DOI:10.12177/emca.2023.146 文章编号:1673-6540(2023)12-0021-11 中图分类号:TM 407 文献标志码:A

基于样本集成学习和 SO-SVM 的变压器故障诊断

刘可真[1], 姚 岳[1*], 赵现平[2], 杨春昊[2], 盛戈皞[3], 王 科[4]
(1.昆明理工大学 电力工程学院,云南 昆明 650500;
2.云南电网有限责任公司,云南 昆明 650200;
3.上海交通大学 电气工程系,上海 200240;
4.云南电网有限责任公司电力科学研究院,云南 昆明 650217)

Transformer Fault Diagnosis Based on Sample Integration Learning and SO-SVM

LIU Kezhen[1], YAO Yue[1*], ZHAO Xianping[2], YANG Chunhao[2], SHENG Gehao[3], WANG Ke[4]
(1. Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming 650500, China;
2. Yunnan Power Grid Co., Ltd., Kunming 650200, China;
3. Shanghai Jiaotong University, School of Electronic Information and Electrical Engineering, Shanghai 200240, China;
4. Electric Power Research Institute of Yunnan Power Grid Co., Ltd., Kunming 650217, China)

Abstract: Aiming at the problem of low accuracy of classification model caused by the unbalanced category of transformer fault samples, a transformer fault diagnosis model based on sample integration learning and snake optimisation algorithm (SO) optimised support vector machine (SVM) is proposed. The model first uses the EasyEnsemble sampler to generate multiple subsets with balanced categories after multiple under-sampling of the samples; then the SVM model optimised by SO with key parameters is trained with the Bagging strategy, and the final fault types are obtained by integrating the results of the classifiers. The validity of the proposed model is verified by the arithmetic example, and the data show that the diagnostic accuracy of SO-SVM′s fault diagnosis is improved by 3.44%, 6.89%, 10.92%, and the AUC value is improved by 0.026 4, 0.042 5, and 0.081 2, respectively, compared with the models of RF, SVM and KNN; in the same classifier, the SO-SVM model is more accurate than the SMOTE and ADASYN sample balancing methods, the diagnostic accuracy is improved by 4.59% and 2.87%, respectively, indicating that the SO-SVM model has better fault diagnosis capability for unbalanced samples.

Key words: transformer; sample ensemble learning; fault diagnosis; snake optimization algorithm

摘 要:针对变压器故障样本类别不平衡造成分类模型准确率偏低的问题,提出一种基于样本集成学习和蛇优化算法(SO)优化支持向量机(SVM)的变压器故障诊断模型。该模型先利用 EasyEnsemble 采样器对样本进行多次欠采样后生成类别平衡的多个子集;然后以 Bagging 策略训练 SO 优化关键参数后的 SVM 模型,综合各个分类器结果得到最终故障类型。通过算例对所提模型有效性进行验证,数据表明,SO-SVM 的故障诊断相比于 RF、SVM、KNN 等模型,诊断准确率分别提高了 3.44%、6.89%、10.92%,AUC 值分别提高了 0.026 4、0.042 5、0.081 2;在同一分类器下,SO-SVM 模型相比于 SMOTE 和 ADASYN 样本平衡方法,诊断准确率分别提高了 4.59%、2.87%,说明 SO-SVM 模型对不平衡样本的故障诊断能力更优。

关键词:变压器;样本集成学习;故障诊断;蛇优化算法

0 引言

能源是人类生存中不可或缺的物质基础,极大地推动着社会经济的高速发展,目前电能在我国终端能源消费中占比高达 26.8%,近十年增幅

基金项目:云南省教育厅科学研究基金资助项目(2022J1279);云南电网有限责任公司科技项目(YNKJXM20180736)

Funded by the Scientific Research Fund of Yunnan Provincial Department of Education (2022J1279); Yunnan Power Grid Co., Ltd., Science and Technology Project (YNKJXM20180736)

在世界主要经济体中最大,电气化整体程度位居世界前列[1]。随着远距离、大规模、高容量的电网发展,对输变电设备的安全稳定性也提出了更高的要求,作为电力系统中变换电压等级的关键设备,变压器可靠的运行极为重要。

变压器内部存在故障时,通常会产生大量的 CH_4、C_2H_6、C_2H_4 和 C_2H_2 等一系列低分子烃类,以及 H_2、CO 和 CO_2 等气体。油中溶解气体分析(DGA)被认为是最主要且有效的变压器故障诊断方法,能够在变压器运行过程中进行测定,不受外界因素的干扰。通过对上述油中溶解气体的类别及浓度进行定性定量分析,可以有效地判断出电力变压器的运行状况,提前发现内部存在潜伏故障,保证变压器能够长期稳定运行[2]。

基于 DGA 理论,国内外研究者提出了诸多的变压器故障诊断方法,早期的比值法、三角形法等,由于存在比值缺失,故障判别边界条件不清晰,无法完全反映特征气体与各类故障之间的隐藏规律[3-4]。随着计算机网络的发展,基于机器学习的智能诊断方法提高了对变压器各种故障类型的识别准确率。目前常用的机器学习算法包括专家系统[5]、SVM[6]、集成学习[7]和神经网络[8]。其中,具有扎实理论基础的 SVM 模型在小样本下泛化性能较好,被广泛应用于变压器故障诊断。然而,变压器实际运行中发生故障的概率较低,导致收集的数据集中各种故障类型样本数量存在较大差异。当类别不平衡的数据集的用于上述分类模型训练时,模型容易忽略少数类样本包含的特征信息,且模型训练时过度依赖少数类样本数据易出现过拟合,造成在新数据上的识别准确率较低[9]。

为降低不平衡数据的影响,从数据角度对数据进行欠采样和过采样。过采样以随机采样来增加样本,常用的方法主要有合成少数类过采样技术(SMOTE)[10]、改进过采样技术(Borderline-Smote)[11]和自适应合成抽样(ADASYN)[12]等。此类算法生成新样本时,存在边缘样本重叠的问题。其中,随机欠采样则以随机丢弃部分多数类样本,会造成故障特性信息的丢失,需要加以改进。

基于此,本文提出了一种基于样本集成学习的 SO-SVM 变压器故障诊断方法。首先,为解决人为设置 SVM 模型关键参数(惩罚系数 c 和核函数系数 g)不合理而导致分类性能降低[13],采用蛇优化算法(SO)对其进行优化。以优化参数后的 SVM 模型作为 Balanced Bagging Classifier 集成学习策略的基分类器,建立变压器故障诊断模型。该方法可通过 EasyEnsemble 采样器对样本集进行多次欠采样获得多个训练子集,然后以 Bagging 策略组合多个基分类器结果,得到最终诊断结果,实现对变压器运行状态的准确分析。

1 基本原理

1.1 支持向量机

SVM 算法的基本原理如图 1 所示,目标是求解一个能有效区分数据的超平面,需要在保证精度较高的同时满足两侧数据到超平面的几何间隔最大。

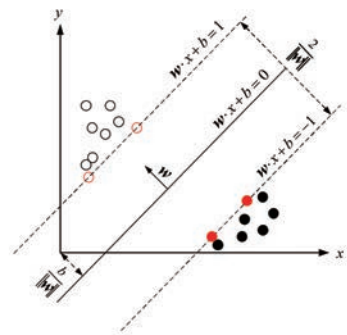

图 1 SVM 分类原理

Fig. 1 SVM classification principle

对于一般的二分类问题,假设正类记为+1,负类记为-1,对于给定的特征空间中的数据集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_k,y_k)\}$,其中 $x_i\in R^k$,代表第 i 个特征所包含的特征向量;$y_i\in\{-1,+1\}$,代表所属类别;$i\in(1,k)$。当数据在空间中线性可分时,记分隔超平面为 $w\cdot x+b=0$,数据集中的任意一点(x_i,y_i)到超平面的几何间隔为[14]

$$\gamma_i=y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot\boldsymbol{x}_i+\frac{b}{\|\boldsymbol{w}\|}\right) \tag{1}$$

式中:$\boldsymbol{w}$ 为超平面权重向量;b 为偏置参数。

根据以上定义,支持向量到超平面的距离可表示为 $\gamma=\min\limits_{i=1,2,\cdots,k}\gamma_i$,则求解 SVM 模型的最大超平面即为求解:$\max\limits_{\boldsymbol{w},b}\gamma$,约束为 s. t. $y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot\boldsymbol{x}_i+\frac{b}{\|\boldsymbol{w}\|}\right)\geqslant\gamma$;对此方程求 γ 的最大值,相当于求$\frac{1}{2}\|\boldsymbol{w}\|^2$的最小值,简化后的求

解问题为：$\min\limits_{\boldsymbol{w},b}\dfrac{1}{2}\|\boldsymbol{w}\|^2$，s. t. $y_i(\boldsymbol{w}\cdot x_i+b)\geqslant 1$。该问题为一个包含不等式约束的凸二次规划问题，为求解此问题，引入 Lagrange 乘子 α 将其构造为无约束目标函数：$L(\boldsymbol{w},b,a)=\dfrac{1}{2}\|\boldsymbol{w}\|^2-\sum\limits_{i=1}^{k}a_i(y_i(\boldsymbol{w}\cdot x_i+b)-1)$，此时求解目标为：$\min\limits_{\boldsymbol{w},b}\max\limits_{a_i\geqslant 0}L(\boldsymbol{w},b,a)$。利用 Lagrange 函数的对偶性，求解目标可转为：$\max\limits_{a_i\geqslant 0}\min\limits_{\boldsymbol{w},b}L(\boldsymbol{w},b,a)$，目标函数的极小值点需要满足函数对 $\boldsymbol{w}$ 和 b 偏导等于 0 的方程，即 $\boldsymbol{w}=\sum\limits_{i=1}^{k}a_iy_i\boldsymbol{x}_i$，$\boldsymbol{w}=\sum\limits_{i=1}^{k}a_iy_i=0$，代入目标函数后可得最终优化问题[15]：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i\alpha_jy_iy_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{i=1}^{k}\alpha_i$$

$$\text{s. t.}\quad\sum_{i=1}^{k}\alpha_iy_i=0\quad\alpha_i\geqslant 0,\ i=1,2,\cdots,k\tag{2}$$

式中：α 为拉格朗日乘子；k 为支持向量数量。

通过计算可得到最优值 $a^*=(a_1^*,a_2^*,\cdots,a_k^*)^{\mathrm{T}}$，此时 $\boldsymbol{w}$ 和 b 的最优值可表示为

$$\boldsymbol{w}^*=\sum_{i=1}^{k}a_i^*y_i\boldsymbol{x}_i\tag{3}$$

$$b^*=y_i-\sum_{i=1}^{k}\alpha_i^*y_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\tag{4}$$

式中：α^* 为拉格朗日乘子。

此时求得分离超平面为：$\boldsymbol{w}^*\cdot x+b^*=0$。相应的分类决策函数为[16]

$$f(x)=sign(\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*)=$$

$$sign\Big[\sum_{i=1}^{k}a_i^*y_i(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)+b^*\Big]\tag{5}$$

而在特征空间中的数据集线性不可分时，需要借助核函数对其作非线性变化映射到高维线性可分的特征空间中，在高维空间中训练线性支持向量机。假设核函数表示为 $K(x,z)$，初始特征在高维空间中的映射关系为 $\phi(x)$，则对初始输入特征空间中的 x、z，有如下变换：

$$K(x,z)=\phi(x)\cdot\phi(z)\tag{6}$$

式中：$K(x,z)$ 为核函数；$\phi(x)$ 为映射函数。

将线性可分下的支持向量机内积用核函数代替的分类决策函数为

$$f(x)=sign\Big[\sum_{i=1}^{k}a_i^*y_iK(\boldsymbol{x},\boldsymbol{x}_i)+b^*\Big]=$$

$$sign\Big[\sum_{i=1}^{k}a_i^*y_i\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x})+b^*\Big]\tag{7}$$

在模型的训练过程中，一个训练样本最终是否出现在模型参数表达式中与其 Lagrange 乘子 α 相关；只有 $\alpha>0$ 时保留，其对应的样本点处在最大间隔边界上，是一个与最终模型相关的支持向量。

1.2 蛇优化算法

蛇优化算法（SO）是 Hashim，F. A 和 Hussien，A. G 在 2022 年提出的新型智能仿生学优化算法[17]，算法模仿蛇的生活习性：分为觅食、战斗和繁殖模式。由于蛇类是冷血动物，其行为与生存的环境温度息息相关。若当前没有食物，雌雄个体都会寻找食物，个体之间会彼此远离以搜寻食物，搜索范围大，找出附近食物充足的区域。当食物充足后，蛇个体之间会通过信息共享，以保证没有得到足够食物的同伴可以快速获得食物，满足需求。若此时食物充足，且周围环境温度较低，雄性个体比较活跃，彼此之间便会有一定概率出现战斗情况吸引雌性个体的注意，交配后雌性可以决定是否产卵；周围环境温度较高，蛇个体会往食物位置靠近，即向全局最优位置靠近。算法的优化原理如图 2 所示。

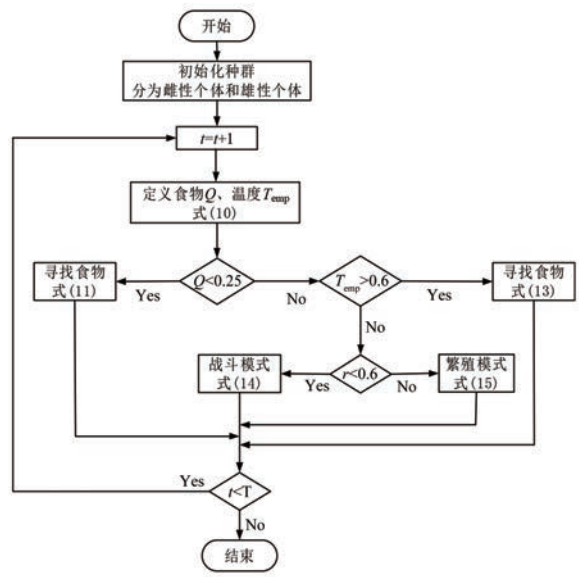

图 2 SO 优化过程

Fig. 2 SO optimization process

具体计算：建立 SO 算法的数学模型，假定蛇的种群数量为 n，待优化问题解的维度为 d，则蛇类个体的位置信息表达为

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\mathrm{d}} \\ x_{21} & x_{22} & \cdots & x_{2\mathrm{d}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{n}1} & x_{\mathrm{n}2} & \cdots & x_{\mathrm{nd}} \end{bmatrix} \tag{8}$$

同所有启发式算法相同,SO 算法需要生成均匀分布的随机种群位置,使得能够进行优化过程,个体位置初始化如式:

$$\boldsymbol{X}_i=\boldsymbol{X}_{\min}+r\times(\boldsymbol{X}_{\max}-\boldsymbol{X}_{\min}) \tag{9}$$

式中:$\boldsymbol{X}_i$ 为第 i 个体在所有维度下的位置集合;r 为[0,1]区间的随机数;$\boldsymbol{X}_{\max}$ 和 $\boldsymbol{X}_{\min}$ 为待优化问题的上下界。然后将初始化后的种群 1∶1划分为雌雄个体,雌性、雄性个体位置分为表示为 $X_{i,\mathrm{m}}$ 和 $X_{i,\mathrm{f}}$。

蛇类活动的环境温度系数 T_{emp} 和食物量 Q 如式(10)所示:

$$\begin{cases} T_{\mathrm{emp}}=\exp(-t/T) \\ Q=c_1*\exp[(t-T)/T] \end{cases} \tag{10}$$

式中:t 为当前迭代次数;T 为最大迭代次数。

当食物量 Q<0.25,算法进行全局搜索;此时蛇类个体随机选择位置去寻找食物,雌雄个体以式(11)更新其位置信息[18]:

$$\begin{cases} X_{i,\mathrm{m}}^{t+1}=X_{\mathrm{rand,m}}^{t}\pm c_2\times A_{\mathrm{m}}\times[(X_{\max}-X_{\min})\times r+X_{\min}] \\ X_{i,\mathrm{f}}^{t+1}=X_{\mathrm{rand,f}}^{t}\pm c_2\times A_{\mathrm{f}}\times[(X_{\max}-X_{\min})\times r+X_{\min}] \end{cases} \tag{11}$$

式中:$X_{t+1\,i,\mathrm{m}}$ 和 $X_{t+1\,i,\mathrm{f}}$ 分别为第 t+1 次迭代雄雌个体的空间位置;$X_{t\,\mathrm{rand,m}}$ 和 $X_{t\,\mathrm{rand,f}}$ 分别为第 t 次时雄雌个体的位置;c_2 为搜索因子,一般取 0.05;A_{m} 和 A_{f} 分别为雄雌个体搜寻食物的能力,计算式如下:

$$\begin{cases} A_{\mathrm{m}}=\exp\left(\dfrac{-f_{\mathrm{rand,m}}}{f_{i,\mathrm{m}}}\right) \\ A_{\mathrm{f}}=\exp\left(\dfrac{-f_{\mathrm{rand,f}}}{f_{i,\mathrm{f}}}\right) \end{cases} \tag{12}$$

式中:$f_{\mathrm{rand,m}}$ 和 $f_{\mathrm{rand,f}}$ 分别为雄雌个体的随机适应度值;$f_{i,\mathrm{m}}$ 和 $f_{i,\mathrm{f}}$ 分别为雄雌搜索代理的适应度值。

当食物量 Q>0.25,算法处于局部搜索状态,此条件下环境温度系数 T_{emp} 大于 0.6 时,蛇类个体只会往食物方向运动,按式(13)更新自身位置:

$$X_{i,\mathrm{j}}^{t+1}=X_{\mathrm{food}}\pm c_3\times T_{\mathrm{emp}}\times r\times(X_{\mathrm{food}}-X_{i,\mathrm{j}}^{t}) \tag{13}$$

式中:$X_{t+1\,i,\mathrm{j}}$ 和 $X_{t\,i,\mathrm{j}}$ 为个体 t + 1 和 t 次迭代时位置信息;X_{food} 为整体的全局最优值;c_3 为更新因子,一般取 2。

当环境温度系数 T_{emp} 小于 0.6 时,蛇类个体处于战斗或繁殖模式,战斗模式下雄雌个体按式(14)更新位置信息,繁殖模式下雄雌个体按式(15)更新位置信息:

$$\begin{cases} X_{i,\mathrm{m}}^{t+1}=X_{\mathrm{rand,m}}^{t}\pm c_3\times F_{\mathrm{m}}\times r\times(X_{\mathrm{best,f}}-X_{i,\mathrm{m}}^{t}) \\ X_{i,\mathrm{f}}^{t+1}=X_{\mathrm{rand,f}}^{t}\pm c_3\times F_{\mathrm{f}}\times r\times(X_{\mathrm{best,m}}-X_{i,\mathrm{f}}^{t}) \end{cases} \tag{14}$$

$$\begin{cases} X_{i,\mathrm{m}}^{t+1}=X_{\mathrm{rand,m}}^{t}\pm c_3\times M_{\mathrm{m}}\times r\times[(Q\times X_{i,\mathrm{f}}^{t})+X_{i,\mathrm{m}}^{t}] \\ X_{i,\mathrm{f}}^{t+1}=X_{\mathrm{rand,f}}^{t}\pm c_3\times M_{\mathrm{f}}\times r\times[(Q\times X_{i,\mathrm{m}}^{t})+X_{i,\mathrm{f}}^{t}] \end{cases} \tag{15}$$

式中:$X_{\mathrm{best,f}}$ 和 $X_{\mathrm{best,m}}$ 分别为雄雌个体的最优位置;M_{m} 和 M_{f} 分别为雄雌个体的繁殖能力;F_{m} 和 F_{f} 分别为雄雌个体的战斗值,计算如式(16)所示:

$$\begin{cases} F_{\mathrm{m}}=\exp\left(\dfrac{-f_{\mathrm{best,f}}}{f_i}\right) \quad F_{\mathrm{f}}=\exp\left(\dfrac{-f_{\mathrm{best,m}}}{f_i}\right) \\ M_{\mathrm{m}}=\exp\left(\dfrac{-f_{i,\mathrm{f}}}{f_{i,\mathrm{m}}}\right) \quad M_{\mathrm{f}}=\exp\left(\dfrac{-f_{i,\mathrm{m}}}{f_{i,\mathrm{f}}}\right) \end{cases} \tag{16}$$

式中:$f_{\mathrm{best,m}}$ 和 $f_{\mathrm{best,f}}$ 分别为雄雌个体的最优适应度值。

如果由后代产生,则由式(9)随机生成一个个体取代全局中适应度最差的个体[19]。

1.3 故障样本的 Ensemble

Balanced Bagging Classifier 是 Ensemble 集成分类器中的一种,原理如图 3 所示,其综合了 EasyEnsemble 采样器与分类器的 bagging 优点[20]。

EasyEnsemble 采样器的基本思路是将存在类别不平衡的数据集以少数类样本为基准,对多数类样本进行随机 k 次欠采样,分别和少数类样本进行组合,最终得到 k 份类别平衡的训练数据集。假设变压器有 m 种故障类型,按其样本数量由多到少进行排序得到样本集合$\{D_1,D_2,\cdots,D_m\}$。若最后一种少数类的样本 D_m 数量为$|S|$,对前 $m-1$ 种多数类样本随机重复 k 次独立的欠采样,每次采样过程中产生的子集记为 $D_{i,k}$,且每一子集在

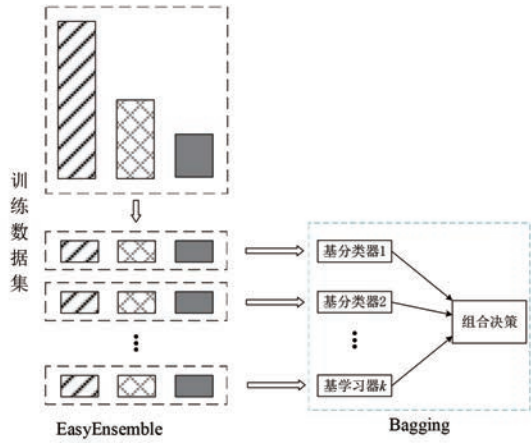

图 3 Balanced Bagging Classifier 原理

Fig. 3 Balanced Bagging Classifier principle

数量上有$|D_{i,k}|=|S|,i=1,2,\cdots,m-1$。

多数类样本 D_i 在以上采样过程下，其样本内部的任意一个样本(x_i,y_i)被抽到一次及以上的概率为$1-(1-1/|D_i|)^{|S|}$，则该样本在采样后形成的 k 个子集中出现一次及以上概率 P_1 和全部出现的概率 P 分别为

$$P_1=1-\left(1-\frac{1}{|D_i|}\right)^{|S|\cdot k} \tag{17}$$

$$P=\left[1-\left(1-\frac{1}{|D_i|}\right)^{|S|}\right]^{k} \tag{18}$$

当最少类样本数量$|S|$确定时，随着采样次数 k 的增大，样本(x_i,y_i)分布在所有训练子集的概率 P 也随之增大，这在一定程度上可减少欠采样方式造成的信息丢失。同时，当变压器其它故障类型样本数量与最少类数量倍数相差不大时，由式(18)可知样本(x_i,y_i)分布在所有训练子集中的概率比倍数相差较大的故障类型更高，在分类模型中的重视程度更高，确保了此类型故障样本的采样质量。将 k 次采样后得到的集合 $D_k=\{D_{1,k},D_{2,k},\cdots,D_{m-1,k}\}$ 与最少类样本集合 D_m 合并，最终故障类别相对平衡的训练子集记为 C_k。

Bagging 是对所有基分类器的结果进行平均化，降低模型过拟合的风险，从而减小输出结果的误差，将上述得到的平衡数据集 C_k 分别输入到基分类器进行学习，假设 $f_i(i=1,2,\cdots,k)$ 为每个分类器的决策结果，变压器的故障类型标识为$L_j(j=0,1,\cdots,7)$，则最终投票结果 f_{end} 为

$$f_{\text{end}}=\arg\max_j\sum_{f_i=L_j} i \tag{19}$$

式中：f_i 为每个分类器的决策结果；L_j 为变压器的故障类型标识。

2 考虑故障案例类别不平衡的 SO-SVM 模型

2.1 数据预处理及特征选取

在基于样本集成学习的 SO-SVM 变压器故障诊断模型中，选择 H_2 和 4 种烃类气体（CH_4、C_2H_6、C_2H_4 和 C_2H_2）共 5 类气体作为诊断模型的输入特征向量。由于特征尺度影响着模型的参数更新，各种特征气体含量量级差异较大，在训练模型过程中，尺度较大特征数据对模型的影响可能远大于尺度小的特征数据。为保证模型能够更好的识别特征中的潜在信息，加快模型的训练速度，采用式（20）将各种特征气体尺度缩放至区间[0,1]。

$$x^*=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{20}$$

式中：x 为原始样本气体序列；x^* 为缩放后的样本序列；$x_{\min}$、$x_{\max}$ 为样本各气体含量的最小值和最大值。

2.2 变压器数据样本划分和故障类型编码

根据 DL/T 722—2000 与 IEC 60599—2015 以及变压器内部故障下放电能量的强弱与温度的高低，划分变压器运行状态及编码如表 1 所示。

本文收集了南方电网公司变压器故障样例库中监测的油中溶解气体数据，将其划分为训练集和测试集，样例中各种故障类型分布如表 2 所示。部分不同故障类型下变压器油中溶解气体数据如表 3 所示。

表 1 变压器状态及编码

Tab. 1 Transformer status and coding

状态类型	简称	编码
正常	N	0
高能放电	D_1	1
低能放电	D_2	2
局部放电	D_3	3
高温过热	T_1	4
中温过热	T_2	5
低温过热	T_3	6
放电兼过热	DT	7

表 2 变压器不同类型数据分布

Tab. 2 Data distribution of different types of transformers

状态类型	总数量	训练集数量	测试集数量
N	650	585	65
D_1	285	256	29
D_2	140	126	14
D_3	128	115	13
T_1	274	247	27
T_2	105	94	11
T_3	82	74	8
DT	76	69	7

表 3 部分不同故障下气体数据

Tab. 3 Gas data under different faults

状态类型	气体浓度/($\mu L \cdot L^{-1}$)				
	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
N	8.01	2.65	0.53	2.12	0.86
N	14.65	3.71	10.52	2.69	0.22
D_1	62.31	11.19	1.86	10.01	14.63
D_1	44.96	8.09	1.09	18.88	26.98
D_2	30.60	29.80	6.82	16.01	99.00
D_3	41.48	4.88	5.76	0.49	10.73
T_1	10.99	111.16	9.44	29.35	4.44
T_1	11.01	102.41	57.99	277.6	0.00
T_2	0.32	13.31	1.99	3.71	0.19
T_3	14.52	8.84	2.33	3.61	0.14
DT	21.75	55.13	12.62	15.87	15.51

2.3 故障诊断模型

基于样本集成学习的 SO-SVM 变压器故障诊断方法如图 4 所示,以 SO 优化的 SVM 模型作为基础分类器,以 Balanced Bagging Classifier 样本的集成学习方法对基础分类器结果进行综合决策。模型的具体步骤为:

Step1:将收集到的油中溶解气体样本进行缩放处理后,按表所示划分为训练集和测试集。

Step2:在训练集中,以少数类样本为基准,对多数类样本进行 k 次采样,分别和少数类样本进行组合,最终得到 k 份类别平衡的训练数据集。

Step3:设置蛇优化算法种群数量 N,优化超参数维度 D,最大迭代次数 T,接着对个体位置进行初始化,计算适应度和个体最优位置,开始训练并计算个体在每一个训练集的适应度。

Step4:计算环境温度系数 T_{emp} 和食物量 Q,依据雄雌个体不同情况更新位置信息,即更新学习参数 c 和 g,得到当前适应度,与上一次迭代适应度相比较取最优适应度,使模型在训练集中分类精度最高。

Step5:当适应度不再变化或达到最大次数时终止迭代,获取当前 SVM 模型最优超参数 c 和 g,否则继续步骤 3。

Step6:使用 SO 算法得到的最优超参数更新 SVM 参数后得到 k 个 SO-SVM 分类器。

Step7:将测试集输入每一个 SO-SVM 模型,投票决定所有基础分类器上的结果,获得最终诊断的变压器故障类型。

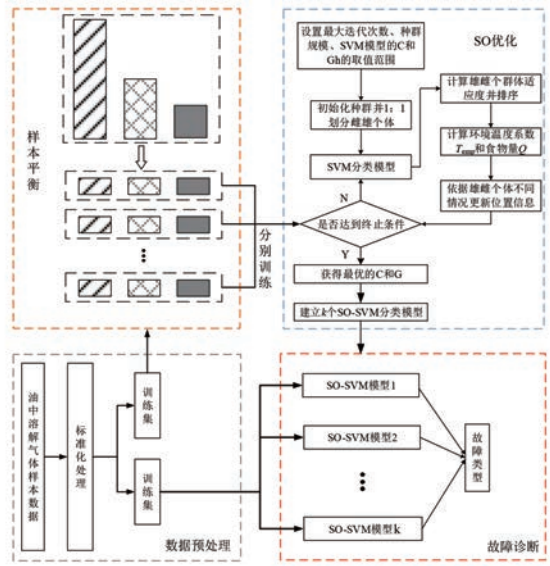

图 4 故障诊断技术路线

Fig. 4 Technical route for fault diagnosis

2.4 诊断结果评价指标

变压器故障诊断模型实质上是分类模型,混淆矩阵可以直观看出模型对各个故障类别的诊断表现,计算出相应准确率。如图 5 所示的二分类混淆矩阵,矩阵的主对角线上矩形块为该类标签被正确预测的个数,与对应类别总数相比即可得到其诊断准确率。

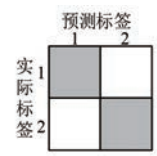

图 5 二分类混淆矩阵图

Fig. 5 Two-category confusion matrix diagram

当数据存在类别不平衡性时,受试者特征曲线(ROC 曲线)可以更为全面的评价模型的性能。计算曲线与横坐标之间的面积可得到分类模型的 AUC 值;同样可衡量模型的分类性能,一般认为

其值大于 0.5 时模型的分类结果具有参考意义，值越大越模型性能越好。

3 算例分析

3.1 SO 优化结果

由于 SVM 模型的分类准确率受到参数 c 和 g 的影响，在模型训练时采用 SO 算法其进行优化。SO 算法的参数设置为：种群数量 t：30；适应度值：模型准确率。算法达到收敛状态时，迭代了 12 次，此时适应度值为 0.891 27，训练过程中的模型参数变化如图 6 所示，最优参数设置和选取结果如表 4 所示。

表 4 最优参数选取结果

Tab. 4 Selection results of optimal parameters

参数	寻优范围	最优参数
c	(0,200)	50
g	(0.1,4)	0.6

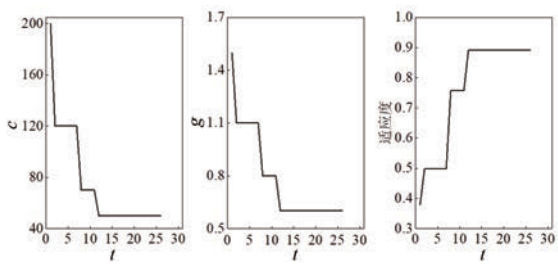

图 6 优化过程中参数变化图

Fig. 6 Parameter variation diagram during optimization process

3.2 模型诊断结果对比

本文在 Window10 系统、核心处理器及频率：AMD Ryzen 5 5600U 2.30 GHz、内存 16 GB、编程语言及版本为 Python3.9.7 环境下以表中数据对模型进行训练。为验证本文所提模型对变压器故障诊断的优越性，选择 RF、SVM、KNN 三种故障诊断模型作为对比模型，四种模型在测试上诊断结果的混淆矩阵如图 7 所示，对应诊断准确率结果见表 5。

从诊断结果可以看出，本文所提故障诊断方法平均准确率最高，可靠性高于三种对比模型。对测试集 174 个数据判断错误个数共 12 个，准确率为 93.10%，分别比 RF、SVM、KNN 模型减少了 6、12、19 个，诊断准确率提高了 3.44%、6.89%、10.92%。

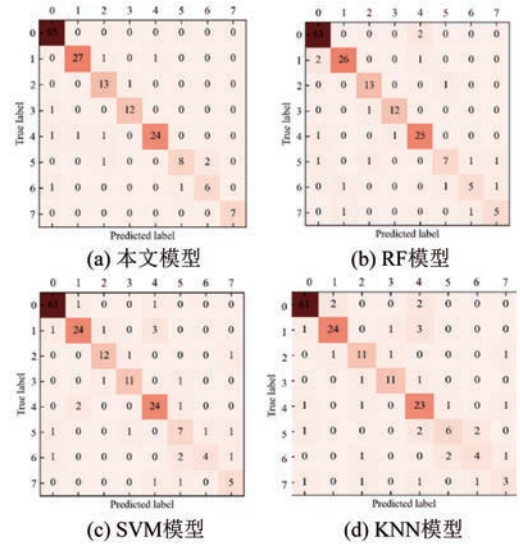

图 7 各模型诊断结果混淆矩阵

Fig. 7 Diagnostic results confusion matrix for each model

表 5 不同模型诊断准确率结果

Tab. 5 Diagnostic accuracy results of different models

故障类型	不同模型诊断准确率/%			
	KNN	SVM	RF	本文模型
N	93.85	96.92	96.92	100
D_1	82.76	82.76	89.66	93.10
D_2	78.57	85.71	92.86	92.86
D_3	84.62	84.62	92.31	92.31
T_1	85.19	88.89	92.59	88.89
T_2	54.55	63.64	63.64	72.73
T_3	50.00	50.00	62.50	75.00
DT	42.86	71.43	71.43	100
平均	82.18	86.21	89.66	93.10

分析少数类的中温过热、低温过热以及属于复合故障的放电兼过热等故障类型的泛化性，由于故障类别间的不平衡性，而 KNN 模型未对其进行有效处理，识别准确率较低，分别判断错误 5、4、4 个。SVM 由于核函数和对判断错误的惩罚系数 C 的存在，对少数类识别率稍有提升，分别判断错误 4、4、2 个。由于 RF 模型是基于决策树的 Bagging 学习策略，虽然随机抽样得到的训练子集仍存在不平衡性，但学习策略在一定程度上提高了对少数类的识别准确率，分别判断错误 4、3、2 个，表明了需要对样本的类不平衡进行一定的处理。而本文提出的基于样本集成学习的 SO-SVM 模型利用 EasyEnsemble 采样器对训练进行欠采样生成多个平衡样本后，以优化后的 SVM 为

Balanced Bagging Classifier 集成学习策略的基分类器,有效降低了样本间的类不平衡性,对少数类样本识别准确率整体最高,三种少数类样本分别判断错误 3、2、0 个。

将不同模型的诊断结果绘制为 ROC 曲线,如图 8 所示。从中可以看出本文模型的 AUC 面积最大为 0.960 8,相较于对比模型分别提高了 0.026 4、0.042 5、0.081 2,表明本文所提模型在测试集上的泛化性能更优,更加趋近完美故障分类器。

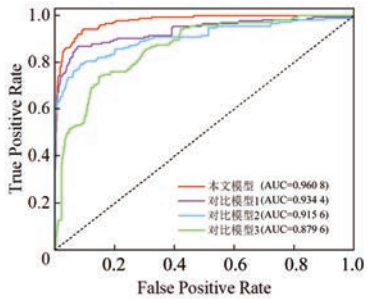

图 8 不同模型诊断结果 ROC 曲线

Fig. 8 ROC curves of diagnostic results for different models

3.3 不同样本平衡方法对比

为验证本文所提样本平衡方法的有效性,本章选择过采样方法 SMOTE、ADASYN 对本文的训练集进行平衡处理,以优化参数后的 SVM 作为分类器进行故障诊断。表 6 中列出了训练集经 SMOTE 和 ADASYN 过采样后各故障类型的数量分布,可以看出,这两种方法处理后,数据集类别同样相对平衡。

将以上两种方法平衡后的训练集分别输入优化后的 SVM 训练后,模型对各类故障的识别结果与本文所提模型对比情况如表 7 所示。从表中可以看出,通过 SMOTE 和 ADASYN 过采样平衡样本后,模型对于少数类样本诊断准确率都得到提高;在本文数据下,模型在 ADASYN 平衡后数据上整体表现更佳,平均准确率较 SMOTE 提高了 1.67%。但由于两种方法采样后都存在边缘样本的重叠问题,提升效果低于本文所提方法。训练集在本文模型下,诊断准确率较 SMOTE 方法和 ADASYN 方法,分别提高了 4.59%、2.87%,验证了所提模型的有效性。分析模型对所有故障类型的诊断结果,可以看出,放电兼过热故障对样本的不平衡性最敏感,在本文所提模型下所有数据均被正确识别。

表 6 不同平衡方式下的样本分布

Tab. 6 Sample distribution under different equilibrium modes

故障类型	原始数量	SMOTE 平衡	ADASYN 平衡
N	585	591	585
D_1	256	591	587
D_2	126	591	591
D_3	115	591	570
T_1	247	591	587
T_2	94	591	585
T_3	74	591	582
DT	69	591	584
总数	1 566	4 728	4 671

表 7 不同平衡方式下的准确率情况

Tab. 7 Accuracy under different balancing methods

故障类型	各平衡方式下准确率/%		
	SMOTE	ADASYN	本文方法
N	98.46	98.46	100
D_1	82.76	86.21	93.10
D_2	85.71	92.86	92.86
D_3	92.31	92.31	92.31
T_1	88.89	85.19	88.89
T_2	72.73	81.82	72.73
T_3	62.50	62.50	75.00
DT	71.43	85.71	100
平均	88.51	90.23	93.10

4 结语

本文针对变压器故障样本类别不平衡造成分类模型准确率偏低的问题,选择 H_2 和 4 种烃类气体(CH_4、C_2H_6、C_2H_4 和 C_2H_2)共 5 类气体作为诊断模型的输入特征向量,提出了一种基于样本集成学习和蛇优化算法(SO)优化支持向量机(SVM)的变压器故障诊断模型。该模型利用 EasyEnsemble 采样器进行多次欠采样,生成类别平衡的多个子集,然后使用 SO 优化关键参数后的 SVM 模型进行训练,并通过 Bagging 策略综合各个分类器结果得到最终故障类型。试验结果表明,相比于其他模型和方法,SO-SVM 的故障诊断准确率和 AUC 值都有明显提高,对不平衡样本的故障诊断能力更优。

1) 针对变压器故障样本类别不平衡问题,建立 Balanced Bagging Classifier 样本集成学习模型,其通过 EasyEnsemble 采样器对数据集进行多次

欠采样后生成多份类别平衡的训练子集，以 Bagging 策略分别训练基分类器后综合输出训练结果，提高了对不平样本的故障识别能力。

2）为降低 SVM 基分类器关键参数对模型性能的影响，采用 SO 算法对其进行优化，避免人为设置参数不合理造成分类准确率的问题。

3）算例分析表明，提出的故障诊断模型相比于 RF、SVM、KNN 模型，诊断准确率分别提高了 3.44%、6.89%、10.92%，AUC 值分别提高了 0.026 4、0.042 5、0.081 2。在同一分类器下，本文模型相比于 SMOTE 和 ADASYN 样本平衡方法，诊断准确率分别提高了 4.59%、2.87%，说明本文方法对不平衡样本的故障诊断能力更优。

参 考 文 献

[1] 杨昆. 新时代我国电力发展成就及展望[J]. 当代电力文化，2022，10：14-17.
YANG K. Achievements and prospects of China′s electric power development in the new era [J]. Contemporary Electric Power Culture，2022，10：14-17.

[2] 汪可，李金忠，张书琦，等. 变压器故障诊断用油中溶解气体新特征参量[J]. 中国电机工程学报，2016，36(23)：6570-6578+6625.
WANG K，LI J Z，ZHANG S Q，et al. New features derived from dissolved gas analysis for fault diagnosis of power transformers [J]. Proceedings of the CSEE，2016，36(23)：6570-6578+6625.

[3] ROGERS R R. IEEE and IEC codes to interpret incipient faults in transformers，using gas in oil analysis [J]. IEEE Transactions on Electrical Insulation，1978，EI-13(5)：349-354.

[4] 江秀臣，盛戈皞. 电力设备状态大数据分析的研究和应用[J]. 高电压技术，2018，44(4)：1041-1050.
JIANG X C，SHENG G H. Research and application of big data analysis of power equipment condition [J]. High Voltage Engineering，2018，44(4)：1041-1050.

[5] 师瑞峰，史永锋，牟军，等. 油中溶解气体电力变压器故障诊断专家系统[J]. 电力系统及其自动化学报，2014，26(12)：49-54.
SHI R F，SHI Y F，MOU J，et al. Power transformer fault diagnosis expert system with dissolved gas analysis in oil[J]. Proceedings of the CSU-EPSA，2014，26(12)：49-54.

[6] 张镱议，焦健，汪可，等. 基于帝国殖民竞争算法优化支持向量机的电力变压器故障诊断模型[J]. 电力自动化设备，2018，38(1)：99-104.
ZHNAG Y Y，JIAO J，WANG K，et al. Power transformer fault diagnosis model based on support vector machine optimized by imperialist competitive algorithm [J]. Electric Power Automation Equipment，2018，38(1)：99-104.

[7] 李楠，马宏忠，张玉良，等. 基于特征筛选和改进深度森林的变压器内部机械状态声纹识别[J]. 电机与控制应用，2022，49(9)：57-65+74.
LI N，MA H Z，ZHANG Y L，et al. Voiceprint recognition of internal mechanical status of transformers based on feature filtering and improved deep forest [J]. Electric Machines & Control Application，2022，49(9)：57-65+74.

[8] 霍浩，马天龙，李宁瑞，等. 基于贝叶斯与深度学习结合的变压器故障诊断[J/OL]. 电力系统及其自动化学报，2023-03-18. https://link.cnki.net/doi/10.19635/j.cnki.csu-epsa.001129.
HUO H，MA T L，LI N R，et al. Transformer fault diagnosis based on bayesian method and deep learning [J/OL]. Proceedings of the CSU-EPSA，2023-03-18. https://link.cnki.net/doi/10.19635/j.cnki.csu-epsa.001129.

[9] 崔宇，侯慧娟，苏磊，等. 考虑不平衡案例样本的电力变压器故障诊断方法[J]. 高电压技术，2020，46(1)：33-41.
CUI Y，HOU H J，SU L，et al. Fault diagnosis method for power transformer considering imbalanced class distribution [J]. High Voltage Engineering，2020，46(1)：33-41.

[10] 刘云鹏，和家慧，许自强，等. 基于 SVM SMOTE 的电力变压器故障样本均衡化方法[J]. 高电压技术，2020，46(7)：2522-2529.
LIU Y P，HE J H，XU Z Q，et al. Equalization method of power transformer fault sample based on SVM SMOTE [J]. High Voltage Engineering，2020，46(7)：2522-2529.

[11] 韩笑，王新迎，韩帅，等. 基于不均衡数据集成学习的大型电力变压器状态评价方法[J]. 电网技术，2021，45(1)：107-114.
HAN X，WANG X Y，HAN S，et al. Ensemble learning method for large-scale power transformer status evaluation based on imbalanced data [J].

Power System Technology, 2021, 45(1): 107-114.

[12] 刘可真, 梁玉平, 王科, 等. 基于数据过采样和深层特征提取的变压器故障诊断[J]. 电力科学与工程, 2022, 38(11): 9-16.
LIU K Z, LIANG Y P, WANG K, et al. Transformer fault diagnosis based on data oversampling and deep feature extraction [J]. Electric Power Science and Engineering, 2022,38(11): 9-16.

[13] 余松, 胡东, 唐超, 等. 基于 TLR-ADASYN 平衡化数据集的 MSSA-SVM 变压器故障诊断[J]. 高电压技术, 2021, 47(11): 3845-3853.
YU S, HU D, TANG C, et al. MSSA-SVM transformer fault diagnosis based on TLR-ADASYN balanced data set [J]. High Voltage Engineering, 2021, 47(11): 3845-3853.

[14] 徐萌. 基于 MA-SVM 方法的短期光伏功率预测[J]. 电机与控制应用, 2022, 49(7): 104-111.
XU M. Short-term photovoltaic power prediction based on MA-SVM method [J]. Electric Machines & Control Application, 2022, 49(7): 104-111.

[15] 李云溟, 咸日常, 张海强, 等. 基于改进灰狼算法与最小二乘支持向量机耦合的电力变压器故障诊断方法[J]. 电网技术, 2023, 47(4): 1470-1478.
LI Y H, XIAN R C, ZHANG H Q, et al. Fault diagnosis for power transformers based on improved grey wolf algorithm coupled with least squares support vector machine [J]. Power System Technology, 2023,47(4): 1470-1478.

[16] 郭艺伟, 谷爱昱, 曹文耀. 基于SVM-MOCDE 算法的永磁同步电机多目标优化[J]. 电机与控制应用, 2021, 48(12): 43-47.
GUO Y W, GU A Y, CAO W Y. Multi-objective optimization for PMSM based on SVM-MOCDE algorithm [J]. Electric Machines & Control Application, 2021,48(12):43-47.

[17] HASHIM F A, HUSSIEN A G. Snake optimizer: A novel meta-heuristic optimization algorithm [J]. Knowledge-Based Systems, 2022, 242: 108320-108353.

[18] 何伟, 陈薄, 贾清健, 等. 基于改进蛇优化算法的永磁同步电机电气参数辨识[J/OL]. 重庆大学学报, 2023-08-25. https://link.cnki.net/urlid/50.1044.N.20230627.1416.004.
HE W, CHEN B, JIA Q J, et al. Electrical parameters identification of permanent magnet synchronous motor based on improved snake optimization algorithm [J/OL]. Journal of Chongqing University, 2023-08-25. https://link.cnki.net/urlid/50.1044.N.20230627.1416.004.

[19] 王永贵, 赵炀, 邹赫宇等. 多策略融合的蛇优化算法及其应用[J/OL]. 计算机应用研究, 2023-08-25. https://link.cnki.net/doi/10.19734/j.issn.1001-3695.2023.05.0197.
WANG Y G, ZHAO Y, ZOU H Y, et al. Multi-strategy fusion snake optimizer and its application [J/OL]. Application Research of Computers, 2023-08-25. https://link.cnki.net/doi/10.19734/j.issn.1001-3695.2023.05.0197.

[20] JI D, WEI Z, TIAN C, et al. Deep transfer ensemble learning-based diagnostic of lithium-ion battery [J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(9): 1899-1901.

收稿日期:2023-06-28
收到修改稿日期:2023-08-28

作者简介:

刘可真(1974-),女,博士,教授,主要研究方向为电力设备状态监测与评估技术,liukzh@foxmail.com;

*通信作者:姚 岳(1999-),男,硕士研究生,主要研究方向为机器学习在电力设备故障诊断中的应用,keeping_go@163.com。

 DOI: 10.12177/emca.2023.146

Transformer Fault Diagnosis Based on Sample Integration Learning and SO-SVM

LIU Kezhen[1], YAO Yue[1*], ZHAO Xianping[2], YANG Chunhao[2], SHENG Gehao[3], WANG Ke[4]

(1. Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming 650500, China;
2. Yunnan Power Grid Co., Ltd., Kunming 650200, China;
3. Shanghai Jiaotong University, School of Electronic Information and Electrical Engineering, Shanghai 200240, China;
4. Electric Power Research Institute of Yunnan Power Grid Co., Ltd., Kunming 650217, China)

Key words: transformer; sample ensemble learning; fault diagnosis; snake optimization algorithm

As the key equipment for transforming the voltage level in the power system, the reliable operation of the transformer is extremely important. Improving the accuracy of power transformer fault diagnosis is of great significance to ensure the safe operation of the power grid. However, traditional fault diagnosis methods have problems such as low accuracy of classification models caused by unbalanced classification of transformer fault samples.

To address the problem of low accuracy of classification model caused by imbalance of transformer fault sample categories, this paper proposes a transformer fault diagnosis model based on sample integrated learning and Snake Optimisation Algorithm (SO) optimisation of Support Vector Machines (SVMs), and the specific steps of this method are shown in Fig.1:

In this paper, the accuracy of the SO-SVM model is verified by taking the oil and gas data of the transformer of the Southern Power Grid as an example. The characteristic gases of the transformer include H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, and eight types of faults are selected.

In order to verify the accuracy of fault diagnosis of the model proposed in this paper, three other similar algorithms are selected for comparison, and the results are shown that the diagnostic accuracy has been improved by 3.44%, 6.89%, and 10.92%, respectively.

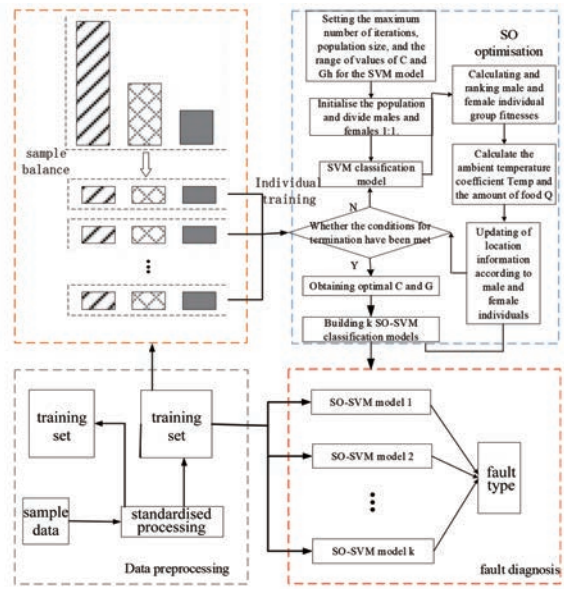

Fig. 1 Technical route for fault diagnosis

In order to verify the effectiveness of the sample balancing method proposed in this paper, two oversampling methods are selected to compare with the method in this paper, and the results are shown the accuracy of the method proposed in this paper is significantly improved.