机器学习在医药统计中的疾病预测研究

李文博
西交利物浦大学　江苏苏州　215123

摘　要：疾病风险预测以其在疾病防控和治疗等方面突出的临床意义一直是医疗卫生领域关注的重要研究课题，传统的基于病例队列研究的疾病预测方法耗时冗长、人力、物力投入巨大，难以满足服务临床应用的需求。随着医疗数据的迅速增长，以及机器学习技术的不断发展，基于统计学的疾病预测模型在医学研究和临床决策中发挥着越来越重要的作用。本文深入探讨机器学习在医药类应用统计学中的重要作用，重点探讨疾病预测模型的构建流程，常用算法和优化策略。本文通过回顾相关研究案例，总结当前疾病预测模型的挑战与未来发展趋势，供医学统计研究者与临床医生参考。

关键词：医药统计学；机器学习；疾病预测；模型构建

引言

近几年来医疗健康数据呈现出的那种爆炸式增长状况，实际上为疾病预测相关研究提供了前所未有的机遇；鉴于电子健康记录EHR、医学影像以及基因组学数据等多种类型信息被加以整合的缘故，基于统计学构建出的疾病预测模型便能够更精准地去识别高危患者进而对临床干预策略予以优化[1]。而在疾病预测领域里，机器学习特别是深度学习这种ML技术则展示出了极为强大的能力[2]。然而需要指出的是，机器学习模型的构建过程涵盖了诸如数据预处理、特征选择还有模型训练与验证等多个至关重要的环节，那么究竟该如何对这些步骤展开优化，进而实现提高模型泛化能力这个目标，这仍然属于当下研究过程当中的重点之处。

一、医药类应用统计学中基于机器学习应用背景

在医药领域，基于机器学习的应用统计学正日益成为推动精准医疗和临床决策革命的核心驱动力[3]。随着医疗信息化建设的快速推进，现代医疗机构每天产生海量异构数据，包括电子健康记录（EHR）中的结构化临床指标、医学影像的像素矩阵、基因组测序的变异位点、可穿戴设备采集的连续生理参数等，这些数据具有维度高、噪声多、时空关联复杂等特点，传统统计方法在处理此类数据时面临显著挑战[4]。机器学习技术凭借其强大的特征学习能力和非线性建模优势，为医药统计学提供了全新的分析范式：在疾病预测方面，通过集成多源异构数据，可构建超越传统风险评分的预测模型（如使用XGBoost实现心血管疾病10年风险预测AUC达0.92）；在药物研发中，深度学习可加速虚拟筛选过程（如用3D-CNN分析分子对接结果，效率提升50倍）；在医学影像分析领域，卷积神经网络（CNN）已实现肺结节检测的敏感度98.5%，达到放射科专家水平；在个性化治疗方面，强化学习算法可优化给药方案（如华法林剂量预测误差降低40%）。特别值得注意的是，这些应用都建立在严格的统计学基础之上[5]。当前研究前沿正聚焦于解决小样本学习（通过迁移学习利用预训练模型）、数据异质性（开发联邦学习框架保护隐私）和模型可解释性（应用SHAP值可视化决策依据）等关键问题，这些突破将进一步推动机器学习在医药统计中的深度应用，最终实现从群体医学到个体化医疗的范式转变[6]。

二、医药类应用统计学中基于机器学习的疾病预测模型构建的必要性

在医药领域当中疾病预测模型的构建，对于精准医疗、公共卫生管理以及临床决策而言都有着重要意义，然而传统的诸如逻辑回归、Cox回归这类统计方法，尽管它们在广泛使用，但是也存在局限性的，所以基于机器学习的疾病预测模型因其能够弥补这些不足而展现出必要性，主要体现于下面三方面：

1.在提升预测精度从而适应具有高维度（例如基因组学、影像学数据）、非线性关系（就像基因–环境交互作用之类）及异质性（比如患者个体差异这种情况）特征的复杂医学数据层面，传统统计模型依赖于线性假设，难以捕捉其中复杂模式，而像随机森林、支持向量机、

深度学习这些机器学习算法，却能够自动学习数据里的非线性关系以提高预测准确性，像在癌症早期筛查中，深度学习模型借助对医学影像（像CT、MRI之类）的分析，就能够识别细微病变进而显著降低漏诊率，并且集成学习方法可有效整合多源数据来优化预测性能，为个性化诊疗提供支持。

2.在实现动态风险评估与早期干预方面，由于疾病发展具备时序性，传统静态模型难以做到及时更新预测结果，而诸如LSTM、时间序列分析这类机器学习模型，则可将患者历史数据（像电子健康记录、穿戴设备监测数据等）进行动态整合来实现实时风险评估，以糖尿病管理为例，连续血糖监测数据与机器学习相结合便能够预测低血糖事件并提前发出预警，而且强化学习还能优化干预策略，这种动态预测能力有助于进行早期干预，达到降低重症发生率以及改善患者预后的目的。

3.针对推动精准医疗与公共卫生资源优化，机器学习模型通过细分患者群体识别出高风险个体，以此助力精准医疗，例如凭借聚类分析把癌症患者划分成不同亚型从而匹配靶向疗法，从公共卫生层面来说，模型能够对疾病流行趋势（例如流感、传染病传播等）进行预测，来指导资源分配，就如同疫情期间，基于SEIR模型的机器学习改进版可更准确预测感染峰值以辅助制定防控政策，此外自动化模型能够整合多中心数据，这样可减少人工分析成本进而提升医疗效率。

三、疾病预测模型的构建流程

（一）数据收集与预处理

医学数据作为疾病预测模型构建的基础，主要来源于四大类数据源：电子健康记录（EHR）、医学影像、基因组学数据和可穿戴设备数据。EHR系统包含结构化的诊断编码、实验室检查结果和非结构化的临床记录，其数据碎片化问题需要通过自然语言处理技术进行信息提取和标准化。医学影像数据（如MRI、CT）以DICOM格式为主，存在扫描参数差异导致的异质性，需进行图像配准和强度校正。基因组学数据包含数百万个SNP位点，具有超高维度特性，需进行质量控制（如去除测序深度<30X的位点）和群体分层校正。

数据清洗是确保数据质量的关键步骤，涉及三个主要环节：缺失值处理需根据缺失机制选择适当方法，当数据完全随机缺失（MCAR）时可采用简单删除，随机缺失（MAR）时推荐多重插补（MICE），其通过建立多个回归模型迭代估算缺失值。异常值检测需结合统计方法和领域知识，除IQR方法外，对时间序列数据可采用动态阈值算法，基于滑动窗口计算局部统计量。

数据标准化处理主要解决特征量纲不一致问题，Z-score标准化（$x'=(x-\mu)/\sigma$）适用于近似正态分布的实验室指标，但对异常值敏感；Min-Max归一化将特征缩放到[0，1]区间，适用于有明确上下界的临床指标（如血氧饱和度）。对于存在偏态分布的数据（如肿瘤标志物浓度），推荐先进行对数变换再进行标准化。特别值得注意的是，所有标准化参数（μ、σ、min、max）必须仅从训练集计算，以避免数据泄露导致模型评估偏差，这一原则在交叉验证时需格外注意。

数据增强技术主要应用于医学影像数据，传统几何变换包括随机旋转（$\pm 15°$）、镜像翻转和弹性形变，可增加数据多样性。针对深度学习的高级增强方法中，生成对抗网络（GAN）可合成逼真的病理图像，但需注意模式坍塌问题；风格迁移技术能实现跨模态增强（如CT到MRI）。最新的自监督学习方法（如SimCLR）通过对比学习构建不变性表示，可有效利用未标注数据。

（二）特征选择与降维

在医学数据分析领域之内，特征选择这一环节对于改进模型性能可以说是起着关键的作用，而过滤法作为一种相对而言最简单的特征选择办法，是凭借计算特征和目标变量的统计相关性来进行筛选的，其中常见会用到的指标包含Pearson相关系数、卡方检验以及互信息等等，要知道互信息具备能够捕捉任意形式统计依赖的特性，所以对于医学数据里经常出现的非线性关系的计算来讲特别适用，通常会利用k近邻算法（此时k取值为3）去估计概率分布。

包装法（Wrapper）通过构建特征子集并评估模型性能来进行选择，递归特征消除（RFE）是典型代表。RFE采用反向消除策略，首先训练包含所有特征的模型（如SVM或随机森林），然后迭代移除权重最小的特征，每次迭代都通过k折交叉验证（通常k=5或10）评估模型性能。这种方法虽然计算成本较高（时间复杂度$O(n^2)$），但能发现特征间的协同效应。

对于超高维医学数据（如10^5维的基因表达数据），降维技术至关重要。主成分分析（PCA）通过正交变换将相关特征转换为线性无关的主成分，通常保留解释95%方差的成分（约10-50个PCs）。t-SNE则擅长可视化高维数据（perplexity通常设为30），但其计算复杂度高（$O(n^2)$），大规模数据推荐使用UMAP（UniformManifold

ApproximationandProjection）。值得注意的是，这些降维方法可能丢失生物学意义，因此需要结合领域知识解释结果。

（三）模型选择与训练

在医学分类任务中，算法选择需综合考虑数据特性和临床需求。逻辑回归因其良好的可解释性（可通过优势比解释特征影响）和计算效率，仍是二分类任务（如良恶性肿瘤判别）的基础算法，其L2正则化版本（C通常设为1.0）能有效防止过拟合。支持向量机（SVM）通过核技巧（RBF核γ参数通常设为1/n_features）处理非线性可分数据，在医学影像分类中表现优异，但计算复杂度随样本量立方增长（O（n^3））。随机森林通过Bootstrap采样和特征随机选择（通常$\sqrt{p}$个特征）构建决策树森林，其内置的特征重要性评估特别适合基因表达数据分析。深度学习方法中，CNN（如ResNet-50）在医学影像分析领域达到专家级水平，而RNN（如BiLSTM）更适合处理电子病历中的时序文本数据。

生存分析算法需要同时考虑事件发生与否和发生时间。Cox比例风险模型通过偏似然函数估计风险比（HR），其前提假设需通过Schoenfeld残差检验（p>0.05）。生存随机森林通过构建生存树（节点分裂准则通常采用log-rank统计量）来估计累积风险函数，能自动处理非线性关系和交互作用。DeepSurv作为深度生存模型代表，采用负对数似然损失函数，通过多层感知机（通常3-5个隐藏层）学习非线性风险函数，在癌症预后预测中C-index可达0.85以上。最新研究将注意力机制引入生存分析，开发的Transformer-Surv模型在多项指标上超越传统方法5-7个百分点。

时间序列预测在疾病进展监测中尤为重要。LSTM通过门控机制（遗忘门、输入门、输出门）解决长期依赖问题，在ICU患者恶化预测中AUROC可达0.92。Transformer模型利用自注意力机制（头数通常设为8）捕捉全局依赖关系，特别适合处理多变量临床时间序列（如生命体征联合实验室指标）。这类模型通常采用滑动窗口策略（窗口大小24-72小时）构建训练样本，并应用教师强制（teacherforcing）技术加速训练过程。值得注意的是，医学时间序列往往存在不规则采样问题，需要引入时间感知的插值层或专门设计的时间编码策略。

模型验证环节是保证临床可靠性的关键环节。通过分层抽样保持类别分布的K折交叉验证（通常K=5或10），其误差估计的标准差一般在0.02以内。对于n<1000的小样本医学数据，推荐采用嵌套交叉验证，外层折5，内层折3，避免参数优化偏差。模型性能评估需结合临床需求：分类任务关注敏感度（避免漏诊）和特异度（避免误诊）；生存分析侧重C-index和校准曲线；时间序列预测则需计算动态AUROC和提前预警时间。

参考文献

[1]张子旋.基于数据挖掘的中医治疗难治性痛风的用药规律和作用机制研究[D].北京中医药大学，2023.

[2]蒋金金，周文君，丁世彬.Seminar法结合案例式教学法在“实用医药统计学”教学中的应用及效果评价[J].科学咨询，2024（10）：81-84.

[3]刘丽，刘海燕，严征，等.基于“混合式教（导）+学”模式“医学统计学”课程思政教学研究与实践[J].首都食品与医药，2024，31（16）：117-120.

[4]黄玉清，白婧.血清肿瘤标志物CA199，CA724，CEA，CA242在胃癌诊断中的应用价值分析[J].中国科技期刊数据库 医药，2023.

[5]Cao Y, Ju X, Chen X, et al.MCL-VD: Multi-modal contrastive learning with LoRA-enhanced GraphCodeBERT for effective vulnerability detection[J]. Automated Software Engineering, 2025, 32(2): 67-67.

[6]Gil M J. Augmenting the Interpretability of GraphCodeBERT for Code Similarity Tasks[J]. International Journal of Software Engineering and Knowledge Engineering, 2025, 35(05):