Modeling and Simulation 建模与仿真, 2023, 12(6), 5600-5607
Published Online November 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.126508

Hans汉斯

不同分类模型的心脏病预测效果分析

杜莹莹

南京信息工程大学，数学与统计学院，江苏 南京

收稿日期：2023年10月11日；录用日期：2023年11月18日；发布日期：2023年11月24日

摘 要

医学统计数据具有类型多样、数据量大等特征，一般的参数回归模型对医学统计数据的预测效果有时不能达到相应的要求。本论文提出了一种基于NW估计的非参数改进logistic回归模型，降低了logistic回归模型的链接函数假设条件；利用所提出的模型对心脏病诊断数据进行了模型拟合，并将其与一般的logistic回归模型的预测ROC曲线进行比较，发现对于样本数据的拟合，非参数改进后的logistic回归模型的预测效果优于一般的logistic回归模型。除此而外，本论文还对比了非参数机器学习方法——支持向量机与上述两种模型的预测效果之间的差异，绘制不同核函数下的支持向量机预测ROC曲线，对比之下，发现polynomial核函数下的支持向量机对患者是否患有心脏病这一问题的预测效果最好。

关键词

Logistic回归，Nadaraya-Watson估计，支持向量机，心脏病

Analysis of Prediction Effect of Different Classification Models on Heart Disease

Yingying Du

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Oct. 11th, 2023; accepted: Nov. 18th, 2023; published: Nov. 24th, 2023

Abstract

Medical statistical data is characterized by various types and large amounts of data, so the prediction effect of general parametric regression model cannot meet the corresponding requirements sometimes. This paper presents a non-parametric improved logistic regression model based on Nadaraya-Watson estimation, which reduces the link function hypothesis of logistic regression model. The proposed model was used to fit the heart disease diagnosis data, and the predictive ROC curve of the general logistic regression model was compared with that of the general logistic

文章引用: 杜莹莹. 不同分类模型的心脏病预测效果分析[J]. 建模与仿真, 2023, 12(6): 5600-5607.
DOI: 10.12677/mos.2023.126508

regression model. It was found that the predictive effect of the non-parametric improved logistic regression model was better than that of the general Logistic regression model. In addition, this paper also compares the difference between the prediction effect of non-parametric machine learning method support vector machine and the above two models, and draws the ROC curve predicted by support vector machine under different kernel functions. By comparison, it is found that support vector machine under kernel function has the best prediction effect on whether patients suffer from heart disease.

Keywords

Logistic Regression, Nadaraya-Watson Estimates, Support Vector Machine, Heart Disease

1. 引言

统计学在医学研究领域一直以来有着广泛的应用，医学统计数据的处理和分析对于医学研究、临床实践和医疗保健的改善至关重要。通过分析医学统计数据，研究人员可以了解疾病的流行趋势、严重程度、治疗方法、预后评估等方面的问题，从而提出新的研究问题、改进治疗方法和提供更好的医疗保健服务。医学统计数据的特点包括数据类型多样、数据量大、缺失值和异常值等问题。因此，对数据分布有着较强假设条件的参数模型对复杂的医学数据拟合效果可能不尽人意。非参数回归模型是一种不考虑数据分布的回归分析方法。与参数回归模型不同，参数回归模型假设数据分布是一个特定类型的分布，例如正态分布或泊松分布。而非参数回归模型不假设数据分布类型，因此它可以处理非正态数据和非二项分布的数据。

非参数回归在处理医学数据方面有着众多研究成果。覃雪纯[1]在非参数 logistic 模型的框架下，研究了基于拟合优度抽样的统计推断问题，利用局部多项式逼近的方法估计未知模型参数并给出了未知模型参数的置信区间的构造方法，并通过数值模型，验证了所提出的方法明显优于简单随机抽样；王晨阳[2]通过一系列算法优化与模型改进，构建出分类准确率较高、泛化能力较强的心脏病诊断模型，通过模型自演进模拟实验可以证明，随着业界医学知识扩充与心脏检查设备的更新，可以借鉴其结论对数据集进行维度扩充，通过重新构建及训练模型来完成心脏病分类模型的自演进功能，提升模型对病情的分类及判断能力；缪琦[3]使用支持向量机对糖尿病数据进行训练得到分类模型，结合随机森林给出的特征重要度对其加以改进，得到适应性更好的预测模型，结果表明该方法可以有效地识别并消除糖尿病数据中冗余的特征。

本论文提出了一种利用核函数优化的 logistic 回归模型，并将其用于预测心脏病数据，进一步对比了其与参数的 logistic 回归模型以及非参数的机器学习方法对 UCI 心脏病数据的拟合预测效果情况。

2. 理论总结

2.1. Nadaraya-Watson 估计

Nadaraya-Watson 回归(下称 NW 估计)是核回归方法的一种，其基本思想是使用样本的局部平均对回归目标进行估计。设 X 为样本自变量，Y 的为样本因变量，n 为样本量，$X_i, i=1,2,\cdots,n$ 为每次实验中样

本自变量的观测值，$Y_i, i=1,2,\cdots,n$ 为每次实验中样本因变量的观测值，则 Nadaraya-Watson 回归模型的形式为：

$$y=f(x)=\frac{\sum_{i=1}^{n}Y_iK_h(X_i-x)}{\sum_{i=1}^{n}K_h(X_i-x)}, \tag{1}$$

其中，$K(\bullet)$ 为对称的概率密度函数，$K_h(X_i-x)=\frac{1}{h}K\left(\frac{X_i-x}{h}\right)$，在 Nadaraya-Watson 估计方法中，常用的核函数为径向基核函数和均匀分布核函数。

2.2. Logistic 回归模型

logistic 回归模型是一种经典的二元分类模型，由条件概率分布 $P(Y=1|X)$ 表示，其模型形式为：

$$P(Y=1|X)=g(\beta^{\mathrm{T}}X+b), \tag{2}$$

其中，X 为样本自变量，$Y\in\{0,1\}$ 为一个二元分类变量，$g(\bullet)=\frac{e^x}{1+e^x}$。logistic 回归模型是一种特殊的广义线性模型，其链接函数为 logit 函数，表示事件“在已知 X 的条件下 $Y=1$”的对数几率。在 logistic 回归模型中，$Y=1$ 的对数几率是 X 的线性函数。

2.3. NW 估计下的 logistic 回归模型

由 2.2 节可知，logistic 回归模型是具有指定链接函数 $g(\bullet)$ 的广义线性模型，为参数模型，由于 logistic 回归模型是一种经典的二元分类模型，其应用范围广，涉及很多不同领域的数据，因此本文对 logistic 回归提出一种非参数改进思路，从而减弱模型使用的假定条件，使其适用范围变得更加广泛。

观察 NW 估计模型的形式可以看出，将 $\sum_{i=1}^{n}K_h(X_i-x)$ 视为自变量 X 的核密度估计，$y=f(x)$ 的形式即为 $P(Y=1|X)$ 的样本估计量，即条件期望 $E(Y|X)$ 的估计量。在此基础上，本文提出一种 logistic 回归模型的非参数改进，主要针对 logistic 回归模型的链接函数 $g(\bullet)$ 的形式进行非参数改进。下面对这种改进模型的形式进行定义，即链接函数 $g(\bullet)$ 的形式的非参数改进：

$$P(Y=1|X)=f(\beta^{\mathrm{T}}X+b)=\frac{\sum_{i=1}^{n}Y_iK_h(\beta^{\mathrm{T}}X_i-\beta^{\mathrm{T}}X)}{\sum_{i=1}^{n}K_h(\beta^{\mathrm{T}}X_i-\beta^{\mathrm{T}}X)}. \tag{3}$$

上述改进方法虽然为 logistic 回归模型的非参数改进，但是由于模型具体是对 logistic 回归模型的链接函数形式进行了核函数估计，在实际应用中仍然需要对参数 β 进行估计，其具体方法与经典的 logistic 回归一致，同样使用极大似然估计方法对参数进行估计。

2.4. 非线性支持向量机

在机器学习方面，非线性支持向量机是一种非常有效的应用非参数核技巧解决二元分类问题的学习方法。其基本思想是求解能够正确划分训练数据集并且几何间隔最大的分离超平面，其基本模型是定义在特征空间上的间隔最大的分类器。假定给定一个特征空间上训练数据集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 $x_i\in R^n, y_i\in\{0,1\}$，$x_i$ 为第 i 个特征向量，y_i 为 x_i 的类标记。从分类训练集，通过核函数与间隔最大化，学习得到分类决策函数

$$f(x)=sign\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x,x_i)+b^*\right), \tag{4}$$

将此过程称为非线性支持向量机，其中，$K(x,z)$为正定核函数，(α_i^*,b^*)为系数向量。

3. 实例分析

本论文所选择的数据来自 UCI Machine Learning Repository 网站上的 Heart Disease Dataset，数据集包括 296 个患者的 13 个预测变量观测值，分别为患者年龄(age)、性别(sex)、胸痛类型(cp)、患者入院时的静息血压(trestbps)、血清胆固醇水平(chol)、空腹血糖(fbs)、静息心电图结果(restecg)、达到的最大心率(thalach)、运动引起的心绞痛(exang)、运动相对于休息引起的 ST 压低(oldpeak)、最高运动 ST 段的斜率(slope)、萤光显色的主要血管数目(ca)、地中海贫血的患病程度(thal)，以及一个响应变量患者是否患有心脏病，记为 target，target 变量为 0~1 分布变量。

3.1. 数据描述以及相关性分析

在初步了解所研究的心脏病数据集的各个变量类型后，我们进一步地对数据进行描述性统计分析。首先，初步观测数据的分类情况，在样本数据中，女性患者人数为 95，男性患者人数为 201；患有心脏病的人数为 137，未患有心脏病的人数为 159。样本数据中患有心脏病和未患有心脏病的患者数量相差不大，未患有心脏病的患者数量稍多于患有心脏病的患者数量，大多数患者为男性。接着我们通过绘制患者年龄直方图来观察所抽选样本的年龄分布特征，图像如图 1 所示。

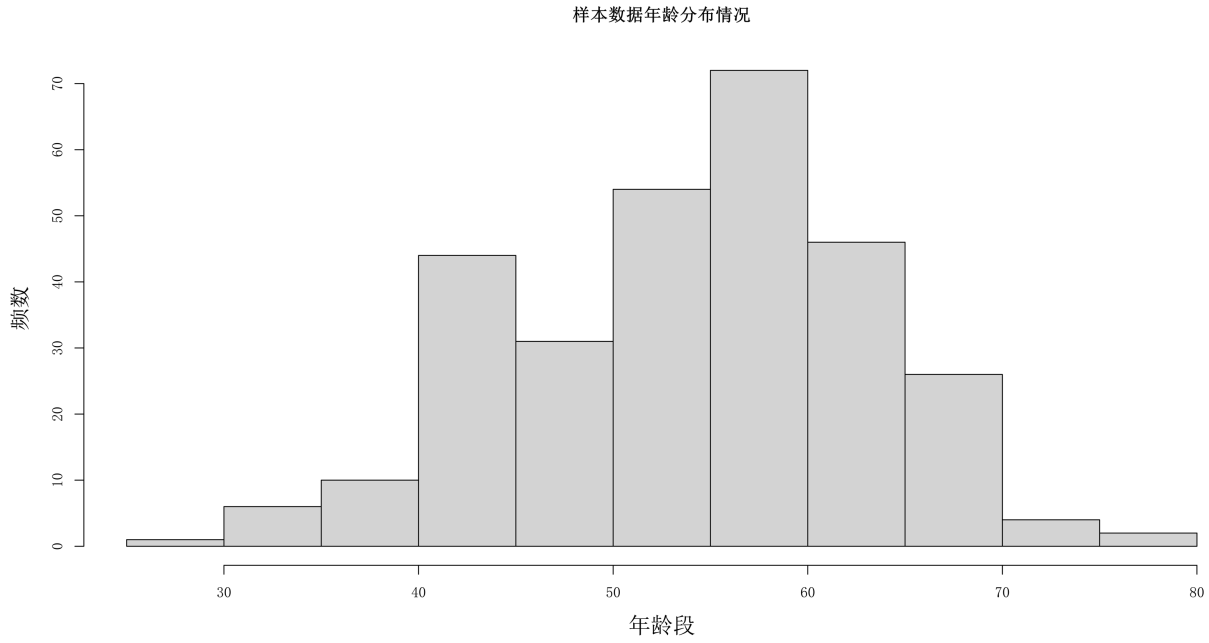

Figure 1. Patient age distribution chart
图 1. 患者年龄分布图

由图 1 可知，本论文所研究的患者样本中年龄集中在 50 至 65 岁，这也同样是普遍认为的心脏病高发的阶段。

在初步分析了我们所研究的样本数据中患者的基本情况后，进一步地，我们对所研究的所有预测变量进行相关性分析，为此，我们绘制了所研究的预测变量的相关性热图，具体图像如图 2 所示。

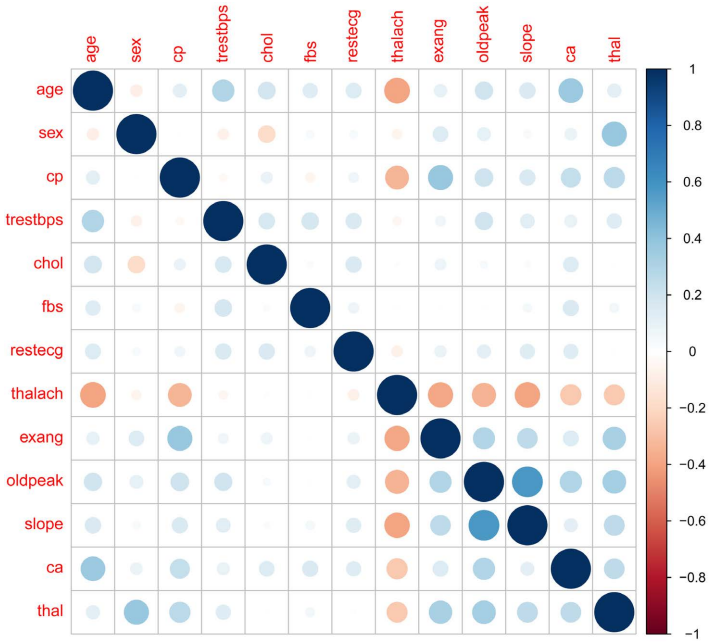

Figure 2. Heat map of the correlation between the studied predictor variables
图 2. 所研究的预测变量之间的相关关系热图

由图 2 可以看出，预测变量“达到的最大心率(thalach)”与多数病理性变量之家存在较为明显的负相关关系；变量“运动相对于休息引起的 ST 压低(oldpeak)”与多数病理性变量之间存在较为明显的正相关关系；变量“运动相对于休息引起的 ST 压低(oldpeak)”与变量“最高运动 ST 段的斜率(slope)”之间的正相关关系最强；变量“达到的最大心率(thalach)”与年龄之间的负相关关系最强。

在对数据进行了相关关系分析后，我们进一步地检查变量之间可能存在的共线性问题，避免共线性所导致的模型估计失真问题，通过计算所得共线性检验的条件数为 8.62 < 100，说明变量数据之间存在的共线性较小。

在对样本数据进行初步的描述分析、相关关系分析并检查变量之间的共线性后，我们将使用一般的 logistic 回归模型、NW 估计下的 logistic 回归模型以及三种不同核函数下的非线性支持向量机模型对样本数据进行进一步地拟合。

3.2. 使用一般的 logistic 回归模型拟合心脏病数据

首先，我们使用一般的逻辑回归模型对样本数据进行模型拟合，具体使用 R 语言 stats 包中的 glm 函数对数据进行参数回归拟合，拟合所得模型参数估计如表 1 所示。

Table 1. Logistic regression model fitting parameters
表 1. Logistic 回归模型拟合参数

变量	Intercept	age	sex	cp	trestbps	chol	fbs
系数估计	−6.95	−0.01	1.31	0.57	0.02	0.01	−1.01
变量	restecg	thalach	exang	oldpeak	slope	ca	thal
系数估计	0.25	−0.02	0.89	0.27	0.60	1.24	0.76

设所得参数估计为向量 $\hat{\beta}_{logistic}^{T}$，则所得 logistic 回归的具体模型形式为

$$P(Y=1|X)=\frac{e^{\hat{\beta}_{logistic}^{T}X}}{1+e^{\hat{\beta}_{logistic}^{T}X}}. \tag{5}$$

在获得拟合模型的基础上，我们进一步考察模型的拟合预测效果，首先计算模型对样本数据的预测精度，所得计算结果为 0.996，说明对于 logistic 回归模型对患者是否患有心脏病的预测精度很好；接着，绘制模型预测的 ROC 曲线，具体图像如图 3 所示。

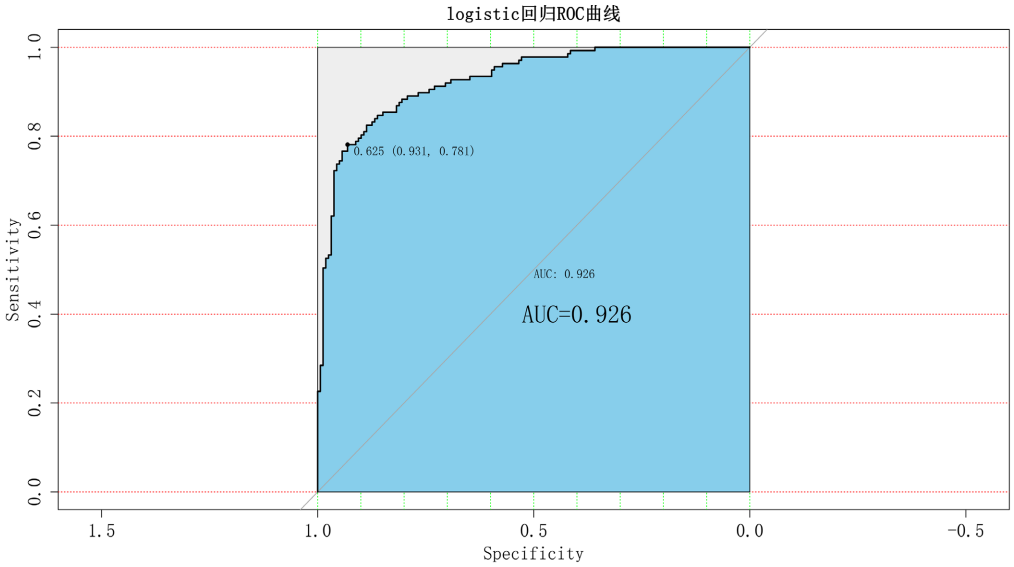

Figure 3. ROC curve of the predictive effect of the logistic regression model
图 3. Logistic 回归模型预测的 ROC 曲线

由图 3 可知，logistic 回归模型的 AUC 值为 0.926，说明该分类模型有 92.6%的概率能够正确地区分一个随机选取的正例比一个随机选取的负例的可能性，表明该模型具有很好的区分能力。

3.3. 使用 NW 估计下的 logistic 回归模型拟合心脏病数据

接着，我们使用非参数改进的 logistic 回归模型对样本数据进行拟合，所选择的核函数为高斯核函数，其表达为

$$K(x)=\frac{1}{h\sqrt{2\pi}}e^{-0.5\left(\frac{x_i-x}{h}\right)^2}, \tag{6}$$

其中窗宽 h 取值为 0.45。具体使用 R 语言中的 optim 函数对模型的极大似然函数进行优化求解，所得拟合模型的参数估计如表 2 所示。

Table 2. Logistic regression model fitting parameters under NW estimation
表 2. NW 估计下的 logistic 回归模型拟合参数

变量	Intercept	age	sex	cp	trestbps	chol	fbs
系数估计	−6.96	−0.01	1.31	0.57	0.02	0.01	−1.01
变量	restecg	thalach	exang	oldpeak	slope	ca	thal
系数估计	0.25	−0.02	0.89	0.62	0.69	1.24	1.01

设所得参数估计为向量 $\hat{\beta}^{\mathrm{T}}_{NW\text{-}logistic}$，则所得改进 logistic 回归的具体模型形式为

$$P(Y=1|X)=\frac{\sum_{i=1}^{n}Y_iK_h\left(\hat{\beta}^{\mathrm{T}}_{NW\text{-}logistic}X_i-\hat{\beta}^{\mathrm{T}}_{NW\text{-}logistic}X\right)}{\sum_{i=1}^{n}K_h\left(\hat{\beta}^{\mathrm{T}}_{NW\text{-}logistic}X_i-\hat{\beta}^{\mathrm{T}}_{NW\text{-}logistic}X\right)}. \tag{7}$$

在获得拟合模型的基础上，我们进一步考察模型的拟合预测效果，首先计算模型对样本数据的预测精度，所得计算结果为 0.997，说明 NW 估计下的 logistic 回归模型对患者是否患有心脏病的预测精度很好；接着，绘制模型预测的 ROC 曲线，具体图像如图 4 所示。

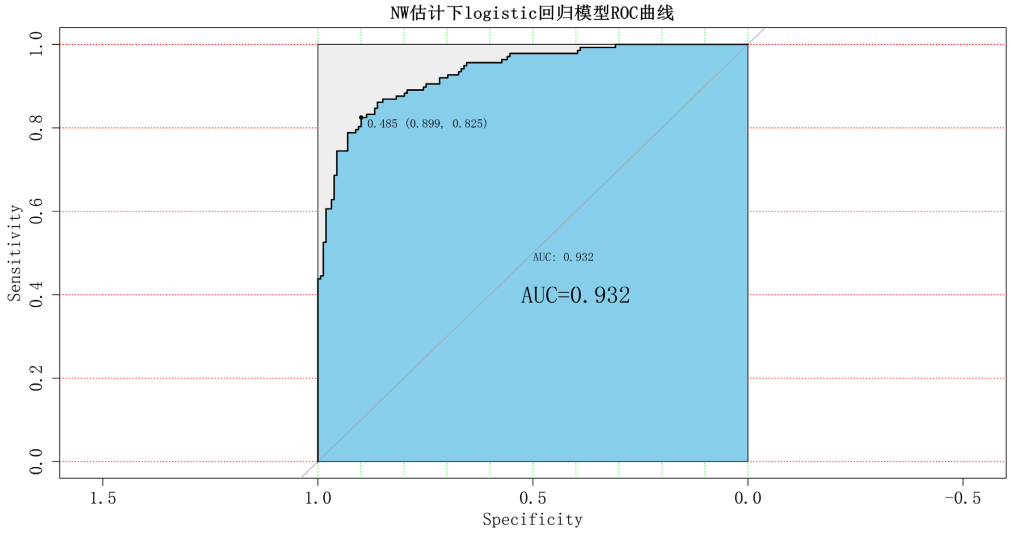

Figure 4. ROC curve of the predictive effect of the logistic regression model under NW estimation
图 4. NW 估计下的 logistic 回归模型预测的 ROC 曲线

由图 4 可知，NW 估计下的 logistic 回归模型的 AUC 值为 0.932 > 0.926，说明该分类模型有 93.2% 的概率能够正确地区分一个随机选取的正例比一个随机选取的负例的可能性，其模型的预测效果将优于一般的 logistic 回归模型。

3.4. 使用支持向量机模型拟合心脏病数据

最后，我们使用非参数的机器学习方法应用核技巧的支持向量机，我们分别使用三种不同核函数对样本数据进行拟合，对比不同核函数下算法的预测效果。具体通过 R 语言 e1071 包中的 svm 函数对样本数据进行拟合，分别拟合在 polynomial 核函数、sigmoid 核函数、radial 核函数下的不同模型，计算三种模型的预测精度，计算所得 polynomial 核函数下的支持向量机预测精度为 0.997；sigmoid 核函数下的支持向量机预测精度为 0.996；radial 核函数下的支持向量机预测精度为 0.997。可见所使用的 5 种模型对患者是否患有心脏病的预测效果相差不大，进一步绘制不同核函数下的支持向量机预测 ROC 曲线进行对比，具体图像如图 5 所示。

由图 5 可以看出，polynomial 核函数下的支持向量机预测 ROC 曲线的 AUC 值为 0.966；sigmoid 核函数下的支持向量机预测 ROC 曲线的 AUC 值为 0.716；radial 核函数下的支持向量机预测 ROC 曲线的 AUC 值为 0.962，说明三种支持向量机分类模型分别有 96.6%、71.6%以及 96.2%的概率能够正确地区分一个随机选取的正例比一个随机选取的负例的可能性；polynomial 核函数下的支持向量机预测效果最好，而 sigmoid 核函数下的支持向量机预测效果较差。

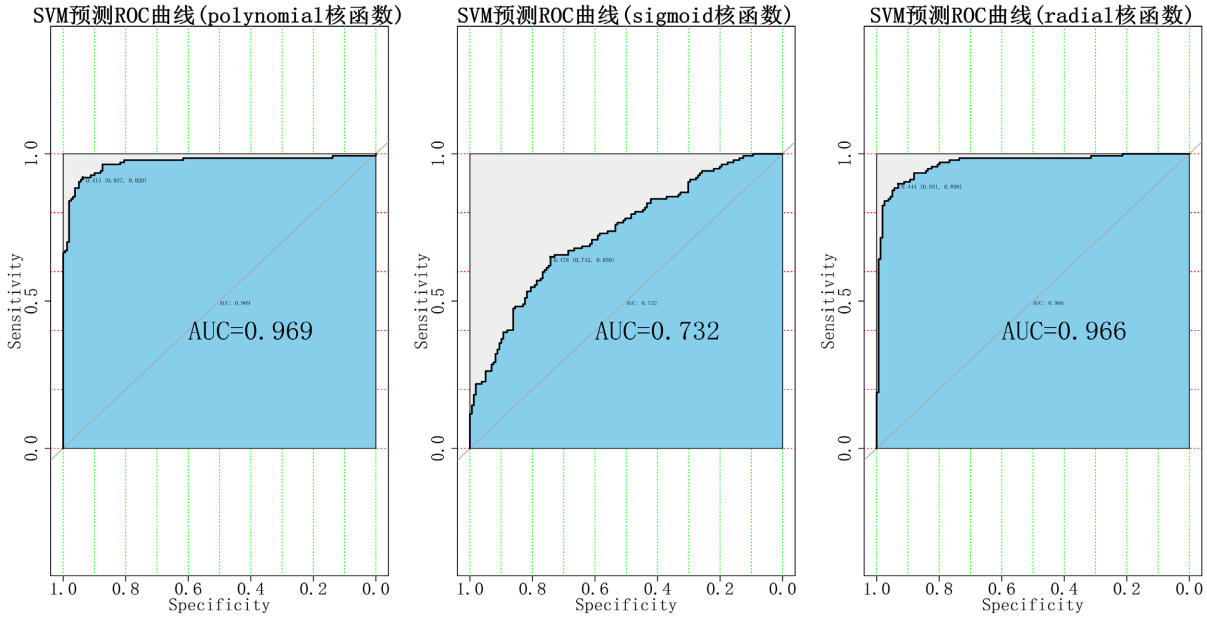

Figure 5. ROC curve of prediction effect of support vector machine under three different kernel functions
图 5. 三种不同核函数下支持向量机预测 ROC 曲线

将三种不同核函数下的支持向量机预测模型与 3.2 节及 3.3 节中拟合的一般 logistic 回归模型以及 NW 估计下的 logistic 回归模型的预测效果(AUC 值)对比可见，预测效果最好的是 polynomial 核函数下的支持向量机，两种 logistic 回归模型的预测效果均略差于 polynomial 核函数、radial 核函数下的支持向量机，但明显优于 sigmoid 核函数下的支持向量机。

4. 总结

本论文主要提出了一种 NW 估计下的 logistic 回归模型，拟合预测了 UCI 的心脏病数据集中“患者是否患有心脏病”这一问题，将所提出的非参数改进模型与一般的 logistic 回归模型以及非参数机器学习方法——支持向量机分别拟合数据，所得结果显示，所使用的 5 种模型对样本数据的预测精度效果都很好，相差并不大；而 5 种模型预测 ROC 曲线的 AUC 值则具有较为明显的差异：polynomial 核函数下的支持向量机对患者是否患有心脏病的预测 ROC 曲线的 AUC 值最大，预测效果最好，sigmoid 核函数下的支持向量机对患者是否患有心脏病的预测 ROC 曲线的 AUC 值最小，预测效果最差；NW 估计下的 logistic 回归模型的预测 ROC 曲线的 AUC 大于一般的 logistic 回归模型，说明本论文所提出的改进模型优化了模型的预测效果；NW 估计下的 logistic 回归模型是对传统 logistic 回归模型的一种非参数优化，减弱了传统 logistic 回归模型的模型假设，使改进后的模型能够适用于更广泛的应用数据，在后续研究中，可以进一步讨论非参数改进的 logistic 回归模型与传统 logistic 回归模型相应的适用范围，从而提高模型利用效率。

基金项目

国家自然科学基金面上项目：超高维复杂数据统计降维研究(11771215)。

参考文献

[1] 覃雪纯. 基于拟合优度抽样下非参数 logistic 模型的统计推断[D]: [硕士学位论文]. 武汉: 华中师范大学, 2022.
[2] 王晨阳. 基于数据驱动的心脏病分类诊断系统设计[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2021.
[3] 缪琦. 基于随机森林和支持向量机的糖尿病风险预测方法研究[D]: [硕士学位论文]. 镇江: 江苏大学, 2019.