基于改进CatBoost的电力系统暂态稳定评估方法

杜一星[1],胡志坚[1],陈纬楠[1],王方洲[1],张翌晖[2]
(1. 武汉大学 电气与自动化学院,湖北 武汉 430072;2. 广西电网有限责任公司电力科学研究院,广西 南宁 530023)

摘要:在实际电网的运行过程中,通过同步相量测量单元实时采集到的电网动态参数通常含有部分噪声,且有时会因通信故障造成数值的随机缺失,对基于人工智能的电力系统暂态稳定评估模型造成很大影响。为此,提出一种基于改进CatBoost的暂态稳定评估方法。通过分箱算法对输入特征数据进行离散化处理,提高模型对噪声的鲁棒性;采用加权的焦点损失函数代替交叉熵损失函数,提升模型的可信度并减少模型对失稳样本的漏判;将量测数据部分缺失的样本划分到单独的节点中继续建模,从而充分挖掘不完整样本中的暂态信息。在新英格兰10机39节点上的实验结果表明,所提方法的准确率和查全率均优于其他几类机器学习算法,而且所提方法对噪声和数值缺失表现出良好的鲁棒性且具有较快的训练速度和预测速度。

关键词:机器学习;人工智能;电力系统;暂态稳定评估;集成学习;CatBoost算法
中图分类号:TM 712　　**文献标志码**:A　　DOI:10.16081/j.epae.202107026

0 引言

大区域电网之间的互联、可再生能源的不断接入以及电力电子变换装备的大规模应用等因素,使得电力系统的网络结构及运行稳定特性日趋复杂,电网的运行也愈加趋近其稳定极限,电力系统的安全稳定运行面临着严峻的考验。研究表明,大面积停电事故通常由电网暂态故障诱发[1-2]。在实际电网的动态监视中发现,故障发生后暂态过程发展十分迅速,调度人员若不能在短时间内对暂态稳定态势做出正确判断并采取相应控制措施,暂态失稳极易引发后续连锁故障从而造成大面积停电事故[2]。因此,如何提高电力系统暂态稳定评估TSA(Transient Stability Assessment)的准确性和实时性对电力系统的安全稳定运行具有重要的意义。

目前,TSA方法主要包括时域仿真法、直接法和机器学习方法。时域仿真法计算精度高但耗时长,无法满足在线评估对实时性的要求。直接法在面对复杂电网拓扑时难以构造出精确的暂态能量函数,普适性较差。随着广域量测系统和同步相量测量单元PMU(Phasor Measurement Unit)在各级电网中的规模化应用,基于数据挖掘的机器学习方法已成为TSA中新的研究热点[3]。

相比传统的TSA方法,基于数据挖掘的机器学习方法脱离了复杂的机电暂态和电磁暂态机理,无需建立具体的数学模型,其将TSA视为分类问题,离线时通过对海量样本数据的学习建立电力系统状态特征集与暂态稳定性之间的映射关系,在模型训练完成后接收各节点PMU的实时测量数据,实现在线评估。该方法具有评估精度高、预测时间短、更新灵活等优点。目前常用于TSA的机器学习模型有人工神经网络ANN(Artificial Neural Network)[4-5]、支持向量机SVM(Support Vector Machine)[6-7]、决策树DT(Decision Tree)[8-9]等,均取得了不错的评估效果。

实际中,由于环境噪声的缘故,通过PMU采集到的实时量测数据与真实值之间会存在随机的量测误差,这给TSA模型带来过拟合的风险,降低了模型的泛化性能。此外,PMU还面临着传感器损坏、数据通信传输失败等风险,在某些时刻模型可能无法接收到一些电气量的具体数值,进而导致模型评估性能下降。文献[10]引入注意力系数提高了模型的召回率,但对上述2个问题均未进行研究。文献[11]利用堆叠变分自编码器的强特征抽取能力,有效地提高了模型的抗噪能力,但其以准确率为唯一评估指标,未考虑样本不平衡问题。文献[12]将所有发电机的功角轨迹视为整体构造输入特征集,在部分发电机信息缺失时仍能保持较高的评估准确率,但其没有考虑不同节点PMU丢失电气特征量不同的情形。文献[13]提出一种基于门控循环单元的预测框架对PMU当前量测缺失值进行修复,并对其未来状态进行预测,但该方法的计算开销较大,难以满足在线评估对实时性的要求。

针对上述不足,本文提出一种基于改进CatBoost模型的电力系统TSA方法。首先,采用Symmetric树为基学习器构建高效集成学习框架,并采用分箱算法对输入特征数据进行离散化处理以提升模型对PMU测量噪声的鲁棒性;其次,使用自动处理缺失值的机制使模型在PMU采集的部分特征数值缺失时也能实现精准预测;然后,针对TSA中样本的难易不平衡及类别不平衡问题引入改进的损失函数进行

收稿日期:2020-08-18;**修回日期**:2021-05-31
基金项目:国家自然科学基金资助项目(51977156)
Project supported by the National Natural Science Foundation of China(51977156)

修正；最后，以新英格兰10机39节点系统为算例验证本文所提方法的有效性。

1 用于TSA的CatBoost模型

1.1 CatBoost模型

梯度提升决策树GBDT（Gradient Boosting Decision Tree）是一种以Boosting策略结合多棵分类与回归树CART（Classification And Regression Tree）的集成学习模型，其基本原理如下。

给定含有m个样本的训练集$\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，其中$\boldsymbol{x}_i=(x_i^1,x_i^2,\cdots,x_i^n)(i=1,2,\cdots,m)$为第$i$个样本，含有$n$维特征值，$x_i^j(i=1,2,\cdots,m;\ j=1,2,\cdots,n)$为其第$j$维特征的第$i$个样本值；$y_i(i=1,2,\cdots,m)$为第$i$个样本对应的标签，其值为0表示稳定，为1表示失稳。总迭代轮数为S，第s轮输出的强学习器为f_s，损失函数为$L(y,f_s)$。

首先按构造CART的方法初始化得到弱学习器$f_1(\boldsymbol{x}_i)$。对于迭代轮数$s=2,3,\cdots,S$，前一轮输出的强学习器损失函数为$L(y_i,f_{s-1}(\boldsymbol{x}_i))$，则本轮迭代学习的目的是从所有可能的CART集合空间H中训练出最优的弱学习器h_s，使得本轮损失函数最小，即：

$$h_s=\underset{H}{\operatorname{argmin}}\sum_{i=1}^{m}L(y_i,f_{s-1}(\boldsymbol{x}_i)+h_s(\boldsymbol{x}_i)) \tag{1}$$

式中：$h_s(\boldsymbol{x}_i)$为输入为$\boldsymbol{x}_i$时弱学习器的输出值。

GBDT算法通过求解损失函数的负梯度确定损失下降的方向，进而拟合出一棵CART，第s轮的第i个样本的损失函数负梯度为：

$$g_s^i=-\frac{\partial L(y_i,f_{s-1}(\boldsymbol{x}_i))}{\partial f_{s-1}(\boldsymbol{x}_i)} \tag{2}$$

利用$\{(\boldsymbol{x}_1,g_s^1),(\boldsymbol{x}_2,g_s^2),\cdots,(\boldsymbol{x}_m,g_s^m)\}$可以拟合出本轮的弱学习器，即得到第$s$棵CART，其对应的叶节点区域为$Q_s^j(j=1,2,\cdots,J)$，其中$J$为叶节点数。

对于任一叶节点中的样本，求出最小化损失函数，即拟合叶节点最佳输出值c_s^j为：

$$c_s^j=\underset{c}{\operatorname{argmin}}\sum_{\boldsymbol{x}_i\in Q_s^j}L(y_i,f_{s-1}(\boldsymbol{x}_i)+c) \tag{3}$$

式中：c为叶节点的输出值。求出所有叶节点最佳输出值后，即可得到第s轮迭代所求CART的拟合函数，如式(4)所示。

$$h_s(\boldsymbol{x}_i)=\sum_{j=1}^{J}c_s^jI\{\boldsymbol{x}_i\in Q_s^j\} \tag{4}$$

式中：$I\{\cdot\}$为示性函数。

从而第s轮迭代后形成的强学习器为：

$$f_s(\boldsymbol{x}_i)=f_{s-1}(\boldsymbol{x}_i)+\sum_{j=1}^{J}c_s^jI\{\boldsymbol{x}_i\in Q_s^j\} \tag{5}$$

经过S轮迭代后，最终可训练得到S棵CART，将它们结合为1个GBDT模型，如图1所示。

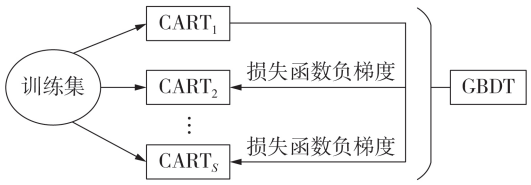

图1 GBDT训练机制

Fig.1 Training mechanism of GBDT

GBDT在绝大多数分类和回归任务中比DT表现更为出色，但仍因存在以下问题而难以在TSA中广泛应用：GBDT采用逐层或逐个建立叶节点的增长策略，最后生成的CART多为非对称树，这种结构容易过拟合，导致泛化能力较弱；预测时需要从根节点遍历整棵树进行计算，耗时较长，而在线稳定评估对实时性要求较高；GBDT在每轮迭代中的近似梯度都是采用与前一轮模型中相同的训练数据集估计得到的，这导致估计梯度在特征空间任何域中的分布与该域中梯度的真实分布存在偏差，使预测发生偏移，难以适应电网复杂多变的运行状态。

CatBoost是俄罗斯的Yandex公司于2017年开源的基于GBDT改进的高效集成学习框架[14]，其采用Symmetric树作为基学习器，该树是完全镜像二叉树，在每次迭代中，在树的整层上应用相同的分割法则，左右子树完全对称而保持平衡。CatBoost利用Symmetric树的对称结构特点，首先将所有特征进行二值化处理，并将每个叶片的索引编码为长度等于树深的二进制向量，预测时通过使用二进制特征计算评估结果。Symmetric树的对称结构使其自由度比普通的DT更小，有效抑制了模型的过拟合，而二进制向量的编码方式使模型在预测时能直接计算出对应类别的索引值，可实现并行分布式计算，大幅提升了预测速度，能够满足TSA在线应用对时效性的要求。此外，CatBoost采用排序提升算法替代GBDT算法框架中的梯度估计方法，为每个样本$\boldsymbol{x}_i$构建一个独立的集成模型M_i，M_i由除$\boldsymbol{x}_i$以外的样本数据集训练得到。CatBoost使用M_i估计损失函数在$\boldsymbol{x}_i$上的梯度，避免了信息泄露，得到了梯度的无偏估计，克服了预测偏移的问题，提升了对数据分布的拟合精度，当电网运行工况变化导致样本值变动时，模型也能表现出良好的适应能力。

1.2 数据分箱

在CatBoost模型中可引入数据分箱法进一步改善模型性能，数据分箱法是一种对连续型特征数据离散化的方法，即按照一定规则将每一维的特征划分为l个区间，得到l个箱，然后遍历该维所有数据，分别存储到对应的箱中。数据分箱大幅减少了寻找DT节点最佳切分值过程中的迭代次数，有效降低了模型的计算开销，从而加快了离线训练的速度。

当PMU采集数据掺杂噪声时，测量值通常会在真实值附近小幅随机波动。当模型对数据敏感程度

很高时，会将噪声作为有用信息进行学习以减少训练误差，从而导致建立的映射过度复杂，即模型过拟合。对任一维特征而言，分箱后会使数据粒度增加，一定区域内的数据会被归集到同一个箱中，消减了噪声造成的偏差，显著提升了模型对噪声的鲁棒性。

1.3　损失函数的改进

在机器学习模型训练过程中，其优化目标是最小化损失函数的值。电力系统TSA是一个二分类问题，在CatBoost中以二元交叉熵作为损失函数，经过p轮迭代后其表达式为：

$$L(\boldsymbol{\theta}_{p-1})=-\frac{1}{N}\sum_{i=1}^{N}[\,y_i\ln\hat{y}_i+(1-y_i)\ln(1-\hat{y}_i)\,] \tag{6}$$

式中：$\boldsymbol{\theta}_{p-1}$为由前$p-1$棵树所有参数组成的向量；$N$为样本总数；$\hat{y}_i$为前$p-1$轮迭代所得到的集成模型对第$i$个样本的预测概率输出，其取值范围为(0,1)，$\hat{y}_i>0.5$时判定为失稳，否则为稳定。定义可信度指标$R$为：

$$R=\max\left\{P\left(C_1\middle|\boldsymbol{x}_i\right),P\left(C_0\middle|\boldsymbol{x}_i\right)\right\} \tag{7}$$

式中：C_1和C_0分别表示事件模型预测样本值为1和0；$P\left(C_1\middle|\boldsymbol{x}_i\right)=\hat{y}_i$为模型预测样本失稳概率；$P\left(C_0\middle|\boldsymbol{x}_i\right)=1-\hat{y}_i$为模型预测样本稳定概率。

可信度反映了模型预测结果的可靠程度。可信度低的样本在特征向量空间中位于稳定边界附近区域，将该类样本称为难学习样本，被模型误分类的样本多为该类样本；可信度高的样本离稳定边界较远且清晰可分，将该类样本称为易学习样本，该类样本的预测结果可靠性很高。

由式(6)可以看出，易学习样本给二元交叉熵带来的损失值小于难学习样本，但在TSA问题中，易学习样本数量通常远多于难学习样本[6,10]，累积后对损失函数起主要贡献作用进而主导梯度的更新方向，使得模型在迭代过程中忽略难学习样本所包含的暂态信息而无法对其进行可靠辨识，制约了模型性能的提升。此外，在实际运行过程中，电力系统绝大多数时候都保持在稳定状态，可用于TSA模型离线训练的失稳样本数量远少于稳定样本。为了在模型训练过程中最小化损失函数，更侧重于学习数量更多的稳定样本而牺牲了对失稳样本的部分识别能力，对失稳情况的漏报会贻误调度人员采取紧急控制措施的有效时机，严重威胁电网的安全稳定性，因此在TSA中应更关注提高模型对失稳样本的识别能力。

为解决TSA问题中难、易学习样本数量不平衡问题，提高模型的可靠性，本文向CatBoost中引入焦点损失FL(Focal Loss)[15]代替二元交叉熵作为损失函数，并进一步加入权重系数α克服失稳样本和稳定样本的数量不均衡问题，以减少对失稳情况的误报。改进后的损失函数为：

$$L_{\text{fl}}(\boldsymbol{\theta}_{p-1})=-\frac{1}{N}\sum_{i=1}^{N}[\,\alpha y_i(1-\hat{y}_i)^{\gamma}\ln\hat{y}_i+(1-y_i)\hat{y}_i^{\gamma}\ln(1-\hat{y}_i)\,] \tag{8}$$

式中：γ为调节系数。

相比原来的二元交叉熵，FL函数分别在正、负样本前添加了调制因子$(1-\hat{y}_i)^{\gamma}$和$\hat{y}_i^{\gamma}$，并设置$\gamma>0$。以$y_i=1$时的失稳情况为例，模型对样本给出的失稳概率$\hat{y}_i$越大，调制因子$(1-\hat{y}_i)^{\gamma}$越小，从而降低了易学习样本损失贡献，使模型更注重于对难学习样本信息的挖掘，从而进一步提升模型总体可信度，$y_i=0$时的稳定情况同理。同时，设置权重系数$\alpha>1$，赋予失稳样本较稳定样本更高的权重，使其在迭代时获得更大的梯度从而被模型更充分地学习，减少模型对失稳样本的漏判。

2　TSA整体框架

2.1　输入特征的选择

基于机器学习的TSA工作的内核是建立输入电气特征量与故障后暂态稳定性之间的映射关系，模型评估性能的优劣很大程度上取决于输入特征的选取。因此，选取一组能准确表征电力系统实时运行状态及安全水平的输入特征集对TSA十分关键。文献[7]以系统全体发电机功角、转速等电气量的统计量为基础构建特征集，但PMU无法直接测量这些发电机机械量，间接计算将带来转换误差和通信延迟，且该方法在PMU部分量测数据丢失情况下的可行性进一步降低。文献[16]指出母线电压水平能迅速、全面地表征电力系统的动态行为，非常适用于TSA。本文选取如表1所示的电气量作为暂态特征集，所有特征数值均可通过PMU直接精确测量，无需进行二次转换，保证了原始数据的精度及数据传输的实时性。特征集维数不固定，具体由实际电力系统的规模和结构决定。

表1　暂态特征集

Table 1　Transient feature set

编号	特征含义
1	稳态时各母线电压幅值
2	稳态时各母线电压相角
3	故障发生瞬间各母线电压幅值
4	故障发生瞬间各母线电压相角
5	故障切除瞬间各母线电压幅值
6	故障切除瞬间各母线电压相角

2.2　TSA流程

本文所提的基于改进CatBoost模型的TSA方法可分为离线训练和在线预测2个阶段，具体流程如附录A图A1所示。各步骤详述如下。

1)在仿真软件中搭建实际电力系统模型，并设置不同的故障时间、故障位置和负荷水平进行批量

时域仿真，获取大量样本数据。

2）根据故障后系统的暂态稳定结果对各样本进行标注，稳定样本的标签标注为0，失稳样本的标签标注为1。本文对电力系统的同步运行稳定性问题进行研究，采用如式（9）所示的功角稳定判据进行判定。

$$\eta=360-\left|\Delta\delta_{\max}\right| \tag{9}$$

式中：$\Delta\delta_{\max}$为仿真结束时刻系统内任意2台发电机的转子最大功角差。当$\eta<0$时，判定系统失稳；否则，判定系统稳定。

3）对采集到的特征集数据进行Yeo-Johnson变换预处理：

$$\hat{x}_i^j=\begin{cases}\dfrac{(x_i^j+1)^{\beta}-1}{\beta} & \beta\neq 0,\ x_i^j\geqslant 0\\ \ln(x_i^j+1) & \beta=0,\ x_i^j\geqslant 0\\ \dfrac{(-x_i^j+1)^{2-\beta}-1}{\beta-2} & \beta\neq 2,\ x_i^j<0\\ -\ln(-x_i^j+1) & \beta=2,\ x_i^j<0\end{cases} \tag{10}$$

式中：$\hat{x}_i^j$为x_i^j变换后的值；β为变换参数，其取值通过最大似然估计法确定。

通过变换可以修正特征数据的偏度，增强其分布的正态性，从而减小特征中不可观测的误差对CatBoost模型构建的影响。

4）将所有样本按一定比例划分为训练集和测试集。训练时，采用自适应矩估计Adam（Adaptive moment estimation）[17]优化算法完成每轮迭代中学习率的自适应更新，加速算法的收敛；采取early-stopping策略，当模型精度在5轮迭代内无改善时提前终止训练。模型训练完成后，根据其在测试集上的评价指标调整超参数再次投入训练，直至达到最优效果。

5）在改进CatBoost模型训练完成后，将其投入在线应用，故障发生后接收系统内所有母线节点处PMU的实时量测数据，输出暂态稳定性预测结果。

2.3 模型性能评估指标

鉴于电力系统TSA中存在的样本类别不平衡问题及2种错误的代价程度不同，本文在评价TSA模型性能时，在准确率的基础上进一步引入查全率作为参考。

定义评估指标准确率A_c和查全率R_e如下：

$$A_c=\frac{T_s+T_u}{T_s+T_u+F_s+F_u} \tag{11}$$

$$R_e=\frac{T_u}{T_u+F_s} \tag{12}$$

式中：T_s和T_u分别为模型预测结果与实际标签相符的稳定样本数和失稳样本数；F_s为模型错判为稳定的失稳样本数；F_u为模型错判为失稳的稳定样本数。

3 算例分析

3.1 样本生成与超参数设置

本文选取新英格兰10机39节点系统作为测试系统，其拓扑如附录A图A2所示。该系统是美国新英兰地区的一个345 kV电网，包括10台发电机、39个节点、12台变压器和46条线路。本文利用PSD-BPA进行故障时域仿真，发电机模型设置为二阶模型，负荷模型采用恒阻抗模型，设置基准负荷的80%、85%、…、120%共9种负荷水平，并相应地调节发电机出力使系统各母线电压偏移保持在0.05 p.u.以内。故障类型设置为三相接地短路故障，对系统进行N-1扫描，先后在不同线路设置故障。故障持续时长分别设置为100、140、180、220 ms。故障点分别设置为线路首端以及距离线路首端为线路总长度的20%、50%、80%的位置。共生成6 624组样本，包括4 240组稳定样本和2 384组失稳样本，比例约为1.78∶1。将全体样本无序重排后，采用分层比例抽样法按0.85∶0.15的比例划分为训练集和测试集，确保训练集和测试集中失稳/稳定样本比例均与样本全集一致。

对改进CatBoost模型性能影响较大的超参数主要有树深、学习率、分箱数以及损失函数超参数α和γ。本文采用贝叶斯优化算法[18]对超参数进行分层推断，以评估准确率和查全率的几何平均值作为优化指标，最终确定的最优超参数组合为：树深为5，学习率为0.58，分箱数为254，α=3.82，γ=2。本文实验在一台Windows 10操作系统的个人电脑上运行，配置为IntelCore i7-9700 CPU/8.00 GB RAM。

3.2 模型预测性能对比

为了验证改进CatBoost模型预测性能的优越性，本文选取常用于电力系统TSA的机器学习模型及原始CatBoost模型进行对比分析。选取的模型包括个体学习器ANN、K最近邻KNN（K-Nearest Neighbor）、DT以及集成学习器随机森林RF（Random Forest）和GBDT。ANN采用3层结构，隐藏层神经元数为100，训练算法采用Adam算法；KNN经调优后取最近邻点数为7；DT、RF、GBDT均采用CART，RF集成100棵树，GBDT和CatBoost的超参数与改进CatBoost的超参数相同。测试结果如表2所示。

表2 不同模型的测试结果

Table 2 Test results of different models

模型	A_c/%	R_e/%	模型	A_c/%	R_e/%
ANN	97.18	95.47	GBDT	97.99	96.68
KNN	95.57	93.07	CatBoost	98.39	97.77
DT	95.67	93.63	改进CatBoost	99.09	99.16
RF	97.48	95.84			

由表2可以看出，集成模型的总体性能明显强于个体模型，这是由于个体模型结构单一，对复杂非线性关系的拟合能力有限，而集成模型能发挥各基学习器结构与超参数多样性的优势，充分学习到训练数据中蕴含的暂态稳定规则，因而具有更高的评估精度以及更强的泛化能力。CatBoost模型由于在GBDT模型的基础上采取了多项改进措施，其评估性能也得到了进一步提升。改进CatBoost模型的准确率最高，为99.09%，相较于CatBoost模型提高了0.7%。此外，除改进CatBoost模型以外，其他模型的查全率相较于准确率均有所降低，主要原因在于样本集中稳定样本数量远多于失稳样本，模型侧重于学习稳定样本而在一定程度上牺牲了对失稳样本的判别能力，导致模型对失稳样本的漏报多于对稳定样本的误报。而改进CatBoost模型很好地克服了该缺陷，查全率达到99.16%，相较于CatBoost模型提升了1.39%，且高于其自身总体预测准确率，这说明改进CatBoost模型对失稳样本的识别率甚至要高于稳定样本，大幅降低了对失稳样本的漏判率。总体而言，改进CatBoost模型兼具极高的预测准确率和查全率，相较于其他机器学习模型具有更卓越的TSA性能。

对于TSA问题，可信度R反映了模型预测结果的可靠程度。为进一步验证本文所提改进方法在模型可靠性方面的提升，对CatBoost模型和改进CatBoost模型在测试集上的预测结果分别进行统计并与时域仿真结果进行比较，引入指标T_s、T_u、R_{Ts}和R_{Tu}，其中指标R_{Ts}和R_{Tu}分别表示模型预测结果与实际标签相符的稳定样本和失稳样本的平均可信度，所得结果见表3。测试集总共含有994个样本，包括636个稳定样本和358个失稳样本。

表3　CatBoost模型和改进CatBoost模型的可信度对比

Table 3　Comparison of confidence between CatBoost model and improved CatBoost model

模型	T_s	T_u	R_{Ts} / %	R_{Tu} / %
CatBoost	628	350	96.57	93.58
改进CatBoost	630	355	97.46	98.12

由表3可以看出，相较于CatBoost模型，改进CatBoost模型不仅同时提高了2类样本的识别能力，而且在平均可信度方面也有显著的提升，原因在于改进CatBoost模型在训练过程中加大了对难学习样本的关注度，在特征向量空间中拟合的稳点边界可分性更高，进而提升了TSA的可靠性。

为了验证所提模型对电力系统复杂多变的运行工况的适应性，对测试集进行重新构造，以验证模型在新样本集中的泛化能力。将测试系统负荷水平分别设置为基准负荷的75%和125%，故障持续时间分别设置为150 ms和200 ms，故障点到线路首端的距离分别设置为线路总长度的30%、60%、90%。通过PSD-BPA时域仿真后得到552个新测试样本，各模型在新测试集上的评估结果如表4所示。

表4　未知工况下各模型评估结果

Table 4　Assessment results of each model under unknown operation condition

模型	A_c / %	R_e / %	模型	A_c / %	R_e / %
ANN	92.03	89.32	GBDT	95.65	92.72
KNN	92.39	87.86	CatBoost	97.10	95.15
DT	88.59	84.46	改进CatBoost	98.19	98.06
RF	94.93	89.81			

由表4可以看出，在面对未经过学习的运行工况时，各模型评估性能均有所下降。其中个体模型相较于集成模型下降更为明显，这是由于它们对稳定判别规则学习不够充分，对样本数据结构变化的适应性较差。RF、GBDT、CatBoost模型的表现相对稳定，但在新工况下的查全率指标表现较差，这说明其对新出现的失稳样本的漏判率较高。相较而言，改进CatBoost模型的表现最为稳健，在2项指标上均表现出最佳性能，说明其具有较强的泛化能力，在面对未经学习的电网运行状态及故障情况时适应性良好。

3.3　不同噪声水平下的模型性能表现

在实际电网的动态监控中发现，由于系统的动态变化等因素，通过PMU获取的量测数据与真实值之间不可避免地存在一定的随机误差。本文通过向训练集和测试集数据添加不同水平的高斯白噪声来模拟测量误差，添加方法如式(13)所示。

$$\hat{x} = x(1+\omega) \quad \omega \sim N(0, \sigma^2) \tag{13}$$

式中：x为原始样本集数据；$\hat{x}$为添加高斯白噪声后的样本集数据；ω为服从均值为0、标准差为σ的正态分布的随机变量。

采用含有高斯白噪声的样本集对上述各模型进行再次训练并测试，σ以0.01为步长，共设置0.01~0.10范围内的10种噪声水平，各模型表现如图2所示。

由图2可见，随着高斯白噪声水平的不断提升，ANN、DT、RF、GBDT模型的准确率和查全率都明显下降，尤其是DT模型的预测性能迅速降低，说明其对噪声的鲁棒性较差。KNN模型的预测性能基本不被噪声抑制，但其准确率和查全率均明显低于改进CatBoost模型。改进CatBoost模型在面对噪声干扰时的表现相对稳健，在噪声含量升高的过程中其准确率和查全率波动微小，并始终在所有模型中保持最高。当σ达到最大值0.1时，相较于RF和GBDT模型，改进CatBoost模型的准确率分别高3.47%和

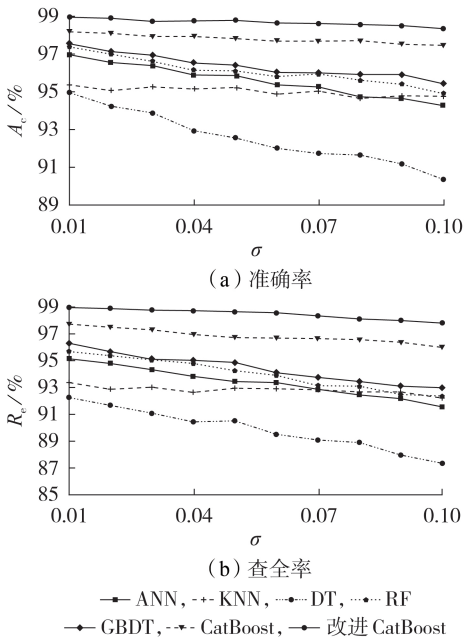

图2 噪声对各模型预测性能的影响

Fig.2 Impact of noise on prediction performance of each model

2.92%，查全率分别高5.44%和4.84%。同时，随着噪声增大，改进CatBoost模型的查全率下降幅度显著小于CatBoost模型，进一步说明了本文改进方法的实用性和有效性。总体而言，改进CatBoost模型在数据含有噪声时的预测性能相较于其他模型具有显著的优越性。

3.4 特征数值缺失对模型性能的影响

虽然PMU是高精密的测量设备，但在实际工作过程中仍然难以规避传感器故障、网络中断、动态响应延时等问题，PMU在某些时刻可能无法采集或传输部分电网参数的具体数值，造成数据的随机缺失。若电力系统TSA模型缺乏良好的缺失值处理机制，则将无法正常进行预测。此外，电网的历史数据中也存在部分电气量数值不完整的样本，这些样本数据中包含的暂态稳定信息仍具有一定价值，训练模型时若直接将其丢弃将会造成大量有用信息流失。

CatBoost对缺失值的处理机制为对其赋予特征的最小值（小于所有其他值）进行处理，确保当前特征缺失的样本都会被划分到单独的一个节点中继续进行接下来的分裂，这实际上类似于对缺失和非缺失样本分开进行建模，从而使模型能有效利用不完全样本中的数据信息。

为了研究特征数值缺失对本文所提方法的影响，分别随机抽取原始样本集中1%、5%、10%、20%的样本，对于抽出的任一样本随机选取20个特征量置为空值并放回原样本集中，对改进CatBoost模型重新进行训练和测试，所得结果如表5所示。

由表5可见，改进CatBoost模型在部分样本特征数值缺失时依然表现出较高水平的评估性能，在总样本集中不完全样本占比为20%的情况下，其准确率和查全率均能保持在97.7%以上，说明本文所提方法具有较高的容错性，离线训练时可以充分挖掘不完全样本中蕴含的暂态稳定规律，并在PMU量测数据缺失时进行可靠的在线预测。

表5 样本存在数值缺失时改进CatBoost模型的评估结果

Table 5 Assessment results of improved CatBoost model when some values of samples missed

不完全样本比例/%	A_c/%	R_e/%
1	98.59	98.89
5	98.29	98.32
10	97.99	98.04
20	97.79	97.77

3.5 模型计算速度

电力系统具有很强的时变性，暂态过程发展非常迅速，而集成模型由于结合了多个基学习器，在训练和预测时往往耗时较长，因此难以在TSA中广泛应用。CatBoost引入数据分箱后可将特征数值离散化以减少内存开销，Symmetric树的独特结构使其寻找叶节点索引时能实现并行式计算，且支持GPU实现加速训练，在模型的构建及预测快速性方面远优于其他集成学习算法。为了验证本文所提方法的快速性，将各集成学习模型基学习器总数统一设置为100，并与各单体学习器进行对比，分别测试各模型的训练时间T_t及在测试集上的总评估时间T_e，所得结果见表6。

表6 不同模型的计算耗时对比

Table 6 Comparison of calculation time consumption among different models

模型	T_t/s	T_e/s	模型	T_t/s	T_e/s
ANN	1.1942	0.0019	GBDT	12.1161	0.0079
KNN	0.0638	0.1113	CatBoost	0.4270	0.0018
DT	0.4714	0.0015	改进CatBoost	0.4244	0.0016
RF	2.4833	0.0170			

表6所示各机器学习模型中，KNN模型为惰性学习算法，基于测试样本向量与训练集各样本向量的近邻距离度量进行预测，因此训练时间开销极小，但在预测时，需逐个计算测试样本向量与训练集中所有样本向量的距离函数以筛选出最近邻值，因此评估耗时较长。除KNN外，其余模型训练时需更多的时间来拟合特征集与暂态稳定性标签之间的映射，预测耗时较少，训练时间明显长于评估时间。集成模型RF和GBDT结合了多个基分类器，相较于个体分类器模型复杂度更高，因此训练耗时和评估耗时均长于ANN和DT模型。CatBoost和改进CatBoost模型的速度相近，训练和预测的效率均远高于RF和

GBDT模型，甚至比个体模型ANN和DT具有更快的训练速度，而预测速度也与其相近。当电网拓扑结构或运行工况发生变化时，改进CatBoost模型能更快地完成模型参数更新，具有很强的适应性。故障发生后，改进CatBoost模型能快速输出预测结果，完全能够满足在线稳定评估对实时性的要求。

4　结论

本文提出一种基于改进CatBoost模型的电力系统TSA方法，并以新英格兰10机39节点系统作为算例对所提模型进行仿真分析，由实验得出以下结论。

1）相较于其他常用机器学习模型，本文改进CatBoost模型能有效地提升评估性能。针对TSA中样本的难、易学习不平衡及类别不平衡问题，向CatBoost算法引入加权的FL函数，提高了模型对难学习样本和失稳样本的重视度，进一步提升了模型的可信度并减少了对失稳样本的漏判。

2）在样本数据含有噪声以及PMU量测数据部分缺失的情况下，改进CatBoost模型依然能保持较高的准确率及查全率，表现出良好的稳健性及适应性。

3）相较于其他集成学习模型，改进CatBoost模型采用多种加速方法对算法进行优化，大幅减少了计算开销，能快速进行模型参数的更新和在线预测，在电网规模更大、特征集更高时优势将更为显著，对提高TSA的实时性具有重要工程价值。

本文所提的TSA方法是以传统的交流电网作为算例进行仿真实验，随着可再生能源接入比例的不断提高，电力系统的动态特性及暂态稳定机理将更为复杂。如何将本文所提方法应用于大规模风电并网系统的TSA，将在后续进一步研究与探讨。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] YU J J Q，HILL D J，LAM A Y S，et al. Intelligent time-adaptive transient stability assessment system[J]. IEEE Transactions on Power Systems，2018，33(1)：1049-1058.

[2] 梁志峰，葛睿，董昱，等. 印度“7.30”、“7.31”大停电事故分析及对我国电网调度运行工作的启示[J]. 电网技术，2013，37(7)：1841-1848.
LIANG Zhifeng，GE Rui，DONG Yu，et al. Analysis of large-scale blackout occurred on July 30 and July 31，2012 in India and its lessons to China's power grid dispatch and operation[J]. Power System Technology，2013，37(7)：1841-1848.

[3] 鞠平，周孝信，陈维江，等. “智能电网+”研究综述[J]. 电力自动化设备，2018，38(5)：2-11.
JU Ping，ZHOU Xiaoxin，CHEN Weijiang，et al. “Smart Grid Plus” research overview[J]. Electric Power Automation Equipment，2018，38(5)：2-11.

[4] KARAMI A. Power system transient stability margin estimation using neural networks[J]. International Journal of Electrical Power & Energy Systems，2011，33(4)：983-991.

[5] 姚德全，贾宏杰，赵帅. 基于复合神经网络的电力系统暂态稳定评估和裕度预测[J]. 电力系统自动化，2013，37(20)：41-46.
YAO Dequan，JIA Hongjie，ZHAO Shuai. Power system transient stability assessment and stability margin prediction based on compound neural network[J]. Automation of Electric Power Systems，2013，37(20)：41-46.

[6] 戴远航，陈磊，张玮灵，等. 基于多支持向量机综合的电力系统暂态稳定评估[J]. 中国电机工程学报，2016，36(5)：1173-1180.
DAI Yuanhang，CHEN Lei，ZHANG Weiling，et al. Power system transient stability assessment based on multi-support vector machines[J]. Proceedings of the CSEE，2016，36(5)：1173-1180.

[7] 刘俐，李勇，曹一家，等. 基于支持向量机和长短期记忆网络的暂态功角稳定预测方法[J]. 电力自动化设备，2020，40(2)：129-139.
LIU Li，LI Yong，CAO Yijia，et al. Transient rotor angle stability prediction method based on SVM and LSTM network[J]. Electric Power Automation Equipment，2020，40(2)：129-139.

[8] RAHMATIAN M，CHEN Y C，PALIZBAN A，et al. Transient stability assessment via decision trees and multivariate adaptive regression splines[J]. Electric Power Systems Research，2017，142：320-328.

[9] 石访，张林林，胡熊伟，等. 基于多属性决策树的电网暂态稳定规则提取方法[J]. 电工技术学报，2019，34(11)：2364-2374.
SHI Fang，ZHANG Linlin，HU Xiongwei，et al. Power system transient stability rules extraction based on multi-attribute decision tree[J]. Transactions of China Electrotechnical Society，2019，34(11)：2364-2374.

[10] 张晨宇，王慧芳，叶晓君. 基于XGBoost算法的电力系统暂态稳定评估[J]. 电力自动化设备，2019，39(3)：77-83，89.
ZHANG Chenyu，WANG Huifang，YE Xiaojun. Transient stability assessment of power system based on XGBoost algorithm[J]. Electric Power Automation Equipment，2019，39(3)：77-83，89.

[11] 王怀远，陈启凡. 基于堆叠变分自动编码器的电力系统暂态稳定评估方法[J]. 电力自动化设备，2019，39(12)：134-139.
WANG Huaiyuan，CHEN Qifan. Transient stability assessment method of electric power systems based on stacked variational auto-encoder[J]. Electric Power Automation Equipment，2019，39(12)：134-139.

[12] 李宝琴，吴俊勇，邵美阳，等. 基于集成深度置信网络的精细化电力系统暂态稳定评估[J]. 电力系统自动化，2020，44(6)：17-26.
LI Baoqin，WU Junyong，SHAO Meiyang，et al. Refined transient stability evaluation for power system based on ensemble deep belief network[J]. Automation of Electric Power Systems，2020，44(6)：17-26.

[13] YU J J Q，LAM A Y S，HILL D J，et al. Delay aware power system synchrophasor recovery and prediction framework[J]. IEEE Transactions on Smart Grid，2019，10(4)：3732-3742.

[14] HUANG G M，WU L F，MA X，et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions[J]. Journal of Hydrology，2019，574：1029-1041.

[15] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42(2)：318-327.

[16] ZHANG R，XU Y，DONG Z Y，et al. Post-disturbance transient stability assessment of power systems by a self-adaptive intelligent system[J]. IET Generation，Transmission & Distribution，2015，9(3)：296-305.

[17] 赵小强，宋昭漾. Adam优化的CNN超分辨率重建[J]. 计算机科学与探索，2019，13(5)：858-865.

ZHAO Xiaoqiang, SONG Zhaoyang. Adam optimized CNN super-resolution reconstruction[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(5): 858-865.
[18] SHAHRIARI B, SWERSKY K, WANG Z Y, et al. Taking the human out of the loop: a review of Bayesian optimization[J]. Proceedings of the IEEE, 2016, 104(1): 148-175.

作者简介：

杜一星(1997—)，男，湖南岳阳人，硕士研究生，主要研究方向为大数据和人工智能在电力系统中的应用(**E-mail**: Artemis@ whu.edu.cn)；


杜一星

胡志坚(1969—)，男，湖北荆州人，教授，博士研究生导师，博士，通信作者，主要研究方向为电力系统稳定分析与控制、新能源与分布式发电等(**E-mail**: zhijian_hu@163.com)。

(编辑 王锦秀)

Transient stability assessment method of power system based on improved CatBoost

DU Yixing[1], HU Zhijian[1], CHEN Weinan[1], WANG Fangzhou[1], ZHANG Yihui[2]
(1. School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China;
2. Electric Power Research Institute of Guangxi Power Grid Co., Ltd., Nanning 530023, China)

Abstract: In the practical operation process of power grid, the dynamic parameters of power grid collected in real time by phase measurement units usually contain some noise, and sometimes the values are randomly deletion due to communication failures, making great influence on the transient stability assessment models of power system based on artificial intelligence, for which, a transient stability assessment method based on improved CatBoost is proposed. The binning algorithm is used to discretize the input feature data for improving the robustness of model to noise. The weighted focal loss function is used for replacing cross-entropy loss function, which improves the confidence of the model and reduces the misjudgment of model to unstable samples. The samples with part of the measurement data missing are divided into separate nodes for continue modeling, thus the transient information can be fully exploited from incomplete samples. The experimental results of New England 10-generator 39-bus system show that the accuracy rate and recall rate of the proposed method are superior than other machine learning algorithms, and the proposed method performs good robustness to noise and values missing and has fast training speed and prediction speed.

Key words: machine learning; artificial intelligence; electric power systems; transient stability assessment; ensemble learning; CatBoost algorithm

附录 A：

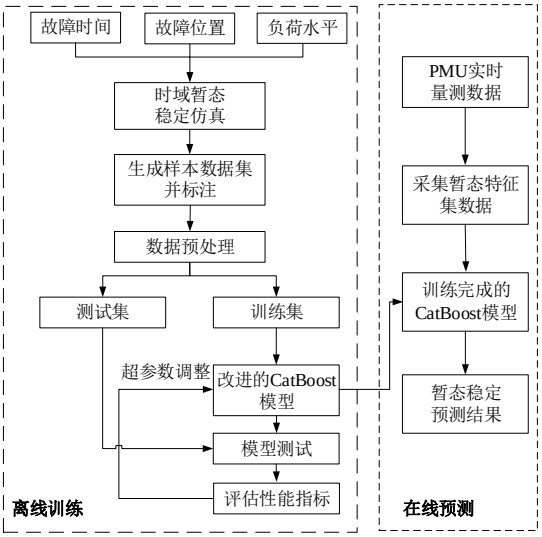

图 A1 基于改进 CatBoost 的暂态稳定评估流程图

Fig.A1 Flowchart of transient stability assessment based on modefied CatBoost

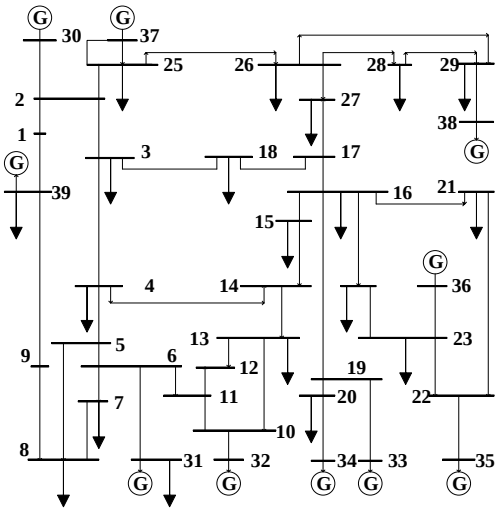

图 A2 新英格兰 10 机 39 节点系统

Fig.A2 New England 10-generator 39-bus system