doi: 10.12068/j.issn.1005-3026.2020.01.014

一种基于 Newton 迭代法的累积 Logistic 回归模型参数估计

印明昂, 王钰烁, 孙志礼, 郭 兵
(东北大学 机械工程与自动化学院, 辽宁 沈阳 110819)

摘 要: 基于 Newton 迭代法提出一种累积 Logistic 回归模型的参数估计方法. 分析了迭代初值选取、常系数大小关系以及迭代过程 Hessian 矩阵奇异等影响算法收敛性的主要问题. 通过自适应地选取迭代初值和控制迭代过程,避免了 Hessian 矩阵奇异的情况. 利用美国凯斯西储大学轴承数据库(CWRU)数据进行验证,实验结果表明,本文方法在模型训练和验证的准确率上均高于统计学软件 SPSS. 并利用 Booststrap 随机试验验证了所提算法的稳健性.

关 键 词: 累积 Logistic 回归;参数估计;Hessian 矩阵;自适应算法;滚动轴承;健康状态评估
中图分类号: TP 312 **文献标志码**: A **文章编号**: 1005-3026(2020)01-0079-05

Parameter Estimation of the Cumulative Logistic Regression Model Based on the Newton Iteration Method

YIN Ming-ang, *WANG Yu-shuo*, *SUN Zhi-li*, *GUO Bing*
(School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China. Corresponding author: YIN Ming-ang, E-mail: yinma@mail.neu.edu.cn)

Abstract: Based on the Newton iteration method, a parameter estimation method of cumulative Logistic regression was proposed. The selection of primary values of iteration, the relationship of sizes of constant coefficients and the singularity of Hessian matrix were analyzed, which are the main problems influencing stypticity in the process of iteration. Through the self-adaptive selection of primary values of iteration and control of the iteration process, the singularity of Hessian matrix was avoided. The data of the bearing database of Case Western Reserve University was used for validation. The experiment results show that the accuracy of model training and verification proposed is higher than that of the statistics software SPSS. Moreover, the robustness of the proposed algorithm is proved by the Booststrap random testing method.

Key words: cumulative Logistic regression; parameter estimation; Hessian matrix; adaptive algorithm; rolling bearing; health state assessment

累积 Logistic 回归又称为有序多分类 Logistic 回归、次序 Logistic 回归[1],其具有模型变量少、分类效率高的特点. 近年来,该方法在医药、金融及制造等领域得到了广泛的应用. 文献[2]选取 2015 年 3 月至 2017 年 4 月接受培非格司亭(Pegfilgrastim)联合化疗的 166 例癌症案例,采用多变量有序 Logistic 回归分析,确定符合 Pegfilgrastim 预防以维持相对剂量强度的预测因素. 文献[3]基于发动机健康变化的 Logistic 回归模型,实现了贝叶斯状态估计的健康状态序列更新,进而预测引擎未来的健康变化. 文献[4]通过审查 2003 至 2011 年外来资本流入的综合影响,检验了次序 Logistic 回归模型应用于基于柯布-道格拉斯生产函数的实证经济增长研究的可行性.

在应用领域日益扩展,数据数量和维度不断

收稿日期: 2018-12-26
基金项目: 国家自然科学基金资助项目(51775097,51875095).
作者简介: 印明昂(1985-),男,辽宁沈阳人,东北大学讲师,博士; 孙志礼(1957-),男,山东巨野人,东北大学教授,博士生导师.

增大的现状下,目前,累积 Logistic 回归模型的参数估计通常采用统计学软件 SPSS 或 SAS 的自带功能实现. 本文基于 Newton 迭代法提出一种累积 Logistic 回归模型的参数估计方法,对其中比较敏感的迭代初值选取及迭代过程 Hessian 矩阵奇异问题进行了自适应控制,最后利用美国凯斯西储大学轴承数据库(CWRU)[5]数据对该算法模型与 SPSS 模型进行了对比验证.

1 参数估计

1.1 迭代格式的构建

设($\boldsymbol{x}$, $\boldsymbol{y}$)为有序 J 分类样本,其中输入观测值 $\boldsymbol{x}$ 为 $n\times M$ 矩阵,n 代表容量,M 代表维度;有序输出观测值 $\boldsymbol{y}$ 为 $n\times 1$ 列向量,其元素 $y_i, i\in\{1,2,\cdots,n\}$, y_i 可取值为 $1,2,\cdots, J$. 则第 i 个输入 $\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{iM}]$ 的前 $j(j\in\{1,2,\cdots,J\})$ 类累积 Logistic 表达式为[6]

$$\ln\left(\frac{p(y_i\leqslant j|\boldsymbol{x}_i)}{1-p(y_i\leqslant j|\boldsymbol{x}_i)}\right)=\alpha_j+\sum_{m=1}^{M}\beta_m x_{im}. \tag{1}$$

其中:$p(y_i\leqslant j|\boldsymbol{x}_i)$ 为累积概率(简记为 p_{ij}),且有 $p(y_i\leqslant J|\boldsymbol{x}_i)=1$;$\alpha_j$ 为截距,对于累积 Logistic 模型共有 $J-1$ 个;β_m 为回归系数,所有分类的系数相同. 由此可得

$$p_{ij}=\frac{1}{1+\exp\left(\alpha_j+\sum_{m=1}^{M}\beta_m x_{im}\right)}. \tag{2}$$

则该输入属于第 j 类的概率为

$$\left.\begin{aligned}&Q_{ij}=p_{ij}-p_{ij-1}, j=1,2,\cdots,J-1;\\&Q_{iJ}=1-p_{iJ-1}.\end{aligned}\right\} \tag{3}$$

最后由概率最大者确定其所属类别. 不妨设 $\boldsymbol{Y}$ 为 $n\times J$ 矩阵,其元素 y_{ij} 为

$$y_{ij}=\begin{cases}1,\text{第 } i \text{ 个输入属于第 } j \text{ 类;}\\0,\text{第 } i \text{ 个输入不属于第 } j \text{ 类}.\end{cases} \tag{4}$$

则它的对数似然函数可表示为

$$\ln[L(\theta)]=\sum_{i=1}^{n}\sum_{j=1}^{J}y_{ij}\ln(p_{ij}-p_{ij-1})=\sum_{i=1}^{n}\sum_{j=1}^{J}y_{ij}\ln\left[\frac{1}{1+\exp(\alpha_j+\sum_{m=1}^{M}\beta_m x_{im})}-\frac{1}{1+\exp(\alpha_{j-1}+\sum_{m=1}^{M}\beta_m x_{im})}\right]. \tag{5}$$

由 Newton 迭代公式得到迭代格式:

$$\boldsymbol{b}_{k+1}=\boldsymbol{b}_k-\boldsymbol{H}(\boldsymbol{b}_k)^{-1}\boldsymbol{f}(\boldsymbol{b}_k). \tag{6}$$

其中,$\boldsymbol{b}_k=[\boldsymbol{\alpha}_k^{\mathrm{T}},\boldsymbol{\beta}_k^{\mathrm{T}}]^{\mathrm{T}}=[\alpha_{1k},\cdots,\alpha_{(J-1)k},\beta_{1k},\cdots,\beta_{Mk}]^{\mathrm{T}}$ 为未知参数的第 k 次迭代值,$k=0,1,2,\cdots$;$\boldsymbol{H}(\boldsymbol{b}_k)$,$\boldsymbol{f}(\boldsymbol{b}_k)$ 分别为

$$\boldsymbol{H}(\boldsymbol{b}_k)=\begin{bmatrix}\frac{\partial^2\ln L}{\partial\alpha_1^2}&\frac{\partial^2\ln L}{\partial\alpha_1\partial\alpha_2}&\cdots&\frac{\partial^2\ln L}{\partial\alpha_1\partial\beta_M}\\\frac{\partial^2\ln L}{\partial\alpha_2\partial\alpha_1}&\frac{\partial^2\ln L}{\partial\alpha_2^2}&\cdots&\frac{\partial^2\ln L}{\partial\alpha_2\partial\beta_M}\\\vdots&\vdots&&\vdots\\\frac{\partial^2\ln L}{\partial\beta_M\partial\alpha_1}&\frac{\partial^2\ln L}{\partial\beta_M\partial\alpha_2}&\cdots&\frac{\partial^2\ln L}{\partial\beta_M^2}\end{bmatrix}, \tag{7}$$

$$\boldsymbol{f}(\boldsymbol{b}_k)=\left(\frac{\partial\ln L}{\partial\alpha_1}\quad\frac{\partial\ln L}{\partial\alpha_2}\quad\cdots\quad\frac{\partial\ln L}{\partial\beta_M}\right)^{\mathrm{T}}. \tag{8}$$

$\boldsymbol{H}(\boldsymbol{b}_k)$ 元素如下

$$\frac{\partial^2\ln L}{\partial\beta_m\partial\beta_s}=-\sum_{i=1}^{n}x_{im}x_{is}\sum_{j=1}^{J}y_{ij}[p_{ij}(1-p_{ij})+p_{ij-1}(1-p_{ij-1})]; \tag{9}$$

$$\frac{\partial^2\ln L}{\partial\beta_m\partial\alpha_j}=-\sum_{i=1}^{n}x_{im}(y_{ij}+y_{ij+1})p_{ij}(1-p_{ij}); \tag{10}$$

$$\frac{\partial^2\ln L}{\partial\alpha_r\partial\alpha_j}=\begin{cases}0, j\neq r, r+1, r-1;\\-\sum_{i=1}^{n}\frac{y_{il}p_{ir}p_{ij}(1-p_{ir})(1-p_{ij})}{(p_{ir}-p_{ij})^2},\\j=r-1, r+1, l=\max(j,r);\\-\sum_{i=1}^{n}p_{ir}(1-p_{ir})\left\{\frac{y_{ir}[p_{ir}^2+p_{ir-1}(1-2p_{ir})]}{(p_{ir}-p_{ij})^2}+\right.\\\left.\frac{y_{ir+1}[p_{ir}^2+p_{ir+1}(1-2p_{ir})]}{(p_{ir+1}-p_{ir})^2}\right\}, r=j.\end{cases} \tag{11}$$

其中,r, $j=1,2,\cdots,J-1$;s, $m=1,2,\cdots, M$.

1.2 迭代初值的选取

Newton 迭代法虽然具有平方收敛速度,但它对迭代初值非常敏感,如果选取不当往往会导致迭代发散. 下面通过数学推导,给出一种初值的选取方法.

首先将数据归一化,使所有输入数据统一映射到[0,1]区间上[7],即

$$x_{im}=\frac{x_{im}-\min(\boldsymbol{x}_m)}{\max(\boldsymbol{x}_m)-\min(\boldsymbol{x}_m)}. \tag{12}$$

其中:x_{im} 表示输入 i 第 m 列的数据;$\boldsymbol{x}_m$ 为所有第 m 列数据组成的向量,$i=1,2,\cdots,n$, $m=1,2,\cdots, M$;$\min(\cdot)$, $\max(\cdot)$ 分别表示向量中的最小、最大值.

根据 Fisher 判别思想,通常各个分类中心差距明显[8]. 以各类中的样本均值表示分类中心,如 j 类中心 $\bar{\boldsymbol{x}}_j=[\bar{x}_{j1},\bar{x}_{j2},\cdots,\bar{x}_{jM}]$ 为所有属于第 j 类输入的样本均值. 计算得出各类中心 $\bar{\boldsymbol{X}}=[\bar{\boldsymbol{x}}_1^{\mathrm{T}},$

$\bar{\boldsymbol{x}}_2^{\mathrm{T}},\cdots,\bar{\boldsymbol{x}}_J^{\mathrm{T}}]^{\mathrm{T}}$ 作为选取初值的特定数据.

设初值为 $\boldsymbol{b}_0=[\boldsymbol{\alpha}_0^{\mathrm{T}},\boldsymbol{\beta}_0^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{\alpha}_0=[\alpha_{10},\cdots,\alpha_{(J-1)0}]^{\mathrm{T}}$为常系数初值,$\boldsymbol{\beta}_0=[\beta_{10},\cdots,\beta_{M0}]^{\mathrm{T}}$ 为变量系数初值. 将 $\bar{\boldsymbol{X}}$ 和 $\boldsymbol{b}_0$ 代入式(2),式(3),得到以未知参数表示的各分类条件预测概率矩阵 $\boldsymbol{Q}$:

$$\boldsymbol{Q}=\begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1J} \\ Q_{21} & Q_{22} & \cdots & Q_{2J} \\ \vdots & \vdots & & \vdots \\ Q_{J1} & Q_{J2} & \cdots & Q_{JJ} \end{bmatrix}. \tag{13}$$

其中,$Q_{ij}=P(y=j|\bar{\boldsymbol{x}}_i,\boldsymbol{b}_0)$. 根据定义,预测条件概率 Q_{jj}应为所在行的最大值,显然,$\boldsymbol{Q}$ 中只有三主对角线上的元素显著不为零. 由此可得下列 J 个关于 $\boldsymbol{b}_0$ 中各元素的不等式:

$$\left.\begin{aligned} &\bar{\boldsymbol{x}}_1\boldsymbol{\beta}_0<-\alpha_{10},\\ &\ln\left(\frac{1}{p_1}-1\right)-\alpha_{10}<\bar{\boldsymbol{x}}_2\boldsymbol{\beta}_0<\ln\left(\frac{1}{p_2}-1\right)-\alpha_{20},\\ &\qquad\vdots\\ &\ln\left(\frac{1}{p_1}-1\right)-\alpha_{(J-2)0}<\bar{\boldsymbol{x}}_{J-1}\boldsymbol{\beta}_0<\ln\left(\frac{1}{p_2}-1\right)-\\ &\alpha_{(J-1)0},\\ &-\alpha_{(J-1)0}<\bar{\boldsymbol{x}}_J\boldsymbol{\beta}_0. \end{aligned}\right\} \tag{14}$$

为使 $\bar{\boldsymbol{x}}_j$ 属于第 j 类的概率最大,推荐取 $p_1=0.3$, $p_2=0.7$. 同时,为避免 $\boldsymbol{\beta}_0$ 各元素间差距过大,导致 Hessian 矩阵奇异,利用系数 $f_{\mathrm{d1}}>0$ 对其进行限定:

$$-f_{\mathrm{d1}}\leqslant\beta_{0i}\leqslant f_{\mathrm{d1}},\ i=1,2,\cdots,M. \tag{15}$$

由式(14)可推出:

$$\left.\begin{aligned} &\alpha_{10}-\alpha_{20}>c>0,\\ &\alpha_{20}-\alpha_{30}>c>0,\\ &\qquad\vdots\\ &\alpha_{(J-2)0}-\alpha_{(J-1)0}>c>0. \end{aligned}\right\} \tag{16}$$

其中,$c=\ln(1/p_1-1)-\ln(1/p_2-1)$. 进而有

$$\alpha_{10}>\cdots>\alpha_{(J-2)0}>\alpha_{(J-1)0}. \tag{17}$$

式(14)~式(17)给出了 $\boldsymbol{b}_0$ 的大致范围. 为进一步压缩区间,提高选取成功率,将 $\boldsymbol{b}_0$ 分为三部分:$\boldsymbol{b}_0=[\boldsymbol{\alpha}_0^{\mathrm{T}},\boldsymbol{\beta}_{01}^{\mathrm{T}},\boldsymbol{\beta}_{02}^{\mathrm{T}}]^{\mathrm{T}}$,其中 $\boldsymbol{\alpha}_0=[\alpha_{10},\cdots,\alpha_{(J-1)0}]^{\mathrm{T}}$,$\boldsymbol{\beta}_{01}=[\beta_{10},\cdots,\beta_{k0}]^{\mathrm{T}}$,$\boldsymbol{\beta}_{02}=[\beta_{(k+1)0},\cdots,\beta_{M0}]^{\mathrm{T}}$,$k\in[1,J]$. 将式(14)的第一和最后一式进行缩放:

$$\left.\begin{aligned} &\alpha_{10}<-\bar{\boldsymbol{x}}_1\boldsymbol{\beta}_0<\sum_{i=1}^{M}\bar{x}_{1i}\cdot f_{\mathrm{d1}}<M\cdot f_{\mathrm{d1}},\\ &\alpha_{(J-1)0}>-\bar{\boldsymbol{x}}_J\boldsymbol{\beta}_0>-\sum_{i=1}^{M}\bar{x}_{Ji}\cdot f_{\mathrm{d1}}>-M\cdot f_{\mathrm{d1}}. \end{aligned}\right\} \tag{18}$$

$$\left.\begin{aligned} &\alpha_{10}<-\bar{\boldsymbol{x}}_1\boldsymbol{\beta}_0=\\ &-((\bar{x}_{11},\bar{x}_{12},\cdots,\bar{x}_{1k})\boldsymbol{\beta}_{01}+(\bar{x}_{1(k+1)},\cdots,\bar{x}_{1M})\boldsymbol{\beta}_{02})=\\ &-(\bar{x}_{11},\bar{x}_{12},\cdots,\bar{x}_{1k})\boldsymbol{\beta}_{01}+f_{\mathrm{d2}}\sum_{i=k+1}^{M}\bar{x}_{1i},\\ &\alpha_{(J-1)0}>-\bar{\boldsymbol{x}}_J\boldsymbol{\beta}_0>\\ &-(\bar{x}_{J1},\bar{x}_{J2},\cdots,\bar{x}_{Jk})\boldsymbol{\beta}_{01}-f_{\mathrm{d2}}\sum_{i=k+1}^{M}\bar{x}_{Ji}. \end{aligned}\right\} \tag{19}$$

其中 $f_{\mathrm{d2}}=M\cdot f_{\mathrm{d1}}$. 设 $\boldsymbol{\alpha}_0$,$\boldsymbol{\beta}_{01}$已知,代入式(14)整理为关于 $\boldsymbol{\beta}_{02}$的不等式组:

$$\left.\begin{aligned} &\boldsymbol{T}_{21}\boldsymbol{\beta}_{02}<-\alpha_{10}-\boldsymbol{T}_{11}\boldsymbol{\beta}_{01},\\ &c_1-\alpha_{10}-\boldsymbol{T}_{12}\boldsymbol{\beta}_{01}<\boldsymbol{T}_{22}\boldsymbol{\beta}_{02}<c_2-\alpha_{20}-\boldsymbol{T}_{12}\boldsymbol{\beta}_{01},\\ &\qquad\vdots\\ &c_1-\alpha_{(J-2)0}-\boldsymbol{T}_{1(J-1)}\boldsymbol{\beta}_{01}<\boldsymbol{T}_{2(J-1)}\boldsymbol{\beta}_{02}<\\ &c_2-\alpha_{(J-1)0}-\boldsymbol{T}_{1(J-1)}\boldsymbol{\beta}_{01},\\ &-\alpha_{(J-1)0}-\boldsymbol{T}_{1J}\boldsymbol{\beta}_{01}<\boldsymbol{T}_{2J}\boldsymbol{\beta}_{02}. \end{aligned}\right\} \tag{20}$$

其中,$c_1=\ln(1/p_1-1)$, $c_2=\ln(1/p_2-1)$; $\boldsymbol{T}_{i1}=[\bar{x}_{i1},\bar{x}_{i2},\cdots,\bar{x}_{ik}]$, $\boldsymbol{T}_{i2}=[\bar{x}_{i(k+1)},\cdots,\bar{x}_{iM}]$, $i=1,2,\cdots,J$. 将式(15),式(20)作为边界条件,通过求解以下最优化问题得到 $\boldsymbol{\beta}_{02}$.

$$\min\sum_{i=k+1}^{M}\beta_{02,i}^2. \tag{21}$$

算法描述见算法 1.

算法 1　迭代初值的选取

输入 $\boldsymbol{x}$,f_{d1},k.

1) 数据 $\boldsymbol{x}$ 归一化;计算分类中心 $\bar{\boldsymbol{X}}$;

2) 区间$[-f_{\mathrm{d1}},f_{\mathrm{d1}}]$内随机生成 $\boldsymbol{\beta}_{01}$;

3) 以式(19)计算区间,在该区间内随机生成 $\boldsymbol{\alpha}_0$;

4) 将 $\boldsymbol{\alpha}_0$ 降序处理,判断式(16)是否成立. 若是,转步骤 5);若否,转步骤 2);

5) 以式(15),式(20)为约束,判断式(21)是否存在最优解. 若是,得到 $\boldsymbol{\beta}_{02}$,结束;若否,转步骤 2).

输出 $\boldsymbol{b}_0=[\boldsymbol{\alpha}_0^{\mathrm{T}},\boldsymbol{\beta}_{01}^{\mathrm{T}},\boldsymbol{\beta}_{02}^{\mathrm{T}}]^{\mathrm{T}}$.

通过此算法得到的 $\boldsymbol{b}_0$ 即可作为 Newton 法的迭代初值.

1.3　迭代过程的控制

虽然得到了较好的迭代初值,但是在迭代过程中可能会出现如下情况:

1) $\boldsymbol{\alpha}_{k+1}$元素之间大小关系错乱;

2) $\boldsymbol{\beta}_{k+1}$与 $\boldsymbol{\beta}_k$之间差距过大.

这些问题会导致 Hessian 矩阵奇异,迭代失败. 通过分析,这些情况通常是由步长 $\boldsymbol{\Delta}_k=\boldsymbol{H}^{-1}(\boldsymbol{b}_k)\cdot\boldsymbol{f}(\boldsymbol{b}_k)=[\Delta\boldsymbol{\alpha}_k;\ \Delta\boldsymbol{\beta}_k]$过大造成. 下面给

出一种自适应控制方法.

当出现情况 1)时,利用下式减小步长:

$$\boldsymbol{b}_{k+1}=\boldsymbol{b}_k-w\cdot\boldsymbol{\Delta}_k. \tag{22}$$

其中,w 为调整系数,初值为 1,一旦常系数在 $k+1$ 次迭代不满足次序条件时便将 w 赋值为 $w/2$,重新计算 $\boldsymbol{b}_{k+1}$.

为防止情况 2)的出现,需重点控制 $\Delta\boldsymbol{\beta}_{k+1}$元素之间的差距. 设置阈值 t_{d},令 $u=\max(|\Delta\boldsymbol{\beta}_{k+1}|)$. 若 u 大于阈值 t_{d},则需要对其进行约束:

$$\boldsymbol{b}_k=\boldsymbol{b}_{k-1}-\frac{\boldsymbol{\Delta}_k}{f_{\mathrm{d3}}\cdot u}. \tag{23}$$

其中:放大系数 f_{d3}推荐为 1.2 ~ 1.5;阈值 t_{d}为1 ~ 2.

算法描述见算法 2.

算法 2 迭代过程的控制

输入 $\boldsymbol{b}_0, f_{\mathrm{d3}}, t_{\mathrm{d}}$,收敛阈值 ε.

1) 初始化参数 $k=0, w=1$;

2) 利用迭代格式(6)计算 $\boldsymbol{b}_{k+1}$;

3) 判断 $\boldsymbol{\alpha}_{k+1}$元素间大小关系是否正确. 若是,转步骤 4);若否,$w=w/2$,通过式(22)调整,转步骤 3);

4) 判断 $u=\max(|\Delta\boldsymbol{\beta}_{k+1}|)$是否在阈值 t_{d}之内. 若是,转步骤 5);若否,通过式(23)调整,转步骤 3);

5) $\|\boldsymbol{b}_{k+1}-\boldsymbol{b}_k\|_2^2$ 是否小于 ε. 若是,结束;若否,$k=k+1, w=1$,转步骤 2).

输出 $\boldsymbol{b}_{k+1}$.

算法 2 中的“=”为赋值符号. 由于算法 1 已保证了 $\boldsymbol{\alpha}_0$的顺序,因此只要 w 调整适当,总会保持 $\boldsymbol{\alpha}_k$的大小关系不变.

2 滚动轴承健康状态评估实例

文献[9]通过非线性时间序列 Logistic 模型,验证了检测滚动轴承非线性突变信号的敏感性和有效性,因此本文也以轴承健康状态评估为例. 选取美国凯斯西储大学轴承数据库[5]中采样频率为 12 kHz,负载为 0,正常状态和滚动体故障(通过电火花加工设置故障,火花点直径为 0.177 8, 0.355 6, 0.533 4 mm,分别模拟轻微磨损、一般磨损及严重磨损)的振动信号数据为原始数据,其中响应类别 1,2,3,4 分别表示正常状态、轻微磨损、一般磨损和严重磨损,模型建立过程如下.

首先,根据文献[10],提取振动信号中的 30 个特征,得到相应特征值,再利用逐步模型选择法,以 Wald 统计量为指标从中逐步筛选出 8 个主要特征:$F_{\mathrm{C}}, F_2, F_5, F_6, F_7, F_8, F_9, F_{12}$;然后,利用文献[11]所述剔除离群点方法,从所有 606 组数据中筛选出有效数据 597 组;最后,通过本文方法和 SPSS 软件分别建立累积 Logistic 模型,所得回归系数如表 1 所示.

表 1 不同模型的回归系数

Table 1 Regression coefficients of different models

模型	常系数			变系数							
	1	2	3	F_{C}	F_2	F_5	F_6	F_7	F_8	F_9	F_{12}
本文	−3.915	−17.839	−31.791	−56.702	−3.054	84.526	8.477	−2 399.118	3 166.694	790.788	19.472
SPSS	0.781	−7.393	−14.495	−25.81	−2.565	39.929	6.188	−1 148.056	1 522.747	385.872	12.574

根据回归结果,两种方法所得各系数在组内所占权重大致相同,且各自变量系数的正负符号相同,说明两种方法对于当前数据具有相同的解释性. 模型评价指标数值结果见表 2.

表 2 不同模型的评价指标

Table 2 Evaluation indicators of different models

评价指标	本文模型	SPSS
AIC	85.898 3	90.644 0
McFadden − R^2	0.946 9	0.922 9
预测准确率	0.974 9	0.966 5
Gamma	0.999 6	0.999 2
Somers' D	0.986 4	0.982 4
Kappa	0.965 0	0.955 6

表 2 中除赤池信息量 AIC 外,其余指标的数值越大越能证明模型的优势. 可以看出,本文方法在各项指标上都更加突出,证明了该方法的有效性. 为进一步验证稳健性,分别利用两种方法进行 200 次 Booststrap 随机试验. 每次试验在正常、轻微磨损、一般磨损及严重磨损这 4 种类别中,分别随机选取 25 组共 100 组数据作为验证样本,其余作为训练样本. 将试验结果制成柱状分布图. 两方法的训练准确率及验证准确率如图 1 ~ 图 4 所示.

由图 1 ~ 图 4 可知,两种方法训练准确率和验证准确率的分布都大致符合正态分布;通过对比可以直观看出,本文方法无论训练准确率还是验证准确率,都处于更高水平,表明该方法具有较强的稳健性.

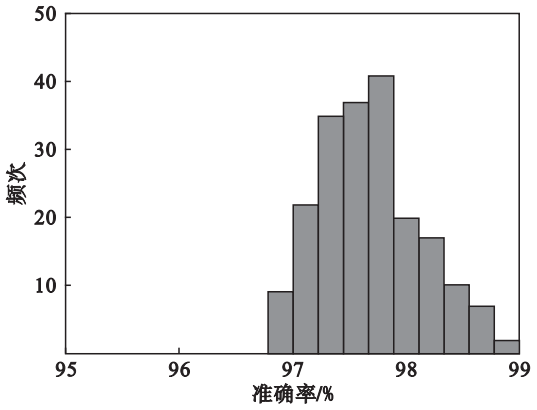

图 1　本文方法训练准确率分布图

Fig. 1　Distribution of training accuracy of the proposed method

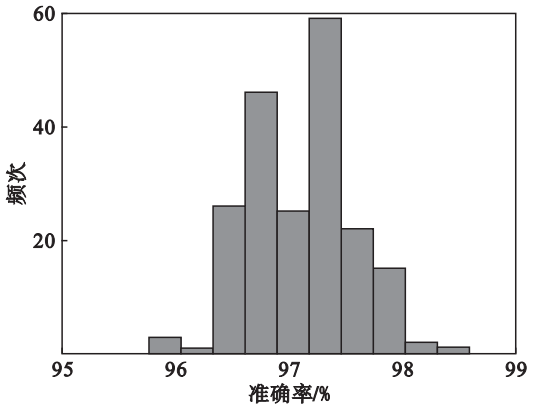

图 2　SPSS 训练准确率分布图

Fig. 2　Distribution of training accuracy of SPSS

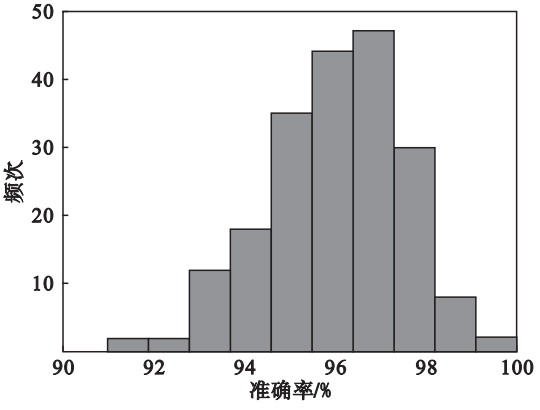

图 3　本文方法验证准确率分布图

Fig. 3　Distribution of verification accuracy of the proposed method

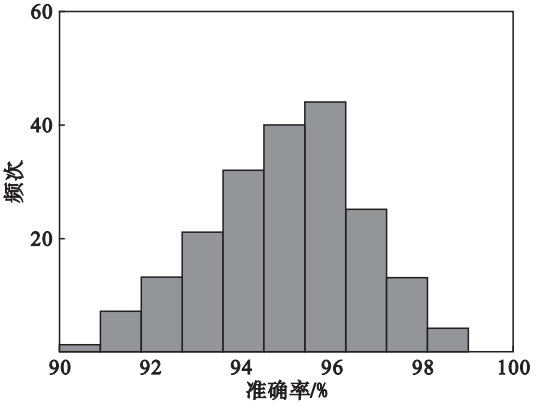

图 4　SPSS 验证准确率分布图

Fig. 4　Distribution of verification accuracy of SPSS

3　结　　论

1）通过自适应地选取迭代初值和控制迭代过程，避免了 Hessian 矩阵奇异的情况，使 Newton 迭代法能够有效地进行累积 Logistic 模型的参数估计.

2）与 SPSS 累积 Logistic 回归模型数值结果对比，本文方法在各项模型评价指标上具有更高的水平，验证了该方法的有效性.

3）基于 200 次的 Booststrap 随机试验对比，表明本文方法对于数据的依赖性较小，具有较强的稳健性.

参考文献：

[1] Nizami N, Prasad N. Multinomial Logistic regression analysis [M]. Singapore: Palgrave Macmillan, 2017: 315 – 319.

[2] Kanbayashi Y, Ishikawa T, Kanazawa M, et al. Predictive factors in patients eligible for pegfilgrastim prophylaxis focusing on RDI using ordered logistic regression analysis [J]. *Medical Oncology*, 2018, 33(5): 55 – 61.

[3] Yu J B. Aircraft engine health prognostics based on Logistic regression with penalization regularization and state-space-based degradation framework [J]. *Aerospace Science & Technology*, 2017, 68: 345 – 361.

[4] Boateng K A. Modeling external capital inflows and economic growth in Africa utilizing ordinal Logistic regression [D]. Minneapolis: Walden University, 2013.

[5] Case Western Reserve University. Bearing data center [EB/OL]. (2013 – 07 – 15) [2018 – 12 – 15]. http://csegroups.case.edu/bearingdatacenter/home.

[6] 王济川，郭志刚. Logistic 回归模型——方法与应用 [M]. 北京：高等教育出版社，2001：237 – 242.
(Wang Ji-chuan, Guo Zhi-gang. Logistic regression model—method and application [M]. Beijing: Higher Education Press, 2001: 237 – 242.)

[7] Wang H Y. More efficient estimation for logistic regression with optimal subsample [EB/OL]. (2018 – 03 – 31) [2018 – 12 – 15]. https://arxiv.org/pdf/1802.02698.pdf.

[8] Chiang L H, Kotanchek M E, Kordon A K. Fault diagnosis based on Fisher discriminant analysis and support vector machines [J]. *Computers & Chemical Engineering*, 2004, 28(8): 1389 – 1401.

[9] 刘永斌. 基于非线性信号分析的滚动轴承状态监测诊断研究 [D]. 合肥：中国科学技术大学，2011.
(Liu Yong-bin. The diagnosis of rolling bearing condition monitoring based on nonlinear signal analysis [D]. Hefei: University of Science and Technology of China, 2011.)

[10] Lei Y, He Z, Zi Y, et al. Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAS [J]. *Mechanical Systems & Signal Processing*, 2007, 21(5): 2280 – 2294.

[11] Yamada M, Liu S, Kaski S. Interpreting outliers: localized Logistic regression for density ratio estimation [EB/OL]. (2017 – 2 – 21) [2018 – 12 – 15]. https://arxiv.org/pdf/1702.06354.pdf.