第49卷第12期 中国科学技术大学学报 Vol. 49, No. 12
2019年12月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Dec. 2019

文章编号:0253-2778(2019)12-0965-09

一种处理非均衡数据的非迭代核逻辑回归方法

崔文泉,余德美,程浩洋

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:针对严重非均衡数据提出一种非迭代核逻辑回归的学习方法.该方法是对经典处理核逻辑回归的迭代加权最小二乘方法的一种改进,不仅减轻了由于迭代所造成的运算负担,而且在模型训练中利用了基准的类别占比信息,避免了使用诸如欠抽样、过抽样、代价敏感学习等通常处理非均衡数据的方式所导致的问题,使得在数据规模大的非均衡数据情形下,可以方便快捷地对核逻辑回归进行建模,构造具有稳健性的修正最小二乘逻辑回归分类器.理论研究表明,所提方法具有一定的优良性质,模拟研究及实证分析显示其分类效果良好.

关键词:核逻辑回归;非迭代方法;非均衡数据;迭代加权最小二乘;稳健

中图分类号:O212.1 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.12.003

2010 Mathematics Subject Classification: Primary 62H30; Secondary 62J12

引用格式:崔文泉,余德美,程浩洋. 一种处理非均衡数据的非迭代核逻辑回归方法[J]. 中国科学技术大学学报,2019,49(12):965-973.

CUI Wenquan, YU Demei, CHENG Haoyang. A non-iterative approach to kernel logistic regression for imbalanced data[J]. Journal of University of Science and Technology of China, 2019,49(12):965-973.

A non-iterative approach to kernel logistic regression for imbalanced data

CUI Wenquan, YU Demei, CHENG Haoyang

(*Department of Statistics and Finance, School of Management, University of Science and of Technology of China, Hefei* 230026, *China*)

Abstract: A non-iterative kernel logistic regression learning method for severely imbalanced data was proposed. The method is an improvement on the iterative weighted least squares method for classical kernel logistic regression. It not only reduces the computational burden caused by iteration, but also utilizes the knowledge of the ratio of the benchmark category, and can avoid problems normally encountered when processing imbalanced data such as undersampling, oversampling and cost-sensitive learning. Thus, this method enables the efficient and fast modelling of kernel based logistic regression in the case of large-scale imbalanced data, through the construction of a robust modified least square logistic classifier. Theoretical research indicates that the proposed method has some excellent properties, and simulation research and empirical studies show that its classification effect is good.

Key words: kernel logistic regression; non-iterative approach; imbalanced data; iterative re-weighted least squares; robustness

收稿日期:2019-04-14;**修回日期**:2019-05-22

基金项目:国家自然科学基金(71873128),安徽省自然科学基金(1308085MA02)资助.

作者简介:崔文泉(通讯作者),男,1964年生,博士/副教授. 研究方向:数理统计. E-mail: wqcui@ustc.edu.cn

0 引言

非均衡数据的分类问题广泛存在于如软件缺陷分析[1]、医疗诊断[2]、电信诈骗[3]、个人信用风险[4]、自然灾害[5]、网络攻击[6]等领域，在此类问题中，对少数类识别错误的代价往往高于多数类. 例如，在某种罕见疾病的诊断中，对患者的诊断错误会严重影响其治疗进程，因而分类模型在少数类上实现更高的准确率是至关重要的. López 等发现经典的分类方法如决策树(Decision Tree)、支持向量机(support vector machines, SVM)、k 最近邻(KNN)在处理非均衡数据时会出现对少数类样本识别不够准确的问题[7]. 近年来大量研究对如何处理非均衡数据提出了各种解决方法[8]，一般归结为数据层面和算法层面. 数据层面最常用的方法是对数据进行欠抽样或者过抽样，从而降低数据的不均衡程度[9-12].

传统的随机欠抽样是随机舍弃多数类中的部分样本，可能存在因去除有意义的模式而导致有效信息损失的问题；而随机过采样通过对少数类样本的简单复制来增加少数类样本量，可能导致过拟合和计算负担的增加. 为此，Chawla 等提出一种过采样方法 SMOTE(synthetic minority over-sampling technique)[13]，通过合成的方法生成新少数类样本点，有效扩大了分类的决策边界，从而提高分类模型的泛化能力以及对少数类样本的识别能力. 然而，SMOTE 方法仅适用于连续输入特征空间的二分类问题. Nguyen 等[14]指出当少数类样本量不够时，会因对真实分布的有效估计而损害 SMOTE 方法的提升效果. 总的来说抽样方法操作简单，但存在有效信息损失或过度拟合的问题. 算法层面中一类重要方法为代价敏感学习(cost sensitive learning). Uyar 等[15]在对体外受精胚胎移植结果的预测中发现通过调整决策边界的阈值可以获得最优的真正确率(true positive)和误报率(false positive). Yang 等[16]提出一种代价敏感的支持向量机，使分类间隔偏移获得无偏的决策边界，在复杂非均衡数据获得较好的分类结果.

在统计方法中，逻辑回归作为广义线性模型中的一种用于构造分类的模型，由于模型简单、易于解释和实现，被广泛使用. West[17]在研究信用风险评估问题中，对比了传统的模型，如线性判别分析、逻辑回归、KNN、核密度估计、决策树，发现其中逻辑回归预测最准确. 然而，现实中大量分类问题属于线性不可分问题，核逻辑回归将核方法引入逻辑回归，在处理非线性分类问题时有良好性质[18]. 相比于同样采用核方法的支持向量机，核逻辑回归不仅给出分类预测的概率而且可以推广应用到多分类问题领域[20-21]. 在求解核逻辑回归问题时，最常使用的方法是牛顿法也即迭代加权最小二乘法(iterative re-weighted least squares，IRLS). IRLS 方法缺点在于迭代过程需要进行多次的矩阵求逆运算，不适用于大规模数据的分类问题. Zhu 和 Hastie[18]提出输入向量机(import vector machine)，使用训练集部分样本点作为输入点构造核基函数，从而达到近似核逻辑回归模型的效果，并降低模型计算复杂度. 2010 年，Maalouf 等[22]将截断牛顿法(truncated Newton method)应用于核逻辑回归，简化为求解非限制正则最优问题(unconstrained regularized optimization). 该方法成功用于中小规模数据集，但对于大规模数据运算费时. 在此基础上，Elbashir 和 Wang[23]于 2015 年提出了 NTR-KLR(Nystrom truncated KLR)算法，借用 Nystrom 方法，通过对核矩阵进行特征分解，并取部分特征向量空间来近似核矩阵，降低计算复杂度，获得与 SVM 相当的分类效果. 然而，以上方法都属于迭代方法，在估计核逻辑回归参数都面临不断迭代进行矩阵求逆计算的问题. 2016 年，Ngufor 和 Wojtusiak[24]提出最小二乘近似法(least squares KLR，LS-KLR)通过在 IRLS 步骤中对逻辑函数进行泰勒展开近似，将原来的迭代求解问题简化为非迭代问题. 该方法原理是将核逻辑回归的学习限定在对最有价值的分类边界点的特征的学习，不仅极大提高速度而且对均衡数据取得和 IRLS 方法相媲美的分类准确度. 然而，该方法假定分类边界处的点属于两类的概率相等，而当分类数据不均衡时该方法不再适用. 2018 年，Maalouf[6]针对抽样方法下的逻辑回归，引入先验信息校正方式解决因抽样导致的参数估计有偏的问题，使其适应于严重非均衡数据的分类问题. 然而，该方法同时从数据层面和算法层面来处理非均衡数据问题，在操作层面较为复杂，并依然存在上面提到的迭代计算问题.

受 LS-KLR 启发，本文提出一种基于非均衡数据的核逻辑回归方法 LS-RKLR. LS-RKLR 引入了基准概率下的广义逻辑函数，并采用泰勒展开近似方法，构造出一种快速、稳健的核逻辑回归方法. 节

1 给出该方法的推导过程以及算法流程，并给出其参数估计性质的相应证明. 节 2 通过模拟数据和 UCI 公共数据实验表明，该方法相比于 LS-KLR 能有效提高非均衡数据的分类表现.

1　方法及性质

1.1　方法实现

本文主要对核逻辑回归方法进行修改，使之对于非均衡数据具有稳健分类效果. 对于二分类问题，记训练集为 $\mathcal{D}=\{\boldsymbol{x}_i,y_i,i=1,\cdots,n\}$，其中 $\boldsymbol{x}_i\in\mathbb{R}^d$ 为输入变量，$y_i\in\{0,1\}$为输出变量. 经典逻辑回归函数的形式为

$$\pi(\boldsymbol{x}_i)=P(y_i=1\mid\boldsymbol{x}_i)=\frac{1}{1+\mathrm{e}^{-f(\boldsymbol{x}_i)}},i=1,\cdots,n.$$

则对于逻辑回归，在 $f(\boldsymbol{x})=0$ 时

$$P(y=1\mid f(\boldsymbol{x})=0)=\frac{1}{2}.$$

本文称 $P(y=1|f(\boldsymbol{x})=0)$为基准概率，显然，对于经典逻辑回归，基准概率为$\frac{1}{2}$. 对于本文所讨论的非均衡数据情形下的问题，当 $y=1$ 代表少数类时，我们将经典的逻辑回归推广至广义逻辑回归情形，此时，我们设定基准概率小于$\frac{1}{2}$，推广情形的广义逻辑函数[25]形式如下：

$$\pi(\boldsymbol{x}_i)=P(y_i=1\mid\boldsymbol{x}_i)=\frac{1}{1+\gamma\mathrm{e}^{-f(\boldsymbol{x}_i)}},$$

其中，$\gamma=\frac{1-\tau}{\tau}\in(0,+\infty)$，$\tau=P(y=1|f(\boldsymbol{x})=0)$ 即为上述定义的基准概率. 则其相应的正则化负对数似然函数为

$$l(f)=\sum_{i=1}^{n}\left[-y_i(f(\boldsymbol{x}_i)-\ln\gamma)+\ln(1+\frac{1}{\gamma}\mathrm{e}^{f(\boldsymbol{x}_i)})\right]+\lambda J(f) \tag{1}$$

式中，$J(f)$为惩罚函数，$f\in\mathscr{H}$且 $\mathscr{H}$为 $J(f)$的定义函数空间. 接下来将核方法引入逻辑回归，假定 $f\in\mathscr{H}_K$，其中 $\mathscr{H}_K$ 为正定核 $K(\cdot,\cdot)$生成的再生核希尔伯特空间(reproducing kernel Hilbert space，RKHS). 则 f 的核表示形式为

$$f(\boldsymbol{x})=\sum_{m=1}^{\infty}\alpha_m K(\boldsymbol{x},\boldsymbol{y}_m)\ ,\boldsymbol{x},\boldsymbol{y}_m\in\mathbb{R}^d.$$

式(1)可写为

$$l(f)=\sum_{i=1}^{n}\left[-y_i(f(\boldsymbol{x}_i)-\ln\gamma)+\ln(1+\frac{1}{\gamma}\mathrm{e}^{f(\boldsymbol{x}_i)})\right]+\lambda\|f\|_{\mathscr{H}_K}^2 \tag{2}$$

因此由表示定理[26]可知，$\min\limits_{f\in\mathscr{H}_K}l(f)$的最优解是有限维的，形式为

$$f(\boldsymbol{x})=\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i).$$

由 $\mathscr{H}_K$ 再生性质，有

$$\|f\|_{\mathscr{H}_K}^2=\langle f,f\rangle_{\mathscr{H}_K}=\langle\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i),\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)\rangle=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

则式(2)等价于

$$\boldsymbol{1}(\boldsymbol{\alpha})=-\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K}\boldsymbol{\alpha})+\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}_n\ln\gamma+\boldsymbol{1}_n^{\mathrm{T}}\ln(1+\frac{1}{\gamma}\mathrm{e}^{\boldsymbol{K}\boldsymbol{\alpha}})+\frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{3}$$

式中，$\boldsymbol{K}$ 是 Gram 矩阵，其元素

$$K_{ij}=\langle\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)\rangle=K(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

最小化式(3)，可以令$\frac{\partial l(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}=0$，并使用 Newton-Raphson 方法求估计方程的解. 该算法也被称为迭代加权最小二乘法(IRLS)，可以得到

$$\boldsymbol{\alpha}^{\mathrm{new}}=(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{K}+\lambda\boldsymbol{K})^{-1}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z} \tag{4}$$

式中，权重矩阵 $\boldsymbol{W}=\mathrm{diag}\{w_1,\cdots,w_n\}$，$w_i=\pi(\boldsymbol{x}_i)\cdot(1-\pi(\boldsymbol{x}_i))$，$i=1,\cdots,n$，$\boldsymbol{z}=\boldsymbol{K}\boldsymbol{\alpha}^{\mathrm{old}}+\boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{\pi})$，$\boldsymbol{\pi}=(\pi(\boldsymbol{x}_1),\cdots,\pi(\boldsymbol{x}_n))^{\mathrm{T}}$. 式(4)可以看成下面加权最小二乘的解：

$$\min_{\boldsymbol{\alpha}}L=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}+\frac{1}{2\lambda}(\boldsymbol{z}-\boldsymbol{K}\boldsymbol{\alpha})\boldsymbol{W}(\boldsymbol{z}-\boldsymbol{K}\boldsymbol{\alpha})^{\mathrm{T}},$$

可进一步理解成约束最优化问题[24]

$$\min_{\boldsymbol{\alpha},\varepsilon_i}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}+\frac{1}{2\lambda}\sum_{i=1}^{n}w_i\varepsilon_i^2$$

$$\text{s.t.}\quad z_i=\sum_{j=1}^{n}\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b-\ln\gamma+\varepsilon_i,\quad\forall i=1,\cdots,n.$$

其中，常数项 b 重新被引入. 引入拉格朗日算子法，可以得到

$$\mathscr{L}(\boldsymbol{\alpha},b,\boldsymbol{\xi},\boldsymbol{\varepsilon})=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}+\frac{1}{2\lambda}\sum_{i=1}^{n}w_i\varepsilon_i^2-\sum_{i=1}^{n}\xi_i(\sum_{j=1}^{n}\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+b-\ln\gamma+\varepsilon_i-z_i)=\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}+\frac{1}{2\lambda}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\varepsilon}-\boldsymbol{\xi}^{\mathrm{T}}(\boldsymbol{K}\boldsymbol{\alpha}+(b-\ln\gamma)\boldsymbol{1}_n+\boldsymbol{\varepsilon}-\boldsymbol{z}),$$

对 $\mathscr{L}(\boldsymbol{\alpha},b,\boldsymbol{\xi},\boldsymbol{\varepsilon})$ 的各分量求导后可得

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\alpha}}=0\Rightarrow \boldsymbol{\alpha}=\boldsymbol{\xi},$$

$$\frac{\partial \mathscr{L}}{\partial b}=0\Rightarrow \boldsymbol{\xi}^{\mathrm{T}}\mathbf{1}_n=0,$$

$$\frac{\partial \mathscr{L}}{\partial \varepsilon}=0\Rightarrow \boldsymbol{\xi}=\frac{1}{\lambda}\boldsymbol{W}\boldsymbol{\varepsilon},$$

$$\frac{\partial \mathscr{L}}{\partial \xi_i}=0\Rightarrow \boldsymbol{K}\boldsymbol{\alpha}+(b-\ln\gamma)\mathbf{1}_n+\boldsymbol{\varepsilon}=\boldsymbol{z}.$$

简化上式得

$$\boldsymbol{K}\boldsymbol{\alpha}+(b-\ln\gamma)\mathbf{1}_n+\lambda\boldsymbol{W}^{-1}\boldsymbol{\alpha}=\boldsymbol{z},$$

$$\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\alpha}=0.$$

写成矩阵形式,有

$$\begin{pmatrix}\boldsymbol{K}+\lambda\boldsymbol{W}^{-1} & \mathbf{1}_n\\ \mathbf{1}_n^{\mathrm{T}} & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ b-\ln\gamma\end{pmatrix}=\begin{pmatrix}\boldsymbol{z}\\ 0\end{pmatrix}\tag{5}$$

且

$$\boldsymbol{z}=\boldsymbol{K}\boldsymbol{\alpha}+(b-\ln\gamma)\mathbf{1}_n+\boldsymbol{W}^{-1}(\boldsymbol{Y}-\boldsymbol{\pi})\tag{6}$$

IRLS 方法迭代过程中,由式(5)更新 $\boldsymbol{\alpha}$ 和 b,再由式(6)更新 $\boldsymbol{z}$,并迭代计算到收敛停止,然而对于大规模数据集,每次迭代都需要对一个 $n\times n$ 矩阵求逆,即每次都需要进行一次复杂度为 $O(n^3)$ 的计算,当迭代次数过多甚至不收敛时,IRLS 方法训练核逻辑回归非常低效[27]. 这种迭代计算是由权重矩阵 $\boldsymbol{W}$ 中的非线性项依赖造成的,$\boldsymbol{W}$ 通过后验概率 $\pi=1/(1+\exp(-(\boldsymbol{K}\boldsymbol{\alpha}+b\mathbf{1}_n)))$ 非线性地依赖于参数 $\boldsymbol{\alpha}$ 和 b. 接下来提出的方法将消除这种非线性依赖并加快训练过程. 本文将 Ngufor 和 Wojtusiak[24] 提出的 LS 近似方法推广至广义逻辑函数情形,进而利用其对大规模的非均衡数据进行分类器构造. 对逻辑函数 $f(z)=\dfrac{1}{1+\gamma\mathrm{e}^{-z}}$ 在 $z=0$ 附近泰勒展开,取其第一项作为近似值,即 $f(0)=\tau$,使得权重矩阵简化为单位矩阵乘以常数的形式,如 $\boldsymbol{W}=\dfrac{1}{\tau(1-\tau)}\boldsymbol{I}_n$,将式(5)变为非迭代的求解线性方程组问题:

$$\begin{pmatrix}\boldsymbol{K}+\dfrac{\lambda}{\tau(1-\tau)}\boldsymbol{I}_n & \mathbf{1}_n\\ \mathbf{1}_n^{\mathrm{T}} & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ b-\ln\gamma\end{pmatrix}=\begin{pmatrix}\dfrac{1}{\tau(1-\tau)}(\boldsymbol{Y}-\tau\mathbf{1}_n)-\ln(\gamma)\,\mathbf{1}_n\\ 0\end{pmatrix}\tag{7}$$

相比 IRLS 方法,该方法只考虑处于极端情况下的分类边界 $f(\boldsymbol{x})=0$ 的点,它们被标记为类别 1 的概率为 τ,而这些点在极端意义下比其他远离分类边界的点对确定分类边界更加重要,因而训练核逻辑回归可以限定为极端意义下对边界点的学习. 这是本文的一个重要技巧,本文的研究显示出此种技巧是有效的,具有对非均衡数据进行分类的良好效果. 只要求解一次线性方程组即可获得相应的参数估计,有效解决迭代方法中可能存在的不收敛问题,因而大大提高求解核逻辑回归的效率.

注 1.1 对 τ 的估计可以通过位于边界附近标记为类别 1 的点所占比例来估计,即 $\hat{\tau}=\sum_{i=1}^{n}y_i\mathbf{1}_{\{|f(\boldsymbol{x}_i)|<d\}}/\sum_{i=1}^{n}1_{\{|f(\boldsymbol{x}_i)|<d\}}$. 为简便,本文用整体训练样本中标记为类别 1 的点所占比例来估计,即 $\hat{\tau}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$. 该方法实现见算法 1.1.

算法 1.1 LS-RKLR

输入:$\mathscr{D}=\{(\boldsymbol{x}_i,y_i)\mid\boldsymbol{x}_i\in\mathbb{R}^d,y_i\in\{0,1\},i=1,\cdots,n\}$,核函数 $K(\cdot,\cdot)$,正则化参数 λ;

输出:参数:$(\boldsymbol{\alpha},b)$;

①计算核函数矩阵 $\boldsymbol{K}_{n\times n}=(K(\boldsymbol{x}_i,\boldsymbol{x}_{i'}))_{i,i'=1}^{n}$;

②计算 $\bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$,并令 $\hat{\tau}=\bar{y}$;

③根据式(7)求解$(\boldsymbol{\alpha},b)$.

1.2 方法的性质

本文给出了基于修改逻辑函数后的核逻辑回归方法,可以将其看作有偏移的核逻辑回归,在 $y\in\{1,-1\}$ 下,其表示为

$$\min_f\sum_{i=1}^{n}\ln(1+\mathrm{e}^{-y_if(\boldsymbol{x}_i)})$$

$$\text{s.t.}\quad \|f\|_{\mathscr{H}_K}\leqslant s,$$

$$f(\boldsymbol{x}_i)=\sum_{k=1}^{n}K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k+B_0(\tau),\ i=1,\cdots,n\tag{8}$$

式中,$B_0(\tau)$ 为偏移项,且

$$B_0(\tau)\begin{cases}\geqslant 0, & \text{当 }\tau\geqslant 0.5;\\ <0, & \text{当 }\tau<0.5.\end{cases}$$

Zhu 和 Hastie[18] 给出无偏移的核逻辑回归的性质,那么对于加入偏移项之后的核逻辑回归,我们也能得到类似的性质,使其适应于非均衡数据的情形. 假设训练数据是可分的,那么偏移的支持向量机模型可表示为

$$\max_f D$$
$$\text{s.t. } y_i f(\boldsymbol{x}_i) \geqslant D(1-\xi_i),\ i=1,\cdots,n,$$
$$f(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k + B_0(\tau),$$
$$\xi_i \geqslant 0,\ \sum_{i=1}^{n}\xi_i \leqslant \lambda,\ \| f \|_{\mathcal{H}_K}^2 = 1 \qquad (9)$$

式中，D 为训练数据离分类超平面的最近距离.

接下来，本文进一步研究上述偏移的核逻辑回归与偏移的支持向量机之间的关系. 首先关注以下两个引理，而为简化记号和证明，在引理及其证明中令

$$G(f)=\sum_{i=1}^{n}\ln(1+\mathrm{e}^{-y_i f(\boldsymbol{x}_i)}).$$

引理 1.1　若数据是可分的，即存在超平面 f 使得对 $\forall i>0$ 有 $y_i f(\boldsymbol{x}_i)>0$. 那么对模型(8)的解 $\hat{f}_s$ 也有 $y_i \hat{f}_s(\boldsymbol{x}_i)>0$，$\forall i>0$，且对 $s>s_0>0$，$\| \hat{f}_s \|_{\mathcal{H}_K}=s$，其中，

$$f(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k + B_0(\tau),$$
$$\hat{f}_s(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\hat{\alpha}_{sk} + B_0(\tau).$$

引理 1.2　存在由 d_1, d_2 构成的数 $s_0=S(d_1,d_2)$，使得对 $\forall s>s_0$，有 $G(sf_1)<G(sf_2)$，其中 S 为某一函数形式，在数据可分的条件下

$$d_1 := \min_i y_i f_1(\boldsymbol{x}_i) > d_2 := \min_i y_i f_2(\boldsymbol{x}_i),$$

且

$$\| f_1 \|_{\mathcal{H}_K} = \| f_2 \|_{\mathcal{H}_K} = 1.$$

引理 1.1 和 1.2 的证明由附录给出. 由上述引理 1.1 和引理 1.2，我们可得如下定理.

定理 1.1　若数据是可分的，即存在超平面 f 使得对 $\forall i>0$ 有 $y_i f(\boldsymbol{x}_i)>0$，那么对模型(8)求得的分类超平面 $\hat{f}_s$，当 $s\to\infty$ 有

$$\frac{\hat{f}_s}{s} \to f^*.$$

式中，$f(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k + B_0(\tau)$，$\hat{f}_s(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\hat{\alpha}_{sk} + B_0(\tau)$. 若 f^* 是唯一的，则 f^* 为模型(9) 中 $\lambda\to 0$ 所求得的分类超平面. 反之，若 f^* 不是唯一的，即 $\frac{\hat{f}_s}{s}$ 有多个极值点，那么这些极值点都能形成最大间隔分类超平面.

证明　假设 f^* 是 $\frac{\hat{f}_s}{s}$ 的一个极值点，并记

$$d^* := \min_i y_i\Big(\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k^* + B_0(\tau)\Big),$$

那么由引理 1.1 可知 $d^*>0$. 现假设存在 $\tilde{f}(\boldsymbol{x})=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x})\tilde{\alpha}_k + B_0(\tau)$ 满足 $\| \tilde{f} \|_{\mathcal{H}_K}=1$ 且有更大的分类间隔 $\tilde{d}>d^*$，那么考虑存在 $\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)^{\mathrm{T}}$ 的邻域 $N_{\alpha^*}=\{\alpha:\| \alpha-\alpha^* \|_2<\delta\}$ 和 $\epsilon>0$，使得

$$\min_i y_i\Big(\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k + B_0(\tau)\Big) < \tilde{d}-\epsilon,\ \forall \alpha \in N_{\alpha^*}.$$

那么由引理 1.2 可知，存在 $s_0=S(\tilde{d},\tilde{d}-\epsilon)$ 使得对 $s>s_0$ 和 $\alpha\in N_{\alpha^*}$ 有 $G(s\tilde{f})<G(sf)$，由此可知 f^* 不应是 $\frac{\hat{f}_s}{s}$ 的极值点，矛盾. 所以，我们可知 $\frac{\hat{f}_s}{s}$ 的极值点为最大间隔分类超平面，如果该最大间隔分类超平面是唯一的，则

$$\lim_{s\to\infty}\frac{\hat{f}_s}{s}=\arg\max_{f,\| f \|_{\mathcal{H}_K}=1}\min_i y_i f(\boldsymbol{x}_i).$$

其中，$f(\boldsymbol{x}_i)=\sum_{k=1}^{n} K(\boldsymbol{x}_k,\boldsymbol{x}_i)\alpha_k + B_0(\tau)$，上式所得 f 即为模型(9)中 $\lambda\to 0$ 所求得的分类超平面，证毕.

需要指出的是，文献[19]已经证明了核逻辑回归泛化误差期望的收敛性，并给出了该泛化误差收敛速度，而定理 1.1 中的收敛不同于样本量趋于无穷大时参数的收敛，而是在固定样本量下，分类超平面随着正则化参数 s 无穷大时收敛到某最大间隔分类超平面. 且当该最大间隔分类超平面唯一时，由定理 1.1 可知，模型(9)和(8)是等价的. 当数据可分时，类似于偏移的支持向量机，偏移的核逻辑回归方法可以看作最大间隔分类超平面的偏移. 对于非均衡数据集，该偏移可以使得分类超平面朝远离少数类样本的方向移动，从而提高少数类样本的分类正确率.

2　实验研究

2.1　模拟研究

2.1.1　模拟数据

二分类模拟数据生成方式[20]：首先，由二元正态分布 $N((0,0)^{\mathrm{T}},3\boldsymbol{I})$ 产生 10 个均值 μ_k 并标记为类别为 1；类似的，由 $N((1,1)^{\mathrm{T}},3\boldsymbol{I})$ 产生 10 个 μ_k，并标记为类别为 0. 其次，对于每个类，以 1/10 的概

率随机选出 μ_k，并由分布 $N(\mu_k, \boldsymbol{I})$ 生成该类别的样本. 模拟样本数为 $n=1\ 500$，其中类别为 1 的样本占比分别为 $r=20\%, 10\%, 5\%, 2.5\%, 2\%, 1.5\%, 1\%$，共产生 7 个模拟数据集.

2.1.2 模型表现

本研究测试 LS-RKLR 在非均衡数据的分类能力. 作为对比的模型有经典的迭代加权最小二乘 IRLS-KLR，没有利用基准类别信息的 LS-KLR，结合随机欠抽样方法的 LS-KLR (under)，结合 SMOTE 方法的 LS-KLR(SMOTE)，不做抽样处理的随机森林(random forest)，不做抽样处理的 SVM. 随机欠抽样中，调整的参数为多数类抽样的比例；SMOTE 方法的调整参数为少数类过抽样后的与原少数类的比例. 随机森林调整参数为树的数目以及每次生成子节点时选择的特征个数. 为方便比较，模型中的核函数都设置成高斯径向基核 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)=e^{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}}$. 模型的评价标准选择 ROC 曲线的面积，即 AUC 值. 实验将 70% 的样本作为训练集，30% 作为测试集，采用 5 折交叉验证(five cross validation)进行参数选择. 实验进行 100 次，获得 AUC 的平均值及标准差. 实验结果如表 1 及图 1 所示.

表 1 模拟数据测试集 AUC 值及模型训练时间

Tab. 1 AUC value of simulation test dataset and model training time

少数类占比 r/%	LS-RKLR	IRLS-KLR	LS-KLR	LS-KLR (under)	LS-KLR (SMOTE)	random forest	SVM
20	0.787(0.005)	**0.787(0.005)**	0.781(0.005)	0.779(0.005)	0.783(0.005)	0.726(0.005)	0.719(0.007)
10	0.774(0.008)	**0.782(0.008)**	0.760(0.007)	0.767(0.008)	0.774(0.008)	0.701(0.008)	0.690(0.014)
5	0.764(0.012)	**0.768(0.009)**	0.721(0.01)	0.739(0.009)	0.752(0.009)	0.676(0.009)	0.667(0.015)
2.5	**0.745(0.019)**	0.737(0.015)	0.676(0.016)	0.722(0.013)	0.719(0.015)	0.641(0.013)	0.648(0.016)
2	**0.734(0.020)**	0.728(0.017)	0.652(0.019)	0.708(0.017)	0.702(0.017)	0.632(0.017)	0.631(0.016)
1.5	**0.729(0.020)**	0.708(0.019)	0.639(0.019)	0.666(0.019)	0.693(0.015)	0.608(0.019)	0.624(0.019)
1	**0.718(0.023)**	0.691(0.021)	0.616(0.017)	0.652(0.018)	0.671(0.023)	0.608(0.018)	0.623(0.021)
训练时间/min	3.2	138.3	3.2	2.7	4.6	2.8	4.5

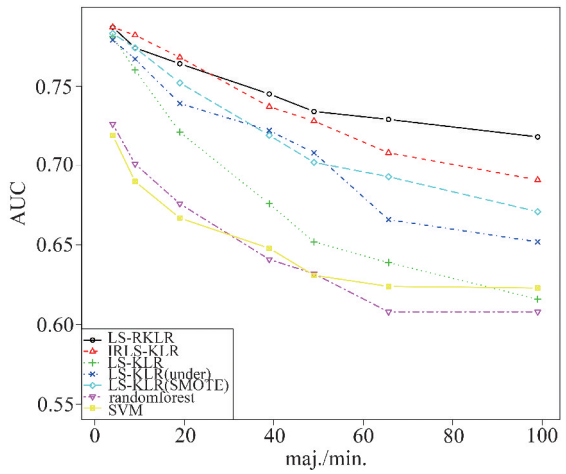

图 1 模拟数据中不同的多数类/少数类比情况下对应的测试集 AUC 值

Fig. 1 AUC value of test set corresponding to different ratio of majority class to minority class of simulated data

实验结果显示，① 在非均衡数据情况下，LS-RKLR 的分类效果相比 LS-KLR 有明显提升，且当不均衡程度增加时，提升的效果越明显. ② LS-RKLR 方法在少数类占比 $r=20\%, 10\%, 5\%$ 时，测试集 AUC 值接近原有的 IRLS-KLR 方法，但在 $r=2.5\%, 2\%, 1.5\%, 1\%$ 时超过 IRLS-KLR 方法，说明 LS-RKLR 通过在边界处用基准概率近似的方式对于严重不均衡的数据具有良好的分类效果. 另外，综合考虑到 LS-RKLR 方法在运行时间远远小于迭代运算的 IRLS-KLR 方法，故 LS-RKLR 方法对于较大规模非均衡数据具有更加良好的潜力. ③ LS-RKLR 在测试集 AUC 值高于 LS-KLR 上做欠抽样以及 SMOTE 过抽样的 AUC 值，说明 LS-RKLR 方法在没有做抽样处理的前提下也能表现出良好的分类效果.

2.2 实证分析

真实数据集包括 5 个经常被用于非均衡分类问题的 UCI 公共数据[28-29]，分别为 yeast2, yeast3, abalone19, abalone9vs18, segment. 数据中少数类样本标记为 1，多数类样本标记为 0，所有数据的输入变量都经过标准化处理. 数据具体信息如表 2 所示.

表 2　真实数据集

Tab. 2　Real dataset

数据集	样本数	输入特征数	类别（少，多）	（少，多）占比/%
yeast2	1 484	8	（ME2，其他）	（3.4，96.6）
yeast3	1 484	8	（ME3，其他）	（10.98，89.02）
abalone19	4 174	8	（19，其他）	（0.77，99.23）
abalone9vs18	731	8	（18，9）	（5.65，94.25）
segment	2 308	19	（sky，其他）	（14.26，85.74）

实验每次随机抽取 70% 的样本作为训练集，30% 作为测试集，采用 5 折交叉验证法进行参数选择. 实验进行 100 次. 表 3 给出测试集 AUC 的平均值和标准差，表 4 为对应的模型训练时间. 由表 3 可知，5 个数据集中，IRLS-KLR 在数据集 yeast2，yeast3，abalone19 中测试集 AUC 值最高，LS-RKLR 在数据集 abalone9vs18 和 segment 中测试集 AUC 值最高. 然而，由表 4 给出的模型训练时间可以看出，本文给出的 LS-RKLR 在训练模型时需要的时间相比于 IRLS-KLR 大大减少，表现出对于较大规模数据优良的训练效果.

表 3　真实数据测试集 AUC 值

Tab. 3　AUC value of real test dataset

数据集	LS-RKLR	IRLS-KLR	LS-KLR	LS-KLR(under)	LS-KLR(SMOTE)
yeast2	0.940(0.005)	**0.955(0.005)**	0.936(0.009)	0.947(0.005)	0.944(0.006)
yeast3	0.981(0.001)	**0.982(0.001)**	0.977(0.002)	0.980(0.001)	0.981(0.001)
abalone19	0.730(0.015)	**0.776(0.019)**	0.692(0.027)	0.730(0.008)	0.731(0.013)
abalone9vs18	**0.958(0.003)**	0.950(0.007)	0.958(0.006)	0.944(0.006)	0.953(0.005)
segment	**0.994(0.001)**	0.994(0.001)	0.993(0.001)	0.994(0.001)	0.994(0.001)

表 4　真实数据集训练时间（单位：min）

Tab. 4　Training time of real dataset (unit: min)

数据集	LS-RKLR	IRLS-KLR	LS-KLR	LS-KLR (under)	LS-KLR (SMOTE)
yeast2	2.9	113.1	2.9	2.3	4.6
yeast3	2.9	113.1	2.9	2.3	4.6
abalone19	50.2	2 364.4	50.2	37.7	111.7
abalone9vs18	0.8	136.6	0.8	0.7	1.4
segment	13.3	591.1	13.3	10.6	17.3

3　结论

本文在已有的核逻辑回归方法基础上，针对非均衡数据情形提出一种新的非迭代的稳健核逻辑回归方法 LS-RKLR. 该方法引入广义逻辑函数，构造出一个偏移的核逻辑回归，在数据可分条件下，能够形成最大间隔分类超平面. 同时，通过在分类边界处进行泰勒展开，将核逻辑回归限定在对更富有信息量的分类边界处样本的学习，使原先的迭代加权最小二乘问题简化为非迭代的求解线性方程组问题. 实验结果表明，与其他抽样方法比，LS-RKLR 方法在对非均衡数据表现出良好的分类效果，同时较经典的 IRLS 方法大大提升了核逻辑回归的训练效率.

本文采用全部训练集中少数类占比来估计基准概率，虽然操作简单，但可能会损失部分精确度，故后续可以针对基准概率的估计进行深入研究. 其次，本研究只采用单个高斯径向基核作为核函数，而核函数的选取是核方法中一类重要问题，故本研究后续将对 LS-RKLR 方法中核函数的选取尤其是多核学习进行深入研究. 最后，虽然本文提出的非迭代方法相对于迭代方法能大大提高速度，但由于本文仍采用对原始的核矩阵求逆方法，其计算复杂度仍为 $O(n^3)$，后续研究可以将该非迭代的方法引入到前文所提到的输入向量机以及 NTR-KLR 方法中，进一步降低运算的复杂度.

参考文献（References）

[1] RODRIGUEZ D, HERRAIZ I, HARRISON R, et al. Preliminary comparison of techniques for dealing with imbalance in software defect prediction [C]// Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM, 2014: 43.

[2] MAZUROWSKI M A, HABAS P A, ZURADA J M, et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance [J]. Neural Networks, 2008, 21(2-3): 427-436.

[3] OLSZEWSKI D. A probabilistic approach to fraud detection in telecommunications[J]. Knowledge-Based Systems, 2012, 26: 246-258.

[4] ABDOU H A, POINTON J. Credit scoring, statistical techniques and evaluation criteria: A review of the literature[J]. Intelligent Systems in Accounting Finance & Management, 2011, 18(2-3):59-88.

[5] MAALOUF M, TRAFALIS T B. Robust weighted kernel logistic regression inimbalanced and rare events data[J]. Computational Statistics & Data Analysis, 2011, 55(1): 168-183.

[6] MAALOUF M, HOMOUZ D, TRAFALIS T B. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods [J]. Computational Intelligence, 2018, 34(4):161-174.

[7] LÓPEZ V, FERNSNDEZ A, GARCÍA S, et al. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics [J]. Information Sciences, 2013, 250: 113-141.

[8] ALI A, SHAMSUDDIN S M, RALESCU A L. Classification with classimbalance problem: A review [J]. Int J Advance Soft Compu Appl, 2015, 7(3): 176-204.

[9] YEN S J, LEE Y S. Cluster-based under-sampling approaches for imbalanced data distributions [J]. Expert Systems with Applications, 2009, 36(3): 5718-5727.

[10] KERDPRASOP N, KERDPRASOP K. On the generation of accurate predictive model from highly imbalanced data with heuristics and replication techniques[J]. International Journal of Bio-Science and Bio-Technology, 2012, 4(1): 49-64.

[11] HASANIN T, KHOSHGOFTAAR T. The effects of random undersampling with simulated class imbalance for big data[C]// 2018 IEEE International Conference on Information Reuse and Integration for Data Science. IEEE, 2018:70-79.

[12] LIN C T, HSIEH T Y, LIU Y T, et al. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(5): 950-962.

[13] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[14] NGUYEN G H, BOUZERDOUM A, PHUNG S L. Learning pattern classification tasks with imbalanced data sets[C]// Pattern Recognition. London: Intech Open, 2009.

[15] UYAR A, BENER A, CIRACY H N, et al. Handling the imbalance problem of IVF implantation prediction [J]. IAENG International Journal of Computer Science, 2006, 37(2): 164-170.

[16] YANG C Y, YANG J S, WANG J J. Margin calibration in SVM class-imbalanced learning [J]. Neurocomputing, 2009, 73(1-3): 397-411.

[17] WEST D. Neural network credit scoring models[J]. Computers & Operations Research, 2000, 27(11): 1131-1152.

[18] ZHU J, HASTIE T. Kernel logistic regression and the import vector machine[J]. Journal of Computational and Graphical Statistics, 2005, 14(1): 185-205.

[19] PARK C. Convergence rates of generalization errors for margin-based classification [J]. Journal of Statistical Planning and Inference, 2009, 139(8): 2543-2551.

[20] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The Elements of Statistical Learning [M]. New York: Springer, 2001.

[21] KARSMAKERS P, PELCKMANS K, SUYKENS J A K. Multiclass kernel logistic regression: A fixed-size implementation [C]// 2007 International Joint Conference on Neural Networks. IEEE, 2007: 1756-1761.

[22] MAALOUF M, TRAFALIS T B, ADRIANTO I. Kernel logistic regression using truncated Newton method [J]. Computational Management Science, 2010, 8(4): 415-428.

[23] ELBASHIR M K, WANG J. Kernel logistic regression algorithm for large-scale data classification [J]. Int Arab J Inf Technol, 2015, 12(5): 465-472.

[24] NGUFOR C, WOJTUSIAK J. Extreme logistic regression [J]. Advances in Data Analysis and Classification, 2016, 10(1): 27-52.

[25] BALAKRISHNAN N. Handbook of the Logistic Distribution[M]. Boca Raton,FL: CRC Press, 1991.

[26] SCHÖLKOPF B, HERBRICH R, SMOLA A J. A generalized representer theorem[C]// Computational Learning Theory. Amsterdam: Springer, 2001: 416-426.

[27] RAMANI S, FESSLER J A. An accelerated iterative reweighted least squares algorithm for compressed sensing MRI [C]// 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, 2010: 257-260.

[28] ASUNCION A, NEWMAN D. UCI machine learning repository[R]. Irvine, CA: University of California,

Irvine, School of Information and Computer Science, 2007.

[29] SHAO Y H, CHEN W J, ZHANG J J, et al. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification[J]. Pattern Recognition, 2014, 47(9): 3158-3167.

附录

引理 1.1 的证明　现设存在 i^*，使得 $y_{i^*}f(\boldsymbol{x}_{i^*})\leqslant 0$，那么

$$G(f)=\sum_{i=1}^{n}\ln(1+\exp(-y_i f(\boldsymbol{x}_i)))\geqslant \ln(1+\exp(-y_{i^*}f(\boldsymbol{x}_{i^*})))\geqslant \ln 2\ .$$

再由引理中的条件可知存在 f^*，$\|f^*\|_{\mathcal{H}_K}=1$，使得 $y_i f^*(\boldsymbol{x}_i)>0$，$\forall i>0$. 那么可构造 s_0 为

$$s_0=\frac{-\ln(2^{1/n}-1)}{\min\limits_i y_i f^*(\boldsymbol{x}_i)}.$$

由此可得对 $s>s_0$，有

$$G(sf^*)=\sum_{i=1}^{n}\ln(1+\exp(-y_i f^*(\boldsymbol{x}_i)s))<\sum_{i=1}^{n}\ln(1+\exp(-y_i f^*(\boldsymbol{x}_i)s_0))<$$

$$\sum_{i=1}^{n}\ln\left(1+\exp(-y_i f^*\frac{-\ln(2^{1/n}-1)}{\min\limits_i y_i f^*})\right)<\sum_{i=1}^{n}\ln(1+\exp(\ln(2^{1/n}-1)))\leqslant\sum_{i=1}^{n}\frac{\ln n}{2}=\ln 2.$$

那么由此可得 $G(\hat{f}_s)\leqslant G(sf^*)<G(f)$，则对 $s>s_0$，有 $y_i\hat{f}_s(\boldsymbol{x}_i)>0$，$\forall i>0$. 现考虑 $s>s_0$ 时，若 $\|\hat{f}_s\|_{\mathcal{H}_K}<s$，我们考虑如下 $\hat{f}_s'$，

$$\hat{f}_s'=\frac{\hat{f}_s'}{\|\hat{f}_s\|_{\mathcal{H}_K}}s.$$

那么由此可得 $\|\hat{f}_s'\|_{\mathcal{H}_K}=s$，且

$$G(\hat{f}_s')=\sum_{i=1}^{n}\ln(1+\exp(-y_i\hat{f}_s'(\boldsymbol{x}_i)))<\sum_{i=1}^{n}\ln(1+\exp(-y_i\hat{f}_s(\boldsymbol{x}_i)))=G(\hat{f}_s).$$

而 $G(\hat{f}_s)$ 应是最小值，矛盾，所以 $\|\hat{f}_s\|_{\mathcal{H}_K}=s$.

引理 1.2 的证明　令

$$s_0=S(d_1,\ d_2)=\frac{\ln n+\ln 2}{d_1-d_2},$$

那么对 $\forall s>s_0$ 有

$$G(sf_1)=\sum_{i=1}^{n}\ln(1+\exp(-y_i f_1(\boldsymbol{x}_i)s))\leqslant n\ln(1+\exp(-sd_1))\leqslant n\exp(-sd_1)<\frac{1}{2}\exp(-sd_2)\leqslant$$

$$\ln(1+\exp(-sd_2))<\sum_{i=1}^{n}\ln(1+\exp(-y_i f_2(\boldsymbol{x}_i)s))=G(sf_2).$$