(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111275204 B
(45)授权公告日 2023.04.07

(21)申请号 202010117648.2

(22)申请日 2020.02.25

(65)同一申请的已公布的文献号
申请公布号 CN 111275204 A

(43)申请公布日 2020.06.12

(73)专利权人 西安工程大学
地址 710048 陕西省西安市碑林区金花南路19号

(72)发明人 黄新波 蒋卫涛 朱永灿 曹雯 田毅

(74)专利代理机构 西安弘理专利事务所 61214
专利代理师 张皎

(51)Int.Cl.
G06N 20/10(2019.01)
G06N 20/20(2019.01)
G06F 18/214(2023.01)
G06F 18/2411(2023.01)

审查员 李锦川

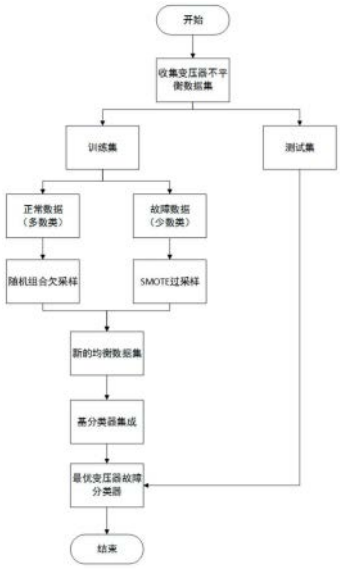
权利要求书3页 说明书6页 附图3页

(54)发明名称

一种基于混合采样和集成学习的变压器状态识别方法

(57)摘要

本发明公开了一种基于混合采样和集成学习的变压器状态识别方法，具体为：步骤1：将收集到的变压器油中溶解气体数据分为两个数据集；步骤2：对步骤1得到的训练集进行SMOTE过采样，将进行SMOTE过采样后的数据集记为新故障训练数据集；步骤3：将得到的新正常训练数据集S_1^*与步骤2得到的新故障训练数据集组合产生新的均衡数据集；步骤4：以最小二乘支持向量机为基分类器，利用步骤3生成的q组均衡子数据集训练q个基分类器；步骤5：将步骤4训练得到的q个基分类器进行集成得到强分类器对变压器进行状态识别；通过组合得到的强分类器即为变压器状态识别最优模型，对模型进行测试。该方法能够对变压器状态进行准确的识别。



CN 111275204 B

1.一种基于混合采样和集成学习的变压器状态识别方法，其特征在于，具体按照以下步骤实施：

步骤1：将收集到的变压器油中溶解气体数据分为两个数据集，正常数据集S_1和故障数据集S_2，S_2数据集中包括：低温过热数据集S_{21}、中温过热数据集S_{22}、高温过热数据集S_{23}、高能放电数据集S_{24}、低能放电数据集S_{25}；

分别将收集得到的6个数据集S_1、S_{21}、S_{22}、S_{23}、S_{24}、S_{25}中的数据个数按5:1的比例分为训练集$S_1^{\ 1}$、$S_{21}^{\ 1}$、$S_{22}^{\ 1}$、$S_{23}^{\ 1}$、$S_{24}^{\ 1}$、$S_{25}^{\ 1}$和测试集$S_1^{\ 2}$、$S_{21}^{\ 2}$、$S_{22}^{\ 2}$、$S_{23}^{\ 2}$、$S_{24}^{\ 2}$、$S_{25}^{\ 2}$；

步骤2：对步骤1得到的训练集$S_{21}^{\ 1}$、$S_{22}^{\ 1}$、$S_{23}^{\ 1}$、$S_{24}^{\ 1}$、$S_{25}^{\ 1}$进行SMOTE过采样，将进行SMOTE过采样后的数据集记为新故障训练数据集S_{21}^*、S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*；

步骤3：随机取出训练集中$S_1^{\ 1}$取w*n个数据，将取得的数据记为新正常训练集记为S_1^*，将得到的新正常训练数据集S_1^*与步骤2得到的新故障训练数据集S_{21}^*、S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*组合产生新的均衡数据集记为S^1，重复上述操作q次，共产生q组均衡数据集记为$S^q=\{x_t^*,y_t^*\}^q,q=1,2,\cdots 10$；其中，w为随机采样采样率，$x_t^*$为输入变量即七种油中溶解气体包含氢气、甲烷、乙烷、乙烯、乙炔、一氧化碳和二氧化碳，y_t^*为输出变量即故障类型包括低温过热、中温过热、高温过热、低能放电和高能放电，t为每一组均衡数据集的数据个数；

步骤4：以最小二乘支持向量机为基分类器，利用步骤3生成的q组均衡子数据集训练q个基分类器；

步骤5：利用Bagging集成算法将步骤4训练得到的q个基分类器进行集成得到强分类器对变压器进行状态识别；采用相对多数投票法对q个基分类器进行组合；通过组合得到的强分类器即为变压器状态识别最优模型，最后利用测试集$S_1^{\ 2}$、$S_{21}^{\ 2}$、$S_{22}^{\ 2}$、$S_{23}^{\ 2}$、$S_{24}^{\ 2}$、$S_{25}^{\ 2}$，对最优模型进行测试。

2.根据权利要求1所述的一种基于混合采样和集成学习的变压器状态识别方法，其特征在于，步骤1中，S_1数据集中的数据个数为n个，S_{21}、S_{22}、S_{23}、S_{24}、S_{25}数据集中的数据个数均为m个，n>6m，数据集S_1中的数据个数多于数据集S_2中的数据个数。

3.根据权利要求2所述的一种基于混合采样和集成学习的变压器状态识别方法，其特征在于，步骤2中对步骤1得到的训练集$S_{21}^{\ 1}$、$S_{22}^{\ 1}$、$S_{23}^{\ 1}$、$S_{24}^{\ 1}$、$S_{25}^{\ 1}$依次进行SMOTE过采样，具体为：

步骤a、对训练集$S_{21}^{\ 1}$进行SMOTE过采样，随机选取一个点$x\in(x_1,x_2,\cdots x_a)$作为训练集中的$S_{21}^{\ 1}$的计算初始点，计算初始点x到训练集$S_{21}^{\ 1}$内除点x外的其他所有点的距离，计算公式如式(1)所示：

$$dist(x,x_j)=\sum_{u=1}^{a}\left|x_u-x_{ju}\right| \qquad (1)$$

其中，x表示初始点，x_j表示训练集$S_{21}^{\ 1}$中的除初始点x外的其他点，x_u表示初始点x的元素，x_{ju}表示训练集$S_{21}^{\ 1}$中的除初始点x外的其他点x_j的元素；

步骤b、选择距离点x最近的b个点，记为邻近点，其中b为SMOTE采样率；

利用选择的b个邻近点与初始点x进行SMOTE插值，每个邻近点与初始点之间只可以插值一次，具体的插值公式如式(2)所示：

$d_k = x + c \cdot (y_k - x)$ (2)

其中，d_k表示第k个插值点，c表示0-1之间的一个随机数，y_k表示第k个邻近点；将得到的插值点与原始数据集合并作为新的数据集记为S_{21}^*；

步骤c、依照步骤a至步骤b的方法分别对S_{22}^1、S_{23}^1、S_{24}^1、S_{25}^1进行SMOTE过采样，将得到的新故障训练数据集分别记为S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*。

4.根据权利要求3所述的一种基于混合采样和集成学习的变压器状态识别方法，其特征在于，步骤4具体按照以下步骤实施：

步骤4.1：以LSSVM为基础建立基分类器，假设二分类的超平面的表达式为：

$w \cdot \phi(x) + b = 0$ (3)

其中，w为权值矢量，b为阈值，$\phi(x)$为输入向量；

将式(3)中寻找最优超平面问题转变为求解线性问题，如式(4)所示：

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + \frac{1}{2}C \cdot \sum_{i=1}^{n} \xi_i^2 \\ \text{约束于：} \ y_i \cdot [w^T \phi(x_i) + b] = 1 - \xi_i \end{cases} \quad (4)$$

其中，C为惩罚参数，ξ_i为非负松弛因子；

步骤4.2：对步骤4.1中的线性问题进行求解，引入拉格朗日乘子并依据KKT条件可求解如下线性问题：

$$\begin{bmatrix} 0 & Y^T \\ Y & \Omega_{ij} + C^{-1} I_N \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ I_V \end{bmatrix} \quad (5)$$

其中：$Y = [y_1, y_2, \cdots y_n]^T$，$\alpha = [\alpha_1, \alpha_2 \cdots \alpha_n]^T$，$\Omega_{ij} = y_i y_j K(x_i, x_j)$，i，j＝1,2,…,n为核矩阵，$I_V = [1,1,\cdots 1]^T$，$K(x_i, x_j)$为核函数，$I_N$为单位矩阵；利用最小二乘法求出α和b后，可得LSSVM的决策函数为：

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + \beta\right) \quad (6)$$

其中，α_i为拉格朗日乘子，β为分类阈值；

步骤4.3：依照步骤4.1至4.2的方法再构造4个分类函数，利用步骤3产生的均衡数据集进行训练，得到最优的分类模型，对变压器的6种状态进行识别；

步骤4.4：重复p次步骤4.1、4.2和4.3得到p个基分类器分别为E_p，其中p＝1,2,…10。

5.根据权利要求4所述的一种基于混合采样和集成学习的变压器状态识别方法，其特征在于，步骤5中采用相对多数投票法对q个基分类器进行组合，具体的组合方法如式(7)所示：

$$H(x)=\begin{cases} C_j & \sum_{q=1}^{10}E_q^r > 0.5\sum_{r=1}^{6}\sum_{q=1}^{10}E_q^r \\ \text{拒绝} & \text{其他} \end{cases} \quad (7)$$

其中，H(x)为最终的强分类器，C_j表示强分类器的最终输出，E_p^r表示第q个分类器的输出结果为r，r＝1,2,3,4,5,6表示变压器的6种状态，分别是正常、低温过热、中温过热、高温过热、低能放电、高能放电。

一种基于混合采样和集成学习的变压器状态识别方法

技术领域

[0001] 本发明属于变压器在线监测与故障诊断领域，具体涉及一种基于混合采样和集成学习的变压器状态识别方法。

背景技术

[0002] 变压器作为电网的关键性设备，其安全稳定运行是保证电力正常供应和电力系统安全的基础，一旦变压器发生状态对周围的经济和生活将产生巨大的影响。因此，变压器的状态识别问题已经成为了国内外学者研究的热点问题。

[0003] 随着人工智能技术的飞速发展，传统的以DGA为基础的例如三比值法、大卫三角形、罗杰斯比值法等方法已经不能满足目前人们对变压器状态识别精度要求了。因此出现了一系列的智能识别方法，如：支持向量机、神经网络、模糊聚类等等。但是这些智能识别方法都有一个共同的特点就是需要大量的训练数据来训练网络，通过训练好的网络来对状态进行识别，因此，智能方法网络训练的效果决定着该方法的最终对变压器的识别准确率。变压器作为重要的设备，其状态的发生概率很低，在变压器的运行过程中状态的数据很少，也会存在大量的正常数据，如果将这种正常数据与状态数据不均衡的数据集作为训练网络的训练数据集，那么在训练的过程中，就会导致分类模型的偏差，会对识别模型的识别准确率产生很大的影响。

[0004] 因此，本发明提出了一种基于混合采样和集成学习的变压器状态识别方法，该方法能够很好地处理变压器训练数据不均衡的情况，最大限度的提高变压器状态识别准确率。

发明内容

[0005] 本发明的目的是提供一种基于混合采样和集成学习的变压器状态识别方法，该方法能够解决数据不平衡的问题，提高变压器状态识别准确率。

[0006] 本发明所采用的技术方案是，一种基于混合采样和集成学习的变压器状态识别方法，具体按照以下步骤实施：

[0007] 步骤1：将收集到的变压器油中溶解气体数据分为两个数据集，正常数据集S_1和故障数据集S_2，S_2数据集中包括：低温过热数据集S_{21}、中温过热数据集S_{22}、高温过热数据集S_{23}、高能放电数据集S_{24}、低能放电数据集S_{25}；

[0008] 分别将收集得到的6个数据集S_1、S_{21}、S_{22}、S_{23}、S_{24}、S_{25}中的数据个数按5:1的比例分为训练集$S_1^{\ 1}$、$S_{21}^{\ 1}$、$S_{22}^{\ 1}$、$S_{23}^{\ 1}$、$S_{24}^{\ 1}$、$S_{25}^{\ 1}$和测试集$S_1^{\ 2}$、$S_{21}^{\ 2}$、$S_{22}^{\ 2}$、$S_{23}^{\ 2}$、$S_{24}^{\ 2}$、$S_{25}^{\ 2}$；

[0009] 步骤2：对步骤1得到的训练集$S_{21}^{\ 1}$、$S_{22}^{\ 1}$、$S_{23}^{\ 1}$、$S_{24}^{\ 1}$、$S_{25}^{\ 1}$进行SMOTE过采样，将进行SMOTE过采样后的数据集记为新故障训练数据集S_{21}^{*}、S_{22}^{*}、S_{23}^{*}、S_{24}^{*}、S_{25}^{*}；

[0010] 步骤3：随机取出训练集中$S_1^{\ 1}$取w*n个数据，将取得的数据记为新正常训练集记为S_1^{*}，将得到的新正常训练数据集S_1^{*}与步骤2得到的新故障训练数据集

S_{21}^{*}、S_{22}^{*}、S_{23}^{*}、S_{24}^{*}、S_{25}^{*}组合产生新的均衡数据集记为S^{1}，重复上述操作q次，共产生q组均衡数据集记为$S^{q}=\{x_t^{*},y_t^{*}\}^{q},q=1,2,\cdots 10$；其中，w为随机采样采样率，$x_t^{*}$为输入变量即七种油中溶解气体包含氢气、甲烷、乙烷、乙烯、乙炔、一氧化碳和二氧化碳，y_t^{*}为输出变量即故障类型包括低温过热、中温过热、高温过热、低能放电和高能放电，t为每一组均衡数据集的数据个数；

[0011] 步骤4：以最小二乘支持向量机为基分类器，利用步骤3生成的q组均衡子数据集训练q个基分类器；

[0012] 步骤5：利用Bagging集成算法将步骤4训练得到的q个基分类器进行集成得到强分类器对变压器进行状态识别；采用相对多数投票法对q个基分类器进行组合；

[0013] 通过组合得到的强分类器即为变压器状态识别最优模型，最后利用测试集S_1^{2}、S_{21}^{2}、S_{22}^{2}、S_{23}^{2}、S_{24}^{2}、S_{25}^{2}，对最优模型进行测试。

[0014] 本发明的特点还在于，

[0015] 步骤1中，S_1数据集中的数据个数为n个，S_{21}、S_{22}、S_{23}、S_{24}、S_{25}数据集中的数据个数均为m个，n＞6m，数据集S_1中的数据个数多于数据集S_2中的数据个数。

[0016] 步骤2中对步骤1得到的训练集S_{21}^{1}、S_{22}^{1}、S_{23}^{1}、S_{24}^{1}、S_{25}^{1}依次进行SMOTE过采样，具体为：

[0017] 步骤a、对训练集S_{21}^{1}进行SMOTE过采样，随机选取一个点$x\in(x_1,x_2,\cdots x_a)$作为训练集中的S_{21}^{1}的计算初始点，计算初始点x到训练集S_{21}^{1}内除点x外的其他所有点的距离，计算公式如式(1)所示：

[0018] $$dist(x,x_j)=\sum_{u=1}^{a}\left|x_u-x_{ju}\right| \qquad (1)$$

[0019] 其中，x表示初始点，x_j表示训练集S_{21}^{1}中的除初始点x外的其他点，x_u表示初始点x的元素，x_{ju}表示训练集S_{21}^{1}中的除初始点x外的其他点x_j的元素；

[0020] 步骤b、选择距离点x最近的b个点，记为邻近点，其中b为SOMTE采样率；

[0021] 利用选择的b个邻近点与初始点x进行SMOTE插值，每个邻近点与初始点之间只可以插值一次，具体的插值公式如式(2)所示：

[0022] $d_k=x+c\cdot(y_k-x)$ (2)

[0023] 其中，d_k表示第k个插值点，c表示0-1之间的一个随机数，y_k表示第k个邻近点；将得到的插值点与原始数据集合并作为新的数据集记为$S_2{}^{*}_1$；

[0024] 步骤c、依照步骤a至步骤b的方法分别对S_{22}^{1}、S_{23}^{1}、S_{24}^{1}、S_{25}^{1}进行SMOTE过采样，将得到的新故障训练数据集分别记为$S_2{}^{*}_2$、$S_2{}^{*}_3$、$S_2{}^{*}_4$、$S_2{}^{*}_5$。

[0025] 步骤4具体按照以下步骤实施：

[0026] 步骤4.1：以LSSVM为基础建立基分类器，假设二分类的超平面的表达式为：

[0027] $w\cdot\Phi(x)+b=0$ (3)

[0028] 其中，w为权值矢量，b为阈值，$\Phi(x)$为输入向量；

[0029] 将式(3)中寻找最优超平面问题转变为求解线性问题，如式(4)所示：

[0030] $$\begin{cases}\min\dfrac{1}{2}\|w\|^2+\dfrac{1}{2}C\cdot\sum_{i=1}^{n}\xi_i^2 \\ 约束于：\ y_i\cdot[w^T\phi(x_i)+b]=1-\xi_i\end{cases} \qquad (4)$$

[0031] 其中,C为惩罚参数,ξ_i为非负松弛因子;

[0032] 步骤4.2:对步骤4.1中的线性问题进行求解,引入拉格朗日乘子并依据KKT条件可求解如下线性问题:

[0033] $$\begin{bmatrix}0 & Y^T \\ Y & \Omega_{ij}+C^{-1}I_N\end{bmatrix}\begin{bmatrix}\beta \\ \alpha\end{bmatrix}=\begin{bmatrix}0 \\ I_V\end{bmatrix} \qquad (5)$$

[0034] 其中:$Y=[y_1,y_2,\cdots y_n]^T$,$\alpha=[\alpha_1,\alpha_2\cdots\alpha_n]^T$,$\Omega_{ij}=y_iy_jK(x_i,x_j)$,i,j=1,2,…,n为核矩阵,$I_V=[1,1,\cdots1]^T$,$K(x_i,x_j)$为核函数,$I_N$为单位矩阵;利用最小二乘法求出α和$_b$后,可得LSSVM的决策函数为:

[0035] $$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_iy_iK(x_i,x_j)+\beta) \qquad (6)$$

[0036] 其中,α_i为拉格朗日乘子,β为分类阈值;

[0037] 步骤4.3:依照步骤4.1至4.2的方法再构造4个分类函数,利用步骤3产生的均衡数据集进行训练,得到最优的分类模型,对变压器的6种状态进行识别;

[0038] 步骤4.4:重复p次步骤4.1、4.2和4.3得到p个基分类器分别为E_p,其中p=1,2,…10。

[0039] 步骤5中采用相对多数投票法对q个基分类器进行组合,具体的组合方法如式(7)所示:

[0040] $$H(x)=\begin{cases}C_j & \sum_{q=1}^{10}E_q^r>0.5\sum_{r=1}^{6}\sum_{q=1}^{10}E_q^r \\ 拒绝 & 其他\end{cases} \qquad (7)$$

[0041] 其中,H(x)为最终的强分类器,C_j表示强分类器的最终输出,E_p^r表示第q个分类器的输出结果为r,r=1,2,3,4,5,6表示变压器的6种状态,分别是正常、低温过热、中温过热、高温过热、低能放电、高能放电。

[0042] 本发明的有益效果是,该方法首先利用混合采样方法处理不平衡数据,可以解决数据不平衡的问题,其次利用最小二乘支持向量机作为基分类器,加快了识别的速度,最后利用bagging集成算法将基分类器进行集成,极大的考虑到所有的训练样本,加快了识别速度,提高了变压器状态识别准确率。

附图说明

[0043] 图1是本发明一种基于混合采样和集成学习的变压器状态识别方法的流程图;

[0044] 图2是本发明一种基于混合采样和集成学习的变压器状态识别方法的原理图;

[0045] 图3是本发明一种基于混合采样和集成学习的变压器状态识别方法中SMOTE过采样示意图;

[0046] 图4是应用本发明一种基于混合采样和集成学习的变压器状态识别方法利用测试集对变压器状态识别最优模型进行测试的结果图。

具体实施方式

[0047] 下面结合附图和具体实施方式对本发明进行详细说明。

[0048] 本发明一种基于混合采样和集成学习的变压器状态识别方法，如图1所示，具体按照以下步骤实施：

[0049] 步骤1：将收集到的变压器油中溶解气体(DGA)数据分为两个数据集，正常数据集S_1和故障数据集S_2，S_2数据集中包括：低温过热数据集S_{21}、中温过热数据集S_{22}、高温过热数据集S_{23}、高能放电数据集S_{24}、低能放电数据集S_{25}；

[0050] 其中，S_1数据集中的数据个数为n个，S_{21}、S_{22}、S_{23}、S_{24}、S_{25}数据集中的数据个数均为m个，n>6m，数据集S_1中的数据个数多于数据集S_2中的数据个数；

[0051] 分别将收集得到的6个数据集S_1、S_{21}、S_{22}、S_{23}、S_{24}、S_{25}中的数据个数按5:1的比例分为训练集S_1^1、S_{21}^1、S_{22}^1、S_{23}^1、S_{24}^1、S_{25}^1和测试集S_1^2、S_{21}^2、S_{22}^2、S_{23}^2、S_{24}^2、S_{25}^2；

[0052] 步骤2：对步骤1得到的训练集S_{21}^1、S_{22}^1、S_{23}^1、S_{24}^1、S_{25}^1进行SMOTE过采样，将进行SMOTE过采样后的数据集记为新故障训练数据集S_{21}^*、S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*；SMOTE过采样的示意图如图3所示；

[0053] 步骤2中对步骤1得到的训练集S_{21}^1、S_{22}^1、S_{23}^1、S_{24}^1、S_{25}^1依次进行SMOTE过采样，具体为：

[0054] 步骤a、对训练集S_{21}^1进行SMOTE过采样，随机选取一个点$x\in(x_1,x_2,\cdots x_a)$作为训练集中的S_{21}^1的计算初始点，计算初始点x到训练集S_{21}^1内除点x外的其他所有点的距离，计算公式如式(1)所示：

[0055] $$dist(x,x_j)=\sum_{u=1}^{a}\left|x_u-x_{ju}\right| \qquad (1)$$

[0056] 其中，x表示初始点，x_j表示训练集S_{21}^1中的除初始点x外的其他点，x_u表示初始点x的元素，x_{ju}表示训练集S_{21}^1中的除初始点x外的其他点x_j的元素；

[0057] 步骤b、选择距离点x最近的b个点，记为邻近点，其中b为SMOTE采样率。

[0058] 利用选择的b个邻近点与初始点x进行SMOTE插值，每个邻近点与初始点之间只可以插值一次，具体的插值公式如式(2)所示：

[0059] $d_k=x+c\cdot(y_k-x)$ (2)

[0060] 其中，d_k表示第k个插值点，c表示0-1之间的一个随机数，y_k表示第k个邻近点；将得到的插值点与原始数据集合并作为新的数据集记为S_{21}^*；

[0061] 步骤c、依照步骤a至步骤b的方法分别对S_{22}^1、S_{23}^1、S_{24}^1、S_{25}^1进行SMOTE过采样，将得到的新故障训练数据集分别记为S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*。

[0062] 步骤3：随机取出训练集中S_1^1取w*n个数据，将取得的数据记为新正常训练集记为S_1^*，将得到的新正常训练数据集S_1^*与步骤2得到的新故障训练数据集S_{21}^*、S_{22}^*、S_{23}^*、S_{24}^*、S_{25}^*组合产生新的均衡数据集记为S^1，重复上述操作q次，共产生q组

均衡数据集记为$S^q=\{x_t^*,y_t^*\}^q,q=1,2,\cdots10$；其中，w为随机采样采样率，$x_t^*$为输入变量即七种油中溶解气体包含氢气、甲烷、乙烷、乙烯、乙炔、一氧化碳和二氧化碳，yt*为输出变量即故障类型包括低温过热、中温过热、高温过热、低能放电和高能放电，t为每一组均衡数据集的数据个数。

[0063] 步骤4：以最小二乘支持向量机(LSSVM)为基分类器，利用步骤3生成的q组均衡子数据集训练q个基分类器；

[0064] 步骤4具体按照以下步骤实施：

[0065] 步骤4.1：以LSSVM为基础建立基分类器，假设二分类的超平面的表达式为：

[0066] $w\cdot\phi(x)+b=0$ (3)

[0067] 其中，w为权值矢量，b为阈值，φ(x)为输入向量；

[0068] 将式(3)中寻找最优超平面问题转变为求解线性问题，如式(4)所示：

[0069] $$\begin{cases}\min\dfrac{1}{2}\|w\|^2+\dfrac{1}{2}C\cdot\sum_{i=1}^{n}\xi_i^2\\ \text{约束于：}\ y_i\cdot[w^T\phi(x_i)+b]=1-\xi_i\end{cases}\qquad(4)$$

[0070] 其中，C为惩罚参数，ξ_i为非负松弛因子；

[0071] 步骤4.2：对步骤4.1中的线性问题进行求解，引入拉格朗日乘子并依据KKT条件可求解如下线性问题：

[0072] $$\begin{bmatrix}0 & Y^T\\ Y & \Omega_{ij}+C^{-1}I_N\end{bmatrix}\begin{bmatrix}\beta\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ I_V\end{bmatrix}\qquad(5)$$

[0073] 其中：$Y=[y_1,y_2,\cdots y_n]^T$，$\alpha=[\alpha_1,\alpha_2\cdots\alpha_n]^T$，$\Omega_{ij}=y_iy_jK(x_i,x_j)$，i，j=1，2，…，n为核矩阵，$I_V=[1,1,\cdots1]^T$，$K(x_i,x_j)$为核函数，$I_N$为单位矩阵；利用最小二乘法求出α和b后，可得LSSVM的决策函数为：

[0074] $$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_iy_iK(x_i,x_j)+\beta)\qquad(6)$$

[0075] 其中，α_i为拉格朗日乘子，β为分类阈值；

[0076] 步骤4.3：依照步骤4.1至4.2的方法再构造4个分类函数，利用步骤3产生的均衡数据集进行训练，得到最优的分类模型，对变压器的6种状态进行识别；

[0077] 步骤4.4：重复p次步骤4.1、4.2和4.3得到p个基分类器分别为E_p(p=1，2，…10)。

[0078] 步骤5：利用Bagging集成算法将步骤4训练得到的q个基分类器进行集成得到强分类器对变压器进行状态识别；采用相对多数投票法对q个基分类器进行组合；步骤5中采用相对多数投票法对q个基分类器进行组合，具体的组合方法如式(7)所示：

[0079] $$H(x)=\begin{cases}C_j & \sum_{q=1}^{10}E_q^r>0.5\sum_{r=1}^{6}\sum_{q=1}^{10}E_q^r\\ \text{拒绝} & \text{其他}\end{cases}\qquad(7)$$

[0080] 其中，H(x)为最终的强分类器，C_j表示强分类器的最终输出，E_p^r表示第q个分类器

的输出结果为r,r＝1,2,3,4,5,6表示变压器的6种状态,分别是正常、低温过热、中温过热、高温过热、低能放电、高能放电。

[0081] 通过组合得到的强分类器即为变压器状态识别最优模型,最后利用测试集S_1^2、S_{21}^2、S_{22}^2、S_{23}^2、S_{24}^2、S_{25}^2,对变压器状态识别最优模型进行测试。

[0082] 图2为本发明一种基于混合采样和集成学习的变压器状态识别方法的原理图,其原理为利用SOMTE过采样和随机欠采样生成均衡数据集,均衡数据集作为集成学习算法的训练数据并进行训练最终得到变压器状态识别最优模型。

[0083] 利用得到的变压器状态识别最优模型对变压器进行识别,如图4所示为利用测试集对变压器状态识别最优模型进行测试的结果。从图4中可以看出基于混合采样和集成学习的变压器状态识别方法可以对变压器的状态进行准确的识别,其识别的准确率可以达到90％。

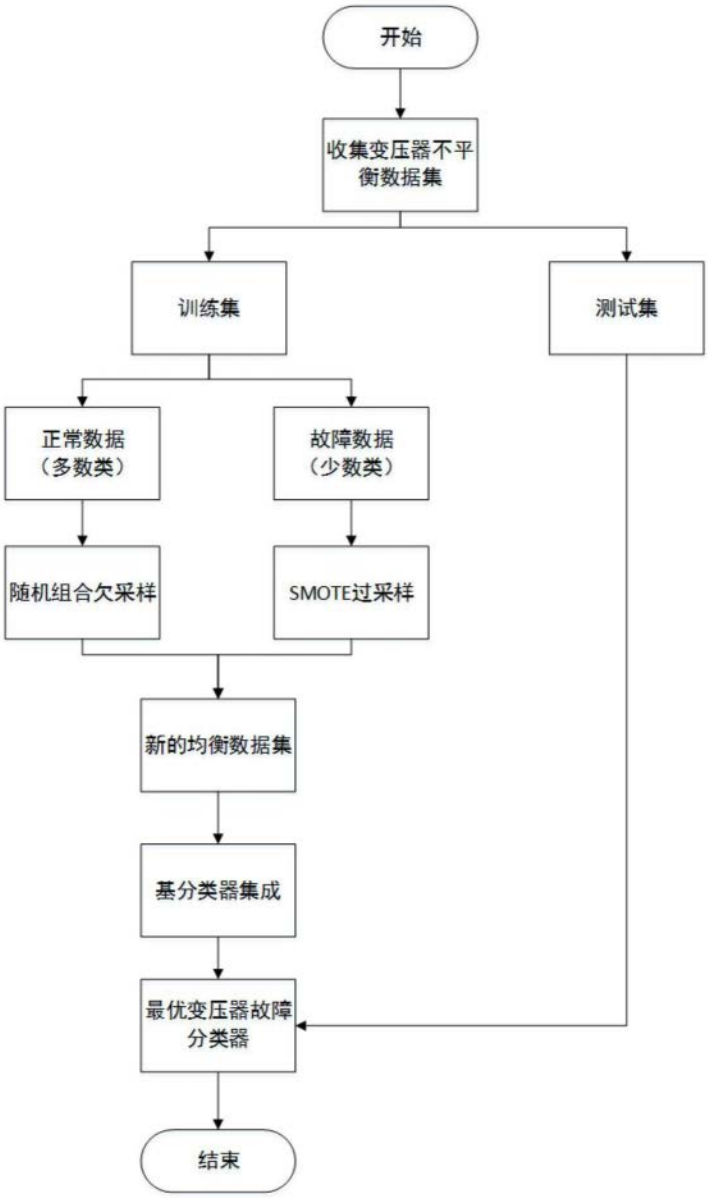

开始
收集变压器不平衡数据集
训练集
测试集
正常数据（多数类）
故障数据（少数类）
随机组合欠采样
SMOTE过采样
新的均衡数据集
基分类器集成
最优变压器故障分类器
结束

图1

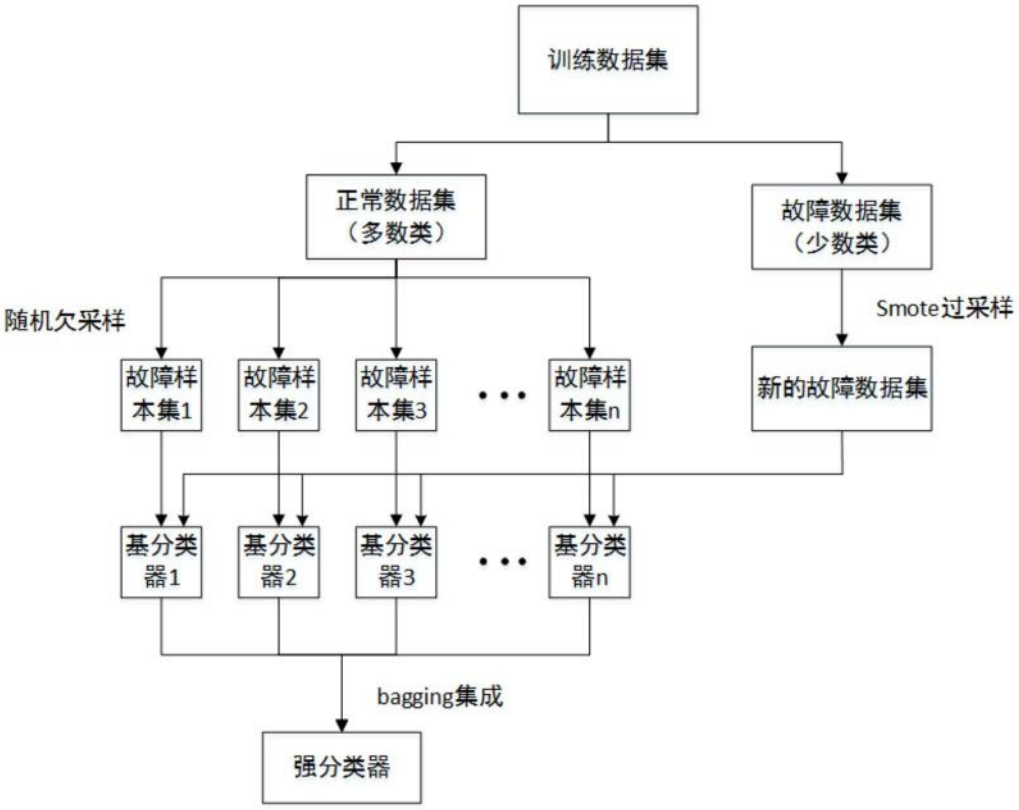

训练数据集
正常数据集
（多数类）
故障数据集
（少数类）
随机欠采样
Smote过采样
故障样本集1
故障样本集2
故障样本集3
故障样本集n
新的故障数据集
基分类器1
基分类器2
基分类器3
基分类器n
bagging集成
强分类器

图2

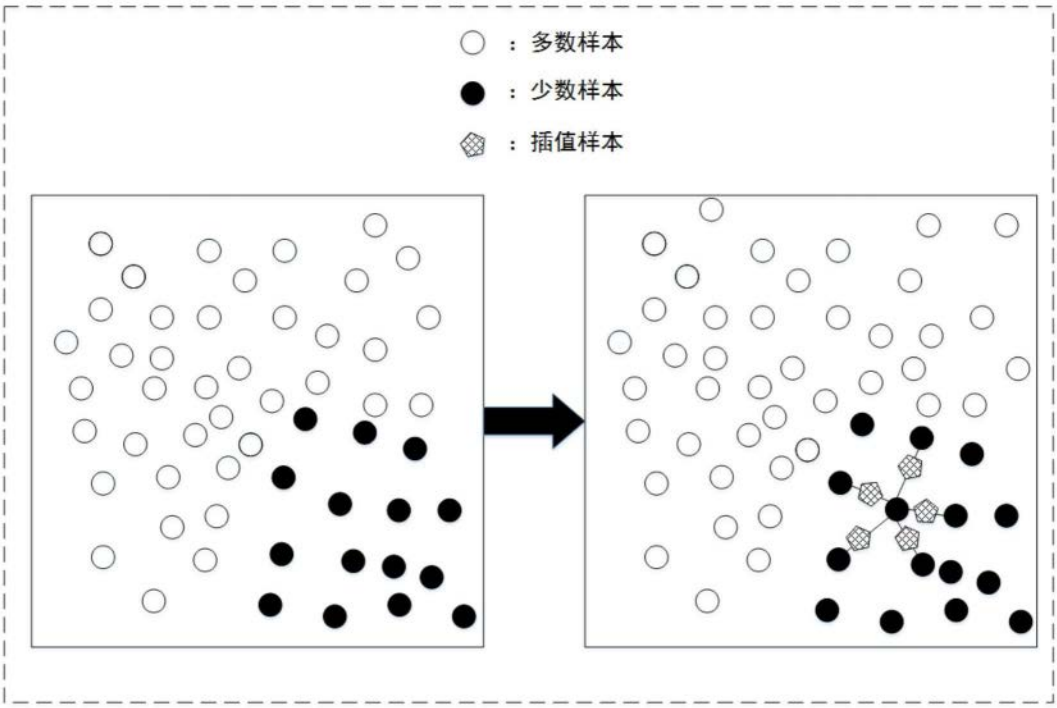

：多数样本
：少数样本
：插值样本

图3

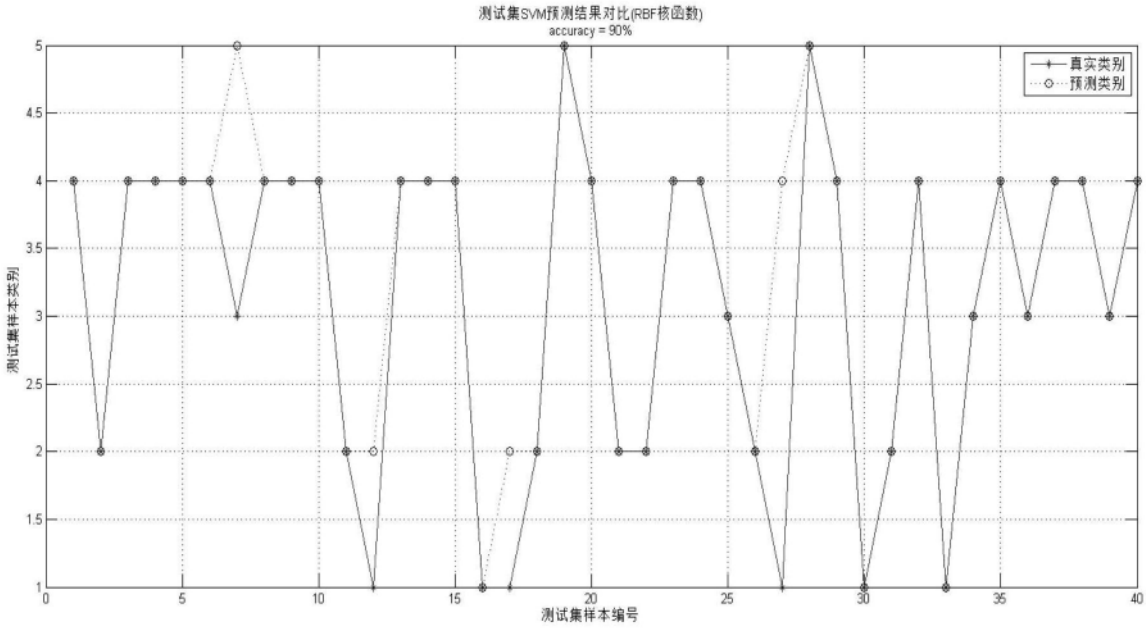

测试集SVM预测结果对比(RBF核函数)
accuracy = 90%
真实类别
预测类别
测试集样本类别
测试集样本编号
1
1.5
2
2.5
3
3.5
4
4.5
5
0
5
10
15
20
25
30
35
40

图4