**Advances in Applied Mathematics** 应用数学进展, 2022, 11(10), 7000-7011
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110743

Hans汉斯

# 基于得分函数的概率分类模型研究

**李佳洁**

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2022年9月12日；录用日期：2022年10月2日；发布日期：2022年10月12日

## 摘　要

**现代统计学中有各种分类方法，在数据研究中，类别分得越精准，得到的结果就越有价值。对于二元分类问题，本文提出了一种基于得分函数的概率分类模型MKL，从理论上证明了所提出的MKL估计的一致性。在实证方面，本文通过拟牛顿算法直接对连续化后的MKL统计量进行优化，给出了模拟研究的分类效果和一个心脏衰竭数据集的实例。该方法考虑了预测能力、计算复杂度和实际可解释性方面的权衡，与现有的分类方法相比具有优势。**

## 关键词

**分类，得分函数，准确率，机器学习**

# Research on Probability Classification Model Based on Scoring Function

**Jiajie Li**

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

Received: Sep. 12th, 2022; accepted: Oct. 2nd, 2022; published: Oct. 12th, 2022

## Abstract

**There are various classification methods in modern statistics, and in data research, the more accurate the classification, the more valuable the results obtained. For the binary classification problem, this paper proposes a probabilistic classification model MKL based on the score function, which theoretically proves the consistency of the proposed MKL estimate. In terms of empirical evidence, this paper directly optimizes the continuous MKL statistics by means of quasi-Newtonian algorithm, and gives the classification effect of the simulation study and an example of a heart failure dataset. This approach takes into account trade-offs in terms of predictive power, computational complexity, and practical interpretability, and offers advantages over existing classification methods.**

文章引用: 李佳洁. 基于得分函数的概率分类模型研究[J]. 应用数学进展, 2022, 11(10): 7000-7011.
DOI: 10.12677/aam.2022.1110743

**Keywords**

**Classification, Scoring Function, Accuracy, Machine Learning**



## 1. 引言

分类问题在社会科学、经济管理以及医学领域中有着非常广泛的应用，例如预测借贷平台用户是否违约、预测客户流失率、诊断不同的脊椎病变等等。在过去的几十年里，大量的分类方法在实际应用中得到了发展。J.S. Cramer [1]系统地回顾了逻辑回归的起源，逻辑回归通过对样本属于某一类的概率进行预测来实现分类。Breiman 等人[2]首次提出了随机森林方法，通过将多棵决策树集成，以及每次用采样的样本和特征分量训练每棵决策树，可以有效地降低模型的方差。Corinna Cortes 和 Vapnik [3]首次提出了支持向量机方法，该方法可以处理标记错误的样本，并证明了支持向量机利用多项式输入变换的高泛化能力。Yann LeCun [4]首次将 BP 算法应用到神经网络结构的训练上，形成了当代卷积神经网络的雏形，只需最少的预处理对高维模式进行分类。

分类作为典型的监督学习方法，目前已发展出许多成熟且优良的分类模型，但随着信息技术的发展与各种新技术的发明，传统的分类模型暴露出了种种不足，于是怎样结合现有的理论知识去拓展分类方法以期更高效地处理数据变得尤为重要。Zhang 和 Li 等人[5]通过训练数据给不同的测试数据点分配不同的 $k$ 值，学习一个相关矩阵来重构测试数据点，对 $k$ 近邻方法进行了改进，该方法在分类、回归、缺失数据归因等数据挖掘应用中，比现有的 kNN 方法更准确、高效。Fang 和 Chen [6]提出了一种直接最大化 Kolmogorov-Smirnov 统计量的信用评分方法对银行客户进行分类，该方法重点展示了预测能力在 KS 统计量上的表现，得到了一个优于传统评分模型的信用评分方法。

本文在现有的理论基础和前人思维的启发下，针对二分类问题，提出了一种新的分类模型，即最大化形如 KL 散度的统计量，记为 MKL，基于得分函数对问题进行分类。本文的其余部分组织如下：第 2 节详细介绍了本文的方法、理论性质和渐进结果，以及最优化目标函数的算法步骤；第 3 节进行了模拟研究，并与传统的分类模型如 Logistic 回归、支持向量机、随机森林和神经网络进行了比较；第 4 节将本文提出的分类方法应用到了真实数据上；第 5 节给出了一些结论。所有的证明都在附录中。

## 2. 方法

### 2.1. 符号和模型

设响应变量为 $Y$，其取值为 0 和 1，$X$ 表示样本特征或协变量。设得分函数 $S(X)$ 是 $X$ 的标量函数，并且 $S(X)$ 与条件概率 $P(Y=1|X)$ 呈正相关。基于得分函数，令 $t$ 表示截止分数即得分函数的一个阈值，也就是说，样本得分不超过 $t$ 的将被分到一类，$P(S(X)\le t|Y=0)$ 为属于 $Y=0$ 的样本被正确分类的百分比，另一方面，$P(S(X)\le t|Y=1)$ 为属于 $Y=1$ 的样本被错误分类的百分比。因此根据正判和误判的概率，我们希望 $P(S(X)\le t|Y=0)\ge P(S(X)\le t|Y=1)>0$，$P(S(X)\le t|Y=0)$ 越大越好，$P(S(X)\le t|Y=1)$ 越小越好，故而本文提出了如下概率分类模型，记为 $MKL$。定义：

$$MKL=\sup_{-\infty<t<\infty}\left\{P\left(S\left(X\right)\le t\,|\,Y=0\right)\log\frac{P\left(S\left(X\right)\le t\,|\,Y=0\right)}{P\left(S\left(X\right)\le t\,|\,Y=1\right)}\right\}. \tag{1}$$

在实践中，基于数据集$\left\{\left(y_i,x_i\right),i=1,\cdots,n\right\}$，则可得到分类模型的样本估计形式：

$$MKL_n=\sup_{-\infty<t<\infty}\left[\frac{1}{n_0}\sum_{y_i=0}I\left\{S\left(x_i\right)\le t\right\}\log\frac{\frac{1}{n_0}\sum_{y_i=0}I\left\{S\left(x_i\right)\le t\right\}}{\frac{1}{n_1}\sum_{y_i=1}I\left\{S\left(x_i\right)\le t\right\}}\right], \tag{2}$$

其中，$n_0=\sum_{i=1}^{n}I\left\{y_i=0\right\}$，$n_1=\sum_{i=1}^{n}I\left\{y_i=1\right\}$，$I\left\{\cdot\right\}$为示性函数，$S\left(x_i\right)$为根据分类模型估计的第$i$个样本的得分。现有的分类方法都是通过优化目标函数来估计样本得分的，例如，逻辑回归通过最大化似然函数估计回归参数，决策树通过最小化熵等损失函数来决定最佳分割。本文采用最大化 MKL 统计量来确定最优分割，假设

$$P\left(Y=1|\,X\right)=f\left(X^{\mathrm{T}}\beta_0\right), \tag{3}$$

其中$f$为0~1之间的未知递增函数，$\beta_0$为$p$维未知常向量，$p$为协变量$X$的维数。为了模型可识别性，设定$\left\|\beta_0\right\|=1$，其中$\left\|\bullet\right\|$为欧几里得范数，对于得分函数$S\left(X\right)=X^{\mathrm{T}}\beta$，也设定$\left\|\beta\right\|=1$。

## 2.2. 模型估计

设$\beta$的参数空间为$\mathrm{B}=\left\{\beta\in\mathbb{R}^p:\left\|\beta\right\|=1\right\}$。对于得分函数$S\left(X\right)=X^{\mathrm{T}}\beta$，其中$\beta\in\mathrm{B}$，则总体水平$MKL$定义为

$$MKL\left(\beta\right)=\sup_{-\infty<t<\infty}\left\{P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=0\right)\log\frac{P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=0\right)}{P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=1\right)}\right\}. \tag{4}$$

定义$MKL\left(\beta,t\right)=P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=0\right)\log\frac{P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=0\right)}{P\left(X^{\mathrm{T}}\beta\le t\,|\,Y=1\right)}$。在假设(3)式下，首先引入以下引理。

**引理 1：**在假设(3)下，如果$X^{\mathrm{T}}\beta$的分布不退化，则$MKL\left(\beta_0\right)=MKL\left(\beta_0,f^{-1}\left(\pi_1\right)\right)$，其中$\pi_1=P\left(Y=1\right)$。

由引理 1 可知，如果$\beta_0$已知，并且$S\left(X\right)=X^{\mathrm{T}}\beta_0$，则最佳分界点为$f^{-1}\left(\pi_1\right)$。注意，$X^{\mathrm{T}}\beta_0\le f^{-1}\left(\pi_1\right)$等价于$f\left(X^{\mathrm{T}}\beta_0\right)\le\pi_1$或$P\left(Y=0\,|\,X\right)>P\left(Y=0\right)$。在实际应用中，$\beta_0$是未知的，这种最佳分类无法实施。但引理 1 是下列定理的基础：当$\beta=\beta_0$时，$MKL\left(\beta\right)$达到其唯一的极大值。

**定理 1：**在假设(3)和引理 1 中的条件下，如果条件分布$F\left(X^{\mathrm{T}}\beta\,|\,X^{\mathrm{T}}\beta_0\right)$对任意$\beta\in\mathrm{B}$且$\beta\ne\pm\beta_0$不退化，则对任意的$\beta\in\mathrm{B}$且$\beta\ne\pm\beta_0$有$MKL\left(\beta_0\right)>MKL\left(\beta\right)$。

根据定理 1，本文提出通过最大化下列统计量来估计$\beta_0$，

$$MKL_n\left(\beta\right)=\sup_{-\infty<t<\infty}\left[\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\log\frac{\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}{\frac{1}{n_1}\sum_{y_i=1}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}\right]. \tag{5}$$

可定义

$$\hat{\beta}=\arg\max_{\left\|\beta\right\|=1}MKL_n\left(\beta\right). \tag{6}$$

得到$\hat{\beta}$后，得分函数则为$S\left(x_i\right)=x_i^{\mathrm{T}}\hat{\beta},i=1,\cdots,n$。下面的定理说明了$\hat{\beta}$的一致性，这一性质支撑了

本文所提出的分类方法。

**定理 2：**在假设(3)和定理 1 中的条件下，当 $n$ 趋于无穷时，(6)中的 $\hat{\beta}$ 在概率上趋于 $\beta_0$。

### 2.3. 算法步骤

因为我们的目标函数是离散的，为了更简便地优化(5)式，本文将其做连续化处理：

$$MKL_n(\beta,t)=\frac{1}{n_0}\sum_{y_i=0}\Phi\left(\frac{t-x_i^{\mathrm{T}}\beta}{h}\right)\log\frac{\frac{1}{n_0}\sum_{y_i=0}\Phi\left(\frac{t-x_i^{\mathrm{T}}\beta}{h}\right)}{\frac{1}{n_1}\sum_{y_i=1}\Phi\left(\frac{t-x_i^{\mathrm{T}}\beta}{h}\right)}, \tag{7}$$

其中，$\Phi\{\cdot\}$ 为标准正态分布函数，$h$ 为带宽。使用拟牛顿法(BFGS)确定参数 $\beta$，具体算法步骤如下：

Step1：设定初始值 $\beta^{(0)}=\left(\beta_1^{(0)},\cdots,\beta_p^{(0)}\right)^{\mathrm{T}}=\frac{\hat{\beta}_{\log}}{\left\|\hat{\beta}_{\log}\right\|}$ 并且满足 $\left\|\beta^{(0)}\right\|=1$，其中 $\hat{\beta}_{\log}$ 是数据集 $(X,Y)$ 拟合逻辑回归模型所得的回归系数，精度要求为 $\varepsilon$；

Step 2：真实模型的回归参数设为 $\beta_0$，则 $t$ 的初始值为 $t_0=x_i^{\mathrm{T}}\beta_0$；

Step 3：给定初始对称正定矩阵 $D_0=I_p$；

Step 4：计算搜索方向 $d^{(s)}=-D_s g_s$（$g_s$ 是 $\beta^{(s)}$ 的梯度)；

Step 5：计算最优步长：$\lambda_s=\arg\max L\left(\beta^{(s)}+\lambda d^{(s)},t_0\right)$，则 $\beta^{(s+1)}=\beta^{(s)}+\lambda_s d^{(s)}$；

Step 6：判断精度，若 $\left\|g_{s+1}\right\|<\varepsilon$，则停止迭代，否则进入下一步；

Step 7：计算 $\Delta g=g_{s+1}-g_s$，$\Delta\beta=\beta^{(s+1)}-\beta^{(s)}$，更新 $D$：

$$D_{s+1}=D_s+\frac{\Delta\beta\Delta\beta^{\mathrm{T}}}{\Delta g^{\mathrm{T}}\Delta\beta}-\frac{D_s\Delta g\Delta g^{\mathrm{T}}D_s}{\Delta g^{\mathrm{T}}D_s\Delta g}; \tag{8}$$

Step 8：令 $s=s+1$，进入 Step 4；

Step 9：通过约登指数确定最佳阈值 $t$。

## 3. 数值模拟

这一节进行了广泛的模拟研究，以验证本文提出的分类模型的效果。考虑以下三种不同的模型：

Model 1：$P(y_i=1\mid X_i)=\frac{1}{1+\mathrm{e}^{-x_i^{\mathrm{T}}\beta_0}}$，

Model 2：$P(y_i=1\mid X_i)=\frac{1}{1+\mathrm{e}^{-x_i^{\mathrm{T}}\beta_0 I\left\{x_i^{\mathrm{T}}\beta_0>0\right\}}}$，

Model 3：$P(y_i=1\mid X_i)=\frac{I\left\{x_i^{\mathrm{T}}\beta_0>-\frac{1}{4}\right\}}{1+\left(1+4x_i^{\mathrm{T}}\beta_0\right)^{-\frac{1}{4}}}$。

模型 1 为 logistic 模型，模型 2 为半 logistic 模型，模型 3 为带参数 1 的 Box-Cox 模型[7]。设 $p=5$，$x_i=\left(x_{i1},\cdots,x_{i5}\right)^{\mathrm{T}}$，其中 $x_i$ 是均值为 0，协方差矩阵为 $\Sigma=\left(\sigma_{ij}\right)_{p\times p}$，$\sigma_{ij}=0.5^{|i-j|}, 1\le i,j\le p$ 的五维多元正态随机向量。真实的回归参数 $\beta_0=(1,1,-1,-1,-1)$，样本量为 $n=200$ 或 1000，每次实验重复 500 次。

本文考虑四种方法进行比较：第一种方法是本文提出的以逻辑回归系数 $\frac{\hat{\beta}_{\log}}{\left\|\hat{\beta}_{\log}\right\|}$ 为初始点的 MKL 方法；第二种方法是 Logistic 回归，是目前最流行的二分类方法；第三种方法是基于高斯核函数的支持向量机

(SVM)；第四种方法是随机森林(RF)；第五种方法是神经网络(Net)，实现了一个包含 3 个隐藏层的神经网络。其中所有方法的分类阈值均通过约登指数来确定，即找到 ROC 曲线的最佳临界点。

对于上述四种方法，本文使用准确率(Accuracy)、精确率(Precision)、查全率(Recall)和 $F_1$ 分数四个定性指标进行评价，如式(8)~(11)所示。准确率能够判断总体的正确率，但是在样本不均衡的情况下，并不能作为很好的指标来衡量结果。精确率是针对预测结果而言的，代表了对正样本结果中的预测准确程度。召回率是针对原样本而言的，以精确率还是以召回率作为评价指标，需要根据具体问题而定，所以可以进一步比较 $F_1$ 分数，$F_1$ 分数同时考虑了精确率和召回率，让两者同时达到最高，取得平衡。此外，还以秒为单位计算时间，计算了各方法 500 次模拟的时间的平均值。所有数值结果见表 1。

$$\text{Accuracy} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{TP}+\text{FN}+\text{FP}}, \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{10}$$

$$F_1 = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{11}$$

**Table 1.** Comparison of results for different sample sizes in five methods
**表 1.** 五种方法不同样本量的结果对比

| 模型 | 方法 | Accuracy | Precision | Recall | $F_1$ | Time |
|---|---|---|---|---|---|---|
| | | | $n$ = 200 | | | |
| Model 1 | MKL | 0.833 | 0.957 | 0.710 | 0.815 | 0.024 |
| | Logistic | 0.850 | 0.923 | 0.774 | 0.842 | 0.002 |
| | SVM | 0.767 | 0.742 | 0.793 | 0.767 | 0.008 |
| | RF | 0.717 | 0.733 | 0.710 | 0.721 | 0.036 |
| | Net | 0.850 | 0.871 | 0.844 | 0.857 | 0.324 |
| Model 2 | MKL | 0.883 | 0.778 | 0.955 | 0.857 | 0.007 |
| | Logistic | 0.900 | 0.786 | 1.000 | 0.880 | 0.003 |
| | SVM | 0.917 | 1.000 | 0.815 | 0.900 | 0.007 |
| | RF | 0.900 | 0.681 | 0.955 | 0.875 | 0.033 |
| | Net | 0.833 | 0.955 | 0.833 | 0.750 | 0.330 |
| Model 3 | MKL | 0.817 | 0.782 | 0.750 | 0.766 | 0.006 |
| | Logistic | 0.767 | 0.632 | 1.000 | 0.774 | 0.003 |
| | SVM | 0.683 | 0.500 | 0.632 | 0.558 | 0.008 |
| | RF | 0.733 | 0.700 | 0.5 83 | 0.636 | 0.038 |
| | Net | 0.783 | 0.833 | 0.690 | 0.755 | 0.235 |

**Continued**

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | $n$ = 1000 | | | |
| Model 1 | MKL | 0.813 | 0.843 | 0.812 | 0.827 | 0.026 |
| | Logistic | 0.843 | 0.801 | 0.952 | 0.870 | 0.004 |
| | SVM | 0.770 | 0.776 | 0.800 | 0.788 | 0.046 |
| | RF | 0.803 | 0.812 | 0.837 | 0.824 | 0.195 |
| | Net | 0.803 | 0.806 | 0.831 | 0.818 | 1.137 |
| Model 2 | MKL | 0.913 | 0.868 | 0.956 | 0.910 | 0.032 |
| | Logistic | 0.923 | 0.865 | 0.985 | 0.922 | 0.004 |
| | SVM | 0.890 | 0.869 | 0.888 | 0.878 | 0.025 |
| | RF | 0.870 | 0.895 | 0.810 | 0.851 | 0.161 |
| | Net | 0.930 | 0.956 | 0.897 | 0.926 | 1.571 |
| Model 3 | MKL | 0.823 | 0.729 | 0.851 | 0.785 | 0.026 |
| | Logistic | 0.803 | 0.687 | 0.886 | 0.773 | 0.003 |
| | SVM | 0.723 | 0.719 | 0.617 | 0.664 | 0.034 |
| | RF | 0.777 | 0.704 | 0.711 | 0.707 | 0.169 |
| | Net | 0.787 | 0.974 | 0.645 | 0.776 | 1.432 |

可以看出，对于不同模型和不同的样本量，本文提出的 MKL 方法和其他传统的四种方法的分类效果很接近，也就是说，MKL 方法是可用的甚至有较强的竞争力。就模型一具体来说，在样本量较小的时候，因为真实模型就是逻辑回归，所以 Logistic 回归方法的分类性能是最好的，其准确率、精确率都是最高的，这符合我们的预期，同时 MKL 方法的分类性能也相对较好，其准确率仅次于 Logistic 回归，且精确率是最高的。三种模型中，支持向量机和随机森林方法的各项指标都是最低的，且出现了较大的偏差，说明模型质量较差，模型稳定性也不是很好。当样本量增大时，应用于三种模型的五种分类方法的性能均有所提升，且 MKL 方法的分类准确度、预测性能和模型质量均较好。同时就计算速率而言，MKL 方法计算速度较快，神经网络计算速度最慢。并且，纵向比较不同模型的方法应用效果，模型二的分类效果是最高的，这就说明对于不同的数据特征，分类方法的效果也是不一样的，因此在实例中通过分析数据和样本特征，或许能更好地应用我们的方法。

## 4. 实例分析

心血管疾病(Cardiovascular disease, CVD)是最常见的死亡原因之一，每年造成约 1700 万人死亡。心血管疾病的主要原因是心肌梗死和心脏不能正常供血。医生可以根据患者的症状和临床实验室调查，通过电子病历诊断心力衰竭(Heart failure, HF)。然而，心力衰竭的准确诊断需要医疗资源和专业人员，而这些资源并不总是可用的，因此诊断具有挑战性。故而，利用机器学习算法来预测患者的病情是节省时间和精力的必要方法。

本文引用了 UCI 数据库中的心力衰竭临床记录数据集[8]，该数据集包含 299 名心力衰竭患者，共 12 个特征，分别为年龄、血清钠、血清肌酐、性别、吸烟、血压、射血分数、贫血、血小板、肌酐磷酸激酶、糖尿病、随访期。因变量为二分类变量，在随访期因心脏衰竭死亡视为 0，未死亡视为 1。针对该数据集，本文仍然使用 MKL、Logistic 回归、支持向量机、随机森林和神经网络这五种方法进行模型拟合，各模型分类效果如图 1 所示。

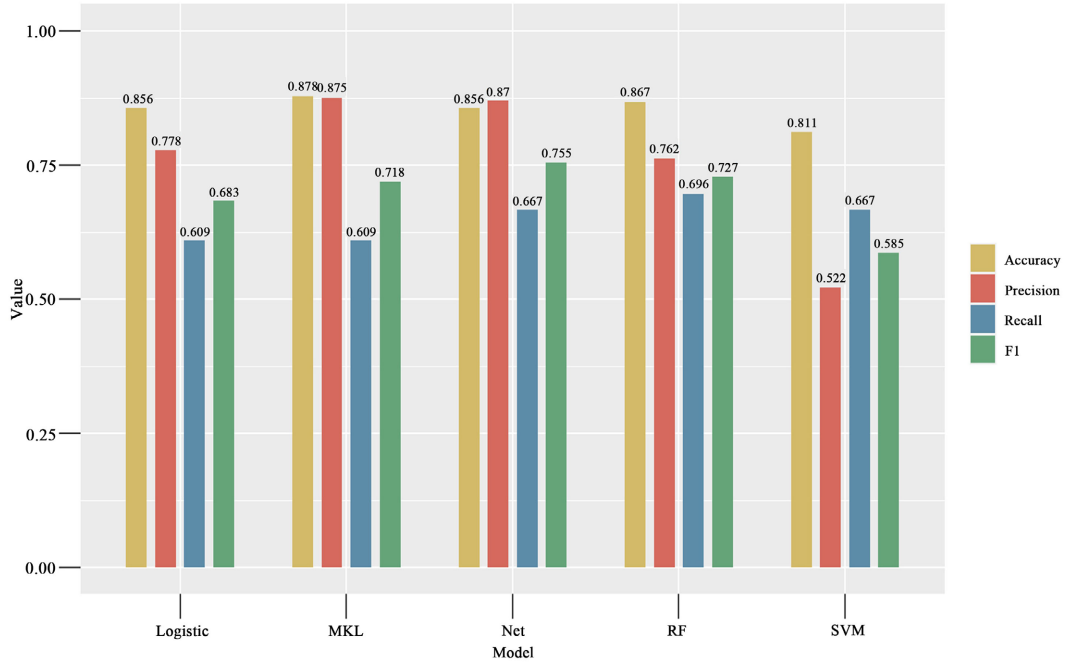


**Figure 1.** Five methods of classification indicator comparison
**图 1.** 五种方法分类指标对比

从图 1 中可以看出，本文所提出的分类模型 MKL 取得了最优的分类效果，其准确率(0.878)和精确率(0.875)不仅是最高的，且两者偏差十分微小，说明模型分类效果和预测性能均较好，同时 $F_1$ 分数(0.71)也相对较高。对于该数据集，神经网络的分类效果也不相上下，然而其余三种方法出现了较大的偏差，尤其支持向量机因为样本不均衡原因导致准确率和精确率相差很大，模型质量最差。

## 5. 结论

本文的贡献在于提出了一种基于得分函数新的概率分类模型 MKL，并通过拟牛顿算法直接对连续化后的 MKL 统计量进行优化，从理论上证明了所提出的 MKL 估计的一致性。该方法在模型效果上比目前最流行的分类模型逻辑回归更有优势，针对不同的数据集，MKL 模型与其他传统的分类模型如支持向量机、随机森林、神经网络相比，该方法在预测能力、计算复杂度和实际可解释性方面具有很强的竞争力。同时本文将 MKL 方法应用到心脏衰竭数据集上，以较高的准确率(0.878)对病人是否因为心脏衰竭死亡进行了分类预测，说明我们提出的方法也可以应用到相似领域，更一般地说，对于任何二分类问题，MKL 方法可能是可用的。在未来，本文所提出的模型可以考虑应用到超高维数据或者多分类问题上，改进优化算法以提高模型分类效果。

## 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)，2018.1~2021.12。

## 参考文献

[1] Cramer, J.S. (2002) The Origins of Logistic Regression. Tinbergen Institute Discussion Papers No. 2002-119/4. https://doi.org/10.2139/ssrn.360300

[2] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[3] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297.

https://doi.org/10.1007/BF00994018

[4] Lecun, Y., Bottou, L., Bengio Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://ieeexplore.ieee.org/document/726791 https://doi.org/10.1109/5.726791

[5] Zhang, S.C., Li, X.L., *et al*. (2007) Learning *k* for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, **8**, 1-19. https://doi.org/10.1145/2990508

[6] Fang, F. and Chen, Y. (2018) A New Approach for Credit Scoring by Directly Maximizing the Kolmogorov-Smirnov Statistic. *Computational Stats & Data Analysis*, **133**, 180-194. https://doi.org/10.1016/j.csda.2018.10.004

[7] Guerrero, V.M. and Johnson, R.A. (1982) Use of the Box-Cox Transformation with Binary Response Models. *Biometrika*, **69**, 309-314. https://doi.org/10.1093/biomet/69.2.309

[8] Chicco, D. and Jurman, G. (2020) Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone. *BMC Medical Informatics and Decision Making*, **20**, 1-16. https://doi.org/10.1186/s12911-020-1023-5

## 附录

引理 1 证明：首先

$$\begin{aligned}P\left(X^{\mathrm{T}}\beta_0>t,Y=0\right)&=E\left\{I\left(X^{\mathrm{T}}\beta_0>t,Y=0\right)\right\}\\&=E\left\{E\left[I\left(X^{\mathrm{T}}\beta_0>t\right)I\left(Y=0\right)\right]\mid X^{\mathrm{T}}\beta_0\right\}\\&=E\left\{I\left(X^{\mathrm{T}}\beta_0>t\right)\left[1-P\left(Y=1\mid X^{\mathrm{T}}\beta_0\right)\right]\right\}\\&=E\left\{I\left(X^{\mathrm{T}}\beta_0>t\right)\left[1-f\left(X^{\mathrm{T}}\beta_0\right)\right]\right\}\\&=\int I\left(X^{\mathrm{T}}\beta_0>t\right)\left[1-f\left(X^{\mathrm{T}}\beta_0\right)\right]\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right).\end{aligned}$$

令 $T=X^{\mathrm{T}}\beta_0$，$\pi_1=P(Y=1)$，$\pi_0=P(Y=0)$。引用 Fang 和 Chen [6]定理一中的结论，由 $KS(t)=1-\left[\int_{-\infty}^{t}\frac{f(T)}{\pi_1}f(T)\mathrm{d}T+\int_{t}^{+\infty}\frac{1-f(T)}{\pi_0}f(T)\mathrm{d}T\right]$，令其导数为 0 即 $KS'(t)=\frac{1-f(t)}{\pi_0}f(t)-\frac{f(t)}{\pi_1}f(t)=0$，可得 $\frac{1-f(t)}{\pi_0}=\frac{f(t)}{\pi_1}$，即 $t=f^{-1}(\pi_1)$ 时 $KS$ 最大。$MKL(\beta_0,t)$ 可做以下变形：

$$\begin{aligned}MKL\left(\beta_0,t\right)&=P\left(X^{\mathrm{T}}\beta_0\le t\mid Y=0\right)\log\frac{P\left(X^{\mathrm{T}}\beta_0\le t\mid Y=0\right)}{P\left(X^{\mathrm{T}}\beta_0\le t\mid Y=1\right)}\\&=\log\left[\frac{1-P\left(X^{\mathrm{T}}\beta_0>t\mid Y=0\right)}{P\left(X^{\mathrm{T}}\beta_0\le t\mid Y=1\right)}\right]^{\left[1-P\left(X^{\mathrm{T}}\beta_0>t\mid Y=0\right)\right]}\\&=\log\left[\frac{1-\int_{t}^{+\infty}\frac{1-f(T)}{\pi_0}f(T)\mathrm{d}T}{\int_{-\infty}^{t}\frac{f(T)}{\pi_1}f(T)\mathrm{d}T}\right]^{\left[1-\int_{t}^{+\infty}\frac{1-f(T)}{\pi_0}f(T)\mathrm{d}T\right]}\\&\triangleq\log\left[\frac{h(t)}{g(t)}\right]^{h(t)}=h(t)\log h(t)-h(t)\log g(t)\end{aligned}$$

其中 $h'(t)=\frac{1-f(t)}{\pi_0}f(t)$，$g'(t)=\frac{f(t)}{\pi_1}f(t)$，所以

$$\begin{aligned}\frac{\mathrm{d}MKL}{\mathrm{d}t}&=h'(t)\log h(t)+h'(t)-h'(t)\log g(t)-\frac{h(t)}{g(t)}g'(t)\\&=\frac{1-f(t)}{\pi_0}f(t)\log\frac{h(t)}{g(t)}+\frac{1-f(t)}{\pi_0}f(t)-\frac{h(t)}{g(t)}\frac{f(t)}{\pi_1}f(t)\end{aligned}$$

令 $\frac{\mathrm{d}MKL}{\mathrm{d}t}=0$ 得 $\frac{1-f(t)}{\pi_0}\log\frac{h(t)}{g(t)}-\frac{h(t)}{g(t)}\frac{f(t)}{\pi_1}+\frac{1-f(t)}{\pi_0}=0$。

当 $t=f^{-1}(\pi_1)$ 时，上式化为

$$\log\frac{h(t)}{g(t)}-\frac{h(t)}{g(t)}+1=\log\left[1-\frac{g(t)-h(t)}{g(t)}\right]+\frac{g(t)-h(t)}{g(t)}=0 \tag{1}$$

此时即证 $t=f^{-1}(\pi_1)$ 时，(1)成立。

因为函数 $y=\log\dfrac{h(t)}{g(t)}=\log h(t)-\log g(t)$，令其导数为 0

$$y'=\frac{h'(t)}{h(t)}-\frac{g'(t)}{g(t)}=\frac{1}{h(t)}\frac{1-f(t)}{\pi_0}f(t)-\frac{1}{g(t)}\frac{f(t)}{\pi_1}f(t)=0,$$

$t=f^{-1}(\pi_1)$ 时，有 $\dfrac{1}{h(t)}-\dfrac{1}{g(t)}=\dfrac{g(t)-h(t)}{h(t)g(t)}=0$，即 $g(t)-h(t)=0$，故(1)成立。所以 $t=f^{-1}(\pi_1)$ 是 $MKL(t)$ 的极大值点。

下证 $t=f^{-1}(\pi_1)$ 是 $MKL(t)$ 的唯一极大值点：

假设 $t=a$ 也是 $MKL(t)$ 的极大值点($a\neq f^{-1}(\pi_1)$)，则有

$\left.\dfrac{\mathrm{d}MKL}{\mathrm{d}t}\right|_{t=a}=\dfrac{1-f(a)}{\pi_0}\log\dfrac{h(a)}{g(a)}-\dfrac{h(a)}{g(a)}\dfrac{f(a)}{\pi_1}+\dfrac{1-f(a)}{\pi_0}=0$，即

$$\left[\frac{1}{f(a)}-1\right]\left[\log\frac{h(a)}{g(a)}+1\right]=\frac{h(a)}{g(a)}\left(\frac{1}{\pi_1}-1\right)\Rightarrow\frac{\dfrac{1}{f(a)}-1}{\dfrac{1}{\pi_1}-1}=\frac{\dfrac{h(a)}{g(a)}}{\log\dfrac{h(a)}{g(a)}+1}\tag{2}$$

因为 $y=\dfrac{1}{f(x)}-1$ 是单调递减函数，$y=\log\dfrac{h(x)}{g(x)}+1\,(x>1)$ 是单调递增函数，所以(2)只有唯一解与假设矛盾，即证 $t=f^{-1}(\pi_1)$ 是 $MKL(t)$ 的唯一极大值点。

定理 1 证明：在假设(3)的条件下，$Y$ 和 $X^{\mathrm{T}}\beta$ 在 $X^{\mathrm{T}}\beta_0$ 的条件下是独立的，

$$\begin{aligned}MKL(\beta_0,t)&=P\left(X^{\mathrm{T}}\beta_0\le t\,|\,Y=0\right)\log\frac{P\left(X^{\mathrm{T}}\beta_0\le t\,|\,Y=0\right)}{P\left(X^{\mathrm{T}}\beta_0\le t\,|\,Y=1\right)}\\&=\log\left[\frac{1-E_{X^{\mathrm{T}}\beta_0}\int\dfrac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}I\left(X^{\mathrm{T}}\beta_0>t\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta\,|\,X^{\mathrm{T}}\beta_0\right)}{E_{X^{\mathrm{T}}\beta_0}\int\dfrac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}I\left(X^{\mathrm{T}}\beta_0\le t\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta\,|\,X^{\mathrm{T}}\beta_0\right)}\right]^{P\left(X^{\mathrm{T}}\beta_0\le t|Y=0\right)}\\&\triangleq\log g\left(X^{\mathrm{T}}\beta_0,X^{\mathrm{T}}\beta,t\right)^{P\left(X^{\mathrm{T}}\beta_0\le t|Y=0\right)},\end{aligned}$$

对于给定的 $X^{\mathrm{T}}\beta_0$，并且 $\beta\neq\pm\beta_0$，若 $\dfrac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}>\dfrac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}$，即 $X^{\mathrm{T}}\beta_0<f^{-1}(\pi_1)$，$\dfrac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}\le g\left(X^{\mathrm{T}}\beta_0,X^{\mathrm{T}}\beta,t\right)\le\dfrac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}$。当 $t=-\infty$ 时，$g$ 达最大值，$t=\infty$ 时，$g$ 达最小值；若 $\dfrac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}<\dfrac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}$，即 $X^{\mathrm{T}}\beta_0>f^{-1}(\pi_1)$，$\dfrac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}\ge g\left(X^{\mathrm{T}}\beta_0,X^{\mathrm{T}}\beta,t\right)\ge\dfrac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}$。当 $t=-\infty$ 时，$g$ 达最小值，$t=\infty$ 时，$g$ 达最大值。

令 $A=\left\{X^{\mathrm{T}}\beta_0<f^{-1}(\pi_1)\right\}$，$B=\left\{X^{\mathrm{T}}\beta_0>f^{-1}(\pi_1)\right\}$，对 $\forall\beta\neq\pm\beta_0$，

$$MKL(\beta_0,t)=\log\left[\frac{1-\int_A\int g\left(X^{\mathrm{T}}\beta_0,X^{\mathrm{T}}\beta,t\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta\mid X^{\mathrm{T}}\beta_0\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right)}{\int_B\int g\left(X^{\mathrm{T}}\beta_0,X^{\mathrm{T}}\beta,t\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta\mid X^{\mathrm{T}}\beta_0\right)\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right)}\right]^{P\left(X^{\mathrm{T}}\beta_0\le t\mid Y=0\right)}$$

$$<\log\left[\frac{1-\int_A\frac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right)}{\int_B\frac{1-f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=0)}\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right)}\right]^{\left[1-\int_A\frac{f\left(X^{\mathrm{T}}\beta_0\right)}{P(Y=1)}\mathrm{d}F\left(X^{\mathrm{T}}\beta_0\right)\right]}$$

$$=MKL\left(\beta_0,f^{-1}(\pi_1)\right)=MKL(\beta_0),$$

其中，是“<”而不是“≤”，因为等式在 $A$ 和 $B$ 条件下不可能同时成立，所以有 $MKL(\beta)<MKL(\beta_0)$（当 $\beta\neq\pm\beta_0$ 时)。

定理 2 证明：即证 $\sup_{\|\beta\|=1}\left|MKL_n(\beta)-MKL(\beta)\right|\xrightarrow{P}0$。对 $\forall\varepsilon>0$，

$$P\left(\sup_{\|\beta\|=1}\left|MKL_n(\beta)-MKL(\beta)\right|>\varepsilon\right)$$

$$\le P\left(\sup_{\beta,t}\left|MKL_n(\beta)-MKL(\beta)\right|>\varepsilon\right)$$

$$=P\left(\sup_{\beta,t}\left|\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\log\frac{\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}{\frac{1}{n_1}\sum_{y_i=1}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\log\frac{P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)}{P\left(X^{\mathrm{T}}\beta\le t\mid Y=1\right)}\right|>\varepsilon\right)$$

$$\le P\left(\sup_{\beta,t}\left|\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\log\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\log P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\right|>\frac{\varepsilon}{2}\right)$$

$$+P\left(\sup_{\beta,t}\left|\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\log\frac{1}{n_1}\sum_{y_i=1}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\log P\left(X^{\mathrm{T}}\beta\le t\mid Y=1\right)\right|>\frac{\varepsilon}{2}\right)$$

$$\triangleq P\left(\sup_{\beta,t}A_1>\frac{\varepsilon}{2}\right)+P\left(\sup_{\beta,t}A_2>\frac{\varepsilon}{2}\right)$$

因为

$$A_1=\left|\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\log\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\log P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\right|$$

$$\le\left|\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-\frac{1}{nP(Y=0)}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}\right|+\frac{1}{P(Y=0)}\left|\frac{1}{n}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\right|$$

$$+\left|\log\frac{1}{n_0}\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}-\log\frac{\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}{nP(Y=0)}\right|+\left|\log\frac{\sum_{y_i=0}I\left\{x_i^{\mathrm{T}}\beta\le t\right\}}{nP(Y=0)}-\log P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\right|$$

$$\le\frac{1}{P(Y=0)}\left|\frac{1}{n}\sum_{i=1}^{n}I\left\{y_i=0\right\}-P(Y=0)\right|+\frac{1}{P(Y=0)}\left|\frac{1}{n}\sum_{i=1}^{n}I\left\{x_i^{\mathrm{T}}\beta\le t,y_i=0\right\}-P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)\right|$$

$$+\left|\log\frac{\frac{1}{n}\sum_{i=1}^{n}I\left\{y_i=0\right\}}{P(Y=0)}\right|+\left|\log\frac{\frac{1}{n}\sum_{i=1}^{n}I\left\{x_i^{\mathrm{T}}\beta\le t,y_i=0\right\}}{P\left(X^{\mathrm{T}}\beta\le t\mid Y=0\right)}\right|$$

所以有

$$P\left(\sup_{\beta,t} A_1 > \frac{\varepsilon}{2}\right) \le P\left(\frac{1}{P(Y=0)}\left|\frac{1}{n}\sum_{i=1}^{n} I\{y_i=0\} - P(Y=0)\right| > \frac{\varepsilon}{8}\right)$$
$$+ P\left(\frac{1}{P(Y=0)}\sup_{\beta,t}\left|\frac{1}{n}\sum_{i=1}^{n} I\left\{x_i^{\mathrm{T}}\beta \le t, y_i=0\right\} - P\left(X^{\mathrm{T}}\beta \le t \mid Y=0\right)\right| > \frac{\varepsilon}{8}\right)$$
$$+ P\left(\left|\log\frac{1}{n}\sum_{i=1}^{n} I\{y_i=0\} - \log P(Y=0)\right| > \frac{\varepsilon}{8}\right)$$
$$+ P\left(\left|\log\frac{1}{n}\sum_{i=1}^{n} I\left\{x_i^{\mathrm{T}}\beta \le t, y_i=0\right\} - \log P\left(X^{\mathrm{T}}\beta \le t \mid Y=0\right)\right| > \frac{\varepsilon}{8}\right) \to 0$$

同理 $P\left(\sup_{\beta,t} A_2 > \frac{\varepsilon}{2}\right) \to 0$，故 $\sup_{\|\beta\|=1}\left|MKL_n(\beta) - MKL(\beta)\right| \xrightarrow{P} 0$ 得证。