(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116822971 A
(43)申请公布日 2023.09.29

(21)申请号 202311104072.6

(22)申请日 2023.08.30

(71)申请人 长江大学武汉校区
地址 430000 湖北省武汉市蔡甸区大学路111号

(72)发明人 杨明合 许楷 李博志 蔡旭龙 何清旖

(74)专利代理机构 武汉蓝宝石专利代理事务所(特殊普通合伙) 42242
专利代理师 范三霞

(51)Int.Cl.
G06Q 10/0635(2023.01)
G06Q 50/02(2012.01)
G06N 3/08(2023.01)

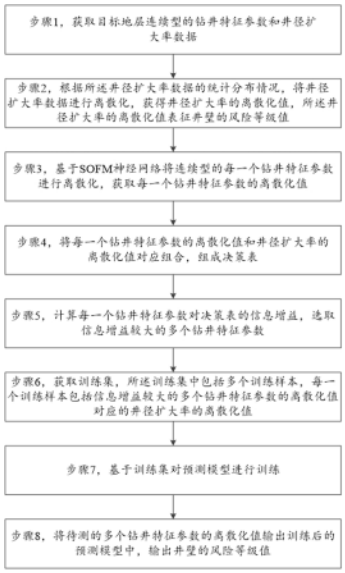
权利要求书4页 说明书11页 附图2页

(54)发明名称

一种井壁风险等级预测方法

(57)摘要

本发明提供一种井壁风险等级预测方法,通过SOFM神经网络对连续型钻井特征参数进行离散,无需预先标记训练数据,通过调整权重向量和邻域来适应输入数据的分布和特点,最大程度保留原始数据的信息;以及通过属性重要性中的信息增益来提取特征参数,显著减少了模型输入参数的数量并提高了预测效果。



CN 116822971 A

1.一种井壁风险等级预测方法,其特征在于,包括:

获取目标地层连续型的钻井特征参数和井径扩大率数据;

根据所述井径扩大率数据的统计分布情况,将所述井径扩大率数据进行离散化,获得井径扩大率的离散化值,所述井径扩大率的离散化值表征井壁的风险等级值;

基于自组织特征映射SOFM神经网络将连续型的每一个钻井特征参数进行离散化,获取每一个钻井特征参数的离散化值;

将每一个钻井特征参数的离散化值和井径扩大率的离散化值对应组合,组成决策表;

计算每一个钻井特征参数对决策表的信息增益,选取信息增益较大的多个钻井特征参数;

获取训练集,所述训练集中包括多个训练样本,每一个训练样本包括信息增益较大的多个钻井特征参数的离散化值和对应的井径扩大率的离散化值;

基于训练集对预测模型进行训练;

将待测的多个钻井特征参数的离散化值输出训练后的预测模型中,输出井壁的风险等级值。

2.根据权利要求1所述的井壁风险等级预测方法,其特征在于,根据所述井径扩大率数据的统计分布情况,将所述井径扩大率数据进行离散化,获得井径扩大率的离散化值,所述井径扩大率的离散化值表征井壁的风险等级值,包括:

根据连续型的井径扩大率的统计分布情况,将连续型的井径扩大率按照等距离划分为多个区间,对每一个区间按照不同风险等级编码赋值,获取每一个区间的井径扩大率对应的风险等级编码值,即为离散化值。

3.根据权利要求1所述的井壁风险等级预测方法,其特征在于,所述基于自组织特征映射SOFM神经网络将连续型的每一个钻井特征参数进行离散化,获取每一个钻井特征参数的离散化值,包括:

A、初始化每一个钻井特征参数对应的离散化值的数目H_0、初始化SOFM神经网络的学习率$\alpha(0)$和邻域$N_c(0)$、设置学习次数T以及初始化每一个钻井特征参数的每一个离散化值的权值向量w_j,其中,j为离散化值编号,j=1,2...,H_0;

B、对于任一个钻井特征参数包括的连续的多个采样点数据,计算任一个采样点数据与每一个离散化值的权值向量w_j之间的距离,并获取最小距离的权值向量w_m;

C、在下一次迭代过程中,对最小距离的权值向量w_m以及SOFM神经网络的学习率和邻域进行迭代更新;

D、重复执行A、B和C,直到迭代更新的次数达到设置的学习次数T,获取与所述任一个采样点数据之间距离最小的权值向量w_m',则权值向量w_m'对应的离散化值为所述任一个采样点数据的离散化值;

E、对每一个钻井特征参数的每一个采样点数据,均执行A、B、C和D,得到每一个钻井特征参数的每一个采样点数据的初始离散化值。

4.根据权利要求3所述的井壁风险等级预测方法,其特征在于,所述B中,对于任一个钻井特征参数包括的连续的多个采样点数据,计算任一个采样点数据与每一个离散化值的权值向量w_j之间的距离,并获取最小距离的权值向量w_m,包括:

计算所述任一个钻井特征参数中采样点数据x_i与各个离散化值的权重向量w_j之间的距离d_{ij}:

$$d_{ij}=\left[\sum_{j=1}^{H_0}(x_i-w_j)^2\right]^{\frac{1}{2}};$$

其中,所述任一个钻井特征参数包括采样点数据($x_1,x_2,...,x_i,...,x_n$),n为采样点个数,j=1,2,...,H_0,i=1,2,...,n;

获取d_{ij}中最小距离d_{im},并获取对应的权值向量w_m,m=1,2,...,H_0。

5.根据权利要求4所述的井壁风险等级预测方法,其特征在于,

所述C、在下一次迭代过程中,对最小距离的权值向量w_m以及SOFM神经网络的学习率和邻域进行迭代更新,包括:

$$w_m(t+1)=\begin{cases}w_m(t)+\alpha(t)[x_i-w_m(t)], & m\in N_c(t)\\ w_m(t), & m\notin N_c(t)\end{cases};$$

其中,

$$\alpha(t)=\alpha(0)(1-t/T);$$

$$N_c(t)=\mathrm{int}\{N_c(0)(1-t/T)\};$$

其中,t为当前迭代次数。

6.根据权利要求3所述的井壁风险等级预测方法,其特征在于,所述E之后还包括:

F、基于每一个钻井特征参数的每一个采样点数据的初始离散化值,确定每一个钻井特征参数对应的离散化值的最优数目;

G、基于每一个钻井特征参数对应的离散化值的最优数目更新A中初始化的每一个钻井特征参数对应的离散化值的数目H_0,执行A、B、C和D,得到每一个钻井特征参数的每一个采样点数据的离散化值。

7.根据权利要求6所述的井壁风险等级预测方法,其特征在于,所述F、基于每一个钻井特征参数的每一个采样点数据的初始离散化值,确定每一个钻井特征参数对应的离散化值的最优数目,包括:

根据初始离散化值确定每个钻井特征参数对井壁风险等级的依赖度γ;

确定依赖度最大值γ_{max}对应的钻井特征参数X_p,并通过迭代计算钻井特征参数X_p对应的离散化值的最优数量H_1;

按照依赖度γ从大到小的顺序依次对H_1减1确定每一个钻井特征参数的离散化值的最优数目。

8.根据权利要求7所述的井壁风险等级预测方法,其特征在于,所述确定依赖度最大值γ_{max}对应的钻井特征参数X_p,并通过迭代计算钻井特征参数X_p对应的离散化值的最优数量H_1,包括:

A',基于每一个钻井特征参数对应的离散化值的初始数目H_0,生成钻井特征参数X_p的H_0个初始聚类中心;

B’，设钻井特征参数$X_p=(x_{p1},x_{p1},...,x_{pn})$，n为采样点的个数，计算采样点数据$x_{pn}$到每一个聚类中心的距离，获取与采样点数据$x_{pn}$的距离最近的聚类中心，将采样点数据$x_{pn}$归类到该簇中；

C’，遍历每一个采样点数据，将每一个采样点数据归类到不同的簇中；

D’，基于每一个簇中的所有采样点数据，重新计算簇中心，利用簇中心更新初始聚类中心；

E’，基于更新后的聚类中心，计算所有簇中所有采样点数据与对应聚类中心的误差平方和SSE；

F’，在下一次迭代过程中，更新钻井特征参数对应的离散化值的初始数目H_0，并执行A’、B’、C’、D’和E’，计算每一次迭代后的误差平方和SSE，直到计算出的误差平方和SSE不发生变化时，获取此次迭代后的钻井特征参数X_p对应的离散化值的最优数目H_1。

9.根据权利要求1所述的井壁风险等级预测方法，其特征在于，所述计算每一个钻井特征参数对决策表的信息增益，包括：

钻井特征参数X_h对决策表的信息增益为：

$$Gain(X_h)=Entropy(L)-Entropy(X_h,L)$$；

式中，$Gain(X_h)$表示钻井特征参数X_h对决策表L的信息增益；

其中，决策表L的信息熵Entropy(L)为：

$$Entropy(L)=-\sum\left(p(a)*\log_2(p(a))\right)$$；

式中，p(a)为决策表L中井径扩大率的离散化值a的概率；

钻井特征参数X_h对Entropy(L)的条件期望为：

$$Entropy(X_h,L)=\sum\left((|L_V|/|L|)*Entropy(L_V)\right)$$；

式中，L_v为在钻井参数X_h上离散化值为v的子集，$|L_v|$为子集L_v的采样点数，|L|为决策表L的样本总数。

10.根据权利要求1所述的井壁风险等级预测方法，其特征在于，所述获取训练集，包括：

获取训练集$S=[(S_1,\ y_1),\ (S_2,\ y_2),\ \ldots,\ (S_{M'},\ y_{M'})]$，其中包括M’个训练样本，其中训练样本$S_{M'}$中包括信息增益较大的g个钻井特征参数，$y_{M'}$为井壁风险等级；

所述基于训练集对预测模型进行训练，包括：

基于训练集训练CatBooat模型建立井壁风险等级预测模型，其中，通过贝叶斯算法优化目标函数$L(y,\hat{y})$寻找CatBoost模型的最佳超参数集，优化目标函数$L(y,\hat{y})$的表达式为：

$$L(y,\hat{y})=-\frac{1}{N}\sum_{i=1}^{N}y_{i\Lambda}\cdot\log p_{i\Lambda}+\frac{\lambda}{2}\sum_{j=1}^{T}\left\|w_j\right\|^2$$；

式中，N为训练样本数量，y为井壁实际风险等级，$\hat{y}$为模型输出的井壁预测风险等级，

$y_{i\Lambda}$ 为样本S_i是否属于第 Λ 类，$p_{i\Lambda}$ 为样本S_i属于第 Λ 类的概率，j是第i个样本的实际类别标签，T为井壁风险等级的最大类别数，λ 是正则化系数。

一种井壁风险等级预测方法

技术领域

[0001] 本发明涉及石油勘探领域，更具体地，涉及一种井壁风险等级预测方法。

背景技术

[0002] 在石油勘探开发过程中，保持井壁的稳定状态对于安全钻井和生产具有至关重要的意义。在实际中，由于地层复杂性、钻头磨损等因素的影响，往往会出现井壁失稳等问题，它不仅会影响钻井速度还会造成严重的安全事故。对井壁稳定性进行预测和分析，可以更好地制定钻井方案，从而降低钻井风险并提高采油效率。

[0003] 井径扩大率是判断井壁失稳比较直观的参数，当井壁失稳时，井眼周围的岩石会发生塑性变形或破裂，导致井径扩大率急剧增加。因此，通过预测井径扩大率的变化来分析井壁的失稳情况。

[0004] 在预测井壁稳定性的方法中，经验公式法易受限于地质条件和岩石类型，无法考虑到不同地区和层位之间的差异；数值模拟方法需要大量的参数和复杂的计算，其结果还会受模型误差的影响。在深部地层中，影响井壁稳定的参数较多，且各参数之间存在复杂的非线性关系，而机器学习方法解决多目标非线性复杂问题的效果较好，可以充分发掘钻井数据与井壁稳定性之间隐藏的潜在关系，实现对井壁稳定性的准确预测。

发明内容

[0005] 本发明针对现有技术中存在的技术问题，提供一种井壁风险等级预测方法，解决现有技术中井壁稳定性预测存在的地质条件限制和模型误差的问题，该预测方法包括：

[0006] 获取目标地层连续型的钻井特征参数和井径扩大率数据；

[0007] 根据所述井径扩大率数据的统计分布情况，将所述井径扩大率数据进行离散化，获得井径扩大率的离散化值，所述井径扩大率的离散化值表征井壁的风险等级值；

[0008] 基于自组织特征映射SOFM神经网络将连续型的每一个钻井特征参数进行离散化，获取每一个钻井特征参数的离散化值；

[0009] 将每一个钻井特征参数的离散化值和井径扩大率的离散化值对应组合，组成决策表；

[0010] 计算每一个钻井特征参数对决策表的信息增益，选取信息增益较大的多个钻井特征参数；

[0011] 获取训练集，所述训练集中包括多个训练样本，每一个训练样本包括信息增益较大的多个钻井特征参数的离散化值和对应的井径扩大率的离散化值；

[0012] 基于训练集对预测模型进行训练；

[0013] 将待测的多个钻井特征参数的离散化值输出训练后的预测模型中，输出井壁的风险等级值。

[0014] 本发明提供的一种井壁风险等级预测方法，通过SOFM神经网络对连续型钻井特征参数进行离散，无需预先标记训练数据，通过调整权重向量和邻域来适应输入数据的

分布和特点，最大程度保留原始数据的信息；以及通过属性重要性中的信息增益来提取特征参数，显著减少了模型输入参数的数量并提高了预测效果。

附图说明

[0015] 图1为本发明提供的一种井壁风险等级预测方法的流程示意图；

[0016] 图2为对预测模型进行训练的流程示意图。

具体实施方式

[0017] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。另外，本发明提供的各个实施例或单个实施例中的技术特征可以相互任意结合，以形成可行的技术方案，这种结合不受步骤先后次序和/或结构组成模式的约束，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时，应当认为这种技术方案的结合不存在，也不在本发明要求的保护范围之内。

[0018] 图1为本发明提供的一种井壁风险等级预测方法流程图，如图1所示，方法包括：

[0019] 步骤1，获取目标地层连续型的钻井特征参数和井径扩大率数据。

[0020] 可理解的是，获取的目标地层的钻井特征参数和井径扩大率数据均为连续型的数据，每一个钻井特征参数包括n个采样点数据，组成连续型的钻井特征参数，同样的，井径扩大率数据也包括n个采样点数据。本发明中获取的钻井特征参数包括14个。

[0021] 对所有的钻井特征参数进行预处理，首先，对所有钻井特征参数，剔除异常值和空缺值。

[0022] 其次，对钻井特征参数进行归一化预处理，采用最大、最小标准化法将特征数据的原始值映射到区间[0,1]之间。

[0023] $$X_i = \frac{X_i' - X'_{min}}{X'_{max} - X'_{min}}$$；

[0024] 式中，X'_i为钻井特征参数；X'_{max}和X'_{min}分别为所有钻井特征参数中的最大值和最小值；X_i为归一化后的钻井特征参数。

[0025] 步骤2，根据所述井径扩大率数据的统计分布情况，将所述井径扩大率数据进行离散化，获得井径扩大率的离散化值，所述井径扩大率的离散化值表征井壁的风险等级值。

[0026] 可理解的是，获取到连续的井径扩大率数据后，对根据连续型的井径扩大率的统计分布情况，将连续型的井径扩大率按照等距离划分为多个区间，对每一个区间按照不同风险等级编码赋值，获取每一个区间的井径扩大率对应的风险等级编码值，即为离散化值。比如，井径扩大率的100个采样点数据为1,2,3,....,100，将其划分为10个区间，1~10为一个区间，11~20一个区间，....,91~100一个区间，那么在1~10之内的采样点数据均离散化为1，在11~20区间的采样点数据被离散化为2，以此类推，得到井径扩大率的每一个采样点数据的离散化值，其中，离散化值代表井壁风险等级。

[0027] 步骤3,基于自组织特征映射SOFM神经网络将连续型的每一个钻井特征参数进行离散化,获取每一个钻井特征参数的离散化值。

[0028] 可理解的是,步骤2对井径扩大率数据进行了离散化,该步骤对每一个钻井特征参数进行离散化。其中,影响井壁稳定性的连续型变量有井斜角、方位角、狗腿度、钻井液密度、漏斗粘度、钻速、泵压、钻压、转速、排量、立管压力、钻头压降、环空压耗和喷射速度共14个钻井参数,对连续型变量的离散化按(1)-(2)步骤进行。

[0029] (1)初始离散。

[0030] 作为实施例,所述基于SOFM神经网络将连续型的每一个钻井特征参数进行离散化,获取每一个钻井特征参数的离散化值,包括:

[0031] A、初始化每一个钻井特征参数对应的离散化值的数目H_0、初始化SOFM神经网络的学习率$\alpha(0)$和邻域$N_c(0)$、设置学习次数T以及初始化每一个钻井特征参数的每一个离散化值的权值向量w_j,其中,j为离散化值编号,j=1,2...,H_0。

[0032] 可理解的是,将每一个钻井特征参数的离散化值的数目均初始化为H_0,设置SOFM神经网络的初始学习率$\alpha(0)$和初始邻域$N_c(0)$以及学习次数T,对每个离散化值的权值向量w_j赋予[0,1]范围内的随机值作为权值向量的初始值。令$X=\left(x_1,x_2,\ldots,x_i,\ldots,x_n\right)$,$\left(x_1,x_2,\ldots,x_i,\ldots,x_n\right)$为每一个钻井特征参数X中的采样点数据,n为采样点个数。

[0033] 对各个参数初始化后,对于每一个钻井特征参数均执行后续的步骤B。

[0034] B、对于任一个钻井特征参数包括的连续的多个采样点数据,计算任一个采样点数据与每一个离散化值的权值向量w_j之间的距离,并获取最小距离的权值向量w_m。

[0035] 作为实施例,所述B中,对于任一个钻井特征参数包括的连续的多个采样点数据,计算任一个采样点数据与每一个离散化值的权值向量w_j之间的距离,并获取最小距离的权值向量w_m,包括:

[0036] 对于每一个钻井特征参数,每个离散化值的归一化权值向量w_j为:

$$w_j=\frac{(\mu_{1j},\mu_{2j},\ldots,\mu_{ej})^T}{\|w_j\|},\quad (j=1,2,\ldots,H_0);$$

[0037] 式中,μ_{ej}为$SOFM$神经元权值,e为神经元个数,H_0为离散化值的初始数目。

[0038] 计算所述任一个钻井特征参数中采样点数据x_i与各个离散化值的权重向量w_j之间的距离d_{ij}:

[0039] $$d_{ij}=\left[\sum_{j=1}^{H_0}(x_i-w_j)^2\right]^{\frac{1}{2}};$$

[0040] 其中,所述任一个钻井特征参数中包括采样点数据$(x_1,x_2,\ldots,x_i\ldots,x_n)$,n为采样点个数,j=1,2,...,$H_0$,i=1,2,...,n;

[0041] 获取d_{ij}中最小距离d_{im},并获取对应的权值向量w_m,m=1,2,...,H_0。

[0042] 可理解的是,对于每一个钻井特征参数中的每一个采样点数据,需要确定采样点数据对应的离散化值。在步骤1中初始化的离散化值的数量为H_0,其对应的权值向量也包括

H_0个，也就是每一个离散化值的权值向量w_j，j=1,2,...,H_0。

[0043] 计算采样点数据与每一个权值向量之间的距离，获取与采样点数据距离最小的权值向量w_m。

[0044] C、在下一次迭代过程中，对最小距离的权值向量w_m进行迭代更新，其它的权值向量不变，以及对SOFM神经网络的学习率和邻域进行迭代更新。

[0045] 可理解的是，本发明是通过迭代求解每一个采样点数据对应的离散化值的，因此，在下一次迭代的过程中，对上一次迭代过程中产生的距离最小的权值向量进行更新，并且对SOFM神经网络的学习率和邻域进行更新。

[0046] 所述C、在下一次迭代过程中，对最小距离的权值向量w_m以及SOFM神经网络的学习率和邻域进行迭代更新，包括：

[0047] $$w_m(t+1)=\begin{cases} w_m(t)+\alpha(t)[x_i-w_m(t)], & m\in N_c(t) \\ w_m(t), & m\notin N_c(t) \end{cases};$$

[0048] 其中，

[0049] $$\alpha(t)=\alpha(0)(1-t/T);$$

[0050] $$N_c(t)=\mathrm{int}\{N_c(0)(1-t/T)\};$$

[0051] 其中，t为当前迭代次数。

[0052] D、重复执行A、B和C，直到迭代更新的次数达到设置的学习次数T，获取与所述任一个采样点数据之间距离最小的权值向量w_m'，则权值向量w_m'对应的离散化值为所述任一个采样点数据的离散化值。

[0053] 基于更新后的权值向量、SOFM神经网络的学习率和邻域，重复执行A、B和C步骤，不断进行迭代更新，直到迭代次数达到设置的学习次数T，获取此时与采样点数据的距离最小的权值向量，该权值向量对应的离散化值为采样点数据的初始离散化值。此时，需要说明的是，之所有称为初始离散化值，是指当初始离散化值的数目为H_0时，采样点数据对应的离散化值，如果离散化值的数目变化时，那么采样点数据对应的离散化值也会变化。

[0054] E、对每一个钻井特征参数的每一个采样点数据，均执行A、B、C和D，得到每一个钻井特征参数的每一个采样点数据的初始离散化值。

[0055] 可理解的是，对于每一个钻井特征参数中的每一个采样点数据均采用相同的方法获取对应的初始离散化值。

[0056] 作为实施例，在E之后还包括：F、基于每一个钻井特征参数的每一个采样点数据的初始离散化值，确定每一个钻井特征参数对应的离散化值的最优数目；G、基于每一个钻井特征参数对应的离散化值的最优数目更新A中初始化的每一个钻井特征参数对应的离散化值的数目H_0，执行A、B、C和D，得到每一个钻井特征参数的每一个采样点数据的离散化值。

[0057] 可理解的是，每一个钻井特征参数对应的离散化值的数目对钻井特征参数的离散化的效果会有影响，进而对后续井壁风险等级的预测也有一定的影响，因此，需要确定每一个钻井特征参数的离散化值的最优数目。

[0058] 作为实施例，所述F、基于每一个钻井特征参数的每一个采样点数据的初始离散化值，确定每一个钻井特征参数对应的离散化值的最优数目，包括：根据初始离散化值确定每

个钻井特征参数对井壁风险等级的依赖度γ；确定依赖度最大值γ_{max}对应的钻井特征参数X_p，并通过迭代计算钻井特征参数X_p对应的离散化值的最优数量H_1；按照依赖度γ从大到小的顺序依次对H_1减1确定每一个钻井特征参数的离散化值的最优数目。

[0059] 其中，根据钻井特征参数不同的数据分布和重要性，由初始离散后的数据确定每个钻井特征参数对井壁风险等级的依赖度γ，确定依赖度最大值γ_{max}对应的钻井参数X_p。

[0060] 依赖度γ计算式如下：

[0061] $$\gamma_{Xp}(D)=\frac{|POS_{Xp}(D)|}{|U|}$$；

[0062] 式中，$\gamma_{Xp}(D)$为钻井参数X_p对井壁风险等级D的依赖度，$POS_{Xp}(D)$为D的X_p正域，|U|为钻井特征参数个数。

[0063] 作为实施例，所述确定依赖度最大值γ_{max}对应的钻井特征参数X_p，并通过迭代计算钻井特征参数X_p对应的离散化值的最优数量H_1，包括：

[0064] A’，基于每一个钻井特征参数对应的离散化值的初始数目H_0，生成钻井特征参数X_p的H_0个初始聚类中心。

[0065] 可理解的是，设钻井特征参数$X_p=(x_{p1},x_{p1},...,x_{pn})$，n为采样点的个数，选择$H_0$个初始聚类中心$U=\{u_1,u_2,...,u_{H_0}\}$，其中$u_j$为1维向量，表示第j个聚类中心位置。

[0066] B’，计算采样点数据x_{pn}到每一个聚类中心的距离，获取与采样点数据x_{pn}的距离最近的聚类中心，将采样点数据x_{pn}归类到该簇中。

[0067] 其中，采样点数据x_{pn}到每个聚类中心u_j的距离d_{nj}为：

[0068] $$d_{nj}=\sqrt{\sum_{j=1}^{H0}(x_{pn}-u_j)^2}$$；

[0069] x_{pn}的聚类中心为：

[0070] $$c_j=\arg\min_j(d_2);$$

[0071] 式中，j为样本x_{pn}的类别。

[0072] C’，遍历每一个采样点数据，将每一个采样点数据归类到不同的簇中。

[0073] 根据上述公式将每一个采样点数据归类到对应的簇中。

[0074] 其中，根据每一个簇中的所有采样点数据重新计算该簇的聚类中心：

[0075] $$c_j=(1/|c_j|)*\sum x_p\in c_j x_p$$；

[0076] 其中，$|c_j|$为簇c_j中采样点数据的个数，$\sum x_p\in c_j x_p$为c_j内所有采样点数据的特征向量总和。

[0077] D’，基于每一个簇中的所有采样点数据，重新计算簇中心，利用簇中心更新初始聚类中心。

[0078] E’，基于更新后的聚类中心，计算所有簇中所有采样点数据与对应聚类中心的误差平方和SSE。

[0079] 根据每一个簇的更新后的聚类中心以及簇中所有采样点数据，计算所有采样点数据到对应聚类中心的误差平方和SSE：

[0080] $$SSE = \sum_{j=1}^{H0}\left(\sum_{x_p \in c_j}^{x_m} d(x_j, c_j)^2\right);$$

[0081] 式中，H_0为初始聚类数，x_j是第j类聚类中心的第j个数据点，c_j为第 j 个聚类中心，$d(x_j, c_j)$ 为x_j到对应聚类中心c_j的距离，x_m为c_j所在簇中的采样点数据的个数。

[0082] F’，在下一次迭代过程中，更新钻井特征参数对应的离散化值的初始数目H_0，并执行A’、B’、C’、D’和E’，计算每一次迭代后的误差平方和SSE，直到计算出的误差平方和SSE不发生变化时，获取此次迭代后的钻井特征参数X_p对应的离散化值的最优数目H_1。

[0083] 根据依赖度 γ 确定不同钻井特征参数离散化值的数目，γ 越大离散化值的数目越大，反之则越小。由于 γ_{max} 对应钻井特征参数X_p对应的离散化值的数目为H_1，则比 γ_{max} 较小的 γ 对应的钻井特征参数的离散化值的数目H_2为：

[0084] $$H_2=H_1-1;$$

[0085] 随着 γ 的逐级减小，对应钻井特征参数离散化值的数目也随之递减，根据每一个钻井特征参数的依赖度，其对应的离散化值的数量依次减1。根据确定的每个钻井特征参数的离散化值的最优数目，重新执行A、B、C、D和E步骤，对所有钻井特征参数分别进行离散并编码赋值。

[0086] 步骤4，将每一个钻井特征参数的离散化值和井径扩大率的离散化值对应组合，组成决策表。

[0087] 可理解的是，将离散后的井径扩大率数据与钻井特征参数的编码数据进行组合，构成决策表L，也就是每一个钻井特征参数的离散化值组合对应的井径扩大率的离散化值，组成决策表L，决策表L中的每一行可表示为$(X_1, X_2, ..., X_{14}, D)$，$X_1, X_2, ..., X_{14}$表示14个钻井特征参数，D表示井壁风险等级。

[0088] 步骤5，计算每一个钻井特征参数对决策表的信息增益，选取信息增益较大的多个钻井特征参数。

[0089] 其中，钻井特征参数X_h对决策表的信息增益为：

[0090] $$Gain(X_h) = Entropy(L) - Entropy(X_h, L);$$

[0091] 式中，$Gain(X_h)$表示钻井特征参数X_h对决策表L的信息增益；

[0092] 其中，决策表L的信息熵Entropy(L)为：

[0093] $$Entropy(L) = -\sum\left(p(a) * \log_2(p(a))\right);$$

[0094] 式中,p(a)为决策表L中井径扩大率的离散化值a的概率;

[0095] 钻井特征参数X_h对Entropy(L)的条件期望为:

[0096] $$Entropy(X_h,L)=\sum\left((|L_V|/|L|)*Entropy(L_V)\right);$$

[0097] 式中,L_v为在钻井参数X_h上离散化值为v的子集,$|L_v|$为子集L_v的采样点数,|L|为决策表L的样本总数。

[0098] 通过上式计算出每一个钻井特征参数对决策表L的信息增益Gain,由于钻井特征参数的信息增益越大,则该钻井特征参数对于井壁风险等级的贡献越大,属性的重要性越高。因此选取Gain较大的多个钻井特征参数为井壁风险等级预测模型的输入参数。

[0099] 步骤6,获取训练集,所述训练集中包括多个训练样本,每一个训练样本包括信息增益较大的多个钻井特征参数的离散化值对应的井径扩大率的离散化值。

[0100] 可理解的是,步骤5中选取了信息增益较大的g个钻井特征参数,其与对应的井径扩大率构成一个训练样本,多个训练样本构成训练集。

[0101] 其中,训练集可表示为$S=[(S_1,\ y_1),\ (S_2,\ y_2),\ \ldots,\ (S_{M'},\ y_{M'})]$,其中包括M'个训练样本,其中训练样本$S_{M'}$中包括信息增益较大的g个钻井特征参数,$y_{M'}$为井壁风险等级。

[0102] 步骤7,基于训练集对预测模型进行训练。

[0103] 可理解的是,基于训练集对预测模型进行训练,包括:

[0104] 基于训练集训练CatBooat模型建立井壁风险等级预测模型,其中,通过贝叶斯算法优化目标函数$L(y,\hat{y})$寻找$CatBoost$模型的最佳超参数集,优化目标函数$L(y,\hat{y})$的表达式为:

[0105] $$L(y,\hat{y})=-\frac{1}{N}\sum_{i=1}^{N}y_{i\Lambda}\cdot\log p_{i\Lambda}+\frac{\lambda}{2}\sum_{j=1}^{T}\|w_j\|^2;$$

[0106] 式中,N为训练样本数量,y为井壁实际风险等级,$\hat{y}$为模型输出的井壁预测风险等级,$y_{i\Lambda}$为样本S_i是否属于第Λ类,$p_{i\Lambda}$为样本S_i属于第Λ类的概率,j是第i个样本的实际类别标签,T为井壁风险等级的最大类别数,λ是正则化系数。

[0107] 以步骤6提取的训练样本为输入,采交叉验证法划分数据集,输入井壁风险等级预测模型进行训练和测试。

[0108] 通过一对多感受性曲线(ROC)来评价预测模型对每个风险等级的预测效果,曲线下面积(AUC)为ROC曲线所覆盖的区域面积,AUC越大分类效果越好。采用正确预测率、分类准确率和风险识别率三类指标对训练后的井壁风险等级预测模型进行评价,计算公式为:

[0109] 正确预测率 $$A=\frac{RO}{RO+AO};$$

[0110] 分类准确率 $B=\frac{RO+RE}{RO+RE+AO+AE}$；

[0111] 风险识别率 $C=\frac{RO}{RO+AE}$；

[0112] 其中，以0级风险等级为正例，1、2、3级风险等级为负例。则，RO，正确预测为0级风险等级的个数；AO，错误预测为0级风险等级的个数；AE，错误预测为1、2、3级风险等级的个数；RE，正确预测为1、2、3级风险等级的个数。

[0113] 步骤8，将待测的多个钻井特征参数的离散化值输出训练后的预测模型中，输出井壁的风险等级值。

[0114] 可理解的是，基于步骤7训练后的预测模型对井壁风险等级进行预测，将待测的多个增益较大的钻井特征参数的离散化值输出训练后的预测模型中，输出井壁的风险等级值。

[0115] 下面以一个具体的算例对本发明提供的井壁风险等级预测方法进行说明。

[0116] 1、获取目标地层的特征参数和目标参数数据；

[0117] 从现场钻井资料中，根据井史情况确定舒善河层组的井深范围，在该范围内选择14个钻井特征参数和1个目标参数的数据，钻井特征参数X_1到X_{14}分别为井斜角、方位角、狗腿度、钻井液密度、漏斗粘度、钻速、泵压、钻压、转速、排量、立管压力、钻头压降、环空压耗和喷射速度，目标参数为井径扩大率。

[0118] 2、对特征参数和目标参数进行预处理；

[0119] 剔除钻井特征参数中异常值和空缺值，进行归一化处理。

[0120] 表1 钻井参数数据集

[0121]

特征	平均值	最小值	最大值
X_1	0.89	0.1	2.29
X_2	177.98	10.24	356.21
X_3	0.19	0.01	1.8
X_4	1.22	1.08	1.3
X_5	44.21	38	62
X_6	21.53	14.16	32.994
X_7	21.47	17	25
X_8	52.13	30	80
X_9	57.47	30	85
X_{10}	45.98	40.33	51.333
X_{11}	21.22	18	25
X_{12}	1.29	0.12	5.74
X_{13}	15.80	0.16	24.44
X_{14}	36.46	11.35	91.71

[0122] 3、构建井壁风险等级决策表；

[0123] 目标参数离散化：根据井径扩大率分布情况，按照等距离将分布区间划分为四个等级，对每一个区间编码赋值，分别为0、1、2、3。

[0124] 钻井特征参数离散化：

[0125] 初始离散：设置SOFM神经网络的学习率为0.1，邻域值为1，训练步数为400，离散类别数H_0为5，对所有钻井特征参数进行离散；

[0126] 确定依赖度γ_{max}对应的钻井特征参数为井斜角和漏斗粘度；确定井斜角和漏斗粘度离散划分的最优离散化值数目H_1为10；

[0127] 确定每个钻井特征参数对井壁风险等级的依赖度γ，根据γ大小对钻井特征参数进行排序，分别为井斜角、漏斗粘度、转速、钻速、钻压、立管压力、方位角、泵压、钻头压降、喷射速度、排量、狗腿度、环空压耗、钻井液密度。

[0128] 重新确定每个钻井特征参数的离散化值数目，其中井斜角和漏斗粘度离散的类别为10，转速离散的类别为9，钻速、钻压和立管压力离散的类别为8，方位角、泵压、钻头压降和喷射速度离散的类别为7，排量离散的类别为6，狗腿度和环空压耗离散的类别为5，钻井液密度离散的类别为4。

[0129] 利用SOFM神经网络对每个钻井特征参数分别进行离散并编码赋值，将离散后的钻井特征参数和离散的井径扩大率数据组合，构建井壁风险等级决策表L。

[0130] 表2 井壁风险等级决策表L

[0131]

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	D
3	6	0	0	0	2	4	1	5	5	5	5	0	0	0
3	6	0	0	0	2	4	1	5	5	5	5	0	0	0
1	2	1	0	4	5	4	1	5	6	5	5	0	0	1
3	3	0	0	0	5	2	7	1	6	3	0	1	2	0
3	3	1	0	4	5	4	1	5	6	5	5	0	0	1
6	3	2	0	0	6	2	7	1	6	3	0	1	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	3	0	6	8	1	3	4	3	0	5	0	0	1	3
3	1	0	6	2	1	2	3	6	0	2	4	2	0	3

[0132] 4、提取特征参数：

[0133] 计算各钻井特征参数与井壁风险等级之间的信息增益，提取其中信息增益较大的钻井特征参数，分别为钻速、排量、立管压力、转速、喷射速度、钻井液密度、方位角。

[0134] 表3 信息增益计算结果

[0135]

特征	X_1	X_2	X_3	X_4	X_5	X_6	X_7
信息增益	0.0021	0.0213	0	0.1759	0	0.1889	0
特征	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
信息增益	0.0196	0.0257	0.2163	0.0493	0.0101	0.0182	0.0096

[0136] 5、建立预测模型；

[0137] 以提取的钻速、排量、立管压力、转速、喷射速度、钻井液密度、方位角为输入参数，不同的井壁风险等级为输出参数建立井壁风险等级预测模型。

[0138] 采用5折交叉验证法划分数据集，利用贝叶斯优化算法对CatBoost模型的超参数进行寻优，得到的最佳超参数集如下。

[0139] 表4 井壁风险等级预测模型最优超参数集

[0140]

参数	范围	最优值
决策树数最大数量	(50,200)	155
学习率	(0.01,0.3)	0.3
树的深度	(3,15)	12
L2 正则化数	(1,10)	5.7

[0141] 6、训练模型；

[0142] 将最优超参数集带入CatBoost模型，采用5折交叉验证法划分数据集，由公式23到25，采用正确预测率、分类准确率和风险识别率三类指标对训练后的井壁风险等级预测模型进行验证和评价。

[0143] 例如，对新疆某油田采集的数据样本进行训练和测试，预测效果较好，表明该模型具有较高的预测准确率和较好的泛化效果。

[0144] 表5 预测效果

[0145]

正确预测率	分类准确率	风险识别率
89.39%	84.72%	82.04%

[0146] 7、井壁风险等级预测

[0147] 基于训练后的预测模型对井壁风险等级进行预测。

[0148] 本发明实施例提供的一种井壁风险等级预测方法，具有以下有益效果：

[0149] (1)本发明通过依赖度来确定不同钻井特征参数的离散类别数，并通过SOFM神经网络方法对连续型特征参数进行离散，无需预先标记训练数据，通过调整权重向量和邻域来适应输入数据的分布和特点，最大程度保留原始数据的信息。

[0150] (2)本发明通过属性重要性中的信息增益来提取特征参数，显著减少了模型输入参数的数量并提高了预测效果。

[0151] (3)本发明通过井径扩大率的大小来表征不同的井壁风险等级，较力学模型而言极大的简化了预测过程，实现过程非常简捷。

[0152] 需要说明的是，在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详细描述的部分，可以参见其它实施例的相关描述。

[0153] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0154] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程

和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式计算机或者其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0155] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0156] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0157] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0158] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包括这些改动和变型在内。

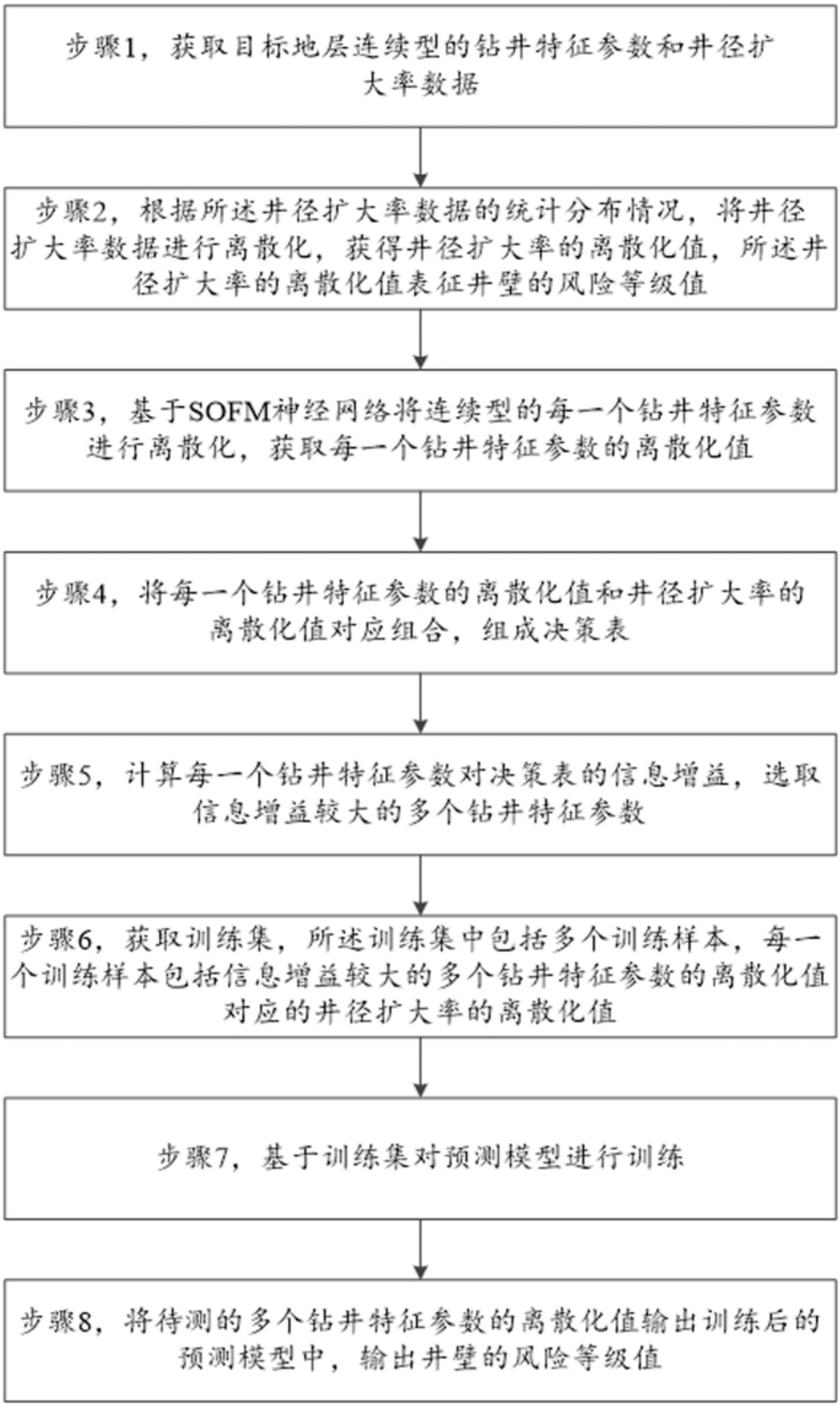

步骤1，获取目标地层连续型的钻井特征参数和井径扩大率数据
步骤2，根据所述井径扩大率数据的统计分布情况，将井径扩大率数据进行离散化，获得井径扩大率的离散化值，所述井径扩大率的离散化值表征井壁的风险等级值
步骤3，基于SOFM神经网络将连续型的每一个钻井特征参数进行离散化，获取每一个钻井特征参数的离散化值
步骤4，将每一个钻井特征参数的离散化值和井径扩大率的离散化值对应组合，组成决策表
步骤5，计算每一个钻井特征参数对决策表的信息增益，选取信息增益较大的多个钻井特征参数
步骤6，获取训练集，所述训练集中包括多个训练样本，每一个训练样本包括信息增益较大的多个钻井特征参数的离散化值对应的井径扩大率的离散化值
步骤7，基于训练集对预测模型进行训练
步骤8，将待测的多个钻井特征参数的离散化值输出训练后的预测模型中，输出井壁的风险等级值

图 1

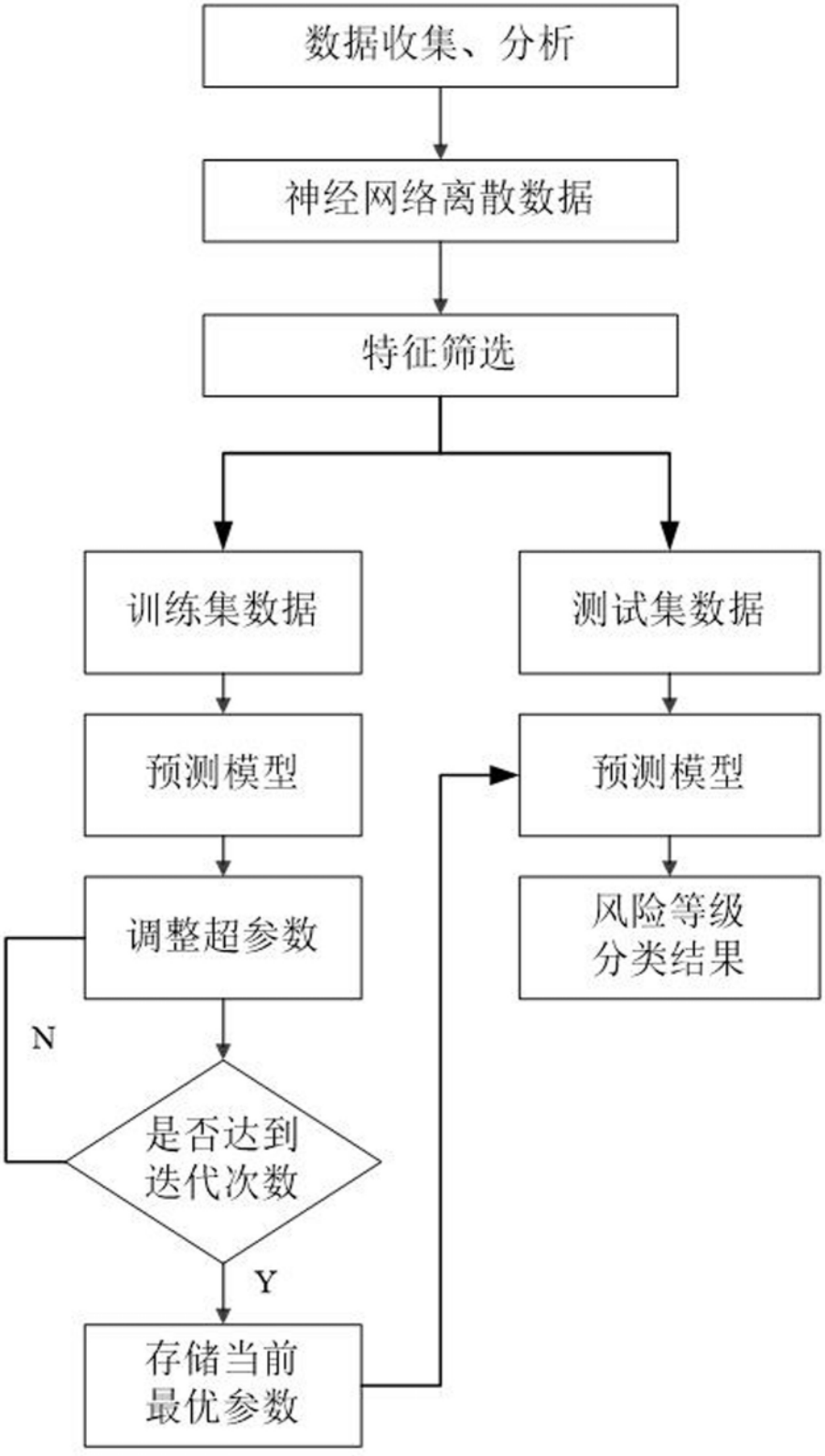

数据收集、分析
神经网络离散数据
特征筛选
训练集数据
测试集数据
预测模型
预测模型
调整超参数
风险等级
分类结果
N
是否达到
迭代次数
Y
存储当前
最优参数

图 2