浙江财经大学

本 科 生 毕 业 论 文（设计）

题目：光伏序列数据异常检测的可视化分析系统

学生姓名：严振宇

学　　号：190110910227

指导教师：张翔

所在学院：信息管理与人工智能学院

专业名称：软件工程

班　　级：19 软件 2 班

2023 年 5 月

光伏序列数据异常检测的可视化分析系统

摘要：光伏 (Photovoltaic,PV) 发电作为一种安全、环保的可再生能源, 近年来受到人们的广泛关注。然而，在光伏发电的过程中会因为各种原因产生大量异常，威胁光伏电组的正常运行，如若不能及时处理可能会对电组造成不可逆的影响。因此，找到一种有效检测 PV 序列数据中异常事件的方法显的尤为重要，同时该方法需要有较好的可解释性能让工作人员更好的理解和解释异常。本文围绕此需求设计了用于检测光伏序列数据异常的可视分析系统，其中包含了一组视觉设计和一组交互设计，用以帮助工作人员完成异常原因分析。该系统先是采用基于长短期记忆和自编码器 (LSTM-AE) 相结合的深度学习算法对无异常标签的 PV 序列数据进行训练，将要进行异常检测的 PV 序列数据放入训练模型中获取对应的预测数据，然后将预测数据与原始 PV 序列数据进行比较以此来设置阈值，并用此阈值来划分异常数据。最后再对比不同 PV 序列数据的基础上分析异常出现的原因并进一步挖掘潜在的相关因素。总而言之，光伏发电异常的检测可以帮助工作人员快速确定异常发生的时间和位置并及时抢修避免发生重大的经济损失，可视化的分析系统可以帮助工作人员分析异常出现的原因并做出进一步的预防。

关键词：可视化分析; 异常检测; PV 序列数据

A Visual Detection System for Anomaly Analytics in Photovoltaic Sequence Data

Abstract: Photovoltaic (PV) power generation, as a safe and environmentally friendly renewable energy source, has received widespread attention in recent years. However, in the process of photovoltaic power generation, a large number of abnormalities may occur due to various reasons, threatening the normal operation of the photovoltaic power unit. If not handled in a timely manner, it may have irreversible effects on the power unit. Therefore, it is particularly important to find an effective method for detecting abnormal events in PV sequence data, and this method also needs to be of great help to staff in understanding and

interpreting anomalies. This paper designed a visual analysis system for detecting abnormal PV sequence data around this requirement, which includes a set of visual design and a set of Interaction design to help staff complete the analysis of abnormal causes. The system first uses a deep learning algorithm based on Long short-term memory and Autoencoder (LSTM-AE) to train the PV sequence data without abnormal labels, puts the PV sequence data to be detected into the training model to obtain the corresponding prediction data, then compares the prediction data with the original PV sequence data to set the threshold, and uses this threshold to divide the abnormal data. Finally, based on comparing different PV sequence data, analyze the reasons for the occurrence of anomalies and further explore potential related factors. In summary, the detection of photovoltaic power generation anomalies can help staff quickly determine the time and location of anomalies and promptly repair them to avoid significant economic losses. A visual analysis system can help staff analyze the causes of anomalies and make further prevention measures.

Key words: Visual Analysis; Anomaly Detection; PV Sequence Data

目 录

1 引 言

光伏 (PV) 发电作为一种新型的发电方式，因其绿色、清洁、无污染、可持续发展而被广泛研究和应用。光伏逆变器作为光伏系统的重要组成部分，具有重要的研究意义，因为其输出功率可以用来表达其基于时间的光伏转化能力[1]。然而，考虑到光照强度、环境温度等其他客观因素，输出的数据具有很大的不确定性和波动性。因此，PV 序列数据在其运行过程中会出现大量异常，严重影响发电设备的管理规划以及电网系统的稳定运行。因此，研究一种检测 PV 序列数据中异常的算法可以帮助工作人员更好的解决发电中的故障，提高光伏发电的效率和稳定性，减少对电站的经济影响[2-3].

在调研过程中发现相关领域已经提出了几种方法用来检测异常，这些方法可以分类分为三类，即基于物理原理的人工仿真模型, 传统的机器学习模型和深度学习模型。基于物理原理的人工仿真模型[4]通常由相关领域的专家创建，然而，随着光伏电站规模的增大，这类方法耗时耗力，无法及时高效地完成异常检测任务。传统机器学习模型[5]对简单的序列数据有较高的检测精度但是对于复杂的序列数据其结果并不理想。此外，此类模型对缺失数据很敏感。相比之下，深度学习模型[6]在处理复杂数据方面更加强大，因此它们在复杂序列数据中能够学习到更清晰的结构特征能够实现更准确的异常检测，且深度学习模型对于数据有更高的容错性。但是，本次毕设将要使用的 PV 序列数据其正常和异常数据的界限并不明确，需要一定的主观判断。引入可视化技术可以将专家领域的知识纳入分析过程有效的解决此问题，并协助用户进行异常识别。总的来说，深度学习模型和可视化技术的有效结合能够使用户快速准确地检测异常事件，并在 PV 序列数据中进一步探索异常发生的原因。

此次毕设的主要目标为深度学习模型和可视化技术的有效结合完成对于 PV 序列数据的异常检测，在此过程中存在着三个主要问题和挑战。**CH1.** 由于 PV 序列数据结构复杂存在噪声数据并且无异常数据标签，在其上进行模型训练是一项艰巨的任务。**CH2.** 另一个挑战是视觉设计，将领域知识纳入分析过程并与模型结果形成相互验证。**CH3.** 由于深度学习方法存在黑盒效应，PV 序列中异常的原因很难探索和解释。

为了应对上述挑战，本毕设使用了一种较为合理的算法模型即基于长短期记忆和自动编码器相结合 (LSTM-AE) 的异常检测模型，用以实现对 PV 序列数据的异常识别和检测。首先，在光照条件下能产生功率的部分序列数据作为训练集，利用 LSTM-AE

模型进行训练，并得出可信的预测数据。再将预测数据与原始数据进行比较，并以此设置一个阈值，并将超过阈值的数据设定为异常 (**CH1**)。然后，使用可视化方法对每个序列数据进行降维和投影，以获得它们的特征分布。同时，设计一个基于时间序列的折线图来详细探索 PV 序列数据的变化，并提供丰富交互来支持用户将异常数据与正常数据进行比较，从而与模型结果形成相互验证 (**CH2**)。另外，影响 PV 发电的三个关键因素：辐照度、环境温度和模块温度。将它们与发电量之间设计一个特征矩阵，以便用户可以比较和分析发电 (**CH3**) 的真正原因。最后，结合上述方法实现一个异常探索框架，将异常检测模型、原因分析以及一组可视化元素交互集在一起，使用户能够轻松直观地分析 PV 序列数据中的异常。

2 相关工作

本次毕设分为三大块，即 PV 序列数据分析，序列数据中的异常检测和异常检测的可视化。

2.1 PV 序列数据分析

PV 序列数据的分析主要集中在三个方面：影响发电的客观因素、序列数据的特征以及 PV 发电的预测。太阳能作为一种特殊的能源，受客观因素的影响很大，因此光伏发电的输出也受其影响。赫雷斯等人[7]将高分辨率气候预测的集合与 PV 发电模型(EURO-98 CORDEX) 结合起来得出结论，由于天气的影响，发电效率按晴天、阴天和雨天的顺序依次下降。陈等[8]构建了一个基于辐照度和组件温度的光伏输出预测模型。他们分析了空气中的尘埃颗粒对 PV 发电的影响并得出俩者成反比关系。Sawadogo 等人[9]基于非洲西部 PV 序列数据分析了温度对 PV 发电量的影响，得出夏季 PV 发电量最高，冬季最低的结论。受客观环境的影响，PV 发电频率不断变化，并表现出周期性、随机性、波动性、不稳定性等特征，给电网安全稳定运行带来了巨大的挑战。这些专家利用 PV 序列数据，分别基于物理建模、统计和人工智能等不同方法，实现了对光伏功率的预测。此外其他创新方法也有被种应用 Jayatissa 等人[10]提出了一种基于传统 SVM 模型的遗传算法优化支持向量机（GA-SVM）模型。马尔沃尼等人[11]采用小波分解和主成分分析结合最小平方支持向量机和时间序列预测方法，实现了光伏发电量的预测。

2.2 序列数据异常检测

序列数据异常是指数据样本的特征分布与正常明显不同。异常检测通常用于网络安全、设备维护和医疗诊断等领域。

目前，现有的异常检测方法大致可分为三类：传统统计方法、机器学习方法和深度学习方法。首先，传统的统计方法中假设正常数据对象是由统计模型生成的，其中的异常被识别为不完全符合模型的对象。Abuzaid 等人[12]提出了一种基于密度的异常检测方法，该方法将距离高密度区较远的低密度区域中的对象识别为异常。然而，传统的统计方法既费时又费力。因此，学者们改变了思路，使用机器学习方法来替代人工统计的方法。KarsligEl 等人[13]设计并实现了一个半监督异常检测系统，该系统使用 k-means 算法

来完成异常检测和识别网络攻击。Murphree 等人[14]使用无监督方法基于传感器数据训练机器模型，以监测故障异常值。与机器学习模型相比，各种神经网络架构也被很好地用于序列结构建模和异常序列检测。Du 等人[15]使用长短期记忆（LSTM）来构建递归神经网络，用于针对系统日志文件的异常序列检测。Kim 等人[16]利用 CNN 在提取空间特征方面的优势，将 CNN 与 LSTM 相结合，检测网络流量信号中的空间和时间异常。Liu 等人[17]进一步开发了一种混合可视化，以揭示每个神经元的多个方面及其之间的相互作用，从而更好地理解、诊断和完善深层细胞神经网络。然而，这些模型通常基于监督学习实现，因此难以为缺乏训练标签的无监督学习获得有效结果。相比之下，在用于异常检测的深度学习工作中，自动编码器（AE）由于能够以无监督的方式识别异常而受到了广泛关注。Zhou 等人[18]根据噪声和异常数据的易重构性，利用 AE 将其从正常数据中分离出来。Lu 等人[19]将 AE 与 RNN 相结合，根据重建误差来识别时间数据中的异常时间窗口。

在目前的调研中，本次毕设使用的 PV 序列数据没有异常识别标签。因此，需要用到无监督学习方法，基于长短期记忆和自动编码器 (LSTM-AE) 的异常检测方法非常符合本次毕设的需求。

2. 3 异常检测的可视化

可视化和分析技术旨在为各种类型的研究带来直观和方便的理解，如机器学习[20-21]、图形感知[22]和采样[23]、对比降维[24]，以及异常检测领域[25]。随着 PV 阵列时间序列监测数据量的不断增加，可视化技术将帮助从业者从庞大而复杂的数据量中提取异常信息，并支持用户识别和分析异常。这些方法对异常处理乃至光伏发电的发展都具有重要意义。

近年来，人们提出了大量的可视化方法来标记事件序列异常。根据它们的表现，这些方法可以分为基于时间线、基于层次、基于流和基于矩阵的可视化方法。基于时间线的可视化方法通常按时间顺序绘制事件序列数据，用颜色、大小或形状以图形方式表示事件，以区分不同的事件。Zhao 等人[26]设计了一个基于时间线的线性圆圈视图，以精确说明每个事件发生的时间点，其中重要的线程参与者显示为圆圈，圆圈的大小和颜色对其细节和异常标记进行编码。基于层次结构的可视化通常用于显示事件序列的层次结构。Guerra Gomez 等人[27]基于树结构描述了数据集随时间的变化，该树结构支持用户选择时间段进行详细比较，同时保留了上下文意义。Wu 等人[28]提出了一种基于轮廓的树状图设计，以说明人类运动的空间和时间特征。基于流的可视化通常强调不同事件

序列之间的转换过程。Jin 等人[29]使用基于流的可视化来描述电子病例记录中的一系列医疗事件，并对不同的事件类型进行不同的颜色编码。基于矩阵的可视化允许对异常进行更详细的表示。Fischer 等人[30]基于矩阵可视化和多个可扩展视图对路由异常的观察进行了跟踪。至于异常比较方面，Phong 等人[31]将每个事件表示为基于时间线的彩色矩形，从而支持用户验证异常日志，并通过比较多个会话之间的异同来探索导致序列异常的事件。Zhao 等人[32]设计了覆盖图，以显示序列中每个连续步骤的过渡矩阵，并明确编码序列之间在颜色和大小方面的差异。

直观的视觉元素可以有效地帮助专家进行异常识别，结合交互式界面，可以帮助专家将领域知识集成到异常分析过程中。因此，需要一个可视化分析系统来集成可视化元素和交互式界面，以帮助本次毕设实现序列数据的异常检测和原因分析。

3 需求分析和系统概述

在本节中，首先描述数据的处理。在和导师以及相关领域专家进行讨论后，总结出分析需求列表。以完成 PV 序列数据的勘探和分析任务。

3. 1 数据描述

在本次毕设中，使用的数据来自印度的两个太阳能发电厂，为期 34 天。数据分为发电数据集和天气数据集。发电数据集是在逆变器层面收集的，每个逆变器都有多条太阳能电池板线路连接。天气数据是从工厂最佳放置的单个传感器阵列读取的。为了对数据进行后续建模和可视化，进行了数据预处理。将这两个数据集按时间组合，形成多属性序列数据。例如，工厂 1 有 22 个逆变器，每个逆变器都与多个序列数据相关联，每个序列数据对应于 15 分钟的测量量，包括环境温度、模块温度、太阳辐照度和 15 分钟间隔期间的直流功率输出。最后，我将整合后的数据前 30% 的选择作为训练集，剩余的数据作为测试数据集。

3. 2 需求分析

在本次毕设定稿之前，与两位相关专家举行了几次会议，讨论了与 PV 发电相关的主要问题，例如，哪些因素影响了光伏发电。通过查阅了相关文献，以了解 PV 阵列的发电量在很大程度上受到客观环境因素的影响，如气候、天气和地理。这些因素使得光伏发电具有高度的昼夜波动性。此外，光伏发电的频率变化很快，几分钟内就会发生变化。因此，从 PV 序列数据中有效地检测异常是一项艰巨的任务。同时，我认位必须为用户或是相关人员提供直观的视觉比较，通过直观地展示异常事件和正常事件之间的内在相关性和差异，有助于解释异常。通过与专家和导师的讨论，我提炼了一系列要求：

R1.PV 序列数据中的异常识别。由于 PV 序列数据的复杂结构以及无法获得清晰的异常标签，传统方法难以捕捉异常事件。为了确保光伏发电事件异常识别的效率和准确性，提出了无监督深度学习方法来处理没有异常标签的序列数据，以提高识别的准确性。

R2. 异常检测的可视化。在分析 PV 序列数据时，需要用户的主观判断，因为正常序列和异常序列之间的边界通常没有明确定义。因此，序列数据的全面直观视图对于 PV

异常检测和误报消除至关重要。通过提供一组视觉提示，可以在未标记的条件下对异常序列数据进行视觉比较和分析。此外，视觉视图可以与模型形成相互验证。

R3. 异常解释和相关因素分析。除了光伏阵列设备的使用和故障外，影响光伏序列数据异常的因素在很大程度上也与潜在的客观因素有关，如环境、温度和日照时间等。然而，深度学习模型在解析数据过程中的黑箱效应给解释结果带来了重大困难。因此，对影响 PV 事件的潜在关键因素的表征，结合可视化设计，帮助专家实现异常序列数据的可解释原因分析。

R4. 交互式异常序列数据分析系统。经过讨论专家们希望提供一个交互式光伏异常事件探测和分析系统，帮助用户快速准确地检测和识别异常序列数据。此外，该系统可以结合领域知识验证模型的识别准确性，消除故障，帮助用户深入了解异常检测结果的相关因素，同时帮助用户比较各种序列之间的差异。因此，它可以更好地支持异常事件的筛选，以进行验证和分析优化。

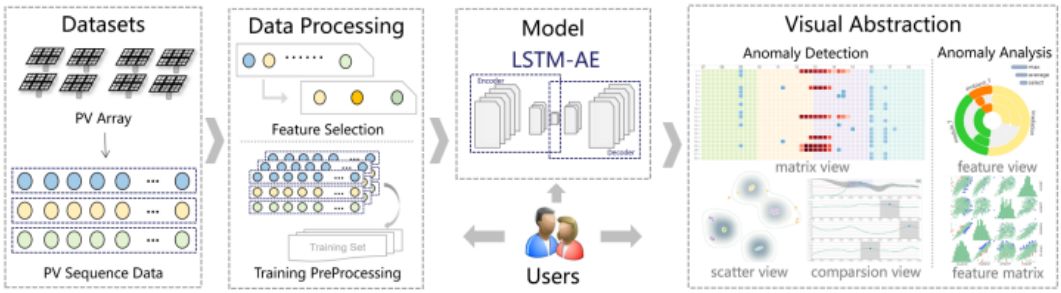

图 3.1　PV 序列数据异常检测和分析系统的流程。

3.3 系统概述

受已确定需求的启发，提出了一个交互式异常检测框架，使用户能够在真实的 PV 序列数据集上检测异常并进行原因分析。本毕设的可视化抽象系统的管道流程如 (图 3.1) 所示。

首先，将 PV 序列数据集加载到系统中进行特征提取，以减少数据集中的干扰信息。然后，基于 LSTM-AE 的检测模型，选择一部分数据作为训练集，并将其放入模型中进行训练和学习。然而，当由自动编码器重建时，数据被压缩，使得正常值往往比异常值具有更小的重建误差。因此，可以通过将预测值与实际值进行比较并将具有大重建误差的数据识别为异常来执行数据异常识别 **(R1)**。序列数据的分布特征通过多维标度（MDS）投影来表达，并使用基于时间序列的折线图来详细显示特定序列数据的变化和

每个周期的异常 **(R2)**。此外，基于特征提取的结果，引入了三个最关键的影响因素：光照强度、模块温度和环境温度，以帮助用户进行潜在的关键原因分析 **(R3)**。将可视化和交互界面集成到一个系统中，供用户检测和分析异常数据，并交互评估关键影响因素 **(R4)**。

4 异常检测模型

在本节中，会详细介绍 LSTM-AE 的工作及其在光伏阵列分析过程中的应用。

4. 1 LSTM-AE 模型

为了解决在具有复杂结构特征的 PV 序列数据 **(R1)** 中获得异常事件标签的问题，我们利用 LSTM-AE 模型来实现 PV 序列数据中的异常事件检测。该模型的第一部分是自动编码器（AE）。AE 是一个层次模型，其目标是使重建数据 X' 无限接近原始数据 X。两者之间的误差越小，AE 的训练效果就越好。AE 由编码器和解码器两部分组成。

编码器将 M 维向量 $X = \{X_1, X_2, \cdots, X_m\}$ 作为输入，并将其抽象为 N 维潜在特征向量 $H = \{H_1, H_2, \cdots, H_n\}$，$M$ 通常大于 N。然后，解码器将 N 维潜在特性向量映射到接近输入向量的预期输出向量 $X' = \{X'_1, X'_2, \cdots, X'_m\}$。

$$\begin{aligned} Encoder &: X_m \to H_n \\ Decoder &: H_n \to X'_m \end{aligned} \tag{4-1}$$

在编码部分，编码器的输出可以表示为方程 (4-2)。μ_1 是从 X 到 H 的对应连接权重矩阵，$\mu_1 \in R^{n*m}$,β_1 是偏置向量，α_1 是激活函数。

$$H = \alpha_1 \left(\mu_1 X + \beta_1\right) \tag{4-2}$$

在解码部分中，解码器的输出可以表示为方程 (4-3)。是从 H 到 X' 的对应连接权重矩阵，$\mu_2 \in R^{m*n}$,β_2 是偏置向量，α_2 是激活函数。

$$X' = \alpha_2 \left(\mu_2 H + \beta_2\right) \tag{4-3}$$

最后，我将把平均绝对误差 (MAE) 最小化，以确保输入向量 X 与输出向量尽可能相似。我们将此值作为检测的默认阈值。因此，AE 的目标函数可以描述为:

$$argminE\left(X, X'\right) = \frac{1}{m}\sum_{i=1}^{m}|X - X'| \tag{4-4}$$

该模型的第二部分是 LSTM 网络，它是 RNN 网络的衍生，旨在避免长序列训练过程中的梯度消失和梯度爆炸。

本毕设中使用的模型由编码器和解码器组成，每个编码器和解码器都是 LSTM。信息被选择性地通过门控机制以增强其时间依赖性。(图 4.1) 显示了 LSTM-AE 模型的体系结构。在编码器阶段，时间 t 的隐藏状态 H_t，其中 $i{-}4 < t \leq i$，受到当前时间的输入向量和前一时间的隐藏状态的影响。然后，时间 i 处的隐式状态被传递到解码器，其中的过程与编码器中的过程正好相反。解码器从时间 i 逐渐遍历到时间 $i{-}4$。因此，时间 $t\,(i{-}4 \leq t < i)$ 的隐藏状态 H_t' 将受到隐含状态和随后时间的重构向量的影响。

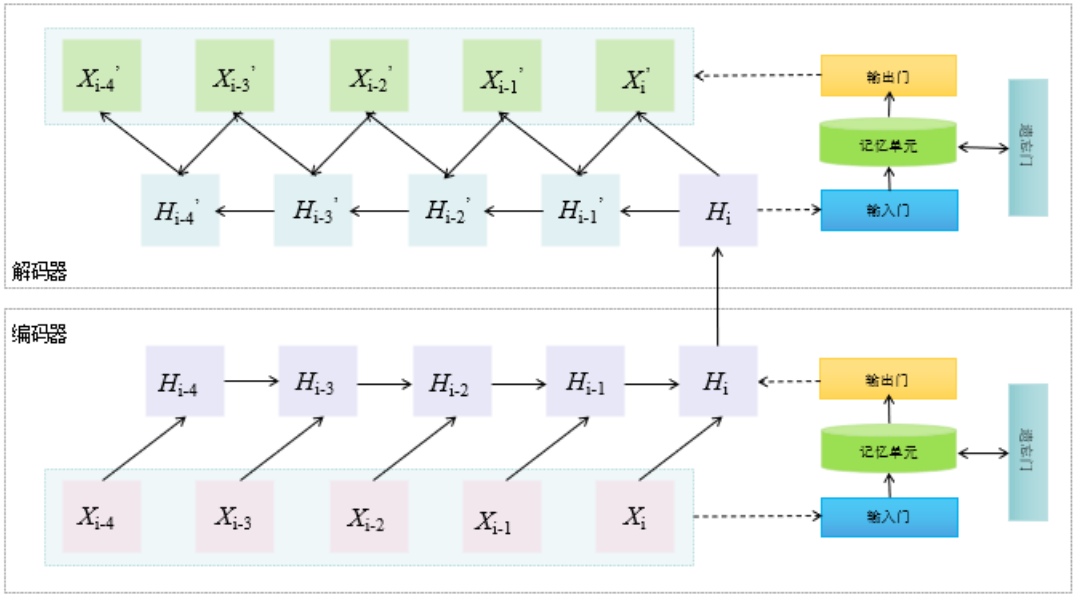

图 4.1　输入序列数据长度为 5 的 LSTM-AE。

4. 2 基于 LSTM-AE 的异常检测

总的来说，如 (图 4.2) 所示，基于 LSTM-AE 模型的 PV 序列数据异常检测主要分为三个步骤，如下所示：

step1. 变量特征重要性计算并排列, 将排名前三的变量（共五个变量）指定用于建模，即环境温度、模块温度和辐照度。

step2. 所选变量与发电量相结合，作为我们模型输入的数据。然后，训练基于 LSTM 的编码器来提取低维特征表示，即潜在向量 H，以表征每个输入序列的进展。潜在向量 H 被馈送到 AE 模型的解码器中进行序列重建，从而恢复输入序列中每个因子的期望值。

step3. 序列重建得到的发电量与真实发电量进行比较来计算偏差。如果偏差超出固定阈值，则定义为会发出异常事件。

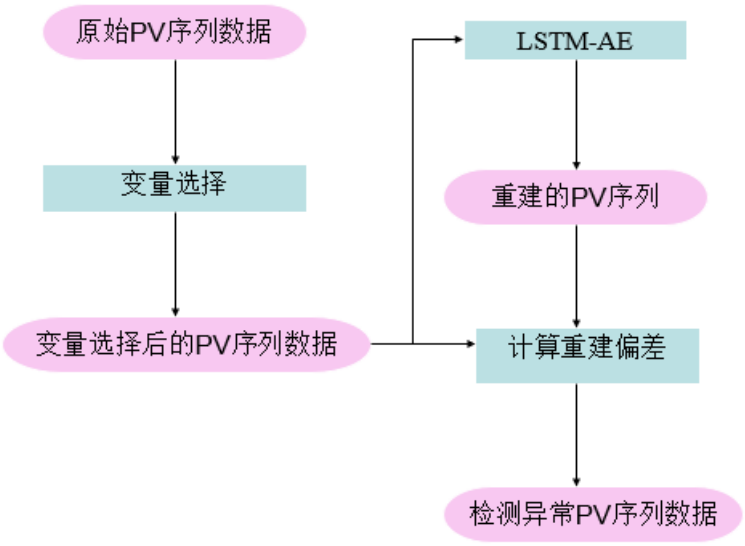

图 4.2　光伏序列数据异常检测的处理。

5 可视化设计

为了满足第 3 节中讨论的设计要求，本次毕设进一步提供了一组可视化设计和交互，以可视化 **(R2)** 和分析 **(R3)** 异常检测和相关因素。

日历视图: 如 (图 5.1(b)) 所示，日历热量视图清楚地反映了一定时期内异常值的分布，从而便于用户进行比较分析。该视图由时间变量和异常级别组成，其中每个色块代表一天。此外，该颜色对 PV 序列数据的异常程度进行编码，较暗的颜色表示当天的序列数据中存在更严重的异常。

矩阵视图: 如 (图 5.1(c)) 所示，在矩阵视图中，纵轴为 22 个逆变器，横轴为一天中发电的时间段，该时间段被划分为每 15 分钟一个单位。每个矩阵框指示该逆变器在相应时间产生的功率是否正常。当发电正常时，显示原始时间类别的颜色。红色表示序列数据的预测值高于实际值，蓝色表示序列数据预测值低于实际值，并且异常程度由颜色的阴影表示。

散点视图: 散点图通过 (MDS) 特征降维呈现了每个逆变器在当前设置的时间步长内的平均特征，其中每个点表示固定时间步长的序列数据，从中展示了不同序列数据在空间中的分布 (图 5.1(d))。其空间分布是通过计算其平均特征获得的，因此，两点之间的距离可以反映它们的相似性。直观地说，多个高密度区域可能意味着多个不同模式的聚集，而位于低密度区域 (异常值) 的序列数据更容易出现异常。此外，我基于内核密度估计（KDE）绘制了等高线，以显示不同的密度区域，高密度区域和低密度区域分别用白色到蓝色渐变绘制。序列 s 位置处的密度定义如下:

$$D_h\left(x_s, x_i\right)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x_s-x_i}{h}\right) \tag{5-1}$$

$D_h\left(x_s, x_i\right)$ 表示核密度函数，K 表示权重函数，x_i 表示除 s 之外的其余序列的位置，h 表示带宽。直观地说，多个高密度区域可能意味着不同模式的聚集, 反之位于具有大半径的低密度区域 (高异常值) 中的序列更有可能具有复杂的异常。

序列比较视图: 选择特定序列数据后，所选数据和同一时间步长的其他序列数据将显示在比较视图中 (图 5.1(e))。此视图是基于一种流的格式来呈现序列数据的细节，横轴为时间刻度，纵轴为发电量。此外，本视图中使用背景阴影来指示相似序列中正常生

成的间隔范围。为了引导用户注意异常，本视图设计了一个直观的曲线来标识异常 (R1，R2) 红色虚线表示序列数据的预测值高于实际值，蓝色虚线表示序列数据预测值低于实际值。此外，为了更好地表达异常和正常序列数据之间的比较，序列显示窗口在选择要分析的异常序列数据段时突出显示与所选段的变化特征具有更高相似度的序列数据。序列数据显示窗口右上角的方框按钮用于选择。

特征视图: 如 (图 5.1(f)) 所示，本次毕设提出了影响输出的三个关键因素 (辐照度、环境温度和模块温度)，以指导用户关联异常原因。将饼图与圆形条形图相结合，以关联所选数据的因素。行业规模旨在表明这三个因素对光伏发电的影响程度。在训练模型的过程中，对每个特征进行预测，并获得损失；每一个损失都是特征的重要性。如果损失更大，则特征更重要。圆形条形图中最内侧、中间和最外侧的条形以两种方式映射，一种表示所选数据点的相关值、不同日期内数据点在同一时间的相关系数的平均值和总数据的相关系数最大值，另一种表示相关系数值。

特征矩阵视图: 如 (图 5.1(g)) 所示，绘制特征矩阵以比较特征与发电之间的相关性。在非对角线位置是不同特征之间的相关性展示，其水平轴和垂直轴分别是相应特征的特定值。对角线位置是每个特征因子与光伏逆变器数量之间关系的直方图。横轴是每个特性本身的特定值的大小，纵轴是相应特性值内的光伏逆变器数量。

信息视图： 如 (图 5.1(h)) 所示，本次毕设根据描绘的时间步长，统计了 22 个逆变器在不同时间类别中的平均发电量和异常事件数量。

序列比较框选图： 如 (图 5.2) 所示，用户可以点加在序列比较图上的框选按钮，在自己感兴趣的序列区域框选一部分，此后系统会根据 DTW 算法在 22 个逆变器的所有时间段中匹配出与框选序列最为相近的序列发电量曲线图，用以比较在相同发电量不同时间不同外界环境因素影响下出现发电异常的原因。同时在框选序列中会出现小型饼状图 (三个环境因素)，可以让用户更加直观的看到环境因素对光伏发电的影响。

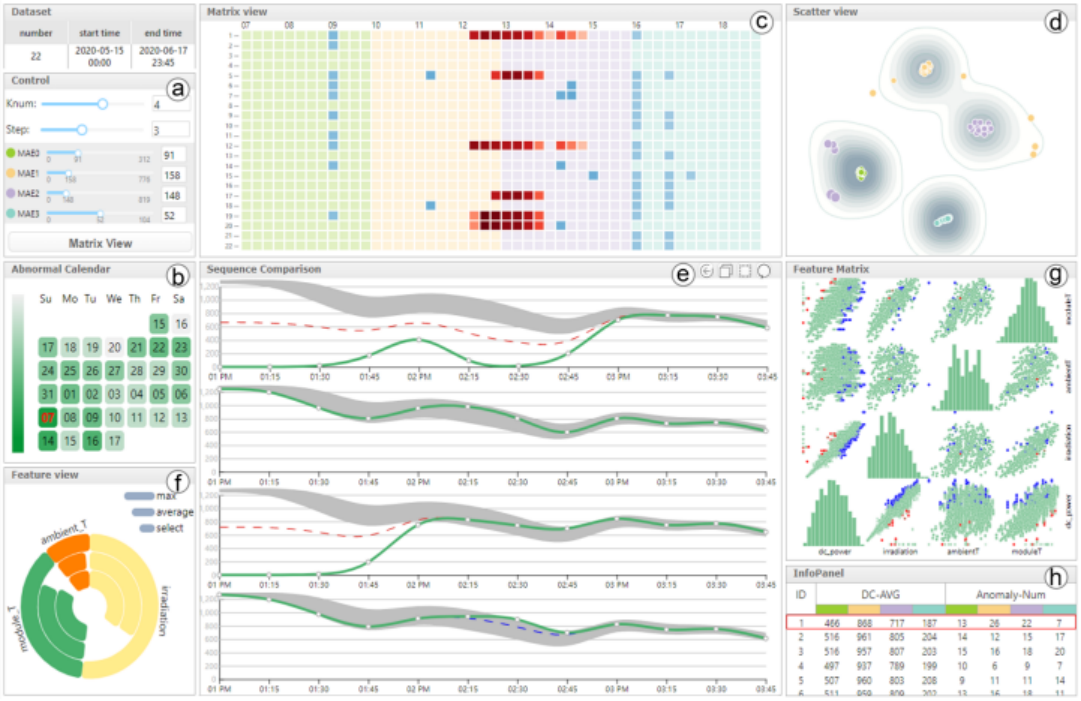

图 5.1　交互式视觉分析系统，用于检测和探索 PV 序列数据的异常。

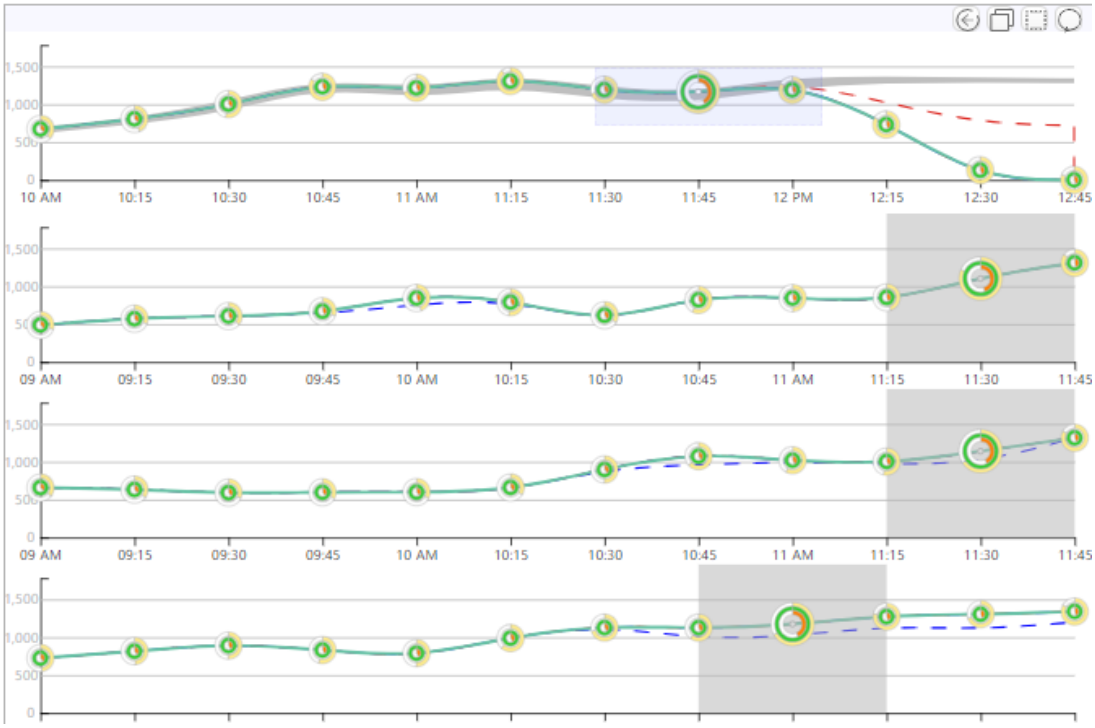

图 5.2　序列比较框选图，比较相同发电量下不同时刻不同环境对光伏发电的影响。

6 评估及案例分析

本节中，将会进一步讨论了本次毕设系统的有效性。

6.1 案例分析

序列数据的特征聚类的数量被设置为四个。除去没有发电的时间段，我们将一天中的主要时间段从早上 7:00 到晚上 19:00 每三个小时划分一次。然后，我们根据它们的发电特征将它们分为上午 7:00 至上午 10:00（绿色标记）、上午 10:00 至下午 13:00（黄色标记）、下午 13:00 至下午 16:00（紫色标记）和下午 16:00 至 19:00（蓝色标记）。根据 LSTM-AE 模型的结果，四种类型的序列数据的异常检测阈值分别为 91、158、148 和 52。

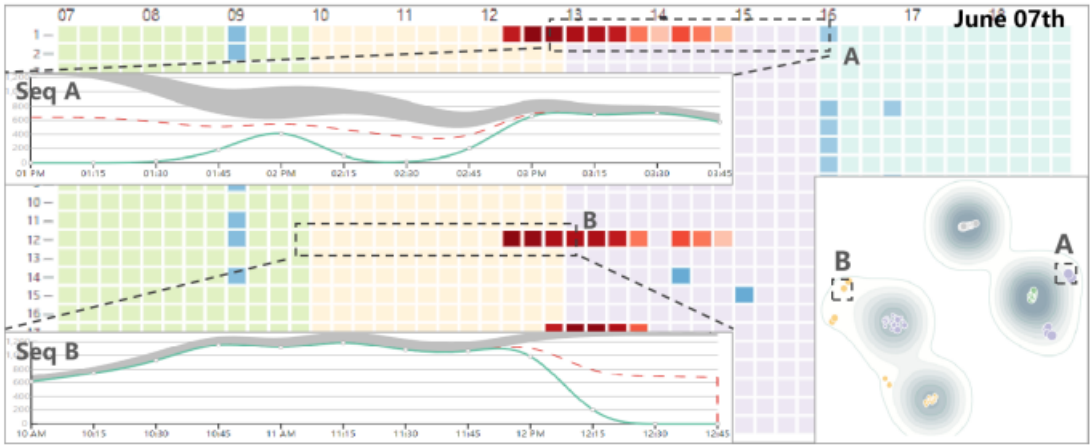

图 6.1　使用两种不同类型的序列数据（序列 A 和序列 B）来验证异常识别的准确性。

案例 1. 异常识别. 对于未标记序列数据的异常检测，结果的准确性将直接影响后续工作中对异常事件的分析和探索。因此，我观察并比较了异常识别结果，用以证明这一结论。本文我选择了一个具有代表性的日子 (6 月 7 日) 来说明这一情况。从图中可以发现一定规律，即白天大量严重异常主要集中在中午 12:00 至 15:00 之间，其中异常可分为俩种序列类别如 (图 6.1)。在 (图 6.1) 中，我选择了包含下午 13:00 至 16:00 期间异常的序列 A。实际光伏发电量（绿色实线）在该期间处于较低水平，重建的序列数据值（红色虚线）显著高于实际值。在散点视图中，所选序列对应于散点 A，该散点 A 位于与类别中的正常序列显著不同的位置，从而表明所选序列的特征结构与正常序列相比显

著不同。同时，在上午 10:00 至下午 13:00 的时间间隔内，另一个选定的异常序列 B 的真实生成值也显著低于其在下午 12:00 至 13:00 之间重建的预测值。在散点图中也与类别中的正常序列显著不同。以此实验结果证明了我们的方法在识别 PV 序列数据异常方面的有效性。

案例 2. 异常情况比较. 如 (图 6.2) 所示，在比较视图中选择了来自同一时期的 1 号逆变器 (序列 A 包含大量严重异常)、4 号逆变器 (序列 B 正常) 和 7 号逆变器 (序列 c 包含少量轻微异常) 的输出序列数据进行分析。在充足的光照条件下，下午 13:00 左右，光伏输出应保持在高范围（灰色背景范围）内。然而，1 号逆变器的实际输出几乎不存在，直到下午 3:00 后才恢复到正常范围。这种异常水平通常不是由外部因素引起的，很可能是由于在此期间其自身设备的故障。而序列 C 在下午 14:30 左右显示出与预测值的偏差略高于真实值。此类异常事件的发生通常是外部因素影响后的结果。具体分析将在案例 3 中进行解释。

此外，本次毕设通过对异常数据进行成帧来匹配和比较其特征。如 (图 6.3) 所示，序列数据与蓝色异常序列具有相似的特征，集中在上午 11:30、下午 1:00 和下午 2:30 左右。请注意，在下午 2:30 前后，匹配的序列段中也存在相应的蓝色异常。这表明当前选择的序列段的特征无法同时与正常序列数据相匹配，进一步验证了系统检测异常的准确性。

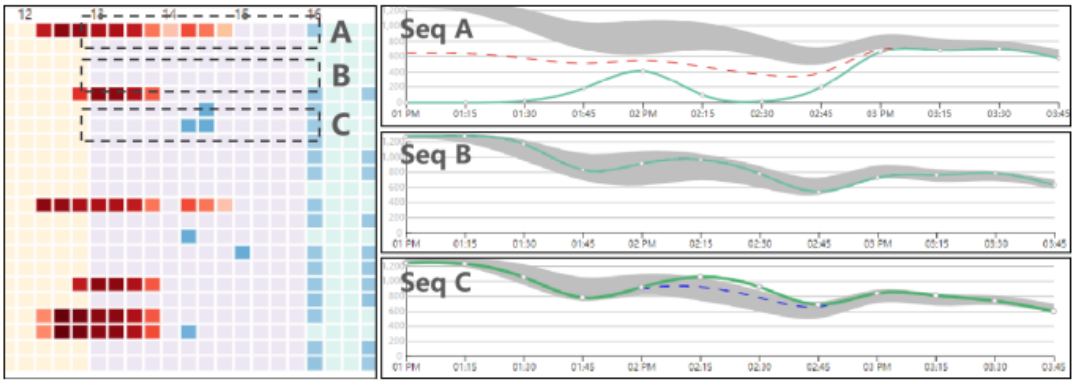

图 6.2　三种不同类型的序列数据（序列 A、序列 B 和序列 C）用于异常检测的比较分析。

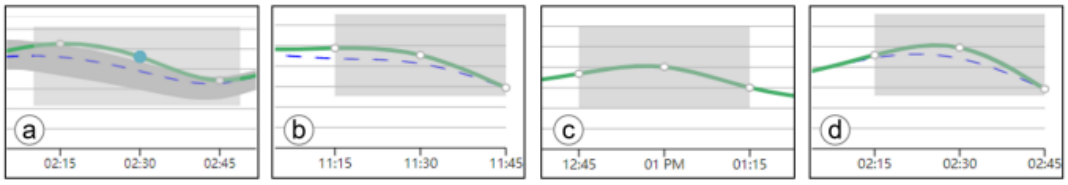

图 6.3　序列段比较示意图。用于分析的序列段（a）和用于比较的匹配序列段（b、c 和 d）。

案例 3. 异常原因分析. 异常事件的产生在很大程度上取决于光伏阵列设备的故障程度。但是，它也很可能受到环境因素的影响。案例 2 表明，1 号逆变器极有可能在下午 13:00（15:00）发生设备故障。同时，7 号逆变器在很大程度上是由于外部因素造成的。因此，基于模型的训练结果，我选择了三个最具影响力的环境特征变量 (辐照度、模块温度和环境温度) 来进一步解释异常的原因。如 (图 6.4) 所示，在对蓝色异常进行比较分析后, 点击异常发生的刻度即点会呈现出三个环境因素的饼状图，最外圈为三个因素的最大值，中间层为三个因素的平均值，最里层则为所选点当前三个因素的数值。在此之后我对该段异常序列进行框选操作，改操作完成后会出现与框选曲线最为类似的正常发电曲线，该正常曲线与框选的异常序列处于不同日期或是时间序列，点击该正常序列曲线对应的点，也会呈现对应的饼状图，通过对比可以得知框选前如 (图 6.4(b)) 所示所选数据的模块温度和环境温度高于平均值，而辐照度低于平均值。框选后的正常序列如 (图 6.4(c)) 所示所选数据的模块温度和环境温度与平均值相同，而辐照度低于平均值。由此，将两者比较分析相结合表明，由于逆变器本身的当前较高模块温度及其所在的环境温度，此时的实际发电量可能高于预测。

同时，我也对所选序列类别进行表征，其结果如 (图 6.4(e)) 所示。发电量与光照量和模块温度呈正相关，即当光照量高和模块温度高时，发电量也高。光发电量与模块温度本身之间也存在正相关关系。值得注意的是，在特征散点图中，存在明显的异常值即在光照量较高或模块温度较高的时候，发电量依然接近于零。这种趋势意味着本次发电量与光照等环境因素无太大关联。因此，极有可能是设备存在问题如短路等。该场景也进一步证实了 (图 6.1) 中的异常值 A 和 B 是由光伏设备故障所引起的。

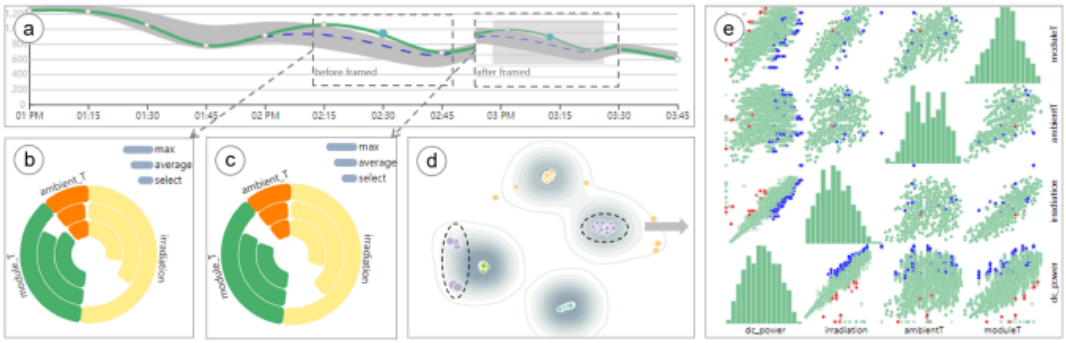

图 6.4　原因分析图解。(a) 呈现要分析的异常序列 (案例 2 中的序列 C)。(b) 以及 (c) 分别呈现在异常序列被框化之前和之后的环境因素的比较。(d) 表示所分析序列的类别，并且 (e) 是相应的特征矩阵。

6. 2　讨论

与传统的异常检测策略相比，本次毕设提出的深度学习模型在识别没有异常标签的 PV 序列数据方面表现更好。它的主要优点包括以下几点：1. 我们使用深度学习 LSTM-AE 模型在没有异常标签的情况下对 PV 序列数据处理更加有效。2. 基于全局序列数据与可视化技术相结合，使用户能够比较和分析异常序列数据，从而帮助他们更好地了解异常特征和可解释的原因。总而言之，本次毕设识别光伏发电序列数据中异常的方法不仅可以在没有异常识别标签的情况下有效地处理序列数据，而且还可以对识别的异常的特征进行比较分析和解释。

然而，本次毕设仍有存在一些不足之处，这些不足将在未来的工作中不断深入解决。(1) 深度学习模型获得的训练结果在某种程度上无法被人类控制，异常序列数据也难以被完全准确地识别。同时，正常序列数据的预测结果也可能受到隐藏在其中的异常序列干扰，从而导致其偏离正常范围。在未来的工作中，我们将进一步优化深度学习模型，以降低异常识别模型的误差，提高预测值的准确性，以满足异常分析的要求。(2) 异常事件的原因受到多种因素的直接或间接影响。本次毕设仅仅有效地引入了三个环境特征变量作为可解释原因分析的影响因素。在未来的工作中，我们将进一步引入更多的影响因素，探索每个特征因素之间的相关性，以掌握它们对异常事件的影响因素从而更有效地探索和分析多方面异常事件的原因。

结　论

针对复杂的光伏发电数据结构，本次毕设使用了 LSTM-AE 模型来检测光伏序列数据中的异常。此外，还开发了一个用于异常检测的可视化分析系统，为 PV 序列监测提供技术支持。系统中提供了一组视觉比较，以在多个视图中描述不同粒度的 PV 序列数据，从而有助于理解异常事件。同时，可以将特征视图组合起来，对光伏发电异常进行相关性分析，支持用户通过属性特征对异常进行进一步解释和探索。基于真实世界数据集的案例研究和定量比较从不同角度对 PV 序列数据的探索证明了我们系统在异常检测性能方面的有效性。

参考文献

[1] Rosato A, Araneo R, Panella M. Multivariate Prediction in Photovoltaic Power Plants by a Stacked Deep Neural Network[C]//2019 Photonics & Electromagnetics Research Symposium - Fall (PIERS - Fall). 2019: 451-457.

[2] Wang K, Qi X, Liu H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network[J]. Applied Energy, 2019, 251: 113315.

[3] Guo S, Jin Z, Chen Q, et al. Interpretable anomaly detection in event sequences via sequence matching and visual comparison[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 28(12): 4531-4545.

[4] Chao K H, Li C J, Ho S H. Modeling and fault simulation of photovoltaic generation systems using circuit-based model[C]//2008 IEEE International Conference on Sustainable Energy Technologies. 2008: 290-294.

[5] Zhao Y, Liu Q, Li D, et al. Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems[J]. Sustainable Energy, IEEE Transactions on, 2019, 10(3): 1351-1361.

[6] Janssens O, Slavkovikj V, Vervisch B, et al. Convolutional neural network based fault detection for rotating machinery[J]. Journal of Sound and Vibration, 2016, 377: 331-345.

[7] Few S. Show Me the Numbers: Designing Tables and Graphs to Enlighten[M]. Show Me the Numbers: Designing Tables, 2012: 371-385.

[8] Chen J, Pan G, Ouyang J, et al. Study on impacts of dust accumulation and rainfall on PV power reduction in East China[J]. Energy, 2020, 194(Mar.1): 116915.1-116915.10.

[9] Sawadogo W, Abiodun B J, Okogbue E C. Impacts of global warming on photovoltaic power generation over West Africa[J]. Renewable Energy, 2020, 151: 263-277.

[10] Jayatissa R, Mahamithawa S, Ranbanda J M. Impact of probabilistic small-scale photovoltaic generation forecast on energy management systems[J]. Solar Energy, 2018, 165: 136-146.

[11] Malvoni M, Giorgi M, Congedo P M. Data on Support Vector Machines (SVM) model to forecast photovoltaic power[J]. Data in Brief, 2016, 9(C): 13-16.

[12] AH A. Identifying density-based local outliers in Medical multivariate circular data[J]. Statistics in medicine, 2020, 39(21): 2793-2798.

[13] KarsligEl M E, Yavuz A G, Güvensan M A, et al. Network intrusion detection using machine learning anomaly detection algorithms[C]//2017 25th Signal Processing and Communications Applications Conference (SIU). 2017: 1-4.

[14] Murphree J. Machine learning anomaly detection in large systems[C]//IEEE Autotestcon. 2016: 1-9.

[15] Murphree J. Machine learning anomaly detection in large systems[C]//2016 IEEE AUTOTESTCON. 2016: 1-9.

[16] Kim T Y, Cho S B. Web Traffic Anomaly Detection using C-LSTM Neural Networks[J]. Expert Systems with Applications, 2018, 106(sep.): 66-76.

[17] Liu M, Shi J, Zhen L, et al. Towards Better Analysis of Deep Convolutional Neural Networks[J]. IEEE Transactions on Visualization & Computer Graphics, 2017, 23(1): 91-100.

[18] Chong Z, Paffenroth R C. Anomaly Detection with Robust Deep Autoencoders[C]//the 23rd ACM SIGKDD International Conference. 2017: 665-674.

[19] Of Agriculture U D. the USDA National Nutrient Database for Standard Reference, Release 19[J]. http://www.ars.usda.gov/main/site$_main.htm?modecode = 12-35-45-00$, 2007: 35-45.

[20] 张润, 王永滨. 机器学习及其算法和发展研究[J]. 中国传媒大学学报: 自然科学版, 2016, 23(2): 10-18.

[21] 刘鲁刘志明. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 048(001): 1-4.

[22] 叶帅男, 储向童, 巫英才. 沉浸式可视化综述[J]. 计算机辅助设计与图形学学报, 2021, 33(4): 11-33.

[23] 夏仁波, 刘伟军, 王越超. 一种改进的基于 DP 原理的分段轮廓采样算法[J]. 计算机工程与应用, 2004, 40(21): 4-17.

[24] 陈军林, 闫岩, 彭润民. 基于 t-SNE 降维算法的区域化探数据中地质体空间分布信息可视化: 以英格兰西南部为例[J]. 地质科技通报, 2021, 40(2): 186-196.

[25] Wang X, Chen W, Xia J, et al. HetVis: A Visual Analysis Approach for Identifying Data Heterogeneity in Horizontal Federated Learning[J]. IEEE Trans. Vis. Comput. Graph., 2023,

29(1): 310-319.

[26] Zhao J, Cao N, Wen Z, et al. FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media[J]. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12): 1773-1782.

[27] Guerra-Gomez J A, Pack M L, Plaisant C, et al. Visualizing Change over Time Using Dynamic Hierarchies: TreeVersity2 and the StemView[J]. IEEE Transactions on Visualization & Computer Graphics, 2013, 19(12): 2566-75.

[28] Wu W, Xu J, Zeng H, et al. TelCoVis: Visual Exploration of Co-occurrence in Urban Human Mobility Based on Telco Data[J]. IEEE Transactions on Visualization & Computer Graphics, 2015, 22(1): 935-944.

[29] Jin Z, Cui S, Guo S, et al. Carepre: An intelligent clinical decision assistance system[J]. ACM Transactions on Computing for Healthcare, 2020, 1(1): 1-20.

[30] Fischer F, Fuchs J, Vervier P A, et al. VisTracer: a visual analytics tool to investigate routing anomalies in traceroutes[C]//Vizsec Symposium on Visualization for Cyber Security. 2012: 80-87.

[31] Nguyen P H, Turkay C, Andrienko G, et al. Understanding user behaviour through action sequences: from the usual to the unusual[J]. IEEE transactions on visualization and computer graphics, 2018, 25(9): 2838-2852.

[32] Jian Z, Liu Z, Dontcheva M, et al. MatrixWave: Visual Comparison of Event Sequence Data[C]//Annual CHI Conferenceon Human Factors in Computing Systems. 2015: 259-268.

在读期间取得的科研成果

1. 本人第二作者.VDAP: A Visual Detection System for Anomaly Analytics in Photovoltaic Sequence Data. 已投稿期刊 Energies.

2. 本人第三作者.Visual Analytics of Multiple Network Ranking Based on Structural Similarity.Pacificvis2022(ccf-c 类会议).(已发表)

3. 本人第三作者.Visual Analytics of Spatio-temporal Urban Mobility Patterns via Network Representation Learning. 期刊 IEEE ACCESS.(已发表)

致　谢

本文的顺利完成，离不开各位老师、同学的支持和帮助。我首先想感谢所有向我提供帮助的同学，为我提供了数据和论文写作的指导。感谢张翔老师不遗余力的帮助和支持，感谢光电的俩位专家给我提供了大量的灵感和知识，在他们的帮助下我的毕业设计才能顺利的完成。

大学四年里，浙江财经大学和信息管理与人工智能学院为我提供了良好的学习环境，在一个良性竞争的环境中不断努力进步，十分感激所有向我提供指导和帮助的老师与同学。