

基于历史数据和深度神经网络的光伏输出功率建模

摘要:

光伏输出功率的随机性和间歇性给电网带来了许多不确定因素，并且当大量光伏电站接入电网时，可能会导致电网的不稳定。为保证电网的稳定运行，对各种光伏电站进行短期和中期的功率预测是必不可少的。本文利用深度神经网络技术在光伏日输出功率建模和预测中的应用，首先采集光伏电站的历史日输出功率，对数据进行预处理和统计分析。利用 k-means 聚类分析确定了四种天气类型，采用深度信念网络 (DBN) 和长短期记忆 (LSTM) 对光伏电站日输出功率进行建模，基于温度的相似日选择算法生成不同天气类型的神经网络初始顺序输入。计算所得 DBN 和 LSTM 预测模型的平均绝对相对误差 (MARE) 和均方根误差 (RMS)，将 DBN 和 LSTM 模型的预测误差与传统 ARIMA 预测模型的预测误差进行比较。实验结果表明，基于 DBN 的光伏电力预测模型在三种模型中实现了最佳预测精度。

关键词:

PV power, Sequence Modeling(序列建模), LSTM, DBN

1.介绍

当前，能源消耗与环境保护之间的争论越来越严重。为了缓解这种情况，许多国家推广了风能，太阳能，核能等可再生能源。目前，光伏 (PV) 功率是利用太阳能的最重要方法之一[1]。但光伏电站的输出功率受昼夜和气候变化的影响，在一定程度上具有随机性和间歇性[2-3]。当大量光伏电站接入电网时，发电量的波动性和不确定性会给电网带来严重的问题[4]。因此，重要的是事先预测光伏站产生的功率输出顺序，以便可以启动适当的功率调度或自动功率调节以降低电网上的风险

近年来，国内外研究了多种方法进行短期或长期光伏发电预测。这些方法可以分为两类:物理方法和数据驱动方法。

物理规律。首先建立太阳辐射强度模型。基于太阳辐射强度模型和光伏电站物理模型，建立了电力预测模型。Wang 等人提出了两种天气类型的辐射预测模型，以实现了 PV 预测模型[6]。Perpiñan 等人根据光伏板的原理和公式，采用了辐照度域积分和统计动量的定义[7]。物理方法需要精确而详细的气象信息，而获取这些信息的成本很高。此外，物理方法很难深入了解相关光伏电站的物理和结构。

数据驱动方法 (Data-driven methods) 以光伏电站的历史数据为训练数据，建立数学预测模型。现在已经有许多线性/非线性数学模型，如用于预测光伏输出的 ARMA, ARIMA, 马尔可夫链, 人工神经网络 (ANN) 和支持向量机 (SVM)

[8-10]。Liu 等人提出了一种结合 EM 算法，小波变换和 SVM 回归预测 PV 功率输出的方法[11]。Shi 等人将测试期 PV 功率输出数据分为四组，根据数据样本建立了四个 SVM 模型[12]。Wang 等人[13]提出了一种结合灰色模型和反向传播（BP）神经网络模型的方法。Ding 等人采用改进的 BP 学习算法和类似的日选择算法进行预测，以克服常规 BP 学习算法的缺点[14]。

为了预测输出，Ramsami 等人初步设计了三个单阶段模型，即广义回归神经网络，前馈神经网络和多元线性回归(Generalized Regression Neural Network, Feed-forward Neural Network and Multiple Linear Regression [15])[15]。然后，利用逐步回归的方法将输入变量输入到单阶段模型中，得到三个混合模型。这些研究证明了仅基于光伏输出数据建立光伏预测模型的可行性。然而，BP 神经网络和 SVM 回归(BP Neural Network and SVM Regression)都是浅层神经网络。用这些网络表示复杂的非线性输入输出函数有一定的局限性。

一些研究人员也尝试了使用深度学习的方法来预测光伏功率。Li et al [16]从定性和定量的角度分析了气象因子(太阳辐射强度、温度、相对湿度和风速)与光伏输出功率之间的关系，然后利用 DBN 建立基于历史数据和深度神经网络的光伏输出功率模型。据报道，该模型具有较高的预测精度。但该方法所需的气象信息过于详细，难以获得。Shi 等人 [17]也采用 DBN 进行光伏发电预测。该方法在晴天或雨天等天气条件稳定的情况下，预测精度较好，但在其他天气条件下，预测精度较差。

以往的研究表明，DBN 在从数据中提取隐含复杂特征（implicated complex features）方面具有很大的潜力。另一方面，长短期记忆(LSTM)神经网络在时间序列建模方面具有优势。基于这些考虑，本文研究了 DBN 和 LSTM 神经网络用于 PV 输出的序列建模，并对这两种神经网络结构在 PV 输出预测方面进行了比较。

2. DEEP NEURAL NETWORKS

2.1 Deep Belief Network

深度信念网络(DBN)是一种由受限玻尔兹曼机(RBM)堆叠而成的神经网络[18]。在顶层增加 Logistic 回归函数或 BP 神经网络，以便可以使用该组合来进行预测。RBM 是一种基于统计热力学的能量模型，它由两层神经元组成，即可视层和隐藏层[19]。一组可视节点和隐藏节点的概率由能量函数定义。对于 RBM，其能量函数可表示为：

$$E(v, h) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} h_i v_j - \sum_{j=1}^m b_j v_j - \sum_{i=1}^n c_i h_i \quad (1)$$

其中 h_i 表示第 i 个隐藏层单元， v_j 表示第 j 个可见层单元， w_{ij} 表示隐藏层和可见层之间的权重， c_i 表示隐藏层的偏置， b_j 表示可见层的偏置。对于 RBM(v, h)的特定状态，其可视节点和隐藏节点的联合概率可以写成：

$$p(v, h) = \frac{e^{-E(v, h)}}{Z} = \frac{e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}} \quad (2)$$

其中 Z 是归一化因子。基于上述联合概率， v 与 h 是特定 Gibbs 分布的概率分布可以表示为：

$$p(v) = \frac{\sum_h e^{-E(v, h)}}{Z} = \frac{\sum_h e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}} \quad (3)$$

$$p(h) = \frac{\sum_v e^{-E(v, h)}}{Z} = \frac{\sum_v e^{-E(v, h)}}{\sum_{v, h} e^{-E(v, h)}} \quad (4)$$

隐藏层单元的打开或关闭基于权重 W 和可见层的神经元值。假设可见层中所有神经元的状态都是已知的。第 J 个隐藏层单元的激活概率可以表示如下：

$$p(h_j = 1/v) = \text{sigmoid} \left(b_j + \sum_i v_i w_{ij} \right) \quad (5)$$

类似地，给定隐藏层的神经元值，第 i 个可见层单元的激活概率可以表示如下：

$$p(v_i = 1/h) = \text{sigmoid} \left(c_i + \sum_j h_j w_{ij} \right) \quad (6)$$

似然函数等于

$$L(\theta) = \prod_v L(v/\theta) = \prod_v p(v) \quad (7)$$

其中 $\theta = (W, b, c)$ 代表网络参数， v 代表训练样本集。假设有 n 个样本，训练数据可以用公式 (8) 的形式编写。

$$X = \{x_1, x_2, \dots, x_n\} = \{v^1, v^2, \dots, v^n\} \quad (8)$$

以最大似然函数为最优函数，通过 RBM 训练过程可以得到一组最优参数 θ 。为了降低计算复杂度，似然函数通常转换为对数形式：

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_v \frac{\partial \ln L(v/\theta)}{\partial \theta} = \sum_v \frac{\partial \ln p(v)}{\partial \theta} \quad (9)$$

对于每个 v^l

$$\frac{\partial \ln p(v^l)}{\partial \theta} = \sum_h p(h/v) \left(-\frac{\partial E(v, h)}{\partial \theta} \right) - \sum_{v, h} p(v, h) \left(-\frac{\partial E(v, h)}{\partial \theta} \right) \quad (10)$$

公式（10）的第一项等于样本本身分布下输入样本数据的自由能函数的期望值。该公式的第二项是在 RBM 网络表示的吉布斯分布下, 自由能函数的偏导数的期望。很难计算第二项, 因为需要遍历 V 和 H 的所有可能值。对比来看, 对比散度 (CD) 算法常用于获得近似值[20]。

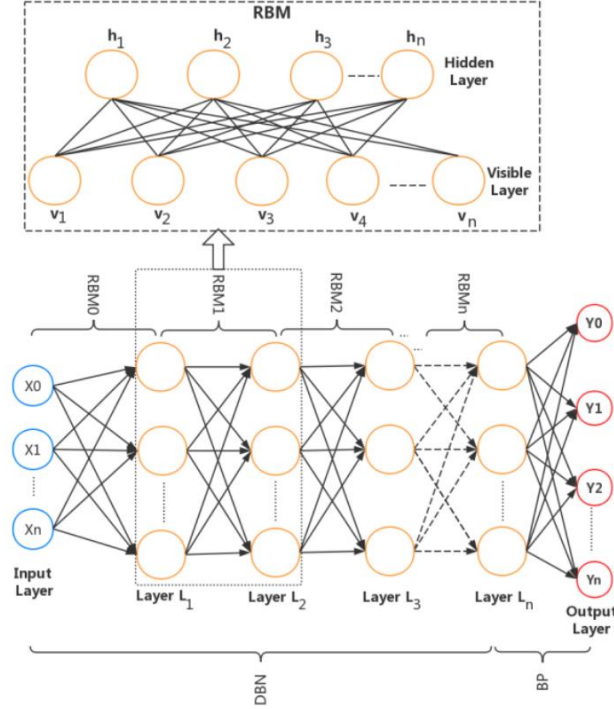


图 1. 基于 RBM 的深度信念网络

2.2 LSTM

递归神经网络 (RNN) 是一种用于处理序列数据的神经网络[21]。与其他神经网络相比, RNN 内各隐层单元是相互关联的。同时, 隐藏层单元的输入与该隐藏层之前时刻的输出有关。这使得 RNN 不仅可以提取当前输入的特性, 而且还依赖于先前的输入信息, 从而实现了序列信息的保存。因此, RNN 在处理序列数据方面具有独特的优势。结构图如图 2 所示。

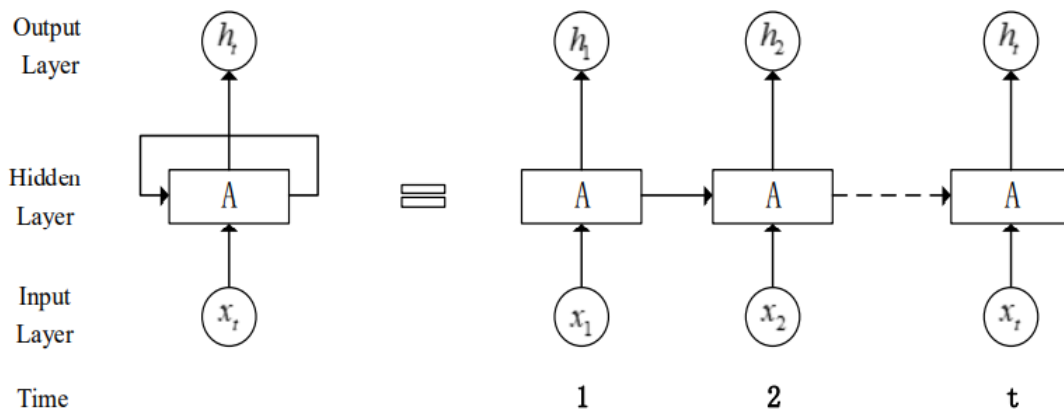


图 2. 递归神经网络结构

其中, $X = \{X_0, X_1, \dots, X_t\}$ 是输入序列, $X_t \in R^k$ 为在该输入序列在第 t 个时间步的 k 维向量, H 为 RNN 在每一个隐藏层在当前时刻的状态。

虽然 RNN 具有记忆能力, 但它的记忆周期很短, 所以它只能记忆短时期的上下文信息。而当序列的信息随着学习量的增多或者学习周期的增长时, 会出现梯度消失的问题。

长期短期记忆 (LSTM) 是一种旨在解决递归神经网络长期依赖问题的结构。LSTM 会在整个生命周期中定义和维护一个内部存储器单元状态[22]。内部存储器单元结构如图 3 所示。存储单元的状态与中间输出和后续输入有关。根据前一阶段的输出和当前阶段的输入, 决定更新、维护或删除哪些内部状态向量元素。除了内部状态之外, LSTM 还引入了三个门单元: 输入门, 遗忘门和输出门。LSTM 通过这三个门来控制信息的舍弃与更新。

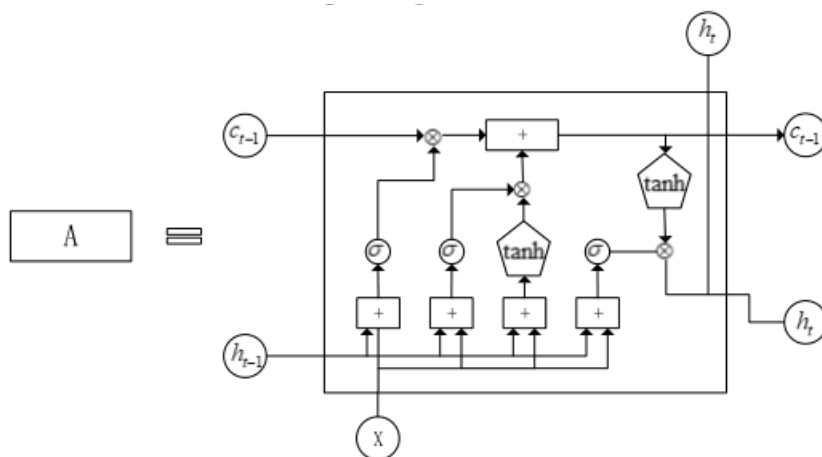


图 3. LSTM 内部存储器单元的结构

每个门的功能如下所示。

1) 遗忘门

遗忘门是一个 sigmoid 函数，它取前一个单元的输出 h_{t-1} ，并将 x_t 作为输入。它为每个单位在 $[0, 1]$ 中生成一个值，并控制上一个单元的遗忘程度。

$$f_t = \sigma(W_f^*[h_{t-1}, x_t] + b_f) \quad (11)$$

2) 输入门

输入门配合 tanh 函数控制加入哪些新信息。该函数生成一个新的候选 \tilde{C}_t ，输入门为每个 \tilde{C}_t 都生成一个值，这个值的取值范围为 $[0, 1]$ ，用来控制添加了多少新信息。

$$i_t = \sigma(W_i^*[h_{t-1}, x_t] + b_i) \quad (12)$$

$$c_t = \tanh(W_c^*[h_{t-1}, x_t] + b_c) \quad (13)$$

3) 输出门

输出门控制当前存储单元状态被过滤的程度。首先，激活单元的状态，然后输出门为每个项目生成 $[0, 1]$ 中的值。因此，过滤控制单元处于被过滤的状态。

$$o_t = \sigma(W_o^*[h_{t-1}, x_t] + b_o) \quad (14)$$

$$e_t = o_t^* \tanh(c_t) \quad (15)$$

其中 W_f ， W_i ， W_c 和 W_o 是权重矩阵； b_f ， b_i ， b_c 和 b_o 为偏置；符号 * 代表矩阵元素相乘； σ 是 sigmoid 函数； \tanh 代表双曲正切函数。

3. 数据聚类 and 天气分类

3.1 数据来源

采用位于华北城市保定的光伏电站作为测试案例。该光伏站的装机容量为 400kW。我们获得了在 2014 年内光伏电站的完整输出数据集。数据集中输出功率的采样周期为 1 分钟。此外，我们还从国家气象信息中心官方网站了解到了 2014 年保定市每日的天气报告。天气信息包括日平均气温，最高和最低气温，降水量，相对湿度，日照时间，平均风速，最大风速和天气类型。

3.2 数据预处理

光伏功率预测的研究基于历史功率数据。由于各种原因，收集的初始数据具有一些奇异点和缺失值。首先，对原始数据进行预处理以找到奇异值，并用插值法表示这些奇异值。实际的光伏发电数据每 1 分钟采样一次。每天测量的开始时间和结束时间并不完全一致。为方便起见，我们选择一个相同的时间窗口（8:00-17:15）。另外，PV 功率输出数据平均每 15 分钟收集一次，因此日序列更短。最后，将所有数据（包括光伏电力数据和气象数据）进行归一化处理，以减少尺度对预测模型的影响。

3.3 数据分析和聚类

光伏功率的输出受两组因素影响。一组因素与光伏站的物理条件有关，如光伏站规模，光伏电池质量，电路连接方式等；另一组因素与地理和气象环境有关。一旦光伏站建成并投入运行，发电站的物理条件或多或少是固定的。光伏站所在地的地理因素也相对稳定，或者说是具有一定的固定模式。但是太阳辐射强度，气温，空气湿度，风速，云层情况等气象因素每天都在变化。图 4 分别显示了在晴天，阴天和雨天光伏站的日输出功率。可以看出，雨天的电力输出与晴天的电力输出有很大的不同。

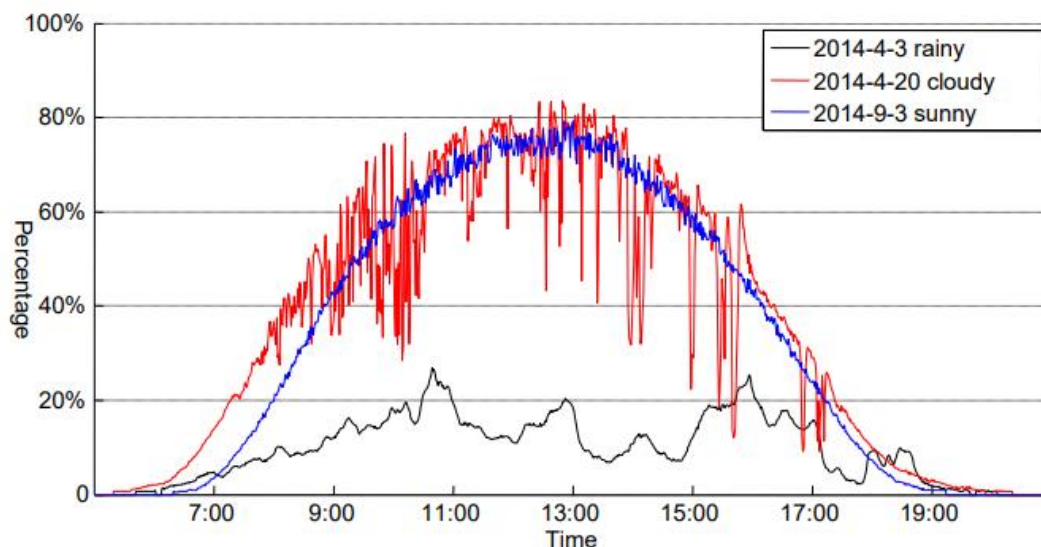


图 4. 不同天气下的光伏站的日输出功率

从数据集中选取三种类型（具体来说是晴天，雨天和多云）的光伏站的历史功率输出，如图 5 所示。对于晴天和多云，日产量序列的整体形状大致相同，但大小和曲率的数量级区别较大。同时，雨天的日产量顺序非常不规则。因此，根据天气类型建立不同的光伏输出预报模型非常重要。

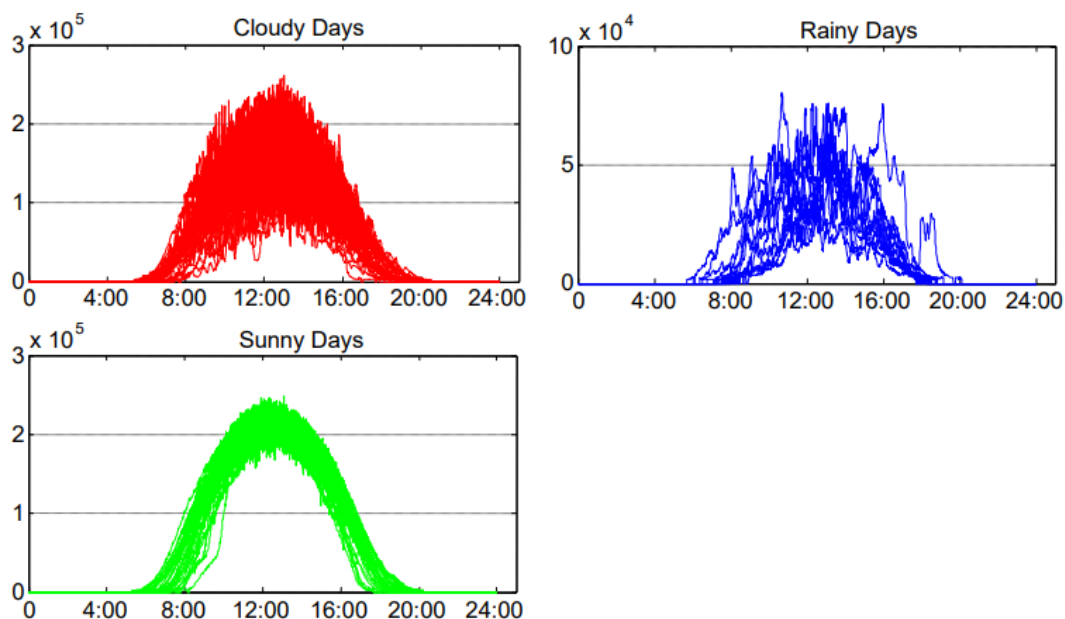


图 5. 三种天气类型下光伏输出的日分布

从国家气象局可以简单的获取一个地区的气象信息。本文提供的气象资料来源于 2014 年保定市每日天气报告。天气报告中约有 10 种天气描述：毛毛雨，小雪，多云，阳光，阵雨，中雨，阴天，雾霾，多云和晴天。如果我们直接使用每日天气描述来对日期进行分类，则将有超过 10 种类型的天气，并且一些天气类型（例如淋浴）在数据集中只有少量样本。这将会对深度神经网络的建模精度带来负面影响。

为了降低建模的复杂性并为每个模型收集更多样本，我们使用 K 均值法对 PV 输出数据进行聚类。我们设置 $k=4$ ，聚类结果如图 6 所示。彩色曲线是指定天气类型下全天 PV 功率输出平均值的曲线图，聚集成相同天气类型的曲线标有相同的标签，四个簇的中心线在图中以黑色虚线绘制。

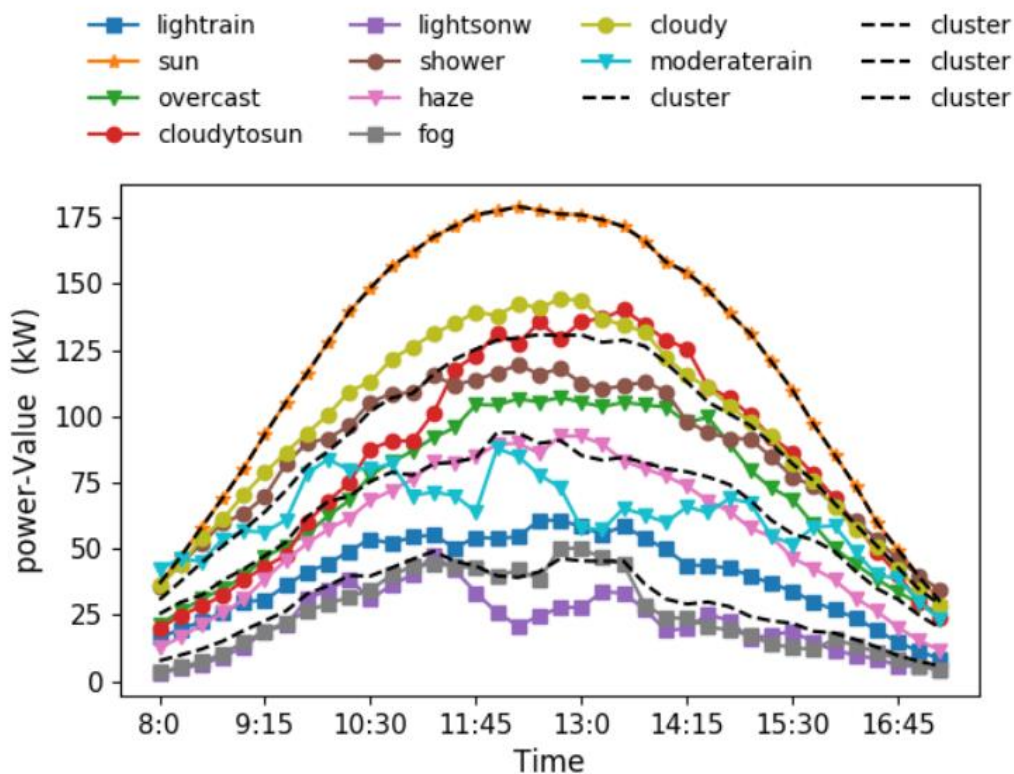


图 6. 不同天气下光伏输出的特点

每个集群的天数如图 7 所示。晴天光伏发电的历史资料属于 Type1，多云阵雨天气光伏发电的历史资料属于 Type2，多云雾霾和阵雨天气光伏发电的历史资料属于 Type3，小雨，雾雪天气光伏发电的历史资料属于 Type4。丢失没有数据的天数，2014 年每种类型的天数分别为 96 天，127 天，45 天和 16 天。值得注意的是，属于 Type4 的日子很少，因为这个光伏站位于中国北部，全年降水量很少。

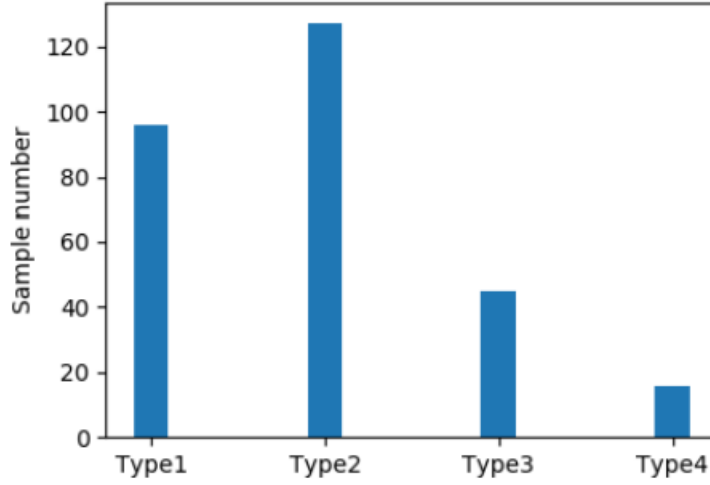


图 7. 每种天气类型的天数

4. 基于 DBN 的预测模型

4.1 相似日的选择

在对历史功率输出进行数据聚类之后，样本数据在时间维度上不再连续。为了选取合适的模型输入，并且为了进一步提高模型预测的准确性，本文提出了一种基于气温的相似日选取算法。

虽然太阳辐射强度在光伏发电功率的输出中起着关键作用[24]，但很难提前获得太阳辐射强度数据。除太阳辐射强度外，温度对光伏发电也起着重要作用。温度以复杂的方式影响光伏电池的转换效率。太阳辐射强度与温度有关，同时温度的升高可能影响电路性能，降低光伏电池的转换效率。因此，温度信息对于光伏输出建模非常重要。此外，预测日的气温信息可以轻易的从相关测量站点获取。因此，气象报告中每天的最高温度，最低温度和平均温度被用作类似日期的选择标准。

类似的日期选择算法如下所示：

1) 计算预报日的温度与动力输出数据集中与预报日的天气类型相对应的历史日温度之间的欧几里得距离。

$$d_i = \sum_{j=1}^3 (Y_j - X_{ij})^2, i = 1, 2 \dots n \quad (16)$$

其中 Y_1, Y_2, Y_3 分别是预报日的最高温度，最低温度和平均温度， $X_{i1}X_{i2}X_{i3}$ 为与预测日天气类型相同的光伏发电功率历史数据集合中的第 i 个数据样本， n 为该集合所有样本的数目。

2) 找到具有最小值 $D = \{d_1, d_2, \dots, d_n\}$ 的数据，其对应的光伏发电功率样本为预测日的相似日光伏发电功率数据。

4.2 模型评估标准

本文采用平均绝对相对误差（MARE）和均方根误差（RMSE）用于评估预测准确性。

$$MARE = \frac{1}{N} \sum_{t=1}^N \frac{|W_{forecasting} - W_{true}|}{W_{total}} * 100\% \quad (17)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^N (W_{forecasting} - W_{true})^2}{N}} \quad (18)$$

其中 $W_{forecasting}$ 是每个采样时间的预测结果， W_{true} 是预测日的真实值， W_{total} 是光伏站的装机容量， t 是样本的时间指数， N 是每个记录的样本数。

4.3. 网络 DBN 的结构和参数

1) 输入层

我们获得的光伏站的输出功率为上午 8 点至下午 5 点，间隔 15 分钟，因此每天有 38 个数据点。预测日的气象数据可以根据当地气象站的气象预测数据获得，预测模型输入($x_1 - x_{46}$)如表 1 所示。

表 1. 输入层

输入	描述
$x_1 - x_{38}$	当天关于预报日的光伏功率历史数据，时间范围 8:00-17:15，间隔 15 分钟，38 点
$x_{39} - x_{46}$	预报日的气象因子包括平均气温，最高气温和最低气温，降水量，相对湿度，日照时间，平均风速，最大风速

2) 输出层

本文的研究目标为光伏站的光伏发电功率短期预测，模型的输出为预测日 8:00-17:15 时间段内每隔 15 分钟的光伏发电功率数据，共计 38 个点。预测日输出层 $y_1 - y_{38}$ 如表 2 所示。

表 2. 输出层

输出	描述
$y_1 - y_{38}$	PV 功率基于时间的 sery，时间范围 8:00-17:15，间隔 15 分钟，总共 38 个离散值

3) 数据集

数据预处理和聚类后，构建数据集的步骤如下所示：

步骤 1：4 种分类类型下预处理后的光伏发电功率历史数据作为标签，形成 4 组标签数据集。

步骤 2：对于 4 个标签数据集内的每个元素，在其所属的标签集内使用相似日算法求出其对应的相似日的光伏发电功率数据序列，同时提取该元素对应的天气信息，将该元素的相似日的光伏发电功率数据序列与天气信息相结合，共同作为该标签元素对应的训练输入。

步骤 3：将输入集合和标签集合的前 70%划分为训练集，剩余的 10%和 20%分别划分为测试集和验证集。

4) 超参数

对于深度信念网络，模型的训练参数包括隐藏层的数目，每个隐藏层中的神经元（节点）数和训练迭代次数。这些参数影响预测模型的收敛程度，训练时间和复杂性，这些对预测模型的预测精度都至关重要。在本文中，使用纵向对比法来进行超参数的选择（即通过在保持其他参数的固定的情况下，改变一个参数，来测试该参数的改变所带来的影响）。4 个测试集的平均 MARE 和预测模型的训练时间用作评估标准。比较结果如表 3 和表 4 所示。

表 3. 输入层

层数					训练时间
	测试集 1	测试集 2	测试集 3	测试集 4	
2	4.84	8.92	5.40	4.05	28.56
3	4.88	9.01	5.41	4.90	32.82
4	4.98	9.44	5.22	4.96	33.30
5	8.15	9.48	5.17	5.33	35.61

表 3 显示了更改隐藏层数对预测准确性和训练时间的影响。在测试期间，每个隐藏层中的神经元数量设置为 128 个。我们可以看到，除了测试集 3 之外，平均 MARE 和训练时间都随着层数的增加而变大。对于测试集 3，平均 MARE 随着层数的增加而缓慢下降。这意味着天气类型 3 下的光伏输出数据具有更复杂的内部特征。这种推断是有意义的，这是因为虽然增加层数会增强神经网络的拟合能力，但是当层数过多时，模型会过于复杂而导致对训练集的过拟合。因此，本文中深度信念网络的层数选定为 2 层。

表 4. 隐藏层节点

节点					训练时间
	测试集 1	测试集 2	测试集 3	测试集 4	
64	4.96	9.16	5.16	3.58	25.87
128	4.82	9.02	5.31	4.00	28.45
256	4.77	8.96	5.18	2.26	28.88
512	4.99	9.34	5.10	2.46	36.93

表 4 显示了隐藏节点数量对预测的影响。在此过程中，隐藏层的数量设置为 2 层。当节点数为 256 时，4 个测试集的 MARE 达到良好的结果。因此，深度信念网络中的节点数量为 256 个。

5. 基于 LSTM 的预测模型

LSTM 的输入通常是时间序列数据，光伏站的光伏输出每 15 分钟采样一次，这些输出值都是按时间顺序排列。我们没有整日的时间气象数据，所以，基于 LSTM 的发电功率预测模型的输入只是预测日对应的相似日的光伏发电功率序列，不再包含预测日的气象数据。基于 LSTM 的光伏预测模型的输入功能如表 V 所示。

表 5. 输入层

输入	描述
$x_1 - x_{38}$	预测日对应的相似日的光伏发电功率历史数据，时间区间为 8:00-17:15，时间间隔 15 分钟，共计 38 个点

在 LSTM 模型中，超参数包括隐藏层的节点数，学习率和迭代次数。这里，LSTM 模型的节点数是 256。训练时间设置为 1000，学习率为 0.06。

6. 评估结果

经过训练和评估，获得了三种预报方法，即 DBN 模型，LSTM 模型和 ARIMA 模型，以及四种天气类型的预测精度。ARIMA 模型是时间序列建模的传统统计模型，样本数据是具有相同天气类型且在预报日之前的 8 天的 PV 输出序列。基于 Akaike 信息准则 (AIC)，使用 Python 内部函数 (auto_arima) 自动获得最佳 ARIMA 模型。四种天气类型的三种模型的预报结果如图 9-12 所示。所有预报样本的平均 MARE 和 RMSE 如表 6 所示。可以看出，对于所有天气类型，DBN 模型比传统 ARIMA 模型具有更好的性能。在大多数样本日，LSTM 模型的性能优于 ARIMA 模型。

在天气类型 1 下，光伏功率峰值明显，变化趋势相对温和。功率随时间增加，中午达到峰值，然后逐渐下降。DBN 模型的预测值与模型的实际值之间的差异很小。LSTM 预测模型在 8 个预测日中的 5 天表现非常好，而在其余 3 天内表现很差。ARIMA 模型的预测结果则落后于实际值。天气类型 1 的 DBN 预测模型的平均 MARE 和 RMSE 分别为 7.7%和 35.7 kW，低于 LSTM 模型和 ARIMA 模型。

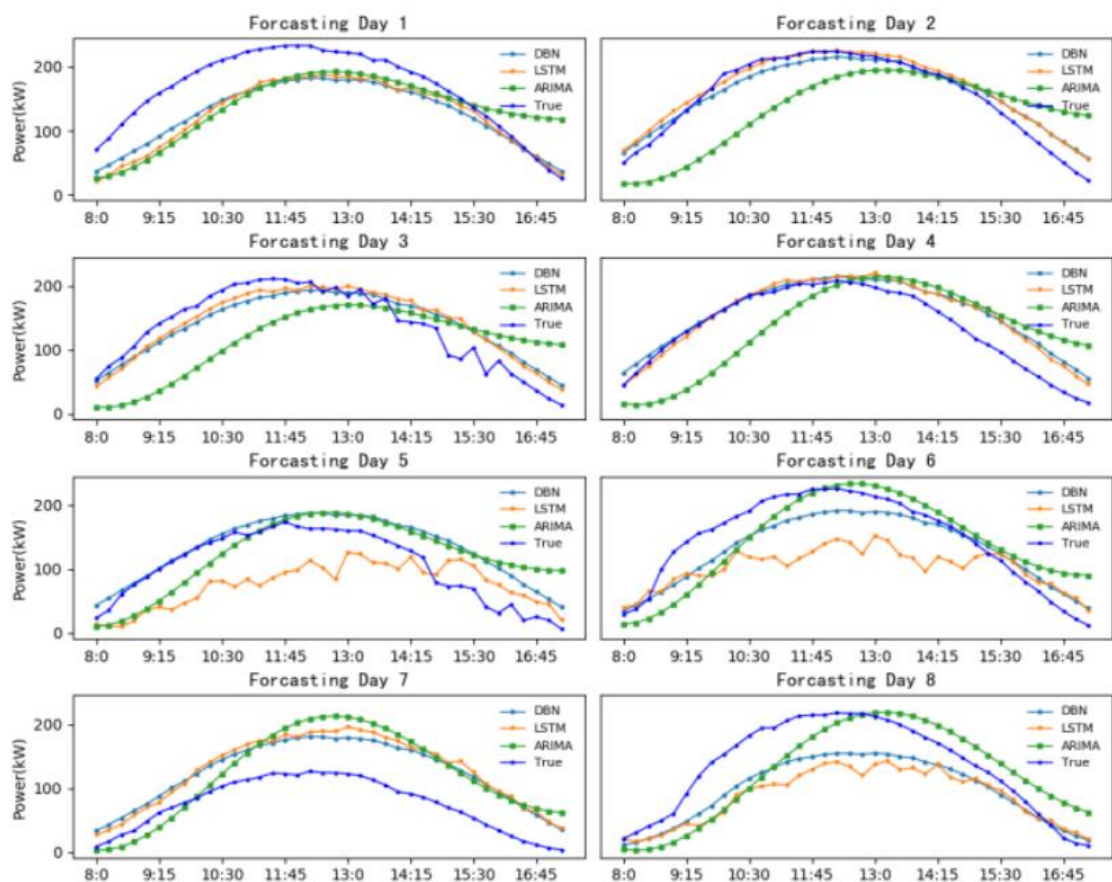


图 9. 基于三种晴天方法的预报比较

2 型天气下的光伏发电峰值也很明显，但不如 1 型天气明显。在样本中，电力的整体变化趋势相对温和，但在日曲线的某个点存在一定的突变峰谷。这是因为 2 型天气下的三种特定天气类型（多云，阵雨和多云转晴）比 1 型天气（晴天）更复杂。例如，当大云覆盖太阳时，它们对太阳辐射强度产生巨大影响，直接影响光伏功率在短时间内的输出值。这种不可预测的情况使其更难以预测。在所有预测样本中，DBN 模型的预测结果都可以很好地跟踪预测日的实际权力趋势。除某些点外，预测结果与实际值之间的误差很小。

表 6. 4 种天气类型下 DBN，LSTM 和 ARMA 预报的平均 MARE 和 RMSE

模型	平均误差	天气类型 1	天气类型 2	天气类型 3	天气类型 4
DBN	MARE (%)	7.7	6.8	7.6	3.3
	RMSE (kW)	35.7	32.2	33.7	17.3
LSTM	MARE (%)	9.3	8.5	6.5	8.9
	RMSE (kW)	43.2	41.4	32.4	42.7
ARME	MARE (%)	11.2	11.3	9.3	4.7
	RMSE (kW)	52.6	38.7	39.6	20.6

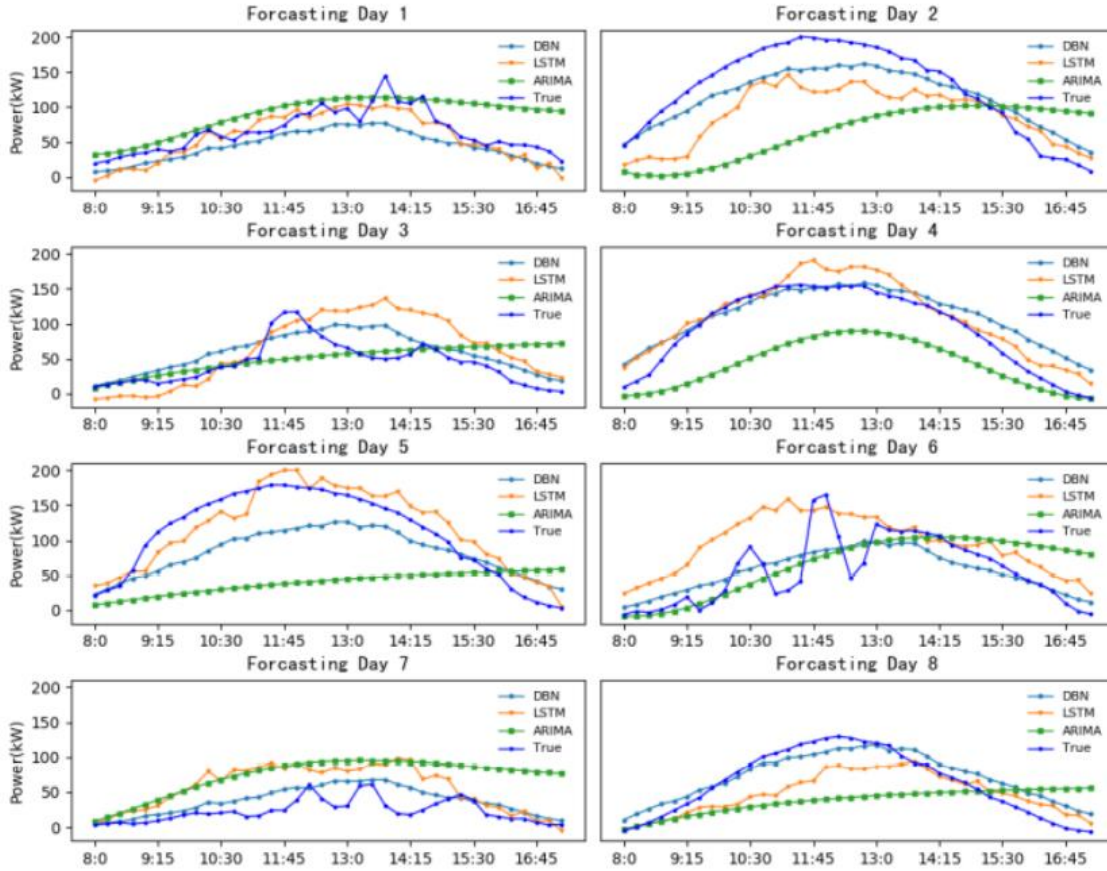


图 10: 基于三种阴天方法的预报比较

3 型天气和 4 型天气下的光伏发电峰值不如 1 型和 2 型天气明显。在天气类型 3 下, 一些样本的总体趋势相对温和, 但在天气类型 4 下, 功率值的变化是不规则的, 几乎每个点的功率值都有突然的变化。总体而言, DBN 预报模型在天气类型 3 和 4 下表现相对较好, 但预测结果与某些样本的实际值之间存在较大差异, 如天气类型 3 下的预报样本 2 和天气类型 4 下的预报样本 4。出现这种情况的原因是, 一方面, 两个预报日的力量波动很大, 3 型天气和 4 型天气的具体天气条件比其他天气类型更复杂, 这使得预报更加困难, 天气类型 3 和天气类型 4 下的历史功率数据训练样本少于其他天气类型的训练样本。此外, 本文使用的气象数据是针对保定市而不是光伏电站, 因此预测的准确性略有下降。

对于天气类型 3, LSTM 预报模型的平均 MARE 和 RMSE 分别为 6.5% 和 32.4kW, 低于 DBN 预测模型。这反映了 LSTM 模型在天气类型 3 下的预测效果略好于 DBN 模型。但 4 型天气下 LSTM 模型的预报结果与预报日的实际产出值差异很大。预测日的最大 RMSE 值为 63.6 kW, 有些预测值甚至为负。这反映了 4 型天气的 LSTM 预报模型的失败。在这种情况下, 除了前面分析中提到的原因, 即 4 型天气下的复杂天气类型外, 另一个原因是对训练样本数量的

要求高于 DBN。此外，虽然 LSTM 预测模型在天气类型 1, 2 和 3 下的验证集上取得了良好的预测结果，但大多数预测样本的 MARE 略高于 DBN 预测模型的相应预测结果，这也揭示了气象因素对光伏输出有重要影响的事实。

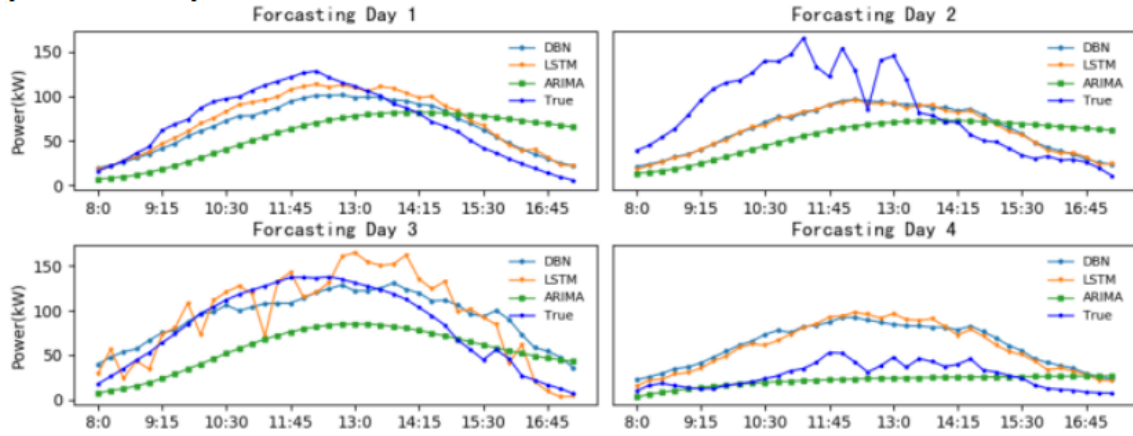


图 8: 基于三种方法的第三类天数预测比较

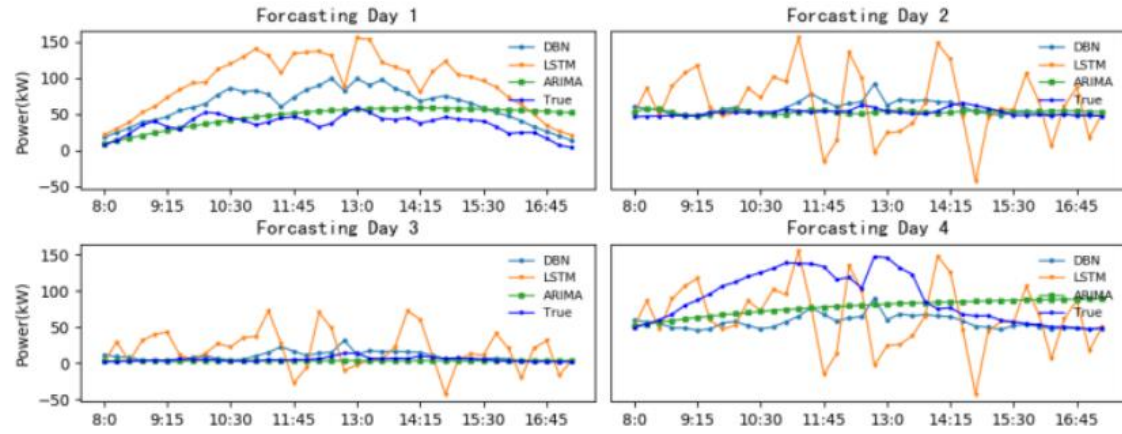


图 9: 基于三种雨天方法的预报比较

7. 结论

提前一天预测光伏发电对于电网的电力调节和日常电力调度非常重要。本文研究了深度信念网络在光伏站功率输出序列建模和日常光伏输出预测中的应用。基于历史数据，建立了三种光伏输出预测模型，包括基于 DBN 的预测模型，基于 LSTM 的预测模型和基于 ARIMA 的预测模型。本文比较了三种模型的预测精度。

由于天气在光伏站输出中起着关键作用，因此对光伏站历史值的日输出序列进行了深入研究，并采用 K 均值聚类算法找出不同的天气类型。基于 k 均值聚类算法，从每日天气报告中发现的 10 种天气描述中最终确定了 4 种天气类型。然后，分别对选出的四种天气，使用所提及的光伏站的历史日常序列，构建和训练基于 DBN 的预报模型以及基于 LSTM 的预报模型。基于 DBN 和 LSTM 的时间序列模型的初始输入值设置为与使用基于温度的相似日算法在历

史数据中找到的每日输出数据相同。该策略解决了集群后光伏电力历史数据在时间维度上不再连续的问题，预测日前一天的电力数据不能作为传统序列建模所需的预测模型输入。

本文提出的案例研究证明了深度神经网络对光伏输出序列建模的有效性。尽管用于训练预测模型的历史数据不大，但基于 DBN 和基于 LSTM 的神经网络都比 ARIMA 模型具有更好的准确性。DBN 预测模型的输入由气象信息以及光伏站历史相似日的日序列组成。另一方面，基于 LSTM 的预测模型仅将历史相似日的日序列作为输入，这可能是基于 DBN 的光伏预报模型在所有四种天气类型中获得比基于 LSTM 的预报模型更好结果的原因。为了提高 LSTM 模型的准确性，一种解决方案是在输入中包含详细的每日温度信息。从这项研究中，只要光伏站可以提前获取足够的历史数据和某些天气信息，我们相信 DBN 和 LSTM 网络都适用于光伏输出预测。

感谢

作者要感谢李建林博士和中国电力研究院（CEPRI）为本研究提供保定光伏发电站的现场数据。这项研究得到了中国国家重点研发计划的部分支持