

基于电子病历时间嵌入的编码器稠密网络冠心病预测模型

陈艳章¹, 韩文静², 刘增光³

(1. 山东环球软件股份有限公司; 2. 潍坊科技学院 艺术学院, 山东 寿光 262700;
3. 山东科技职业学院 信息工程系, 山东 潍坊 261053)

摘要: 冠心病早期发现可显著提高治愈率, 但受限于其前驱症状不明显、患者自我认识不足等原因, 冠心病早期发现率较低。通过构建基于电子病历时间嵌入的编码器稠密网络冠心病预测模型E-TEED-CHD, 可以有效使用居民从出生起跨区域、跨机构、跨病种的历次非结构化电子病历数据进行冠心病早期预测。该模型将居民历次电子病历分别嵌入时间位信息, 并行输入多个编码器进行加速处理; 采用稠密网络模型对编码器输出进行解析, 实现冠心病的早期预测。进一步地, E-TEED-CHD模型使用一种新颖的病历掩膜技术来忽略电子病历间的数量差异, 以精简的自身结构实现不同数量电子病历的统一处理。消融实验证明了时间嵌入方法及编码器稠密网络模型的有效性, 同时评估实验表明, 经数据预处理、超参数调整后E-TEED-CHD模型的冠心病早期预测准确率为98.71%, 千名居民电子病历单轮平均训练时间为17.752 s。E-TEED-CHD模型的准确率及训练速度均优于其他先进模型。

关键词: 电子病历; 时间嵌入; 稠密网络; 编码器; 冠心病预测

DOI: 10.11907/rjtk.241496

开放科学(资源服务)标识码(OSID):



中图分类号: TP181

文献标识码: A

文章编号: 1672-7800(2025)009-0048-07

EMR-Based Time Embedding Encode-Dense Network Model for Coronary Heart Disease Prediction

CHEN Yanzhang¹, HAN Wenjing², LIU Zengguang³

(1. Univalsoft Joint Stock Co. Ltd.; 2. School of Art, Weifang University of Science and Technology, Shouguang 262700, China;
3. School of Information Engineering, Shandong Vocational College of Science and Technology, Weifang 261053, China)

Abstract: Early detection of coronary heart disease can significantly improve the cure rate, but the early detection rate of coronary heart disease is low due to the inconspicuous prodromal symptoms and lack of self-awareness of patients. By constructing the E-TEED-CHD early prediction model based on the EMR-based time embedding encode-dense network, it is possible to effectively use the non-structured electronic medical record data from residents across regions, institutions, and diseases for early prediction of coronary heart disease. The E-TEED-CHD model embeds the residents' electronic medical records with time position information and parallelly inputs them into multiple encoders for accelerated processing. A dense network model is used to parse the outputs of the encoders and achieve early prediction of coronary heart disease. Furthermore, the E-TEED-CHD model uses a novel medical record masking technique to ignore the quantity differences of electronic medical records among patients, achieving unified processing of different numbers of EMRs with a streamlined structure. Ablation studies demonstrate the effectiveness of the time embedding method and the encode-dense network model. Evaluation experiments show that the early coronary heart disease prediction accuracy of the E-TEED-CHD model, after data preprocessing and hyperparameter tuning, is 98.71%. And the average training time for electronic medical records of 1,000 residents is 17.752 seconds. Both accuracy and training speed of E-TEED-CHD model outperform the other state-of-the-art models.

Key Words: electronic medical record; time embedding; dense network; encoder; coronary heart disease prediction

收稿日期: 2024-05-29

扫描二维码阅读全文:



基金项目: 潍坊市科学技术发展计划项目(2023GX063)

作者简介: 陈艳章(1984-), 男, 山东环球软件股份有限公司总裁, 研究方向为软件项目管理、智慧医疗; 韩文静(1982-), 女, 博士, 潍坊科技学院艺术学院讲师, 研究方向为虚拟现实; 刘增光(1982-), 男, 博士, 山东科技职业学院信息工程系副教授, 研究方向为深度学习、智慧医疗。本文通讯作者: 刘增光。

0 引言

冠心病是由冠状动脉粥样硬化斑块积聚引起的一种心血管疾病。该疾病导致流向心肌的血流量减少,临床表现出胸痛或心绞痛、呼吸短促、心悸和心力衰竭等症状。在严重的情况下,冠心病可能会引起心脏病发作,导致心肌的永久性损伤,对个人的生活质量产生深远影响。据全球冠心病负担报告显示,全球估计每年有914万人死于冠心病,现有冠心病患者为1.97亿人^[1]。研究表明,冠心病早期发现可显著提高治愈率^[2]。但由于冠心病前驱症状不明显、患者自我认识不足等原因,致使冠心病早期发现率较低。因此,冠心病早期发现已成为全球公共卫生系统面对的棘手难题和关键挑战。

随着医院健康信息系统(Health Information System, HIS)及基于国家统一互联互通标准的卫生管理机构区域卫生系统的建立,实现了医疗机构间电子病历(Electronic Medical Record, EMR)的共享和交换,从而建立起居民全生命周期的电子病历信息体系^[3]。该电子病历信息体系包含居民从出生起跨区域、跨机构、跨病种的历次诊疗记录、体检记录、健康服务记录等,为开展冠心病早期预测与研究提供了强有力的数据支持,最终可为居民提供精准的个性化冠心病医疗服务和健康管理方案。因此,通过深度学习等技术,根据非冠心病专项的大量非结构化诊疗、体检数据,提取出关于冠心病早期症状的隐藏知识,开发出冠心病预测模型,及早预测冠心病发生的风险指数,对于提高人民健康水平、保障国民安全具有重要意义。因此,本文针对居民不同病种诊疗期间形成的多份非结构化电子病历,提出基于电子病历时间嵌入的编码器稠密网络冠心病预测模型,以解决冠心病早期发现率较低的问题。

1 相关工作

冠心病是一种危险的疾病,不能完全治愈,因此准确地发现早期冠状动脉疾病可以帮助医生治疗患者。早在上世纪,研究人员即已开展相关研究。在20世纪90年代,Shen等^[4]设计了基于多层感知器网络的判定系统,根据用户问卷调查表的数据判定高危个体。实验表明,该系统对有症状和无症状的患者均具有检测能力。但受限于当时的计算机性能及数据采集难度,该系统并未开展临床应用。Krstacic等^[5]提出基于兴趣子群的归纳机器学习模型,实现对冠心病危险因子的对比统计分析以及对结果的解释,从而诱导模型描述冠心病患病率异常高的人群。Lim等^[6]克服现有深度信念网(Deep Belief Net, DBN)无全局最优层数和节点数的缺点,应用遗传算法提出一种更先进的深度信念网络冠心病风险预测模型Optimized-DBN,该模型在特异性、敏感度方面均优于同期成果。Zhang

等^[7]提出堆叠多个分类器的集成机器学习模型,该模型使用64个二维斑点追踪超声心动图特征和7个临床特征来预测冠心病。借助于多分类器的优势,冠心病分类准确度可提高到70%以上。Huang等^[8]采用GA-PSO混合元启发算法,并建立SVM参数优化预测模型来预测冠心病。黄敏等^[9]构建冠心病预测模型TPE_XGBoost,并采用贝叶斯优化算法优化模型参数,对各特征进行分析解释,获得冠心病的高致病因素。Doppala等^[10]基于遗传算法和径向基函数构建混合模型GA-RBF,并采用特征选择机制仅优选14个关键属性,从而提高冠状动脉疾病检测的准确性和效率。Yang等^[11]提出冠心病预测模型HY_OptGBM,该模型基于焦点损失改进了损失函数,并采用Optuna超参数优化框架优化预测模型的超参数,使结果的精确率、查全率、F评分等方面都得到提升。机器学习的分类方法虽然取得了一定成功,但上述算法均基于结构化的常规体检数据,试图解释冠心病与某种属性之间的相关性,对临床数据的关注程度相对较少。

随着电子病历的出现和深度学习技术的发展,利用深度学习模型分析电子病历等数据获得了越来越多研究者的关注。Jahmunah等^[12]基于卷积神经网络开发GaborCNN模型,该模型使用心电图信号诊断冠心病、心肌梗死和心力衰竭等疾病,具有良好的性能和较低的计算复杂度。Yang等^[13]提出一种新型的两通道混合卷积网络THC-Net,该模型包含经典的相关分析—主成分分析卷积网络、独立成分分析—主成分分析卷积网络,从心电图信号中提取包含导联相关和导联特定信息的深度特征,预测效率和准确率更高。Han等^[14]基于VGG19全卷积网络与U-Net网络,诊断冠心病和心绞痛发病概率。Yang等^[15]提出多模态数据融合方法M-US-EMRs,将超声图像与电子病历(EMR)相结合用于自动筛查冠心病(Coronary Heart Disease, CHD),具有成本低、对人体无害、易于实施等优点。Zhang等^[16]基于生物网络的深度学习方法构建冠心病相关基因模型,通过基因之间的关联信息和表达特点预测冠心病。Wiharto等^[17]利用遗传算法和支持向量机进行特征选择,并利用深度神经网络进行冠心病诊断决策,具有较低的成本和较高的分类性能。Fajri等^[18]结合蜂群优化算法与Q-learning强化学习算法进行特征选择,以改善冠心病预测效果。倪凌^[19]基于长短期记忆网络设计电子病历分析诊断系统(Data Mining for EMR of Hospital, DM-EMR),该系统结合注意力机制对电子病历的上下文语义进行分析,识别病情并追踪诊疗进度。蒋林甫等^[20]提出一种基于注意力可解释表格学习神经网络的冠心病预测框架PCHD-TabNet,同时使用自监督学习帮助模型加速收敛,并保持稳定性。

然而,以上方法都是基于居民自我认识到可能存在心血管疾病后的检查、检验数据,并未将居民历史病历中的患者主诉数据及历次健康体检数据作为输入特征。区域

卫生系统的建立为自动解释和分析医疗系统中存档的居民全生命周期电子病历奠定了基础。本文提出一种基于电子病历时间嵌入的编码器稠密网络冠心病预测模型(EMR-based Time Embedding Encode-Dense Network Model for Coronary Heart Disease Prediction, E-TEED-CHD),其主要贡献及创新之处如下:

(1)E-TEED-CHD模型将居民不同时期的非结构化电子病历嵌入时间位信息,并行输入多个编码器进行处理。同时,采用稠密网络模型对不同编码器的输出进行解析,并实现冠心病的早期预测。

(2)提出一种病历掩膜技术,通过该技术可以忽略居民电子病历的数量差异,实现对不同数量电子病历的统一处理,简化了E-TEED-CHD模型结构。

(3)采用数据预处理、超参数调整以及消融实验等技术获得最佳参数,从而准确预测早期冠心病。

2 E-TEED-CHD 模型介绍

为了更好地预测冠心病,本文提出一种基于电子病历时间嵌入的编码器稠密网络冠心病预测模型(E-TEED-CHD),如图1所示。首先,同一居民不同年月日的EMR经过电子病历掩膜方法形成固定长度的时序系列,并将其非结构化的部分经由嵌入层处理,心电图等图像部分经由VGG19全卷积网络处理;其次,拼接不同网络的输出,导入线性层进行维度变化,并通过 $relu$ 函数进一步引入非线性;再次,通过电子病历时间嵌入方法将时间因素融入待处理向量中,并将序列数据转为并行数据,加速多份EMR处理过程;最后,经多层多头注意力编码器后,将不同年月日的处理结果连接并送入稠密网络中,在保持少量参数情况下进行分类处理,从而提高分类效率。

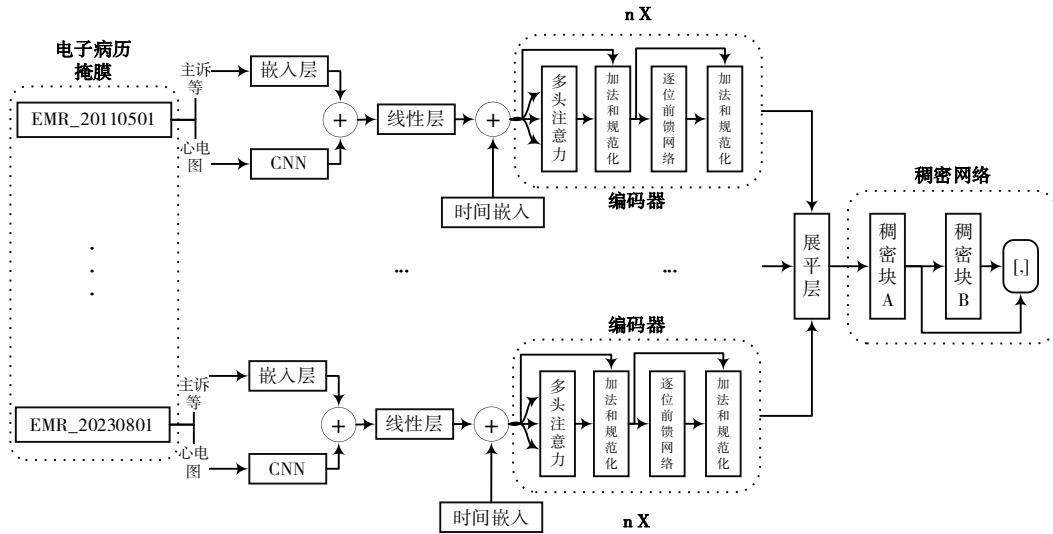


Fig. 1 EMR-based time embedding encode-dense network model for coronary heart disease prediction

图1 基于电子病历时间嵌入的编码器稠密网络冠心病预测模型

2.1 电子病历掩膜方法

每位居民的历史电子病历总数并不一致,而编码器—稠密网络模型要求输入数据(即历史电子病历)具有固定的维度长度。因此,为了把不同居民电子病历放在一起进行处理,可以通过基于 $-\infty$ 的填充方法($-\infty$ -based Padding Method, PAD)补齐电子病历数量,并使用掩膜技术屏蔽掉伪电子病历。

本文在计算注意力之前,将掩膜 M 添加到式(1)中。由于采用电子病历PAD方法,为防止模型错误地认为填充值“ $-\infty$ ”是电子病历的有效信息,本文通过找到 QK 矩阵中PAD位置并进行屏蔽的方法来解决此问题。

$$\text{softmax}(QK^T + M)V \quad (1)$$

假定居民有3份电子病历 $[a, b, c]$,而模型期望的长度为4份,因此可以用 $-\infty$ 将其补齐成为 $[a, b, c, D]$ 。 M 矩阵表示如式(2)所示。在掩码矩阵 M 表示中, D 列是被屏蔽的,下一步将 M 添加到 QK 中。任何数值与 $-\infty$ 相加结果仍

为 $-\infty$,结合式(1)可得式(3)。由式(3)可知,Softmax后的 D 都是0,不会影响到权重矩阵中的其他权重。

$$M = \begin{matrix} & a^K & b^K & c^K & D^K \\ \begin{matrix} a^Q \\ b^Q \\ c^Q \\ d^Q \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & -\infty \end{bmatrix} \end{matrix} \quad (2)$$

$$\text{softmax}(QK^T + M)V = \begin{bmatrix} (a^Q a^K + a^Q b^K + a^Q c^K)s & 0 \\ (b^Q a^K + b^Q b^K + b^Q c^K)s & 0 \\ (c^Q a^K + c^Q b^K + c^Q c^K)s & 0 \\ (D^Q a^K + D^Q b^K + D^Q c^K)s & 0 \end{bmatrix} \begin{bmatrix} a_1^V & a_2^V \\ b_1^V & b_2^V \\ c_1^V & c_2^V \\ D_1^V & D_2^V \end{bmatrix} \quad (3)$$

屏蔽后的 QK^T 矩阵为 $(QK^T + M)$,根据式(3),将其与 V 矩阵相乘,得到的结果为自注意力结果。该结果矩阵如式(4)所示。由式(4)可知,结果矩阵中每一行的 D 分量均为0。因此,该电子病历掩膜方法保留了电子病历的自回归属性,确保训练时仅依赖当前时间之前的真实电子病历。

$$\begin{bmatrix} a^Q a^K a_1^V + a^Q b^K b_1^V + a^Q c^K c_1^V + 0 & a^Q a^K a_2^V + a^Q b^K b_2^V + a^Q c^K c_2^V + 0 \\ b^Q a^K a_1^V + b^Q b^K b_1^V + b^Q c^K c_1^V + 0 & b^Q a^K a_2^V + b^Q b^K b_2^V + b^Q c^K c_2^V + 0 \\ c^Q a^K a_1^V + c^Q b^K b_1^V + c^Q c^K c_1^V + 0 & c^Q a^K a_2^V + c^Q b^K b_2^V + c^Q c^K c_2^V + 0 \\ D^Q a^K a_1^V + D^Q b^K b_1^V + D^Q c^K c_1^V + 0 & D^Q a^K a_2^V + D^Q b^K b_2^V + D^Q c^K c_2^V + 0 \end{bmatrix} \quad (4)$$

2.2 电子病历时间嵌入方法

在处理电子病历序列时,因电子病历间存在时间顺序关系,通常需要逐个进行重复处理。这将增加处理时间,影响处理速度。由于病情的发展具有时间上的依存关系,但不同居民电子病历记录的时间间隔也不尽相同。为了使用电子病历序列的顺序信息,通过在电子病历输入表示中嵌入时间编码来注入相对时序信息。本文采用基于正弦函数和余弦函数的固定时间编码。

假设电子病历掩膜后的线性层输出表示为 $X \in \mathbb{R}^{n \times d}$, 则时间嵌入编码使用形状相同的时间嵌入矩阵 $E \in \mathbb{R}^{n \times d}$ 输出 $X + E$ 。其中, E 由时间方阵 $T \in \mathbb{R}^{n \times n}$ 和位置嵌入矩阵 $P \in \mathbb{R}^{n \times d}$ 矩阵相乘得到。矩阵 P 的第 i 行、第 $2j$ 列和第 $2j + 1$ 列上的元素如式(5)所示。

$$\begin{aligned} p_{i,2j} &= \sin\left(\frac{i}{10\,000^{2j/d}}\right) \\ p_{i,2j+1} &= \cos\left(\frac{i}{10\,000^{2j/d}}\right) \end{aligned} \quad (5)$$

2.3 编码器—稠密网络

在E-TEED-CHD的编码器部分,编码器是由 n 个相同的层叠加而成,每个层又有多头自注意力汇聚子层和基于时间嵌入的前馈网络子层。借鉴残差连接的优点,子层间采用残差连接。在编码器中,任何输入 $x \in \mathbb{R}^d$ 都要求经过子层变换后仍满足 $sublayer(x) \in \mathbb{R}^d$, 以便残差连接满足 $x + sublayer(x) \in \mathbb{R}^d$ 。经过残差计算后,通过层规范化对输出进行规范化处理,以改善参数过大或过小的情况,进而加快网络收敛速度。

在E-TEED-CHD的稠密网络部分,将不同编码器的输出在第0维扩展,并按第0维进行拼接,形成带时间嵌入信息的矩阵。然后将在稠密网络内部不同层的输出按照式(6)进行连接,从而保留不同视野下的所有信息,增强特征重用效果,提高模型的泛化能力。

$$x \rightarrow \left[x, f_1(x), f_2([x, f_1(x)]), f_3([x, f_1(x), f_2([x, f_1(x)])]) \right], \dots \quad (6)$$

3 实验与结果分析

3.1 实验环境及数据集

本文模型采用Python 3.6、sklearn、keras 2.1.6以及tensorflow 1.5.0实现,训练和测试在配置为GTX 3800ti GPU、16G内存的Ubuntu的系统上完成。

因缺乏模型实验所需的公共数据集,本文使用的私有电子病历数据集来自于某市区域卫生平台1 000位居民的全部电子病历,共3 764份。其中,罹患冠心病的居民为500位,正常或罹患其他疾病的居民500位,属于均衡数据

集。电子病历中包含的字段如表1所示。为了与机器学习类模型^[6,9,11,20]的性能进行对比,本文将上述非结构化数据按照Framingham心脏病研究项目数据集中的格式规定进行处理。经清洗后的数据共有2 942条,每条数据中包含15个特征和1个标签。

Table 1 Fields contained in the dataset of EMR

表1 电子病历数据集中包含字段

字段	描述	字段类型
性别	居民性别:男、女	布尔型
年龄	居民年龄:[0, 120]	整型
职业	居民现从事工作情况:[职业列表]	整型
婚姻	已婚、未婚	整型
民族	居民民族:[民族列表]	整型
主诉	就诊主要症状、持续时间等	文本型
现病史	患者自述某时因某种原因出现某种症状,以及症状特点、变化情况与伴随症状	文本型
既往史	慢性病、手术史、外伤史、输血史、过敏史等	文本型
个人史	居住史、吸烟史等	文本型
婚姻史	已婚者爱人情况	文本型
生育史	月经及生育情况	文本型
家族史	父母、兄弟姐妹、子女健康情况及遗传病情况	文本型
体格检查	体温、脉搏、呼吸、血压、身高、体重等检查结果,以及身体各部位视诊、触诊、叩诊、听诊情况	文本型
辅助检查	心电图等	影像数据
病历摘要	初步诊断、依据及诊疗计划等	文本型

3.2 评价指标

为了评估E-TEED-CHD模型的分类能力,本文采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)作为分类评价指标,计算公式如式(7)~式(9)表示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

其中, TP 表示E-TEED-CHD模型预测冠心病发生,且实际发生的情况; FP 表示预测冠心病发生,但实际未发生的情况; TN 表示预测冠心病未发生,且实际未发生的情况; FN 表示预测冠心病未发生,但实际发生的情况。

为了评估E-TEED-CHD模型的训练和测试时间情况,本文将在训练集上的10次实际训练或测试时间的平均值作为衡量标准,单位为秒。

3.3 超参数调优

为获得最优模型,本文针对部分超参数进行调优实

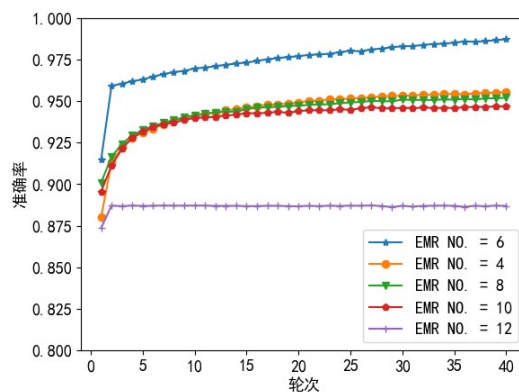
验,包括电子病历数调优和编码器数调优两部分。对于激活函数和学习率等超参数,则采用实验常用设置。为了在线性层、卷积层等中间层中学习任意复杂的特征表示,本文采用非线性变换函数 $relu$ 作为激活函数来引入非线性计算。同时,该激活函数在正数区域的梯度恒为1,使中间层不会因输入值过大或过小而导致梯度消失问题,从而加速中间层训练。在分类层中,则采用 $softmax$ 激活函数。为找到极值点,学习率采用较小的0.001,设置迭代轮次为40次,未设置早停函数。

在电子病历数调优过程中,居民电子病历数中值为3.27份,最小值为1份,最大值为12份。因此,本文将输入电子病历的范围设置为[4, 6, 8, 10, 12]。图2显示了在不同电子病历数下40轮次迭代的准确率和损失值情况。由此可知,当电子病历数较小时,如4等,并不能很好地反映居民的真实诊疗记录情况,准确率较低而损失值较高;而当电子病历数较大时,如8、10和12等,由于绝大多数居民的病历采用电子病历掩膜方法进行填充,也在一定程度上影响了结果的准确率和损失值。因此,本文将调优后的电子病历数设置为6。

在编码器数调优过程中,由于采用 $relu$ 函数,多层 $relu$ 函数叠加也可能带来梯度问题。因此,编码器数量一般不超过6。本文将编码器数量范围设置为[2, 3, 4, 5, 6]。图3显示了在不同编码器数量下40轮次迭代的准确率和损失值情况。由此可知,当编码器数量较小时,如2、3和4等,对居民电子病历真实情况的信息提取仍显不足,准确率较低而损失值较高;而当编码器数较大时,如6等,受多层 $relu$ 函数叠加的影响,梯度传导易出现问题,从而在一定程度上影响了结果的准确率和损失值。因此,本文将调优后的编码器数量设置为5。

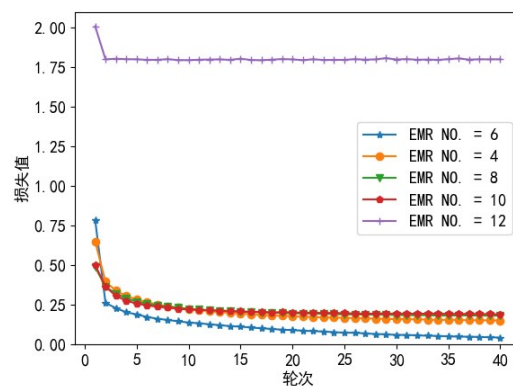
3.4 实验结果

经上述参数调优后,本文在某市区域卫生平台的1 000位居民电子病历数据集上进行E-TEED-CHD模型训练与验证。



(a) The accuracy of model

(a) 模型准确率情况



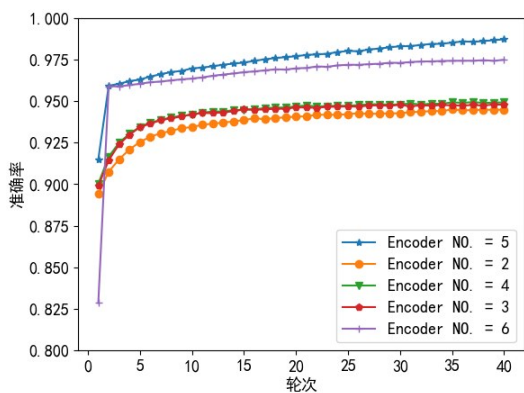
(b) The loss of model

(b) 模型损失情况

Fig. 2 Influence of number of EMR on results

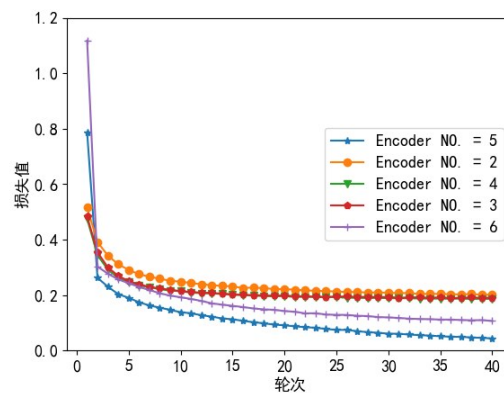
图2 电子病历数对结果的影响

练与验证。如图4所示,经过40轮训练迭代,模型的准确率呈上升趋势,并逐渐趋于稳定,稳定值为98.71%。同时,模型的损失值逐步下降,直至低于0.05。在测试集上,模



(a) The accuracy of model

(a) 模型准确率情况



(b) The loss of model

(b) 模型损失情况

Fig. 3 Influence of number of encoders on results

图3 编码器数量对结果的影响

型的准确率为 98.66%, 损失值为 0.045。综上, E-TEED-CHD 模型具有良好的预测能力及泛化能力, 具备实际应用价值。

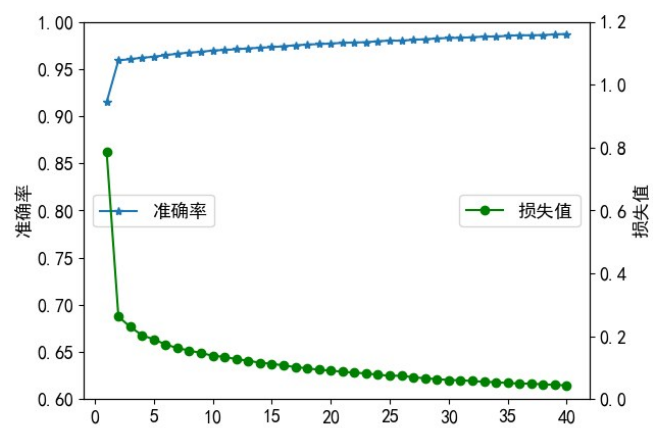


Fig. 4 Experiment results of E-TEED-CHD model on accuracy and loss under different epochs

图 4 E-TEED-CHD 模型不同迭代次数下准确率及损失实验结果

进一步地, 为验证本文提出的电子病历掩膜方法、时间嵌入方法以及稠密网络的有效性, 本文设计了消融实验, 实验结果如表 2 所示。

Table 2 Ablation experiment results of E-TEED-CHD model			
表 2 E-TEED-CHD 模型消融实验结果 (%)			
实验方法	准确率	精确率	召回率
电子病历掩膜	87.37	88.75	98.01
时间嵌入	91.49	94.40	97.82
稠密网络, +softmax	93.37	96.11	97.85
E-TEED-CHD	98.71	98.60	98.82

首先, 去掉电子病历掩膜环节, 将所有电子病历一次性输入。即允许 E-TEED-CHD 模型具有未卜先知的能力, 知道居民后续所有可能的患病情况。从训练阶段来看, 该模型的准确率获得了较大提升, 但在测试集上的准确率只有 87.37%, 低于其他非并行类深度学习模型的预测结果。因此, 电子病历掩膜方法对最终预测结果具有较大影响。其次, 去掉电子病历的时间嵌入环节。从测试结果来看, 准确率下降了约 8% 左右, 但仍高达 91.49%, 且具有较高的召回率(97.82%)。这说明在居民非结构化电子病历的描述中具有一部分时间相关信息, 导致电子病历时间嵌入方法发挥的作用没有达到预期效果。同时, 应考虑居民主诉等环节的部分时间描述可能与真实值存在偏差, 下一步应考虑如何清洗 EMR 中不准确时间因素的影响, 同时进一步发挥时间嵌入方法的作用。最后, 用 *softmax* 函数代替稠密网络。相对而言, 稠密网络利用编码器后各个层级的张量表示, 在分类结果上具有更高的准确性。而 *softmax* 函数仅基于概率进行一次性的最大值输出, 在信息利用上略有欠缺, 致使准确率下降至 93.37%。

同时, 本文还与最新的机器学习模型^[6, 9, 11, 20]和深度

学习模型^[13, 15, 19]进行了对比实验, 实验结果如表 3 所示。为与机器学习模型对比, 本文将上述数据集按照 Framingham 心脏病研究项目中的格式要求进行整理, 每条数据中包含 15 个特征和 1 个标签。在此数据集上, 尽管 Optimized-DBN^[6]、TPE_XGBoost^[9]、HY_OptGBM^[11]、PCHD-TabNet^[20]等模型一个轮次的训练耗时均在 1s 左右, 具有较好的时间性能, 但在准确率方面与深度学习模型存在较大差距。尤其是在召回率方面, Optimized-DBN^[6]、TPE_XGBoost^[9]、HY_OptGBM^[11]等模型尚不足 20%。也即是说, 存在众多漏检的情况, 具有较大的风险系数。而 PCHD-TabNet^[20]虽具有一定的召回能力, 但也是在牺牲部分准确率的基础上实现的, 即该方法存在较高的误报率。相对而言, 深度学习模型充分利用了如检验检查、心电图等电子病历中的结构化和非结构信息, 均具有较好的性能。在准确率、精确率和召回率方面, DM-EMR^[19]基于长短期记忆网络并结合注意力机制对电子病历的上下文语义进行分析, 具有较好的效果, 但其训练耗时明显比 E-TEED-CHD 模型长。THC-Net^[13]、M-US-EMRs^[15]等模型具有较少的参数量, 具有更快的训练速度, 但在准确率、精确率、召回率等方面相比 E-TEED-CHD 模型有着较大差距。

Table 3 Comparative experimental results of E-TEED-CHD model

表 3 E-TEED-CHD 模型对比实验结果

实验方法	准确率/%	精确率/%	召回率/%	训练耗时/s
Optimized-DBN ^[6]	82.45	21.43	12.5	0.917
TPE_XGBoost ^[9]	84.32	34.88	6.33	0.934
HY_OptGBM ^[11]	83.27	20.90	9.72	1.011
PCHD-TabNet ^[20]	60.05	20.28	69.44	0.982
THC-Net ^[13]	90.02	92.05	97.79	3.583
M-US-EMRs ^[15]	90.09	92.06	97.96	3.427
DM-EMR ^[19]	94.16	96.63	98.06	21.295
E-TEED-CHD	98.71	98.60	98.82	17.752

综上, E-TEED-CHD 模型在准确率、精确率、召回率等方面优于其他最新模型, 在训练和预测时间上在可接受的范围之内。因此, E-TEED-CHD 模型具有较高的应用价值和广阔的应用前景。

4 结语

近年来, 冠心病患者数量逐年攀升, 且呈现出年轻化、重度化等趋势。为提高冠心病预测的准确率, 减少医疗诊断开支, 将深度学习模型应用于冠心病风险预测中以辅助临床决策, 对冠心病早筛早防有着重要意义。本文使用居民在区域卫生平台历年积累的电子病历数据, 提出基于电子病历掩膜方法、电子病历时间嵌入方法及编码器—稠密网络的 E-TEED-CHD 模型, 并通过超参数优化方法确定最佳模型结构。通过消融实验、对比实验等实验手段证明了 E-TEED-CHD 模型的优越性。其准确率高达 98.71%, 且千名居民电子病历单轮训练时间低于 17.752 s。同时, 由于 E-TEED-CHD 模型采用的数据均来自于区域卫生平

台,数据获取更加方便,节约了居民去医院检查的时间和费用,便于其自行分析和调整身体状态。

在实验过程中,本文发现在主诉等描述中存在与时间嵌入方法冲突的时间因子,导致电子病历时间嵌入方法发挥的作用没有达到预期。在下一步研究中,将充分考虑并设计消除主诉等非结构数据中时间因子与真实值偏差的影响,从而完善时间嵌入方法,进一步提高E-TEED-CHD模型的准确率。

参考文献:

- [1] SAEID S, NAHID K, KULJIT S, et al. Burden of ischemic heart disease and its attributable risk factors in 204 countries and territories, 1990–2019 [J]. *European Journal of Preventive Cardiology*, 2022, 29(2): 420–431.
- [2] KATTA N, LOETHEN T, LAVIE C J, et al. Obesity and coronary heart disease: epidemiology, pathology, and coronary artery imaging[J]. *Current Problems Cardiology*, 2021, 46(3): 100655.
- [3] National Health Commission of the People's Republic of China. WS/T447–2014. Technical specification for hospital information platform based on EMR[S]. Beijing: Standards Press of China, 2014.
中华人民共和国国家卫生和计划生育委员会. WS/T447–2014 基于电子病历的医院信息平台技术规范[S]. 北京: 中国标准出版社, 2014.
- [4] SHEN S, CLARKE M, JONES R W, et al. A new neural network structure for detection of coronary heart disease [J]. *Neural Computing & Applications*, 1995, 3(3): 171–177.
- [5] KRSTACIC G, GAMBERGER D, SMUC T. Coronary heart disease patient models based on inductive machine learning[C]//*Proceedings of 8th Conference on Artificial Intelligence in Medicine in Europe*, 2001: 113–116.
- [6] LIM K, LEE B M, KANG U G, et al. An optimized DBN-based coronary heart disease risk prediction[J]. *International Journal of Computers Communications & Control*, 2018, 13(4): 492–502.
- [7] ZHANG J, ZHU H, CHEN Y, et al. Ensemble machine learning approach for screening of coronary heart disease based on echocardiography and risk factors [J]. *BMC Medical Informatics and Decision Making*, 2021, 21(1): 187.
- [8] HUANG H, ZHANG R, JIANG Q. An SVM CHD prediction study with optimal parameters[C]// *Qingdao: International Conference on Applied Statistics, Computational Mathematics, and Software Engineering*, 2022.
- [9] HUANG M, LANG X F, ZHOU Z J, et al. Risk assessment and causative factors of coronary heart disease based on TPE_XGBoost [J]. *Software Guide*, 2023, 22(5): 42–49.
黄敏, 郎许锋, 周作建, 等. 基于TPE_XGBoost的冠心病风险评估与致病因素研究[J]. *软件导刊*, 2023, 22(5): 42–49.
- [10] DOPPALA B P, BHATTACHARYYA D, CHAKKRAVARTHY M, et al. A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset[J]. *Distributed Parallel Databases*, 2023, 41(1): 1–20.
- [11] YANG H, CHEN Z, YANG H, et al. Predicting coronary heart disease using an improved LightGBM model: performance analysis and comparison [J]. *IEEE Access*, 2023, 11: 23366–23380.
- [12] JAHMUNAH V, NG E, TAN R S, et al. Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals[J]. *Computers in Biology and Medicine*, 2021, 134: 104457.
- [13] YANG W, SI Y, ZHANG G, et al. A novel method for automated congestive heart failure and coronary artery disease recognition using THC-Net [J]. *Information Sciences*, 2021, 568: 427–447.
- [14] HAN X, LIANG G. Echocardiographic features of patients with coronary heart disease and angina pectoris under deep learning algorithms[J]. *Scientific Programming*, 2021(1): 8336959.
- [15] YANG B, ZUO Y, YANG S. M-US-EMRs: a multi-modal data fusion method of ultrasonic images and electronic medical records used for screening of coronary heart disease[C]// *International Symposium on Bioinformatics Research and Applications*, 2022: 88–89.
- [16] ZHANG T, LIN Y, HE W, et al. GCN-GENE: a novel method for prediction of coronary heart disease-related genes [J]. *Computers in Biology and Medicine*, 2022, 150: 105918.
- [17] WIHARTO T, SURYANI E, SETYAWAN S, et al. The cost-based feature selection model for coronary heart disease diagnosis system using deep neural network[J]. *IEEE Access*, 2022, 10: 29687–29697.
- [18] FAJRI Y A, WIHARTO T, SURYANI E, et al. Hybrid model feature selection with the bee swarm optimization method and Q-learning on the diagnosis of coronary heart disease[J]. *Information*, 2023, 14(1): 1–15.
- [19] NI L. Data mining for electronic medical records of hospital based on bi-directional long short memory network [J]. *Computer Application and Software*, 2023, 40(6): 70–76.
倪凌. 基于双向长短期记忆网络的医院电子病历数据挖掘[J]. *计算机应用与软件*, 2023, 40(6): 70–76.
- [20] JIANG L F, YUAN Z M, ZHANG X W, et al. Ten-year prediction of coronary heart disease based on PCHD-TabNet [J]. *Data Analysis and Knowledge Discovery*, 2023, 7(5): 133–144.
蒋林甫, 袁贞明, 张那炜, 等. 基于PCHD-TabNet的十年冠心病预测[J]. *数据分析与知识发现*, 2023, 7(5): 133–144.

(责任编辑:黄健)