计算机科学
COMPUTER SCIENCE

基于对比学习的疾病诊断预测算法

王明霞, 熊赟

引用本文

王明霞, 熊赟. 基于对比学习的疾病诊断预测算法[J]. 计算机科学, 2023, 50(7): 46-52.

WANG Mingxia, XIONG Yun. Disease Diagnosis Prediction Algorithm Based on Contrastive Learning [J]. Computer Science, 2023, 50(7): 46-52.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200216

基于对比学习的疾病诊断预测算法

王明霞　熊　赟
复旦大学计算机科学技术学院　上海 200433
上海市数据科学重点实验室　上海 200433
(wangmx20@fudan.edu.cn)

摘　要　疾病诊断预测旨在利用电子健康数据建模疾病进展模式,预测患者未来的健康状况,其在辅助临床决策、医疗保健服务等领域得到广泛应用。为了进一步发掘就诊记录中有价值的信息,提出了一种基于对比学习的疾病诊断预测算法。对比学习通过衡量样本间相似度为模型提供自监督训练信号,提升模型的信息捕捉能力。所提算法通过对比训练挖掘相似患者之间的共性知识,增强模型学习患者表征的能力;为了捕获更加全面的共性信息,还进一步挖掘了目标患者相似群体的信息作为辅助信息刻画患者健康状态。在公开数据集上的实验结果表明,相比 Retain,Dipole,LSAN 和 GRASP 算法,所提算法在再入院预测任务的 AUROC 和 AUPRC 指标上分别提升 2.9%和 8.1%以上,在诊断预测任务的 Recall@10 和 MAP@10 指标上分别提升 2.1%和 1.8%以上。

关键词:诊断预测;深度学习;对比学习;聚类;相似患者
中图法分类号　TP311

Disease Diagnosis Prediction Algorithm Based on Contrastive Learning

WANG Mingxia and XIONG Yun
School of Computer Science,Fudan University,Shanghai 200433,China
Shanghai Key Laboratory of Data Science,Shanghai 200433,China

Abstract　Disease diagnosis prediction aims to use electronic health data to model disease progression patterns and predict the future health status of patients,and is widely used in assisting clinical decision-making,healthcare services and other fields. In order to further explore the valuable information in the medical records,a disease diagnosis prediction algorithm based on contrastive learning is proposed. Contrastive learning provides self-supervised training signals for the model by measuring the similarity between samples,which can improve the information capture ability of the model. The proposed algorithm excavates the common knowledge between similar patients through contrastive training,and enhances the ability of the model to learn patient representations. In order to capture more comprehensive common information,the information of similar groups of the target patient is further explored as auxiliary information to characterize the health status of the target patient. Experimental results on the public dataset show that compared with the Retain,Dipole,LSAN and GRASP algorithms,the proposed algorithm improves AUROC and AUPRC of the readmission prediction task by more than 2.9% and 8.1% respectively,and Recall@10 and MAP@10 of the diagnosis prediction task by 2.1% and 1.8%,respectively.

Keywords　Diagnosis prediction,Deep learning,Contrastive learning,Clustering,Similar patients

1　引言

得益于数字医疗系统的广泛应用,大量的电子健康记录(Electronic Health Records,EHRs)被搜集整理,其涵盖患者全面的医疗健康信息,包含疾病诊断信息、药物处方情况和采取的治疗手段等。电子健康记录是带有时间戳的动态数据,通过长期跟踪患者病情的进展情况获得。它可以为增强临床决策和促进医疗保健服务提供海量有价值的信息。在健康分析领域,根据患者历史电子健康记录预测患者未来的健康状态已经引起广泛的关注,其中诊断预测是根据患者的历史病历建模疾病进展,预测患者未来可能的患病情况,是临床预测领域研究的重点问题之一。

随着深度学习的发展逐渐成熟,很多工作基于深度学习模型[1-3]开展诊断预测研究。序列模型考虑到患者历史就诊记录是按顺序排列的,采用基于循环神经网络(Recurrent Neural Network,RNN)的方法建模历史就诊记录。Retain[4]采用基于注意力机制的两个反向 RNN 识别出关键变量和重要就诊。Dipole[5]在双向递归 RNN 的基础上引入多种注意力机制对电子病历数据进行建模。CONTENT[6]采用混合主题的 RNN 捕获局部和全局上下文信息。但这类方法忽略了

到稿日期:2023-02-28　返修日期:2023-04-17
通信作者:熊赟(yunx@fudan.edu.cn)

就诊记录中重要的时间信息，例如短时间内同一事件的突然爆发可能预示着一种严重疾病的发生，而事件之间长时间无明显变化可能表明它们对诊断没有影响。

为了处理就诊记录间的时间间隔信息，具有时间感知能力的方法应运而生。Doctor AI[7]将就诊中的医疗编码和时间戳合并在一起作为输入信息。T-LSTM[8]设计了具有时间感知能力的长短期记忆网络单元(Long Short Term Memory，LSTM)建模时间间隔，采用时间衰减函数使长期记忆对当前输出的影响降低。RetainEX[9]在 Retain 基础上采用双向 RNN 建模就诊序列，并将同一时间间隔的 3 种不同时间表示作为输入向量的附加特征。Timeline[10]使用基于时间的疾病特定进展函数建模每种疾病有多少信息流入 RNN 模型中。HiTANet[11]设计了对时间敏感的 Transformer，将时间信息嵌入就诊表示中，学习每次就诊的局部注意力权重，从全局角度识别出历史就诊中关键的时间步。Men 等[12]基于 LSTM 建模电子健康数据，提出一种时间感知机制来处理临床就诊之间时间间隔的不规则性。这类方法仅建模患者自身的电子病历，其稀疏性和不完整性的特点给诊断预测带来了巨大挑战。

为了缓解健康数据稀疏性问题，一些方法引入相似患者信息作为目标患者个体电子健康数据的补充，更加全面地描绘患者健康状态，增强预测性能。Suo 等[13]首先采用基于成对训练的融合时间信息的卷积神经网络(Convolutional Neural Networks，CNN)框架学习患者表征并测量患者对之间的相似度，根据相似性分数进行疾病预测。Suo 等[14]使用 CNN 学习患者表示，采用三重损失函数或交叉熵损失学习患者间相似性，使用 K 最近邻分类器进行预测。GRASP[15]提出了一种融合相似患者知识学习患者健康状态表示的框架，通过聚类对不同类型的患者进行分组，并使用图神经网络(Graph Neural Networks，GNN)获得增强后的患者分组表示，自适应地融合患者表示和相似患者的辅助信息以执行预测任务。DK-CNN[16]采用基于不同大小卷积核的 CNN 方法建模就诊记录间短期、中期和长期进展状况，计算成对患者的匹配分数，基于相似性标签对模型进行有监督训练。AHGCN-PS[17]根据电子健康数据构建具有医学属性的异构信息网络，采用异构图卷积网络学习患者节点表征，基于该表征计算患者相似性。MERGE[18]提出了一个基于多图注意力机制的患者表征学习框架，从多个角度获取目标患者的相似群体，构建对应的相似性图学习患者群体表示。M3Care[19]面向多种模态的医疗健康数据提出了一个端到端的缺失模态学习算法，利用相似患者辅助信息插补目标患者缺失的模态，缓解模态缺失带来的问题。这类方法对相似患者的学习过程仅和预测任务相关，没有深层次挖掘患者历史就诊记录中相似患者的共性特征，对 EHR 数据的利用不够充分。

为了弥补现有研究存在的不足，本文提出了一种基于对比学习的疾病诊断预测算法(Disease Diagnosis Prediction Algorithm Based on ContrastiveLearning，DCL)，不仅利用来自标签的监督信号识别出和预测任务相关的相似性群体，而且引入对比损失的无监督信号深入挖掘相似患者潜在的共性知识，增强模型学习患者表征的能力。此外，算法利用多次聚类操作融合多个不同相似程度的患者群体，使学习到的相似患者信息更加全面，将其作为患者自身电子健康记录的补充信息。

DCL 算法不需要引入额外医疗本体知识或先验知识构建模型，而是通过深入挖掘电子健康数据中相似患者的共性特征辅助刻画患者健康状态。该算法不依赖于电子病历的具体构成内容，而是在抽象语义空间学习患者表征，对于复杂的健康状态进行概括凝练。相比现有算法，DCL 不依赖于某种特定疾病的特征，可用于多种临床预测任务，具有更强的泛化能力和抽象表示能力。本文在公开医疗数据集上设计了丰富的实验，与现有研究工作进行对比分析，结果表明了 DCL 算法的有效性和优势。

2 DCL 算法

2.1 任务描述

用 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_m\}$ 代表所有的诊断编码，m 表示诊断编码的总数，$\boldsymbol{c}_j=[0,\cdots,1,\cdots,0]^{\mathrm{T}}$ 是独热编码形式的列向量，1 出现在第 j 行，T 是转置符号。一次历史就诊记录用 $\boldsymbol{v}$ 表示，$\boldsymbol{v}\in\{0,1\}^m$，如果第 j 个诊断编码 $\boldsymbol{c}_j$ 出现在就诊记录中，那么 $\boldsymbol{v}_j=1$，否则 $\boldsymbol{v}_j=0$，就诊记录 $\boldsymbol{v}$ 对应的时间戳为 t。按照时间排序的历史就诊记录为 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n]$，对应的时间戳为 $\boldsymbol{T}=[t_1,t_2,\cdots,t_n]$，$n$ 为就诊总次数。

本文在两个预测任务上进行实验：1)再入院预测任务，根据历史就诊记录预测未来 30 天内患者是否会再入院，即 $t_{n+1}-t_n$ 是否小于 30，该任务是一个二分类任务；2)诊断预测任务，预测患者下一次就诊时可能出现的诊断编码，该任务常被视为一个多标签分类任务。

2.2 模型概述

DCL 算法由 4 部分组成，分别是个体表征学习模块、对比学习预训练模块、相似患者群体学习模块和预测模块。

如图 1 所示，个体表征学习模块基于现有模型建模患者历史就诊记录。

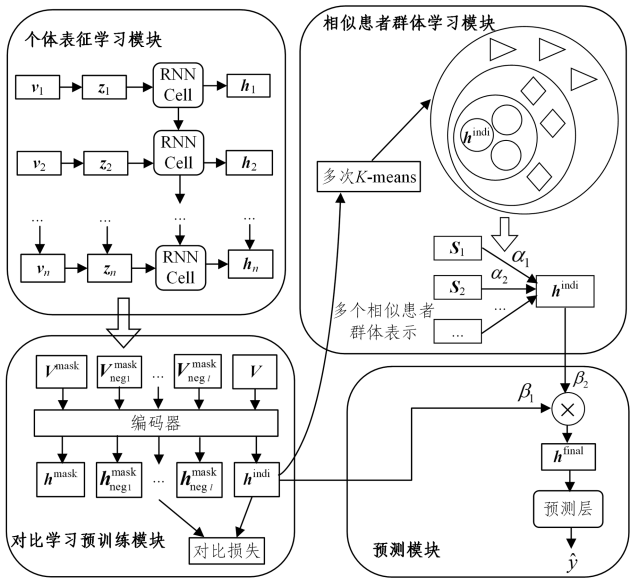

图 1 DCL 算法图

Fig. 1 DCL algorithm diagram

学习患者个体自身的健康状况；对比学习预训练模块通过引入对比损失对模型进行预训练，深入挖掘患者电子健康数据的内在联系，捕捉患者之间潜在的共性特征，提高编码器学习患者表征的能力；相似患者群体学习模块通过发掘多种相似程度的相似患者群体，捕获更为全面的相似患者知识，作为辅助信息增强患者表征的表达能力；预测模块将从患者历史就诊记录中学习到的个体表征和相似群体表征自适应地融合在一起，获得经过相似患者信息增强后的表征向量，用于预测任务。

2.3 个体表征学习模块

个体表征学习模块是对患者过往就诊记录中的医疗编码信息进行建模，学习患者个体表征向量，刻画患者自身的健康状态。编码器 Encoder 可以选择现有研究工作中的任一序列模型或具有时间感知能力的模型，此处以 RNN 为例进行阐述。首先获得每次就诊记录 $\boldsymbol{v}_i$ 的低维嵌入表示：

$$\boldsymbol{z}_i=\sigma(\boldsymbol{W}_0\boldsymbol{v}_i+\boldsymbol{b}_0) \tag{1}$$

其中，$\boldsymbol{W}_0$ 和 $\boldsymbol{b}_0$ 是需要学习的参数，$\boldsymbol{z}_i\in R^d$，d 是向量维度，σ 是激活函数。之后将嵌入表示输入 RNN 模型中，生成每一个时间步 t 的隐状态表示：

$$\boldsymbol{h}_t=RNN(\boldsymbol{z}_t,\boldsymbol{h}_{t-1}),t=1,2,\cdots,n \tag{2}$$

其中，$\boldsymbol{h}_0=0$。选择最终时间步的隐状态 $\boldsymbol{h}_n$ 作为患者个体表征向量。上述过程可以简化为：

$$\boldsymbol{h}^{\text{indi}}=Encoder(\boldsymbol{V}) \tag{3}$$

2.4 对比学习预训练模块

引入对比损失对模型进行预训练，深入挖掘患者电子健康记录潜在的内在联系。对比学习的关键步骤是构造正实例对和负实例对。针对每个患者，使用由患者个体表征学习模块构成的编码器生成查询向量。

$$\boldsymbol{q}=Encoder(\boldsymbol{V}) \tag{4}$$

然后从历史就诊记录中随机掩盖一次就诊记录，记作 $\boldsymbol{V}^{\text{mask}}$，用于构造正实例。

$$\boldsymbol{k}=Encoder(\boldsymbol{V}^{\text{mask}}) \tag{5}$$

对于目标患者 p 而言，正实例对是 $(\boldsymbol{q},\boldsymbol{k})$，再从剩余患者中选择 l 个患者，构造 l 个负实例对 $(\boldsymbol{q},\boldsymbol{k}_{\text{neg}_1})$，$(\boldsymbol{q},\boldsymbol{k}_{\text{neg}_2})$，$\cdots$，$(\boldsymbol{q},\boldsymbol{k}_{\text{neg}_l})$。采用 InfoNCE[20] 损失函数训练模型。

$$L_{\text{con}}=\sum_{i=1}^{|P|}-\log\frac{\exp\left(\boldsymbol{q}_i\cdot\frac{\boldsymbol{k}_i}{\tau}\right)}{\exp(\boldsymbol{q}_i\cdot\boldsymbol{k}_i)+\sum_{j=1}^{l}\exp\left(\boldsymbol{q}_i\cdot\frac{\boldsymbol{k}_{\text{neg}_j}}{\tau}\right)} \tag{6}$$

其中，τ 是温度超参。通过最小化 InfoNCE 损失函数，识别出相似患者的共性，保留不相似患者的判别信息，更好地学习患者表征。

2.5 相似患者群体学习模块

考虑到患者历史就诊记录可能存在稀疏性和不完整性的特点，单纯使用患者电子健康数据学习患者表征可能无法准确反映患者的整体健康状况。除了患者自身的历史就诊记录外，相似患者中也蕴含着丰富的信息。为了能够将相似患者信息以更加直观的方式引入模型中，首先利用相似患者群体学习模块针对目标患者找出多个具有不同相似程度的相似患者群体，进而捕获相似患者的共性知识，作为描述患者健康状态的补充信息，增强患者表征的表达能力。

本文采用多次聚类对患者进行分组，为目标患者学习多个不同相似程度的患者群体[21]。具体来说，对上述个体表征学习模块学习到的所有患者表征 $\boldsymbol{H}^{\text{indi}}$ 进行多次 k-means 聚类操作，每次选择不同数目的簇，从而获得目标患者的多个相似患者群体。当簇的数目为 o 时，聚类后簇心表示如下：

$$\boldsymbol{s}=k\text{-}means(o,\boldsymbol{H}^{\text{indi}}) \tag{7}$$

目标患者 p 在经过多次聚类后得到的所属簇的质心表示记为 $\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_{|O|},\boldsymbol{s}_j\in R^d$。

为了能够从相似患者群体中尽可能提取出和患者自身相似的信息，此处采用键值查询(key-query)注意力机制获得多个相似患者群体表征的权重值，用于决定不同相似患者群体信息在多大程度上影响患者健康状况的描述。首先计算个体表征和相似患者群体表征之间的相似性。

$$\boldsymbol{sim}_i=\boldsymbol{h}^{\text{indi}^{\mathrm{T}}}\cdot\boldsymbol{s}_i,\ i\in\{1,2,\cdots,|O|\} \tag{8}$$

然后引入 softmax 对相关性得分进行归一化处理。

$$\alpha_i=\text{Softmax}(\boldsymbol{sim}_i)=\frac{\boldsymbol{sim}_i}{\sum_i\exp(\boldsymbol{sim}_i)} \tag{9}$$

之后将权重值和相似患者群体表征融合在一起获得相似患者群体的最终表示。

$$\boldsymbol{h}^{\text{simi}}=\sum_{i=1}^{|O|}\alpha_i\boldsymbol{s}_i \tag{10}$$

2.6 预测模块

采用自适应融合机制将患者个体表征和相似患者群体表征组合在一起，进行后续预测。首先计算出二者的权重。

$$\beta_1=\text{Sigmoid}(\boldsymbol{W}_1\boldsymbol{h}^{\text{indi}}+b_1) \tag{11}$$

$$\beta_2=\text{Sigmoid}(\boldsymbol{W}_2\boldsymbol{h}^{\text{simi}}+b_2) \tag{12}$$

其中 Sigmoid 公式为：

$$\text{Sigmoid}(x)=\frac{1}{1+\exp(-x)} \tag{13}$$

然后通过如下公式获得患者最终表征。

$$\boldsymbol{h}^{\text{final}}=\beta_1\boldsymbol{h}^{\text{indi}}+\beta_2\boldsymbol{h}^{\text{simi}} \tag{14}$$

再入院预测任务是一个二分类任务，因此通过 softmax 获得预测结果。

$$\hat{y}_r=\text{Sofmax}(\boldsymbol{W}_3\boldsymbol{h}^{\text{final}}+b_3) \tag{15}$$

其中，$\hat{y}_r\in[0,1]$ 是预测概率。用 $y_r\in\{0,1\}$ 代表真实标签，计算交叉熵作为损失函数。

$$L_r=-\sum(y_r^{\mathrm{T}}\log\hat{y}_r+(1-y_r)\log(1-\hat{y}_r)) \tag{16}$$

诊断预测任务是一个多标签分类任务，由于不同的诊断编码之间并不是互斥的，因此使用 Sigmoid 层进行预测。

$$\hat{y}_d=\text{Sigmoid}(\boldsymbol{W}_4\boldsymbol{h}^{\text{final}}+b_4) \tag{17}$$

其中，$\hat{y}_d\in[0,1]^m$ 是每个编码在下次就诊中出现的概率。用 $y_d\in\{0,1\}^m$ 表示真实标签，计算交叉熵作为损失函数。

$$L_d=-\sum(y_d^{\mathrm{T}}\log\hat{y}_d+(1-y_d)\log(1-\hat{y}_d)) \tag{18}$$

3 实验及结果

3.1 数据集

实验部分使用公开的大规模电子健康记录数据集 MIM-

IC-III(Medical Information Mart for Intensive Care)，其中诊断编码遵循国际疾病分类(International Classification of Diseases，ICD)中 ICD-9 高维编码体系。本文按临床分类软件(Clinical Classification Software，CCS)中的单层诊断编码分组对 ICD-9 诊断编码进行分组，并将原始 ICD-9 诊断编码替换为分组编码，在诊断预测任务中预测这一分组编码，可以在为每个诊断编码保留足够精度的同时加快模型的训练速度。对数据集进行预处理，提取至少含有两次就诊记录的患者作为实验样本，删除所有不常见的诊断编码，根据经验将阈值设置为 5[22]。最终一共提取 7486 名患者用于实验。表 1 列出了经过预处理后数据集的详细信息。

表 1 数据集统计信息

Table 1 Dataset statistics

数据集	MIMIC-III
患者数	7486
就诊总数	19884
平均就诊次数	2.66
最大就诊次数	42
最小就诊次数	2
ICD-9 编码数	2437
就诊平均 ICD-9 编码数	13.05
就诊最大 ICD-9 编码数	39
就诊最小 ICD-9 编码数	1
CCS 分组编码数	243
就诊平均 CCS 分组编码数	11.07
就诊最大 CCS 分组编码数	34
就诊最小 CCS 分组编码数	1

3.2 对比算法

实验使用 4 种模型(GRU，Retain，Dipole，LSAN[23])作为基准模型建模患者历史就诊记录信息，获得患者的个体表征向量，此外，还对比了另一个通用表征学习算法 GRASP。在 4 种基准模型的基础上，分别利用 GRASP 算法和 DCL 算法进行实验。

GRU 采用 GRU 模型对患者历史就诊记录进行编码，取最后一个时间步的输出作为患者表征向量。Retain 使用两个反向 RNN 建模就诊记录序列，利用就诊级别和变量级别的注意力机制分别识别出关键的就诊记录和变量。Dipole 使用双向递归 RNN 学习患者历史就诊信息，设计 3 种不同的注意力机制计算每次就诊的重要程度。LSAN 采用分层注意力机制分别识别出就诊中的关键疾病和历史就诊记录中的重要就诊，此外，还考虑就诊之间的长期依赖性和短期相关性特征建模疾病进展过程。GRASP 是一种通用表征学习算法，不仅考虑了当前患者的健康信息，而且采用聚类方法从相似患者中提取辅助信息，增强了患者表征的表达能力。

3.3 实验设置

DCL 和对比算法用 Python 实现，版本是 3.8.13；使用的框架是 PyTorch，版本是 1.9.1。在 GeForce GTX 1080 Ti GPU 上进行实验，使用 Adam 优化器训练模型参数。将数据集按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集，批大小(*batch size*)为 64，患者表征维度(d)为 128。

3.4 评价指标

针对再入院预测任务使用 3 个常用的指标，分别是 AUPRC，AUROC 和 Min(Se，P+)。AUPRC(Area under Precision-Recall Curve)常用来处理高度不平衡数据集的性能：AUROC(Area under Receiver Operator Curve)利用 ROC 曲线与坐标轴之间的面积大小反映分类器的性能：Min(Se，P+)是准确率和灵敏度的最小值。这 3 个评估指标的数值越大，代表预测结果的表现越好。

针对诊断预测任务使用 2 个常用的指标 Recall@k 和 MAP@k。Recall@k 表示排序后前 k 个返回结果中预测正确的编码数占所有正确编码数的比例。MAP@k(Mean Average Precision)是信息检索领域广泛使用的指标，利用该指标可以衡量预测的准确性。k 取 5，10 和 15。这两个评估指标的数值越大，表明模型的性能越好。

3.5 实验结果与分析

3.5.1 性能比较实验

为了研究 DCL 算法对基准模型产生的影响，在现有基准模型的基础上分别使用 GRASP 算法和 DCL 算法进行实验，+G 代表使用 GRASP 算法，+D 代表使用 DCL 算法。

再入院预测任务和诊断预测任务的性能实验结果分别如表 2 和表 3 所列。实验结果中最佳结果用加粗表示。对比分析实验结果，可以得到以下结论：

(1)与现有基准模型相比，在再入院预测任务上使用 DCL 算法后性能均有提升，在诊断预测任务上使用 DCL 算法后各项指标都达到最优结果，表明了 DCL 算法的有效性。通过引入对比学习深入挖掘数据中潜在的信息，使用相似患者信息增强患者表征的表示能力，共同促进了预测性能的提升。

(2)和 GRASP 相比，DCL 对模型的预测性能的提升效果更为显著，在再入院预测任务的 AUROC 和 AUPRC 指标上都达到最优结果，仅在 LSAN 模型的 Min(Se，P+)指标略低；在诊断预测任务的所有指标上都获得了最佳结果，表明 DCL 算法具有一定的优势。

表 2 再入院预测性能结果

Table 2 Performance results on readmission prediction

Method	*AUROC*	*AUPRC*	*Min(Se，P+)*
GRU	0.508	0.247	0.256
GRU+G	.513	0.258	0.259
GRU+D	**0.571**	**0.291**	**0.315**
Retain	0.534	0.265	0.268
Retain+G	0.546	0.271	0.298
Retain+D	**0.582**	**0.310**	**0.322**
Dipole	0.507	0.258	0.269
Dipole+G	0.520	0.263	0.276
Dipole+D	**0.590**	**0.296**	**0.333**
LSAN	0.540	0.270	0.273
LSAN+G	0.558	0.282	0.314
LSAN+D	**0.574**	**0.305**	0.303

表 3 诊断预测性能结果

Table 3 Performance results on diagnosis prediction

Method	*Recall@k*			*MAP@k*		
	Recall@5	*Recall*@10	*Recall*@15	*Recall*@5	*Recall*@10	*Recall*@15
GRU	0.257	0.398	0.500	0.220	0.306	0.357
GRU+G	0.264	0.416	0.518	0.229	0.326	0.379
GRU+D	**0.276**	**0.430**	**0.533**	**0.241**	**0.342**	**0.398**
Retain	0.263	0.413	0.513	0.228	0.325	0.379
Retain+G	0.269	0.416	0.522	0.235	0.332	0.388
Retain+D	**0.281**	**0.436**	**0.545**	**0.245**	**0.348**	**0.407**
Dipole	0.269	0.421	0.517	0.232	0.330	0.382
Dipole+G	0.271	0.423	0.522	0.233	0.331	0.384
Dipole+D	**0.277**	**0.432**	**0.533**	**0.243**	**0.345**	**0.400**
LSAN	0.276	0.420	0.519	0.242	0.336	0.389
LSAN+G	0.278	0.424	0.522	0.244	0.341	0.394
LSAN+D	**0.281**	**0.434**	**0.529**	**0.247**	**0.347**	**0.399**

分析原因主要有以下两点：1)DCL 算法引入对比学习预训练模块对模型进行预训练，相比 GRASP，预训练模块可以充分利用数据中潜藏的有价值的信息，深入捕捉患者之间的共性特征，并且尽可能将相异的患者个体区分开，增强模型学习患者表征的能力；2)相比 GRASP 算法只使用一次 k-means 获得一种相似程度的相似患者群体，DCL 算法通过改变簇的数目，采用多次聚类操作捕获目标患者的多个不同相似程度的相似群体知识，使得学习到的患者相似性信息更加完善，有助于描述患者的健康状态。

3.5.2 消融实验

为了检验 DCL 算法各个部分的作用，本文设计了 DCL 算法的两个变体：1)+S 代表使用单次聚类引入相似患者群体信息，且不对模型进行预训练；2)+M 表示采用多次聚类学习多个不同相似程度的相似患者群体，同时去除对比学习预训练模块。

再入院预测任务和诊断预测任务的消融实验结果分别如表 4 和表 5 所列。观察结果可以得到以下结论：

(1)引入一种相似程度的相似患者群体信息(+S)在两个任务的所有性能指标上都有所提高，这表明使用患者相似性信息能够增强患者表征的表达能力，更好地展现患者健康状况，最终促进预测过程。

表 4 再入院预测消融实验结果

Table 4 Results of ablation studies on readmission prediction

Method	*AUROC*	*AUPRC*	*Min*(*Se*,*P*+)
GRU	0.508	0.247	0.256
GRU+S	0.518	0.269	0.270
GRU+M	0.539	0.271	0.289
GRU+D	**0.571**	**0.291**	**0.315**
Retain	0.534	0.265	0.268
Retain+S	0.548	0.275	0.311
Retain+M	0.549	0.297	0.307
Retain+D	**0.582**	**0.310**	**0.322**
Dipole	0.507	0.258	0.269
Dipole+S	0.549	0.273	0.280
Dipole+M	0.555	0.287	0.284
Dipole+D	**0.590**	**0.296**	**0.333**
LSAN	0.540	0.270	0.273
LSAN+S	0.558	0.286	0.300
LSAN+M	0.568	0.295	**0.304**
LSAN+D	**0.574**	**0.305**	0.303

表 5 诊断预测消融实验结果

Table 5 Results of ablation studies on diagnosis prediction

Method	*Recall@k*			*MAP@k*		
	Recall@5	*Recall*@10	*Recall*@15	*Recall*@5	*Recall*@10	*Recall*@15
GRU	0.257	0.398	0.500	0.220	0.306	0.357
GRU+S	0.264	0.416	0.518	0.229	0.326	0.379
GRU+M	0.275	0.425	0.526	0.237	0.335	0.388
GRU+D	**0.276**	**0.430**	**0.533**	**0.241**	**0.342**	**0.398**
Retain	0.263	0.413	0.513	0.228	0.325	0.379
Retain+S	0.269	0.416	0.522	0.235	0.332	0.388
Retain+M	0.275	0.429	0.530	0.245	0.341	0.394
Retain+D	**0.281**	**0.436**	**0.545**	**0.245**	**0.348**	**0.407**
Dipole	0.269	0.421	0.517	0.232	0.330	0.382
Dipole+S	0.271	0.423	0.522	0.233	0.331	0.384
Dipole+M	0.267	0.425	0.527	0.233	0.335	0.390
Dipole+D	**0.277**	**0.432**	**0.533**	**0.243**	**0.345**	**0.400**
LSAN	0.276	0.420	0.519	0.242	0.336	0.389
LSAN+S	0.278	0.424	0.522	0.244	0.341	0.394
LSAN+M	**0.282**	0.431	0.528	0.245	0.344	0.396
LSAN +D	0.281	**0.434**	**0.529**	**0.247**	**0.347**	**0.399**

(2)和只引入单个相似患者群体信息相比，引入多个相似患者群体信息(+M)后，在再入院预测任务的 AUROC 和 AUPRC 指标上都有所提升。Min(Se,P+)指标仅在 Retain 上略低，在诊断预测任务的大多数指标上都有所提升，这表明通过多次聚类操作获得的目标患者的多个不同相似程度的相似患者群体信息更为全面，捕获了更加完善的患者健康状态，

对模型效果产生了积极的影响。

(3)使用 DCL 算法(+D)在引入多个相似患者群体信息的基础上增加对比学习预训练模块，在再入院预测任务的 AUROC 和 AUPRC 指标上都取得了最优结果。Min(Se，P+)指标仅在 LSAN 模型上略低，在诊断预测任务的几乎所有指标上都获得了最佳结果，这表明添加对比学习预训练模块能够增强模型的表现，通过对模型进行预训练，可以充分挖掘数据中潜在的共性信息，提升模型的预测能力。对比学习预训练模块和相似患者群体学习模块两部分组合在一起，相互促进，使 DCL 算法发挥了最佳效果。

3.5.3 参数敏感性实验

本文涉及几个重要的超参数，包括对比学习预训练模块中的负实例对个数 l、批大小(batch size)。为了探究这些超参数对 DCL 算法性能的影响，本文设计了多个超参数敏感性分析实验，接下来将进行详细阐述。

(1)负样本个数参数分析

为了探究选取不同数量的负样本对如何影响预测结果，基于 Dipole 模型在诊断预测任务上进行实验，采用不同数量的负样本对预先训练模型，基于该模型执行诊断预测任务获得最终结果。实验结果如图 2 所示，分析可知，$k=5$ 时两个指标随着负样本个数的增加展现的变化趋势类似，$k=10$ 和 $k=15$时指标和负样本个数基本呈现正相关性，表明增加负样本个数在一定程度上可以促进模型的预训练过程，提高模型的表征学习能力，对结果产生积极影响。

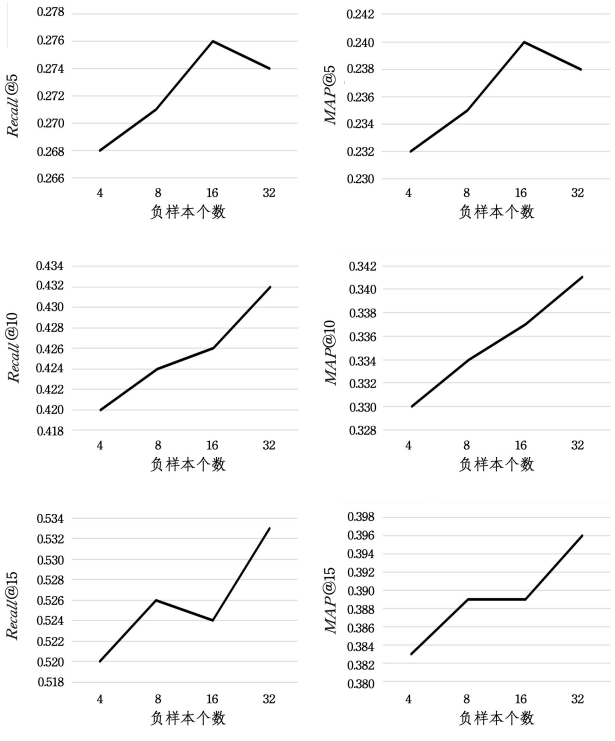

图 2 负样本个数参数实验结果

Fig. 2 Experiment results of negative instance pair parameters

(2)批大小参数分析

为了探究不同的批大小(batch size)对模型效果的影响，本文基于 GRU 模型在诊断预测任务上进行实验，固定预训练模型，在预测阶段更改批大小。实验结果如图 3 所示，结果表明，当批大小低于一定阈值时，模型性能表现和批大小呈现正相关性，但是当批大小大于一定阈值时，模型的预测性能呈现下降趋势。因此，为使模型获得最佳效果，需要选择合适的批大小。

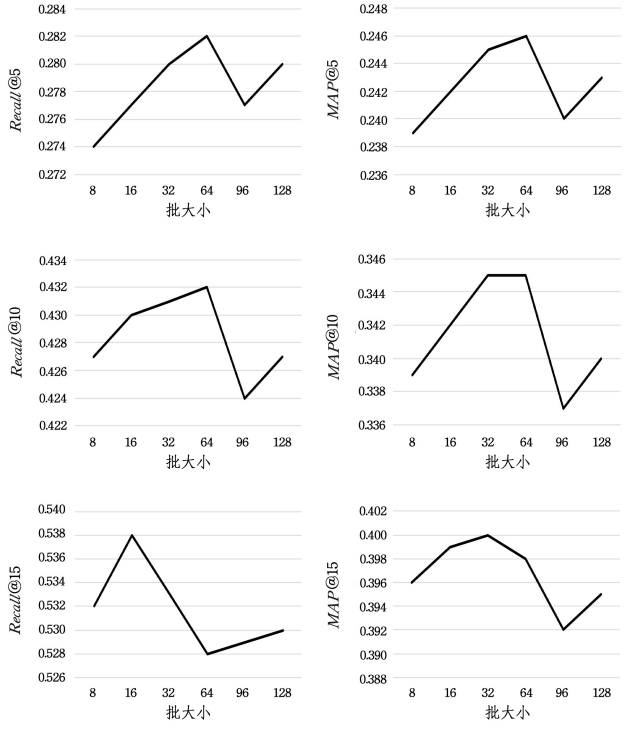

图 3 批大小参数实验结果

Fig. 3 Experiment results ofbatch size parameters

结束语 本文提出了基于对比学习的疾病诊断预测算法 DCL，通过对比学习对患者历史就诊记录中潜藏的丰富信息进行深度挖掘，捕捉患者相似性知识，利用对比损失对模型进行预训练，增强了模型学习患者表征的能力。通过多次聚类操作识别出目标患者的多个不同相似程度的患者相似性群体，引入更为全面的患者相似性知识，辅助描述患者健康状况，促进预测结果的提升。在再入院预测任务和诊断预测任务上的实验结果表明了 DCL 算法的有效性和优越性。与现有方法相比，DCL 算法的预训练过程和多次聚类操作会增加模型训练的时间。下一步将使用电子健康数据中的图片、文本等多种模态信息学习患者表征，构建更为丰富的患者画像。

参 考 文 献

[1] LI Y J，ZHENG R L，YANG X M. Diagnosis and prediction model of coronary heart disease based on data mining technology [J]. Medical Information，2020，33(24)：14-17.

[2] ZHU X T，PANG C Y，ZHU H. Cardiovascular disease prediction model based on deep learning [J]. Journal of Computer Applications，2021，41(S2)：346-350.

[3] LI M，MA L Y，YAO Z. Study on an intelligent diagnosis prediction model based on deep neural network[J]. Medical Information，2022，43(8)：52-55，75.

[4] CHOI E，BAHADORI M T，KULAS J A，et al. Retain：An interpretable predictive model for healthcare using reverse time attention mechanism[C]// Proceedings of the 30th International

Conference on Neural Information Processing Systems. 2016: 3512-3520.
[5] MA F, CHITTA R, ZHOU J, et al. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1903-1911.
[6] XIAO C, MA T, DIENG A B, et al. Readmission prediction via deep contextual embedding of clinical concepts[J]. PLOS ONE, 2018, 13(4): 1-15.
[7] CHOI E, BAHADORI M T, SCHUETZ A, et al. Doctor AI: Predicting clinical events via recurrent neural networks[C]// Proceedings of the 1st Machine Learning for Healthcare Conference. 2016: 301-318.
[8] BAYTAS I M, XIAO C, ZHANG X, et al. Patient subtyping via time-aware LSTM networks[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 65-74.
[9] KWON B C, CHOI M J, KIM J T, et al. RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records [J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 299-309.
[10] BAI T, ZHANG S, EGLESTON B L, et al. Interpretable representation learning for healthcare via capturing disease progression through time[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018: 43-51.
[11] LUO J, YE M, XIAO C, et al. HiTANet: Hierarchical time-aware attention networks for risk prediction on electronic health records [C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2020: 647-656.
[12] MEN L, ILK N, TANG X, et al. Multi-disease predictionusing LSTM recurrent neural networks[J]. Expert Systems with Applications, 2021, 177: 114905.
[13] SUO Q, MA F, YUAN Y, et al. Personalized disease prediction using a CNN based similarity learning method[C]// 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017: 811-816.
[14] SUO Q, MA F, YUAN Y, et al. Deep patient similarity learning for personalized health care[J]. IEEE Transactions on NanoBioscience, 2018, 17(3): 219-227.
[15] ZHANG C, GAO X, MA L, et al. GRASP: Generic framework for health status representation learning based on incorporating knowledge from similar patients [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 715-723.
[16] OEI R W, HSU W, LEE M L, et al. Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: a domain knowledge infused convolutional neural network approach[J]. Journal of the American Medical Informatics Association, 2022, 30(2): 273-281.
[17] LI Y, YANG D, GONG X. Patient similarity via medical attributed heterogeneous graph convolutional network[J]. IAENG International Journal of Computer Science, 2022, 49(4): 1152-1161.
[18] AN Y, LI R, CHEN X. MERGE: A multi-graph attentive representation learning framework integrating group information from similar patients[J]. Computers in Biology and Medicine, 2022, 151: 106245.
[19] ZHANG C, CHU X, MA L, et al. M3Care: Learning with missing modalities in multimodal healthcare data[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD, 2022: 2418-2428.
[20] VAN DEN OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv: , 1807. 03748, 2018.
[21] LI J, ZHOU P, XIONG C, et al. Prototypical contrastive learning of unsupervised representations [J]. arXiv: 2005. 04966, 2020.
[22] PENG X, LONG G, SHEN T, et al. Self-attention enhanced patient journey understanding in healthcare system[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2020: 719-735.
[23] YE M, LUO J, XIAO C, et al. LSAN: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction[C]// Proceedings of the 29th ACM International Conference on Information and Knowledge Management. 2020: 1753-1762.

WANG Mingxia, born in 1999, postgraduate. Her main research interests include big data and medical data mining.

XIONG Yun, born in 1980, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include data science and data mining.

(责任编辑:何杨)