(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117540277 B
(45)授权公告日 2024.06.21

(21)申请号 202311587126.9
(22)申请日 2023.11.27
(65)同一申请的已公布的文献号
申请公布号 CN 117540277 A
(43)申请公布日 2024.02.09
(73)专利权人 西南石油大学
地址 610000 四川省成都市新都区新都大道8号
(72)发明人 许成元 周杰 康毅力 郭昆 谢军 郝克桃
(74)专利代理机构 深圳峰诚志合知识产权代理有限公司 44525
专利代理师 吴林
(51)Int.Cl.
G06F 18/241(2023.01)
G06F 18/2431(2023.01)
G06F 18/213(2023.01)
G06F 18/214(2023.01)
G06F 18/2113(2023.01)
G06F 18/10(2023.01)
G06N 3/0475(2023.01)
G06N 3/094(2023.01)
G06N 3/0455(2023.01)
G06Q 10/04(2023.01)
G06Q 50/02(2024.01)

(56)对比文件
CN 110766192 A,2020.02.07
CN 116244657 A,2023.06.09

审查员 狄希

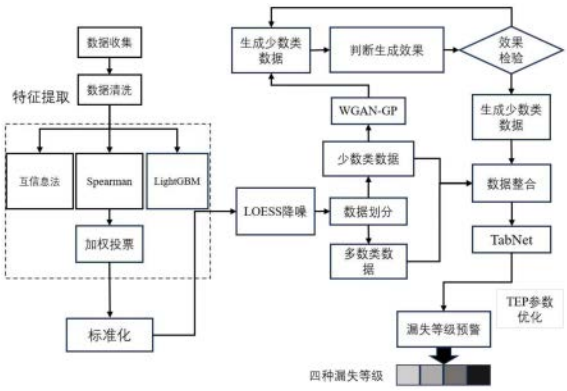
权利要求书2页 说明书8页 附图3页

(54)发明名称

一种基于WGAN-GP-TabNet算法的井漏预警方法

(57)摘要

本发明公开了一种基于WGAN-GP-TabNet算法的井漏预警方法,属于井漏预测技术领域。包括:收集现场数据;筛选与漏失流量相关性强的特征参数并删除现场数据中相关性不强的参数,从而形成初始参数;根据漏失流量对初始参数进行分级,根据分级后各级参数的数量将各等级分为多数类和少数类;将初始参数输入WGAN-GP模型,生成少数类数据;利用初始参数和生成的少数类数据来训练并评估井漏预警模型TabNet;采集现场数据并利用通过训练的TabNet模型预测漏失程度。本发明有效解决了深度学习防漏堵漏中类不平衡导致的召回率和少数类预测精度不高的问题,具有稳定可靠,准确率高,便于操作,反应速度快,可迁移性强等优势。



CN 117540277 B

1.一种基于WGAN-GP-TabNet算法的井漏预警方法,特征在于,包括以下步骤:

步骤1、从现场收集钻井工程数据,建立防漏堵漏大数据库,并对数据进行预处理;

步骤2、基于预处理后的数据中各特征参数与漏失流量的相关性筛选相关性强的特征参数,以预处理后的数据中相关性强的特征参数作为初始参数,并对初始参数进行LOESS降噪;

步骤3、根据漏失流量对LOESS降噪后的初始参数进行分级,并根据分级后各个等级中初始参数的数量对各个等级进行分类,分为多数类和少数类;

步骤4、将步骤3处理后的数据输入WGAN-GP模型,生成少数类数据;

所述WGAN-GP模型如下:生成器具有4层隐藏层,隐藏层神经元数目为256,128,64,64;判别器具有5层隐藏层,隐藏层神经元数目为256,128,64,64,32,判别器输出层为1个神经元;在生成器和判别器隐藏层后均设置Dropout层,丢弃率为0.25,判别器输出层有一个节点,用于判断输入样本的真实性,采用Adam作为优化函数;

步骤5、利用步骤2经过LOESS降噪后的初始参数以及步骤4生成的少数类数据来训练并评估井漏预警模型TabNet;

TabNet模型由多个决策步堆叠而成,每个决策步由Feature transformer和Attentive transformer、Mask层、Split层和ReLU组成;输入的样本特征如有离散型特征,TabNet首先利用训练嵌入的方式,将离散化特征映射成连续性数值特征,然后保证每个决策步数据输入形式都是B×D矩阵,其中B代表batch size的大小,D代表井漏漏失参数的维度;每个决策步的特征由上一个决策步的Attentive transformer输出,最后将决策步输出结果集成到整体决策中;

步骤6、采集现场数据并利用通过训练的TabNet模型预测漏失程度,所述现场数据中包括所述相关性强的特征参数。

2.根据权利要求1所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,根据漏失流量对LOESS降噪后的初始参数进行分级,具体的分级标准如下标所示:

如果漏失速率为0m^3/h,则漏失等级为无漏失,漏失分级为0;

如果漏失速率大于0m^3/h且小于或等于5m^3/h,则漏失等级为少量漏失,漏失分级为1;

如果漏失速率大于5m^3/h且小于或等于15m^3/h,则漏失等级为中等漏失,漏失分级为2;

如果漏失速率大于15m^3/h,则漏失等级为严重漏失,漏失分级为3。

3.根据权利要求1所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,根据分级后各个等级中初始参数的数量对各个等级进行分类具体包括:统计LOESS处理后的初始参数中各类漏失等级的数据的比例,以漏失数据比例最高的漏失等级为基准,对于所有漏失等级,如果其所占比例与比例最高的漏失等级所占比例的比值低于设定的归类阈值,则该漏失等级属于少数类,否则属于多数类。

4.根据权利要求3所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,所述设定的归类阈值为20%。

5.根据权利要求1所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,步骤2中筛选相关性强的特征参数包括如下步骤:

S1、分别利用斯皮尔曼相关系数、互信息法、LightGBM三种模型处理预处理后的数据,确定各模型中各特征参数与漏失流量的相关性;

S2、根据各特征参数与漏失流量的相关性确定不同模型中各特征参数的重要程度;

S3、综合同一特征参数在不同模型中的重要程度确定该特征参数的综合重要程度,筛选综合相关性强的特征参数。

6.根据权利要求4所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,步骤S2包括对于同一模型的特征参数按照相关性递增的顺序对各个特征参数进行排序,各个特征参数的分值等于其位次的值,步骤S3包括计算各个特征参数的总分,并根据总分排序确定各个特征参数的综合重要程度;

其中,各特征参数总分的计算式如下:

$$D_j = \sum_{i=1}^{n} f_i D_{ij}$$

式中,D_j为第j个特征参数的总分,n为表征相关性的模型的总数,f_i为第i种相关性模型的分值在总分值中的权重;D_{ij}为第i种相关性模型中第j特征参数的分值。

7.根据权利要求1所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,

步骤4包括如下步骤:

在生成器输入随机噪声时加入类别标签进行类别指导,在判别器输出端判别数据所属类别,实现按类别生成数据;

将真实样本和生成样本一同输入判别器,使其学习捕捉数据分布的特征,以便让生成样本更接近真实样本;在数据分布逐渐接近的同时,引入分等级的任务,进一步训练判别器以区分真实样本和生成样本,并确保生成样本在分类任务上也具有良好的性能;迭代过程中,利用Adam方法,用当前迭代时WGAN-GP网络中判别器的损失值、生成器的损失值依次更新WGAN-GP网络中判别器、生成器的参数直至生成器和判别器误差。

8.根据权利要求1所述的一种基于WGAN-GP-TabNet算法的井漏预警方法,其特征在于,步骤4中生成的少数类数据数量满足如下条件:经过LOESS降噪后的初始参数以及步骤4生成的少数类数据合并后,数据数量最多的等级的数据数量与少数类中各个等级的数据数量与之比均为5:1。

一种基于WGAN-GP-TabNet算法的井漏预警方法

技术领域

[0001] 本发明涉及钻井过程中井漏预测技术领域,也涉及数据驱动的深度学习技术领域,具体为一种基于WGAN-GP-TabNet算法的井漏预警方法。

背景技术

[0002] 数字信息化时代,使用计算机和自动化技术处理油气钻井中处理油气钻井中遇到的各种问题越来越成为一种趋势。面对井下渗漏问题,处理措施不当会导致封堵成功率低,钻井液持续渗漏,现场工作时间损失增加,甚至会导致弃井发生。频繁的井漏问题耗费了大量的施工时间,堵漏施工会增加钻井周期,大大增加了钻井成本,而且不能满足低成本发展的战略需要。

[0003] 基于数据驱动的机器学习或者深度学习方法似乎是一种解决方案。基于数据驱动的机器学习模型依赖于油田钻井时的所收集的各类参数,包括钻井参数、地质参数、工程参数、钻井液参数等,通过数据处理,特征提取,模型训练,模型评价等步骤,建立漏失层位的预测模型。深度学习模型在处理数据量大的情形下比机器学习模型具有优势,因为深度学习模型能够自动学习数据的高级特征表示,这意味着它们可以从原始数据中提取有用的特征,而无需手动进行特征工程。深度学习模型通常由多个层次组成,可以处理大量的参数,这使得它们能够适应大规模数据并学习复杂的关系。机器学习模型可能会在面对大规模数据时变得不够灵活或不足以捕获数据中的模式,而且深度学习模型通过多层次的表示逐渐提取数据的抽象信息。

[0004] 但是在最近年深度学习模型(如RNN,LSTM、一维CNN等)在漏失预测方面的效果不如传统机器学习模型。在钻井时发生严重漏失所占的时间比漏失不发生的时间少得多,即严重漏失类别数量远小于无漏失的情况,导致深度学习模型中预测中等和严重漏失类别的准确率以及召回率不高。

发明内容

[0005] 为了解决上述问题中,本发明提供了一种基于WGAN-GP-TabNet算法的井漏预警方法,其采用甚多学习模型TabNet来判断漏失等级,并基于WGAN-GP模型丰富了TabNet模型进行训练、测试的少数类数据,所得的WGAN-GP-TabNet模型在预测井漏上具有较好的效果。

[0006] 本发明的具体方案如下:

[0007] 一种基于WGAN-GP-TabNet算法的井漏预警方法,包括如下步骤:

[0008] 步骤1、从现场收集钻井工程数据,建立防漏堵漏大数据库,并对数据进行预处理;

[0009] 步骤2、基于预处理后的数据中各特征参数与漏失流量的相关性筛选相关性强的特征参数,以预处理后的数据中相关性强的特征参数作为初始参数,并对初始参数进行LOESS降噪;

[0010] 步骤3、根据漏失流量对LOESS降噪后的初始参数进行分级,并根据分级后各个等级中初始参数的数量对各个等级进行分类,分为多数类和少数类;

[0011] 步骤4、将步骤3处理后的数据输入WGAN-GP模型,生成少数类数据;

[0012] 步骤5、利用步骤2经过LOESS降噪后的初始参数以及步骤4生成的少数类数据来训练并评估井漏预警模型TabNet,如果训练不合格则返回步骤2-4的任一步骤修改相关参数后继续训练,训练合格则确定TabNet模型并进入下一步;

[0013] 步骤6、采集现场数据并通过TabNet模型预测其漏失程度,所述现场数据中包括所述相关性强的特征参数。

[0014] 作为本发明的一种具体实施方式,根据分级后各个等级中初始参数的数量对各个等级进行分类具体包括:统计LOESS处理后的初始参数中各类漏失等级的数据的比例,以漏失数据比例最高的漏失等级为基准,对于所有漏失等级,如果其所占比例与比例最高的漏失等级所占比例的比值低于设定的归类阈值,则该漏失等级属于少数类,否则属于多数类。

[0015] 进一步,所述设定的归类阈值为20%。

[0016] 作为本发明的一种具体实施方式,筛选相关性强的特征参数包括如下步骤:

[0017] S1、分别利用斯皮尔曼相关系数、互信息法、LightGBM三种模型处理预处理后的数据,确定各模型中各特征参数与漏失流量的相关性;

[0018] S2、根据各特征参数与漏失流量的相关性确定不同模型中各特征参数的重要程度;

[0019] S3、综合同一特征参数在不同模型中的重要程度确定该特征参数的综合重要程度,筛选综合相关性强的特征参数。

[0020] 进一步,步骤S2包括对于同一模型的特征参数按照相关性递增的顺序对各个特征参数进行排序,各个特征参数的分值等于其位次的值,步骤S3包括计算各个特征参数的总分,并根据总分排序确定各个特征参数的综合重要程度;

[0021] 其中,各特征参数总分的计算式如下:

[0022] $$D_j = \sum_{i=1}^{n} f_i D_{ij}$$

[0023] 式中,D_j为第j个特征参数的总分,n为表征相关性的模型的总数,f_i为第i种相关性模型的分值在总分值中的权重;D_{ij}为第i种相关性模型中第j特征参数的分值。

[0024] 作为本发明的一种具体实施方式,所述WGAN-GP模型如下:生成器具有4层隐藏层,隐藏层神经元数目为256,128,64,64;判别器具有5层隐藏层,隐藏层神经元数目为256,128,64,64,32,判别器输出层为1个神经元;在生成器和判别器隐藏层后均设置Dropout层,丢弃率为0.25,判别器输出层有一个节点,用于判断输入样本的真实性,采用Adam作为优化函数。

[0025] 作为本发明的一种具体实施方式,步骤4包括如下步骤:

[0026] 在生成器输入随机噪声时加入类别标签进行类别指导,在判别器输出端判别数据所属类别,实现按类别生成数据;

[0027] 将真实样本和生成样本一同输入判别器,使其学习捕捉数据分布的特征,以便让生成样本更接近真实样本;在数据分布逐渐接近的同时,引入分等级的任务,进一步训练判别器以区分真实样本和生成样本,并确保生成样本在分类任务上也具有良好的性能;迭代过程中,利用Adam方法,用当前迭代时WGAN-GP网络中判别器的损失值、生成器的损失值依次更新WGAN-GP网络中判别器、生成器的参数直至生成器和判别器误差。

[0028] 作为本发明的一种具体实施方式,步骤4中生成的少数类数据数量满足如下条件:经过LOESS降噪后的初始参数以及步骤4生成的少数类数据合并后,数据数量最多的等级的数据数量与少数类中各个等级的数据数量与之比均为5:1。

[0029] 作为本发明的一种具体实施方式,所述TabNet模型由多个决策步堆叠而成,每个决策步由Feature transformer和Attentive transformer、Mask层、Split层和ReLU组成;输入的样本特征如有离散型特征,TabNet首先利用训练嵌入的方式,将离散化特征映射成连续性数值特征,然后保证每个决策步数据输入形式都是B×D矩阵,其中B代表batch size的大小,D代表井漏漏失参数的维度;每个决策步的特征由上一个决策步的Attentive transformer输出,最后将决策步输出结果集成到整体决策中;

[0030] 与现有技术相比,具有以下优点:

[0031] (1)本发明提出了一种基于WGAN-GP-TabNet模型预测井漏的方法,该方法基于对钻井数据的深度学习进行预测,准确度高。

[0032] (2)针对钻井数据中漏失类数据不平衡情况,本发明使用生成数据模型平衡样本分布增强了数据特征。

[0033] (3)本发明提出了一种基于相关系数、互信息、LightGBM的井漏特征参数组合提取方法。相关性有线性相关性和非线性相关性,而非线性关系要比线性关系复杂的多且更加难以描述。目前,单一特征筛选方法难以准确描述出变量之间的所有相关关系,并且每一种方法衡量相关性的指标各不相同,因此我们考虑利用三种包含线性和非线性特征筛选方法的进行综合选择。

[0034] (4)本发明采取了多种方式(LOESS降噪,平衡漏失样本分布)增强了模型的鲁棒性,具有稳定可靠,准确率高,便于操作,反应速度快,可迁移性强等优势,为场工程人员根据漏失预警采取相关的堵漏措施,维护钻井人员的安全和提高钻井过程的效率有积极的影响。

附图说明

[0035] 图1是WGAN-GP-TabNet井漏预测系统流程图;

[0036] 图2是特征筛选流程图;

[0037] 图3是WGAN-GP生成器和判别器损失函数变化曲线;

[0038] 图4是TabNet结构流程图;

[0039] 图5是不同比例生成数据WGAN-GP-TabNet性能图。

具体实施方式

[0040] 下面结合实施例及附图,对本发明作进一步地的详细说明,但本发明的实施方式不限于此。

[0041] 实施例1

[0042] 图1是WGAN-GP-TabNet井漏预测系统流程图,具体实施方式将结合本图给与说明,如下所示:

[0043] 步骤1:从现场收集钻井工程数据,建立防漏堵漏大数据库,并对数据进行预处理。本实施例收集西南油气田一口井数据,其特征参数有30余种。本实施例的数据预处理即为

数据清洗,其包括缺失值处理,数据插补,异常值“3σ”法则检测剔除、数据整合、数据归一化。

[0044] 具体而言,可以用python中Pandas和NumPy两个库进行数据分析与处理,缺失值处理可用isnull()函数检测和fillna()函数填补,“3σ”法则是对于正态分布N(μ,)的数据,约有99.73%的数据落在μ正负3σ范围内,3σ范围外基本是缺失值,删除;数据清洗还包括删除空白、错误异常字符,检查数据行列结构,行列标题是否有误,是否存在重复的数组;数据归一化的计算式如下:

[0045] $$\overline{x_{ij}} = \frac{2\left(x_{ij} - \min(X_i)\right)}{\max(X_i) - \min(X_i)} - 1$$

[0046] 式中,$\overline{x_{ij}}$为第i特征第j个数据的归一化值,x_{ij}为第i特征第j个参数的值;X_i为第i特征所有参数的集合,$x_{ij} \in X_i$。

[0047] 步骤2:基于预处理后的数据中各特征参数与漏失流量的相关性筛选相关性强的特征参数,以预处理后的数据中相关性强的特征参数作为初始参数,对初始参数进行LOESS降噪。

[0048] 预处理后的数据中,相关的特征参数有30余种,周知的,各特征参数对漏失的影响是不同的,因此,有必要筛选出相关性最强的特征参数作为后期模型中的特征参数,本实施例具体采用如下方法筛选相关性强的特征参数:

[0049] S1、分别利用斯皮尔曼相关系数、互信息法、LightGBM三种模型处理预处理后的数据,确定各模型中各特征参数与漏失流量的相关性;

[0050] S2、根据各特征参数与漏失流量的相关性确定不同模型中各特征参数的重要程度,本实施例以打分法表征特征参数的重要程度,比如,对于同一模型的特征参数按照相关性递增的顺序对各个特征参数进行排序,各个特征参数的分值等于其位次的值,即第一位次的特征参数为1分,第m位次的特征参数为m分;

[0051] S3、综合同一特征参数在不同模型中的重要程度确定该特征参数的综合重要程度,筛选综合相关性强的特征参数,比如本实施中,计算各个特征参数的总分,最后根据总分排序确定各个特征参数的综合重要程度,然后对分值进行归一化(即所有特征参数的总分之和为1),筛选归一化后数值大于0.01的特征参数作为目标特征参数,最终获得16个特征因素,如表1所示。

[0052] 各特征参数总分的计算式如下:

[0053] $$D_j = \sum_{i=1}^{n} f_i D_{ij}$$

[0054] 式中,D_j为第j个特征参数的总分,n为表征相关性的模型的总数,本实施例为3,f_i为第i种相关性模型的分值在总分值中的权重,本实施例均为1/3;D_{ij}为第i种相关性模型中第j特征参数的分值。

[0055] 特征参数数值归一化的计算式如下:

[0056] $$\overline{D_j}=\frac{D_j}{\sum_{j=1}^{m}D_j}$$

[0057] 式中，$\overline{D_j}$为第j个特征参数的归一化分值，m为特征参数的总数。

[0058] 表1三种模型最终结果图

[0059]

变量名称	钻头直径	钻头转速	屈服点	平均钻速	滞后的钻井液返回深度	钻井液进流量	钻井液出流量	钻井液密度

[0060]

归一化总分	0.343	0.162	0.032	0.091	0.058	0.023	0.022	0.101
归一化总分	破裂压力	立管压力	井口转速	平均扭矩	测量深度	垂直深度	钻头压力	胶凝强度
重要度	0.038	0.035	0.012	0.015	0.014	0.012	0.017	0.015

[0061] S4：以预处理后数据中相关性强的特征参数作为初始参数，对初始参数使用LOESS算法降噪，在每个数据点附近使用一个局部的加权回归来拟合数据，从而得到更平滑的曲线或曲面，而不受全局拟合的影响；对于每个数据点x_i，LOESS使用一个权重函数$w(x_i,x)$来调整附近数据点的影响。

[0062] 步骤3：根据漏失流量对漏失程度进行分级，具体分级的数量可以根据需要选择，比如本实施确定为4级，具体分级标准如表2所示，由表可知，各个等级的数据之间是高度不平衡的，大部分数据集中在0类和1类，2类和3类一共也少于2%，不采取其他措施，只采用机器学习或者深度学习模型是难以准确且全面的识别出漏失最为严重的2类和3类。因此，统计LOESS处理后的初始参数中各类漏失等级的数据的比例，以漏失数据比例最高的漏失等级为基准，对于其余的漏失级别，如果其所占比例与比例最高的漏失等级所占比例的比值低于设定的归类阈值，则该漏失等级属于少数类，否则属于多数类，此处的归类阈值可以认为规定，但在机器学习或深度学习中，一般来说少数类为多数类的20%以下即可认为存在类不平衡问题，类不平衡问题会导致模型对少数类数据学习不够，使得少数类预测精确率和查全率不高，因此，此处，将归类阈值定为20%，其分类如表2所示。

[0063] 表2某井漏失等级分级表

[0064]

漏失等级	漏失速率（m3/h）	漏失分级	比例	类别
无漏失	0	0	88.19%	多数类
少量漏失	0<Qdl≤5	1	9.32%	少数类
中等漏失	5<Qdl≤15	2	1.67%	少数类
严重漏失（包含失返）	Qdl > 15	3	0.82%	少数类

[0065] 步骤4：将步骤3处理后的数据输入WGAN-GP模型，WGAN-GP是带有梯度惩罚项Wasserstein距离的生成对抗网络，把少量漏失、中等漏失、严重漏失类别对应的随机噪声输入生成器，生成器输出端生成相应类别数据，在判别器输出端判别数据误差，以此生成少

数类数据。

[0066] 本实施例构建WGAN-GP网络如下:其生成器具有4层隐藏层,隐藏层神经元数目为256,128,64,64,生成器输入层接受输入变量,输出层输出变量;判别器具有5层隐藏层,隐藏层神经元数目为256,128,64,64,32,判别器接受变量作为输入,判别器输出层为1个神经元,用于判断真假。此外,为了避免过拟合,在生成器和判别器隐藏层后均设置Dropout层,丢弃率为0.25,判别器输出层有一个节点,用于判断输入样本的真实性,采用Adam作为优化函数。

[0067] 利用WGAN-GP生成数据的方法如下:

[0068] (2a)在生成器输入随机噪声时加入类别标签进行类别指导,在判别器输出端判别数据所属类别(多数类或少数类),实现按类别生成数据。

[0069] (2b)将真实样本和生成样本一同输入判别器,使其学习捕捉数据分布的特征,以便让生成样本更接近真实样本。在数据分布逐渐接近的同时,引入分等级的任务,进一步训练判别器以区分真实样本和生成样本,并确保生成样本在分类任务上也具有良好的性能。

[0070] (3c)利用Adam方法,用当前迭代时WGAN-GP网络中判别器的损失值、生成器的损失值依次更新WGAN-GP网络中判别器、生成器的参数(附图3)。

[0071] 由图3可以看出,在6000轮次的训练中,在5500轮左右生成器和判别器误差达到最佳,可以提前停止训练,进行下一步:

[0072] $L_G=1-D(G(z))$

[0073] $$L(G,D)=\min_G\max_D\left\{\mathop{E}_{x\sim P_t(x)}[D(x)]-\mathop{E}_{x\sim P_t(z)}[D(G(z))]+\lambda\mathop{E}_{\hat{x}\sim P_{\hat{x}}(\hat{x})}[\nabla_{\hat{x}}D(\hat{x})_p-1]^2\right\}$$

[0074] 式中L_G为生成器的损失函数;G(·)为生成函数;z为输入噪声;D(·)为判别器函数;L(G,D)为判别器损失函数;G为生成器;D为判别器;E(·)为期望函数;x为归一化后的输入数据;P_t(·)为真实数据分布;λ为惩罚项系数;$\cdot_p$为p范数;∇为梯度算子;$\hat{x}$为真实数据和生成数据之间的随机插值,$\hat{x}=\zeta x+(1-\zeta)G(z)$,ζ服从[0,1]范围的均匀分布;$P_{\hat{x}}(\cdot)$为从真实数据分布和生成数据分布抽样点之间的均匀取样。

[0075] 步骤5:利用步骤2经过LOESS降噪后的初始参数以及步骤4生成的少数类数据来训练井漏预警模型1000轮,然后在测试集上对模型进行性能评估;如果评估合格则确定模型参数得到TabNet模型,否则,则继续训练模型。

[0076] 将生成的少数类数据和经过LOESS处理后的初始参数合并作为特征加强的数据,以特征加强过后的数据样本训练TabNet井漏预警模型训练。将数据集按照8:2比例随机划分为训练集和测试集,把数据输入TabNet模型,因TabNet模型超参数对于性能有着重要影响,使用TPE算法对模型参数进行调优,建立井漏预警模型。

[0077] 本实施例的TabNet模型如下:

[0078] (4.1)TabNet由多个决策步堆叠而成,每个决策步由Feature transformer和Attentive transformer、Mask层、Split层和ReLU组成。输入的样本特征如有离散型特征,TabNet首先利用训练嵌入的方式,将离散化特征映射成连续性数值特征,然后保证每个决策步数据输入形式都是B×D矩阵,其中B代表batch size的大小,D代表井漏漏失参数的维度。每个决策步的特征由上一个决策步的Attentive transformer输出,最后将决策步输出

结果集成到整体决策中,如附图3所示。

[0079] (4.2)Feature transformer的功能是实现了决策步的特征计算。Feature transformer由BN层、门控线性单元(GLU)层和全连接层组成,GLU的目的是在原始FC层的基础上增加一个门单元,计算如公式如下所示:

[0080] $$h(X) = (W * X + b) \oplus \sigma(V * X + c)$$

[0081] 其中,h(X)为特征变换器的输出;X为输入特征;W,b分别为全连接层的权重和偏置;*表示矩阵乘法(矩阵点积);⊕表示元素级别的异或操作;σ是sigmoid激活函数;V,c分别为GLU层的权重和偏置。特征转换层由两部分组成。该层前半部分属于共享决策步骤,每个决策步的feature transformer的共享决策步骤的参数共享,而后半部分是独立决策步骤,参数需要在每一决策步上单独训练的。

[0082] (4.3)Split层的作用是对Feature transformer输出的向量切割,计算如下所示:

[0083] $[d[i],a[i]]=f_i(M[i].f)$

[0084] 上式中d[i]表示计算模型的最终输出,a[i]表示下一个决策步的Mask层;f_i为函数,用于处理Feature Transformer输出的向量的第i个元素;M[i]为Feature Transformer模型输出的向量的第i个元素;f为Feature Transformer模型。

[0085] (4.4)Attentive transformer根据上一个决策步的输出结果获取当前决策步的Mask层矩阵,并使Mask矩阵是稀疏且不重复的。

[0086] 此外,为了克服数据类不平衡的问题,本实施例进一步探究了少数类中各等级数据生成量对模型准确度的影响,具体而言:把少数类样本数据使用WGAN-GP算法训练,A中16个井漏特征因素按照表3分组,设计8组不同比例的数据,以TabNet作为分类模型,在测试集上使用精确率,召回率,F1值和G-mean四个评价指标,寻找最佳的少数类数据生成量,结果如图5所示(图中,横坐标编号表示表3中实验编号),当特征加强的数据(将生成的少数类数据和经过LOESS处理后的初始参数合并)中少数类数据中各等级数据数量与基准数据(0泄漏等级)数量比值为5:1时,效果最好,有效克服类不平衡,因此,设定WGAN-GP-TabNet模型中生成少数类数据时,以最终特征加强的数据中少数类数据中各等级数据与基准等级数据量比值为5:1为准。

[0087] 表3加入WGAN-GP生成数据后的数据集

[0088]

实验编号	0类	1类	2类	3类	生成样本数量1、2、3类	多数类/少数类各等级的比值		
Origin	13100	1384	247	121	0、0、0	9：1	53：1	108：1
WGAN-GP1	13100	1384	247	121	1236、189、42	5：1	30：1	80：1
WGAN-GP2	13100	1384	247	121	1236、1062、140	5：1	10：1	50：1
WGAN-GP3	13100	1384	247	121	1236、2372、311	5：1	5：1	30：1
WGAN-GP4	13100	1384	247	121	1236、2372、1188	5：1	5：1	10：1
WGAN-GP5	13100	1384	247	121	1236、2372、2499	5：1	5：1	5：1
WGAN-GP6	13100	1384	247	121	2982、4119、4245	3：1	3：1	3：1
WGAN-GP7	13100	1384	247	121	11715、12853、12979	1：1	1：1	1：1

[0089] 步骤:6:将数据进行步骤1相同数据处理并提取16个已筛选好的井漏特征因素,输入到训练好的TabNet模型,就能够预测漏失程度等级。

[0090] 本发明使用模型测试15个不同深度的样本数据,间隔取样,从数据开始记录的500m开始,750m、1000m、1250m、1500m、1750m、2000m、2250m、2500m、2750m、3000m、3250m、3500m预测结果如下表所示。

[0091] 表4漏失预测结果表

[0092]

深度(*10m)	50	75	100	125	150	175	200	225	250	275	300	325	350
真实值	0	0	0	0	0	1	2	0	3	1	1	0	0
预测值	0	0	0	0	0	1	2	0	2	1	1	0	0

[0093] 以上所述,仅为本发明较佳的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明实施例揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应该以权利要求的保护范围为准。

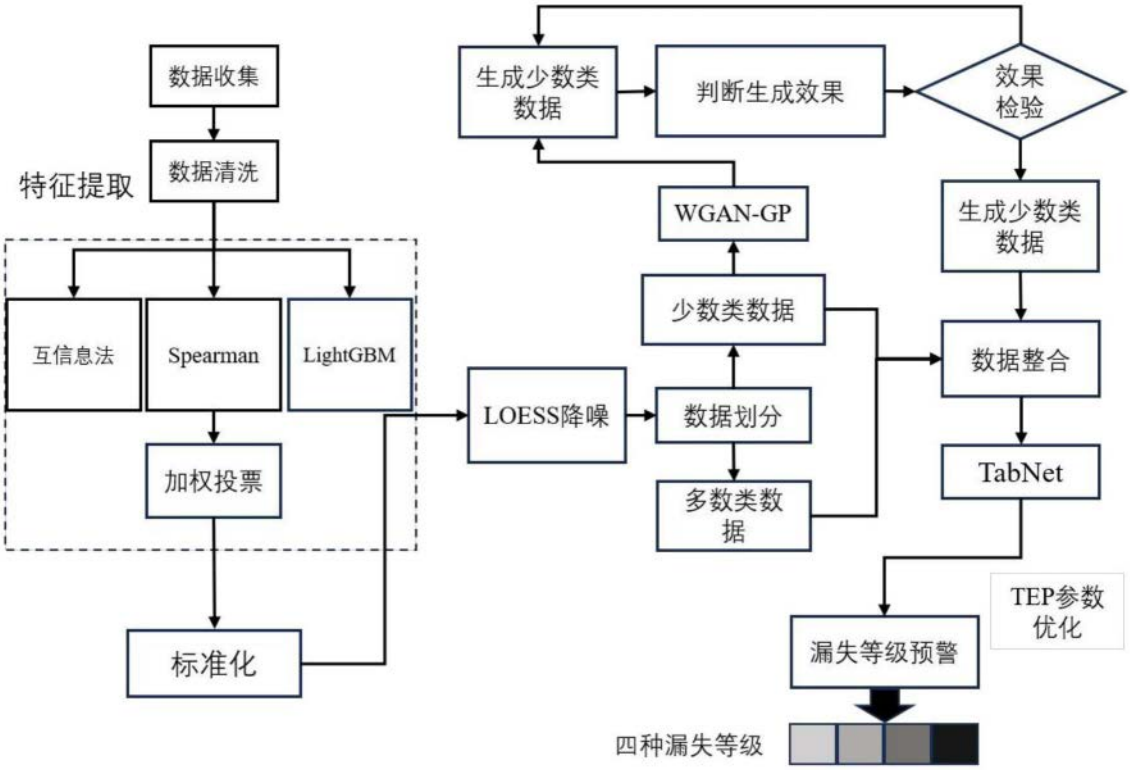

数据收集
数据清洗
特征提取
互信息法
Spearman
LightGBM
加权投票
标准化
LOESS降噪
生成少数类数据
判断生成效果
效果检验
WGAN-GP
少数类数据
数据划分
多数类数据
生成少数类数据
数据整合
TabNet
TEP参数优化
漏失等级预警
四种漏失等级

图1

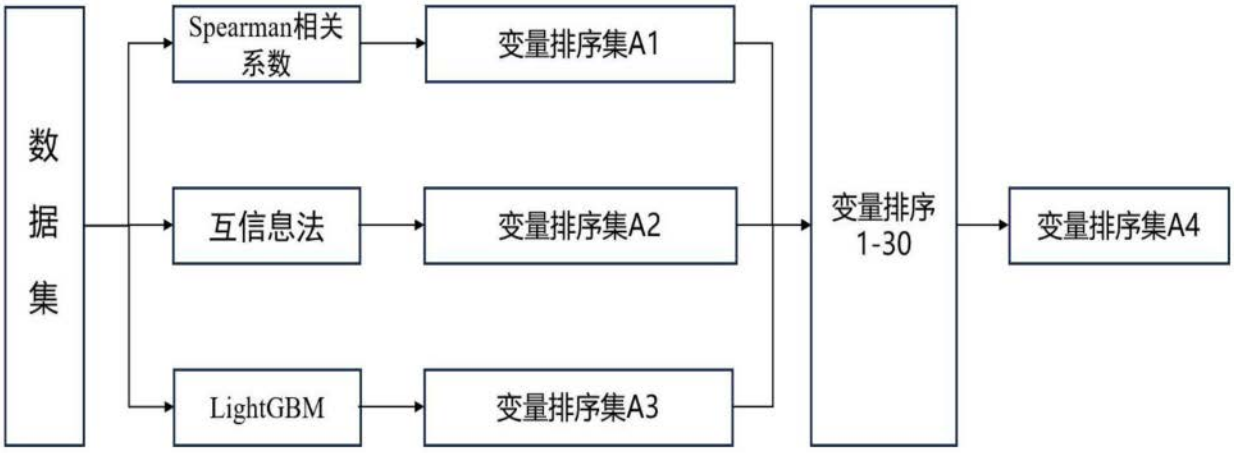

数据集
Spearman相关系数
互信息法
LightGBM
变量排序集A1
变量排序集A2
变量排序集A3
变量排序1-30
变量排序集A4

图2

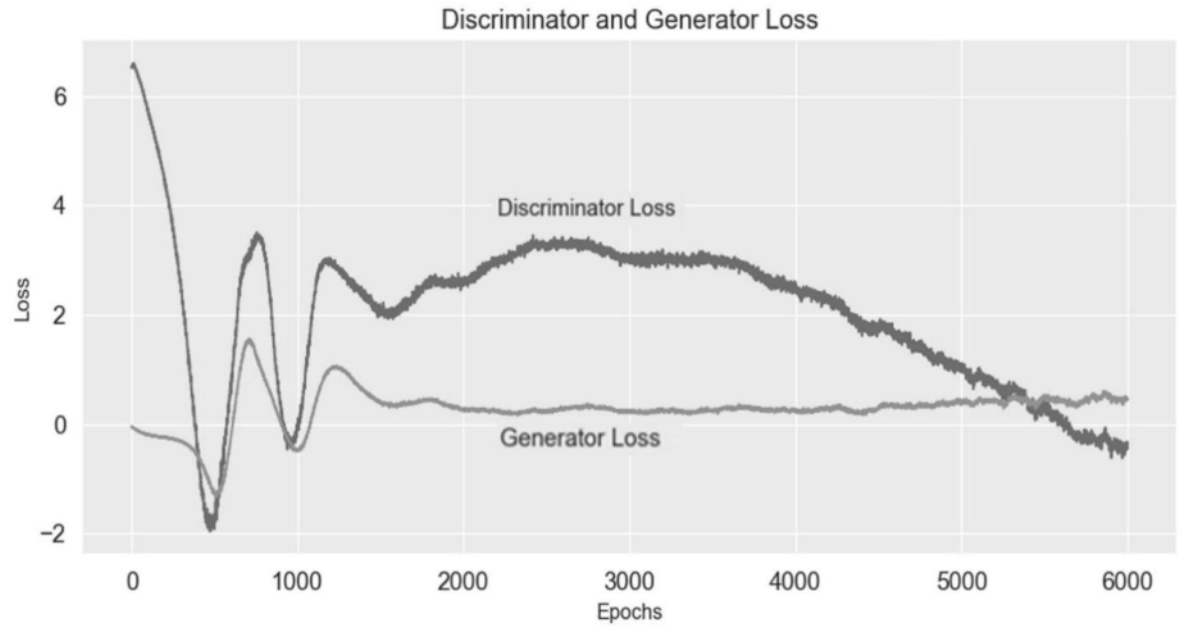

Discriminator and Generator Loss
Discriminator Loss
Generator Loss
Loss
Epochs
6
4
2
0
-2
0
1000
2000
3000
4000
5000
6000

图3

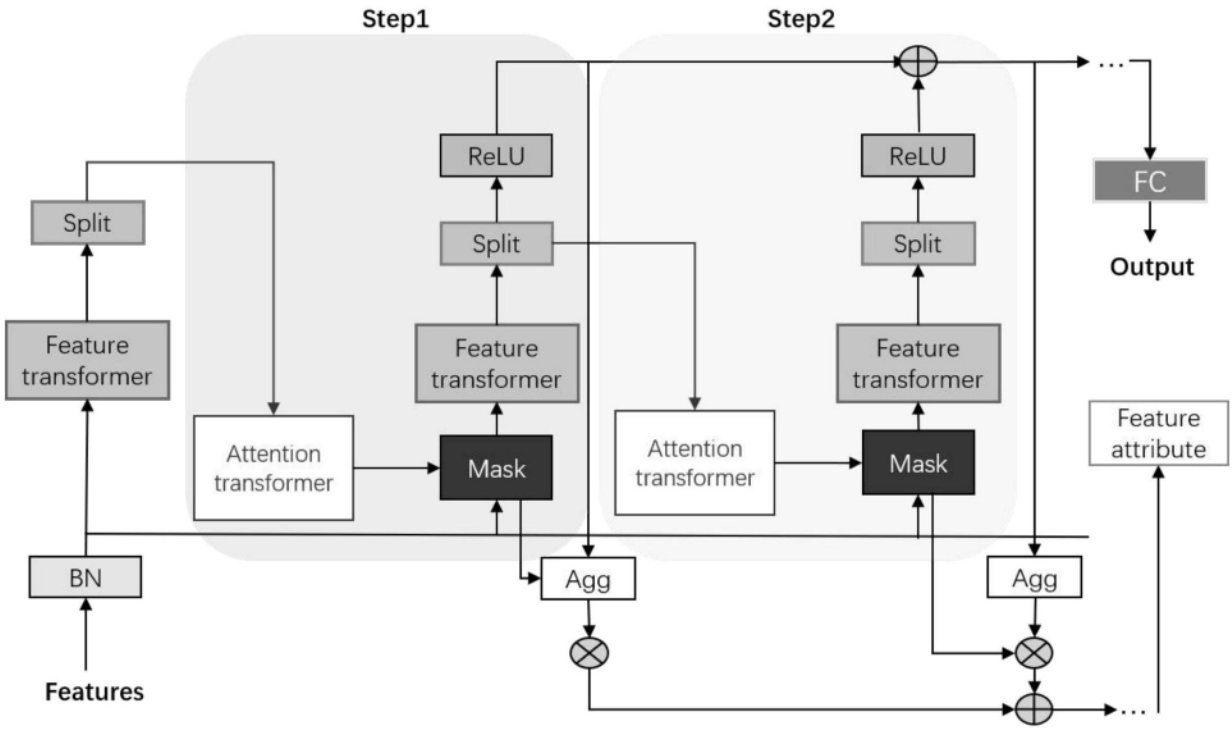

Step1
Step2
Split
Feature transformer
BN
Features
Attention transformer
ReLU
Split
Feature transformer
Mask
Agg
Attention transformer
ReLU
Split
Feature transformer
Mask
Agg
FC
Output
Feature attribute

图4

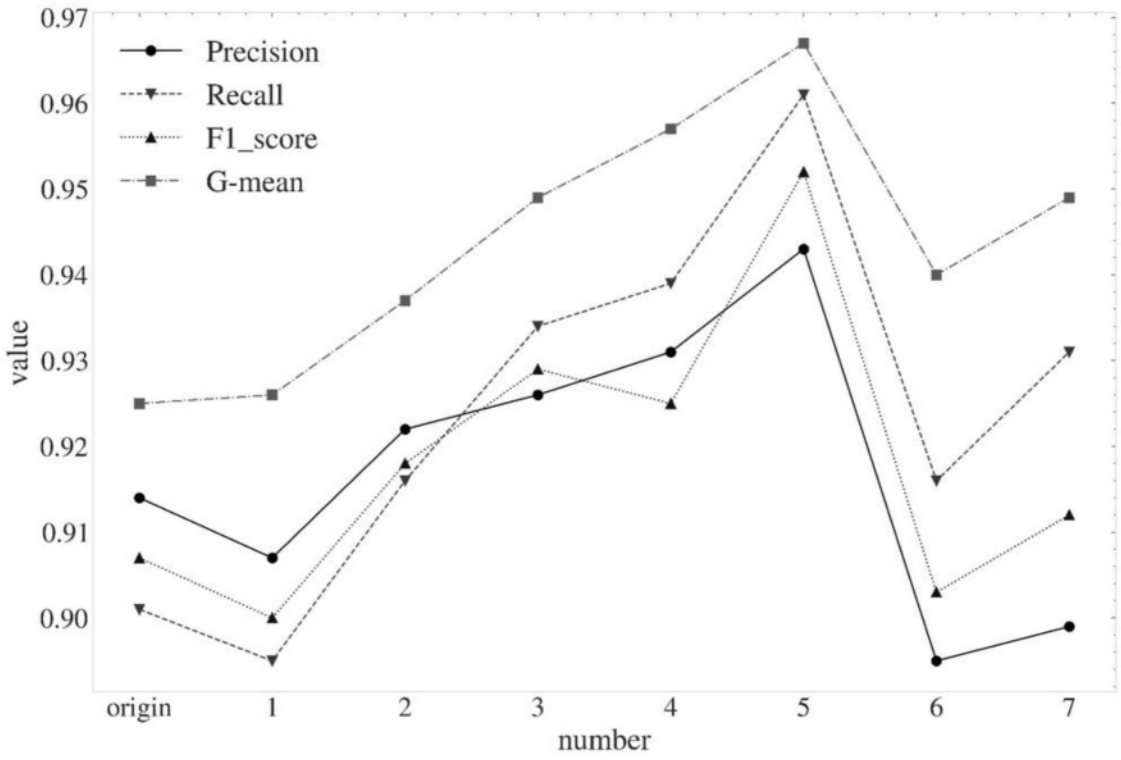

Precision
Recall
F1_score
G-mean
value
number
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
origin
1
2
3
4
5
6
7

图5