南开大学

本科生毕业论文（设计）

中文题目：基于可预测气象因素的超短期太阳辐照度研究

外文题目：Ultra-short-term Prediction of Solar Irradiance Based on Predictable Meteorological Factors

学　号：1810055
姓　名：刘汉青
年　级：2018 级
专　业：统计学
系　别：概率统计系
学　院：数学科学学院
指导教师：　阮吉寿
完成日期：　2022.05

关于南开大学本科生毕业论文（设计）的声明

本人郑重声明：所呈交的学位论文，是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或没有公开发表的作品内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

本人声明：该学位论文是本人指导学生完成的研究成果，已经审阅过论文的全部内容，并能够保证题目、关键词、摘要部分中英文内容的一致性和准确性。

学位论文指导教师签名：

年　　月　　日

摘　要

光伏发电是当前利用太阳能的重要途径。通过对辐照度的精确预测，可以间接预测出光伏发电原件在未来一段时间内的出力数据。本文通过风速、风向、温度、湿度、压强等历史数据，预测未来较短时间内的辐照度。本文使用模式识别方法将相似的样本进行重组，并利用每一维气象因素的功率谱分析结果，分析其在信号学角度的可预测性。为满足对于采样频率的要求，使用线性插值方法对含有高频分量的气象因素进行加细。经测试，使用重组样本后的优化 MMMLP 模型相比原模型取得了显著更好的预测结果。

关键词：辐照度预测；模式识别；功率谱分析；多层感知机

Abstract

Photovoltaic power generation is an important way to utilize solar energy and one of the important clean energy sources in China. Using the accurate prediction of irradiance, the output data of photovoltaic power generation elements in the future can be indirectly predicted. Based on the historical data of wind speed, wind direction, temperature, humidity and pressure, the solar irradiance in a short time in the future is predicted. In this paper, the pattern recognition method is used to reunion the similar samples, and the power spectrum analysis results of each meteorological factor are used to analyze their predictability in terms of signal science. In order to meet the requirement of sampling frequency, the meteorological factors with high frequency components are fertilized by linear interpolation method. After being tested, the optimized MMMLP model using the recombined samples achieved significantly better prediction results than the original model.

Key words: Irradiance prediction; Pattern recognition; Power spectrum analysis; Multilayer perceptron

目 录

摘 要 4
Abstract 5
一、绪论 7
（一）课题研究的背景和意义 7
1．光伏发电系统的现状与特性 7
2．光伏发电预测的要求与困难 7
3．辐照度预测和光伏发电功率预测的关联 9
（二）现有的辐照度预测方法综述 9
1．基于历史数据的数理预测方法 10
2．基于图像处理的预测方法 11
3．基于 WRF 模式气象预报的预测方法 12
二、基于模式识别的天气类型识别和昼夜识别 14
（一）利用聚类方法的天气类型探索和相似日识别 14
1．天气类型设计和预处理 14
2．聚类分析 16
（二）基于支持向量机的昼夜识别 17
（三）样本数据重组 20
三、基于功率谱计算的可预测性分析 21
（一）基于 SFGPS 和 AFAGPS 功率谱分析与可预测性分析 21
（二）采样频率确定与样本插值 23
四、利用 MMMLP 的辐照度预测 26
（一）算法简述 26
（二）预测性能分析 26
五、结语 29
参考文献 30

一、绪论

（一）课题研究的背景和意义

1． 光伏发电系统的现状与特性

我国作为世界上太阳能资源最为丰富的几个国家之一，资源开发的前景相当广阔。中国每年的太阳能理论储量相当于 149.5 亿吨标准煤燃烧所产生的能量[1]，而光伏发电作为利用太阳能的重要方式，其应用形式灵活、维护简单等特性为其提供了广阔的应用前景[2]。根据我国在可再生能源领域的发展规划，到 2050 年，太阳能发电装机容量将达到 600GW，届时光伏发电装机容量将占全国电力装机容量的 5%。同时，中国的太阳能装机容量的复合增长率将在未来十几年达到 25%以上。如何相对准确地对光伏发电的效率进行预测，从而对能源调度进行合理的指导，以提升电网系统在消纳利用电能方面的能力，就变成了一个关键的问题。

光伏发电系统根据其是否接入电网可以分为并网型和离网型两种，其中离网型发电系统一般应用在海岛等相对偏僻的地区，发电功率较小，其装机容量占光伏发电总容量的比重也很小，主要用于满足当时当地的小规模用电需求。并网型光伏电站则通常分为大规模并网型光伏电站和分布式并网光伏发电系统，都需要接入电能调度管理系统与不同来源的电能进行协调。这也是目前光伏发电的主要方式。由于绝大多数的并网型光伏电站都不配备储能设备，因此其产生的电能通常直接接入电网中，并根据昼夜变化来进行关闭和运行。

2． 光伏发电预测的要求与困难

火力发电、水力发电、核电等常规的电力来源通常都具备连续、可控的特征[3]，但根据文献[4]提出的光伏发电效率方程式，光伏发电的出力很大程度上取决于太阳辐照度和其他气候状况，如温度、云量、相对湿度、晴空指数、风速等[4]。这些气象因素对光伏组件的发电性能会产生剧烈的影响。文献[2]对光

伏发电机制的研究中就指出，局部的阴影可能会导致严重的功率损失。如果阴影面积占组件总面积的 10%，光伏出力的功率可能减少至原来的 20%；而当遮挡面积达到 20%以上时，组件的发电功率几乎为 0，因此云层、水汽等因素会给光伏发电带来严重的不确定性[2]。这些因素不仅随时间发生着周期性变换，还很容易出现突变。这使得光伏发电的功率存在显著的非线性性和间接波动性。

当光伏发电所占据的比例较小时，其波动对于发电——输电——用电的平衡状态所带来的影响并不明显。随着光伏电站装机容量的快速增加，其出力波动需要电力调度系统提供与其容量匹配的备用电源来平抑其波动，以保证功率和频率的稳定，这不仅意味着需要一定装机容量的电源长期处于准备状态，同时意味着需要进行频繁的并网操作。如果光伏发电的功率持续处于不能被稳定、精确预测的状态，不仅会带来能源的浪费，还有可能给电力系统的安全造成威胁。而如果能够对光伏出力功率进行准确、及时的预测，一方面可以为电网管理和决策提供重要的数据支撑，帮助其进行电网的功率调控和调峰调频，同时还可以减少备用容量，降低燃料的运行成本。从光伏发电运营商的角度来看，有效的预测方法可以改善其经济效益和投资回报，帮助其合理安排发电单元和逆变器的维护检修等工作，降低由于不确定性导致的惩罚和损失，从而进一步提升光伏电站在经济上的竞争力。

对光伏发电功率的预测工作近几年成为了能源行业和电力行业关注的重点领域，并出台了相关的行业标准。根据《光伏发电站功率预测技术要求》（征求意见稿）和《光伏发电功率预测系统功能规范》（征求意见稿）中所提出的要求，超短期光伏预测应能够以 15 分钟的分辨率预测未来 15 分钟至 4 小时的输出功率，且预测应 15 分钟执行一次，单次计算时间应少于 5 分钟。短期预测的时间跨度应延长至由次日零时起到未来 72 小时，每日执行两次，且单次执行时间应小于 5 分钟。这要求预测机制需要在充分研究变化规律的基础上，还需要具备时间复杂度低、可迭代、可自我修正等特征。本文在兼顾了模型对于短期预测的适应性的情况下，主要按照超短期预测的相关要求开展研究。

3. 辐照度预测和光伏发电功率预测的关联

在影响光伏发电功率的因素中，太阳的辐射是最重要的气象因素。太阳辐照度指物体在单位时间、单位面积上接受到的太阳辐射能，通常由气象观测设备直接测得。太阳的辐照直接引起了光生伏打效应，照射到光伏电池板的太阳光促使单晶硅的电子发生定向移动，进而产生内建电场，在外接负载时就会形成电流同路。在 25 摄氏度下，一定范围内太阳辐照度越强输出电流越大，输出功率也相应越大，且最大功率点的位置也相应增高[1]。因此太阳辐照度的大小直接影响了光伏电池组件输出功率的大小。同时，太阳辐照度和光伏电池组件输出功率之间存在物理意义上可拟合的二次曲线关系[5]。如果能够较为精确地预测太阳辐照度，再利用不同光伏电站发电单元的有关物理模型，就可以精确地对光伏电站的发电功率进行间接预测。

通常，根据天文学公式，大气层外切平面的辐射强度唯一取决于大气上界的太阳辐照强度和方向[5]，且可以精确地求出相关数值。这一数值也决定了某地区接受太阳辐照度的理论最大值。但是，太阳辐射在大气层中进行持续传输时，会直接受到许多自然环境因素的干扰，比如云、气溶胶和细颗粒、水汽和臭氧等。如何在这些复杂且混沌的气候系统中寻找潜在规律并进行预测，是太阳辐照度预测所面临的主要困难。

（二）现有的辐照度预测方法综述

目前，主要的辐照度预测方案都集中于超短期预测和短期预测，从方案逻辑上讲主要分为三个类别：基于历史数据的数理预测方法、基于图象处理的预测方法和基于 WRF 模式气象预报的预测方法。其中，基于历史数据的数理预测方法主要利用历史中的气象数据和辐照度数据，使用统计学、机器学习或经验方法训练出一个预测模型，其中部分方法还基于季节、地理位置、天气类型对样本进行了重组，或利用气象学、天文学概念对气象数据进行了前处理和特征提取，或融合了相关站点的地理信息和天文信息。基于图像处理的预测方法则主要考虑移动的云团对辐照度造成的影响，主要利用卫星图像或地基云图，对光伏发电站点上方的云团大小、厚度、高度和运动趋势进行分析，并结合地

外的太阳辐照强度计算其衰减程度，从而预测所在站点的太阳辐照度。基于WRF模式气象预报的预测方法则是利用成熟的天气预报系统提供的数值和相关的天气模型来对辐照度进行预测。

1． 基于历史数据的数理预测方法

基于历史数据的数理预测方法包括持续预测法、马尔科夫链预测法、小波分解法、神经网络法、随机时间序列法和组合预测法等[1]。

持续预测法假设未来时间的辐照度和此前较短时间内的辐照度存在一定程度上的一致性，因此采用此前一段时间的历史辐照度数据的滑动平均值来作为预测值。这一方法虽然计算简单，但对于日出、日落和天气突变极其不敏感，不适合作为短期、高精度的预测方法。多元线性回归算法将多元气象数据和时间作为自变量，经过变量筛选、预处理后训练出一个多元线性模型。但是辐照度本身存在强非线性性，尤其是在日落后持续为0，存在显然的截断现象，因此预测效果不理想。马尔科夫链预测方法则根据当前辐照度和其变化的趋势，使用状态概率向量和状态概率转移矩阵来描述辐照度在未来处于不同状态的可能性。但马尔科夫预测的基本假设也不能适应辐照度预测的要求，即辐照度发生状态改变的趋势和概率并不相同，除周期性的日出、正午、日落变化外，还存在频繁的随机变化。因此在跨昼夜预测、以及用于天气变化较为剧烈的场景下时会出现较大的误差[6]。

小波分解通常作为神经网络的辅助方法进行使用。通过对太阳辐照度的历史数据进行小波分解，拆分其中的高频分量和低频分量，再利用神经网络对各个分量进行单独的预测，最后通过小波重构得到太阳辐照度的预测值。神经网络算法的相关研究中采用了多种不同的网络模型，包括使用BP神经网络以气象数据作为输入，预测辐照度，或使用极限学习机（ELM）算法加快训练速度以适应短期预测对于时间复杂度的要求。为了更好地适应长序列，有研究使用LSTM长短期记忆网络，利用门结构控制对旧信息的遗忘，较为长期地保存一部分信息，从而增加对于时间序列的敏感性。通过对历史样本进行重组，可以对辐照度取得较好的预测效果。但由于网络深度相对较大，且如果试图投入较长的序列进行预测，则需要较强大的计算资源运行执行较长时间。在超短期预

测中，需要每 15 分钟投入新的迭代样本后并重新训练网络结构，可能不能符合单次计算不得超过五分钟的行业标准。

对于随机事件序列法，Pandit 等学者指出自回归滑动平均模型（ARMA）足以预报非线性时间序列，并进行预测，但预测的效果仍不够理想。组合预测方法则试图将上述方法与变量选择方法、样本组合方法进行不同的搭配，从而实现取长补短，包括使用小波分解与 LSTM 相结合，以强化训练数据的规律性；或使用层次聚类或支持向量机分类对训练样本进行聚合后单独训练；或将辐照度的周期性变化规律和神经网络预测结果进行加权平均；或尝试使用不同时间步长的训练样本来适应天气模式的转换等等。

2． 基于图像处理的预测方法

基于图像处理的预测方法主要包括基于卫星数据的预测方法和基于地基云图的预测方法。其中卫星数据指气象卫星在大气层外捕获的气象卫星云图，地基云图指使用包括 TSI 全天空成像仪在内的地面设备对云图进行采集[7]。获得相关图像后，通过图像去噪算法去除云图中由于绘制等高线而在不同云团间存在的干扰信息，并利用最大类间方差自适应阈值分割算法提取出其中的云团。虽然云团的运动和自身形态的改变非常复杂，但对于超短期预测来说，云团在超短时间内不容易发生剧烈的形态变化，可以基本假设为云团保持原有形状在图像上发生平移[5]。再利用地理学手段计算其在地球表面构成阴影的深度、形状、范围和变化趋势，并据此判断云团是否对光伏发电元件构成遮挡及其遮挡发电元件的程度。利用遮挡系数的数值和地外太阳辐照强度的数据来判断光伏电站未来的辐照度。

这种策略理论上可以非常好地预判由云团及其运动导致的辐照度衰减，但仍然存在一定的硬性缺陷。例如文献[8]中利用气象卫星云图的辐照度预测中，所使用的卫星云图在分辨率上不能满足单一光伏电站的要求。其时间分辨率为 30 分钟，空间分辨率为 2.5 平方千米。而根据国土资源部发布的关于光伏发电项目用地的相关规定，在充分考虑纬度和转化效率的前提下，我国的大部分光伏电站的占地面积不应大于每十兆瓦 0.5 平方千米，这也就意味着卫星云图提供的最小空间单位仍在光伏电站占地面积的五倍及以上，其时间分辨率也不符

合行业标准。在此基础上，如果采用地基云图进行预测，可以进行一定弥补。比如常用的 TSI-880 天空成像仪，可以通过一个较小的旋转式半球形镜面对天空图像进行捕捉和拍摄[8]。但其有效检测范围为 5 千米，几乎不能监测 5 千米外的云团移动。而在超短期和短期预测的最长时间尺度中，均有可能出现云团已经完成了跨域监测范围的移动的情况，容易导致预测模型的失效。

3． 基于 WRF 模式气象预报的预测方法

WRF（The Weather Research and Forecasting）模型是一个用于大气研究和数值天气预报的系统，其具有先进的数值计算技术，可以在中尺度的天气预报业务中取得较好的应用效果[9]。通过将大气划分为从地面垂直向上直到高空的柱体，并研究柱体内部的温度、风向、风速、湿度和压强等物理量来描述一个气象单位内的天气状况。它可以将多个影响辐照度的物理过程进行参数化而完成描述，包括微物理过程（水汽凝结等）参数化、积云参数化等。

但是 WRF 模型仍不是一个非常可靠的气象预测模型，尤其表现在其对于不常见的天气状况和突发天气的不敏感性[10]。对于辐照度预测，其在低云量天气下的辐照度预测的准确率较高，在高云量天气下，预测准确度出现明显的下降，且随着云量增多，辐照雨预测误差也会增大[9]。研究中通常采用按天气类型进行分类后预测、利用多元线性回归进行变量筛选等方案进行订正，但实际作用有限。这是因为 WRF 模型在离散化过程中存在一定问题。由于其在时间积分方面采用了三阶 Runge-Kutta 时间分割方案：

$$\Phi^{*} = \Phi^{t} + \frac{\Delta t}{3} R\left(\Phi^{t}\right) \tag{1a}$$

$$\Phi^{**} = \Phi^{t} + \frac{\Delta t}{2} R\left(\Phi^{*}\right) \tag{1b}$$

$$\Phi^{t+\Delta t} = \Phi^{t} + \Delta t R\left(\Phi^{**}\right) \tag{1c}$$

而固有 WRF 模型中对于时间步长Δt采用定值，规定其时间步长（单位秒）不超过 6 倍的最外层网格空间步长（单位千米）。从信号学角度来说，由于采用了固定的采样间隔，其对于高频分量和低频分量分别进行离散化所取得的效果显然不同。

根据采样定理的要求，如果采用充分小的时间步长就可以满足高频分量的要求，但如何取值、要求的采样频率是否会高于实际分辨率等等问题不能被解决。如果不能实现针对不同气象维度、不同时间的动态调整，那么其必然会存在较强的不稳定性。应当通过功率谱分解出不同气象因素的组成分量后，再根据 Shannon 采样定理确定一个最大的采样间隔以提升离散化的质量。这是本文的主要任务之一。

除本文提到的上述三类辐照度预测方案之外，研究者还将三种方案进行组合使用或交叉印证来进一步缩小误差，有的取得了较好的成果。例如文献[5]通过利用数值方法预测的辐照度综合地基云图分析计算出的遮挡参数，来预测辐照度，并利用 WRF 模型提供的气象数据进行矫正，从而尽可能利用不同方法的优势，提高预测的准确度。

二、基于模式识别的天气类型识别和昼夜识别

模式识别，指对表征信息进行分析处理，再进行描述、解释、辨别和分类等工作。应用中，可以利用统计模式识别的方法进行同类样本的聚合和不同类别样本之间的区分。而客观上，气候因素在发生变化的过程中，存在显著的周期性、日间相似性、昼夜区别和季节区分。本节中试图利用统计模式识别的方法，对历史样本之中的相似性和区分度进行评估，从而对其进行辨别和分类，为设计和组合训练样本提供标准。

（一）利用聚类方法的天气类型探索和相似日识别

1． 天气类型设计和预处理

为了能够利用天气状况对气象数据进行区分，相关研究者根据需要提出了多种天气分类方法。文献[1]主张以晴、晴到多云、阴、阴有小雨、小雨转大雨以及雪等日类型进行划分，而文献[11]则认为过多的分类可能导致部分类中数据间断明显、样本量不足等问题，从而主张按表 1 划分为四类[11]：

类	天气状况
A	晴、晴转多云、多云转晴
B	多云、阴、阴转多云、多云转阴、雾
C	阵雨、雷阵雨、雨夹雪、小雨、小雪、阵雪、冻雨、小到中雨、小到中雪
D	中雨、大雨、暴雨、大暴雨、特大暴雨、中雪、大雪、暴雪、中到大雨、大到暴雨、暴雨到大暴雨、中到大雪、大到暴雪、沙尘

表 1 天气状况四类分类法

但上述两个研究没有提出如何解决某一日内发生类别转移的情况。为了避免这一问题，本文忽略天气类型的标签，选择使用无监督的聚类方法针对气象

数据进行聚类。为了保证每一类天气状况中都有充足的样本数量，采用文献[5]中的建议，设定聚类数为 4。

由于样本数据为 15 分钟分辨率的气象观测数据，而进行气象分类的对象应为气象学意义下的一日，因此对每日气象数据进行重组，用于形成描述每日气象状况的向量。对于一个自然日，按照气象预报对每日气象状况的表示方法，将其气象状况的描述为

$$X_i = [T_{min}, T_{max}, H, W_S, W_D, R, F]^T \tag{2}$$

其中各变量对应意义和计算方式如表 2：

变量名	物理意义	计算方法
T_{min}	最低气温	$T_{min} = min\ Temprature$
T_{max}	最高气温	$T_{max} = max\ Temprature$
H	平均湿度	$H = \frac{1}{96}\sum_{n=1}^{96} Humidity$
W_S	平均风速	$W_s = \frac{1}{96}\sum_{n=1}^{96} Wind\ Speed$
W_d	主要风向	$W_d = \frac{1}{96}\sum_{n=1}^{96} Wind\ Direction$ 并将其划归至东、东北、北等八个方向之一
R	日出时间	辐照度由 0 转为非零的第一个时刻对应的时间
F	日落时间	辐照度由非零转为 0 的第一个时刻对应的时间

表 2 气象状况向量的物理意义和计算方法

为了适应聚类分析中计算样本间距离的需要，需要将不同维度数据转化至同一量纲下。因此对于以上的各维度$x \in X$，采用$x = \frac{x}{max\ x}$将其限制$[0,1]$区间内。

2． 聚类分析

聚类是一种典型的无监督学习方法。其训练数据中不含有关于其类别的信息，而完全根据其属性之间的关系，发掘其中潜在的类别。聚类任务中，通常将样本中性质相近的样本归为同类，将性质差异较大的分在不同的类别当中。数学上，对于不同的变量常用相关系数、相似系数和距离来衡量其关系。对于无序的样本集，常用的聚类方法包括系统聚类法、一次形成聚类法、K 水准逐步形成聚类法、移动中心聚类法等[12]。对于本文的数据集中存在的样本容量很大的情况，有必要选择计算量较小的方案，以适应快速预测中对于运算时间的要求。移动中心聚类法被经验证明对于大数据组很有效[12]，且计算更为简单。因此采用移动中心聚类法。

对前文提及描述每日气候状况的特征向量X_i，使用欧式距离d反映不同样本之间存在的差别，并作为聚类的依据。其中第i个样本与第j个样本之间的欧式距离定义为：

$$d_{ij} = \sqrt{\sum_{x \in X} \left(x_i - x_j\right)^2} \tag{3}$$

根据目标分类数为 4 类, 移动中心聚类法的实现步骤如下：

Step 1: 用随机挑选的方法从n个样本中选择 4 个样本作为类的中心；

Step 2: 对于I中的所有样本，计算其到每一个类的距离d，并将其划分给距离最近的类别；

Step 3: 根据新划分的分类，计算类中样本在各维度的算术平均值，并以此作为新的类中心；

Step 4: 重复 Step 2 和 Step 3 的步骤，对所有样本反复进行分类。

循环停止时，需满足下列两个条件之一：

（1）达到实现规定的迭代次数；

（2）相继的两次迭代得到相同的划分。

循环停止后，以最后一次的分类情况作为聚类结果。

根据上述算法执行聚类任务，以自然日为单位将样本划分至四个自然聚类中进行保存。

（二）基于支持向量机的昼夜识别

将感光元件不能感测到阳光的时段定义为夜间时段，将感光数值不为 0 的时段定义为日间时段，则光伏电站采集的气象和辐照度数据可以被划分为昼间数据和夜间数据两部分。其中夜间时段观测到的辐照度稳定为 0，因此对夜间辐照度进行预测没有实际意义。但如果昼夜之间的气象数据存在显著差异，则将夜间气象数据加入训练样本时，有助于训练天气——辐照度之间的潜在映射；如果差异不足以区分昼间和夜间数据，则意味着训练样本中的夜间时段数据会影响模型的可靠性。因此尝试通过分类任务，来判断其气象数据间存在的差距是否足够实现实现昼——夜之间的区分。

分类任务的本质是通过学习得到一个目标函数f，使得对于每一个属性集x都可以将其映射到一个类标号y[13]。分类模型通过拟合输入数据中的类标号和属性集之间潜在的映射关系，从而预测未知样本的类标号。本任务中，这一模型可以用于描述类间差距，用于区分不同类中所包含的对象。

对于一组线性可分的样本，即存在一个超平面，使得其中正例分布在分离超平面的正类一侧。其中，样本点和超平面之间的间距可用于表示对其分类为正类的确信度。如图 1，对于由直线f表示的分类法则，点A距离超平面的距离更远，预测其为正类时的确信程度就越大。

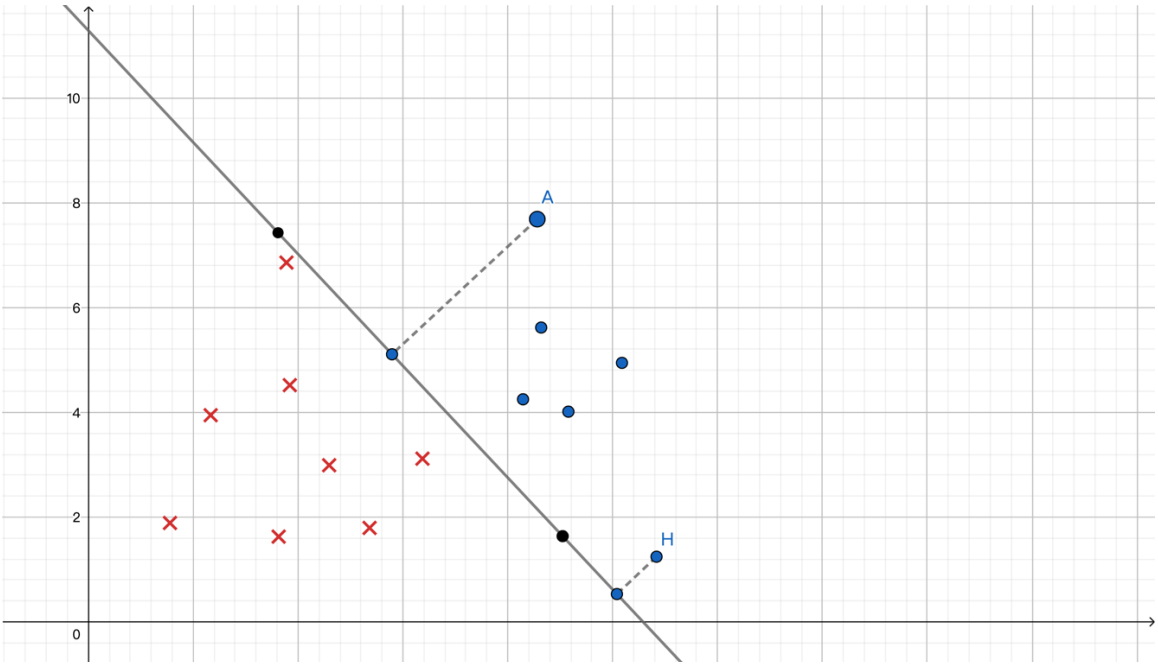

图 1 分类法则和确信度的示例

在设定超平面$w \times x + b = 0$保持确定的情况下,$|w \times x + b|$可以相对表示其距离超平面的距离，而其符号与类标记y的符号是否一致则表示其分类的正确性。考虑到(w, b)发生成比例变化时，上述结果也会发生改变，因此用

$$r_i = \frac{w}{||w||} \cdot x_i + \frac{b}{w} \tag{4}$$

表示样本点x和超平面(w, b)之间的距离。此时如果(w, b)发生成比例改变，样本点到超平面的距离不变。如果将超平面关于训练集的几何间隔定义为样本集中的全部样本点到超平面的最小几何距离，即

$$r = min\ r_i \tag{5}$$

那么对于支持向量机方法，其目的就在于寻找一个能够正确划分训练集并取得尽可能大的几何间隔的分离超平面的方法，其中，将距离超平面最近的样本点称为支持向量。在寻找超平面的过程中，只有少数的支持向量对其起到约束作用。

考虑到昼夜的区分本质上是气象状态的区分，而{风速，风向，温度，湿度，压强}这五维数据作为对气象状态的一种低维度描述，且数据中有噪声和特异点的存在，因此数据在原始维度上是线性不可分的，使得线性可分问题中的支持向量机（SVM）学习方法不再适用，应试图采用一个非线性模型，对样本数据进行正确分类。SVM 学习方法在这一问题上，利用核函数设计一个非线性变换ϕ将原空间下的样本$x \in X$映射到新空间中的样本$z \in Z$满足$z = \phi(x)$。经过变换，样本是在新的高维空间中是线性可分的，从而可以将原空间的线性不可分问题转换成新空间中的线性可分问题。考虑到特异点和噪声的存在，引入松弛变量ϵ和惩罚系数C，使得对于某些特异点，将函数间隔与松弛变量相加后可以获得正确分类。而惩罚系数$(C > 0)$越大，对于错误分类的情况的惩罚越大，训练出的超平面越严格。

基于这样的分类方法，将夜间观测数据赋予标签 1，日间观测数据赋予标签 0。选取全部样本数据中 70%的样本作为训练集，在其上利用支持向量方法试图训练一个尽可能好的分类超平面，并在测试集中评估其分类效能。为排除所选气象数据维度不同带来的影响，设定相同的随机数种子，随后取出温度、风速、风力、湿度、压强五维数据中的 22 个组合进行评估。为寻找较强的分类标

准，设定惩罚系数为 5.0，采用线性核函数，进行二分类训练。分类准确率如图 2。

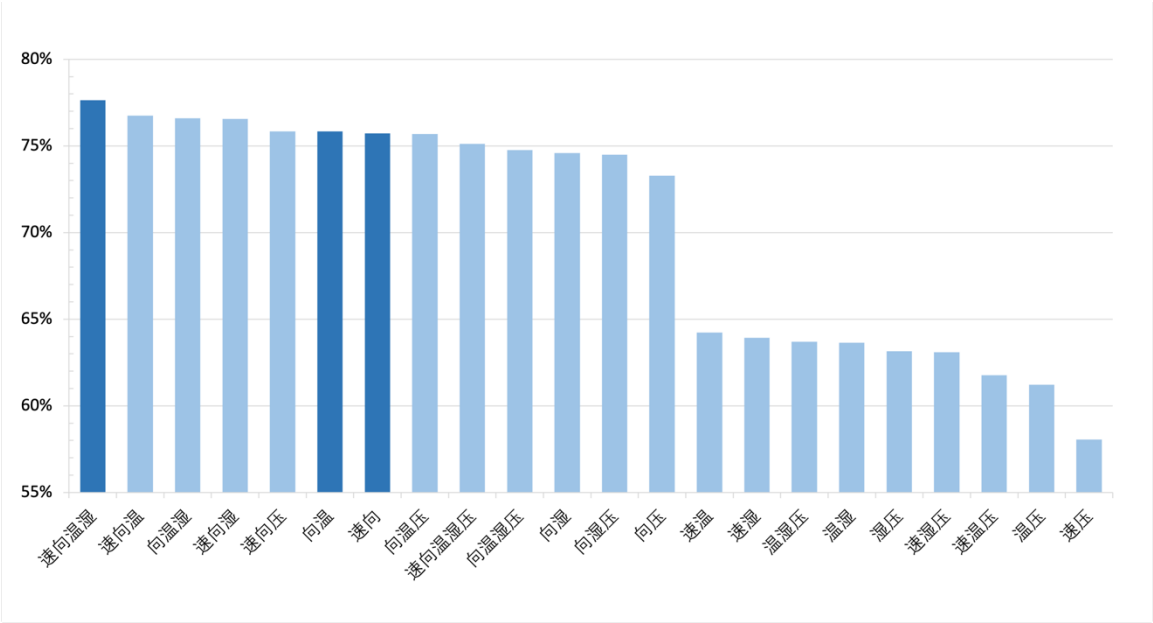

图 2 不同数据组合下的分类准确率

训练结果显示，在这 22 种组合方式中，选用{风速，风向，温度，湿度}四个维度的数据时，训练出的分类器的分类效果最好，准确率达到 77.64%；只采用{风速、压强}两个维度的数据时，分类效果最差，准确率仅为 58.05%。同时考虑到训练样本和测试样本中昼间和夜间样本的数量均相等，因此这一模型的分类效果并不显著好于随机猜测。

同时，有 8 种组合方式下的训练准确率可以保持在 75%以上，其中使用{风速，风向}和{风向、温度}可以实现在最少的样本数据的情况下获得较好的分类效能，准确率分别达到 75.72%和 75.84%。这说明，风和气温是昼夜之间气候最显著的区别，而压强变化会对昼夜区分起到干扰作用。湿度对于昼夜区分的任务带来的作用不显著。这说明在试图进行昼夜区分任务时，应尽可能选取与描述风、温度有关的物理量，而忽略描述气压的物理量。

但值得注意的是，在剔除气压这一干扰因素的情况后，{风速，风向，温度，湿度}这四个维度的数据在分类时的准确率仅取得有限的提升，由 75.12%提升至 77.64%。无论采用何种组合方式，分类准确率均不能突破 78%，这说明试图利用已有的气象数据，支持向量机方法使得不少于 22%的样本在分类时出现多值和错误分类情况，进行昼时和夜时的准确分割是较为困难的，只能通过

人工添加标签的方法进行区分。在训练预测模型时，则应当将夜间的气象数据进行剔除。

（三）样本数据重组

原始数据为全天 24 小时内以 15 分钟为分辨率统计的{风速，风向，温度，湿度，压强，辐照度}的六维气象数据。根据上述的聚类分析和支持向量机结果，将每日的昼间数据从全体数据中提取，并根据每日的类别标签将样本分在不同的类别当中。根据其相对时间顺序，将其组成连续的信号序列，组成不同天气类型下的气象数据样本。

三、基于功率谱计算的可预测性分析

将气象数据视为一列随机信号，根据平稳性对信号的要求，对于任意的s，序列$(y_t, y_{t+1}, y_{t+2}, \dots, y_{t+s})$的分布与$t$无关。对于气象数据而言，其季节性变化可能影响其平稳性[14]。但考虑到上一节中聚类结果对于季节性变化的分解，重组后的气象数据序列仍具有一定的平稳性。为了进一步改善其平稳性以适应平稳信号处理的需求，将信号序列拆分成多个较短的平稳序列，切割为长度不大于104个自然日的共计9984条记录。此时信号不出现由于季节改变导致的显著的变化趋势。而由于任意信号总可以被表示成谐波信号和白噪声的叠加，因此认为切割后的信号序列适合研究其被有限多个简单谐波拟合的理论可能性。

（一）基于SFGPS和AFAGPS功率谱分析与可预测性分析

将切割后的平稳序列视为一系列谐波序列与白噪声的叠加，即气象信号$\tilde{X}(t)$可以被

$$X(t) = \sum_{k=1}^{K} A_k \ cos(\omega_k t + \phi_k) + e_t \tag{6}$$

拟合，其中$\phi_k \in [-\pi, \pi]$满足均匀分布，e_t为白噪声过程。对于序列$X(t)$，其功率谱为

$$P(\omega) = \frac{\pi}{2} \sum_{k=1}^{K} A_k^2 \ \delta(\omega - \omega_k) + \sigma^2 \tag{7}$$

其中σ^2为 e_t的方差。

因此，通过对平稳序列$\tilde{X}$进行功率谱分析，可以根据分析结果确定出组成$X(t)$的谐波过程的功率和频率参数。

SFGPS(Simple and Fast algorithm to Genetrate Power Spectrum)算法指出，相比于Welch方法计算功率谱时，将整体信号使用滑动窗进行分段处理的方法，如果对周期图直接进行如下方式的滤波：

$$A^2(m) = \begin{cases} A^2(m), & if A^2(m) > \overline{A^2} + \sigma^2 \\ \sigma^2, & else \end{cases} \tag{8}$$

可以将计算功率谱的时间复杂度由$O(N^2logN)$降为$O(NlogN)$。且针对离散的谐波过程的叠加，与 Welch 方法的效能一致。需要注意的是，滤波方法对于波动非常明显的序列存在一定的不适应情况，如方差σ^2过大时，滤波反而会出现适得其反的状况。要解决这一问题，可以对原信号进行一定程度的放缩，从而实现理想的效果[15]。

此外，AFAGPS（Accurate and fast algorithm to Generate Power Spectrum）也可以非常准确地捕捉功率谱。AFAGPS 算法做出以下三个重要改进：

1. 绕开对于 Yule-Walker 方程的直接求解，而转而对其进行改写，使用快速傅立叶变换和逆变换获取所需的关于振幅的参数，将时间复杂度由求解 Yule-Walker 方程的$O(N^3)$降至$O(NlogN)$。

2. 将估计相关函数的方法由传统的渐进无偏估计

$$\hat{R}(m) = \frac{1}{N-|m|}\sum_{k=0}^{N-|m|-1}(x(k)-\mu)(x(m+k)-\mu), \quad m = 0,1,2,\ldots \tag{9}$$

改为圆周卷积

$$\hat{R}(k) = \frac{1}{N-k}z(k), \quad ,k = 0,1,2,\ldots,K < \left[\frac{N}{2}\right] \tag{10}$$

其中$z(n)$是$x(n)$和其逆序序列$x(-n)$的圆周卷积。此时时间复杂度由$O(N^2)$降为$O(NlogN)$，且可以适应数据长度较长的情况。

3. 在设定$K\left(\leq\left[\frac{N}{2}\right]\right)$足够大的情况下，用$\left(\hat{R}(0),\hat{R}(1),\ldots,\hat{R}(K)\right)$估计出的近似功率谱即可用于估计真实功率谱[15]。

用上述两种算法对五个气象维度的信号序列分别进行功率谱分析，截取的信号长度为 9984，所得结果如图 3。根据功率谱分析结果，由于其最大截止频率均远离中心π，且在两种算法下的功率谱结果具有较强的一致性，所以五维气象数据确实可以用离散或连续的谐波过程进行拟合。在 AFAGPS 算法下，风速和风向表现出可能由连续的谐波构成，温度、湿度、压强可能由离散谐波构成；SFGPS 算法下的结果则验证了温度、湿度、压强三个维度可能由离散谐波构成。对于风速、风向维度，如果用较多个离散的谐波，可能可以在容忍一定误差的情况下实现拟合。

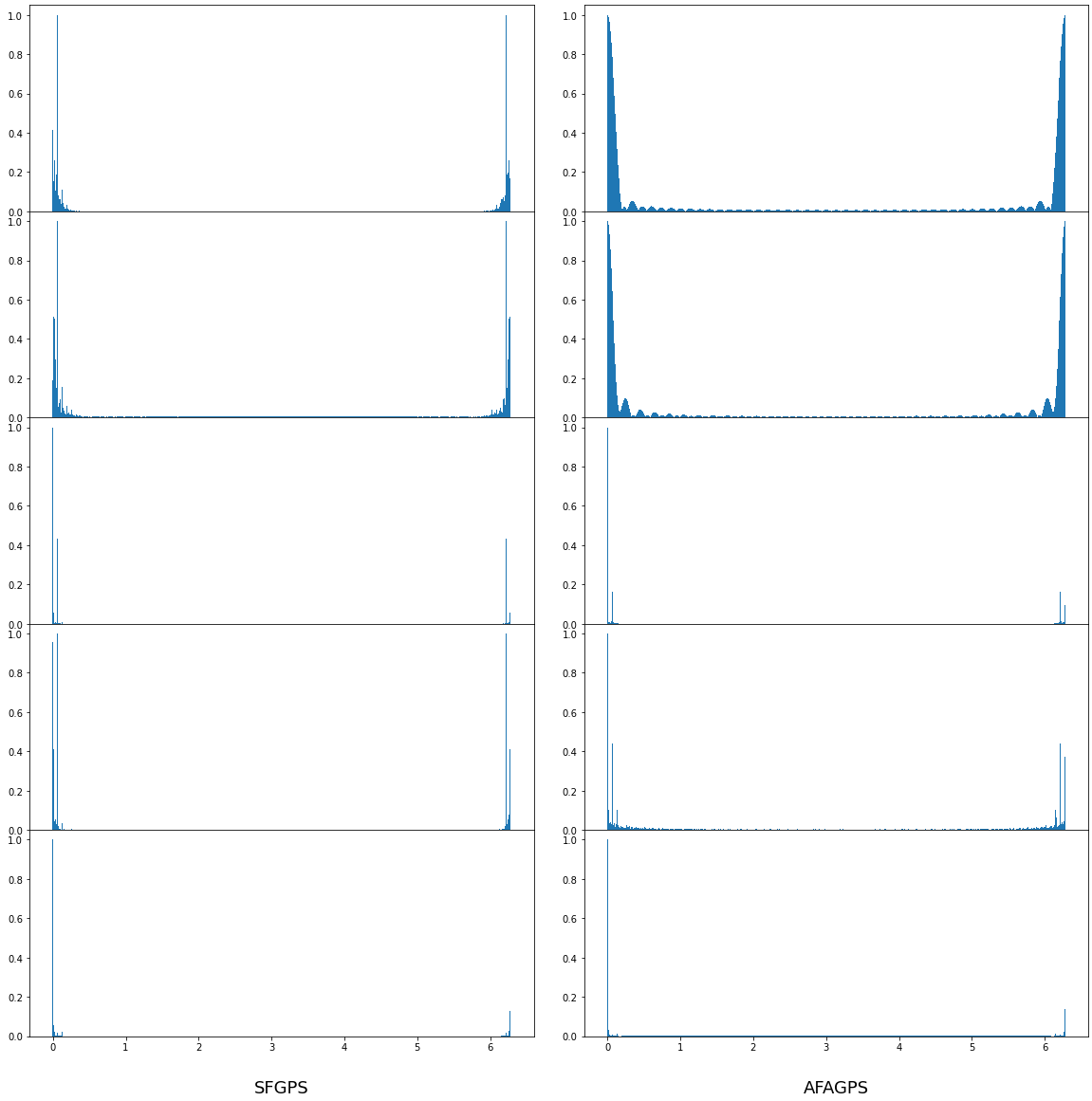

图 3 SFGPS 和 AFAGPS 算法下五维气象数据的功率谱图像

（二）采样频率确定与样本插值

对于一个有限长的信号，当其频谱的支撑有界时，认定其频带有限，也就是存在一个最大截止频率。对于时间域上的信号，经典的 Shannon 采样定理指出，如果其存在最大截频f_c，那么要想根据信号的采样序列恢复出原序列，需要采样列信号$x(n\Delta)$的采样间隔$\Delta \leq \frac{1}{2f_c}$，且$f_s = \frac{1}{\Delta}$。此时可以利用对信号进行如下重构

$$x(t)=\sum_{n=-\infty}^{+\infty}x(n\Delta)\frac{sin\frac{\pi}{\Delta}(t-n\Delta)}{\frac{\pi}{\Delta}(t-n\Delta)} \tag{11}$$

换句话说，从信号学的观点出发，对于一个平稳的随机序列，可以被一个谐波过程逼近。而对于一个谐波过程的观测，其采样频率满足 Shannon 采样定理的要求时，这个观测才可以被恢复成完整的谐波过程，从而利用这一谐波过程对未来数据进行合理的预测。

五个气象变量拆解形成的谐波分量中，高频的分量所需要的采样频率更高，如果采样频率低于这些高频分量对于最小采样频率的要求，那么这些高频分量在预测中极有可能出现丢失。这些高频的谐波分量，在气象学中往往对应天气的突变和非常规现象的出现。现有的预测模型对此类天气状况普遍效能较差，这很大程度上是由于其使用的观测数据频率不能满足恢复高频分量的要求，只能对低频分量进行较好的还原，从而只能对常规天气状况进行合理的预测。如果想要对极端天气进行及时、准确的预测，就需要将其中的高频分量进行合理的恢复，这要求研究者采用更高频率的气象观测方法，或对现有数据进行插值和加细，人为提高其采样频率。

图 4 中展示了连续的三天中五维气象数据的分布情况，y轴的起始坐标均设定为 0。气候中风相比温度、压强更具随机性和不连续性，更加容易出现突变，因此风速和风向两个维度所需要的采样频率较高。而本文使用数据所在的电站处于我国北方地区，其湿度的日内变化非常明显，因此应需要较高频率的采样数据，以精细描述其发生的突变。相比之下，一天中温度的变化是相对平滑，变化幅度小、速度慢，压强则始终保持在相似的数值，这两者对应的采样频率应相对较低。

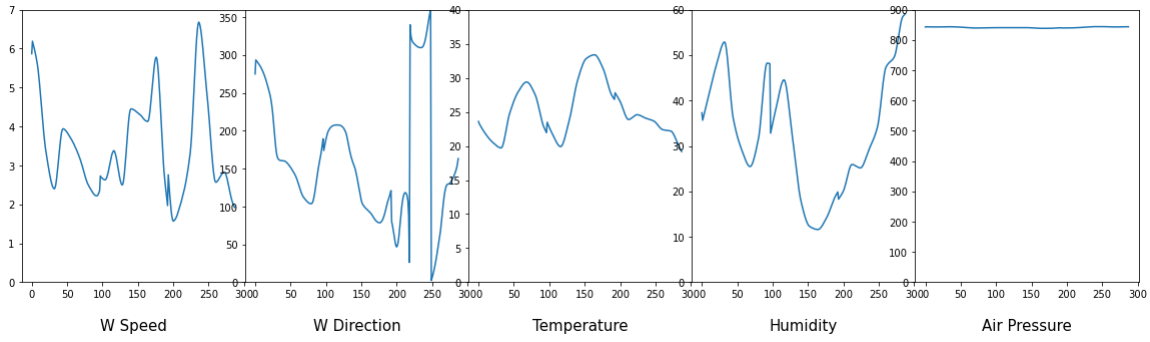

图 4 三天内五维气象数据折线图

利用 AFAGPS 和 SFGPS 计算的功率谱数据，对不同维度的信号序列所需的采样频率进行定量计算。过滤掉功率小于最大功率的1%的谐波后，最后一个功率不为 0 的谐波对应的频率为截止频率。在两种算法下，某一站点的五维气象数据的信号序列的最大截止频率、最小采样频率、最低采样分辨率（采样间隔）和插值后分辨率如表 3。

维度	最大截频（AFAGPS）	最大截频（SFGPS）	最小采样频率	最低采样间隔	插值后间隔
风速	0.454π	0.065π	0.908π	2.203min	1.5min
风向	0.525π	0.115π	1.050π	1.904min	1.5min
温度	0.023π	0.021π	0.046π	43.478min	15min
湿度	0.242π	0.042π	0.484π	4.132min	3min
压强	0.042π	0.042π	0.084π	23.810min	15min

表 3 五维气象数据的最大截屏与最小采样间隔

SFGPS 和 AFAGPS 在{风速，风向，湿度}上计算出的最大截频存在较大差异，而对于{温度，压强}，计算出的最大截频则存在一定误差，这在一定程度上是由于两种算法选取的滤波标准存在差异而导致的。对于信号序列，采样频率只要高于采样定理限定的最低频率，都可以用采样序列还原出原信号。因此选用两种算法下较大的截止频率作为该信号的截频，并由此计算出最小采样频率和最低采样分辨率。结果显示，{风速、风向、湿度}这三个维度的采样分辨率要求均高于原数据的 15 分钟分辨率，有必要进行插值和加细处理。为了与原数据的分辨率相适应，选择了不大于最低分辨率且可以被原分辨率整除的插值分辨率。相比之下，温度和压强的采样分辨率大于 15 分钟的要求，不需要进行特别的插值处理。对于需要进行插值加细的位置，考虑相邻的观测数据(x_0, y_0)与(x_1, y_1)，按照分辨率要求使用线性插值法进行数据加细

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{12}$$

重新组成满足 Shannon 采样定理要求的信号序列。

四、利用 MMMLP 的辐照度预测

（一）算法简述

多层感知机（Multilayer Perception）是由大量神经元组成的复杂计算网络，具有强大的非线性性。其中利用误差反向传播训练出的神经网络，通过将样本数据通过隐藏层的权值处理后在输出层与学习样本进行比较[16]。根据误差，进入反向传播过程，沿误差下降的最快方向调整各个单元的权值，并持续循环至达到训练轮次要求，或误差低于希望的训练误差。

传统的多层感知机在一次输入中只能传入单个向量，且无法考虑到训练数据间的时间序列关系。为了使得多层感知机能够适应辐照度预测中，对于输入多个向量、保留其时间顺序的需求，对其输入向量的方式进行了改进，得到了多元多步多层感知机模型（MMMLP）。其中，多元指选用多个维度的自变量，多步指预测时使用多个时间步长的训练数据，得到多个时间步长的预测数据。对于标准的 MLP 模型。输入层需要输入单个向量，因此多元多步 MLP 将输入数据进行展平，从而将多个时间步长的数据融合在一次输出中。选取预测时间前 8 小时的数据，对未来 4 小时的辐照度进行预测。选用 Adam 优化器和 ReLU 激活函数，执行 500 次循环的学习过程，利用均方误差（MSE）估计其训练误差。

为了公平地比较预测性能，将未经过处理的训练数据与本文重组后的训练数据分别作为 MMMLP 的学习样本，保持其他训练参数相同，在相同的测试数据中评估其预测效能。

（二）预测性能分析

图 5 中显示优化后的 MMMLP 训练结果和原始 MMMLP 训练结果，截取前 11 日内的辐照度预测情况。从图 5 中注意到，两种方法均在总体上实现了较好的预测，两种模型在峰值预测上均较为保守，且都对异常天气有一定的探测能力。

从二者的比较中来看，优化 MMMLP 在预测效能上整体优势明显。第一日内，两种算法的预测结果均与实测结果差异明显。这很大程度上是由于第一日采用的气象样本数量较少导致的。在第二日预测中，优化后的 MMMLP 的预测误差显著小于原始 MMMLP 的预测结果，说明优化方法对于数据量小的预测状况适应更加良好。第三日作为辐照度峰值下降的特殊天气，原始 MMMLP 在面对此类特殊天气时表现出预测准确度严重不足的问题，出现了违背自然规律的波动；相比之下优化后的 MMMLP 和实测数据拟合程度高，对当日峰值的预测准确。

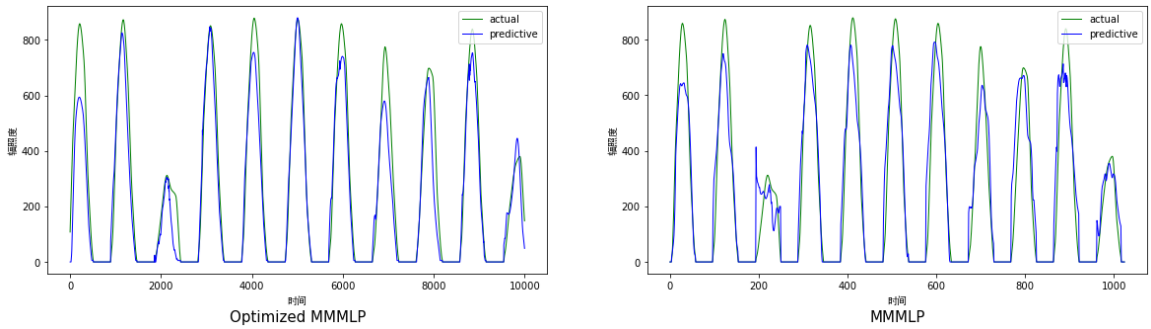

图 6 优化 MMMLP 和原始 MMMLP 的预测结果与实际值

此外，在如第四日、第六日的气象状况下，优化 MMMLP 方法的预测准确率极高，相比原始方法有明显提升，尤其在峰值的大小和时间预测方面有明显改善。整体上，原始方法所预测的峰值都出现明显的提前现象，而优化 MMMLP 方法对峰值时间的预测基本准确。

从数据分析的数值结果来看，优化 MMMLP 的性能相比 MMMLP 也提升明显。使用平均绝对误差评估两种算法的预测准确性，优化 MMMLP 方法的平均绝对误差为 44.732，原始 MMMLP 方法的平均绝对误差为 56.002，优化方法在原有方法基础上，平均绝对预测误差减小了 20.12%。定义每天辐照度的最大值为当日辐照度峰值，优化 MMMLP 方法的峰值平均绝对误差为 94.863，远低于原始 MMMLP 方法的峰值平均绝对误差 451.94。图 6 是两种算法下峰值预测误差的频数分布直方图。

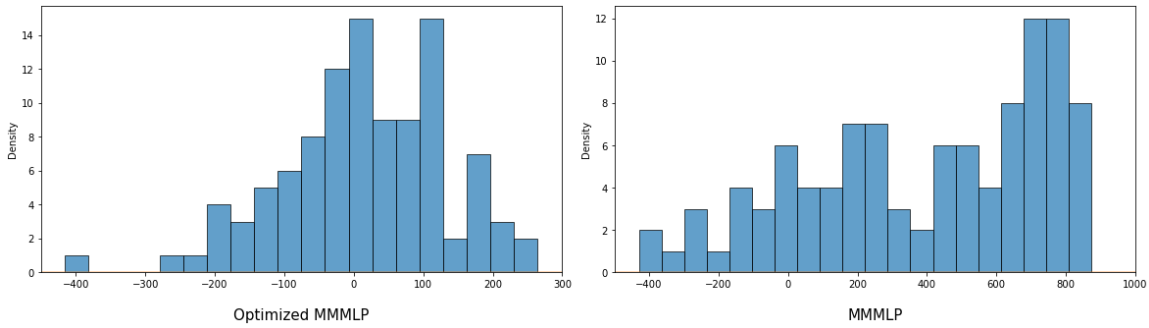

图 6 优化 MMMLP 和原始 MMMLP 峰值预测误差的直方图

显然，原始 MMMLP 中呈现出明显的保守预测倾向，存在大量样本的峰值预测误差高于 600，即存在较多的峰值存在而没有成功预测的情况。且大部分样本预测误差为正，说明模型在预测峰值时有明显的保守倾向。在模型优化后的 MMMLP，误差更加接近以 0 为中心的正态分布，且预测误差普遍较小。因此在峰值预测方面，优化 MMMLP 的效能也远远强于原始 MMMLP 方法。

五、结语

基于对辐照度预测背景和对于常见的三类经典算法的总结，本文利用模式识别、信号处理和神经网络方法实现了一种基于历史数据的较好的辐照度预测方案。利用聚类方法，将存在相似性的日间样本重组为连续的序列，并利用支持向量机方法对昼间气象数据和夜间气象数据的差异性进行分析，得出了昼夜间数据需要进行分别处理的结论。随后，从信号学的视角出发，探索不同维度的气象数据作为平稳信号时具备可预测性的条件，并根据其功率谱分析中揭示的对于采样频率的要求，对样本数据进行补齐和加细。考虑到辐照度预测的强非线性性，运用重组数据优化的多元多步多层感知机方法对辐照度进行预测，并取得了显著好于原始预测效果的预测模型。值得注意的是，本文的预测方案是完全由基于历史数据所训练的数理模型得出的，没有参考卫星云图、气象预报数据等于辐照度预测高度相关的其他数据来源，在以后的工作中可以利用更加丰富多元的数据，为模型效能进一步的提高创造可能性。

参考文献

[1] 吴硕.光伏发电系统功率预测方法研究综述[J].热能动力工程,2021,36(08):1-7.DOI:10.16146/j.cnki.rndlgc.2021.08.001.

[2] 王飞. 并网型光伏电站发电功率预测方法与系统[D].华北电力大学,2013.

[3] 武天雨. 考虑气象影响因素的光伏出力预测研究[D].燕山大学,2021.DOI:10.27440/d.cnki.gysdu.2021.001196.

[4] Skoplaki E , Boudouvis A G , Palyvos J A . A simple correlation for the operating temperature of photovoltaic modules of arbitrary mounting[J]. Solar Energy Materials & Solar Cells, 2008, 92(11):1393-1402.

[5] 朱想,居蓉蓉,程序,丁宇宇,周海.组合数值天气预报与地基云图的光伏超短期功率预测模型[J].电力系统自动化,2015,39(06):4-10+74.

[6] 杨留锋.光伏发电预测中人工智能算法的应用研究综述[J].太阳能,2020(08):30-35.

[7] 周海,朱想,金山红,朱婷婷,张雪松,魏海坤.超短期太阳辐照度多模型预测[J].中国科技论文,2017,12(23):2695-2700.

[8] Kassianov E , Long C N , Ovtchinnikov M . Cloud Sky Cover versus Cloud Fraction: Whole-Sky Simulations and Observations[J]. Journal of Applied Meteorology, 2005, 44(1):págs. 86-98.

[9] 叶林. 基于 WRF 模式输出的光伏发电量预测研究[D].宁夏大学,2018.

[10] 李晶,许洪华,赵海翔,彭燕昌.并网光伏电站动态建模及仿真分析[J].电力系统自动化,2008,32(24):83-87.

[11] 钟志峰,张艺,张田田,杨晨茜,苏勇.一种简单的短时辐照度预测研究[J].计算机测量与控制,2017,25(07):181-185.DOI:10.16526/j.cnki.11-4762/tp.2017.07.045.

[12] 张润楚. 多元统计分析[M]. 科学出版社, 2006:266.

[13] Tweedale J W , Jain L C . Advances in Modern Artificial Intelligence[M]. Springer International Publishing, 2014.

[14] Hyndman R J , Athanasopoulos G . Forecasting: Principles and Practice[J]. London: Bowker-Saur. Pharo, 2014.

[15] 阮吉寿. 信号学讲义, 2022.

[16] Heaton J . Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks. 2015.

致　谢

明日世界，人类对他的好奇与迷恋已经足够写成一部与科学有关的秘史。用数学和统计学借给我的眼睛，我看到那些已被证实的工具，是如何帮我们在随机的汪洋里捉到一颗确定性的苇草。但再强大的工具也无法面对我们在第四维中无法逃避的现实：我们正在踏入未知，时间不会因你是否看清未来而改变它的脚步。

我能想起在某个春天的下午，出神时看到二主楼外的草地亮闪闪；我也记得漆黑的夜里，我胸口的火花如何跳动在无人的街道；我还会回忆出真诚和温情是如何流淌在安静的房间里，让一切死物都得以发芽。它们都提醒我，我曾努力地耕种于脚下的现实，把我的全副的热情、全身的力量都用于擎起生活的大笔，在我的书页上努力书写。刺痛的失落、发现的喜悦、等待的焦灼、跳跃的幸福……那些在我的田野上曾发生的，我都不曾让它漫不经心地流去。这是我能交给时间的最好的答案。

这旅途比我曾揣度的样子更加孤单，又比我全部的想象更加美丽。没有在阮吉寿教授的办公室里度过的时光、没有和我敬爱的几位老师的坦诚的交谈、没有南开数学和南开统计给我在科学和经验上的武装，我的力量和信念都将受到折损。我将永远感激他们为我的独立思想而带来的保护与支持。

如何预测未来是我的科学问题，而如何面对已知与未知的鸿沟则是我持久的人生命题。在我将要不慎落入失意的洋流当中时，她把我的快要尘封的光芒重新擦亮，珍重地陈列起来；我的父母攥紧我的手，引着我把失败装进行囊，永远重新出发；我的朋友，他们总是用信任的涓流牵引我的身体。这些连接让我重获险些腐朽的意义，让我在脆弱时刻得以锚定。他们共同绘制了我的人生底色。

回看高中时代留下的自我期许时，我的完整的故事，它的情节早已超出我的全部想象。而我终于可以不至于太过惭愧地说，我守住了多年前留给自己的关于真诚和勇气的承诺。我在卫津路 94 号和 92 号的生活，都将轻轻折叠好，放进衬衣的口袋。对我的国家、我的时代与我的生活，在为你们播种的时候，我把我最深沉的心愿也埋在了土地，我将为你们毫不犹豫地把手弄脏。未来的迷雾里有我的火把，你们将随我一起远行。