DOI:10.20033/j.1003-7241.(2025)07-0066-05

基于网格搜索优化逻辑回归的配电物联协议检测

王立旭[1,2]，何鸣一[1,2,3]，吕 非[1,2]，周 福[1,2]

（1.南瑞集团有限公司（国网电力科学研究院有限公司），江苏 南京 211000；
2.国电南瑞科技股份有限公司，江苏 南京 211000；
3.智能电网保护和运行控制国家重点实验室，江苏 南京 211000）

摘 要：为了保证电力工程现场能够高效地验证配电物联设备的接入能力，以及解决目前传统协议检测方法的效率低、依赖人工、缺乏自适应性等问题，提出运用机器学习理论实现适用于配电物联协议一致性检测的方法。该方法采用逻辑回归（logistic regression，LR）模型，结合改进的网格搜索（grid search，GS）穷举法对学习步长和正则项系数值进行参数寻优，以此构建出泛化能力较强的检测模型，有效改善了逻辑回归模型的调优效率和提升了预测准确率。通过对比分析，该方法实现的协议一致性检测准确率可达到97.84%，AUC值可达到0.97，误差低至2.7，整体性能优于其他分类模型，能够使得检测系统具备自动化和智能化，极大地提升检测效率，可以为配电物联设备批量接入检测提供可靠保障。

关键词：逻辑回归；协议一致性检测；配电物联网；MQTT

中图分类号：TP277 **文献标识码：**A **文章编号：**1003-7241（2025）07-0066-05

Power Distribution IoT Protocol Detection Based on Grid Search Optimization with Logistic Regression

WANG Lixu[1,2], HE Mingyi[1,2,3], LV Fei[1,2], ZHOU Fu[1,2]

(1. NARI Group Corporation (State Grid Electric Power Research Institute), Nanjing 211000, China;
2. NARI Technology Co., Ltd., Nanjing 211000, China;
3. State Key Laboratory of Smart Grid Protection and Control, Nanjing 211000, China)

Abstract: In order to ensure that power engineering sites can efficiently verify the access capability of power distribution IoT devices, as well as to solve the current problems of low efficiency, dependence on manual labor, and lack of adaptability of traditional protocol detection methods, it proposes to use machine learning theory to realize a method applicable to the consistency detection of power distribution IoT protocols. The method adopts the logistic regression (LR) model, combined with the improved grid search (GS) exhaustive method for parameter optimization of the learning step and regular term coefficient values, to construct a detection model with strong generalization ability, which effectively improves the tuning efficiency of the logistic regression model and enhances the prediction accuracy. Through comparative analysis, the accuracy of protocol consistency detection achieved by this method can reach 97.84%, the AUC value can reach 0.97, and the error is as low as 2.7. The overall performance is better than other classification models, and it can make the detection system have automation and intelligence, greatly improve the detection efficiency, and provide a reliable guarantee for the detection of batch access to the distribution IoT equipment.

Keywords: logistic regression; protocol conformance testing; power internet of things; message queuing telemetry transport

0 引言

在国家电网智慧物联体系的建设背景下，配电物联网消息队列遥测传输协议（message queuing telemetry transport，MQTT）协议一致性检测能力决定了国网统一技术标准是否能够快速地落地和应用。随着各类智能设备的批量接入，工程现场急需一个高效的接入检测方式，能够快速实现对不同配电设备通信交互的MQTT报文判别出是否遵循协议规范，对配电物联协议一致性进行智能检测，摆脱人工低效而又重复的工作，提高工程现场批量接入的效率。

目前在电力行业对协议检测方法的研究都是基于传统方式。文献[1]设计了物联全场景仿真检测系统，提升了场景检测能力，但检测过程中需要手动导入模型，自动化程度不足。文献[2]提出针对用采协议的检测方法，提升了检测扩展性，但协议内容相对单一，检测效率较低。

*基金项目：南瑞研究院科技项目资助（524608230033）；南瑞研究院科技项目资助（524608220017）

收稿日期：2023-11-21

文献[3]提出了电力通用服务协议一致性检测，该方法节省了大量检测时间也避免了人工测试的随机性，但主要是对变电站的通信接口做远程交互检测。可以看出，传统的电力物联协议检测只能一定程度上改善某个专业方向的检测能力，随着配电物联设备的逐渐多样化，通信交互的MQTT协议报文类型和内容也随之增多和复杂，对协议检测方法的准确性、智能度以及检测性能都有了更高的要求。

基于此，本文研究逻辑回归(logistic regression，LR)模型在配电物联协议一致性检测中的应用，相比传统的协议检测方式，该方法不但具有更高效的检测能力，而且还具备了检测智能性。

1 检测方法

1.1 配电物联协议检测系统

配电物联MQTT协议[4-5]一致性检测系统采用分层架构设计，如图1所示。主要包括5部分：数据系统、数据预处理、离线训练、设备接入和在线检测。数据系统是物联平台获取已投运设备业务交互的协议报文数据，通过不同业务主题分类作为模型离线训练的历史数据源，如图1(a)所示。数据预处理是对投运设备的历史数据和检测设备的实时数据进行统一解析处理，利用标准化手段生成样本数据，如图1(e)所示。离线训练针对不同的业务场景利用数据集对相应的检测项模板分别寻优得出预测模型，并存储至模型库，如图1(b)所示。设备接入是通过前置服务完成待检设备自动注册，设备接入上线后，检测系统通过检测项模板和检测设备进行业务交互，输出实时检测数据，如图1(c)所示。在线检测主要是根据待检设备选择的检测项模板自动从模型库中选择相应的预测模型，通过对设备进行实时检测和过程记录，最终系统会将各检测项的预测结果整合输出，如图1(d)所示。

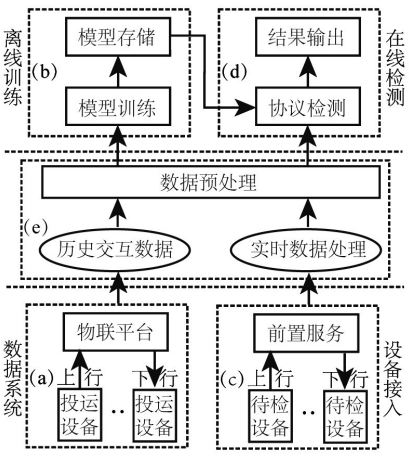

图1 协议检测系统结构图

1.2 逻辑回归模型

逻辑回归是解决分类问题的一种常用模型[6-8]，它根据样本数据特征，计算其归属于某一类别的概率$P(x)$，根据概率值判断其所属类别。假设输入样本的特征向量为$x=(x_1,x_2,\cdots,x_i)$，在给定特征x_i条件下输出真实标签y_i的概率如式(1)所示，通用损失函数如式(2)所示。

$$P(\boldsymbol{x}_i)=\boldsymbol{\Phi}(z_i)^{y_i}\cdot(1-\boldsymbol{\Phi}(z_i))^{1-y_i} \tag{1}$$

$$l(\boldsymbol{w})=\sum_{i=1}^{n}\begin{bmatrix} y_i\ln(\boldsymbol{\Phi}(z_i))+ \\ (1-y_i)\ln(1-\boldsymbol{\Phi}(z_i)) \end{bmatrix} \tag{2}$$

式中，z_i是$\boldsymbol{w}$的线性变换，$\boldsymbol{w}$是系数向量，即$z=\boldsymbol{w}^T\boldsymbol{x}+b$，$l(\boldsymbol{w})$为对数极大似然函数。利用梯度上升法得出求和项最大的$\boldsymbol{w}$值的过程就是模型的学习过程，步骤如下：

(1) 初始化回归系数向量$\boldsymbol{w}$；

(2) 设定最大迭代次数；

(3) 循环计算出每次迭代梯度，即$grad=\partial l(\boldsymbol{w})/\partial \boldsymbol{w}$。

迭代中最新的$\boldsymbol{w}$值为梯度乘以步长系数值α，加上前一次循环的$\boldsymbol{w}$值，如下式：

$$\boldsymbol{w}_{new}=\boldsymbol{w}+\alpha\cdot grad \tag{3}$$

式中，$\boldsymbol{w}$的值在每一次迭代中被$\boldsymbol{w}_{new}$更新，即梯度为

$$\begin{aligned}\frac{\partial l(\boldsymbol{w})}{\partial \boldsymbol{w}}&=\sum_{i=1}^{n}\left(y_i-\frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}_i+b)}}\right)\cdot\boldsymbol{x}_i \\ &=\sum_{i=1}^{n}\mathrm{err}_i\cdot\boldsymbol{x}_i=\mathbf{err}\cdot\boldsymbol{x}\end{aligned} \tag{4}$$

式中，y_i是$\boldsymbol{x}_i$真实标签，$1/1+e^{-(\boldsymbol{w}^T\boldsymbol{x}_i+b)}$是预测值，两者之差为误差$\mathrm{err}_i$，所以梯度表示为$\mathrm{err}\cdot\boldsymbol{x}$。根据迭代$\boldsymbol{w}$的过程，结合式(3)，$\boldsymbol{w}$值最终为

$$\boldsymbol{w}:=\boldsymbol{w}+\alpha\cdot\mathbf{err}\cdot\boldsymbol{x} \tag{5}$$

实际中，为了保证模型的泛化能力和避免过拟合，对损失函数进行正则化来调整模型的复杂度。本文结合样本特征避免系数向量值过于稀疏，采用L2范数正则化[9]，如下式：

$$l(\boldsymbol{w})_{L2}=l(\boldsymbol{w})-\frac{\lambda}{2}\|w\|^2 \tag{6}$$

根据式(4)计算过程对$l(\boldsymbol{w})_{L2}$求梯度再结合式(5)，最终引入正则项的$\boldsymbol{w}$值表示为

$$\boldsymbol{w}_{L2}:=\boldsymbol{w}+\alpha\cdot(\mathbf{err}\cdot\boldsymbol{x}-\lambda\boldsymbol{w}) \tag{7}$$

因此，对式(7)中采用网格搜索法(grid search，GS)[10-12]选取到合理的学习步长α和正则项系数λ值，构建出表现能力最优的模型。

1.3 GS-LR预测模型

为了保证构建出的模型更加准确且避免对数据集过拟合。将训练集分成N个子集，利用交叉验证(cross validation，CV)的方式进行分组训练，如图2所示。每组依次

选择N-1个子集作为训练集(Trn)和1个子集作为验证集(Val),分别作为模型训练和参数调优。进行N轮次训练,不符合预期的模型则重新调参训练,测试集不参与模型创建,只作为衡量最终模型性能。

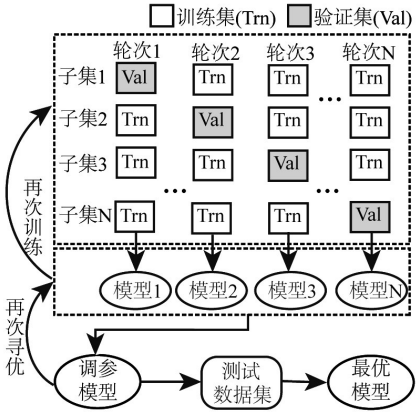

图2 交叉验证详细图

训练中手动指定参数过程繁杂且效率低下。为了能够实现自动寻参,本文结合GS法对模型预设参数范围,结合交叉验证进行评估,选出最优参数组合建立模型对象。实际中,为了解决GS法穷举搜索到的值在范围边界而错过最优值以及出现局部最优解的问题,本文提出一种改进的GS寻参法,即通过关系判断来扩展搜索区间进行二次搜索,步骤如下:

(1) 初次搜索时设定穷举范围$[J,K]$,设定步长P和最大迭代次数I;

(2) 交叉验证对范围内的所有参数值进行训练评估,得出均方根误差最小的参数值Q;

(3) 比较$|Q\text{-}J|$和$|Q\text{-}K|$绝对值与步长P之间大小,若$|Q\text{-}J|$小于P,范围设定为$[2J,J+P]$;若$|Q\text{-}K|$小于P,范围设定为$[K+P,2K]$,步长均为P,再次交叉验证寻找到新的参数值,且均方根误差和准确率都满足预期,直至不接近边界值;

(4) 比较穷举范围内最优参数值Q_1和次优参数值Q_2的差绝对值;若小于P,则再次进行交叉验证,范围设定为$[Q_1+P,Q_2+P]$,步长为$P/2$,寻找到新的参数值,若新值接近范围边界转到第3步,直至寻找到参数值不在边界,也不局部过于集中,则将寻优到的各参数值输出生成模型对象,即GS-LR预测模型。

由于逻辑回归模型性能取决于学习步长和正则项系数的值,通过改进的GS法寻参获得的最优参数组合在均方根误差和准确率上都可以满足精度和达到预期,并且较其他寻优算法收敛快、整体开销小。

2 GS-LR模型构建

2.1 数据获取

为了实验本文算法模型的可靠性和适用性,以江苏现场已投运设备历史交互数据进行分析。设备的类型主要包括边缘计算框架、融合终端和智能终端等,通过收集现场工程人员反馈的各类设备常见异常因素分析,将MQTT协议规范性上的异常错误作为本文数据集的主要特征,如表1所示。配电设备通过MQTT规约与平台通信交互,按照不同业务主题进行分类,为了避免不同主题的协议报文不一致性,检测系统会对采集的实际数据进行预处理,将主要特征参数和报文体中的遥信、遥测等各类数值作为模型训练的数据特征,根据主题值区分出交互类型和业务用例,检测系统通过数据集训练出不同业务用例的检测模型。

表1 主要特征描述

特征	描述	取值
TOPIC	交互主题	业务编码值
MID	报文请求号	消息标识值
DEVICEID	设备标识	出厂编号值
TYPE	消息类型	消息业务标识值
CODE	应答返回码	响应状态值
PARAM	报文体	遥信、遥测等值
DIRECTION	消息流向	报文标识值

2.2 特征处理

获取到的历史数据都是MQTT报文消息,需要通过预处理对每条消息特征解析。协议检测用到的数据分为2种类型,一是上行消息,即设备上送给平台的响应消息;二是下行消息,即平台下发到设备的指令消息。设备上行消息会存在重复或类型异常等数据,而下行消息会出现冗余字段或缺失值的情况。因此,首先通过数据清洗去除无效数据和多余特征,再对分开特征和缺失值以及标签值单独处理,保证数据的一致性和正确性。

为了避免不同特征数值差别大影响数据分析的结果,将训练集和测试集数据映射到同一尺度进行标准化处理。部分特征值没有明显边界可能存在极端值,因此利用均值方差归一化统一将数据归一到均值为0方差为1的分布中,完成特征值标准化,如式(8)所示。

$$x_{scale}=\frac{x-x_{mean}}{s} \tag{8}$$

式中,x为某特征值,x_{mean}为特征值对应均值,s为特征值对应方差。

测试集是模拟真实数据对模型做性能评判,因为无法得出其均值和方差,因此测试集进行归一化采用训练集的每个特征均值和方差进行处理,从而保证最终训练出的模型泛化能力更强,如所(9)示。

$$x_{test\_scale}=\frac{x_{test}-x_{train\_mean}}{s_{train\_std}} \tag{9}$$

式中,x_{test}为测试集某特征值,x_{train_mean}为训练集中特征值对

应均值，s_{train_std}为训练集特征值对应方差。

2.3 模型构建

本文通过构建GS-LR模型用于对配电物联协议一致性进行预测。将特征处理后的数据集作为输入，相应的标签类型作为输出，模型构建过程如下：

（1）数据集划分，训练集和验证集进行模型训练，测试集做模型性能评判；

（2）利用改进的GS法对模型的学习步长和正则项系数迭代寻优；

（3）将寻优结果输入到模型，利用测试集对模型性能评判。符合评判的模型存储到模型库，反之返回第（2）步继续迭代寻优；

（4）使用最优的GS-LR模型进行预测。

因此，根据以上模型构建过程，下面给出GS-LR算法伪代码。GS寻优参数步骤如算法1所示。

算法1 GS寻优算法

输入：训练集；验证集；测试集
输出：最优参数组合

1）设定参数穷举范围、步长P、迭代次数I；
2）while 寻参未达到评判指标 do
3）交叉验证得出均方根误差最小的参数组合；
4）判断参数组合是否接近边界值或过于集中，若满足其一重新迭代循环；
5）利用验证集评判参数组合；
6）end do

逻辑回归步骤如算法2所示。

算法2 逻辑回归算法

输入：最优参数；训练集；验证集；测试集
输出：最优检测模型对象

1）while 性能未达到评判指标 do
2）调用参数GS寻优算法；
3）根据最新调优参数生成模型对象；
4）利用测试集评判最新模型对象；
5）end do

3 实验及结果分析

3.1 数据集

本文算例以江苏现场2023年3月1日至5月30日共90天获取到的投运设备与平台交互的MQTT历史报文，以及每条报文正确性的标识值。通过预处理对原始数据解析和标准化后，一共选取了3 379个样本数据，如表2所示。其中正常报文1 720个，异常报文1 659个，0为符合规约的正常数据，1为异常数据，按照70%训练集，15%验证集和15%测试集划分。

表2 数据集分布

数据集	异常	训练集	验证集	测试集	总计
协议报文	0	1 204	258	258	1 720
	1	1 162	248	249	1 659

3.2 性能指标

模型性能评判选用准确率（Accuracy）、精准率（Precision）、召回率（Recall）、F1-score、AUC（area under curve）以及均方根误差（root mean square error，RMSE），计算公式分别为

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

$$Recall=\frac{TP}{TP+FN} \tag{12}$$

$$F1-score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{13}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{14}$$

式（10）~式（13）中，TP（True Positive）为真正例，TN（true negative）为真反例，FP（false positive）为伪正例，FN（false negative）为伪反例；式（14）中y_i为真实值，$\hat{y}_i$为预测值，N为预测总数。

3.3 结果分析

根据数据特征，对表2中训练集进行5轮分组交叉验证进行训练，再用测试集对模型预测评估。

首先，使用改进的GS法对模型参数进行寻优。设定最大迭代次数为500，选取2组学习步长和3组正则项系数区间，通过数据集边训练边寻优，直至GS算法达到最大迭代次数后输出最优值。如表3所示，学习步长范围在[0.001，0.01]之间时模型表现较优，测试集准确率达0.93；正则项系数范围在[0.1，1]之间时模型表现达到最优，测试集准确率达0.94。最终完成后输出迭代总次数211，学习步长为0.002，正则化系数为0.209，将这组最优参数输入逻辑回归算法生成GS-LR预测模型。

表3 参数寻优结果

超参数	寻优范围	最优值	准确率	
			训练集	测试集
学习步长	[0.001,0.01]	0.002	0.97	0.93
	[0.01,0.1]	0.019	0.96	0.90
正则项系数	[0.001,0.01]	0.004	0.94	0.89
	[0.01,0.1]	0.083	0.95	0.90
	[0.1,1]	0.209	0.97	0.94

为了验证GS-LR预测模型的分类性能，选取了4种常用的分类模型进行对比，包括默认参数逻辑回归、KNN、

Gaussian NB和随机森林，所有模型均在相同数据集上进行训练和测试，结果如图3。可以看出，本文GS-LR模型在训练集和测试集上的平均准确率较默认参数的逻辑回归提升了5%以上，且优于其他模型，说明本文的GS寻优方式可以有效地提升模型准确率。

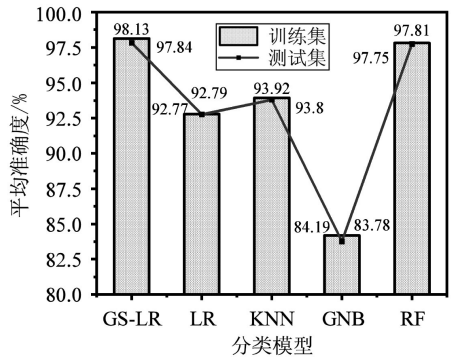

图3 各类模型平均准确率

为了进一步地展现GS-LR模型的性能，对比不同评判指标结果如图4所示。可以看出，GS-LR模型在4类指标均表现较优，分类效果较好，说明本文提出的改进GS法对逻辑回归模型优化能有效提升性能，使得模型表现更佳。

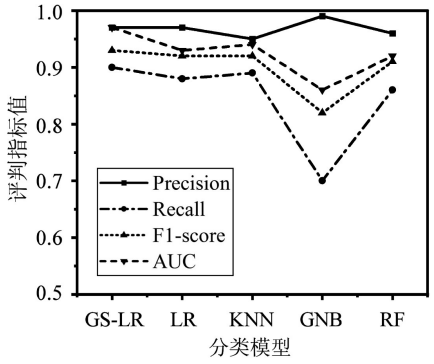

图4 不同分类模型预测结果指标

最后为了验证GS-LR模型泛化能力以及应用性，从误差和耗时上对比分析，结果如表4所示。可以看出，由于KNN和Gaussian NB都无需训练，所以时间成本相对较小，但Gaussian NB误差较大，KNN预测速度慢，两者算法代价相对较高；随机森林容易出现过拟合结果；本文GS-LR在模型构建时间上要比默认参数的逻辑回归略长，但预测误差较低，拟合效果更好，所以整体预测效率和性能更为理想。因此，本文GS-LR模型的泛化能力和应用性能均优于其他模型，具有较好的实用性。

表4 各模型预测性能评判

性能指标	本文GS-LR	逻辑回归	KNN	Gaussian NB	随机森林
均方根误差	2.705	2.926	2.861	3.959	3.012
建模耗时/s	2.779	1.802	0.000	0.000	2.029
预测时间/s	0.001	0.179	0.031	0.002	0.201

4 结束语

本文介绍了一种改进的GS法和逻辑回归相结合的配电物联协议一致性检测方法。为了验证其有效性和应用性，通过对现场投运设备历史交互数据进行处理得到数据集，利用改进的GS穷举法进行参数寻优，得出泛化能力较强的GS-LR预测模型。经测试，该模型整体预测性能和适用性均优于其他常用模型。因此，本文GS-LR模型方法适用于配电物联协议一致性检测，能够改善传统检测的自适应力和摆脱人工测试的随机性，为智能化检测提供了新思路。此外，模型构建耗时略长，这也将是后续优化重点。

参考文献：

[1] 刘冬兰，张昊，王睿，等. 面向能源互联网的智慧物联体系全场景仿真检测系统设计与应用[J]. 山东电力技术，2022，49(2)：1-6.

[2] 巫钟兴，阿辽沙·叶，刘兴奇，等. 面向对象的用电信息采集通信协议一致性测试设计[J]. 电测与仪表，2018，55(15)：65-70.

[3] 彭志强，徐春雷，张琦兵，等. 电力系统通用服务协议一致性测试技术[J]. 电力系统保护与控制，2020，48(3)：84-91.

[4] TANTIEHARANUKUL N，OSATHANUNKUL K，HANTRAKUL K，et al. MQTT-topics manage ment system for sharing of open data[C]//Proceedings of the 2017 Inter national Conference on Digital Atrs，Medic and Technology，2017：62-25.

[5] 董大兴，刘捍植，武泽，等. 基于人工智能的PCB缺陷检测系统[J]. 自动化技术与应用，2023，42(3)：129-133.

[6] HOOSHMAND A. Accurate diagnosis of prostate cancer using logistic regres sion[J]. Open Medicine，2021，16(1)：459-463.

[7] 张小秋，周超，徐晴. 基于逻辑回归的增量式异常用电行为检测方法[J]. 科学技术与工程，2019，19(29)：144-149.

[8] 严晓明. 一种逻辑回归学习率自适应调整方法[J]. 福建师范大学学报(自然科学版)，2019，35(3)：24-28.

[9] 王德贤，何先波，贺春林，等. 结合L_1和L_2正则化约束的隐语义预测模型研究[J]. 计算机工程与应用，2019，55(19)：121-127.

[10] 方琪琦，黄河. 基于多传感器信息融合的出入段全方位车辆状态检测系统[J]. 自动化技术与应用，2023，42(12)：150-154.

[11] 秦春林，石建刚，任帅. 基于大数据分析的轴承退卸工装智能化研究[J]. 自动化技术与应用，2023，42(11)：60-63.

[12] 杨婧，续婷，白艳萍，等. 基于网格搜索与支持向量机的轴承故障诊断[J]. 科学技术与工程，2021，21(22)：9360-9364.

作者简介：王立旭(1991—)，男，硕士，工程师，研究方向：配电物联网及计算机技术应用。