面向医疗缺失数据的选择性学习

韩宗博[1],西雨晗[2],盛培卓[3],廖 芸[2*],张长青[3]
(1. 北京邮电大学信息与通信工程学院,北京 100876;2. 天津科技大学人工智能学院,天津 300457;
3. 天津大学人工智能学院,天津 300354)

摘 要: 现实世界中的数据普遍存在缺失,此类数据缺失问题会显著劣化机器学习模型的性能表现。应对数据缺失的代表性方法为数据插补,即基于已观测到的数据对缺失数据进行估计,进而利用插补后的完整数据开展模型训练。然而,若插补过程中存在估计偏差,则可能引入额外噪声,往往会阻碍模型的学习过程。例如,低质量的插补往往带有估计偏差,这会影响数据的真实分布。模型若基于这些有偏样本建立映射关系,将难以捕捉数据的本质规律,进而导致泛化性能显著衰减。为此,本文提出了一种新的处理缺失数据的方法——面向缺失数据的选择性学习(Selective Learning for Incomplete Data,SLID),旨在有效利用缺失数据进行学习。SLID致力于构建一个能够动态感知数据质量的可靠学习框架。该方法在训练过程中引入动态评估机制,将被动接受插补数据转变为主动筛选:对于高置信度的插补样本,模型将其作为有效特征进行充分学习;而对于低置信度的样本,则引入正则化策略将其视为不可靠信息进行平滑处理。通过这种选择性学习机制,有效地规避了对有偏分布的过拟合,从而显著提升了泛化性能。基于多个数据集上的大量实验结果表明,与以往方法相比,SLID在准确性与可靠性方面均取得了显著提升。

关键词: 选择性学习;缺失数据;插补;低质量样本
基金项目: 国家重点研发计划(No.2025YFF0515300);国家自然科学基金(No.624B2100)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2026)04-1775-14
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260303

Selective Learning for Incomplete Medical Data

HAN Zongbo[1], XI Yuhan[2], SHENG Peizhuo[3], LIAO Yun[2*], ZHANG Changqing[3]
(1. *School of Information and Communication Engineering*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*;
2. *College of Artificial Intelligence*, *Tianjin University of Science and Technology*, *Tianjin* 300457, *China*;
3. *College of Artificial Intelligence*, *Tianjin University*, *Tianjin* 300354, *China*)

Abstract: In real-world applications, data commonly suffer from missing values, which can substantially degrade the performance of machine learning models. A widely adopted approach to handling missing data is data imputation, where missing entries are estimated from observed data and models are subsequently trained on the completed dataset. However, imputation inevitably introduces estimation errors, and inaccurate imputations may inject additional noise that hinders effective learning. In particular, low-quality imputations often exhibit systematic bias, distorting the true underlying data distribution. When models are trained on such biased samples, they struggle to capture the intrinsic data-generating patterns, resulting in a pronounced decline in generalization performance. To address this issue, we propose a novel framework for learning with missing data, termed selective learning for incomplete data (SLID), which aims to more effectively exploit incomplete data during model training. SLID is designed as a reliable learning paradigm that dynamically accounts for data quality. Specifically, it introduces a dynamic assessment mechanism that transforms the conventional passive reliance on imputed data into an active selection process. Imputed samples with high confidence are treated as reliable and are fully leveraged for learning, whereas low-confidence samples are regarded as unreliable and are accordingly smoothed through regularization strategies. Through this selective learning mechanism, retaining informative components while suppressing misleading ones, SLID effectively avoids overfitting to biased distributions and substantially enhances generalization performance. Extensive experimental results on multiple datasets demonstrate that, compared with existing approaches, SLID achieves significant improvements in both accuracy and reliability.

Keywords: selective learning; incomplete data; imputation; low-quality sample
Foundation Item(s): National Key Research and Development Program of China (No.2025YFF0515300); National Natural Science Foundation of China (No.624B2100)

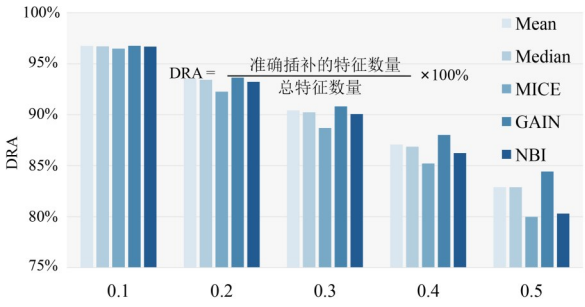
收稿日期:2026-03-27;录用日期:2026-04-08;责任编委:唐旭
*通信作者:廖芸

0 引言

在实际应用场景中，受限于数据收集过程的不完善及隐私保护等因素，数据缺失现象普遍存在，这种问题在现实应用中也广泛存在[1-3]。不完整的数据集往往包含复杂的缺失模式，严重制约了模型的性能上限[4]。为应对训练和测试阶段的缺失数据挑战，数据插补(data imputation)被广泛采用[5]。这类方法基于观测数据对缺失特征进行估计与填充，旨在提高数据利用率并尽最大程度抑制缺失对模型性能带来的负面影响[6]，进而利用插补后的完整数据进行模型训练。然而，插补过程不可避免地会引入噪声特征，进而影响下游预测模型的性能表现[7]。特别是当原始数据缺失率较高时，插补引入的估计偏差往往会扭曲原始数据的真实分布[1,8]。

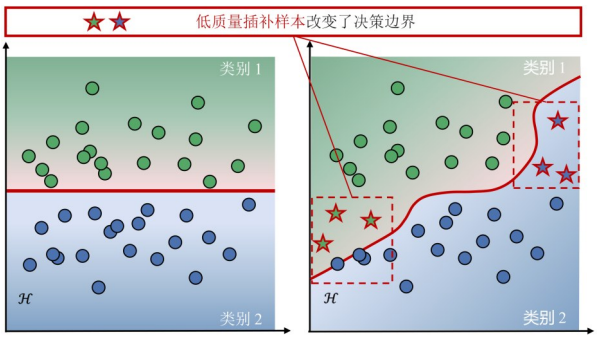
为验证上述观点，本文在心脏病(Heart Disease，HD)数据集[9]上，针对不同缺失率设定采用了多种插补方法，对其数据重建精度(Data Reconstruction Accuracy，DRA)[10]进行了系统性检验与分析。如图1(a)所示，插补后的数据包含大量不准确特征，且随着缺失率升高，DRA呈现逐步下降趋势。进一步分析表明，当利用这些插补数据训练机器学习模型时，这些不准确特征会对模型产生负面影响。具体而言，如图1(b)所示，当使用理想的完整数据进行训练时，模型能够有效习得合理的分类决策边界；而如图1(c)所示，受低质量插补样本的干扰，模型在训练过程中极易受到噪声样本负面影响，使得分类决策边界偏离理想状态。此类偏差在实际应用中可能造成严重后果，导致决策失误[11]。因此亟需提出一种新方法，使模型能够在面对因插补带来的不确定性时，能更有效地利用插补数据，显著削弱不准确插补对模型的负面影响，从而实现更可靠的预测。

应对上述噪声干扰的一种经典策略是通过正则化方法抑制模型对低质量样本的过拟合，即通过限制模型复杂度，迫使其关注数据的整体趋势，而非捕捉由插补偏差导致的虚假关联[12]。例如，标签平滑(Label Smoothing，LS)[13]通过软化原本绝对的“硬标签”，将不确定性融入分类损失中；置信度惩罚(Penalty Confidence，PC)方法[14]则通过对模型的过高置信度输出施加惩罚，抑制深度神经网络置信度。这两者本质上都是为了抑制模型对(潜在有偏的)样本产生盲目的高置信度。然而，这些方法虽然在一定程度上缓解了对低质量样本的过拟合，却因采用“无差别”的统一正则化策略而引发了新问题[15]：它们忽略了样本间显著的质量差异。类似在伪标签学习场景中，若不区分伪标签质量而进行统一优化，容易导致噪声标签误导模型训练[16]。一方面，对于观测到的高质量样本，模型本应保持高置信度以充分提取特征，却被正则化机制强行抑制，导致有效信息流失；另一方面，对于低质量的插补样本，尽管数据本身存在严重缺陷，这些方法依然强制模型基于此计算分类损失。这种在低质量数据上的强行优化，即便辅以正则化，也难以从根本上消除噪声对模型训练的误导。



(a) 插补数据中包含大量不准确特征

(a) Data imputation introduces a significant number of inaccurate values into the dataset



(b) 理想的决策边界 (c) 有偏的决策边界

(b) Optimal decision boundary (c) Biased decision boundary

图1 设计SLID方法的动机

Figure 1 Motivation for the SLID method

为解决上述问题，本文提出一种简单高效的缺失数据的处理方法——面向缺失数据的选择性学习(Selective Learning for Incomplete Data，SLID)，旨在从插补数据集中实现样本级的选择性学习。具体而言，SLID在训练过程中嵌入了基于不确定性的动态样本质量评估机制，能够对每个插补样本的质量进行判别式评估。对于高质量样本，采用经验风险最小化(Empirical Risk Minimization，ERM)[17]以充分挖掘其特征信息；而对于高概率属于低质量插补的样本，引入针对性正则化项抑制模型的置信度，避免模型在此类样本上发生过拟合。这种机制不仅维持了模型在高质量数据上的性能，还能赋予模型识别并规避不可靠样本的能力，从而在测试阶段避免潜在的决策错误。尤其对于与目标分布相近但实际并不可靠的样本，模型对非理想样本辨识能力的提升有助于增强预

测结果的稳健性与可靠性[18-19]。大量实验结果显示,SLID显著提升了模型的预测准确性与可靠性。

1　相关工作

1.1　缺失数据学习

现实场景中,受数据采集缺陷、隐私保护限制等因素影响,低质量数据现象广泛存在。数据质量下降会干扰模型有效表征的学习,并削弱预测结果的可靠性[20]。在医疗等高风险领域,该问题往往表现为数据缺失,多模态医学图像也常因采集条件受限而出现模态缺失,进而显著削弱模型的预测稳定性与实际可部署性[21]。不完整数据会严重制约下游模型的性能上限[1-2]。缺失数据处理的核心目标,在于提升数据可用性、保障分析结果的准确性与可靠性,最大限度削减乃至消除缺失数据对模型训练与推理全流程的负面影响[22-23]。现有缺失数据处理方法可主要划分为两大范式[4]:删除法(deletion methods)与插补法(imputation methods)。删除法的核心逻辑是直接剔除数据集中含缺失值的样本或特征维度[24],该方法在低缺失率的大规模数据集场景下具备简洁高效的优势。但当数据缺失不满足完全随机缺失(Missing Completely At Random, MCAR)假设时,极易造成有效信息的大量流失,甚至引入系统性偏差,严重破坏数据的原始分布特性[25]。相较而言,插补法通过基于观测数据对缺失特征进行估计与填充,最大限度地保留数据集的样本量与特征完整性[26],是当前工业界与学术界应用最广泛的缺失数据处理方案。根据插补机制的差异,现有插补方法可分为单次插补(Single Imputation, SI)与多重插补(Multiple Imputation, MI)两大类别[1],方法的选型通常需适配数据分布特性、缺失模式与下游业务场景的核心需求[27]。尽管插补法能显著降低缺失带来的信息损耗[28],但其性能高度依赖插补精度。不准确的插补结果会不可避免地引入噪声特征与估计偏差,尤其在高缺失率场景下,插补偏差会严重扭曲原始数据的真实分布,进而对下游模型的训练与决策产生不可逆的负面影响[1,8,29]。为应对这一挑战,最新研究尝试将扩散模型与期望最大化算法相结合,通过学习数据的无条件分布实现任意缺失模式下的鲁棒插补[30]。此外,已有研究开始探索基于大语言模型推理的零样本插补方法,通过多智能体协同建模时空依赖关系,在无需额外训练数据的条件下实现鲁棒插补,为缺失数据处理提供了新的思路[31]。然而,这类方法通常需要复杂的模型设计和大量的计算资源,且其插补结果仍可能存在不确定性。

1.2　正则化方法

针对插补数据带来的噪声干扰与过拟合风险,正则化(regularization)是机器学习领域抑制模型过拟合、提升泛化性能的核心技术手段[32-34]。主流正则化方法可分为隐式正则化(implicit regularization)与显式正则化(explicit regularization)两大范式[35-36]。隐式正则化无需在目标函数中引入显式约束项,而是通过优化学习流程与训练策略间接调控模型复杂度[37],典型代表包括LS[13]与PC[14]。其中,LS通过软化分类任务中的“硬标签”,将标签不确定性融入损失计算;PC则通过对模型的过高置信度输出施加惩罚,抑制深度神经网络的过置信问题。二者核心逻辑一致,均是通过全局约束抑制模型对潜在有偏样本产生盲目高置信度[15]。显式正则化则通过在目标函数中引入额外的约束项与惩罚项,直接控制模型复杂度以降低过拟合风险[38]。该类方法具备更强的针对性,更易实现对不同样本的差异化约束,典型方法包括聚焦难例样本优化的焦点损失(Focal Loss, FL)[39]、双焦点损失(Dual Focal Loss, DFL)[40]等。近年来,样本选择策略受到广泛关注。该类方法在训练过程中基于损失值动态筛选高质量样本进行优化,有效避免了模型对噪声标签的过拟合[41]。例如,有研究提出样本类不确定性抽样策略,利用模型输出的不确定性动态调整样本使用概率,以减少简单样本的重复训练并增强对边界样本的关注,从而提升训练效率与模型性能[42];也有研究在多模态命名实体识别任务中引入置信学习机制,通过评估多模态表征质量并对低置信度标签进行替换与融合,减弱模态偏差带来的错误监督,提升模型鲁棒性[43]。然而,现有研究多聚焦于标签噪声场景,对插补特征噪声的适配性仍有待探索。国内学者近年来也从鲁棒优化角度研究缺失数据学习问题。本文提出的SLID即属于显式正则化的技术范畴。

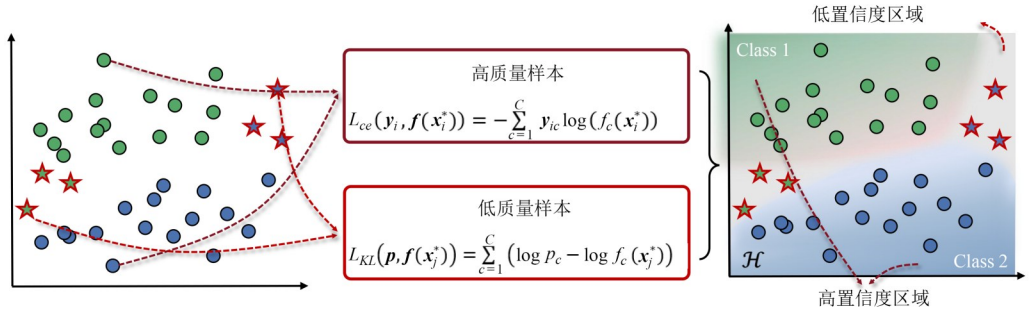
1.3　与现有方法比较

尽管以MI为代表的先进插补方法,已在插补流程中纳入了对数据不确定性的量化与考量,可在一定程度上缓解不准确插补带来的负面效应,但仍无法从根源上消除低质量插补样本对下游模型训练的误导,因此仍需依赖正则化方法抑制模型对低质量插补样本的过拟合风险。然而,现有正则化方法难以适配插补数据集的样本质量异质性,无法充分解决不准确插补带来的核心技术难题。一方面,以LS、PC为代表的隐式正则化方法,对所有样本采用“无差别”的统一正则化策略[15,44],既会强行抑制模型对高质量观测样本本应保持的高置信度,造成有效特征信息的流失,也无法规避模型在低质量插补样本上的强行优

化，难以从根本上消除噪声样本对模型的误导；另一方面，现有显式正则化方法多聚焦于样本的分类难易程度，而非插补带来的样本质量差异，在处理插补数据集时，无法精准区分高质量与低质量样本，易造成模型有效信息提取不足、泛化性能劣化等问题。针对上述技术瓶颈，本文提出SLID，构建了基于不确定性的动态样本质量评估机制，实现了训练过程中对高质量与低质量插补样本的自适应区分。该方法对高质量样本采用经验风险最小化策略以充分挖掘有效特征信息，仅对低质量样本施加针对性的正则化约束，在保障模型充分学习高质量样本核心信息的同时，显著削弱低质量插补样本对模型的误导，最终实现对不完整数据的高效、鲁棒选择性学习。

2 方法

针对前述插补噪声引发的模型过拟合、决策边界偏移，以及现有无差别正则化方法存在的有效信息流失、噪声误导无法根除的核心问题，本文提出SLID，使模型能够对插补后的不完全数据实现样本级的差异化学习，最终输出更具可靠性与泛化性的预测结果。本节将首先阐述缺失数据与插补任务的基本设定，随后对插补数据的分布特性进行建模分析，最后详细拆解SLID的核心机制与完整实现流程，SLID的整体框架如图2所示。



注：SLID可对低质量样本动态施加正则化约束。针对高质量样本，模型通过最小化其交叉熵损失优化参数，以构建稳定的分类能力；反之，针对低质量样本，模型通过执行正则化，避免模型对这类样本过拟合并抑制其过度自信的预测行为。

图2 SLID方法的整体框架

Figure 2 The overall framework of the SLID method

任务基本设定。本文针对分类任务展开研究，给定训练样本为$(\boldsymbol{x},\boldsymbol{y})$，其中$\boldsymbol{x}=\left[x^1,x^2,\cdots,x^D\right]\in\mathbb{R}^D$为$D$维特征向量，$\boldsymbol{y}$为样本对应的类别标签。当样本存在特征缺失时，引入与特征维度一一对应的掩码向量$\boldsymbol{s}=\left[s^1,s^2,\cdots,s^D\right]\in\{0,1\}^D$标识各特征的观测状态。当$s^d=1$时，代表第$d$维特征$x^d$为真实观测值；当$s^d=0$时，代表第$d$维特征$x^d$存在缺失。基于上述定义，给定包含$n$个样本的原始不完全数据集$\mathcal{D}=\left(\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{s}_i\right)_{i=1}^n$，缺失数据插补的核心目标是通过构建数据生成函数，对所有样本的缺失特征进行合理估计与填充。当$\mathcal{D}$中所有含缺失的样本完成插补后，即可得到用于模型训练的完整插补数据集$\mathcal{D}^*=\left(\boldsymbol{x}_i^*,\boldsymbol{y}_i\right)_{i=1}^n$，其中$\boldsymbol{x}_i^*$为第$i$个样本插补后的完整特征向量。需要特别说明的是，本文的研究重点为如何降低低质量插补样本对模型训练的负面影响，构建更鲁棒的预测模型。

基于邻域信息的多重插补方法。为构建完整的实验验证流程，本文首先提出了一种基于邻域信息的多重插补方法（Neighbor-Based Imputation，NBI），以完成缺失特征的插补，其核心思想是利用同类样本的邻域信息对缺失特征进行统计估计，具体流程如下：针对待插补样本$(\boldsymbol{x},\boldsymbol{y})$，首先通过随机森林模型[45]计算数据集全量特征的重要性权重ω，基于该权重计算样本间的加权距离，以更精准地刻画样本间的特征相似性。随后，在与$(\boldsymbol{x},\boldsymbol{y})$同类别样本中筛选出$K$个最近邻样本，作为缺失特征估计的参考集。最终，针对样本中的每个缺失特征，基于K个近邻样本的对应特征值构建概率分布并完成插补。对于连续型特征，构建高斯分布[46]并以其均值作为插补值；对于离散型特征，构建多项分布[47]并以其众数作为插补值。所有缺失特征完成插补后，即可得到插补后的完整样本$(\boldsymbol{x}^*,\boldsymbol{y})$。需要注意的是，插补算法不会对SLID的核心性能产生影响。SLID具备良好的算法通用性，可与任意现有插补方法无缝结合，适配不同场景下的插补数据处理需求。

插补数据的分布特性建模。如引言所述，插补过程不可避免地会引入估计偏差，因此插补后的数据集$\mathcal{D}^*$本质上由两类样本构成：一类是原始无缺失样本与插补结果准确的高质量样本，其特征分布与真实数据分布一致；另一类是插补结果存在显著偏差的低质量样本，其特征分布被插补噪声扭曲，与真实数据分布存在偏离。基于此，本文采用Huber的η-污染模

型[48]对插补样本的整体分布P^*进行形式化建模:

$$P^*=(1-\eta)P^*_{\text{acc}}+\eta P^*_{\text{inac}} \tag{1}$$

其中,P^*_{acc}为高质量样本的真实特征分布,P^*_{inac}为受插补噪声污染的低质量样本分布;η为污染率参数,表征数据集中低质量插补样本的占比。对于特定的数据集与插补方法,η为固定但未知的常数,在实际模型训练中,可将其作为可调节的超参数进行优化。从模型学习的角度来看,低质量插补样本所携带的噪声与虚假关联,会误导模型的训练过程。若不加区分地使用所有插补样本训练模型,模型极易对低质量样本产生过拟合,习得偏离真实规律的决策边界,最终导致模型在测试集上的泛化性能与决策可靠性显著下降。这也是本文设计SLID的核心出发点。

本节提出的面向缺失数据的选择性学习(SLID),旨在解决数据集中低质量插补样本引发的负面影响。SLID的核心目标在于,对插补数据集$\mathcal{D}^*=\{(\boldsymbol{x}_i^*,\boldsymbol{y}_i)\}_{i=1}^n$进行有效学习,抑制低质量样本对模型训练的干扰。在训练阶段,SLID能自动区分两类样本:一类是遵循P^*_{acc}分布的高质量样本,另一类是遵循P^*_{inac}分布的低质量插补样本。对高质量样本,采用经验风险最小化(Empirical Risk Minimization, ERM)准则对模型参数进行优化;对低质量样本,引入正则化约束,促使模型对该类样本输出较低的预测置信度。

基于采用经验风险最小化的标准训练。对于插补后的完整数据集$\mathcal{D}^*=(\boldsymbol{x}_i^*,\boldsymbol{y}_i)_{i=1}^n$,主流的模型训练范式均基于经验风险最小化准则,其优化目标为

$$\min_f \frac{1}{n}\sum_{i=1}^{n} L\left(\boldsymbol{y}_i,\boldsymbol{f}(\boldsymbol{x}_i^*)\right) \tag{2}$$

其中,$f:\boldsymbol{x}\to\boldsymbol{y}$为待训练的预测模型;$L$为损失函数。针对分类任务,本文采用领域内广泛使用的交叉熵损失(Cross-Entropy Loss,CEL)[49]作为基础损失函数,其定义为

$$L_{\text{ce}}\left(\boldsymbol{y}_i,\boldsymbol{f}(\boldsymbol{x}_i^*)\right)=-\sum_{c=1}^{C}\boldsymbol{y}_{ic}\log\left(f_c(\boldsymbol{x}_i^*)\right) \tag{3}$$

其中,C为分类任务的总类别数;$\boldsymbol{y}_{ic}$为样本i对应类别c的真实标签(one-hot编码);$f_c(\boldsymbol{x}_i^*)$为模型对样本i属于类别c的预测概率。然而,插补数据集$\mathcal{D}^*$中包含一定占比的低质量噪声样本。若直接基于ERM准则对全量样本进行无差别优化,模型会被迫拟合低质量样本中的虚假特征关联,进而产生过拟合,最终导致决策边界偏离真实分布,严重削弱模型的泛化能力与决策可靠性。

正则化训练局限性。为抑制模型对噪声样本的过拟合,现有研究通常引入正则化策略对ERM训练范式进行修正。这些方法可大致分为隐式正则化与显式正则化两类,但在插补数据场景下均存在显著的局限性。隐式正则化方法以LS为代表,通过对全量样本的硬标签进行统一软化,将不确定性融入分类损失,从而抑制模型对样本的盲目高置信度拟合。但这类方法采用“无差别”的正则化策略,对高质量样本与低质量样本施加同等强度的约束,导致模型本应充分拟合的高质量真实样本的特征信息被正则化抑制,造成有效信息流失,反而限制了模型的性能上限。显式正则化方法以FL、DFL为代表,通过在损失函数中引入重加权机制,强化模型对高损失样本的关注,从而提升对难分样本的学习能力。但在插补数据场景中,低质量样本因特征偏差通常表现为高损失难分样本,这类方法会进一步放大模型对低质量样本的拟合权重,加剧模型对噪声的过拟合与决策边界的偏移,无法从根本上消除插补噪声的负面影响。综上,现有正则化方法无法实现“高质量样本充分学习、低质量样本有效规避”的目标,难以适配插补数据的学习需求。

基于损失排序的低质量样本动态识别机制。针对现有方法的核心缺陷,SLID的核心设计思路是实现样本级的差异化学习:对高质量样本充分拟合以提取真实特征规律,对低质量样本施加针对性正则化以规避噪声误导。要实现这一目标,首要步骤是在训练过程中对样本质量进行动态、高效的判别。受“低质量插补样本因特征与标签的关联被噪声扭曲,模型对其拟合难度更高,通常会产生更高的交叉熵损失”[50]这一核心特性的启发,本文基于训练过程中样本的实时交叉熵损失,设计了一种非参数化的动态样本质量识别机制,具体流程如下:在训练过程中,给定n_b个样本,依据式(4)计算每个样本的交叉熵损失L_{ce},随后按照交叉熵损失值升序对样本进行排列:

$$I=\text{argsort}\left(L_{\text{ce},1},L_{\text{ce},2},\cdots,L_{\text{ce},n_b}\right) \tag{4}$$

其中,argsort(·)为升序排序函数,排序越靠前的样本,交叉熵损失越低,模型拟合效果越好,属于高质量样本的概率越高;反之则属于低质量样本的概率越高。然后,基于式(1)所描述的Huber的η-污染模型假设,预先设定低质量样本占比参数η,将损失值最小的$k_b=(1-\eta)\times n_b$个样本判定为高质量样本,并将其归为高质量样本集合S_{acc}:

$$S_{\text{acc}}=\left\{\left(\boldsymbol{x}_j^*,\boldsymbol{y}_j\right):j\in I_{[0:k_b]}\right\} \tag{5}$$

相应地,剩余样本构成低质量插补样本集合S_{inac}:

$$S_{\text{inac}}=\left\{\left(\boldsymbol{x}_j^*,\boldsymbol{y}_j\right):j\in I_{[k_b:n_b]}\right\} \tag{6}$$

本文采用上述二元划分策略对样本质量进行判别,而非连续型的质量权重赋值,核心原因在于:其

一，该策略具备极简的实现逻辑与极高的计算效率，无需额外引入可学习参数，不会增加模型的训练开销，适配大规模数据集与实时训练场景；其二，二元划分显著降低了超参数调优的复杂度，仅需调整 η 一个超参数即可实现样本质量的有效判别，避免了复杂阈值设计带来的调优负担；其三，硬划分机制能更彻底地切断低质量样本对模型分类损失的误导，从根源上避免模型拟合噪声样本中的虚假关联，相比连续权重赋值更能保障模型对高质量样本的充分学习。

选择性学习。将样本划分为高质量样本 S_{acc} 和低质量样本 S_{inac} 后，SLID 的核心目标是使模型能够从高质量样本集合 S_{acc} 中习得有效信息，同时最大限度削弱低质量插补样本 S_{inac} 对模型训练的干扰效应。为此，针对高质量样本，直接采用交叉熵损失对模型进行训练，其损失函数定义为

$$L_{\text{acc}}=\sum_{\left(\boldsymbol{x}_j^*,\boldsymbol{y}_j\right)\in S_{\text{acc}}} L_{\text{ce}}\left(\boldsymbol{y}_j,\boldsymbol{f}\left(\boldsymbol{x}_j^*\right)\right) \tag{7}$$

与之相对，针对低质量样本集合 S_{inac}，需施加正则化约束。具体而言，计算这些样本的先验类别概率分布 $\boldsymbol{p}$ 与模型预测类别概率 $\boldsymbol{f}\left(\boldsymbol{x}_j^*\right)$ 之间的 KL 散度（Kullback-Leibler divergence）[51]，则低质量样本集合 S_{inac} 的正则化损失可表示为

$$L_{\text{inac}}=\sum_{\left(\boldsymbol{x}_j^*,\boldsymbol{y}_j\right)\in S_{\text{inac}}} L_{\text{KL}}\left(\boldsymbol{p},\boldsymbol{f}\left(\boldsymbol{x}_j^*\right)\right) \tag{8}$$

其中，L_{KL} 表示 KL 散度，用于量化两个分布之间的差异，其数学计算公式为

$$L_{\text{KL}}\left(\boldsymbol{p},\boldsymbol{f}(\boldsymbol{x}_j^*)\right)=\sum_{c=1}^{C} p_c\left(\log p_c-\log f_c(\boldsymbol{x}_j^*)\right) \tag{9}$$

先验类别概率分布 $\boldsymbol{p}=\left(p_1,p_2,\cdots,p_c\right)$ 表示各类别的先验概率分布，$\boldsymbol{f}(\boldsymbol{x}_j^*)=\left(f_1(\boldsymbol{x}_j^*),f_2(\boldsymbol{x}_j^*),\cdots,f_c(\boldsymbol{x}_j^*)\right)$ 表示模型输出的预测的类别概率分布，p_c 和 $f_c(\boldsymbol{x}_j^*)$ 分别对应类别 c 的先验概率和预测概率。此处，先验概率基于数据集中的类别频率进行估计，用于表征在观测到特定样本之前各类别的相对可能性。类别频率通过计算每个类别样本占数据集中总样本数的比例得到。采用该先验概率估计方式的核心优势在于，它能够直观地捕捉类别分布的特征，即使对于类别不平衡数据集，也能提供合理的先验估计，从而增强模型的稳定性和鲁棒性。此外，基于类别频率计算先验概率的过程简洁高效，且具有数据集无关性，适用于多种应用场景。基于上述定义，一个训练批次中所有样本的总损失函数可计算为

$$L=L_{\text{acc}}+\alpha L_{\text{inac}} \tag{10}$$

其中，α 为正则化强度超参数，用于调控正则化项对模型训练的约束程度。需要特别说明的是，为确保该方法对低质量样本的精准识别效果，通常先采用交叉熵损失 L_{ce} 对模型进行 e 个训练轮次的预训练，待完成该预训练阶段后，再执行前述正则化策略对低质量样本施加约束。SLID 的完整执行逻辑如算法 1 所示。

算法 1 SLID 训练的伪代码

输入：插补后训练数据集 $\mathcal{D}^*$，低质量数据比例超参数 η，正则化强度 α

```
FOR epoch = e to max_epochs DO
    FOR 样本批次 b DO
1. 根据式(3)计算每个样本的交叉熵损失 L_ce;
2. 根据式(4)对交叉熵损失值升序对样本进行排列;
3. 将样本划分为高质量样本(式(5))和低质量样本(式(6));
4. 针对高质量样本集合 S_acc，用式(7)计算其损失;
5. 针对低质量样本集合 S_inac，用式(8)计算其损失;
6. 通过最小化总损失函数 L = L_acc + αL_inac 训练模型 f;
    END
END
```

输出：训练完成的模型 f

3 实验

为验证 SLID 的有效性与优越性，本文在多个真实医疗数据集上开展系统性实验研究。通过针对性设计实验解答以下 5 个关键问题，全面评估 SLID 在不同场景下的性能表现。Q1 有效性验证：SLID 在预测准确率和置信度校准效果方面是否显著优于其他主流正则化方法？Q2 鲁棒性验证 1：面对不同特征缺失率时，SLID 能否保持稳定的预测性能？Q3 鲁棒性验证 2：在不同插补方法的适配场景下，SLID 是否仍能保持优异性能？Q4 消融实验：若不引入低质量样本的正则化机制，模型性能会产生何种变化？Q5 超参数敏感性分析：超参数 η 对实验结果有何影响，其调整规律是否与理论预期一致？

3.1 数据集说明

医疗领域普遍存在数据缺失问题，且预测错误可能引发严重临床风险，因此本文选择医疗数据集作为实验对象，以贴合实际应用场景。为全面验证 SLID 的有效性与泛化能力，实验选用了四个不同临床场景的医疗数据集，分别为 HD 数据集[9]、科英布拉乳腺癌（Coimbra Breast Cancer，CBC）数据集[52]、脊柱（Vertebral Column，VC）数据集[53]和克利夫兰-匈牙利心脏病（Cleveland Hungarian Statlog Heart，CHSH）数据集[54]。上述数据集均可公开获取，并在发布前均已由原始提供方进行了严格的匿名化和脱敏处理，去除了所有可能识别患者身份的敏感信息。为模拟真实临床数据的缺失场景，实验通过随机掩码方式为各数据集引入不同缺失率的缺失值（缺失模式遵循完全随机缺失假设）。

3.2 评估指标与对比方法

为系统性评估SLID的综合性能,本文重点围绕预测准确率和置信度校准效果两大核心维度设计评估指标,各指标定义与优化方向如下:

(1)准确度(Accuracy,ACC)(↑)[55]。量化模型正确预测样本占总样本的比例,直观反映整体分类效果。

(2)受试者工作特征曲线下面积(Area Under the ROC Curve,AUC)(↑)[56]。评估模型区分不同类别的判别能力,尤其适用于类别不平衡场景。

(3)尤登指数(Youden's Index,YI)(↑)[57]。计算为灵敏度与特异度之和减一,综合表征模型的诊断效能。

(4)F1分数(F1 Score,F1)(↑)[58]。计算模型精确度与灵敏度的调和平均值,用于平衡模型对正负样本的分类性能。

此外,通过以下两个指标评估置信度校准性能:

(1)预期校准误差(Expected Calibration Error,ECE)(↓)[59]。用于评估模型的校准效果,反映预测概率与实际概率的一致性。

(2)Brier分数(Brier Score, Brier)(↓)[60]。通过计算预测概率与实际概率之间的平方误差,衡量概率预测的准确性。

其中,↑表示数值越高模型性能越优,↓表示数值越低模型性能越优。

对比方法。为验证SLID的性能优越性,本文选取7种主流基线方法进行对比,涵盖传统训练范式、经典正则化方法及消融对照方法,具体如下:

(1)经验风险最小化(ERM)[17]。通过最小化训练数据集上的平均交叉熵损失来确定最优模型参数。

(2)标签平滑(LS)[13]。通过在训练过程中为非目标类别分配较小概率值以软化标签,增强模型的泛化能力,对预测置信度过高的样本施加惩罚。

(3)罚则置信(PC)[14]。通过对预测置信度过高的样本施加惩罚,提升模型的泛化性能,促使模型做出更谨慎的预测。

(4)焦点损失(FL)[39]。通过降低易分类样本的权重占比、提升难分类样本的贡献度,强化模型对少数类样本的识别能力。

(5)双焦点损失(DFL)[40]。融合型损失函数,将焦点损失与边界损失相结合,旨在同时优化分类任务中正负样本的平衡性和边界样本的准确性,从而提升模型性能。

(6)样本依赖焦点损失(Sample-Dependent Focal Loss,SDFL)[39]。一种权重动态调整的损失函数,根据每个样本的分类难度自适应分配权重,增强模型对难分类样本的关注,同时削弱易分类样本的冗余影响,有效解决类别不平衡问题。

(7)移除低质量样本(Remove Low-Quality Samples,RLS)。消融实验对照组,在训练过程中直接丢弃低质量样本集合,用于验证SLID正则化策略的必要性与有效性。

插补方法。除上述基于邻域的插补(NBI)方法外,为验证SLID在不同插补方法下的适配性与鲁棒性,本文还选取以下五类主流插补方法对缺失数据进行处理:均值插补(mean imputation)[1]、中位数插补(median imputation)[1]、链式方程多元插补(Multivariate Imputation by Chained Equations, MICE)[61]、生成对抗网络插补(Generative Adversarial Imputation Nets,GAIN)[62]和变分自编码器多重插补(Multiple Imputation With variational Autoencoders,MIWAE)[63]。

3.3 实验结果与分析

为全面评估SLID的性能,我们开展了一系列实验以解答本节提出的关键问题。

Q1有效性验证。为验证SLID在预测准确率和置信度校准方面的有效性,实验在缺失率设定为50%的四个医疗数据集上,均采用NBI方法处理缺失数据,将SLID与各类基线方法进行全面性能对比,实验结果如表1~表3所示。

从预测准确率来看,SLID在所有数据集的ACC指标上均表现稳定,显著优于其他所有基线方法。在四个数据集上,SLID的ACC较次优方法平均提升3.21%,其中在CBC数据集上的性能提升幅度最大,达5.42%,这一结果表明,SLID在乳腺癌诊断这类高风险场景中具有更强的实用价值。在YI和F1指标方面,SLID较次优方法的平均提升幅度分别达到4.19%和3.19%,这一现象说明SLID不仅具备优异的整体分类性能,其对各类别样本的识别能力也具有较强的鲁棒性,能够有效适配医疗数据中可能存在的类别不平衡场景。

在置信度校准方面,SLID在ECE和Brier两项指标上均表现最优,其中Brier分数较次优方法平均降低1.29%,充分凸显了SLID在概率预测精准度和置信度校准能力方面的显著优势。SLID在四个不同临床场景医疗数据集上的一致出色表现,充分证明其具备良好的泛化能力。核心原因在于,SLID通过对低质量样本施加针对性正则化约束,既能够充分挖掘此类样本中潜在的有效信息,又能为模型提供更明确的置信度监督信号,最终实现了"精准预测"与"合理置信度估计"的双重目标,完美解决了传统方法"要么无差别拟合、要么盲目丢弃"的核心痛点。

Q2鲁棒性验证1。为验证SLID在不同数据缺失率下的鲁棒性,本文以不同缺失率(μ)的HD数据集

表1 在HD和CBC数据集上的预测性能对比 单位:%

Table 1 Predictive performance comparison on HD and CBC datasets unit: %

数据集	HD				CBC			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	65.08$_{(6.91)}$	71.02$_{(6.76)}$	29.99$_{(14.18)}$	67.12$_{(6.62)}$	50.83$_{(7.56)}$	51.22$_{(11.34)}$	−2.10$_{(13.44)}$	58.68$_{(14.88)}$
LS	63.77$_{(3.04)}$	71.28$_{(5.13)}$	26.97$_{(6.39)}$	66.61$_{(2.97)}$	53.33$_{(2.64)}$	60.04$_{(11.45)}$	1.68$_{(5.94)}$	64.64$_{(5.51)}$
PC	67.05$_{(3.82)}$	73.57$_{(4.09)}$	33.52$_{(7.51)}$	69.68$_{(4.17)}$	52.08$_{(7.92)}$	50.70$_{(2.98)}$	0.07$_{(17.33)}$	61.44$_{(9.22)}$
FL	66.72$_{(5.01)}$	72.86$_{(3.64)}$	33.29$_{(10.03)}$	68.47$_{(5.81)}$	47.50$_{(8.38)}$	50.98$_{(16.94)}$	−8.67$_{(16.06)}$	56.83$_{(11.74)}$
DFL	66.07$_{(8.33)}$	71.75$_{(5.69)}$	31.54$_{(15.69)}$	68.18$_{(10.97)}$	50.00$_{(11.45)}$	57.83$_{(19.19)}$	−5.18$_{(22.08)}$	62.24$_{(11.17)}$
SDFL	60.49$_{(7.10)}$	64.11$_{(4.80)}$	19.77$_{(13.58)}$	64.29$_{(9.09)}$	52.50$_{(4.03)}$	57.17$_{(6.53)}$	1.54$_{(8.45)}$	60.38$_{(10.11)}$
RLS	65.57$_{(6.37)}$	70.43$_{(4.99)}$	31.11$_{(11.94)}$	66.87$_{(9.27)}$	53.33$_{(8.29)}$	54.09$_{(14.52)}$	3.78$_{(15.73)}$	57.76$_{(21.77)}$
OUR	69.18$_{(5.88)}$	75.21$_{(5.15)}$	36.65$_{(12.42)}$	73.48$_{(4.53)}$	58.75$_{(9.31)}$	63.36$_{(13.20)}$	14.62$_{(17.91)}$	66.29$_{(7.77)}$

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

表2 在VC,CHSH数据集上的预测性能对比 单位:%

Table 2 Predictive performance comparison on VC and CHSH datasets unit: %

数据集	VC				CHSH				AVG			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	69.35$_{(3.65)}$	73.63$_{(2.51)}$	35.12$_{(4.21)}$	75.92$_{(4.63)}$	73.11$_{(2.57)}$	78.98$_{(1.07)}$	45.99$_{(5.33)}$	74.68$_{(2.24)}$	64.59	68.71	27.25	69.10
LS	68.87$_{(4.30)}$	75.15$_{(3.10)}$	34.14$_{(7.48)}$	75.46$_{(5.03)}$	72.98$_{(1.67)}$	79.32$_{(1.43)}$	45.91$_{(3.32)}$	74.18$_{(1.92)}$	64.74	71.45	27.18	70.22
PC	69.19$_{(4.65)}$	72.94$_{(5.19)}$	29.38$_{(7.63)}$	77.10$_{(4.32)}$	72.56$_{(2.29)}$	77.96$_{(1.31)}$	44.66$_{(5.01)}$	74.64$_{(1.65)}$	65.22	68.79	26.91	70.71
FL	69.52$_{(3.08)}$	74.49$_{(3.15)}$	31.17$_{(11.00)}$	77.28$_{(2.22)}$	73.19$_{(2.10)}$	78.57$_{(1.99)}$	46.46$_{(3.81)}$	74.09$_{(2.52)}$	64.23	69.23	25.56	69.17
DFL	69.52$_{(4.06)}$	74.23$_{(2.07)}$	32.48$_{(6.22)}$	76.62$_{(5.06)}$	73.40$_{(1.34)}$	78.54$_{(1.43)}$	46.88$_{(2.61)}$	74.32$_{(1.63)}$	64.75	70.59	26.43	70.34
SDFL	64.84$_{(3.96)}$	65.27$_{(6.56)}$	17.46$_{(13.05)}$	74.00$_{(5.43)}$	72.86$_{(1.65)}$	78.73$_{(1.07)}$	45.62$_{(3.23)}$	74.12$_{(2.09)}$	62.67	66.32	21.10	68.20
RLS	67.58$_{(3.44)}$	69.29$_{(2.85)}$	32.24$_{(5.17)}$	74.32$_{(4.48)}$	72.27$_{(2.79)}$	77.99$_{(2.24)}$	44.13$_{(5.63)}$	74.24$_{(2.64)}$	64.69	67.95	27.82	68.30
OUR	70.81$_{(2.57)}$	73.20$_{(2.62)}$	27.81$_{(6.31)}$	79.77$_{(2.10)}$	74.83$_{(1.58)}$	79.59$_{(1.51)}$	49.38$_{(3.25)}$	76.42$_{(1.65)}$	68.39	72.84	32.11	73.99

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

表3 四个医疗数据集上与置信度校准相关指标对比 单位:%

Table 3 Confidence calibration metrics comparison across four medical datasets unit: %

数据集	HD		CBC		VC		CHSH		AVG	
指标	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier
ERM	29.23$_{(5.48)}$	58.88$_{(8.87)}$	48.88$_{(7.72)}$	96.10$_{(4.21)}$	24.39$_{(3.76)}$	51.58$_{(4.21)}$	17.00$_{(1.22)}$	42.64$_{(1.92)}$	29.87	62.30
LS	27.84$_{(4.86)}$	58.57$_{(4.79)}$	45.21$_{(4.94)}$	89.58$_{(7.48)}$	23.38$_{(4.24)}$	50.93$_{(7.48)}$	16.62$_{(2.84)}$	42.54$_{(3.11)}$	28.26	60.41
PC	26.70$_{(5.11)}$	56.30$_{(8.35)}$	47.23$_{(8.20)}$	93.04$_{(16.34)}$	24.19$_{(3.64)}$	51.06$_{(5.67)}$	15.55$_{(2.68)}$	43.05$_{(2.63)}$	28.42	60.86
FL	28.34$_{(3.70)}$	57.34$_{(7.09)}$	32.28$_{(16.41)}$	73.00$_{(24.58)}$	24.00$_{(3.72)}$	49.68$_{(4.95)}$	14.57$_{(2.07)}$	41.47$_{(2.64)}$	24.80	55.37
DFL	29.22$_{(8.35)}$	59.32$_{(15.64)}$	48.67$_{(6.22)}$	94.54$_{(21.28)}$	23.31$_{(3.18)}$	50.06$_{(5.75)}$	17.76$_{(2.49)}$	44.02$_{(2.68)}$	29.74	61.98
SDFL	10.27$_{(6.63)}$	48.22$_{(2.62)}$	41.05$_{(10.05)}$	82.38$_{(15.97)}$	10.71$_{(5.58)}$	45.42$_{(3.44)}$	9.18$_{(2.01)}$	39.24$_{(1.19)}$	17.80	53.82
RLS	30.24$_{(6.13)}$	61.48$_{(8.18)}$	45.49$_{(10.03)}$	88.63$_{(19.84)}$	31.54$_{(3.32)}$	62.31$_{(6.17)}$	22.57$_{(5.05)}$	48.11$_{(6.05)}$	32.46	65.38
OUR	16.13$_{(5.68)}$	46.00$_{(7.98)}$	38.33$_{(10.66)}$	75.72$_{(21.37)}$	19.34$_{(4.47)}$	46.10$_{(5.58)}$	11.55$_{(4.27)}$	39.19$_{(2.60)}$	22.48	52.53

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

为实验对象,系统评估了各对比方法的分类性能,其中缺失率设置为10%、20%、30%、40%和50%,各方法在不同缺失率下的ACC结果如表4和表5所示。基于实验数据可得出以下核心结论:(1)随着缺失率的逐步提升,所有方法的测试准确率(ACC)逐渐下降。这一现象的本质原因在于,缺失数据会阻碍模型学习真实的数据分布,由于插补特征存在固有误差,无法彻底抵消数据缺失引发的负面影响。例如,ERM方法在各缺失率下均表现出较低的性能,其准确率(ACC)从缺失率10%时的80.66%下降到50%时的65.08%,这一结果凸显了对插补数据进行选择性学习的必要性。(2)SLID在所有缺失率下均保持优异的分类性能,这表明SLID具备良好的鲁棒性,能够适配不同程度的缺失数据场景。其核心优势在于,通过动态识别低质量插补样本并施加针对性正则化约束,有效抑制了该类样本对模型训练的负面影响。

表4　通过NBI方法插补后的HD数据集在10%和20%缺失率(μ)下各类方法的指标结果　　单位:%

Table 4　Results of metrics for various methods on the HD dataset at at 10% and 20% missing rates (μ), with missing values imputed via the NBI method　　unit: %

μ	10% (η = 0.05)				20% (η = 0.05)			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	80.66$_{(3.52)}$	87.29$_{(3.16)}$	60.89$_{(6.61)}$	82.17$_{(3.86)}$	80.00$_{(3.85)}$	85.55$_{(2.07)}$	59.30$_{(7.44)}$	81.88$_{(3.96)}$
LS	82.46$_{(3.10)}$	87.69$_{(4.63)}$	67.36$_{(8.36)}$	84.01$_{(2.96)}$	79.84$_{(3.94)}$	84.66$_{(2.30)}$	59.21$_{(7.48)}$	81.43$_{(4.30)}$
PC	83.11$_{(4.70)}$	86.44$_{(3.56)}$	65.97$_{(9.06)}$	84.28$_{(5.51)}$	78.52$_{(4.54)}$	84.50$_{(5.90)}$	56.89$_{(8.56)}$	79.79$_{(5.09)}$
FL	83.44$_{(5.43)}$	89.79$_{(3.24)}$	66.53$_{(10.47)}$	84.66$_{(5.61)}$	76.23$_{(7.18)}$	82.59$_{(6.89)}$	52.22$_{(13.63)}$	77.58$_{(7.80)}$
DFL	78.85$_{(4.19)}$	86.45$_{(3.64)}$	57.55$_{(8.15)}$	80.18$_{(4.51)}$	79.67$_{(4.58)}$	84.21$_{(4.89)}$	58.64$_{(9.12)}$	81.55$_{(4.47)}$
SDFL	83.28$_{(3.96)}$	89.57$_{(3.50)}$	66.33$_{(7.17)}$	84.40$_{(4.25)}$	79.18$_{(3.63)}$	83.21$_{(5.09)}$	58.21$_{(7.13)}$	80.52$_{(3.76)}$
RLS	79.34$_{(4.52)}$	85.98$_{(3.88)}$	59.06$_{(8.51)}$	79.96$_{(5.40)}$	75.90$_{(4.70)}$	81.88$_{(2.80)}$	51.61$_{(9.15)}$	77.44$_{(5.00)}$
OUR	84.59$_{(2.47)}$	90.56$_{(1.92)}$	68.59$_{(5.13)}$	86.05$_{(2.18)}$	80.33$_{(3.79)}$	84.80$_{(3.45)}$	59.52$_{(7.45)}$	82.59$_{(3.67)}$

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

表5　30%和40%缺失率(μ)下各类方法的指标结果　　单位:%

Table 5　Results of metrics for various methods at 30% and 40% missing rates (μ)　　unit: %

μ	30% (η = 0.10)				40% (η = 0.20)			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	76.39$_{(4.39)}$	82.92$_{(3.76)}$	51.69$_{(9.21)}$	78.78$_{(3.44)}$	71.48$_{(8.44)}$	76.88$_{(8.95)}$	40.73$_{(17.13)}$	76.12$_{(6.96)}$
LS	77.87$_{(4.32)}$	84.92$_{(3.17)}$	54.06$_{(8.40)}$	80.92$_{(4.21)}$	74.92$_{(5.84)}$	81.95$_{(5.04)}$	47.74$_{(11.92)}$	78.89$_{(4.78)}$
PC	77.05$_{(3.46)}$	81.87$_{(3.99)}$	52.65$_{(7.26)}$	80.08$_{(2.72)}$	71.80$_{(7.64)}$	76.95$_{(7.55)}$	41.87$_{(15.10)}$	75.41$_{(8.06)}$
FL	74.26$_{(5.13)}$	82.07$_{(4.81)}$	47.45$_{(9.82)}$	76.92$_{(5.65)}$	73.11$_{(5.58)}$	78.86$_{(4.46)}$	44.40$_{(11.24)}$	76.90$_{(5.14)}$
DFL	78.69$_{(5.57)}$	84.56$_{(4.51)}$	56.44$_{(10.74)}$	80.81$_{(5.98)}$	75.41$_{(6.37)}$	81.53$_{(6.31)}$	49.08$_{(13.27)}$	79.00$_{(5.10)}$
SDFL	77.70$_{(5.74)}$	82.27$_{(7.05)}$	54.08$_{(11.46)}$	80.46$_{(5.32)}$	74.59$_{(7.01)}$	79.38$_{(8.74)}$	47.13$_{(14.01)}$	78.48$_{(6.27)}$
RLS	76.23$_{(4.18)}$	81.29$_{(4.23)}$	52.30$_{(9.72)}$	79.12$_{(3.27)}$	67.70$_{(7.33)}$	74.26$_{(5.74)}$	33.11$_{(15.43)}$	73.07$_{(5.48)}$
OUR	78.69$_{(5.01)}$	83.24$_{(4.64)}$	55.52$_{(10.44)}$	81.94$_{(3.95)}$	75.25$_{(8.62)}$	80.47$_{(9.31)}$	48.13$_{(17.54)}$	79.48$_{(7.10)}$

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

Q3 鲁棒性验证2。为验证SLID在不同插补方法处理后数据上的鲁棒性与适配能力,本文以缺失率设定为50%的HD数据集为实验对象,采用5种不同的主流插补方法处理缺失数据,并系统对比了各方法的分类性能,实验结果如表6~表8所示。基于实验数据可得出以下核心结论:(1)ERM方法在所有插补方法处理后的数据上均表现出较差的分类性能,这一现象表明,无论采用何种插补方法,生成的插补数据中均会包含一定比例的低质量样本,此类样本会严重制约模型的性能表现,也进一步验证了插补噪声对模型训练的负面影响。(2)聚焦难分类样本的传统正则化方法(如FL),其性能往往劣于ERM方法。例如,在均值插补和MIWAE插补的数据上,FL方法的ACC分别比ERM方法下降了1.97%和6.39%。这一结果表明,

表6　通过均值插补,中位数插补方法插补后的HD数据集在50%的缺失率下,各类方法的指标结果　　单位:%

Table 6　Performance of various methods on the HD dataset (50% missing rate) under mean imputation and median imputation　　unit: %

方法	Mean				Median			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	70.66$_{(6.35)}$	76.42$_{(8.68)}$	41.21$_{(13.61)}$	72.36$_{(5.57)}$	70.82$_{(4.49)}$	76.90$_{(6.56)}$	42.16$_{(9.58)}$	71.34$_{(5.00)}$
LS	68.36$_{(6.37)}$	76.95$_{(4.49)}$	37.56$_{(11.38)}$	67.56$_{(10.83)}$	70.82$_{(4.15)}$	77.49$_{(3.29)}$	42.65$_{(8.14)}$	70.52$_{(5.36)}$
PC	68.69$_{(9.70)}$	76.23$_{(7.78)}$	38.12$_{(18.75)}$	68.39$_{(11.38)}$	67.21$_{(7.65)}$	73.72$_{(5.37)}$	35.82$_{(14.92)}$	65.72$_{(10.35)}$
FL	68.69$_{(5.27)}$	75.88$_{(5.54)}$	37.68$_{(11.90)}$	69.83$_{(3.68)}$	69.18$_{(4.88)}$	75.85$_{(3.59)}$	39.35$_{(8.93)}$	68.57$_{(7.04)}$
DFL	71.15$_{(6.09)}$	77.85$_{(3.58)}$	41.58$_{(10.86)}$	72.91$_{(9.50)}$	69.18$_{(6.08)}$	76.02$_{(5.32)}$	38.59$_{(12.50)}$	70.32$_{(5.90)}$
SDFL	70.33$_{(5.32)}$	77.44$_{(5.77)}$	40.82$_{(10.22)}$	71.34$_{(6.66)}$	68.36$_{(3.79)}$	75.65$_{(4.01)}$	36.70$_{(6.58)}$	69.68$_{(5.83)}$
RLS	71.64$_{(5.36)}$	78.27$_{(7.74)}$	41.79$_{(12.13)}$	75.43$_{(3.18)}$	70.00$_{(3.71)}$	75.22$_{(4.47)}$	40.22$_{(7.14)}$	71.13$_{(4.52)}$
OUR	72.79$_{(5.01)}$	76.99$_{(4.41)}$	45.75$_{(13.33)}$	73.42$_{(8.57)}$	71.64$_{(6.78)}$	74.24$_{(7.41)}$	42.22$_{(13.88)}$	74.51$_{(6.72)}$

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

表7 通过MICE和GAIN方法插补后的结果 单位:%

Table 7 Performance of various methods on the HD dataset (50% missing rate) under the MICE and the GAIN unit: %

方法	MICE				GAIN			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	65.41$_{(6.11)}$	74.06$_{(7.17)}$	31.84$_{(12.17)}$	64.95$_{(6.79)}$	64.10$_{(8.05)}$	69.29$_{(6.23)}$	26.33$_{(16.60)}$	68.97$_{(7.52)}$
LS	67.54$_{(4.56)}$	76.06$_{(6.21)}$	35.94$_{(8.67)}$	67.05$_{(7.84)}$	66.07$_{(5.13)}$	73.45$_{(7.41)}$	30.24$_{(11.03)}$	70.17$_{(6.98)}$
PC	67.05$_{(6.11)}$	73.74$_{(5.98)}$	34.65$_{(12.49)}$	67.50$_{(7.84)}$	69.34$_{(4.02)}$	74.78$_{(2.98)}$	36.62$_{(8.57)}$	73.98$_{(3.50)}$
FL	67.38$_{(7.14)}$	74.36$_{(7.76)}$	34.34$_{(15.52)}$	69.95$_{(5.20)}$	65.90$_{(5.72)}$	69.27$_{(7.85)}$	30.10$_{(11.03)}$	70.17$_{(6.98)}$
DFL	67.21$_{(5.73)}$	72.62$_{(5.67)}$	34.69$_{(11.02)}$	68.09$_{(6.86)}$	69.83$_{(2.91)}$	75.02$_{(4.38)}$	37.86$_{(5.61)}$	73.94$_{(3.73)}$
SDFL	55.08$_{(9.72)}$	58.46$_{(10.09)}$	10.96$_{(18.91)}$	53.71$_{(13.65)}$	63.28$_{(7.46)}$	68.77$_{(7.54)}$	24.87$_{(13.73)}$	67.07$_{(11.90)}$
RLS	70.98$_{(7.77)}$	77.73$_{(7.21)}$	42.74$_{(14.93)}$	70.86$_{(8.86)}$	66.39$_{(5.14)}$	72.02$_{(5.77)}$	31.11$_{(9.76)}$	70.62$_{(6.23)}$
OUR	71.31$_{(7.58)}$	76.90$_{(7.61)}$	42.69$_{(15.20)}$	72.63$_{(7.83)}$	70.16$_{(6.50)}$	74.37$_{(5.27)}$	38.25$_{(12.82)}$	74.59$_{(6.45)}$

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

表8 通过MIWAE方法插补后的结果 单位:%

Table 8 Performance of various methods on the HD dataset (50% missing rate) under the MIWAE unit: %

方法	MIWAE				AVG			
指标	ACC	AUC	YI	F1	ACC	AUC	YI	F1
ERM	66.72$_{(6.56)}$	74.59$_{(4.73)}$	34.81$_{(12.66)}$	65.37$_{(8.42)}$	67.54	74.25	35.27	68.60
LS	67.38$_{(5.27)}$	71.69$_{(6.57)}$	35.42$_{(10.27)}$	67.65$_{(6.31)}$	68.03	75.13	36.36	68.69
PC	61.97$_{(6.11)}$	68.43$_{(5.55)}$	23.69$_{(13.97)}$	64.08$_{(6.51)}$	66.85	73.38	33.78	67.93
FL	60.33$_{(7.64)}$	63.73$_{(10.51)}$	20.61$_{(15.16)}$	61.94$_{(8.76)}$	66.30	71.82	32.42	68.09
DFL	66.23$_{(5.73)}$	71.41$_{(4.10)}$	32.76$_{(8.19)}$	67.35$_{(4.21)}$	68.72	74.58	37.10	70.52
SDFL	67.38$_{(6.06)}$	71.28$_{(6.08)}$	34.50$_{(12.03)}$	69.44$_{(6.10)}$	64.89	70.32	29.57	66.25
RLS	65.90$_{(8.49)}$	69.36$_{(8.17)}$	31.83$_{(17.25)}$	67.42$_{(8.25)}$	68.98	74.52	37.54	71.09
OUR	68.69$_{(4.91)}$	74.32$_{(4.65)}$	37.31$_{(9.62)}$	70.31$_{(5.50)}$	70.92	75.37	41.24	73.09

注:使用红色突出显示每个数据集的最佳结果,蓝色突出显示次优结果。

此类方法未能区分难分类样本的“质量属性”,将低质量插补样本误判为有价值的难分类样本,在训练过程中过度为低质量样本分配学习权重,导致模型进行有偏学习,进而造成性能下降,这也凸显了现有正则化方法在应对低质量插补样本挑战时的固有局限性。(3)SLID在所有插补方法处理后的数据上均表现出优异的鲁棒性和适配能力,其ACC均显著高于其他对比方法。例如,在MICE和GAIN两种插补方法处理后的数据上,SLID的ACC较ERM方法分别提升了5.90%和6.06%。这一结果表明,SLID具备出色的鲁棒性,其核心的样本质量识别机制能够跨越不同插补方法的差异,精确识别各类插补数据中的低质量样本,从而避免模型过拟合于有偏样本,维持稳定的分类性能,验证了SLID与各类插补方法的良好适配性。

Q4 消融实验。结合各表的实验结果可知,SLID在所有实验场景下的综合性能均持续优于ERM和RLS两种基线方法(其中RLS为本文设置的消融对照组,其核心策略为直接丢弃低质量插补样本集合)。具体而言,ERM对所有插补样本进行无差别利用(未区分样本质量),导致模型拟合低质量样本中的虚假关联,性能较差;RLS则直接丢弃低质量样本,虽能避免噪声干扰,但造成了有效数据的浪费,同样无法充分发挥插补数据的价值。而SLID通过对低质量插补样本施加针对性正则化约束,实现了对插补数据的高效且合理利用——既规避了低质量样本的噪声误导,又挖掘了其中潜在的有效信息,最终使模型的预测结果更准确、置信度估计更可靠。这一对比结果充分证明了SLID核心设计模块的必要性与优越性。

Q5 超参数敏感性分析。低质量样本的比例是影响该SLID性能的关键因素。表4和表5展示了SLID在不同缺失率的HD数据集上学习时所使用的参数η的值。实验结果表明,在HD数据集上,随着特征缺失率从10%逐步增加到50%,代表低质量样本比例的参数η也从0.05增加到0.25。此外,当特征缺失率为50%时,实验所用四个医疗数据集对应的η值均处于相对较高水平。这一实验现象与理论预期高度一致,特征缺失率越高,插补后产生的低质量样本占比越大,因此需要通过更高的η取值扩大正则化覆盖范围,以抵消此类样本的负面影响。这些结果进一步增强了SLID的可解释性,超参数η的动态调整机制

能够自适应匹配不同数据缺失场景下的低质量样本分布,证明该方法通过针对性正则化策略有效抑制了低质量插补样本对模型训练的干扰,最终实现模型预测准确性与可靠性提升。

4 结论

本文针对插补过程中引入的噪声会显著增加模型学习复杂度,导致模型在训练数据上出现过拟合、在测试数据上泛化性能衰减这一核心问题展开研究。为突破这一技术局限,本文提出一种简洁高效的选择性学习方法——面向缺失数据的选择性学习(SLID)。该方法的核心在于,能够精准区分插补数据中的高质量样本与低质量样本,并对低质量样本动态施加正则化约束,从而避免模型训练过程中产生系统性偏差。通过这一核心设计,模型既能从高质量样本中高效挖掘有效特征信息,又能有效抑制低质量样本带来的偏差干扰,避免习得虚假特征关联。在四个真实医疗数据集上的系统性实验,充分验证了SLID的有效性,SLID在预测准确率和置信度校准效果上均显著优于现有主流正则化方法。同时,实验也证明SLID在不同缺失率、不同插补方法等多种场景下均能保持稳定优异的性能,具备良好的鲁棒性和泛化能力。本文提出的SLID为医疗等对决策可靠性要求较高领域的缺失数据处理,提供了一种高效可行的解决方案。

参考文献

[1] Little R J A, Rubin D B. Statistical analysis with missing data[M]. 3rd ed. Hoboken: John Wiley & Sons, 2019.

[2] Marlin B M, Kale D C, Khemani R G, et al. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models[C]//Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. New York: ACM, 2012: 389-398.

[3] 韩悦, 臧晓尧, 李海, 等. 面向数据缺失场景的网络性能评估方法[J]. 通信学报, 2025, 46(12): 199-212.
Han Yue, Zang Xiaoyao, Li Hai, et al. Network performance evaluation method for scenarios with missing data[J]. Journal on Communication, 2025, 46(12): 199-212. (in Chinese)

[4] Schafer J L, Graham J W. Missing data: Our view of the state of the art[J]. Psychological Methods, 2002, 7(2): 147-177.

[5] Schafer J L. Analysis of incomplete multivariate data[M]. New York: CRC Press, 1997.

[6] Van Buuren S. Flexible imputation of missing data[M]. 2nd ed. Boca Raton: CRC Press, 2018.

[7] Wabina R S, Looareesuwan P, Sonsilphong S, et al. Uncertainty-aware approach for multiple imputation using conventional and machine learning models: A real-world data study[J]. Journal of Big Data, 2025, 12(1): 95.

[8] White I R, Royston P, Wood A M. Multiple imputation using chained equations: Issues and guidance for practice[J]. Statistics in Medicine, 2011, 30(4): 377-399.

[9] Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease[J]. The American Journal of Cardiology, 1989, 64(5): 304-310.

[10] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6): 520-525.

[11] Sarker I H. Machine learning: Algorithms, real-world applications and research directions[J]. SN Computer Science, 2021, 2(3): 160.

[12] Du Mengnan, Liu Ninghao, Yang Fan, et al. Learning credible deep neural networks with rationale regularization[C]//Proceedings of the 2019 IEEE International Conference on Data Mining. Piscataway: IEEE, 2019: 150-159.

[13] Müller R, Kornblith S, Hinton G E, et al. When does label smoothing help [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019: 422.

[14] Pereyra G, Tucker G, Chorowski J, et al. Regularizing neural networks by penalizing confident output distributions[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net, 2017.

[15] Wang Dengbao, Feng Lei, Zhang Minling, et al. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2021: 903.

[16] 贾洁茹, 张硕蕊, 钱宇华, 等. 基于伪标签正则化损失的无监督行人重识别[J]. 电子学报, 2024, 52(5): 1743-1758.
Jia Jieru, Zhang Shuorui, Qian Yuhua, et al. Unsupervised person re-identification with pseudo label regularization loss[J]. Acta Electronica Sinica, 2024, 52(5): 1743-1758. (in Chinese)

[17] Vapnik V N. Statistical learning theory[M]. New York: Wiley, 1998.

[18] Bai Yichen, Han Zongbo, Cao Bing, et al. ID-like prompt

learning for few-shot out-of-distribution detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 17480-17489.

[19] Han Ruisong, Han Zongbo, Zhang Jiahao, et al. Retrieval-augmented prompt for OOD detection[PP/OL]. V1. arXiv (2025-08-14)[2026-03-20]. https://arxiv.org/abs/2508.10556.

[20] Zhang Qingyang, Wei Yake, Han Zongbo, et al. Multimodal fusion on low-quality data: A comprehensive survey[PP/OL]. V3. arXiv (2024-04-27) [2026-03-20]. https://arxiv.org/abs/2404.18947.

[21] 贾熹滨, 杨川旭, 范超, 等 . SM3RNet: 一种基于空间 Mamba 的模态缺失鲁棒病灶分割网络[J]. 电子学报, 2025, 53(11): 3943-3955.
Jia Xibin, Yang Chuanxu, Fan Chao, et al. SM3RNet: Spatial Mamba based missing modality robust lesion segmentation network[J]. Acta Electronica Sinica, 2025, 53(11): 3943-3955. (in Chinese)

[22] De Leeuw E D. Reducing missing data in surveys: An overview of methods[J]. Quality and Quantity, 2001, 35(2): 147-160.

[23] Glance L G, Osler T M, Mukamel D B, et al. Impact of statistical approaches for handling missing data on trauma center quality[J]. Annals of Surgery, 2009, 249(1): 143-148.

[24] Allison P D. Missing data[M]//The SAGE handbook of quantitative methods in psychology. London: SAGE Publications, 2009: 72-89.

[25] Young W, Weckman G, Holland W. A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits[J]. Theoretical Issues in Ergonomics Science, 2011, 12(1): 15-43.

[26] Bertsimas D, Pawlowski C, Zhuo Y D. From predictive methods to missing data imputation: An optimization approach[J]. Journal of Machine Learning Research, 2017, 18(1): 7133-7171.

[27] Graham J W. Missing data analysis: Making it work in the real world[J]. Annual Review of Psychology, 2009, 60: 549-576.

[28] Pigott T D. A review of methods for missing data[J]. Educational Research and Evaluation, 2001, 7(4): 353-383.

[29] Acuña E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy[M]//Classification, clustering, and data mining applications. Berlin: Springer, 2004: 639-647.

[30] Cheng Lyu, Antoniou C. A diffusion-based expectation-maximization framework for probabilistic traffic data imputation[J]. IEEE Internet of Things Journal, 2025, 12(20): 43227-43240.

[31] 梅雅欣, 秦慧玲, 梁玉珠, 等 . 基于大语言模型的时空数据零样本插补[J]. 电子学报, 2025, 53(9): 3047-3059.
Mei Yaxin, Qin Huiling, Liang Yuzhu, et al. LLM-based zero-shot imputation of spatiotemporal data[J]. Acta Electronica Sinica, 2025, 53(9): 3047-3059. (in Chinese)

[32] Hastie T, Tibshirani R, Friedman J H. The elements of statistical learning: Data mining, inference, and prediction[M]. 2nd ed. New York: Springer, 2009.

[33] Moradi R, Berangi R, Minaei B. A survey of regularization strategies for deep models[J]. Artificial Intelligence Review, 2020, 53(6): 3947-3986.

[34] Han Zongbo, Zhang Changqing, Fu Huazhu, et al. Trusted multi-view classification with dynamic evidential fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2551-2566.

[35] Zhang Chiyuan, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107-115.

[36] Hernández-García A, König P. Data augmentation instead of explicit regularization[PP/OL]. V5. arXiv (2018-06-11)[2026-03-20]. https://arxiv.org/abs/1806.03852v5.

[37] Zheng Qinghe, Yang Mingqiang, Yang Jiajie, et al. Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process[J]. IEEE Access, 2018, 6: 15844-15869.

[38] Tian Yingjie, Zhang Yuqi. A comprehensive survey on regularization strategies in machine learning[J]. Information Fusion, 2022, 80: 146-166.

[39] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2999-3007.

[40] Tao Linwei, Dong Mingjing, Xu Chang. Dual focal loss for calibration[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023: 33833-33849.

[41] Yu Xiaotong, Sun Shiding, Tian Yingjie. Sample selection for noisy partial label learning with interactive contrastive learning[J]. Pattern Recognition, 2025, 166: 111681.

[42] 贺前华, 陈永强, 郑若伟, 等. 基于样本类不确定性抽样的端到端语音关键词检测训练方法[J]. 电子学报, 2024, 52(10): 3482-3492.
He Qianhua, Chen Yongqiang, Zheng Ruowei, et al. End-to-end speech keyword spotting training method based on sample's class uncertainty[J]. Acta Electronica Sinica, 2024, 52(10): 3482-3492. (in Chinese)

[43] 王海荣, 王彤, 徐玺, 等. CLGLF: 置信学习引导标签融合的多模态命名实体识别方法[J]. 电子学报, 2024, 52(7): 2429-2437.
Wang Hairong, Wang Tong, Xu Xi, et al. CLGLF: Confidence learning guided label fusion for multimodal named entity recognition method[J]. Acta Electronica Sinica, 2024, 52(7): 2429-2437. (in Chinese)

[44] Han Zongbo, Yang Yifeng, Zhang Changqing, et al. Selective learning: Towards robust calibration with dynamic regularization[PP/OL]. V2. arXiv (2024-02-13)[2026-03-20]. https://arxiv.org/abs/2402.08384.

[45] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[46] Goodman N R. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)[J]. The Annals of Mathematical Statistics, 1963, 34(1): 152-177.

[47] Bloch D A, Watson G S. A Bayesian study of the multinomial distribution[J]. The Annals of Mathematical Statistics, 1967, 38(5): 1423-1435.

[48] Huber P J. Robust estimation of a location parameter[M]// Breakthroughs in statistics: Methodology and distribution. New York: Springer, 1992: 492-518.

[49] Shannon C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[50] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[51] Kullback S, Leibler R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[52] Patrício M, Pereira J, Crisóstomo J, et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer[J]. BMC Cancer, 2018, 18(1): 29.

[53] Ma Yuzhe, Zhu Xiaojin, Hsu J. Data poisoning against differentially-private learners: Attacks and defenses[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019: 4732-4738.

[54] Uddin M N, Halder R K. An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach[J]. Informatics in Medicine Unlocked, 2021, 24: 100584.

[55] Neyman J, Pearson E S. On the problem of the most efficient tests of statistical hypotheses[J]. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character, 1933, 231(694/695/696/697/698/699/700/701/702/703/704/705/706): 289-337.

[56] Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29-36.

[57] Youden W J. Index for rating diagnostic tests[J]. Cancer, 1950, 3(1): 32-35.

[58] Lewis D D. A sequential algorithm for training text classifiers: Corrigendum and additional data[J]. ACM SIGIR Forum, 1995, 29(2): 13-19.

[59] Guo Chuan, Pleiss G, Sun Yu, et al. On calibration of modern neural networks[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017: 1321-1330.

[60] Brier G W. Verification of forecasts expressed in terms of probability[J]. Monthly Weather Review, 1950, 78(1): 1-3.

[61] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R[J]. Journal of Statistical Software, 2011, 45(3): 1-67.

[62] Yoon J, Jordon J, van der Schaar M. GAIN: Missing data imputation using generative adversarial nets[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018: 5675-5684.

[63] Mattei P A, Frellsen J. MIWAE: Deep generative modelling and imputation of incomplete data sets[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 4413-4423.

作者简介

韩宗博 男,1997年7月出生于山东省德州市。北京邮电大学信息与通信工程学院助理教授、硕士生导师。主要研究方向为多模态机器学习和智能医疗。中国电子学会会员编号:E190188742M。

E-mail: zongbo@bupt.edu.cn

西雨晗 女,2005年9月出生于河南省三门峡市。现为天津科技大学人工智能学院本科生。主要研究方向为机器学习。

E-mail: xiyuhan@mail.tust.edu.cn

盛培卓 女,2000年5月出生于天津市。天津大学人工智能学院硕士研究生。主要研究方向为机器学习与数据挖掘。

E-mail: peizhuosheng@tju.edu.cn

廖 芸 女,1991年5月出生于天津市。2024年博士毕业于天津大学,现任天津科技大学人工智能学院讲师。主要研究方向为在线学习、多模态医疗图像融合、大模型推理与自演化智能体。

E-mail: yliao@tust.edu.cn

张长青 男,1982年8月出生于河南省安阳市。天津大学人工智能学院教授、博士生导师。主要研究方向为机器学习、计算机视觉、智能医疗。中国电子学会会员编号:E190188589M。

E-mail: zhangchangqing@tju.edu.cn