文章编号:1671-4598(2025)07-0243-09 DOI:10.16526/j.cnki.11-4762/tp.2025.07.030 中图分类号:TP274 文献标识码:A

基于时变隐马尔科夫模型的风电机组SCADA数据异常状态智能判定方法

陈秀康[1,2]
(1.广西农业职业技术大学,南宁 535427;2.广西民族大学,南宁 530006)

摘要:由于风电机组运行受风速变化、恶劣环境和人为限电等影响,其SCADA系统包含多种异常数据,直接使用原始数据进行分析会导致偏差,从而降低异常状态检测精度;为此,提出基于时变隐马尔科夫模型的风电机组SCADA数据异常状态智能判定方法;使用滑动窗口技术和数据增广状态矩阵对风电机组SCADA数据进行平滑处理得到去噪数据,为后续的数据异常状态智能判定提供高质量数据;根据去噪后数据的特征进行数据分类处理,获得不同子集;利用时变隐马尔科夫模型对比正常数据属性和分类后子数据集属性,若二者一致表明数据状态正常,若不一致则表明数据异常,从而实现风电机组SCADA数据异常状态智能判定;实验结果表明,该方法可以精准判定风电机组SCADA数据异常状态,召回率为0.98,漏报率最大值为4.13%,F_1值最大达到0.96,能够确保SCADA系统的安全稳定运行。

关键词:时变隐马尔科夫模型;风电机组SCADA数据;数据异常状态;状态智能判定;滑动窗口技术

Intelligent Determination Method for Scada Data Abnormal States of Wind Turbines Based on Time-Varying Hidden Markov Model

CHEN Xiukang[1,2]
(1. Guangxi Vocational University of Agriculture, Nanning 535427, China;
2. Guangxi Minzu University, Nanning 530006, China)

Abstract: Due to the influences of wind speed changes, harsh environments, and human power restrictions on the operation of wind turbines, there are various abnormal data in supervisory control and data acquisition (SCADA) systems. Directly analyzing raw data can lead to deviations and reduce the accuracy of abnormal state detection. Therefore, an intelligent determination method for abnormal states in the SCADA data of wind turbines based on time-varying hidden Markov model is proposed. sliding window technology and data augmentation state matrix are used to smooth the SCADA data of wind turbines and obtain denoised data, providing high-quality data for the intelligent determination of abnormal data states in the future. Classify and process the data based on its features after denoising to obtain different subsets. By using a time-varying hidden Markov model, a comparsion of the attributes of normal data with the attributes of classified sub datasets is made, which shows that the data state is normal for the consistent of data attributes, and the data state is abnormal for the inconsistent, thus achieving the intelligent determination of abnormal states in wind turbine SCADA data. Experimental results show that this method can accurately determine the abnormal status of wind turbine SCADA data, with a recall rate of 0.98, a maximum false alarm rate of 4.13%, and an F_1 value of up to 0.96, ensuring the safe and stable operation of the SCADA system.

Keywords: time-varying hidden Markov model; wind turbine SCADA data; abnormal data status; intelligent state determination; sliding window technology

0 引言

在现代工业控制系统中,数据监控及采集系统(SCADA, supervisory control and data acquisition)扮演着至关重要的角色[1]。作为连接现场设备与企业信息

收稿日期:2024-12-24; **修回日期**:2025-02-14。
基金项目:国家自然科学基金(0710FD221054080)。
作者简介:陈秀康(1987-),男,硕士,工程师。
引用格式:陈秀康.基于时变隐马尔科夫模型的风电机组SCADA数据异常状态智能判定方法[J].计算机测量与控制,2025,33(7):243-251.

管理系统的桥梁，SCADA 能够实时、准确地采集生产过程中的各类数据[2]，并通过图形化界面展示给操作人员，从而实现对生产全过程的全面监控与远程控制。特别是在风电机组监控领域，SCADA 系统能够实时记录风电机组的各项关键运行参数，如功率、转速、温度、风速等，为风电场的运维管理和优化提供了宝贵的数据支持。然而，随着工业规模的不断扩大和复杂性的增加，风电机组运行环境多变、设备复杂度高，SCADA 数据量呈爆炸式增长，数据质量的问题也日益凸显，其中数据异常状态的判定尤为关键。数据异常状态是指在数据采集、传输或处理过程中，由于设备故障、环境干扰、人为误操作或系统内部逻辑错误等原因，导致的数据值偏离正常范围或预期规律的现象[3]。这些异常数据不仅会影响操作人员对生产过程的准确判断，还会误导决策制定[4]，甚至引发安全事故。如何高效、准确地判定 SCADA 数据中的异常状态，成为了保障工业生产安全、提高生产效率、优化资源配置的重要课题。

针对数据异常状态智能判定的研究已经取得了一定的进展，国内文献［5］收集了历史数据与主站调度端时序数据，通过极限学习方法实现数据异常状态智能判定。而该方法并未对于数据进行任何处理，那么原始数据中的噪声、错误值会直接影响极限学习机模型的训练效果和判定准确性。文献［6］通过对每台设备运行特性之间的 Pearson 相关系数进行统计分析，剔除具有较强相关性的特征，将最大比例系数确定为数据运行特征，结合局部离群因子（LOF，local outlier factor）算法完成数据异常状态智能判定。但是在使用基于 Pearson-LOF 的方法之前没有充分识别和处理数据中的极端值、错误值或噪声等，这些数据会显著影响 Pearson 相关系数的计算结果，进而影响特征选择的准确性和后续 LOF 算法的效果，导致异常数据判定方法的表现不佳。文献［7］结合降噪自编码器和深度神经网络，构建深度自编码结构学习数据特征，并在误差函数中增加约束项优化网络参数。通过训练数据的重构误差设定异常阈值，对比测试数据的重构误差与阈值得出检测结果。由于该方法并未对于原始数据进行的整理、校验等操作，即使模型具有强大的特征提取和学习能力也很难从这些特征中学习到有效的信息，从而影响异常判定的效果。文献［8］提出一种带噪声基于密度的空间聚类（DBSCAN，density-based spatial clustering of applications with noise）模型的风电机组 SCADA 异常数据识别方法。通过引入预测误差与分类精度，在保证系统精度的前提下，选择最优聚类参数的邻域半径及近邻最小样本点数进行数据聚类，结合不同数据特征完成异常数据判定。然而，DBSCAN 等密度聚类算法主要关注于数据点之间的密度关系，通过识别高密度区域来形成聚类，如果原始数据中存在大量的噪声或随机波动，这些特征会被错误地解释为数据的正常分布特性，从而影响聚类结果和异常判定的准确性。国外文献［9］提出基于生成对抗网络的同步生成和异常检测框架，以解决不平衡数据集下的异常检测。通过生成器在合成真实数据的同时完成传感器信号之间的传输，并引入了一个分类鉴别器，以促进有利于异常检测的数据合成，同时该鉴别器也作为异常检测器使用。但生成对抗网络训练过程通常比较不稳定，可能会出现模式崩溃的问题。这意味着生成的样本可能会在训练过程中逐渐失去多样性，最终生成固定的模式，从而影响异常检测的准确性。文献［10］提出基于最优深度学习的流数据分类模型。该模型首先进行预处理，针对流数据不平衡的特点，采用支持向量机－合成少数过采样技术进行过采样处理。随后，利用双向长短期记忆网络进行异常检测和分类。为了优化模型性能，采用均方根传播优化器对双向长短期记忆网络模型进行超参数调整。但该方法通常需要大量的计算资源和时间来训练和推理。在处理大规模流数据时，可能影响模型的实时性和可扩展性。

人工智能技术的飞速发展，特别是机器学习、深度学习等算法的广泛应用，为 SCADA 数据异常状态的智能判定提供了新的思路和解决方案。这些算法能够自动学习数据的内在规律和特征，通过训练和优化模型，实现对异常状态的精准识别和预测。其中，时变隐马尔科夫模型作为一种强大的时间序列分析工具，因其能够捕捉数据的动态行为和状态转移特性，在 SCADA 数据异常状态智能判定中展现出巨大的潜力。针对以上方法并未关注数据中的噪声和随机波动并进行处理，从而导致 SCADA 数据异常状态智能判定精度下降的问题，提出了基于时变隐马尔科夫模型的风电机组 SCADA 数据异常状态智能判定方法。

1 风电机组 SCADA 系统

风电机组 SCADA 系统是一个集成了数据监控、采集与处理功能的关键系统，广泛应用于风电场的远程监控与管理中。该系统通过传感器网络和通信技术，实时收集风电机组的各项运行参数，如风速、风向、功率输出、温度、振动等，为风电场的运维人员提供全面的设备运行信息。风电机组 SCADA 系统通常由以下几个核心模块组成。

1）数据采集模块：负责从风电机组的各个传感器和控制器中实时采集数据，确保数据的准确性和完整性。

2）数据传输模块：将采集到的数据通过有线或无线方式传输至中央监控中心，实现数据的远程访问和分析。

3）数据存储模块：对接收到的数据进行存储和管理，提供历史数据查询和趋势分析功能。

4）数据处理与分析模块：对采集到的数据进行处理和分析，包括数据清洗、异常检测、故障诊断等，为运维人员提供决策支持。

5）人机界面模块：提供友好的用户界面，展示风电机组的实时运行状态和历史数据，支持报警和预警功能。

为了进一步提升风电机组 SCADA 系统的性能，特别是在异常状态智能判定方面，文章引入了时变隐马尔科夫模型对数据处理与分析模块进行改进。这一改进设计旨在通过动态捕捉风电机组运行状态的时变特性，提高异常检测的准确性和及时性。首先，针对 SCADA 数据中存在的噪声和异常值，采用滑动窗口技术进行平滑处理，以提高数据的可靠性和准确性。其次，对平滑处理后的数据进行标准化处理和特征提取，根据数据特征进行分类，以减少数据冗余，优化处理效率，并为后续异常判定奠定基础。最后，利用时变隐马尔科夫模型对不同子集中的数据属性进行识别，从而实现数据异常状态的智能判定。具体过程如下。

2　风电机组 SCADA 数据平滑处理

在风电机组 SCADA 系统中，由于传感器精度、环境因素或通信问题等原因，采集到的数据往往包含噪声和异常值。滑动窗口技术[11-12]有效去除噪声，平滑数据曲线，提高数据的可靠性和准确性，这有助于后续更准确地识别出真正的异常数据。

滑动窗口[13]是一种数据处理和分析技术，其核心思想在于创建一个可移动的、固定大小的窗口，在一系列数据上进行各种计算和分析操作。这个窗口在数据序列上移动，每次处理一定数量的数据点，然后移动到下一个位置，继续处理下一组数据点。该技术特别适用于处理风电机组 SCADA 系统所采集的连续时间序列数据，通过灵活调整窗口大小和移动步长，能够很好地适应不同的数据特性和应用场景。滑动窗口技术[14]通过计算窗口内数据的平统计量，有效平滑了数据中的随机波动，提高了数据的质量和可用性。此外，该技术允许在数据不断输入时实时进行计算和分析，满足了风电机组 SCADA 系统对实时性的要求。同时，滑动窗口技术具有较高的计算效率，能够迅速处理大规模数据集，确保风电机组 SCADA 系统在面对大量数据时依然保持高效性能。

基于滑动窗口法的风电机组 SCADA 数据平滑处理过程为：

假设风电机组 SCADA 获取的原始数据为 $X=\{x_i^{(j)}\}, i=1,2,\cdots,n, j=1,2,\cdots,m$；$n$ 表示传感器个数，m 表示采集到的样本数量，$X_i=[x_i^{(1)}x_i^{(2)}\cdots x_i^{(m)}]$ 表示第 i 个传感器所采集到的样本数据，也就是 $X_i=[x_i^{(1)}x_i^{(2)}\cdots x_i^{(m)}]$。假设滑动窗口的宽度是 χ，窗口每个时刻都会移动一次，针对风电机组 SCADA 数据 X 有 $m-\chi+1$ 个滑动窗口，将 $S_i^{(l)}$ 作为第 l 个滑动窗口采集的第 i 个监控参数的数据：

$$\boldsymbol{S}_i^{(l)}=[x_i^l, x_i^{l+1}, \cdots, x_i^{l+\chi-1}]^T \tag{1}$$

式中，T 表示采集周期。

使用滑动窗口对风电机组 SCADA 数据进行平滑处理，其所使用的数据增广状态矩阵如下：

$$\boldsymbol{Y}=[S^{(1)} \quad S^{(2)} \quad \cdots \quad S^{(m-\chi+1)}]^T \tag{2}$$

采用滑动窗口技术将输入数据从原来风电机组 SCADA 数据 X 的 n 维增加到 $\boldsymbol{Y}$ 的 $n\times\chi$ 维度，采样数据也从 m 到 $m-\chi+1$，从而实现数据平滑处理，具体的计算公式如下：

$$X'=\alpha Y+(1-\alpha)E_{t-1} \tag{3}$$

式中，E_{t-1} 表示是时间点 t 的指数加权移动平均值；α 表示平滑系数。

利用传感器采集风电机组 SCADA 数据，包括能耗数据、事件与报警数据以及运行状态数据等，然而其中也掺杂了少量的异常数据和噪声数据，其原因是传感器在长时间运行后因老化、损坏或精度下降等问题导致采集数据不准确或产生异常值，同时，外部环境因素如温度、湿度、电磁场等的干扰使得传感器的测量精度和稳定性下降。此外，数据在传输过程中可能受到通信干扰、信号衰减、丢包等问题的影响，导致接收数据不完整或产生误差，传输协议或设备配置错误也会导致数据被错误解析或处理。不仅如此，风电机组 SCADA 系统内部的软件或硬件故障、系统更新或升级过程中的兼容性问题也可能导致数据处理错误或异常值的产生。在数据采集和处理过程中，人为因素如操作员误操作、数据录入错误或故意篡改数据等同样会导致数据异常，这些数据严重影响到了风电机组 SCADA 数据质量，因此需要对于其进行处理，以此为后续的分析奠定重要的数据基础。

利用滑动窗口技术对于数据样本进行平滑处理，具体的结果如图 1 所示。

图 1 中，x、y 轴表示数据的空间范围。分析图 1 可知，经过基于滑动窗口法的风电机组 SCADA 数据平滑处理后，噪声数据被有效去除，说明该方法的数据平滑处理效果好，其原因在于滑动窗口法能够根据需要设置不同的采样窗口和滤波窗口大小，从而平衡数据的实时性和平滑度。在需要快速响应的场合，可以减小窗口大小以提高实时性；在需要更平滑数据的场合，可以增大窗口大小以减少噪声，因此该方法能够有效去除风电机组 SCADA 数据中的噪声。

3　基于特征提取的风电机组 SCADA 数据分类

对于平滑处理后的风电机组 SCADA 数据进行标准化处理后提取其特征，根据数据特征进行风电机组

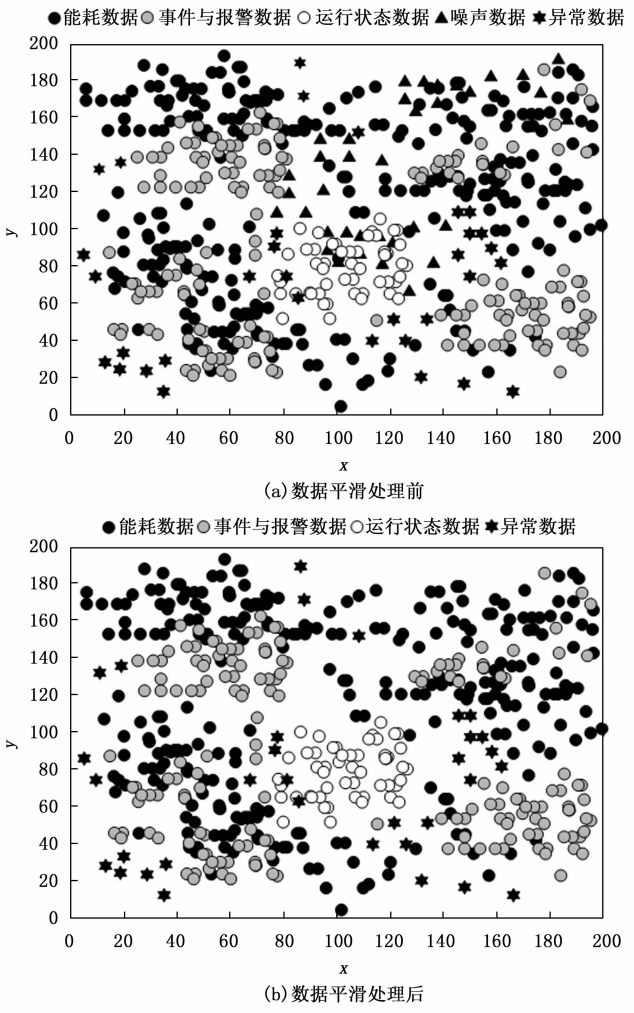

图 1 数据平滑处理结果

SCADA 数据分类，结合数据分类结果判定数据是否存在异常，从而实现数据异常判定。

风电机组 SCADA 系统采集的数据通常包含丰富的信息，如能耗数据、事件与报警数据、运行状态数据等，这些数据在时间上具有连续性，且往往呈现出一定的规律和模式[15]。通过特征提取技术，可以从这些数据中提取出关键的信息和特征，这些特征能够反映数据的本质属性和类别信息。通过提取数据特征，可以减少数据的冗余，从而降低分类算法的复杂度和计算量。同时，特征能够更好地反映数据的类别信息，使得分类方法能够更准确地识别数据的类别，显著提高分类的准确性和效率。

对于平滑处理后的风电机组 SCADA 数据进行标准化处理，具体的计算公式为：

$$G = \frac{X' - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

式中，G 表示风电机组 SCADA 数据标准化处理结果；$X_{\min}$ 表示 X' 的最小值；$X_{\max}$ 表示 X' 的最大值。

分析标准化处理后的数据变化规律，得出数据间的相似性，从而求出数据特征分布函数，具体为：

$$P(y) = \sum_{i=1}^{h} \delta_i \varphi_i(y, u, \sum G) \tag{5}$$

式中，y 表示数据标签；δ_i 表示数据特征参量；$\varphi_i\left(y, u, \sum G\right)$ 表示概率密度函数；u 表示特征累积分布参数；h 表示数据分块个数。

根据一定的次序划分风电机组 SCADA 数据，并归为一个数据集。假设风电机组 SCADA 数据集为 D，样本数量设为 M，特征数量设为 N。采用流量分类器对数据集 D 的所有样本集中式处理，将其分割成若干个子集；在分割过程中，聚合样本数量大的数据，得到新数据集，反复进行以上步骤，最后得到随机数据集。由于监控采集序列特征的波动性、随机性大，直接输入到时变隐马尔科夫模型难以对 SCADA 数据异常精确判定，也不能确保结果的可靠性。为此，根据特征将不同种类的数据划分到同一子集，从而实现数据分类，数据分类正确率的计算公式如下：

$$p = \left(1 - \frac{A}{DN}\right)^{M} \tag{6}$$

式中，A 表示分类错误次数。

通过对数据集的分类正确率和权值的分析，判定目前的分类与风电机组 SCADA 实际情况是否一致。若满足，则以该分类准确度为依据，对最后的风电机组 SCADA 数据分类结果执行操作；若不满足，则再次计算分类准确度，直至所求的分类准确度与实际状况相符为止。然后，通过 KNN 算法[16-17]分类并统计分析，得出最终的风电机组 SCADA 数据的分类结果如下：

$$w = \max\left[\sum_{t=1}^{T} T_{tx} P(y)\right], p \leqslant 0.01 \tag{7}$$

式中，T_{tx} 表示风电机组 SCADA 数据属性。

提取平滑处理后的数据特征，并利用该特征进行数据分类，将具有相似特征的数据聚集到同一子集。通过将数据划分为不同子集，有效减少了数据冗余，增强了数据集的纯净度和代表性，同时聚合大样本数据提升了数据集的稳定性和统计意义。数据分类还优化了处理效率，支持并行处理庞大数据集，显著提高了模型训练与预测的速度。此外，归类处理相似特征的数据在一定程度上平滑了数据的波动性和随机性，使模型能更好地捕捉内在规律和模式，从而进一步提高了异常判定的准确性。

4 基于时变隐马尔科夫模型的数据异常状态智能判定

对于风电机组 SCADA 数据进行分类处理后，利用时变隐马尔科夫模型[18]对于不同子集中的数据属性进行识别，从而实现数据异常状态智能判定。

时变隐马尔科夫模型是隐马尔科夫模型的一种扩

展，它允许状态转移概率随时间变化，从而更准确地描述系统的动态行为[19]。SCADA 系统广泛应用于工业领域，用于实时监测和控制设备的运行状态，其中异常状态的智能判定对于保障设备的安全运行和及时排除故障至关重要。时变隐马尔科夫模型能够捕捉 SCADA 数据的时变特性，即设备运行状态随时间的变化，通过状态转移概率矩阵的动态调整来反映这种变化。在异常状态检测方面，该模型通过比较观测序列与模型预测序列之间的差异，当差异超过设定阈值时，即可判定为异常状态。设定异常检测阈值时，需平衡灵敏度和特异性，避免误报和漏报。阈值依据数据特性和应用场景来设定，如时间序列的变化幅度、网络安全中的流量波动。同时，考虑到数据分布可能变化，阈值需适时动态调整，以适应新数据特性，确保检测有效性。文章将数据集划分为训练集、验证集和测试集。使用训练集来训练 HMM 模型，验证集则用于确定异常检测的阈值。具体做法是，将验证集中的样本输入到训练好的 HMM 模型中，计算每个样本的输出概率。在验证集样本输出概率的最大值与最小值之间均匀划分多个值作为异常检测的阈值候选值，通过评估指标（如 F_1 分数）来选择最优的阈值。

此外，模型参数（包括初始状态概率分布、状态转移概率矩阵和观测概率矩阵）可通过 SCADA 系统的历史数据进行训练，并根据新数据进行更新，以确保模型的准确性和适应性。最重要的是，时变隐马尔科夫模型能够实时处理 SCADA 数据，对设备运行状态进行判定，并在检测到异常状态时触发预警机制，及时通知相关人员进行处理。

时变隐马尔科夫模型是建立在隐马尔科夫过程的理论基础上，但状态和时间均存在离散特征，也被被称之为马尔科夫链。在概率空间 (Ω,F,Q) 中，其中，Ω 表示样本空间，F 表示事件集合，Q 表示概率测度，$Z=(z_a,a>0)$ 为随机过程，而状态空间是可数集合 S，当任意的非负整数 a、$q_0,\ldots,q_a,q_{a+1}\in S$，则：

$$Q(Z_{a+1}=o\mid Z_a=q)=Q(Z_{a+1}=q_{a+1}\mid Z_0=q_0,Z_1=q_1,\ldots,Z_a=q_a) \tag{8}$$

离散时间马尔科夫链可用随机过程 Z 描述，即马氏链。

马尔科夫链没有记忆，即过程当前的状态是已知的，则过程之前的状态不会对过程将来呈现的状态产生任何影响[20]。

将马氏链 $Z=(z_a,a>0)$ 的状态空间设为 S，对任何一个整数 $a\geqslant 0$，$k\geqslant 0$ 和任何状态 $q,o\in S$，其存在的条件概率的计算公式如下：

$$Q_{q,o}^{k}(a)=Q(Z_{a+k}=o\mid Z_a=q) \tag{9}$$

式中，$Q_{q,o}^{k}(a)$ 表示 a 时间变化序列马尔科夫链 Z 通过 k 步转移从状态 q 到状态 o 的可能性，即 a 时变马尔科夫链 Z 的 k 步转移可能性，$k=1$ 则为一步转移可能性。

状态转移矩阵 $\boldsymbol{V}$ 表达式为：

$$\boldsymbol{V}=\begin{bmatrix}v_{11}&v_{12}&v_{1a}\\v_{21}&v_{22}&v_{2a}\\v_{a1}&v_{a2}&v_{aa}\end{bmatrix},0\leqslant v_{qo}\leqslant 1,\sum_{o=1}^{a}v_{qo}=1 \tag{10}$$

式中，v_{qo} 表示状态 q 向状态 o 转移的概率值，并且是一次转移概率值。

初始状态分布 μ 表达式为：

$$\mu=(\mu_1,\ldots,\mu_a),\quad \mu_q=Q(Z_q=q_q),0\leqslant\mu_q\leqslant 1,\sum_{q=1}^{q}\mu_q=1 \tag{11}$$

通过状态转移矩阵 $\boldsymbol{A}$ 与初始状态分布 μ，能够获得完整的马尔科夫链。

隐马尔科夫模型是一类能够描述其统计性质的概率模型[21-22]，基本特征是一阶双重随机过程，其主要结构包括：状态转移概率即不同转移状态的马尔科夫链，其中的主体部分是初始状态概率 z_0、状态转移矩阵 $\boldsymbol{V}=(z_{qo})_{G*G}$，$G$ 表示状态量的数量，z_{qo} 表示隐藏状态 q 转移到当前状态 o 的概率值，$z_{qo}=Q(s_t=o\mid s_{t-1}=q),q,o\in S$，$s_t$ 表示 t 时数据属性状态。对异常数据状态进行判定时，1 表示正常，0 表示异常。$S=\{0,1\}$ 表示正常数据与异常数据集合，$\boldsymbol{V}=(z_{qo})_{2*2}$ 表示状态转移矩阵。

时变隐马尔科夫模型的另外构成要素就是刻画时间变化下的数据变化随机过程。假设数据属性变化概率矩阵为 $\boldsymbol{Q}=(\boldsymbol{Q}_{tk})_{1*2}$，其中 $k=0,1$ 表示 t 时的数据属性值为 0 或 1 的可能性。Q_{t1} 表示观测值概率：

$$Q_{t1}=P[U(t,f)s_t=1]=R[U(t,f)\mid U_z,U_\sigma]=\exp\left\{-\frac{1}{2}[U(t,f)-U_z]^T[U(t,f)-U_z]\right\} \tag{12}$$

式中，$R(\cdot\mid\cdot)$ 表示高斯分布函数，U_σ 表示数据属性方差。

则正常数据属性和子数据集属性的计算公式如下：

$$\vartheta_t(1)=\mu\left(\frac{z_{q1}*Q_{t1}}{V}\right) \tag{13}$$

$$\vartheta_t(0)=\omega z_{q0}Q_{t0}=\omega z_{q0}(1-Q_{t1}) \tag{14}$$

其中：z_{q1} 表示从前一时间点的状态 q 向 t 时状态为 1 的状态转移概率值。

当 $\vartheta_t(1)=\vartheta_t(0)$ 时，说明 t 时数据是正常的；当 $\vartheta_t(1)\neq\vartheta_t(0)$ 时，说明 t 时数据是异常的，从而实现 SCADA 数据异常状态智能判定.

综上所述，时变隐马尔科夫模型在风电机组 SCADA 数据异常检测中，相较于传统隐马尔科夫模型，展现出显著优势。它通过引入时间变化因素，使状态转移概率矩阵能够动态调整，更准确地捕捉风电机组运行状态的动态特性。这种模型不仅继承了隐马尔科夫模型处理序列数据的优势，如通过隐藏状态来刻画观测数据的

统计规律，而且智能判定机制更加先进，能自动比较观测序列与预测序列的差异，无需人工设定复杂规则即可判定数据异常。此外，时变隐马尔科夫模型具备实时处理性能，能迅速响应 SCADA 数据，及时触发预警。模型参数训练与更新机制也更为灵活，能利用历史数据训练，并根据新数据更新，确保模型准确性。同时，通过验证集动态调整异常检测阈值，增强了模型的鲁棒性。总之，时变隐马尔科夫模型在风电 SCADA 数据异常检测中，实现了更高自动化、准确性和适应性的智能判定，为风电运维提供了有力支持。

5 风电机组 SCADA 数据异常状态智能判定实验

5.1 实验设计

为了验证基于时变隐马尔科夫模型的方法对风电机组 SCADA 数据异常状态智能判定的准确性，进行了实验测试。具体的实验环境如图 2 所示。

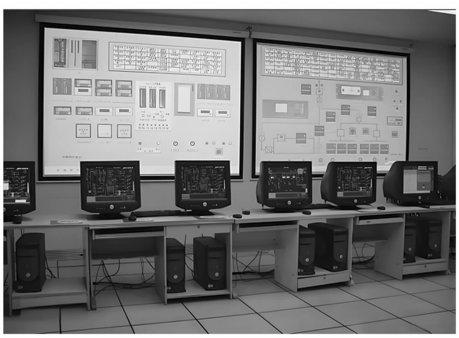

图 2 具体的实验环境

根据图 2，实验采用高性能计算机进行测试，计算机配置为：CPU Intel Core i7-93000H，GPU NVIDIA TITAN，内存 32 G，选用 2T 硬盘，利用 DS-2CD 2T25I5 摄像头采集实验数据。软件环境：Ubuntul 6.04 操作系统，Mysql5.0 数据库，SpringBoot3.5 开发框架，Flask Pytorch 服务器。实验测试的数据为风电机组 SCADA 系统采集的数据，包括能耗数据、事件与报警数据以及运行状态数据等。其中异常数据包括传感器在长时间运行后因老化、损坏或精度下降等问题导致采集数据不准确或产生异常值；数据在传输过程中可能受到通信干扰、信号衰减、丢包等问题的影响，导致接收数据不完整或产生误差；风电机组 SCADA 系统内部的软件或硬件故障、系统更新或升级过程中的兼容性问题也可能导致数据处理错误或异常值的产生。共采集数据 25 864 条，将风电机组 SCADA 数据划分为训练集、验证集和测试集，比例设置为 7：1：2。

实验过程如下：首先进行数据预处理，包括滑动窗口平滑处理去除 SCADA 数据中的噪声和异常值，以及标准化处理统一数据量纲和分布。接着，从预处理后的数据中提取关键特征，并使用特征分布函数描述数据相似性和累积分布参数。然后，将特征数据划分成数据集，使用流量分类器分割成子集，并通过 KNN 算法等分类方法将数据分类。进入模型训练阶段，使用历史数据初始化时变隐马尔科夫模型的参数，包括初始状态概率分布、状态转移概率矩阵和观测概率矩阵。随后，用训练集数据训练模型，捕捉数据时变特性。训练完成后，使用验证集数据确定异常检测阈值，通过评估指标选择最优阈值。最后，将测试集数据输入训练好的模型中，计算输出概率，并根据阈值判定数据是否异常。

实验过程参数设置如下：滑动窗口宽度为 5，平滑系数为 0.7，迭代次数为 100，学习率为 0.01，数据分块个数为 3，数据采集周期为 10 s。

对于风电机组 SCADA 数据进行分类处理后，利用时变隐马尔科夫模型对于数据异常状态进行智能判定，判定结果如图 3 所示。

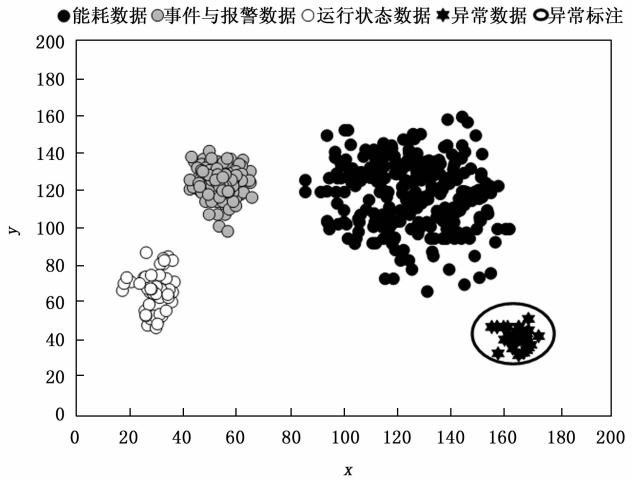

图 3 数据异常状态智能判定结果

分析图 3 中的结果可知，利用时变隐马尔科夫模型对于数据异常状态进行精准判定，实际应用效果好，其原因在于风电机组 SCADA 数据中的异常状态往往是由一些潜在的、不可直接观测的因素引起的，这些因素可以看作是隐藏状态。时变隐马尔科夫模型通过引入隐藏状态的概念，能够对这些潜在因素进行建模和推断。模型假设观测到的数据是由隐藏状态生成的，并且隐藏状态之间存在转移概率，通过学习这些转移概率和观测概率，模型可以从观测数据中推断出隐藏状态的变化，从而实现对异常状态的检测。

选择文献［6］的基于 Pearson-LOF 和文献［8］的基于空间聚类的方法的方法作为实验对比方法，选用假阳性率、召回率以及 ROC 曲线作为实验指标。

1）假阳性率：

在风电机组 SCADA 系统中，数据异常状态的判定对于确保系统的安全稳定运行至关重要。假阳性率作为评估方法性能的重要指标之一，在 SCADA 数据异常状

态判定中具有广泛应用。通过计算假阳性率，可以了解不同方法在识别正常数据方面的表现，提高其在 SCADA 系统中的实用性和可靠性，其计算公式如下：

$$\sigma_{FP} = \frac{n_{FP}}{n_{FP} + n_{TN}} \tag{15}$$

式中，n_{FP} 表示阳性样本中被错判定成阴性样本个数；n_{TN} 表示阴性样本中被正确判定成阴性样本个数。

2）召回率：

在风电机组 SCADA（监控与数据采集）数据异常状态智能判定实验过程中，召回率是一个至关重要的评估指标，其主要是用于评估模型在识别正样本（即异常状态）方面的能力，具体指方法正确识别出的异常数据占所有实际异常数据的比例。通过计算召回率，可以了解模型在识别异常数据方面的表现，进而对模型进行优化和调整，提高其在 SCADA 系统中的实用性和可靠性。其计算公式如下：

$$\sigma_{re} = \frac{n_{TP}}{n_{TP} + n_{FN}} \tag{16}$$

式中，n_{TP} 表示阳性样本中被正确判定成阳性样本个数；n_{FN} 表示阴性样本中被错判定成阳性样本个数。

3）F_1 值：

F_1 值作为综合评估算法性能的重要指标，在 SCADA 数据异常状态判定中具有广泛应用。F_1 值是精确率（Precision）和召回率（Recall）的调和平均数，用于综合评估算法的性能。在二分类问题中，精确率是指方法预测为正样本的实例中真正为正样本的比例，而召回率是指所有真正为正样本的实例中被模型正确预测出来的比例。F_1 值通过综合考虑这两个指标，为方法性能提供了一个更为全面的评估视角。通过计算 F_1 值，可以了解 不同在识别异常状态和正常状态方面的性能，进而对模型进行优化和调整，提高其在 SCADA 系统中的实用性和可靠性。这一指标的计算公式如下：

$$F_1 = \frac{2 \times (\sigma_{re} \times \zeta_{ac})}{\sigma_{re} + \zeta_{ac}} \tag{17}$$

式中，ζ_{ac} 表示精确率。

4）ROC 曲线：

ROC 曲线下面积是评价判定性能的关键指标之一。ROC 曲线，即受试者工作特征曲线（ROC，receiver operating characteristic curve），以假阳性率（1-特异度）为横坐标，真阳性率（灵敏度）为纵坐标绘制而成。ROC 曲线下面积（AUC）则是指 ROC 曲线与 x 轴、（0，0）－（1，1）围成的面积。AUC 的值域为［0，1］，AUC 越大，表示方法的性能越优越，其具体的计算公式如下：

5）漏报率：

漏报率是指不同方法未能检测出实际存在的异常状态的比率。在风电机组 SCADA 数据异常状态判定中，漏报率反映了不同方法在识别异常状态时的遗漏情况，低漏报率意味着方法能够更准确地识别出实际存在的异常状态，减少遗漏情况。其计算公式如下：

$$L_W = \left(\frac{l_i}{l_z}\right) \times 100\% \tag{18}$$

式中，l_i 表示应被识别为异常但未被识别的样本数量，l_z 表示理论上应当被识别为异常的样本总数。

5.2　实验结果

3 种方法的假阳性率曲线测试结果如图 4 所示。

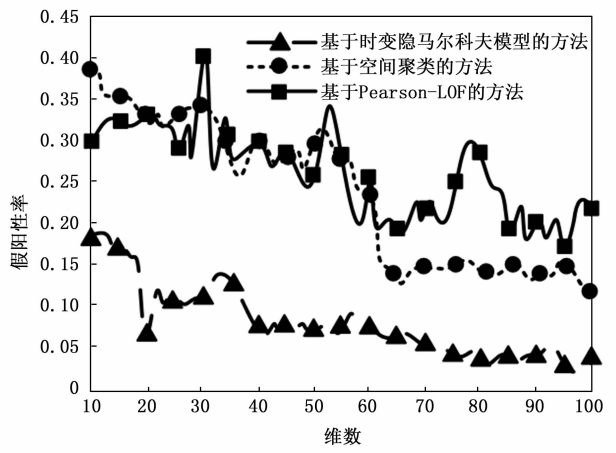

图 4　假阳性率曲线

从图 4 中可以看出，无论数据的维数如何变化，基于时变隐马尔科夫模型的方法都能较为准确地识别出真正的异常数据，同时减少将正常数据误判为异常的情况。基于时变隐马尔科夫模型的方法能够更有效地捕捉数据中的异常特征，降低误报率与漏报率，从而准确判定数据的异常状态。基于空间聚类的方法与基于 Pearson-LOF 的方法的假阳性率曲线虽然随维数增加而下降，但其整体水平相对较高，这表明这两种方法在判定数据异常状态时，容易将正常数据误判为异常，从而导致错判和漏判的情况增多。综上所述，基于时变隐马尔科夫模型的方法能够准确判定数据异常状态，降低误报率和漏报率及时发现并排除潜在的安全隐患，防止因数据异常导致的系统故障。

3 种方法的召回率曲线测试结果如图 5 所示。

通过图 5 能够得知，基于时变隐马尔科夫模型的方法具有很高的召回率，均值为 0.98，漏报率低，可找出所有存在的异常数据。基于空间聚类的方法、基于 Pearson-LOF 的方法的召回率相对较低，在判定数据异常状态时不够准确，增加了误报和漏报风险。综上所述，基于时变隐马尔科夫模型的方法能够准确判定数据的异常状态，从而降低误报率和漏报率，保证 SCADA 的稳定性。

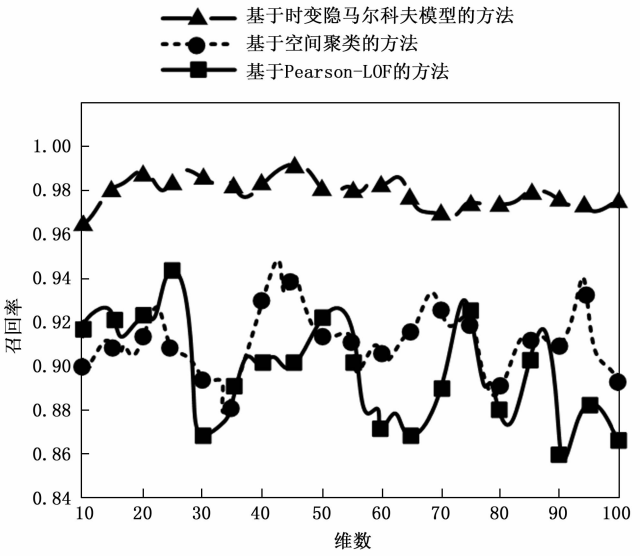

图 5 召回率曲线

3 种方法的 F_1 值测试结果如表 1 所示。

表 1 F_1 值测试结果

实验次数	基于时变隐马尔科夫模型的方法	基于空间聚类的方法	基于 Pearson-LOF 的方法
10	0.91	0.75	0.81
20	0.95	0.73	0.85
30	0.96	0.52	0.82
40	0.94	0.66	0.79
50	0.96	0.64	0.75
60	0.94	0.72	0.67
70	0.96	0.68	0.73
80	0.93	0.55	0.81
90	0.94	0.68	0.76
100	0.95	0.71	0.77
平均值	0.94	0.66	0.78

分析表 1 中的数据可知，基于时变隐马尔科夫模型的方法的 F_1 值最大达到了 0.96，基于空间聚类的方法的 F_1 值最大达到了 0.75，基于 Pearson-LOF 的方法的 F_1 值最大达到了 0.85，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法高出了 0.21、0.11。基于时变隐马尔科夫模型的方法的 F_1 值最小为 0.91，基于空间聚类的方法的 F_1 值最小为 0.52，基于 Pearson-LOF 的方法的 F_1 值最小为 0.67，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法高出了 0.39、0.31。基于时变隐马尔科夫模型的方法的平均 F_1 值为 0.94，基于空间聚类的方法的平均 F_1 值为 0.66，基于 Pearson-LOF 的方法的平均 F_1 值为 0.78，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法高出了 0.28、0.16。经过对比可知，基于时变隐马尔科夫模型的方法的 F_1 值高，意味着这一方法在处理不同类型的异常数据时表现出较强的适应性，能够应对各种复杂的异常情况。

3 种方法的 ROC 曲线测试结果如图 6 所示。

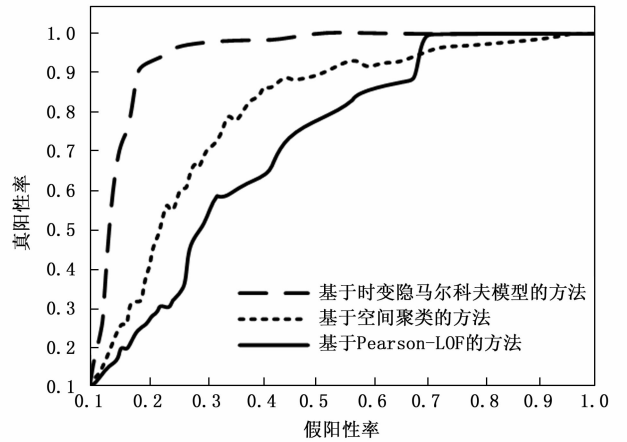

图 6 ROC 曲线

从图 6 可以看出，基于时变隐马尔科夫模型的方法的 ROC 曲线位于最上方，在判定数据异常状态时具有较高的准确性，识别出异常数据。与基于时变隐马尔科夫模型的方法相比，基于空间聚类的方法与基于 Pearson-LOF 的方法的 ROC 曲线位于下方，说明判定数据异常状态效果不佳，容易将异常数据误判为正常，或者漏检部分真正的异常数据。为此，证明基于时变隐马尔科夫模型的方法能够准确判定数据异常状态，有助于管理员快速响应并处理潜在的问题，确保 SCADA 稳定运行。

3 种方法的漏报率测试结果如表 2 所示。

表 2 漏报率测试结果

实验次数	基于时变隐马尔科夫模型的方法/%	基于空间聚类的方法/%	基于 Pearson-LOF 的方法/%
10	2.45	21.47	12.69
20	2.36	22.36	14.78
30	2.58	25.81	15.85
40	3.41	19.63	17.63
50	3.25	15.78	15.28
60	3.69	20.31	14.96
70	2.58	21.78	15.28
80	2.66	22.63	17.31
90	3.47	24.79	14.86
100	4.13	26.33	19.67
平均值	3.06	22.09	15.83

分析表 2 中的结果可知，基于时变隐马尔科夫模型的方法的漏报率最大值为 4.13%，基于空间聚类的方法的漏报率最大值为 26.33%，基于 Pearson-LOF 的方法的漏报率最大值为 19.67%，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法低 22.2%、15.54%。基于时变隐马尔科夫模型的方法的漏报率最小值为 2.45%，基于空间聚类的方法的漏报率最小值为 15.78%，基于 Pearson-LOF 的方法的漏报率最小值为 12.69%，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法低 13.33%、

10.15%。基于时变隐马尔科夫模型的方法的漏报率平均值为 3.06%，基于空间聚类的方法的漏报率平均值为 22.09%，基于 Pearson-LOF 的方法的漏报率平均值为 15.83%，分别比基于空间聚类的方法、基于 Pearson-LOF 的方法低 19.03%、12.77%。经过对比可知，基于时变隐马尔科夫模型的方法能够较为精准地识别出 SCADA 数据中的异常状态，这意味着该方法对正常数据模式和异常数据模式的区分能力较强，能够准确捕捉到数据中的微小变化和偏离正常范围的特征，实际应用效果好。

6　结束语

在工业自动化领域中，SCADA 扮演着至关重要的角色。由于 SCADA 的稳定性和可靠性直接影响到生产效率和安全性，所以准确地识别 SCADA 数据中的异常状态成为该领域的重要研究课题之一。为此，提出了基于时变隐马尔科夫模型的风电机组 SCADA 数据异常状态智能判定方法。利用滑动窗口对于风电机组 SCADA 数据进行平滑处理，结合数据分类结果与时变隐马尔科夫模型获取状态转移矩阵及初始状态分布，完成异常状态智能判定。实验结果表明，该方法的风电机组 SCADA 数据异常状态智能判定效果好，可以在实际中得到广泛应用。当前人工智能在 SCADA 数据异常状态智能判定中发挥着重要作用，其准确性和效率得到了显著提升。未来，随着人工智能技术的不断发展和完善，相信其在 SCADA 系统中的应用将会更加广泛和深入，为工业安全和生产效率的提升做出更大的贡献。

参考文献：

[1] 田银磊，刘书伦. 基于神经网络的船舶通信网络异常数据识别 [J]. 舰船科学技术，2022，44 (17)：148－151.

[2] 任其亮，徐　韬，刘　媛，等. 考虑载客状态的改进孤立森林浮动车异常数据检测算法 [J]. 交通运输系统工程与信息，2024，24 (1)：124－131.

[3] 邱　阳，李　盛，金　亮，等. 基于统计特征混合与随机森林重要性排序的桥梁异常监测数据识别方法 [J]. 传感技术学报，2022，35 (6)：756－762.

[4] 张志昂，廖光忠. 改进变分自编码器的工业时序数据异常检测 [J]. 计算机工程与设计，2024，45 (1)：17－23.

[5] 李一鹏，王亚军，栗维勋，等. 基于极限学习机的变电站监控系统负荷异常信息的辨识方法 [J]. 电子器件，2022，45 (5)：1219－1224.

[6] 石玉亮，王　呈. 基于 Pearson-LOF 算法的梯联网数据采集端异常帧检测 [J]. 控制工程，2022，29 (8)：1457－1463.

[7] 满雯妍，李红娇. 基于深度降噪自编码神经网络的 SCADA 系统异常检测 [J]. 计算机工程与设计，2023，44 (7)：1977－1984.

[8] 李　特，王荣喜，高建民. 风电机组数据采集与监控系统异常数据识别方法 [J]. 西安交通大学学报，2024，58 (3)：106－116.

[9] ZHAO P，DING Z，YANG W Y. SGAD-GAN：Simultaneous Generation and Anomaly Detection for time-series sensor data with Generative Adversarial Networks [J]. Mechanical Systems & Signal Processing，2024，210 (Mar.)：111141.1－111141.16.

[10] RAJAKUMAR R，DEVI S S. An efficient modelling of oversampling with optimal deep learning enabled anomaly detection in streaming data [J]. China Communications，2024，21 (5)：249－260.

[11] 王　翔. 基于滑动窗口的流式 RDF 数据的模式匹配方法 [J]. 计算机工程与设计，2024，45 (5)：1458－1464.

[12] 叶阿勇，孟玲玉，赵子文，等. 基于预测和滑动窗口的轨迹差分隐私保护机制 [J]. 通信学报，2020，41 (4)：123－133.

[13] QU S，ZHOU Q，WANG Q. Sliding Window-Based machine learning for environmental inspection resource allocation [J]. Environmental Science & Technology：ES&T，2023 (44)：57.

[14] 周　琴，周凡颖，丁友东. 基于多尺度滑动窗口自注意力网络的交互动作识别 [J]. 工业控制计算机，2025，38 (1)：111－112.

[15] 胡龙舟，李韬睿，吴　頔，等. 基于 SCADA 系统的风电机组 KNN 故障状态监测研究 [J]. 机械设计与制造工程，2025，54 (1)：91－94.

[16] 张书瑶，王梓齐，刘长良. 基于改进集成 KNN 回归算法的风电机组齿轮箱状态监测 [J]. 动力工程学报，2023，43 (6)：759－767.

[17] 曹　宇，鲁明旭. 基于动态调参 KNN 分类算法的股票涨跌预测模型分析 [J]. 微型电脑应用，2024，40 (4)：1－4.

[18] 宋玉琴，赵　攀，周琪玮，等. 基于时变隐马尔科夫模型的连锁故障预测 [J]. 电测与仪表，2023，60 (1)：146－153.

[19] FRANCESCO L，MARCO M. Nonhomogeneous hidden semi-Markov models for toroidal data [J]. Journal of the Royal Statistical Society Series C：Applied Statistics，2024 (1)：1.

[20] 王　芳. 基于隐马尔科夫链的企业产品营销绩效评估模型 [J]. 山西师范大学学报（自然科学版），2024，38 (4)：133－139.

[21] 黄　荷，鲍凌翔. 基于隐马尔科夫模型的电费抄核误差校正系统 [J]. 自动化技术与应用，2024，43 (9)：155－158.

[22] 王　欣，杨宏伟. 基于隐马尔科夫模型的变电站继电保护设备故障声源定位方法 [J]. 微型电脑应用，2024，40 (4)：173－177.