-occurrence task is constructed,and the multi-task form is used to make the prediction result independent of the sequential relationship between labels,and the propagation of error prediction will not be carried out. For the relationship between symptoms and symptoms,the comparative learning is used to obtain meaningful representations and learn the consistency of symptoms in the same clinical text. Through the combination of the above tasks,the framework of ICD automatic coding model based on multi-task learning is constructed. The experiment on the MIMIC-Ⅲ dataset shows that the proposed method has improved the Micro-f1 index by 1.0%,the Micro-auc index by 0.3%,and the P@5 index by 0.7% compared with the excellent model.

Key words:ICD encoding;multi-task learning;code co-occurrence;contrastive learning;natural language processing

责任编辑:梁　晓