Statistics and Application 统计学与应用, **2025, 14(9), 216-227**
Published Online September 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.149270

Hans汉斯

选择性共形预测在乳腺癌高风险人群风险评估中的方法研究

王子雯，彭诗淇，曹雅琦*

中央民族大学理学院，北京

收稿日期：2025年8月16日；录用日期：2025年9月6日；发布日期：2025年9月18日

摘　要

在医学风险预测中，为个体化预测结果提供可靠的不确定性量化，是评估模型临床实用性的关键环节。本文聚焦于选择性共形预测方法在乳腺癌高风险人群风险评估中的应用，旨在通过构建覆盖率可控的预测区间，实现对风险模型预测性能的稳健评估，并提升其在高风险人群中的可解释性。我们首先系统回顾了分裂共形预测的基本原理，进一步引入选择性共形预测方法，在保证理论覆盖率的同时，实现高风险个体的有效识别与重点覆盖，并针对该人群给出最大置信度与预测可信度等单指标不确定性度量，为临床干预提供参考信息。结合基于临床指标的风险预测模型，本文利用威斯康星州乳腺癌数据开展实证分析。结果表明，所提出的方法在高风险个体的预测准确率与不确定性控制方面均优于分裂共形预测，具有良好的实际应用潜力。本研究为医学诊断场景下的不确定性建模提供了新的统计工具，并为高风险人群的精准防控提供了方法上的支持。

关键词

选择性共形预测，不确定性度量，覆盖率，高风险人群，乳腺癌风险预测

Selective Conformal Prediction for Risk Assessment in High-Risk Breast Cancer Subgroups

Ziwen Wang, Shiqi Peng, Yaqi Cao*

School of Science, Minzu University of China, Beijing

Received: Aug. 16th, 2025; accepted: Sep. 6th, 2025; published: Sep. 18th, 2025

*通讯作者。

文章引用: 王子雯, 彭诗淇, 曹雅琦. 选择性共形预测在乳腺癌高风险人群风险评估中的方法研究[J]. 统计学与应用, 2025, 14(9): 216-227. DOI: 10.12677/sa.2025.149270

Abstract

In biomedical research, reliable quantification of predictive uncertainty is a key criterion for evaluating the clinical utility of risk prediction models. This study focuses on the application of selective conformal prediction to risk assessment in high-risk breast cancer populations. The objective is to construct prediction sets with controllable coverage, enabling robust evaluation of model performance and enhancing interpretability for high-risk subgroups. We first review the split conformal prediction method and then introduce a selective conformal prediction framework that achieves both theoretical coverage guarantees and effective identification with prioritized coverage of high-risk individuals. For these individuals, we provide uncertainty measures such as maximum confidence and credibility scores to inform targeted clinical interventions. Using a risk prediction model incorporating clinical indicators, we conduct real data analysis based on the Wisconsin Breast Cancer dataset. Results demonstrate that the proposed method outperforms standard split conformal prediction in terms of predictive accuracy and uncertainty control for high-risk cases, showing strong potential for practical applications. This work offers a novel statistical tool for uncertainty modeling in medical diagnostics and provides methodological support for precision prevention and control in high-risk populations.

Keywords

Selective Conformal Prediction, Uncertainty Quantification, Coverage Probability, High-Risk Group, Breast Cancer Risk Prediction

1. 引言

随着人工智能与统计建模方法在医学、金融及社会科学等领域的迅速发展，复杂预测模型尤其是深度学习模型在多项任务中有良好表现。然而，在医学等高风险决策场景中，模型的“黑箱”特性带来了严重挑战。以临床诊断为例，错误预测可能导致严重后果，使得仅依赖单点预测结果远远不够。虽然深度学习等模型具备强大的拟合能力和预测性能，但其可解释性不足、不确定性难以量化等问题，制约了其在实际医疗决策中的广泛应用。因此，如何在保证预测表现的同时，为黑箱模型提供可靠的不确定性度量工具，成为当前统计建模与人工智能交叉研究的重要方向。一方面，不确定性度量有助于识别预测结果的可信程度，为医疗人员提供额外判断依据；另一方面，它也是实现风险控制、保障模型安全使用的关键机制。

共形预测(Conformal Prediction)作为一种形式化、模型不可知的框架，正为这一问题提供了新的解决思路。它是一种与所选模型无关的方法，为分类与回归任务提供了形式化的预测区间，并在有限样本设定下提供可控置信水平的统计保障。该方法通过对模型预测结果进行校准，构建满足覆盖率要求的预测集合，为评估个体级别的预测不确定性提供了一种通用而严谨的手段。因此，近年来共形预测在学术界受到高度关注，金璋等人[1]认为多数学者主要围绕其自身理论性质展开研究，或是致力于探索各种统计模型与共形预测的融合，并把该方法用于各学科的具体问题研究。与此同时，还有一部分学者在此基础上对共形预测方法进行理论拓展和创新。2021 年 Jung 等人就提出了基于共形预测框架的新型聚类方法

[2]；2022 年 Angelopoulos 等人[3]通过控制单调损失函数的期望值来对共形预测进行拓展，使其能够应用于更广泛的场景；2023 年 Barber 等人[4]使用加权残差分布来修改共形预测，使其在数据不可交换的情况下，仍能提供有效预测；同年，Hu 等人[5]尝试用共形预测来检验超出可交换性之外的统计假设，Zaffran 等人[6]研究了协变量缺失情况下的共形预测并证明了共形预测的边际覆盖保证对任何缺失值都成立；为使预测结果提供更可靠的置信区间 Yang 等人在 2024 年提出了效率优先共形预测和有效性优先共形预测[7]。然而，现有文献多聚焦于连续型或多分类响应变量，对于医学问题中常见的二值响应下如何构建预测集合及度量预测不确定性的问题研究仍相对较少。

选择性共形预测(Selective Conformal Prediction, SCOP)是共形预测的拓展，通过动态筛选高置信度样本进行预测，在保证统计可靠性的前提下提升效率。因其可优先输出高确定性结果、避免对模糊病例误判，该方法常用于医学风险预测中。在以往的研究中，Zhu 等人利用符合 p 值进行个体选择来对罕见病进行更好的预测[8]，Feng 等人将选择性共形预测方法用于预测病人死亡率和 ICU 住院时间[9]，Bao 等人将该方法用于识别与特定靶点具有高亲和力的药物[10]。

本文围绕二值响应变量的预测任务，提出一种基于共形预测的不确定性度量方法，并进一步发展出适用于高风险人群的选择性共形预测方法。具体而言，本文的主要贡献包括：首先，针对二值响应变量，构建基于共形预测的预测集合，并引入最大置信度(maximum confidence)与预测可信度(credibility)两个指标，对单点预测结果的不确定性进行量化，提升模型解释性与临床可用性；第二，本文使用选择性共形预测方法，使模型在高风险子集上的预测准确性显著提升；第三，将所提出方法应用于乳腺癌风险预测问题中，通过对威斯康星州乳腺癌数据的实证分析，验证方法在高风险人群医疗干预中的实用价值。

2. 理论基础

2.1. 分裂共形预测

共形预测是一种模型无关方法框架，能够在保持有限样本可覆盖性(finite-sample validity)的前提下，为分类任务构建具有置信保障的预测集合。设有输入变量 x_j 和二值响应变量 $y_j \in \{0,1\}$，共形预测的目标是针对给定的 x_j，构造一个预测集合 Γ_j^{ε}，使得在置信水平 $1-\varepsilon$ 下满足 $P\left(y_j \in \Gamma_j^{\varepsilon}\right) \geq 1-\varepsilon$。其也属于共形预测中的一种，近年来在医学风险控制领域展现出重要应用价值。这种方法能够为机器学习模型的预测生成具有理论保证的置信区间或预测集。核心思想是将数据分为训练集和校准集，利用校准集来量化预测的不确定性，确保覆盖率满足预设要求。因仅需一次模型训练，故该方法在运算速度上具有显著优势，适合处理大规模数据。

给定样本量为 $2n$ 的数据集 D，样本空间为 $X \times Y$，其中 Y 表示标签，$z_i \in D$，$z_i = (x_i, y_i)$。将数据集 D 分成两个互不相交的子集—校准集 D_c 和训练集 D_t。此时，校准集 $D_c = \{(x_i, y_i)\}_{i=1}^{n}$ 对应的样本数为 n，训练集 $D_t = \{(x_i, y_i)\}_{i=n+1}^{2n}$ 对应的样本数也为 n。在 D_t 上使用适当的机器学习的方法训练出模型 $\mu(\bullet)$，并将 D_c 代型 $\mu(\bullet)$ 中，计算其不一致分数 α_i：

$$\alpha_i = \Delta\left[\mu(x_i), y_i\right], i = 1,2,...,n,$$

其中，Δ 是 $\mu(\bullet)$ 针对 (x_i, y_i) 的预测误差。对新数据 x_j 其可按如下方式预测：确定显著性水平 $\varepsilon \in \{0,1\}$，接下来将每个可能的预测值 $\tilde{y}$ 代入 $\Delta\left[\mu(x_j), \tilde{y}\right]$ 得到不一致分数 $\alpha_j^{\tilde{y}}$ 并计算 $P_j^{\tilde{y}}$：

$$P_j^{\tilde{y}} = \frac{\left|\left\{i = 1,2,...,n : \alpha_i \geq \alpha_j^{\tilde{y}}\right\}\right|}{n+1}.$$

最后得到预测区域 $\Gamma_j^{\varepsilon} = \left\{\tilde{y} \in Y : P_j^{\tilde{y}} > \varepsilon\right\}$。若样本 (x_i, y_i) 独立同分布则该预测区域满足覆盖概率不等

式[11]：

$$P\left(y_j \in \Gamma_j^{\varepsilon}\right) \geq 1-\varepsilon,$$

且当残差 α_i 具有连续的联合分布时，覆盖概率的上界为：

$$P\left(y_j \in \Gamma_j^{\varepsilon}\right) \leq 1-\varepsilon+\frac{2}{n+2}.$$

由此可见，这一方法不仅计算高效，而且在内存需求上也具有优势，尤其是当模型 $\mu(\bullet)$ 涉及变量选择时，仅需存储选中的变量即可完成新数据点的预测和残差计算。此外，预测区域 Γ_j^{ε} 还提供了近似样本内覆盖保证，便于基于现有数据直观解释和验证其有效性。根据以上定义与计算，图 1 系统阐述了分裂共形预测的实施流程。

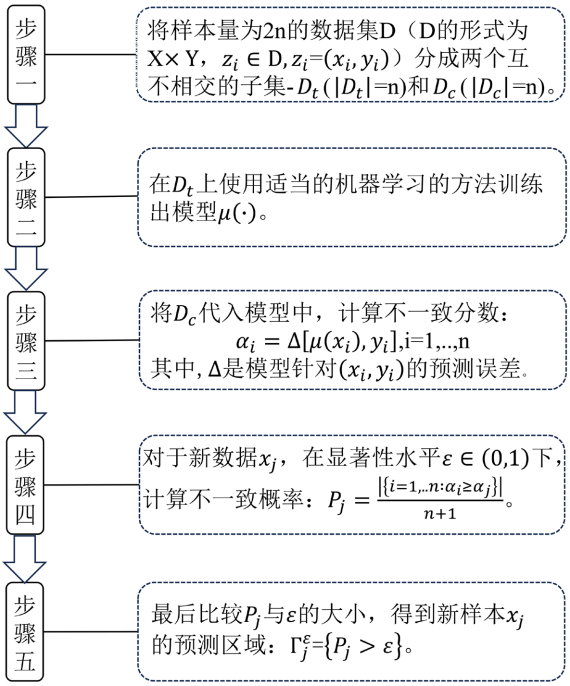

Figure 1. Flowchart of split conformal prediction
图 1. 分裂共形预测流程图

2.2. 选择性共形预测

选择性共形预测是一种扩展的共形预测，其在保证统计覆盖率的同时，允许模型在某些情况下选择性地放弃预测，以提高预测集的精确度或效率。其具有计算高效，灵活性、精确度较高的优势，在高风险决策等应用领域中展现出良好的适用性。

已有数据集 $D_l = \left\{\left(x_i, y_i\right) \in R^d \times R\right\}_{i=1}^{2n}$ 和测试集 $D_u = \left\{x_i\right\}_{i=2n+1}^{2n+m}$ 其中 $\left\{y_i\right\}_{i=2n+1}^{2n+m}$ 未知。将 D_l 分成大小相等的训练集 D_t 和校准集 D_c，并将 D_t 视为固定值同时在其上拟合预测模型 $\mu(\bullet)$。接下来指定一个得分函数 $g(\bullet)$，计算得分 $T_i = g\left(x_i\right)_{i \in c \cup u}$，并确定置信水平 ε 和阈值 $\hat{\tau}$，阈值 $\hat{\tau}$ 可以被归纳为三种类型：(1) 由用户指定或仅依赖于训练集 D_t 获得：(2) 仅取决于得分 $g\left(x_i\right)_{i \in u}$ 即选择测试集中得分最小的一定比例个体：(3) 完全或部分依赖于校准集 D_c。获得新校准集 $S_c = \left\{i \in c : T_i \leq \hat{\tau}\right\}$ 和新测试集 $S_u = \left\{i \in u : T_i \leq \hat{\tau}\right\}$。最后计算：

$$R_i = \left|Y_i - \mu\left(x_i\right), i \in S_c\right|,$$

$$PI_j^{SCOP} = \mu(x_j) \pm Q_\varepsilon\left(\{R_i\}_{i\in S_c}\right), j \in S_u,$$

其中 $Q_\varepsilon\left(\{R_i\}_{i\in S_c}\right)$ 表示 $\{R_i\}_{i\in c}$ 中从小到大排序后第 $\left[(1-\varepsilon)(|c|+1)\right]$ 个值。进而得到预测区间 $\left\{PI_j^{SCOP} : j \in S_u\right\}$。根据以上定义与计算，图 2 系统地阐述了选择性共形预测的实施流程。

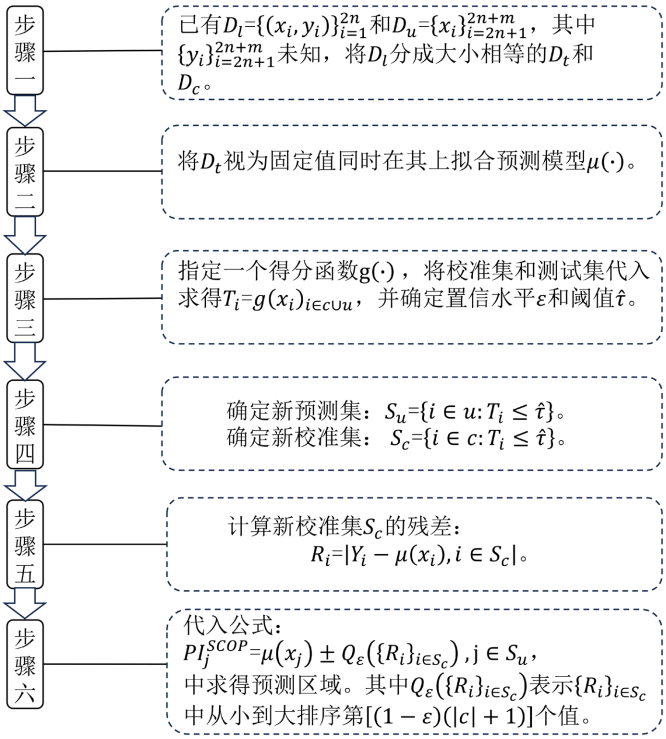

Figure 2. Flowchart of selective conformal prediction
图 2. 选择性共形预测流程图

若样本 (x_i, y_i) 独立同分布且 $\hat{\tau}$ 关于 $\{T_i\}_{i\in c\cup u}$ 可交换则其条件错误覆盖率满足：

$$P\left(Y_j \notin PI_j^{SCOP} \mid j \in S_u\right) \le \varepsilon.$$

当 $\{R_i\}_{i\in c\cup u}$ 几乎必然取不同值，且 $P(|S_u| > 0) = 1$ 时，条件错误覆盖率的下界为：

$$\varepsilon - E\left\{\left(|S_c|+1\right)^{-1}\right\} \le P\left(Y_j \notin PI_j^{SCOP} \mid j \in S_u\right).$$

如果不满足 $\{T_i\}_{i\in c\cup u}$ 可交换的条件，可以适当调整新校准集和新测试集，比如选定排序阈值 $k \in [m]$，将新测试集重写为 $S_u^+ = \left\{j \in u : T_j \le T_{(k)}\right\}$，其中 $T_{(k)}$ 表示 $\{T_i\}_{i\in u}$ 从小到大排序后第 k 个值并将新校准集重写为 $S_c^+ = \left\{i \in c : T_i \le T_{(k+1)}\right\}$。

2.3. 针对二值响应变量的风险预测模型

本研究采用核密度估计方法建立针对二值响应变量的风险预测模型。给定训练集 $D_t = \left\{(x_i, y_i)\right\}_{i=1}^n$ 其中 x_i 表示经过降维后的主成分特征(PC1 和 PC2)，$y_i \in \{0,1\}$ 为类别标签(0 表示良性，1 表示恶性)。对于每个类别c都有 n_c 作为该类别的样本量，则其核密度估计函数表达式为：

$$\hat{f}_c(x) = \frac{1}{n_c h_c^2} \sum_{i:y_i=c} K\left(\frac{x - x_i}{h_c}\right),$$

其中高斯核函数 $K(\bullet)$ 和带宽参数 h_c 分别表示为 $K(u)=\frac{1}{2\pi}\exp\left(-\frac{u^2}{2}\right)$ 和 $h_c=h_0\cdot\frac{1}{\sqrt{n_c}}$ 其中 h_0 为初始带宽。在预测阶段，类别c的条件概率为 $P(y=c|x)=\frac{\hat{f}_c(x)w_c}{\sum_k \hat{f}_k(x)w_k}$ 其中类别权重 $w_c=\frac{n_c}{n}$。对于新样本 x_{test} 的预测集为：

$$\Gamma^{\alpha}(x_{test})=\{y:P(y\mid x_{test})\geq 1-q_{\alpha}\},$$

其中 q_{α} 为校准残差 $s_i=1-P(y=y_i|x_i)$ 的 $1-\alpha$ 分位数。

2.4. 基于最大置信度和可信度的不确定性度量

在共形预测的框架下，应确保真值不取任意一个候选预测值的概率较低。给定数据集 $Z=X\times Y$ ($z_i=(x_i,y_i),i=1,...,n-1$)，其中特征用 X 表示，二分类标签用 Y 表示($y=0$ 或 $y=1$)和新样本 x_j 时，我们需要确定预测值 y_j 的可能取值范围。在此情境下，候选预测值共有 4 种，分别为∅、$\{0\}$、$\{1\}$、$\{0,1\}$。当预测集仅包含 1 个元素时，便将该元素确定为本次预测的结果 y_j。为了精准衡量此次预测的效果，需使用最大置信度和预测可信度来量化此次预测的不确定性。接下来，运用不一致性度量公式 $\alpha_i=A\left((z_1,...,z_{n-1})/(z_i),z_i\right)$，分别计算当情况 $y_j=0$ 和 $y_j=1$时的不一致性分数 $\alpha_j^{y_j=0}$ 和 $\alpha_j^{y_j=1}$。然后计算：

$$P_0=\frac{\#\left\{i=1,...,n-1\middle|\alpha_i\geq\alpha_j^{y_j=0}\right\}}{n-1},$$

$$P_1=\frac{\#\left\{i=1,...,n-1\middle|\alpha_i\geq\alpha_j^{y_j=1}\right\}}{n-1}.$$

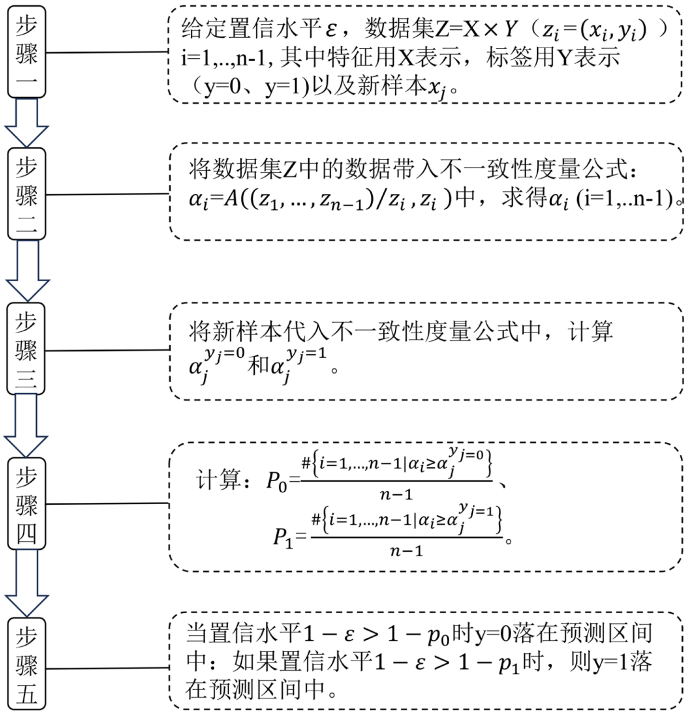

Figure 3. Conformal prediction process for binary outcome
图 3. 针对二值响应变量的共形预测过程

当置信水平$1-\varepsilon>1-p_0$时，$y_j=0$落在预测区间中。如果置信水平$1-\varepsilon>1-p_1$，则$y_j=1$落在预测区间中。具体计算步骤如图 3 所示。

假设经过图 3 中计算得到$P_0=0.89$、$P_1=0.04$。则说明当$1-\varepsilon>1-0.89$时，$y_j=0$处于预测区域，当$1-\varepsilon>1-0.04$时，$y_j=1$处于预测区域。在$\varepsilon=0.89$的情况下，若预测集呈现为空集，这表明把置信度下调至 11%时，预测集合将会是空集；在$\varepsilon=0.04$的情况下，若预测集为$\{0\}$，则意味着有 96%的信心预测$y_j=0$会落在预测区间中；在$\varepsilon=0.03$的情况下，若预测集为$\{0,1\}$，则说明如果将置信度升至 97%，那么预测集中将会既包含 0 也包含 1。

定义预测的可信度为使得预测集为空集的最小ε，最大置信度为使得预测集为单指标的最大$1-\varepsilon$。因此，在上述$P_0=0.89$、$P_1=0.04$的情况下，预测$y_j=0$的最大置信度为 96%而可信度为 89%。在医学常见的二分类问题中，如若预测的可信度较低，则说明没有足够的信息支持这个预测需结合其它手段进行综合诊断。而当最大置信度和可信度均较高时则说明该预测很有可能是准确的，可以使用该结果进行后续诊断治疗。因此，在二分类问题中引入最大置信度和可信度，本质是通过量化不确定性来实现更精细的风险控制，提升预测的可靠性。

3. 实证分析

本文采用康斯星州乳腺癌数据集[12]进行研究，该数据集汇聚了 699 例样本信息，详尽涵盖了 9 个标准化处理的病理特征：肿块厚度、细胞大小均匀性、细胞形态均匀性、边缘粘连、单上皮细胞大小、裸核、乏味染色体、正常核和有丝分裂，同时附有相应的诊断结果(分为良性和恶性)。因其数据质量高、特征解释性强常被用于机器学习和医学统计领域。早在 2010 年 Basu 等人就开始使用加权似然估计法对该数据集的诊断结果进行预测，旨在提升恶性病例的识别准确率[13]。随后，2017 年 Barrett 等人利用该数据集验证了选择性招募设计在提高统计功效、得出更可靠结论方面的有效性[14]，这项研究充分展示了在数据集中寻找合适子集进行预测的重要性。因此，本次研究选取高风险人群作为预测对象，分别使用分裂共形预测和选择性共形预测对其进行预测，通过对比二者的预测成效，选择更适合高风险人群的预测方法，从而确保获得更为稳健、可靠的预测结果。在之前的研究中，我们研究了分裂共形预测[15]发现其在高风险人群的预测中表现一般，因此进一步研究选择性共形预测是否在高风险人群中表现良好。

数据集中的 699 例样本有 458 例为良性，其余 241 例为恶性。针对数据中 16 个缺失样本，采用删除策略，最终保留 683 个完整数据进行模型训练。本研究对 683 个有效样本构建 KDE 模型并进行 PCA 降维，以便可视化及提取关键特征。按照诊断结果将数据分为良性组和恶性组。从良性样本中随机抽取 70 例，从恶性样本中随机抽取 30 例将两者合并构成 100 例测试集D_u。将剩余样本按 6:4 的比例划分为训练集D_t和校准集D_c。按照诊断结果$y=0$时为良性，$y=1$时为恶性，选取肿块厚度(x_1)，边缘粘连(x_2)，裸核(x_3)和乏味染体(x_4)作为风险预测因子，使用 logistic 回归来构建风险预测模型：

$$\operatorname{logit}(p)=-10.1+0.8x_1+0.4x_2+0.5x_3+0.7x_4,$$

转换该函数得到概率$p=\dfrac{1}{1+e^{-\operatorname{logit}(p)}}$，将$p\geq 0.5$个体的集合称为高风险人群，并构建新测试集$S_u=\{i\in u:p_i\geq 0.5\}$和新校准集$S_c=\{i\in c:p_i\geq 0.5\}$。此时，$S_u$对应的样本数为$n_u$，$S_c$对应的样本数为$n_c$，$D_t$对应的样本数为$n_t$。

3.1. 分裂共形预测在高风险人群乳腺癌患病预测中的应用

设类别$k\in\{0,1\}$，其中 0 为良性，1 为恶性。对训练集D_t选定机器学习训练模型，生成 KDE 核密度

估计模型作为预测器。因此，类别 k 的核密度估计函数为 $\hat{f}_k(x)=\frac{1}{n_k h^2}\sum_{i=1}^{n_k}K\left(\frac{|x-x_i|}{h}\right)$，其中带宽 h 设定为 0.35、高斯核函数 $K(u)$ 表示为 $\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$。计算校准集的不一致性分数 $\alpha_i=\frac{1}{\hat{f}_k(x_i)}$，$i\in n_c$。在 $\alpha=0.1$ (90%置信水平)下，$q_{1-\alpha}$ 表示校准集不一致性分数的$(1-\alpha)$分位数。对于 S_u 中的任意样本 x_j 构建的预测集为：

$$\Gamma(x_j)=\left\{y\in\{0,1\}:\hat{f}_y(x_j)\geq\frac{1}{q_{1-\alpha}}\right\}.$$

之后，进行实际覆盖率的计算 $\text{Coverage}=\frac{1}{n_u}\sum_{j=1}^{n_u}\mathrm{I}\left(y_j\in\Gamma(x_j)\right)$。其中 y_j 为样本 x_j 的真实类别，$\mathrm{I}(\cdot)$ 为指示函数即括号中的条件为真则为 1，否则为 0。

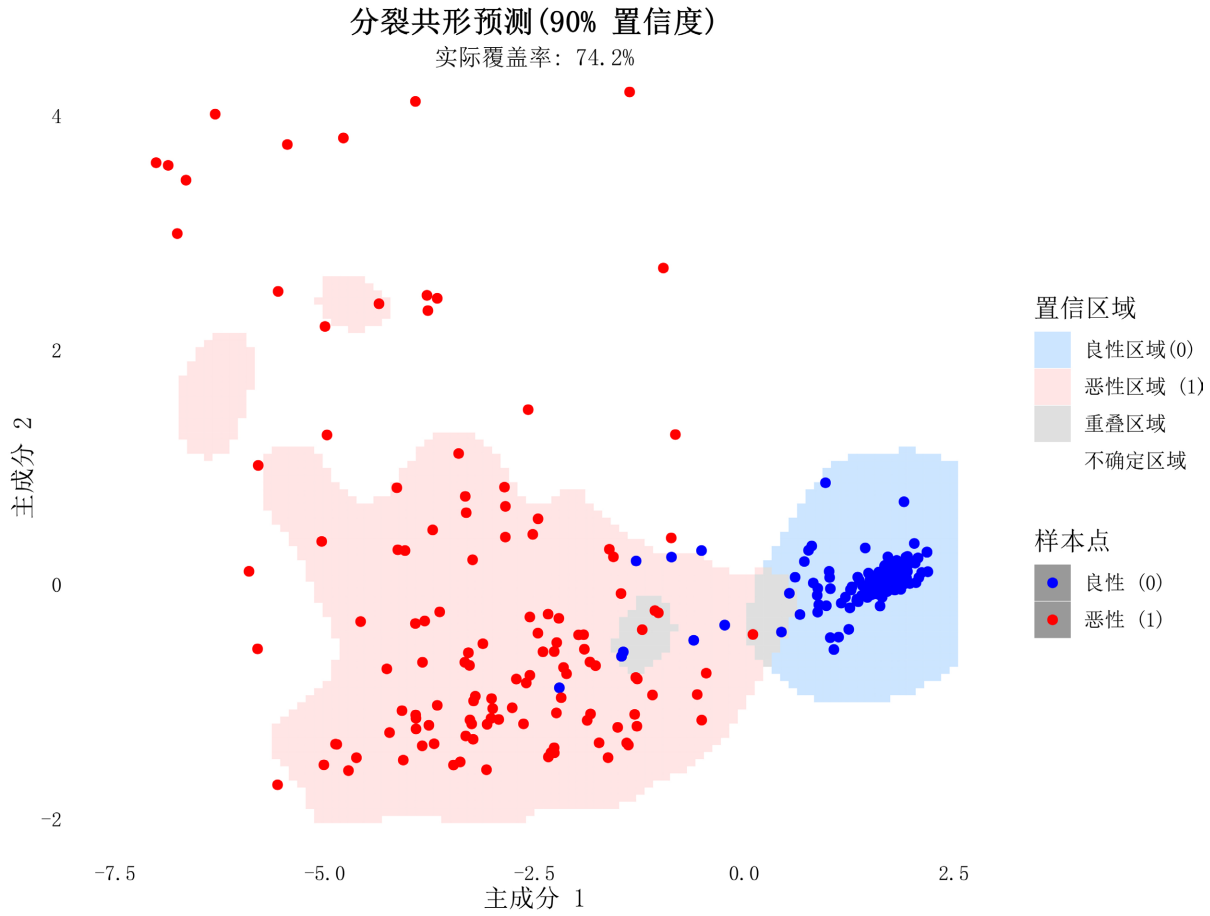

Figure 4. Classification and results of split conformal prediction (90% confidence) in high-risk populations
图 4. 分裂共形预测(90%置信度)在高风险人群中的分类及结果

散点图图 4 呈现了高风险人群中的良性样本(以蓝点标识)和恶性样本(以红点标识)在主成分分析空间中的分布情况。图中蓝色区域代表高风险人群在 90%置信水平下预测为良性的置信区域：红色区域代表高风险人群在 90%置信水平下预测为恶性的置信区域：重叠区域表示无法明确分类的区域：白色区域为不确定区域表示无法做出可靠性预测的区域。经过计算可得分裂共形预测在高风险人群中预测的实际覆

盖率为 74.2%。

3.2. 选择性共形预测在高风险人群乳腺癌患病预测中的应用

对于已有的类别标签集 $k \in \{0,1\}$，其中 0 为良性，1 为恶性，类别 k 的核密度估计函数表达式为 $\hat{f}_k(x) = \frac{1}{n_k h^2} \sum_{i=1}^{n_k} K\left(\frac{|x - x_i|}{h}\right)$，其中带宽 h 设定为 0.35、高斯核函数 $K(u)$ 表示为 $\frac{1}{\sqrt{2\pi}} e^{-u^2/2}$。计算校准集残差 $r_i = 1 - \hat{p}_{y_i}(x_i), x_i \in S_c$，其中类别概率 $\hat{p}_k(x) = \frac{\hat{f}_k(x) w_k}{\sum_{l \in \{0,1\}} \hat{f}_l(x) w_l}$，类别权重 $w_k = \frac{n_k}{n}$。对于新样本 x_j 构建的预测集 $\Gamma(x_j) = \left\{k \in \{0,1\} : \hat{p}_k(x_j) \geq \frac{1}{r_{1-\alpha}}\right\}$。其中 $r_{1-\alpha}$ 表示在 $\alpha = 0.1$ (90%置信水平)下的校准集残差 r_i 的 $1-\alpha$ 分位数。经计算得选择性共形预测在高风险人群中预测的实际覆盖率为 90.3%。

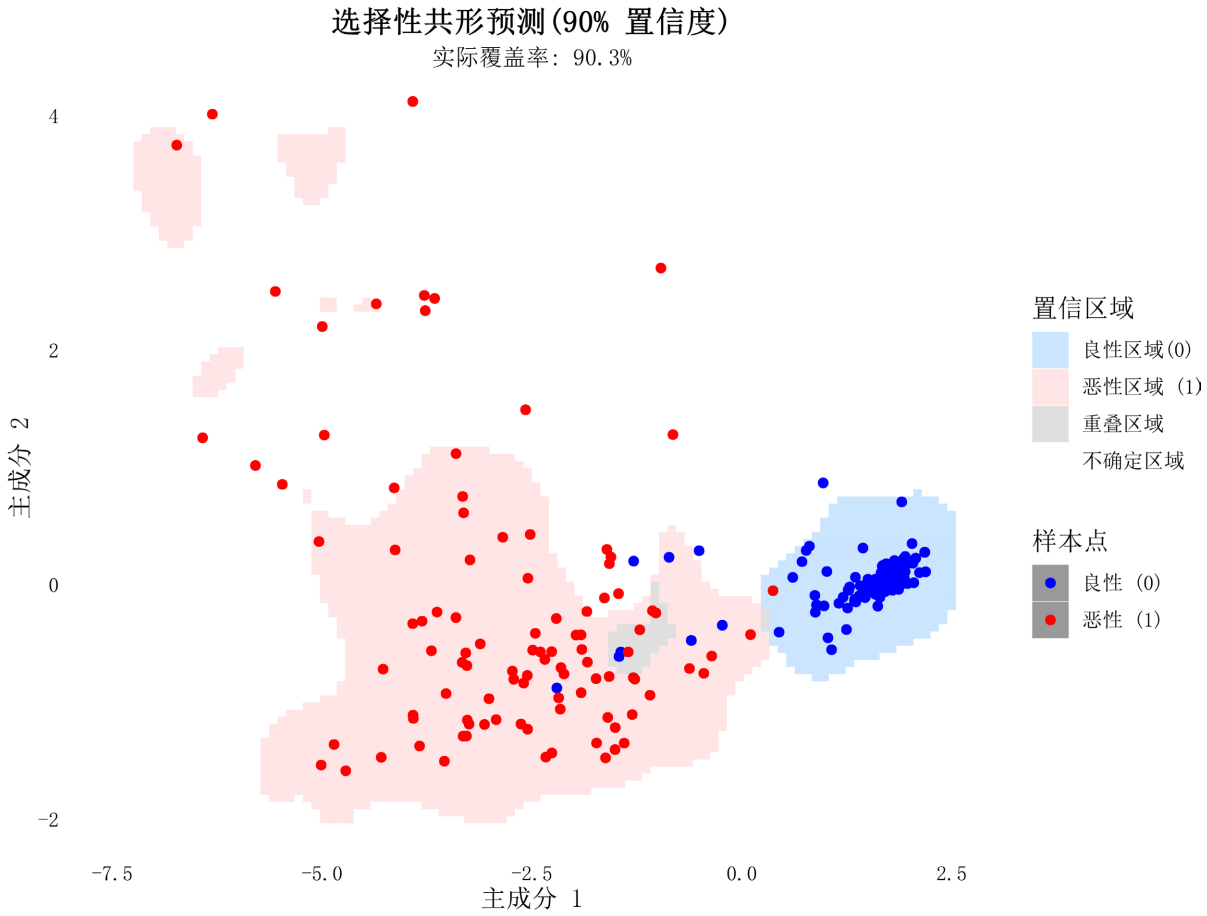

Figure 5. Classification and results of selective conformal prediction (90% confidence) in high-risk group
图 5. 选择性共形预测(90%置信度)在高风险人群中的分类及结果

针对已有结果，我们可以发现在高风险人群中，选择性共形预测的实际覆盖率(图 5)相比分裂共形预测的实际覆盖率(图 4)提升 16.1 个百分点。这在乳腺癌早期诊断中尤为关键，可降低漏诊风险。进一步与完全共形预测方法对比显示，选择性共形预测在高风险人群中的覆盖率仍高出 19.3 个百分点。在高风险人群上的实际覆盖率提升 19.3 个百分点。完全共形预测方法的计算效率较低，在本数据集中的运行结果

相较于分裂共形预测没有显著提升。本研究也将类别条件共形预测作为基线方法进行了对比，试验结果表明，其在恶性样本上的覆盖率相较于分裂共形预测有所提升，但该方法没有聚焦高风险样本，划分依据为诊断标签，在真实数据分析中，我们往往需要对高风险人群进行进一步检查和医疗干预，因此本文采用选择性共形预测方法。下面，我们将对分裂共形预测与选择性共形预测在高风险人群中的预测结果进行对比，举例说明预测结果不同的样本点，从而更直观地观测两者之间的区别。

表 1 中的结果显示，在两种预测方法于高风险人群中预测结果存在差异的样本点中，对于大部分恶性样本，分裂共形预测的输出结果为空集，说明该方法在该情境下失效，这表明分裂共形预测对高风险人群中恶性样本识别能力有限，存在漏诊的潜在风险。而选择性共形预测能够成功识别出分裂共形预测无法识别的恶性样本，进而提升了预测结果的准确性，为高风险人群的疾病预测提供更可靠的依据。因此，在疾病的风险预测中选择性共形预测能够更好地适应高风险人群数据的复杂性和不确定性，减少误判和漏判的情况，从而为后续的决策提供更为可靠的依据。在疾病早期筛查方面，它可以帮助医生更准确地识别出高风险患者，提高筛查的效率和准确性，使患者能够尽早接受进一步的诊断和治疗，从而提升治疗效果和患者的生存率。

Table 1. Differences in prediction results between split conformal prediction and selective conformal prediction in high-risk populations
表 1. 分裂共形预测与选择性共形预测于高风险人群中预测结果存在差异的样本点举例

序号	PC1	PC2	真实标签	分裂共形预测结果	选择性共形预测结果
1	−3.7706	2.3425	1	空集	1
2	−1.0639	−0.2172	1	0 和 1	空集
3	−1.0229	−0.2362	1	0 和 1	空集
4	−3.7815	2.4727	1	空集	1
5	−5.5563	2.5064	1	空集	1
6	−4.5273	4.4832	1	空集	1
7	−3.6589	2.4470	1	空集	1
8	−4.9971	2.2071	1	空集	1

3.3. 选择性共形预测的乳腺癌风险预测不确定性度量

对于测试集 S_u 中的任意样本 x_j 其真实标签未知，则该样本属于真实标签“1”的预测概率为 $\hat{p}_1 = \dfrac{\hat{f}_1(x_j)w_1}{\hat{f}_1(x_j)w_1 + \hat{f}_0(x_j)w_0}$。其中核密度估计函数 $\hat{f}_k(x) = \dfrac{1}{n_k h^2}\sum_{i=1}^{n_k} K\left(\dfrac{|x - x_i|}{h}\right)$，带宽 h 取值为 0.35、类别权重 w_k 的具体表达式为 $w_k = \dfrac{n_k}{n}$。针对该预测概率定义不一致性度量 $\sigma_j^{y=i} = |i - \hat{p}_1|$ (I = 1 或 0)，同时比较其与校准集残差 $r_i = 1 - \hat{p}_{y_i}(x_i), x_i \in S_c$ 的大小并计算：

$$p_k = \frac{\sum_{i=1}^{n_c} I\left(r_i \geq \sigma_j^{y=k}\right)}{n_c}, k \in \{0,1\},$$

其中 $I(\bullet)$ 为指示函数。通过比较置信度 $1-\alpha$ $(\alpha = 0.1)$ 与 $1-p_0$、$1-p_1$ 的大小生成预测集并得到预测可信度 $credibility = \max\{p_0, p_1\}$。下面我们以威斯康星州乳腺癌数据集为例，从预先定义的高风险人群中抽取 10 个样本，来对比运用分裂共形预测和选择性共形预测方法，计算所得的 90%置信度下的预测集、预

测单指标集的具体信息以及预测可信度。

由表 2 可看出，选择性共形预测和分裂共形预测给出了单指标预测集的最大置信度和预测可信度。其中，样本 1 在两种预测方法中均出现预测错误，但可以看出其在两种预测中的可信度较低分别为 44%和 32%，则说明对于该患者没有足够的信息来进行预测需结合其他手段进行诊断。而样本 6 在两种预测方法中得到的预测是准确的，而且单指标预测集的最大置信度均为 96%，预测可信度分别为 70%和 64%，因此我们可以将该患者判定为高风险患者并采取相应的医疗干预。除此之外，在多数情况下，选择性共形预测的预测可信度要高于分裂共形预测的预测可信度，这说明选择性共形预测在对高风险人群进行预测时具备更高的准确性。对于分裂共形预测中预测错误的样本，选择性共形预测可给出较为灵活的预测 $\{0,1\}$ (样本 1 和样本 8)，甚至是正确的预测(样本 3)，这说明与分裂共形预测相比，选择性共形预测的选择性机制可以动态调整预测灵活性与可信度。

Table 2. Application of selective conformal prediction (SCOP) and split conformal prediction (SCP) in Wisconsin breast cancer dataset
表 2. 选择性共形预测(SCOP)与分裂共形预测(SCP)在康斯星州乳腺癌肿瘤预测数据中的应用

序号	PC1	PC2	SCOP-90%置信度下预测集	SC0P 单指标预测集(最大置信度)	SCOP 预测可信度	SCP-90%置信度下预测集	SCP 单指标预测集(最大置信度)	SCP 预测可信度	真值
1	−0.0472	−0.1782	0 和 1	1 (81%)	44%	1	1 (96%)	32%	0
2	0.8521	0.1389	0	0 (96%)	85%	0	0 (96%)	56%	0
3	−3.7706	2.3425	1	1 (99%)	96%	空集	0 (96%)	20%	1
4	−1.5077	−1.2140	1	1 (96%)	74%	1	1 (96%)	60%	1
5	1.1206	−0.1588	0	0 (96%)	89%	0	0 (96%)	92%	0
6	−2.5978	−0.8346	1	1 (96%)	70%	1	1 (96%)	64%	1
7	−1.9052	−0.5472	1	1 (93%)	67%	1	1 (96%)	56%	1
8	−1.2880	0.2060	0 和 1	0 (81%)	41%	1	1 (96%)	40%	0
9	−1.0928	−0.9376	1	1 (96%)	67%	1	1 (96%)	56%	1
10	−2.1814	−0.9608	1	1 (96%)	74%	1	1 (96%)	56%	1

4. 结论与展望

4.1. 研究总结

本文提出了选择性共形预测方法并研究了选择性共形预测在乳腺癌高风险人群风险评估中的应用。通过与分裂共形预测相比，选择性共形预测在高风险人群中预测的实际覆盖率从 74.2%显著提升至 90.3%，进一步比对两种方法预测结果存在差异的样本点，发现选择性共形预测能够更好地识别恶性样本点，提升高风险人群的预测可靠性。同时，通过对比两种预测方法的最大置信度和预测可信度，发现对于分裂共形预测预测错误的边界病例，选择性共形预测可给出更为灵活、准确的结果。综上所述，与分裂共形预测相比，选择性共形预测在高风险人群的预测中具有更高的准确性，可用于癌症早期筛查确保高风险患者获得更及时的治疗。

4.2. 局限性与未来工作

尽管本研究取得了有意义的成果，但仍存在一些局限性。本研究基于威斯康星州乳腺癌单一数据集

进行方法验证，虽然该数据集质量较高且解释性强，但未来研究仍需要在更多样化的医疗数据集上验证该方法的一般性和稳健性。另外，本研究主要基于 logistic 回归模型给出的风险评分进行样本选择，未来工作可以探索更复杂的选择性策略，如基于多模态数据的整合选择机制，或者开发自适应阈值选择方法，以进一步提升预测性能。

共形预测的其他方法也可以应用于针对高风险人群的医疗干预和预测不确定性度量中，比如类别条件共形预测方法通过对于不同诊断类别构造预测集合，能够针对每个类别有一定的覆盖率保证，后续研究可以将分类扩展至基于传统模型的不同风险等级。未来研究需要进一步优化算法计算复杂度，开发更高效的实现方案，同时加强与临床应用的结合。当前研究主要针对二分类问题，未来可以扩展至多分类或生存分析等更复杂的医疗预测场景，探究选择性共形预测在这些领域的应用潜力。通过这些后续研究的深入开展，选择性共形预测方法将在医疗风险预测领域发挥更大的价值。

基金项目

国家自然科学基金青年项目(No. 12301369)，中央民族大学理学院 URTP 项目(URTP2025110352)。

参考文献

[1] 金璋, 王学钦. 共型预测方法的发展和应用[J]. 中国科学: 数学, 2024, 54(12): 2121-2140.

[2] Jung, S., Park, K. and Kim, B. (2021) Clustering on the Torus by Conformal Prediction. *The Annals of Applied Statistics*, **15**, 1583-1600. https://doi.org/10.1214/21-aoas1459

[3] Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L. and Schuster, T. (2022) Conformal Risk Control. arXiv, 2208.02814.

[4] Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2023) Conformal Prediction Beyond Exchangeability. *The Annals of Statistics*, **51**, 816-845. https://doi.org/10.1214/23-aos2276

[5] Hu, X. and Lei, J. (2024) A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum. *Journal of the American Statistical Association*, **119**, 1136-1154. https://doi.org/10.1080/01621459.2023.2177165

[6] Zaffran, M., Dieuleveut, A., Josse, J. and Romano, Y. (2023) Conformal Prediction with Missing Values. arXiv, 2305.19034.

[7] Yang, Y. and Kuchibhotla, A.K. (2024) Selection and Aggregation of Conformal Prediction Sets. *Journal of the American Statistical Association*, **120**, 435-447. https://doi.org/10.1080/01621459.2024.2344700

[8] Zhu, K., Yang, S. and Wang, X. (2024) Enhancing Statistical Validity and Power in Hybrid Controlled Trials: A Randomization Inference Approach with Conformal Selective Borrowing. arXiv, 2410.11713.

[9] Feng, J., Sondhi, A., Perry, J. and Simon, N. (2021) Selective Prediction-Set Models with Coverage Rate Guarantees. *Computing Research Repository*, **79**, 811-825. https://doi.org/10.1111/biom.13612

[10] Bao, Y., Huo, Y., Ren, H. and Zou, C. (2024) Selective Conformal Inference with False Coverage-Statement Rate Control. *Biometrika*, **111**, 727-742. https://doi.org/10.1093/biomet/asae010

[11] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J. and Wasserman, L. (2018) Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, **113**, 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

[12] Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1995) Breast Cancer Wisconsin (Diagnostic) Dataset. University of Wisconsin-Madison.

[13] Basu, A., Bose, S. and Purkayastha, S. (2004) Robust Discriminant Analysis Using Weighted Likelihood Estimators. *Journal of Statistical Computation and Simulation*, **74**, 445-460. https://doi.org/10.1080/00949650310001609458

[14] Barrett, J.E. (2017) Information-Adaptive Clinical Trials with Selective Recruitment and Binary Outcomes. *Statistics in Medicine*, **36**, 2803-2813. https://doi.org/10.1002/sim.7353

[15] 夏露竹, 曾文君, 曹雅琦. 共形预测框架下的预测区间估计方法及乳腺癌应用研究[J]. 应用统计与数据科学, 2025, 1(3): 90-95.