基于 KPH-GeoMAN 的光伏功率预测方法

刘宏博[1,2] 李寒[1,2]

1 北方工业大学信息学院，北京市，100144；

2 大规模流数据集成与分析技术北京市重点实验室，北京市，100144；

摘要：随着全球工业的快速发展，国内外对于能源的需求与日俱增,使用太阳能的光伏发电有着不稳定性，因此对光伏发电功率进行准确的预测是能源管理的重中之重。为满足这样的预测需求，本文考虑光伏功率预测的时空相关性特征，提出了基于 KPH-GeoMAN 光伏功率预测方法。该方法首先在 GeoMAN 模型基础上，构建了融合面向光伏功率预测的特征工程方法的光伏功率预测框架，由额外因素模块、编码器、解码器和特征工程部分构成；然后通过融合地理信息等多源时空数据加强模型对空间相关性的提取能力；最后，通过加入基于半正矢公式的临近场站选择方法进一步增强了模型对于空间相关性的学习能力，降低了低相关场站信息对模型的干扰，提升了模型的预测准确度。实验表明 KPH-GeoMAN 方法能够大幅度提高短期预测准确率。

关键词：光伏功率预测；深度学习；特征工程；时空相关性

DOI：10.69979/3041-0673.25.09.009

引言

在现代电网环境中，往往是传统火电结合其他种类的新型能源联合发电，并网输送。光伏发电功率的预测是为了保证电网合理调度,以降低对电网造成的冲击[1]，对光伏发电的功率进行预测是十分重要的。

由于光伏系统的发电功率取决于许多高度不确定的气象变量，如太阳辐照度、温度、相对湿度、云层厚度、风速等，这使得光伏发电功率具有很强的波动性和不可控性[2]，从而大大提高了光伏发电功率预测的难度。光伏发电功率的预测按照建模方式可划分为三类：物理模型、统计学模型和人工智能模型。人工智能模型是目前应用最广泛的一种，它将人工智能的算法与历史数据相结合，用现代计算机的机器算力进行迭代训练，最终建立历史数据与功率间的模型。基于人工神经网络的建模方法往往都有着较好的效果。

GeoMAN 模型以编码器-解码器(Encoder-Decoder)结构为框架，将长短时记忆网络(LSTM)作为基础单元，并加入多级注意力机制，在空气质量与水质数据集上的预测效果良好。本文在其基础上做出更进一步的工作：

（1）使用皮尔逊相关系数的方法，对原始数据集进行特征选择，选取保留相关性较高的特征因素。

（2）使用半正矢公式计算临近场站与目标场站间的地理位置距离，优化站间空间注意力机制。

（3）对历史数据进行天气类型的聚类，根据预测日天气类型进行样本筛选以求能更好的提取相似的特征。

1 相关工作

针对光伏电站时空相关性的研究，文献[3]提出一种深度时空特征提取的光伏发电功率预测模型，针对邻近区域的光伏电站进行图建模，使用 LSTM 模型进行时间特征提取，使用图卷积原理提取电站的空间特征。但是由于分布式光伏电站具有强空间性[4]，图机器学习还应该结合地理方位和云层运动，以达到更加精准的预测效果。

在利用人工神经网络进行时间序列预测时，对数据集的处理以及特征工程往往有着不可或缺的作用，能使模型更好地拟合。文献[5]利用皮尔逊相关系数（Pearson）逐一计算各影响因素与光伏输出功率的相关程度，选取辐照度和组件温度为主要影响因素作为预测模型的输入变量，通过实验得出新混合模型算法使得预测精度得到提高。文献[6]使用变分模态分解(Variational Mode Decomposition, VMD)结合斯皮尔曼相关系数(Spearman Correlation Coefficient, SCC)处理无关序列和异常值，筛除相关性不高的特征，通过大量实验证明，与其他方法相比具有更好的预测结果。考虑到光伏数据集的特征特点，皮尔逊系数计算相关度是一个合适的选择。

综上所述，上述方法都使用各种手段不同程度的加

强了光伏功率预测的准确度，但同时也各有各的缺陷，如在使用的基础神经网络模型时，经典的神经网络模型没有充分的利用到光伏数据的时空特征，可以在此方面进行提升。

2 方法

2.1 方法总体架构

本文首先对数据集进行数据预处理后，对数据集进行面向光伏功率预测的特征工程方法，包括利用基于皮尔逊(Pearson)相关性系数的方法进行特征选择和基于天气类型聚类的样本筛选方法将样本聚类成3种天气类型；然后，将样本输入到融合了时空数据的、加入多级注意力机制和基于半正矢优化的空间注意力机制模型中；最后对数据集进行训练与测试，方法总体架构如图1所示。

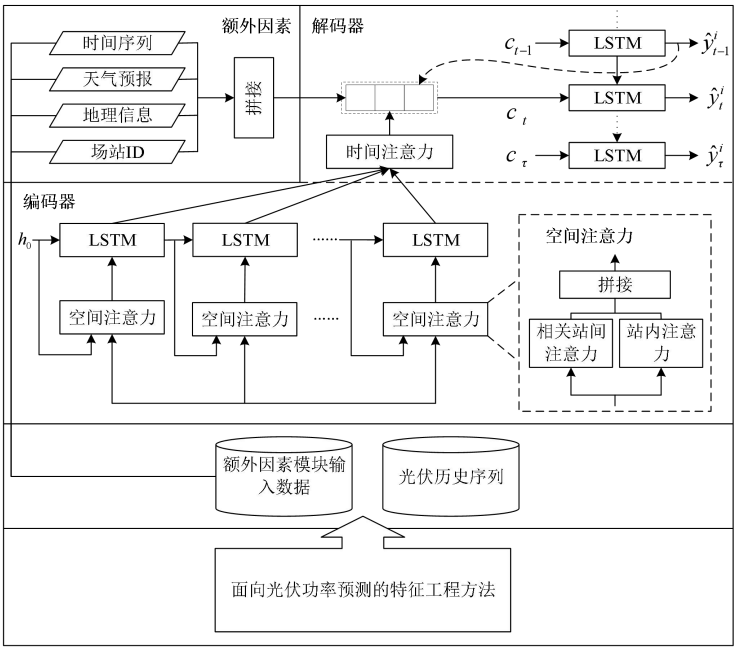

图 1 方法总体架构

2.2 面向光伏功率预测的特征工程方法

2.2.1 基于皮尔逊相关系数的特征选择

在一些数据集中，为了表征数据的某些特征，往往会有一些统计数据。此外，在对光伏发电功率的影响因素进行分析时，发现有的属性对于功率的影响微乎其微，于是，为了排除一些冗余属性，利用皮尔逊相关系数进行相关性分析，计算出各个属性与功率之间的相关度系数。皮尔逊相关系数公式如式(1)所示。

$$r=\frac{n\sum_{i=1}^{n}xy-\sum_{i=1}^{n}x\sum_{i=1}^{n}y}{\sqrt{n\sum_{i=1}^{n}x^2-\left(\sum_{i=1}^{n}x\right)^2}\times\sqrt{n\sum_{i=1}^{n}y^2-\left(\sum_{i=1}^{n}y\right)^2}} \quad (1)$$

其中：r 表示相关系数范围，$r\in[-1,1]$，其值大于0则表示该属性与目标属性之间正相关，反之则表示负相关的。n 表示属性的总个数。x 表示相关因素属性，y 表示目标属性。

使用式(1)计算出目标场站各属性与目标属性之间的相关度，选取相关度较高的几个属性，并将相关度不高的属性删去。

2.2.2 基于天气类型聚类的样本筛选

为了提升预测的准确度，可以对训练样本进一步筛选。根据历史数据集中所提供的辐照度的特征，可以使用 K-means++[7]算法将每天的天气类型聚类成 3 种不同的类型。K-means++ 算法通过以下步骤改进初始化过程：

（1）选择第一个中心：从数据点中随机选择第一个聚类中心。

（2）概率选择后续中心：对于数据集中的每个点，根据其到已选择的聚类中心的最近距离的平方来选择下一个中心。距离越远的点被选为中心的概率越大。

（3）重复：重复步骤（2），直到选择了 k 个聚类中心。

（4）聚类：使用这 k 个中心进行标准的 K-means 聚类。

对于光伏发电数据集，将目标场站根据每日辐照度特征均值作为聚类特征，使用 K-mean++算法划分为 3

种天气类型。而后基于已经划分的天气类型的类簇中心，计算出预测日的辐照度特征与 4 个类簇中心的欧几里得距离，并将其划归为其中一类。最后，保留目标类型的样本作为训练数据集输入到模型中，完成样本筛选。

2.3 基于 KPH-GeoMAN 的光伏功率预测方法

GeoMAN 模型是由 Yuxuan Liang 等人在文献[8]中发布的一种用于地理传感器时间序列预测的多层次注意力网络模型。该模型基于编码器-解码器架构，其中采用了两个独立的 LSTM（长短时记忆网络），分别用于输入时间序列和预测目标序列。本文在其基础上，首先融合了面向光伏功率预测的特征工程方法，构建基于 KPH-GeoMAN 的光伏功率预测框架，然后在原本的站间空间注意力基础上，加入基于半正矢公式的临近场站选择来进一步优化空间注意力机制；最后在额外因素模块中，融合了多源光伏功率预测数据，包括未来时刻的时间序列、数值天气预报 NWP、地理信息和场站 ID 用以增强预测准确性和提取时空相关性。

2.3.1 编码器

（1）目标场站的站内局部注意力机制

对于目标场站来说，影响光伏发电功率的因素有很多，但每个因素的影响力却又不尽相同。在这里，对于目标场站内的不同属性引入了一种局部特征注意力机制，设编解码器中的时间步长分别为 m, n，可根据式(2)、(3)计算出第 i 个场站的第 k 个特征向量的注意力系数。

$$e_t^k = v_l^T \tanh\left(W_l[h_{t-1}; s_{t-1}] + U_l x^{i,k} + b_l\right), \quad (2)$$

$$\alpha_t^k = \frac{\exp\left(e_t^k\right)}{\sum_{j=1}^{N_l} \exp\left(e_t^j\right)} \quad (3)$$

其中，h_{t-1}和s_{t-1}表示上一时刻编码器单元的隐藏状态和细胞状态，[x;y]表示按列拼接操作；$v_l \in \mathbb{R}^T$、$W_l \in \mathbb{R}^{T\times 2m}$、$U_l \in \mathbb{R}^{T\times T}$和$b_l \in \mathbb{R}^T$是可训练参数。$x^{i,k}$表示第 i 个场站的第 k 个特征向量。经过式(2)计算得到每个特征的注意力系数后，再经过式(3)的 softmax 归一化后即可得出每一个特征的权重系数。最后将权重系数代入到局部特征的输入序列中，则目标场站 i 在 t 时刻的局部特征输入向量如式(4)所示。

$$\tilde{x}_t^{local} = \left(\alpha_t^1 x_t^{i,1}, \alpha_t^2 x_t^{i,2}, ..., \alpha_t^{N_l} x_t^{i,N_l}\right)^T \quad (4)$$

（2）加入基于半正矢公式优化的站间空间注意力机制

目标场站的天气状况往往与临近场站有着一定的相似性，使得目标场站的发电功率与临近场站之间存在空间相关性。半正矢公式是一种根据经纬度计算两点在球体上距离的公式，如式(5)所示。

$$d = 2r \times \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_1)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (5)$$

其中，d 表示两点之间的距离。r 表示地球半径，由于地球是一个近似球体，各处半径不尽相同，难免会有一定的误差存在，本文中采用地球平均半径为 r = 6371km。φ_1、φ_2表示两个点的纬度（以弧度制度量），λ_2、λ_1表示两个点的经度。根据式(5)和所有场站的经纬度数据，求得其余场站与目标场站之间的地理距离后，筛选保留其中 4 个作为临近场站。

影响目标场站的光伏发电功率不光与该场站的因素有关联，在空间范围内，临近场站的一些因素也与之有关。根据这种空间相关性，可以将临近场站的特征向量作为模型的一部分输入，并加入全局空间注意力机制，对于给定目标场站 i，根据式(6)可以计算出不同临近场站 l 的注意力系数。

$$g_t^l = v_g^T \tanh\left(W_g[h_{t-1}; s_{t-1}] + U_g y^l + W'_g X^l u_g + b_g\right) \quad (6)$$

其中，h_{t-1}和s_{t-1}表示上一时刻编码器单元的隐藏状态和细胞状态；$v_g \in \mathbb{R}^T$、$W_g \in \mathbb{R}^{T\times 2m}$、$U_g \in \mathbb{R}^{T\times T}$、$W'_g \in \mathbb{R}^{T\times N_l}$、$u_g \in \mathbb{R}^T$和$b_l \in \mathbb{R}^T$是可训练参数。$y^l$和$X^l$分别临近场站 l 的目标向量和局部特征向量。在这个式子中，将编码器中前一时刻的隐藏状态和细胞状态以及临近场站的属性考虑进去，能够自适应地分配给临近场站以不同的注意力。

此外为了充分利用空间上的相关性，将地理距离数据引入其中，根据式(7)可以计算出临近场站 l 的注意力权重系数。

$$\beta_t^l = \frac{\exp\left((1-\lambda)g_t^l + \lambda P_{i,l}\right)}{\sum_{j=1}^{N_g} \exp\left((1-\lambda)g_t^j + \lambda P_{i,j}\right)} \quad (7)$$

其中，$P_{i,j}$表示场站 i 和 j 之间的地理位置距离系数，如地理距离的倒数；λ是一个权衡超参数。与局部注意力相同地，经过 softmax 函数归一化后，得到每一个临

近场站的注意力权重系数，最后得到全局空间注意力输入特征向量如式(8)所示。

$$\tilde{x}_t^{global} = \left(\beta_t^1 y_t^1, \beta_t^2 y_t^2, ..., \beta_t^{N_g} y_t^{N_g}\right)^T \tag{8}$$

2.3.2 解码器

为了提取时序特征并给予编码器中不同时间步的隐藏状态不同的权重，在解码器中也加入了一层注意力机制，即时间注意力机制。对于解码器中的时间步 t'，可根据式(9)、(10)计算出编码器中每一个时刻 o ∈ [1,T]的时间注意力权重系数。

$$u_{t'}^o = v_d^T \tanh\left(W_d'\left[d_{t'-1}; s_{t'-1}\right] + W_d h_o + b_d\right) \tag{9}$$

$$\gamma_{t'}^o = \frac{\exp\left(u_{t'}^o\right)}{\sum_{j=1}^{T} \exp\left(u_{t'}^j\right)} \tag{10}$$

$$c_{t'} = \sum_{o=1}^{T} \gamma_t^o h_o \tag{11}$$

其中，$d_{t'-1}$和$s_{t'-1}$表示上一时刻解码器单元的隐藏状态和细胞状态；$v_d \in \mathbb{R}^m$、$W_d \in \mathbb{R}^{m\times m}$、$W'_d \in \mathbb{R}^{m\times 2n}$和$b_d \in \mathbb{R}^m$是可训练参数；$h_o$是编码器时刻 o 的隐藏状态。根据式(10)，对于给定的解码器时刻 t'，每个编码器时刻都能计算得出一个注意力权重系数γ_t^o。最后，将编码器每个时刻的注意力权重系数与该时刻的隐藏状态h_o做加权求和，得到解码器时刻 t'的语义向量（即输入向量）$c_{t'}$，如式(11)所示。

2.2.3 额外因素融合模块

由于光伏功率预测具有时空相关性、并对地理和气象等信息具有较强的相关性，设计了各位因素融合模块，用于引入并融合多源数据，将它们按列拼接得到在解码器时刻t'的额外因素向量 $ex_{t'} = (t'; lat; lon; X_{nwp}^T; i_{emb}) \in \mathbb{R}^{N_e}$，其中，$N_e$表示额外因素的属性个数，lat、lon 分别表示经纬度信息，X_{nwp}^T表示数值天气预报若干个属性的一组数据的转置，i_{emb}表示经过 Embedding 的场站 ID。

3 实验与分析

为了评估方法的有效性，本文在公开数据集上对本文方法和一些经典的时间序列预测模型的预测准确性进行了实验和分析。

3.1 数据集

本文选择了两个真实工业数据集，分别是 PVOD 数据集和北方光伏电场数据集。归一化计算方法如公式(16)所示。

$$z = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{16}$$

3.2 评价指标

本文采用平均绝对误差（Mean Absolute Error，MAE）和均方根误差（Root Mean Square Error，RMSE）作为评测指标，通过计算预测值和真实值之间的均方误差，衡量预测准确性。其计算方法分别如公式(17)和公式(18)所示。

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(y_i - \hat{y}_i)| \tag{17}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{18}$$

其中，n 代表样本数量，y_i代表第 i 个时刻的真实值，$\hat{y}_i$代表第 i 个时刻的预测值。MAE 和 RMSE 的值越小代表预测值越接近真实数据分布，即预测的准确度越高。

3.3 参数设置

本文方法涉及的模型参数有训练次数（epoch）、学习率（learning_rate）和批量大小（batch_size）。两个数据集的模型参数设置情况如表 1 所示：

表 1. 超参数设置

超参数	数值
epoch	100
batch_size	16
learning_rate	0.0001

对于本文的所有实验，我们选取数据集的前 80%作为训练集，中间 10%作为验证集，后 10%作为测试集。

3.4 消融实验

为了验证额外因素模块和优化空间注意力机制对预测效果的影响，进行了消融实验：Ours-ne 表示在本文所提方法的基础上去除额外因素模块的模型，GeoMAN+半矢量为加入优化空间注意力机制的模型。

表 2. PVOD 数据集的消融实验 1

方法	MAE	RMSE
GeoMAN	1.738094	2.393243
GeoMAN+半正矢	1.534638	2.181221
Ours-ne	1.378816	1.768793
Ours	1.362557	1.694247

表 3. 北方光伏电场数据集的消融实验 1

方法	MAE	RMSE
GeoMAN	1.765213	2.362321
GeoMAN+半正矢	1.577342	2.423451
Ours-ne	1.312621	1.742345
Ours	1.312345	1.512321

如表 2 所示，在 PVOD 数据集上对比 MAE 指标，Ours 比 GeoMAN 低约 21.61%，比 GeoMAN+半正矢低约 11.21%，比 Ours-ne 低约 1.18%；对比 RMSE 指标，Ours 比 GeoMAN 低约 29.21%，比 GeoMAN+半正矢低约 22.32%，比 Ours-ne 低约 4.21%。

如表 3 所示，在北方光伏电场数据集上对比 MAE 指标，Ours 比 GeoMAN 低约 25.66%，比 GeoMAN+半正矢低约 16.82%，比 Ours-ne 低约 0.02%；在 RMSE 指标上，Ours 比 GeoMAN 低约 35.98%，比 GeoMAN+半正矢低约 37.6%，比 Ours-ne 低约 13.2%。其中对比基础 GeoMAN 方法的提升最高。

可以看出，本文所提出方法是十分有效的。一是加入基于半正矢公式的临近场站选择的空间注意力优化方法增强了模型对临近场站的空间特征提取能力；二加入的融合多源光伏功率预测数据的额外因素融合模块能够增强时空相关性的提取。综上所述，本文所提出的融合多源数据、加入半正矢公式优化的空间注意力机制能够增强模型的预测能力。

4 结论

本文提出一种结合面向光伏功率预测的特征工程方法的 KPH-GeoMAN 光伏发电功率预测方法。该方法首先对原始数据集做数据预处理；然后通过计算各个特征与功率之间的皮尔逊相关系数进行特征选择并通过 K-means++聚类算法将目标数据集分成 3 种不同的天气类型，并基于不同的天气类型进行样本筛选；之后将数据输入到融合多源时空数据和使用半正矢公式选择优化空间注意力的多级注意力模型中；最后在两个数据集中进行训练、测试。实验表明 KPH-GeoMAN 方法能够降低 MAE 指标约 0.02%~21%，降低 RMSE 指标 4%~37%，对比基线模型 MAE 低约 6%~23.4%，RMSE 指标低约 0.7%~15%。

参考文献

[1]DENG Xinyi, AI Xin. Comprehensive benefit assess-ment and incentive mechanism of distributed photovol-taic energy storage system [J]. Power Generation Technology, 2018, 39 (1): 30-36 (in Chinese).

[2]陈嘉铭. 基于深度学习和强化学习的光伏发电功率预测研究 [D]. 广州: 广东工业大学, 2022

[3]阚博文,刘广一,KHODAYAR Mahdi,等 基于图机器学习的分布式光伏发电预测[J].供用电,2019,36(11):20-27.

[4]焦田利，基于时空关系的广域分布式光伏发电群出力预测关键模型研究[D].杭州电子科技大学,2019.

[5]李争,张杰,徐若思,等.基于相似日聚类和 PCC-VMD-SSA-KELM 模型的短期光伏功率预测[J].太阳能学报,2024,45(02):460-468.DOI:10.19912/j.0254-0096.tynxb.2022-1608.

[6]陈君,郭立颖,赵小会,等. 基于MPBiLSTM的短期光伏发电功率预测[J/OL].计算机技术与发展,1-8[2024-08-02].https://doi.org/10.20165/j.cnki.ISSN1673-629X.2024.0204.

[7]Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[R]. Stanford, 2006.

[8]Liang Y, Ke S, Zhang J, et al. Geoman: Multi-level attention networks for geo-sensory time series prediction[C]//IJCAI. 2018, 2018: 3428-3434.

作者简介：刘宏博（2000-05）男，黑龙江省哈尔滨市，北方工业大学,硕士研究生在读，研究方向人工智能、云计算。

基金项目：北京市自然科学基金项目“面向时空流式大数据融合型计算的平台服务及保障方法研究”（No.4192020）资助.