文章编号:2096 - 5389(2022)02 - 0088 - 05

Logistic 回归联合 ROC 曲线模型在雷电潜势预报中的应用

吴安坤[1],郭军成[2],黄天福[3]

(1. 贵州省气象灾害防御技术中心,贵州　贵阳　550081;2. 贵州省安顺市气象局,
贵州　安顺　561000;3. 贵州省六盘水市气象局,贵州　六盘水　553000)

摘　要:利用贵阳站 2020 年 1—10 月逐日探空资料和闪电监测资料,逐一选取 72 个物理量参数纳入单因素逻辑回归,筛选具有显著性统计学意义(P <0.001)的指标进入多因素逻辑回归模型,选取满足检验条件 P <0.05 的参数得到雷电潜势预报模型,通过 Logistic 回归联合 ROC 曲线模型开展雷电潜势预报研究。结果表明:①多因素逻辑回归模型预警效果优于单因素模型,预警准确度从 75.4% 提高到 79.5%;②联合 ROC 曲线确定预报模型的概率阈值为 0.611,雷电潜势预报的命中率 *POD* 为 84.01%,虚假警报率 *FAR* 为 26.05%,临界成功指数 *CSI* 为 69.38%,准确率较高,雷电潜势预报具有较好的预报能力;③Logistic 回归模型联合 ROC 曲线法在气象预测预报,特别是非线性预测中有一定的应用价值。

关键词:雷电潜势预报;物理量参数;Logistic 回归;ROC 曲线

中图分类号:P457.9　**文献标识码**:B

Application of Logistic Regression Combined with ROC Curve Model in Lightning Potential Prediction

WU Ankun[1], GUO Juncheng[2], HUANG Tianfu[3]

(1. Guizhou Meteorological Disaster Prevention Technology Center, Guiyang 550081, China;
2. Anshun Meteorological Bureau of Guizhou Province, Anshun 561000, China;
3. Liupanshui Meteorological Bureau of Guizhou Province, Liupanshui 553000, China)

Abstract: Selects 72 physical parameters one by one into the single factor Logistic regression model based on the daily radiosonde data and lightning monitoring data of Guiyang station from January to October 2020, selects the indexes with significant statistical significance ($P < 0.001$) into the multi factor Logistic regression model, and selects the parameters that meet the test condition $P < 0.05$ to obtain the lightning potential forecast model, Logistic regression combined with ROC curve model is used to study the lightning potential prediction. The results show that the warning effect of multi factor Logistic regression model is better than that of single factor model, and the warning accuracy is improved from 75.4% to 79.5%. In addition, the probability threshold of prediction model determined by joint ROC curve is 0.611, the hit rate pod of lightning potential prediction is 84.01%, the false alarm rate far is 26.05%, and the critical success index is 69.38%; Finally, Logistic regression model combined with ROC curve method has a certain application value in meteorological forecasting, especially in nonlinear forecasting.

Key words: lightning potential prediction; physical parameters; Logistic regression; ROC curve

收稿日期:2021 - 05 - 25

第一作者简介:吴安坤(1986—),男(苗族),副高,主要从事雷电灾害防御技术研究工作,E - mail:wak - mail@163.com。

资助项目:贵州省科技支撑项目(黔科合支撑[2021]一般 510):雷电地电位致灾机理研究及雷电应急避险装置研制;贵州省科技基金项目(黔科合(基础[2022]一般 245):基于 FY - 4A 气象静止卫星的雷暴云识别及预警技术研究。

0　引言

雷暴活动作为常见的强对流天气过程,造成的灾害是联合国公布的十大最严重的自然灾害之一。随着社会经济不断发展,每年因强对流天气过程造成的损失越加严重。因此,加强雷暴活动的预测预报,对防灾减灾有十分重要的指导意义。雷暴云的发生发展伴随着不稳定环境中气团的抬升,探空资料观测大气中的温湿压、水汽和抬升等物理量参数,对研究局地雷电潜势预报具有很好的指示作用[1-4]。目前采用探空对流参数开展的雷暴预报研究,大多直接采用多元统计线性回归方法,需解决雷暴发生与否的非线性与探空资料之间的线性回归问题。线性回归模型要求因变量是连续的正态分布变量,且自变量和因变量之间呈现线性关系。当因变量为分类型变量,且自变量与因变量没有线性关系时,线性回归模型的假设条件就会遭到破坏。而采用 Logistic 回归分析模型可以很好地解决此类问题,它对因变量的分布没有要求,巧妙地避开了分类型变量的分布问题。Logistic 回归作为一种非线性概率性预测模型,可实现对研究观察结果进行分类、处理协变量之间的多变量分类分析[5],被广泛用于流行病学的病因研究中,分析疾病与危险因素间联系,所观测的因素常以二分变量取值,如生存与死亡、是否发病等,即因变量为 0 或者 1。如罗蒙等[6]将具有统计学意义的检查指标纳入多因素 Logistic 回归分析,预测新型冠状病毒肺炎患者发生危重症的风险。而 ROC 曲线是目前学术界公认的诊断价值最佳的方法,其操作简便,且具有通过图形就能够判断分析的诊断性能[7-9]。宗迎迎等[10]应用 Logistic 回归和 ROC 曲线研讨血清 Dickkopf、高尔基体糖蛋白 73 和甲胎蛋白对原发性肝癌的诊断价值。张宇等[11]应用 Logistic 模型联合 ROC 曲线法对新型冠状病毒肺炎严重程度进行判别,具有较高的正确率。引入 Logistic 回归联合 ROC 曲线模型采用探空物理量资料开展雷电潜势预报研究,分析雷电活动有、无问题,目前未见相关技术研究。因此,本文选取闪电监测资料和探空观测参数,筛选数据样本纳入单因素逻辑回归模型,选取有统计学意义的参数纳入多因素逻辑回归模型,采用 ROC 曲线联合二分类 Logistic 回归模型开展雷暴活动潜势预报研究。

1　数据来源及处理

探空资料来源于 Micaps 系统提供的 T - lnp 探空数据,提取贵阳站 2020 年 1—10 月逐日 08 时和 20 时的修正总指数、K 指数、沙氏指数、Faust 指数、最大抬升指数、对流稳定度指数等 72 种物理参数。为保证数据的可靠性,采用四分位检测异常值,剔除上四分位 +1.5 IQR 距离、下四分位 -1.5 IQR 距离以外时刻的数据。闪电资料来源于贵州省闪电监测网,考虑实际业务中 T - lnp 探空数据每天主要包括 08 时和 20 时 2 个时次,以及探空站之间的距离。规定该站当日 08 时或 20 时以后 12 h 内、100 km 范围内若发生 50 次以上的闪电,则将当日 08 时或者 20 时对应的物理参量作为 1 个雷暴天气样本,反之为非雷暴天气。本文通过筛选得到 294 个雷暴、238 非雷暴天气样本以及对应的 72 个物理参数值、闪电活动次数。

2　分析方法

以筛选的样本中 72 个对流参数为因变量,纳入单因素逻辑回归模型,筛选变量,将有显著性统计学意义($P<0.001$)的变量纳入多因素逻辑回归模型,选取满足一定检验条件的参数代入模型计算概率预测值。以此概率预测值为检验变量,样本雷暴活动情况为状态变量,绘制 ROC 曲线,以敏感度与特异性之和最大所对应的概率值作为截断值,纳入气象预报质量评分检验。

2.1　Logistic 模型联合 ROC 曲线法

假设雷暴发生情况 y,发生为 1,未发生为 0。影响雷暴发生情况 y 的 m 个对流参数分别为 $x_1,x_2,\cdots,x_m$。雷暴活动发生的概率记为 $P(y=1\mid x_i)=P_i$,发生与否的 2 个概率分别为:

$$P_i=\frac{1}{1+e^{-(\alpha+\sum_{i=1}^{m}\beta_i x_i)}}=\frac{e^{\alpha+\sum_{i=1}^{m}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{m}\beta_i x_i}} \tag{1}$$

$$1-P_i=1-\frac{e^{\alpha+\sum_{i=1}^{m}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{m}\beta_i x_i}}=\frac{1}{1+e^{\alpha+\sum_{i=1}^{m}\beta_i x_i}} \tag{2}$$

其中 P_i 代表在第 i 个观测中雷暴发生的概率,$1-P_i$ 对应雷暴未发生的概率,均为对流参数 x_i 构成的非线性函数。雷暴发生与不发生的概率之比 $P_i/(1-P_i)$,称为事件的发生比(Odds),对 Odds 取对数变换,得到逻辑回归模型的线性模式如下:

$$\ln\left(\frac{p_i}{1-p_i}\right)=\alpha+\sum_{i=1}^{m}\beta_i x_i \tag{3}$$

得到雷暴活动发生概率 P 如下所示:

$$p=\frac{1}{(1+e^{-z})},z=\alpha+\sum_{i=1}^{m}\beta_i x_i \tag{4}$$

ROC 曲线(receiver operating characteristic

curve)分析被认为是一种诊断试验评价中理想和经典的方法。其思路是以逻辑回归模型所得的预测概率为基础,选取不同截断值按照表 2 描述的雷暴预报混淆矩阵进行统计,计算不同截断值下的敏感度与特异度。

表 1　混淆矩阵

Tab. **1**　Confusion matrix

雷暴发生情况	预测发生雷暴	预测未发生雷暴
实际发生雷暴	*TP*	*FP*
实际未发生雷暴	*FN*	*TN*

其中敏感度 = $\frac{TP}{TP + FN}$、特异性 = $\frac{TN}{TN + FP}$、准确度 = $\frac{TP + TN}{TP + TN + FP + FN}$,以敏感度、特异性为坐标轴,将曲线图形化,获得更直观的评估检验结果。曲线面积(AUC) 用于验证敏感性模型的准确性,较高的值表明模型具有更好的预测能力。一般认为,AUC < 0.5,预测失败;0.5 ~ 0.7 之间预测较好;0.7 ~ 0.9 之间预测很好;AUC > 0.9 表示模型预测效果十分好[12]。

2.2　气象预报检验

对于雷电等强对流天气的小概率时间检验采用 Donaldson[13] 提出的方法,计算命中率(*POD*)、虚警率(*FAR*)、临界成功指数(*CSI*)、失误率(*FOM*)衡量预报方程的准确率和进行预报质量评分。在表 1 混淆矩阵的基础上,计算 *POD*、*FAR*、*CSI*、*FOM* 如下所示:

$$POD = \frac{TP}{TP + FP} \times 100\% \tag{5}$$

$$FAR = \frac{FN}{FN + TP} \times 100\% \tag{6}$$

$$CSI = \frac{TP}{TP + FP + FN} \times 100\% \tag{7}$$

$$FOM = \frac{FP}{TP + FP} \times 100\% \tag{8}$$

表 2　探空物理量参数为因变量构建单因素、多因素逻辑回归模型

Tab. **2**　Single factor and multi factor Logistic regression models with sounding physical parameters as dependent variables

对流	单因素分析		多因素分析	
参数	*OR*(95% *CI*)值	*P* 值	*OR*(95% *CI*)值	*P* 值
SWISS12	0.84(0.81 ~ 0.87)	< 0.001	1.01(0.86 ~ 1.17)	0.930
*SWISS*00	0.84(0.81 ~ 0.87)	< 0.001	0.74(0.58 ~ 0.94)	0.015
LI	0.83(0.80 ~ 0.86)	< 0.001	0.65(0.45 ~ 0.96)	0.031
BLI	0.69(0.64 ~ 0.75)	< 0.001	0.81(0.71 ~ 0.92)	0.002
IL	0.91(0.90 ~ 0.93)	< 0.001	1.13(0.97 ~ 1.32)	0.109
SI	0.84(0.81 ~ 0.88)	< 0.001	1.30(0.81 ~ 2.08)	0.277
TQG	1.00(1.00 ~ 1.00)	< 0.001	1.00(1.00 ~ 1.00)	0.346
TMJ	1.03(1.02 ~ 1.03)	< 0.001	1.03(0.98 ~ 1.08)	0.270
Faust	1.17(1.14 ~ 1.21)	< 0.001	1.01(0.65 ~ 1.55)	0.972
DCI	1.06(1.05 ~ 1.07)	< 0.001	0.89(0.65 ~ 1.22)	0.462
mK	1.12(1.09 ~ 1.14)	< 0.001	0.97(0.92 ~ 1.03)	0.327
TCL_T	1.19(1.15 ~ 1.23)	< 0.001	0.92(0.57 ~ 1.48)	0.719
CCL_T	1.19(1.15 ~ 1.23)	< 0.001	1.13(0.82 ~ 1.56)	0.443
IntegralQ	1.00(1.00 ~ 1.00)	< 0.001	1.00(0.90 ~ 1.20)	0.004

3　结果分析

3.1　Logistic 回归分析模型

通过单因素分析显示瑞士第二雷暴指数(*SWISS*12)、瑞士第一雷暴指数(*SWISS*00)、抬升指数(*LI*)、最大抬升指数(*BLI*)、条件对流稳定度指数(*IL*)、沙氏指数(*SI*)、通气管指数(*TQG*)、修正杰弗逊指数(*TMJ*)、Faust 指数(*Faust*)、修正对流指数(*DCI*)、修正 K 指数(*mK*)、抬升凝结处温度(*TCL_T*)、对流凝结高度处温度(*CCL_T*)、整层比湿积分(*IntegralQ*)14 个对流参数,有统计学意义(均有 *P* < 0.001),即以上 14 个参数对雷暴活动趋势有指示作用,涉及大气热力因子、动力条件及综合指数等,可综合反映中低层热动力稳定度特性。其中 *SWISS*12、*SWISS*00、*LI*、*BLI*、*IL*、*SI* 6 个参数 *OR* < 1,表征参数越小,发生雷暴活动的可能性越大;反之其他 8 个参数值越大,发生雷暴活动的可能性越大。进一步对有统计学意义的 14 个参数采用多因素分析,结果显示 *SWISS*00(*OR* = 0.74,95% *CI*: 0.58 ~ 0.94, *P* < 0.05)、*LI*(*OR* = 0.65,95% *CI*: 0.45 ~ 0.96, *P* < 0.05)、*BLI*(*OR* = 0.81,95% *CI*: 0.71 ~ 0.92, *P* < 0.05)、*IntegralQ*(*OR* = 1.00,

95% *CI*:0.90 ~ 1.20,$P < 0.05$)4 个参数为雷电潜势预报多参数逻辑回归指标,即 $\ln(p/(1-p)) = 0.306 \times SWISS00 + 0.424 \times LI + 0.214 \times BLI - 0.001 \times IntegralQ$。

3.2　ROC 曲线

在概率截断值为 0.5 水平下,如表 3 所示,单因素构建逻辑回归模型预准确度介于 68.2% ~ 75.4%,ROC 曲线的曲线下面积介于 0.751 ~ 0.793,以整层比湿积分相对最好、沙氏指数相对最差。若以整层比湿积分(*IntegralQ*)作为单因素指标开展雷电潜势预报,准确度为 75.4%。

采用多因素逻辑回归模型,ROC 曲线的曲线下面积为 0.839(0.804 ~ 0.875),$P < 0.001$,预测能力较单因素模型有所提高,具有较好的预测价值(图 1)。当 Logistic 回归分析模型得到的预测值为 0.611 时,其敏感度为 0.789,特异度为 0.799,二者之和最大,因此将该值作为最佳临界点将研究对象分为 2 组,即 Logistic 回归分析模型预测概率值≥ 0.611 认为有雷电天气过程,在此条件下,准确度由单因素的 75.4% 提高到 79.5%。

表 3　单因素指标 ROC 曲线下的面积

Tab. **3**　Area under ROC curve of single factor index

对流参数	准确率/%	面积	标准差[a]	渐进 Sig.[b]	渐近 95 置信区间	
					下限	上限
SWISS12	71.40	0.792	0.020	0.000	0.753	0.830
SWISS00	70.10	0.786	0.020	0.000	0.747	0.824
LI	70.10	0.769	0.020	0.000	0.729	0.809
BLI	70.90	0.766	0.021	0.000	0.726	0.807
IL	69.70	0.760	0.021	0.000	0.719	0.800
SI	68.20	0.751	0.021	0.000	0.709	0.792
TQG	71.10	0.770	0.021	0.000	0.730	0.811
TMJ	73.10	0.766	0.021	0.000	0.725	0.808
Faust	71.80	0.759	0.021	0.000	0.717	0.801
DCI	72.00	0.780	0.020	0.000	0.741	0.820
mK	74.10	0.786	0.019	0.000	0.768	0.814
TCL_T	73.50	0.775	0.021	0.000	0.734	0.816
CCL_T	72.60	0.776	0.021	0.000	0.735	0.818
IntegralQ	75.40	0.793	0.019	0.000	0.775	0.821

注:a. 在非参数假设下;b. 零假设:实面积 = 0.5

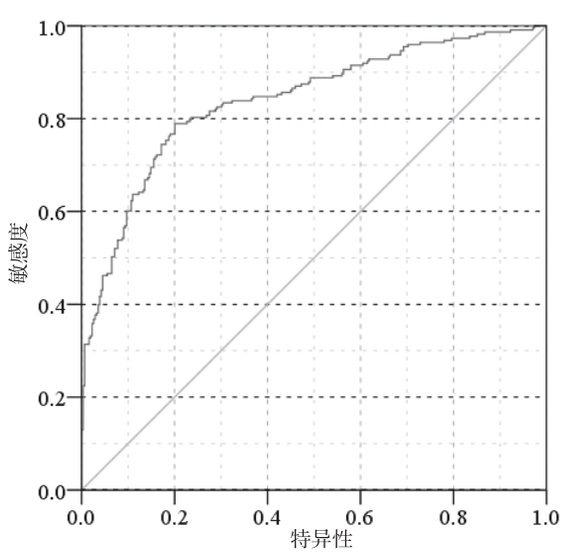

图 1　多因素逻辑回归 ROC 曲线

Fig. 1　Multivariate Logistic regression ROC curve

3.3　预报质量校验

根据确定的概率截断值 0.611,在混淆矩阵的基础上统计 *TP* 为 247、*FP* 为 47、*FN* 为 62、*TN* 为 176,采用气象预报评分计算命中率(*POD*)为 84.01%、虚警率(*FAR*)为 26.05%、临界成功指数(*CSI*)69.38%、失误率(*FOM*)为 20.06%。

4　结论与讨论

本文选取 72 个探空物理量参数作为自变量,闪电监测系统探测是否发生闪电作为因变量,将单因素指标逐一纳入逻辑回归模型,筛选具有显著性统计学意义($P < 0.001$)的指标进入多因素回归模型,选取满足检验条件 $P < 0.05$ 的参数得到雷电潜势预报模型 $\ln(p/(1-p)) = 0.306 \times SWISS00 + 0.424 \times LI + 0.214 \times BLI - 0.001 \times IntegralQ$。得到结论如下:

①多因素逻辑回归模型预警效果优于单因素模型,预警准确度从 75.4% 提高到 79.5%。

②联合 ROC 曲线确定预报模型的概率阈值为 0.611，雷电潜势预报的命中率 *POD* 为 84.01%，虚假警报率 *FAR* 为 26.05%，临界成功指数 *CSI* 为 69.38%。准确率较高，雷电潜势预报具有较好的预报能力。

Logistic 回归模型处理“二分类”问题，旨在拟合结果的“有”“无”问题，有效弥补了线性回归的缺陷；同时结合 ROC 曲线对模型进行检验，确定合适的预测概率值，可进一步提高预警准确率。Logistic 回归模型联合 ROC 曲线法在气象预测预报，特别是非线性预测中有一定的应用价值。

参考文献

[1] Solomon R, Baker M. Electrification of New Mexico thunderstorms [J]. Mon Wea Rew, 1994, 122: 1878－1886.

[2] Robert A Mazany, Steven Businger, Srth I Gutman, et al. A Lightning Prediction Index that Utilizes GPS integrated Precipitable Water Vapor[J]. Weather and Forecasting, 1998, 17(5): 1034－1047.

[3] 李迪，吴安坤，陆扬. 下垫面对贵州闪电活动规律影响的分析[J]. 气象水文海洋仪器，2021，38(4)：44－46.

[4] 毕波，高兵，杨航. 大理机场雷暴特征及潜势预报分析[J]. 中低纬山地气象，2020，44(6)：71－75.

[5] Yu F F, Liu H, Guo X. Integrative multivariate Logistic regression analysis of risk factors for Kashin－Beck disease[J]. Biol Trace Elem Res, 2016, 174: 274－279.

[6] 罗蒙，黄晓东，江波，等. Logistic 回归联合 ROC 曲线模型预测新型冠状病毒肺炎患者发生危重症的风险[J]. 中草药，2020，51(20)：5287－5292.

[7] 李红艳. ROC 曲线评价血清胱抑素 C 对早期肾功能损伤的临床诊断价值[J]. 中国实验诊断学，2017，21(6)：958－960.

[8] 刘洋，刘宁. 受试者工作特征(ROC)曲线在超声诊断中的应用[J]. 中外医学研究，2012，10(9)：149－151.

[9] 陈卫中，潘晓平，倪宗瓒. Logistic 回归模型在 ROC 分析中的应用[J]. 中国卫生统计，2007(1)：22－24.

[10] 宗迎迎，徐浩，许伟，等. Logistic 回归和 ROC 曲线分析血清 DKK1、GP73 和 AFP 在原发性肝癌诊断中的价值[J]. 检验医学，2015，30(6)：559－563.

[11] 张宇，舒晓利，钟波，等. Logistic 模型联合 ROC 曲线法和 Bayes 判别函数对新型冠状病毒肺炎严重程度的鉴别[J]. 中华疾病控制杂志，2020，24(7)：851－855.

[12] 周超，方秀琴，吴小君，等. 基于三种机器学习算法的山洪灾害风险评价[J]. 地球信息科学学报，2019，21(11)：1679－1688.

[13] Donaldson R J, Dyer R M, Keauss M J. An objective evaluator of techniques for predicting severe weather events[C]. Amer Meteor Soc eds. 9 th Conf Severe Local Storms, 1975: 321－326.