

# 基于稀疏贝叶斯极限学习机的光伏电站设备故障诊断研究

孙 莉<sup>1</sup>, 李 静<sup>1</sup>, 李继云<sup>1</sup>, 王 磊<sup>2</sup>

(1. 东华大学计算机科学与技术学院, 上海 201620; 2. 上海肇星信息科技有限公司, 上海 200126)

**摘 要:** 基于运维数据针对光伏(PV)电站逆变器的故障诊断进行研究,提出一种基于稀疏贝叶斯极限学习机(SBELM)的精准的光伏逆变器故障诊断方法。首先分析逆变器故障数据特征,将该问题转化为一个多分类问题;然后,采用合成少数类过采样技术(SMOTE)方法人工生成数据,解决数据不平衡问题,根据环境和逆变器实时监控数据提取特征向量,并通过SBELM训练模型,可给出输出的概率分布,自动修剪冗余的隐藏节点,在不影响性能的前提下实现用部分节点进行多故障分类。通过实验分析,相比于其他故障诊断方法,SBELM诊断速度快且精度高,更适用于诊断光伏逆变器的故障。

**关键词:** 光伏电站设备; 故障诊断; 逆变器; 稀疏贝叶斯极限学习机; SMOTE; 机器学习

**中图分类号:** TK513.5

**文献标志码:** A

## 0 引 言

太阳能在日常生产生活中的应用已越来越普遍,光伏(PV)发电是其中最主要的一种<sup>[1]</sup>。光伏逆变器是光伏发电系统的重要环节,若逆变器发生故障,却没有及时诊断出来并处理,会造成不可挽回的经济损失和安全隐患。随着对光伏发电系统安全性和可靠性要求的不断提高,及时的故障诊断已经成为一个亟待解决的问题。因此,研究采用何种方式进行有效的光伏逆变器的故障诊断就变得异常重要<sup>[2]</sup>。

目前已有的光伏逆变器的故障诊断研究主要针对的是逆变器开关电路故障的研究,研究三相电流与故障之间的关系,通过三相电流判断逆变器的故障模式。文献[2]提出将隐马尔科夫模型(HMM)用于诊断光伏逆变器的故障,使其可适用于非线性系统的研究。在此基础上,文献[3]提出结合遗传算法与HMM的方法。文献[4]提出结合小波分解和自适应梯度下降动量BP神经网络实现故障诊断。文献[5]提出采用多分辨率信号分解技术提取特征,并利用模糊推理系统判断故障发生与否。文献[6]基于信号均值与 $rms$ 值,将组合逻辑和模糊逻辑应用于诊断故障,该方法能智能的

检测出逆变器故障。而随着数据量的增多,如何加快计算的速度,目前的方法并未考虑。此外,传统学习算法往往都存在着训练速度慢,局部最优之类的问题<sup>[7]</sup>,这也是需要解决的问题之一。

而对大量运维数据的分析发现,在实际运营中,逆变器的故障不仅仅局限于开关电路的故障,运维监控数据包含大量其他与逆变器的故障相关的数据。此外,故障数据存在样本分布不平衡问题,在现有的研究中并未考虑这一问题。因此基于运维数据,根据运维数据的特性研究如何有效实现精准的光伏逆变器的故障诊断十分重要。

基于以上问题,本文提出一种基于稀疏贝叶斯极限学习机(SBELM)方法针对运维数据的精准的光伏逆变器故障诊断方法,并将SVM、ELM、BELM与本文方法进行了比较。

## 1 问题定义

通过以上分析,整个故障诊断过程中有以下2个问题需解决:1)数据不平衡问题;2)基于运维数据进行精准的故障诊断。

定义从光伏逆变器数据中提取出的特征向量为

收稿日期: 2018-03-15

通信作者: 孙 莉(1962—),女,博士、副教授,主要从事数据库技术、数据挖掘、面向对象分析与设计方面的研究。sli@dhu.edu.cn

$X$ ,  $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbf{R}^n$ , 其中  $x_{ij}$  是一条故障中的一个特征数据,  $X_i$  代表一条故障数据; 故障的标签定义为  $t$ ,  $t_i \in \mathbf{R}^m$ , 一条  $X_i$  对应一个  $t_i$ ; 定义故障诊断的模型函数为  $f(x)$ , 因此光伏逆变器故障诊断的模型可表示为  $t = f(X)$ 。

## 2 光伏逆变器故障诊断方法

$f(x)$  的求解模型主要包括 2 个模型: 数据不均衡模型和故障诊断模型。针对数据不均衡问题, 本文采取 SMOTE 方法人工合成数据; 针对故障诊断问题, 采取 SBELM 模型。整个故障诊断的流程如图 1 所示。

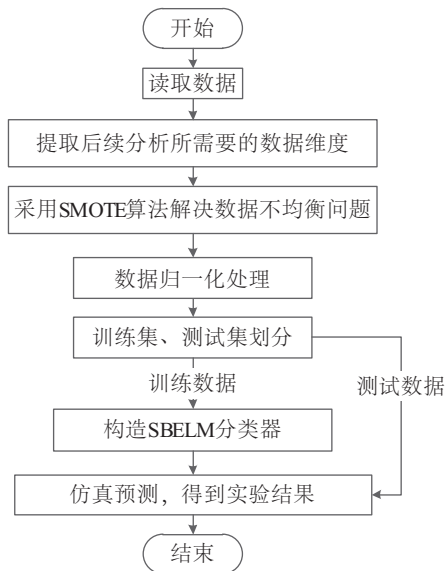


图 1 故障诊断流程图

Fig. 1 Flow chart of fault diagnosis

### 2.1 基于 SMOTE 的数据不平衡解决方法

由于数据不均衡, 采用常规方法则会导致小样本类别无法被识别出来。针对数据不平衡问题, 目前已有不少研究。文献[8]总结了针对不同问题不同的分类评价指标。比较常用的有过采样、欠采样等对数据进行重采样的方法[9]。合成少数类过采样技术(synthetic minority oversampling technique, SMOTE)是一种常用的人工生成新数据的方法[10]。根据文献[11], SMOTE 与 C4.5 决策树、集成学习等方法的结合, 都能有效提高该方法在不平衡数据集上的性能, 但是这些方法本身存在计算量大, 计算耗时长等问题。因此, 在本研究中, 采用 SMOTE 解决数据不平衡问题。

SMOTE 主要针对数据集中的数据量少的样本并为

这些样本合成新数据从而实现现有数据集的扩展, 该算法的主要过程如下。

假设训练集的一个样本数为  $T$  的少数类, 取其中一个样本  $i$ , 记其特征向量为  $x_i, i \in \{1, 2, \dots, T\}$ :

1) 从全部  $T$  个样本中找到该样本的  $k$  个近邻, 记为  $x_{i(near)}, near \in \{1, 2, \dots, k\}$ ;

2) 在这  $k$  个近邻中随机选择一样本  $x_{i(nm)}$ , 从而可计算出一个新样本为  $x_{i1} = x_i + s_i \cdot (x_{i(nm)} - x_i)$ , 其中  $s_i$  为 0~1 的随机数;

3) 将步骤 2) 重复进行  $N$  次, 合成  $N$  个新样本:  $x_{i1}, new \in \{1, 2, \dots, N\}$ 。

根据提取出的所需要数据维度, 即采集到实时的逆变器监控数据与环境监控数据, 构建故障的特征向量。从中找出少数类  $X_T$ , 由此可获得 SMOTE 的输入, 通过 SMOTE 人工合成数据, 可获得最终的数据集  $(X, t)$ 。

### 2.2 基于 SBELM 的光伏逆变器故障诊断方法

对不平衡数据集进行处理后, 基于获得的数据集  $(X, t)$  可采用 SBELM 实现精准的光伏逆变器的故障诊断。通过分析发现, 在故障编码中存在单一故障, 也存在组合故障, 根据组合故障的特性, 可通过 label powerset(LP)方法[12], 将该问题转化为多分类问题。根据故障及其对应的特征向量, 可构造基于 SBELM 模型的分分类器。

SBELM 在 ELM 的基础上, 通过贝叶斯方法学习获得输出权重, 因此首先介绍 ELM。

#### 2.2.1 ELM

ELM 是由黄广斌教授提出的一种单隐层前馈神经网络算法, 不需要进行迭代调参, 且具有学习耗时短, 高泛化能力的特点[13]。

定义一神经网络, 隐层数为 1, 其节点数为  $L$ , 定义数据集  $(X, t)$ ,  $X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbf{R}^n$ ,  $t_i = [t_{i1}, t_{i2}, \dots, t_{im}]^T \in \mathbf{R}^m$ 。该数据集在该神经网络中可表示为:

$$\sum_{i=1}^L \beta_i h(W_i \cdot X_j + b_i) = t_j, \quad j = 1, 2, \dots, N \quad (1)$$

式中,  $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$  —— 输出权重;  $W_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$  —— 输入权重;  $h(x)$  —— 激活函数;  $b_i$  —— 第  $i$  个隐层单元的偏置。

式(1)可用矩阵表示为:

$$H\beta = T \quad (2)$$

式中,  $H$  ——隐层输出;  $\beta$  ——输出权重;  $T$  ——期望输出。

若  $W_i$  和  $b_i$  被确定,  $H$  就被唯一确定, 网络的输出权重也将被确定, 即:

$$\hat{\beta} = H^+ T \quad (3)$$

式中,  $H^+$  ——  $H$  的 Moore-Penrose 广义逆。

### 2.2.2 SBELM

SBELM 使用贝叶斯方法获得输出权重, 能有效避免过拟合问题<sup>[14]</sup>。因而 SBELM 的输入将变为隐层的输出  $H$ 。根据特征向量  $X$ , 随机选取的  $W_i$  和  $b_i$ , 可得出 SBELM 的输入, 再结合标签  $t$ , 根据 SBELM 的公式求解出输出权重  $\hat{\beta}$  以及合适的隐层神经元数, 即可构建故障诊断模型。

在 SBELM 中, 每个样本都被视作一个独立的 Bernoulli 试验, 其概率分布  $p(t|\alpha)$  就可表示为 Bernoulli 分布。似然函数可表示为:

$$p(t|\beta, H) = \prod_{i=1}^N \sigma\{y(h_i; \beta)\}^{t_i} [1 - \sigma\{y(h_i; \beta)\}]^{1-t_i} \quad (4)$$

式中,  $t = [t_1, t_2, \dots, t_N]^T$ ,  $t_i \in \{0, 1\}$ ;  $\sigma(\cdot)$  ——激活函数,  $\sigma[y(h_i; \beta)] = \frac{1}{1 + e^{-y(h_i; \beta)}}$ ,  $y(h_i; \beta) = h_i \beta$ ;  $\beta = [\beta_1, \beta_2, \dots, \beta_L]^T$ 。

对于每一个  $\beta_i$  都有一个独立的自动关联性测定 (ARD) 超参数  $\alpha_i$  与之相对应, ARD 先验通过控制一些  $\beta_i$  使其为 0 来选择显著的隐藏神经元。

对  $\beta$  积分得到  $t$  的边际分布,  $\alpha$  的值则可根据最大化边际似然来确定, 表示为:

$$p(t|\alpha, H) = \int p(t|\beta, H) p(\beta|\alpha) d\beta \quad (5)$$

通过 Laplace 近似和迭代加权最小二乘法 (IRLS),  $\hat{\beta}$  和  $\Sigma$  ( $\beta$  上的高斯分布的协方差矩阵) 可表示为:

$$\hat{\beta} = \sum H^T B \hat{\beta} \quad (6)$$

$$\Sigma = (H^T B H + A)^{-1} \quad (7)$$

式中,  $B$  ——  $N \times N$  的对角矩阵,  $b_i = y_i(1 - y_i)$ ;  $\hat{\beta} = H \beta + B^{-1}(t - y)$ ,  $y = [y_1, y_2, \dots, y_N]^T$ ;  $A = \text{diag}(\alpha)$ 。

通过边际似然, 可求得  $\alpha$  为:

$$\alpha_k^{\text{new}} = \frac{1 - \alpha_k \Sigma_{kk}}{\hat{\beta}_k^2} \quad (8)$$

初始化  $\beta_k$  和  $\alpha_k$ ,  $\hat{\beta}$ 、 $\Sigma$  和  $\alpha$  可通过式 (6) ~ 式 (8) 进行更新, 直至收敛。一旦  $\hat{\beta}$  收敛, 模型的输出权重即可确定, 当有新数据  $x_{\text{new}}$  到来时, 模型的输出  $p(t_{\text{new}} | x_{\text{new}}, \hat{\beta})$ , 即故障的诊断结果则可求得。

ARD 先验机制通过将部分输出权重调整为 0, 达到

修剪相关的隐藏神经元的目的。根据文献[15], 通过对边际似然函数取对数, 就可得到超参数  $\alpha_k$ , 根据该参数可实现节点的稀疏。

$\alpha_k$  的静态点可表示为:

$$\hat{\alpha}_k = \begin{cases} \frac{s_k^2}{q_k^2 - s_k}, & q_k^2 > s_k \\ \infty, & q_k^2 \leq s_k \end{cases} \quad (9)$$

式中,  $s_k = h_k^T C_k^{-1} h_k$ ,  $C = B + H A H^T$ ;  $q_k = h_k^T C_k^{-1} \hat{t}$ ;  $h_k$  ——  $H$  的第  $k$  列。

通过对  $\alpha$  的不断迭代, 使得边际似然最大化, 存在一些  $\alpha_k$  趋向于无穷大, 这将对均值  $\hat{\beta}$  和协方差  $\Sigma$  产生如下影响:

$$\lim_{\alpha_k \rightarrow \infty} \Sigma_{kk} = \lim_{\alpha_k \rightarrow \infty} \{h_k^T b_i h_k + \alpha_k\}^{-1} = 0 \quad (10)$$

$$\lim_{\alpha_k \rightarrow \infty} \hat{\beta}_k = \lim_{\alpha_k \rightarrow \infty} \{\varepsilon_k h_k^T b_i \hat{t}\} = 0 \quad (11)$$

式中,  $\varepsilon_k$  ——  $\Sigma$  的第  $k$  行, 式 (10) 和式 (11) 表明  $\hat{\beta}_k \sim N(0, 0)$ , 相当于 0。因此, 相对应的神经元得到修剪, 从而达到了稀疏的结果。

根据求解出的  $\hat{\beta}$  和选择出的显著神经元, 可构建 SBELM 模型。当有新的故障数据到来时, 根据构造出的 SBELM 模型就可实现故障诊断。

## 3 实验分析

### 3.1 实验环境

实验的硬件环境是 CPU 为 3.40 GHz 的 Intel® Core i7-3770M, 8G 的 RAM, 软件环境是 Windows7 操作系统、Matlab R2016a。SVM 的执行环境为 Python 3.5, 相应的算法调用 sklearn 中的 SVM 算法。

### 3.2 数据集

本文中所用到的数据集为 2015 年和 2016 年实际光伏电站运营中产生的运维数据, 主要有: 逆变器的监控数据 (5 min 采集 1 次) 包括直流输入电流电压, 交流输出电流电压, 当前 PV 功率, 当前 PV 工作时长, 采集时间等实时监控数据; 环境监测数据 (5 min 采集 1 次) 包括逆变器组件温度, 发电量, 功率, 温度, 湿度, 辐射, 风速, 采集时间等实时监控数据; 逆变器故障数据包括故障设备, 故障编码, 故障发生时间等实时监控数据。本文主要针对 4 种常用的逆变器进行实验, 分别为: Sungrow、Delta、Huawei 和深圳古瑞瓦特。表 1 列出 4 种品牌中具体的故障类别数与故障总数目。表 2 给出具体的故障及其数目 (以 Delta 为





确率。图 3 给出了 SBELM 的初始节点数量和剩余隐藏节点数量的变化图,图中横坐标表示初始设置的隐层节点数量,纵坐标表示经过模型稀疏化后剩余的隐层节点数量。

在实验中,通过 SBELM 方法,可得出每种故障发生的概率,实验的部分结果如表 4 所示。在本实验中选取了 13 种重度故障,因而有 13 列数据,每一列为不同的故障类别,共有 13 类,每一行为一个测试样例属于这 13 种不同故障的概率。给出每种故障发生的概率,而不是单一给出一种预测结果,这在实际生产生活中有更大的现实意义。在光伏逆变器的故障诊断问题中,SBELM 相比于其他方法显然具有更好的实用性。

表 4 每种故障发生的概率

Table 4 Probability of each fault

编号	1	2	3	4	5	6	7	8	...	11	12	13
1	1.00	$5.17 \times 10^{-6}$	$3.89 \times 10^{-6}$	$1.81 \times 10^{-7}$	$2.70 \times 10^{-6}$	$3.27 \times 10^{-6}$	$1.50 \times 10^{-6}$	$1.82 \times 10^{-6}$	...	$1.41 \times 10^{-6}$	$3.45 \times 10^{-8}$	$2.05 \times 10^{-7}$
2	1.00	$4.74 \times 10^{-6}$	$3.44 \times 10^{-6}$	$1.82 \times 10^{-7}$	$2.66 \times 10^{-6}$	$3.18 \times 10^{-6}$	$1.53 \times 10^{-6}$	$1.95 \times 10^{-6}$	...	$1.49 \times 10^{-6}$	$3.44 \times 10^{-8}$	$1.93 \times 10^{-7}$
3	1.00	$5.32 \times 10^{-6}$	$4.06 \times 10^{-6}$	$1.74 \times 10^{-7}$	$2.72 \times 10^{-6}$	$3.28 \times 10^{-6}$	$1.51 \times 10^{-6}$	$1.83 \times 10^{-6}$	...	$1.41 \times 10^{-6}$	$3.65 \times 10^{-8}$	$2.13 \times 10^{-7}$
4	$7.07 \times 10^{-7}$	$1.41 \times 10^{-6}$	$4.59 \times 10^{-6}$	$9.53 \times 10^{-5}$	$4.52 \times 10^{-6}$	$1.72 \times 10^{-6}$	$2.55 \times 10^{-6}$	$1.28 \times 10^{-5}$	...	$1.02 \times 10^{-5}$	$1.41 \times 10^{-4}$	$9.99 \times 10^{-1}$
5	$6.68 \times 10^{-5}$	$8.17 \times 10^{-1}$	$1.65 \times 10^{-1}$	$5.24 \times 10^{-5}$	$2.20 \times 10^{-3}$	$5.10 \times 10^{-3}$	$9.10 \times 10^{-3}$	$8.73 \times 10^{-5}$	...	$7.15 \times 10^{-5}$	$1.11 \times 10^{-4}$	$5.41 \times 10^{-5}$
...	...	...	...	...	...	...	...	...	...	...	...	...
26661	1.00	$4.93 \times 10^{-6}$	$3.56 \times 10^{-6}$	$2.18 \times 10^{-6}$	$2.64 \times 10^{-6}$	$3.24 \times 10^{-6}$	$1.52 \times 10^{-6}$	$2.02 \times 10^{-6}$	...	$1.49 \times 10^{-6}$	$3.25 \times 10^{-8}$	$1.81 \times 10^{-7}$

## 4 结 论

基于电站运维数据,为实现精准的光伏逆变器故障的及时准确诊断,本文选取逆变器周围环境监控数据,逆变器监控数据构建特征向量;采用 SMOTE 为数据集的小样本数据人工合成新样本来解决运维数据中存在的样本不平衡问题;根据运维数据的特性,采用 label powerset(LP)方法,将该问题转化为多分类问题;针对传统学习算法存在训练速度慢,局部最优等问题,将稀疏贝叶斯极限学习机引入光伏逆变器的故障诊断中。实验结果表明,相比于其他故障诊断方法,SBELM 方法的诊断精度最高,比 SVM 方法高近 7%,可达到 98.01% 的准确率,在相同测试集上,相较于 BELM 和 SVM,SBELM 的测试时间只有这 2 种方法耗时的 1/2 和 1/4,诊断耗时较短,其性能明显优于其他常用故障诊断方法。

### [参考文献]

[1] 于静,车俊铁,张吉月. 太阳能发电技术综述[J]. 世

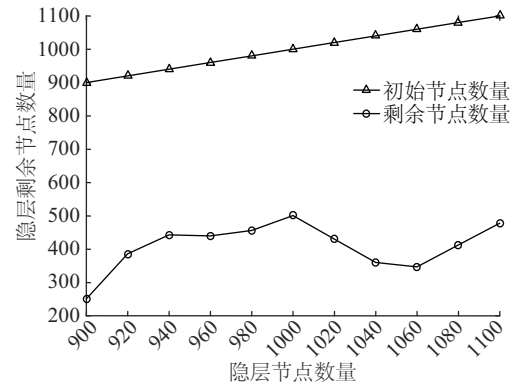


图 3 SBELM 中初始和剩余隐藏节点数量的变化图

Fig. 3 Change diagram of the number of initial hidden nodes and number of remaining hidden nodes in SBELM

界科技研究与发展, 2008, 30(1): 56-59.

YU J, CHE J T, ZHANG J Y. The overview of the technique of solar energy generating electricit[J]. World sci-tech R&D, 2008, 30(1): 56-59.

- [2] ZHENG H, WANG R, WANG Y, et al. Fault diagnosis of photovoltaic inverters using hidden Markov model [C]//Control Conference, Dalian, China, 2017.
- [3] ZHENG H, WANG R Y, WANG Y F, et al. Combining a HMM with a genetic algorithm for the fault diagnosis of photovoltaic inverters[J]. Journal of power electronics, 2017, 17(4): 1014-1026.
- [4] WAN X F, LIU Q, DU L P, et al. Fault diagnosis of two-level inverter in photovoltaic grid-connected system [J]. Electrical measurement & instrumentation, 2016, 53(10): 122-128.
- [5] YI Z, ETEMADI A. Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems [J]. IEEE transactions on smart grid, 2017, 8(3): 1274-1283.
- [6] CHEN W, BAZZI A. Logic-based methods for

- intelligent fault diagnosis and recovery in power electronics [J]. IEEE transactions on power electronics, 2017, 32(7): 5573-558.
- [7] WITCZAK M, KORBICZ J, MRUGALSKI M, et al. A GMDH neural network-based approach to robust fault diagnosis: Application to the DAMADICS benchmark problem [J]. Control engineering practice, 2006, 14(6): 671-683.
- [8] 王成, 刘亚峰, 王新成, 等. 分类器的分类性能评价指标[J]. 电子设计工程, 2011, 19(8): 13-15, 21.  
WANG C, LIU Y F, WANG X C, et al. Appraisal identification of classifier's performance [J]. Electronic design engineering, 2011, 19(8): 13-15, 21.
- [9] CHAWLA N V. Data mining for imbalanced datasets: An overview [M]. Boston: Springer US, 2005: 853-867.
- [10] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002, 16(1): 321-357.
- [11] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches [J]. IEEE transactions on systems man & cybernetics part C applications & reviews, 2012, 42(4): 463-484.
- [12] TSOUMAKAS G, VLAHAVAS I. Random k-Labelsets: An ensemble method for multilabel classification [J]. IEEE transactions on knowledge & data engineering, 2011, 23(7): 1079-1089.
- [13] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications [J]. Neurocomputing, 2006, 70(1): 489-501.
- [14] SORIA-OLIVAS E, GOMEZ-SANCHIS J, MARTIN J D, et al. Belm: Bayesian extreme learning machine [J]. IEEE transactions on neural networks, 2011, 22(3): 505-509.
- [15] LUO J, VONG C M, WONG P K. Sparse bayesian extreme learning machine for multi-classification [J]. IEEE transactions on neural networks & learning systems, 2014, 25(4): 836-843.

## RESEARCH ON FAULT DIAGNOSIS OF PHOTOVOLTAIC POWER STATION EQUIPMENT BASED ON SPARSE BAYESIAN EXTREME LEARNING MACHINE

Sun Li<sup>1</sup>, Li Jing<sup>1</sup>, Li Jiyun<sup>1</sup>, Wang Lei<sup>2</sup>

(1. School of Computer Science and Technology, Donghua University, Shanghai 201620, China;

2. Shanghai Bo Xing Information Technology Co. Ltd., Shanghai 200126, China)

**Abstract:** Based on operation and maintenance data and aiming at the fault diagnosis of PV inverter, a precise PV inverter fault diagnosis method based on Sparse Bayesian extreme learning machine (SBELM) is proposed in this paper. First, by the analysis of inverter fault data characteristics, the problem is transformed into a multi classification problem. Then, using Synthetic Minority Oversampling Technique (SMOTE) method to generate data, the imbalanced data problem is solved, according to the environment data and the inverter monitoring data to extract the feature vector, and train model through the SBELM. This method can estimate the probability distribution of the model output can be estimated, the model can automatically eliminate useless hidden nodes, and realize the purpose of multiple fault classification with some hidden nodes without affecting the performance. The experimental results show that, compared with other fault diagnosis methods, SBELM is fast and accurate, and is more suitable for the fault diagnosis of photovoltaic inverters.

**Keywords:** PV power station equipment; fault diagnosis; photovoltaic inverters; sparse Bayesian extreme learning machine; SMOTE; machine learning