北京交通大学

硕士专业学位论文

基于 Informer 的中长期风电功率预测研究

Research on medium and long term wind power prediction based on Informer

作者：王文贵

导师：黄华

北京交通大学

2022 年 9 月

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

基于 Informer 的中长期风电功率预测研究

Research on medium and long term wind power prediction based on Informer

作者姓名：王文贵 学　　号：20140081

导师姓名：黄华 职　　称：副教授

专业领域：软件工程 学位级别：硕士

北京交通大学

2022 年 9 月

摘要

时下能源的变革是社会的热点话题，风力发电在新能源发电行业处于至关重要的地位。中长期风电功率预测可以为风电场的生产运行、计划检修等工作提供重要的数据支撑，提高风电并网安全性，保障电力系统的安全稳定运行。

风能的不稳定性、随机波动性等特点使得风力发电具有不确定性。尽管风电机组功率历史数据庞大，但因其存在一定的异常值和缺失，造成风电功率预测的困难。尽管目前短期风功率预测的研究已相对成熟，但中长期的风电功率预测仍然存在较大的挑战。为此，本文探索了利用深度学习模型进行中长期风电功率预测的方法，提出了一种基于期望最大算法的风电功率缺失值填充方法，考虑风力发电机组时空关联关系，构建了基于 Informer 和图卷积的组合模型风功率预测方法，并设计开发了风功率预测可视化平台。具体研究内容如下：

（1）针对风功率数据缺失问题，提出了一种基于工程先验知识和期望最大算法的风功率缺失值填充方法。利用风功率工程先验设置将期望最大算法初始化参数控制在合理范围内，提高算法的收敛速度，降低了缺失值填充的误差。最后与回归填充、随机森林填充、多重插补填充进行实验对比，验证了所提方法的有效性。

（2）提出了基于 Informer 和图卷积的组合模型风功率预测算法。通过 Informer 模型捕获风机功率长序列数据间的依赖关系，利用图卷积捕获风机功率的空间依赖关系，最后通过方差-协方差的组合预测模型进行时空关系的风功率预测。与基准模型在某风电场真实数据集上开展的实验结果表明，本文提出的模型预测效果更佳。

（3）针对风电功率相关数据分散、数据查询繁琐，缺少风功率中长期预测功能等问题，基于上述研究成果使用 Vue 和 SpringBoot 设计开发了一个风功率中长期预测平台。系统可展示风机功率相关历史数据，并集成了上述中长期风功率预测模型和其它基准预测模型算法，具有预测结果可视化展示功能，为风电场运营人员提供了一个专业的可视化平台。

关键词：风功率预测；缺失值填充；时空预测模型；中长期预测

ABSTRACT

Nowadays, the transformation of energy is a hot topic in society, and wind power plays a vital role in the new energy power generation industry. Medium- and long-term wind power forecasting can provide important data support for the production and operation of wind farms, planned maintenance and other work, improve the safety of wind power grid connection, and ensure the safe, stable operation of the power system.

The instability and random fluctuation of wind energy make wind power generation uncertain. Although the historical data of wind turbine power is huge, it is difficult to forecast wind power due to the existence of certain outliers and missing values. Although the research on short-term wind power prediction is relatively mature, there are still great challenges in medium and long-term wind power prediction. To this end, this paper explores a method of using deep learning models for mid- and long-term wind power prediction, and proposes a method for filling missing values of wind power based on the Expectation-Maximum algorithm. A combined model wind power prediction method based on Informer and graph convolution is proposed, and a wind power prediction visualization platform is designed and developed. The specific research contents are as follows:

(1) Aiming at the problem of missing wind power data, a method for filling missing values of wind power based on engineering prior knowledge and the expectation-maximization algorithm is proposed. The EM initialization parameters are controlled within a reasonable range by using the prior setting of wind power engineering, which improves the convergence speed of the EM algorithm and reduces the error of missing value filling. Finally, experiments are compared with regression filling, random forest filling and multiple imputation filling to verify the effectiveness of the proposed method.

(2) A combined model wind power prediction algorithm based on Informer and graph convolution is proposed. The Informer model is used to capture the dependence between long series data of wind turbine power, and the spatial dependence of wind turbine power is captured by graph convolution. Finally, the wind power prediction of the spatiotemporal relationship is carried out through the combined prediction model of variance-covariance. Experimental results carried out on the real data set of a wind farm with a benchmark model show that the model proposed in this paper has better prediction effect.

(3) Aiming at the problems of scattered wind power related data, cumbersome data query, and lack of medium and long-term wind power forecasting functions, based on the above research results, a medium and long-term wind power forecasting platform was designed and developed using Vue and SpringBoot. The system can display the historical data related to wind turbine power, and integrates the above-mentioned mid- and long-term wind power prediction model and other benchmark prediction model algorithms, and has the function of visualizing the prediction results, providing a professional visualization platform for wind farm operators.

KEYWORDS： WindPower Prediction; Missing Value Filling; Spatio-Temporal Prediction Model; Medium And Long Term Forecast

目　录

1 绪论

本章节从研究背景和选题意义展开，介绍了风功率中长期预测的背景及现实意义，重点介绍了缺失值填充、时空数据预测和风功率中长期预测的国内外研究现状。最后对本文的研究难点、主要研究内容以及论文组织结构进行了说明。

1.1 研究背景及意义

随着我国社会经济和科技的快速发展，能源的需求和消耗逐渐变大，人们逐渐意识到节能减排和可持续发展的重要性，因此新能源的发展日益得到重视。大力推进太阳能发电、风力发电等新能源建设，逐渐成为能源转型[1]的主要趋势。而风能作为清洁能源的一种，风力发电无污染、可再生，且取之不尽用之不竭，基建周期短，装机规模灵活，运行和维护成本低，也可与其它的发电系统互补，是一种具备大规模商业开发条件的发电方式。如图 1-1 所示，随着风力发电技术日益成熟，我国的风电机组并网的装机容量逐年提升。

图 1-1 风电装机容量

Figure 1-1 Installed capacity of windPower

但受气候条件影响，风力发电具有动态性、随机性和不确定性，风电并网运行对电力电量平衡和电网安全运行产生较大影响，因此需要对风电功率进行有效预测来维持电力系统的稳定运行。及时且准确的对中长期风功率进行预测，不仅有利于风电场的自身管理和调节，对机组维护及检修工作进行提前计划安排，而且在与电网调度之间起到了积极的协调作用，减少风电场对电网系统产生的冲击，显著降

低风电接入电网产生的影响。因此，研究风电功率中长期预测对于支撑风电场经济运行决策和优化电网调度具有重要的意义和实用价值。

1.2 国内外研究现状

本文的研究对象是风力发电机组的功率数据，每台风机的功率数据是抽象的时间序列数据。由于风功率数据由设备采集并传输，若设备本身发生故障或数据在传输过程中网络出现异常状况，就会存在某段时间的数据丢失，数据呈现不完整性，从而导致预测的精度会在一定程度上有所下降，因此风功率数据的缺失填充是功率预测中较为重要的环节。本节首先介绍了缺失值的研究现状，再介绍时空数据预测的研究现状，最后介绍中长期风功率预测研究现状。

1.2.1 缺失值填充

数据预处理中的数据缺失问题一直是需要克服的困难，由于模型的训练依赖于数据，因此数据的质量直接影响模型训练的输出，不可靠的输出会对工程造成一定的损失。因此，国内外学者针对缺失值问题处理进行了深入的研究，并贡献了大量的科研成果。缺失值填充可以分为基于统计学填充方法、基于机器学习填充方法以及基于深度学习的方法[2]。

基于统计学缺失值填充包括均值插补法、回归插补法、多重插补法[3]等。均值填充方法没有充分的利用数据之间的关系学习，会造成数据的分布出现一定的偏差。回归插补法存在训练量过大，且无法处理非线性复杂数据的缺失和分类属性的数据缺失。Zhao[4]等研究学者提出了一种贝叶斯套索回归的多重插补法，在高维的数据上，验证了缺失填充的有效性。基于统计学的填充方法利用了数据集的共性进行填充，但没有考虑数据间的差异性，可能会导致填充后的数据同质化，使得数据集的整体方差变小，降低了数据的多样性，不适用于缺失比例较大的数据，填充的准确率较低。

基于机器学习的填充方法代表有加权重的最近邻（K Nearest Neighbors, KNN）算法[5]、集成聚类填充法[6]、基于矩阵分解填充[7]等。传统的最近邻算法需要遍历整个样本的空间，时间和空间的复杂度较高，因此效率较低，而且 k 是一个超参数，k 值的选取直接影响预测的精度。聚类填充方法没有考虑到数据的整体分布，仅考虑了数据的局部情况，且初始中心点的选定和类簇的个数选择都会对填充结果造成影响。加权重的 K 近邻填充法在权重的计算过程中引入了核函数，这样 K 近邻算法的鲁棒性得到了很大的提升。基于集成聚类的方法，使用每个聚类的均值

作为缺失属性的填充值，然后应用聚类质心的扰动分别寻找最优的填充。基于矩阵分解的填充方法，将原始的数据集当作一个矩阵，用两个矩阵的乘积来代表数据集的矩阵，利用矩阵乘积的结果来填充缺失值。

基于深度学习的填充方法近年来不断被提出，涌现出大量的研究成果。Beaulieu[8]等人通过去噪自编码器对缺失值进行填充，此方法可以处理分布复杂的数据情况，但是缺点是不适合时间序列的填充，仅适用于对非时序数据的填充。Silva[9]利用基于多层感知器和 K 近邻对缺失值填充，利用未缺失的数据集进行训练，得到三层的 MLP 神经网络，再结合 K 近邻对缺失数据进行填充。作为生成式的神经网络，生成式对抗网络也被应用到了缺失数据的填充工作。比如 Yoon[10]等人在 2018 年采用生成对抗网络的方法提出了一种生成对抗填补网络，主要根据观察值来计算出缺失分量并且输出到一个完整向量，但是没有考虑到时间序列。我国的研究学者 Luo 等人使用门控循环神经单元生成对抗网络模型 GRUI-GAN[11]，用来对时序的数据进行填充，本质上是将 GAN 与循环神经网络结合应用到时序数据的填充工作。时隔一年，Luo 团队又利用噪声自编码器来对 GRUI-GAN 进行改进，提出了 E^2GAN[12]模型，进一步提高了填充的精度。

1.2.2 时空数据预测

时空数据预测是利用历史的数据来预测未来一段时间的数据，包括连续型变量预测和离散型变量预测，具有较高的研究价值。时间序列数据预测在时间轴上体现事物的特征或状态随着时间变化而产生改变。与时间序列数据预测相比，时空序列数据预测更加复杂，不仅需要考虑随时间事物变化的抽象化表示，而且要考虑事物空间区域关系。空间区域又包含社会空间和地理位置等，且空间区域关系又会随着时间的发展不断的产生改变。因此可以得出结论，时空数据的预测复杂度要远高于时间序列数据的预测。通常将时空序列预测方法分为三种，基于统计学方法、基于机器学习方法和基于深度学习方法[13][14]。

基于统计学方法已较成熟，其模型简单且参数易于理解，在小数据集上实验效果较好，缺点是在规模较大的数据集上难以捕捉数据的非线性的规律，也较难捕获空间关联关系。ARIMA[15]模型反映了时间点越近，对当前时间点影响越大，该模型将时空序列数据看成多条独立的时间序列分别进行预测，但是忽略了时间序列之间的空间特征。STARIMA[16]模型通过使用加权矩阵反映了地点相邻从而影响更大，不过只能解决时间与空间平稳的序列，对于不平稳的序列，预测效果较差。向量自回归模型（Vector Auto Regression，VAR）[17]是一种非结构性方程组模型，是自回归模型的推广，通常用于描述时间序列数据多变量之间的动态关系，但此模型

需要较多的预测参数，样本不足时，较多参数的估量误差较大。决策树法[18]进行预测的思想是利用滑动窗口将时空序列预测问题划分成一系列固定长度的子问题，然后再分别进行预测，但是此预测方法的缺点是方差较大，数据的分布细微改变极可能导致树结构的大相径庭，且难以捕获时空特征的动态变化。

基于机器学习方法也被大量学者应用到时空预测中的任务中。Kobayashi[19]等人提出基于隐马尔可夫模型的时空序列预测方法，该算法通过隐马尔科夫链生成一系列状态的随机序列，再由各个状态生成预测。王佳璆等人[20]受到时空核函数启发，提出时空支持向量机模型（Spatio-Temporal Support Vector Regression, STSVR），其创新点是在传统支持向量机基础上引入了时空核函数，模型中对权重的分配采用折扣最小平方法，分配更大的权重值给距离更近的数据，其缺点是不适用小样本数据，对样本数据量需求较大，并且计算的时间和空间复杂度较高。柳娇娇等人[21]基于马尔可夫模型上引入时空密度聚类进行时空预测，此模型首先采用 PLR 方法将数据序列分段，再依次对分段后的数据聚类操作，最后使用隐马尔科夫模型进行时空预测。DeepAR[22]模型由亚马逊公司 2017 年提出，已集成到 Amazon 开源时序预测库中，该算法是一个自回归循环神经网络，结合递归神经网络和自回归进行时间序列预测，可从时间序列中对全局模型进行有效的学习，模型的输出是可选时间的多步预测结果，单节点的预测为概率预测。这样做的好处有两点，第一很多过程本身就具有随机属性，因此输出的概率分布更加贴近实际，预测的精度反而更高。第二可以对预测的不确定性及相关风险进行评估。基于机器学习的预测方法能在一定程度上对时空数据特征的捕获有一定效果，但是针对大数据量规模的中长期预测，效果欠佳。

基于深度学习方法在大规模时空数据的中长期预测工作中，取得了不错的效果，有着较好的预测精度。Zhang 等[23]提出了一种基于深度网络的时空序列数据预测模型，该模型将时间序列分为三个子序列，依次表示数据的邻近性、周期性和趋势性，接着分别对子序列进行卷积融合操作，再对融合后的数据进行卷积操作，最后与全连接层进行特征融合，输出最终预测值。该方法有效对时空数据的时空特征进行提取，与卷积神经网络等模型进行比较，预测结果更好。Guo[24]等提出一种 ASTGCN 时空预测模型，该模型先在空间维度进行图卷积，再在时间维度进行一维卷积，缺点是未同时对时间和空间特征进行特征捕获。Wu[25]等提出了一个针对多元时间序列数据预测的框架，通过图模块学习变量之间的关系，对时间上空洞卷积层和空间图卷积层进行进一步改进来捕获时空的依赖关系。Zhao[26]等提出了时间图卷积网络预测模型（T-GCN），该模型使用图卷积用于捕获复杂的空间依赖关系，使用门控循环单元捕获动态变化的时间依赖关系。ConvLSTM[27]模型不仅能够建立如 LSTM 的时序关系，而且可以拥有 CNN 的空间特征提取能力。实验证明

ConvLSTM 在获取时空预测上有着不错的效果，而且 ConvLSTM 还能够解决其他时空序列的预测场景问题。STGCN[28]由北大团队提出，该模型由多个时空卷积模块构成，每个模块包含一个空间卷积模块和两个门控序列卷积层，类似一个三明治的结构，优点是参数少，而且训练速度更快。

时空序列数据通常相对复杂，空间和时间特征难以捕获，如何充分利用各种各样影响因素，提升预测精确度依旧是可以持续关注的研究话题。

1.2.3 风功率中长期预测

国内外于上世纪六七十年代就陆续开展了风功率预测相关工作，近年来，随着全球能源的不断紧缺，电力负荷预测研究课题愈来愈受到专家学者关注和研究。经过长期的研究过程，业内已经形成很多预测方法，总结归纳包括以下三类：基于统计模型的传统预测方法、基于神经网络为代表的机器学习方法和组合预测方法。

传统预测方法以回归分析法[29]和灰色预测模型[30]（Grey Model, GM）为代表。回归分析法思想通过设定回归方程，将电力负荷当作方程因变量，求出方程的回归系数，通过建立电力负荷的函数方程关系进行负荷预测。回归分析法的优点包括计算时间复杂度较低、参数较少和可解释性强等，但是鲁棒性较差、容易过拟合，在非线性数据中难以获得较好的预测精度。灰色预测模型的特点是用灰色数学处理不确定量，使其量化，充分地利用已知信息挖掘数据规律，此模型适用于历史样本量较小的数据集，计算的复杂度较低，不需对历史数据的分布规律考虑。针对灰色模型，一些学者进行了优化。鲁宝春等[31]人提出了基于优化 NGM（1,1,k）灰色预测模型，增加一个修正量，对灰色模型的系数进行修正，并使用缓冲算子对原始数据进行预处理，该模型具有不错的预测效果。

机器学习算法对风功率中长期预测以支持向量机[32]、人工神经网络[33]为代表。支持向量机利用非线性映射将数据映射到高位特征空间，并在此特征空间进行预测。在中长期预测方面，张健美[34]采用灰色 Elman 神经网络模型进行预测，将灰色理论依赖样本小、计算复杂度小的特点和神经网络特点结合，进一步提高预测速度和预测精度。Li Bowen 等[35]提出了基于小波分解的二阶灰色神经网络预测模型，首先对负荷序列通过小波变换进行分解，再对小波变换分解后的各个分量采用二阶灰色预测模型进行预测。

单一的负荷预测方法适用场景固定，较难满足负荷预测需求，因此较多研究的重点集中在采用组合预测方法[36]进行电力负荷预测。通常组合预测方法分为两种思路，一种是通过赋予模型权重的方式，也称为加权组合预测模型；第二种是利用优化算法组合模型，计算单一模型的最优参数，通过赋权方式将单个预测模型的结

果二次拟合，得到新的组合预测模型。这种模型能获得单一预测模型的优点，而且预测结果对单一模型的敏感度降低，提高预测模型的抗干扰能力，但是权重系数直接影响最终模型的预测结果。周淙等[37]提出了层次结构的变权组合预测方法，采用熵值法来计算模型权重，最终确定组合预测中的组合权重参数。也有学者通过生物界的群体智能行为建模计算权重系数，这种数学建模个体逻辑不但具有简单性，而且群体逻辑具有复杂性。常见的群智能优化算法包括：Wang Xiping[38]提出的粒子群优化算法，Wang Jianjun[39]提出差分进化算法等。Li 等[40]分别构建差分整合移动平均自回归模型、Elman 神经网络模型和相似日预测模型三个预测模型，权重系数使用改进粒子群优化变权重组合模型确定，最终实验获得了较好的预测结果。赵芝璞等[41]提出一种基于关联模糊神经网络和改进型蜂群算法的负荷预测方法，实验证明在某些数据集上，该方法有较高预测精度。

时空序列数据内部之间复杂性较高，如何更好的捕获时间和空间特征，提升预测的精度依旧是值得研究的热点问题。

1.3 研究难点与研究内容

风功率中长期预测对我国风力发电发展具有很高的实际应用价值，对于风电场基础设施建设和规划起到一定的辅助作用。尽管风机功率历史数据庞大，但数据受气候、环境等多因素影响，呈现不平稳性和间隙性，又因风功率历史数据中存在一定比例的缺失值和异常值，对风功率预测造成了一定困难。如何高精度的填充缺失数据是需要解决的难点。目前中长期风功率预测研究基本都是基于时序序列的数据关系进行建模，鲜有融合空间特征，所以融合风电机组空间关联特征进行预测也是本文研究的一个难点和重点。由于目前的模型都是定期更新，时效性不强，少有利用大数据平台计算组件的强大计算能力来提高模型的时效性，如何利用大数据实时计算组件提高模型的时效性也是研究的一个难点。针对以上问题，本文的研究内容如下：

（1）提出了基于工程先验知识和期望最大算法的缺失值填充方法。

针对风电功率历史数据的缺失问题，在期望最大算法填充的基础上引入风功率工程先验知识。通过缺失时刻的风速计算出风机的理论功率，结合聚类和 K 近邻算法，进一步求得期望最大算法初始化参数，将其控制在合理的范围内，提高期望最大算法收敛的速度，减少缺失值填充的误差。最后在各缺失比例下的数据集进行实验验证分析。

（2）提出了基于 Informer 和图卷积的组合模型预测算法。

针对风功率中长期预测不仅与历史时间序列变化相关，各个风机由于地理分

布位置、风机本身产生风速等影响，在空间上也具有一定的相关性，提出了基于Informer 模型和图卷积的组合预测模型。首先通过 Informer 捕获长序列时序关系，图卷积捕获空间特征依赖关系，再通过组合预测模型融合空间和时间的特征关系，得到最终预测结果。最后在某风电场数据集上对算法效果进行了分析验证。

（3）实现了一种风功率预测可视化平台

针对风电场繁琐的数据查询工作，缺少可靠的中长期预测功能等问题，实现了一个风功率中长期预测平台。平台展示风功率相关指标数据，集成了本文所提出的风功率预测模型和其它基准预测模型算法，并利用 Spark 组件的实时计算能力，提高模型的实效性，具有预测结果可视化展示功能。为发电厂工作人员提供可靠稳定的风功率预测可视化平台。

1.4 章节安排

本论文分为六大章节，各章节内容如下。

第一章：绪论。本章从研究背景及意义开始，介绍了风功率中长期预测对风电场的重要性。接着对缺失值研究现状、时空数据预测现状和风功率中长期预测的研究现状进行了充分调研。最后介绍了全文的主要研究内容、研究难点和章节安排。

第二章：相关理论基础。本章首先阐述了时空数据挖掘的概念，接着分别论述了时序数据预测模型 Informer、图神经网络概念及相关模型和组合模型。最后介绍了本文采用的评价指标及计算公式。

第三章：基于工程先验知识和期望最大算法的风功率缺失值填充研究。首先介绍数据采集的相关工作，再介绍工程先验知识与期望最大算法结合的算法流程，最后通过实验对比证明此工作的有效性，为之后的工作做好数据支撑。

第四章：提出了基于 Informer 和图卷积的风功率组合预测模型。首先进行空间相关性研究，确定空间相关性研究的正确性，接着提出预测方案，介绍 Informer 模型和图卷积组合预测的算法流程，最后在数据集上进行实验，与基准模型进行对比，证明提出的研究模型的有效性。

第五章：风功率预测可视化平台的设计和实现，介绍了平台的系统架构设计、技术栈、系统功能，并展示平台功能页面。

第六章：总结与展望。本章对本文的科研内容和实验结果进行了分析总结，指出工作中的不足之处以及亟需优化之处，并对未来的研究内容进行展望。

2 相关理论基础

本章首先介绍了时空数据挖掘的基本知识，接着介绍了本文中时序预测模型 Informer 和空间预测模型图卷积的基本概念和网络模型结构，最后介绍了方差-协方差的组合模型和本文中所用到的评价指标。

2.1 时空数据挖掘

时空数据指的是同时具备空间和时间特性的数据，从空间特性上看，不同的对象动态地分布在复杂的地理空间中，譬如交通路线、气候科学、城市轨道交通路线等，这些对象由于会在空间维度上呈现不断的运动轨迹，因此出现了较为复杂的空间相关性。从时间特性来看，对象的运动呈现较强的周期性、随机性等时间序列特征。本文研究的风功率中长期预测，属于时空问题的范畴，接下来将从数据特征、数据类型以及时空轨迹数据展开介绍。

2.1.1 数据特征

时空序列数据是同时具有时间戳信息和位置信息的一系列数据序列，数据的内部特征将随时间不断变化。数据的来源多种多样，数据之间的噪声，数据的精度不一致，数据采样的时间维度也难以对齐，这些问题都增加了时空数据预测的难度。在所有的时空数据中，都具有时间相关性、空间相关性、时空相关性这三个特征。

时间相关性指的是每个数据点在不同时间节点存在较强的数据依赖关系。这种时间相关性又可以分为时间周期性、近邻性、趋势性。周期性体现在比如说风电场每天的中午的风速较大，风机的输出功率较大，而早上和晚上的风速较小，因此风机的输出功率较小；近邻性体现在譬如位置相邻的风机风速接近，故产生的风机输出功率也接近；趋势性体现在例如随着四季的变化，气候条件会随之改变，也会影响风机功率的输出变化。

空间相关性指的是数据点在不同的空间维度上会相互影响，空间维度上的数据大部分带有地理位置信息，因此不同地理位置类型的数据有着不同特点。比如处于海拔较高的风机、风速较大处的风机输出功率越大，反之越小。

时空相关性指的是时空序列数据在时间维度和空间维度是会产生相互影响作用的，而非独立的。同一位置坐标的时间序列数据会有时序关联，不同位置坐标的时间序列也会产生一定的影响。比如相邻的风机，风机转速产生的风速会对附近其他风机带来影响，进而影响附近风机的功率。

2.1.2 数据类型

时空数据在不同场景下有着不同的处理方式，随之也产生了不同形式的时空数据，根据时空数据的特点，分为时空网格数据、时空图数据、时空轨迹数据三种。

时空网格数据是将计算机视觉的技术运用到时空预测问题中来，提出 n 个地点的空间关系。将每个区域看作图片中的像素点，将区域特征看作图片的通道，即目标空间 χ 划分为包含了 n 个点 $P\times Q$ 的网格图，通道数 D 对应每个采样点取值的特征维度。此类数据方便使用二维矩阵和三维矩阵进行表示，因此通常二维卷积和三维卷积[42]的方法进行数据建模。时空网格数据如图 2-1 所示：

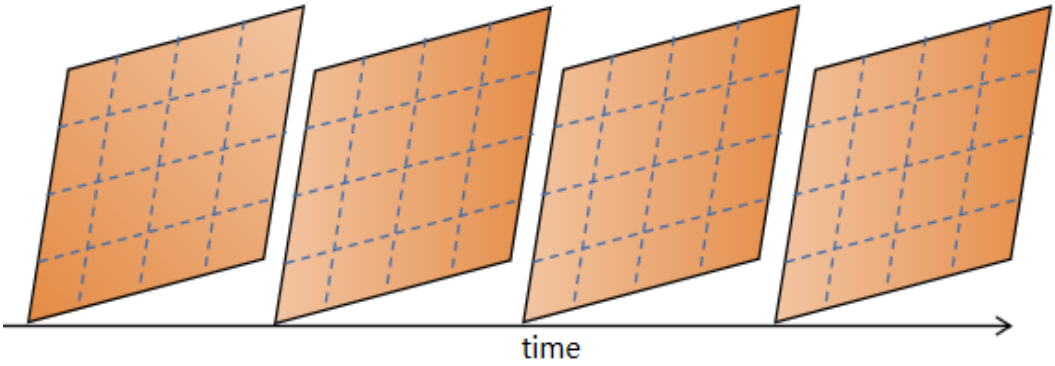

图 2-1 时空网格数据示例

Figure 2-1 An example of spatial-temporal grid data

时空网格数据只适用于网格数据，但是实际场景下绝大部分都是非欧几里得空间数据，比如社交网络、信息网络和知识图谱等，时空图数据就是解决非欧几里得空间数据问题。时空图数据的采样方法与时空网格数据类似，但与时空网格数据不同的是，时空图数据在空间分布上是不均匀的，而这种不均匀的数据，在真实场景中的应用更为广泛。t 时刻节点 i 的变量 $\boldsymbol{X}_{i,t}$ 与图中其它节点都是相连接，故 $\boldsymbol{X}_{i,t}$ 可以作为节点 i 定义在图 G 上的图信号。t 时刻图可以表示为 $G_t=\left(X_{:,t},\xi,W\right)$，其中 $X_{:,t}$ 是 t 时刻所有节点特征的集合，ξ 是边的集合，表示节点的连通性，$\boldsymbol{W}\in\mathbf{R}^{n\times n}$ 是图 G_t 对应的加权邻接矩阵。例如地铁中将每一个站点作为数据点，将两个站点之间作为边。面向时空图数据的预测算法，包括 STGCN、T-GCN 等。 时空网格数据如图 2-2 所示：

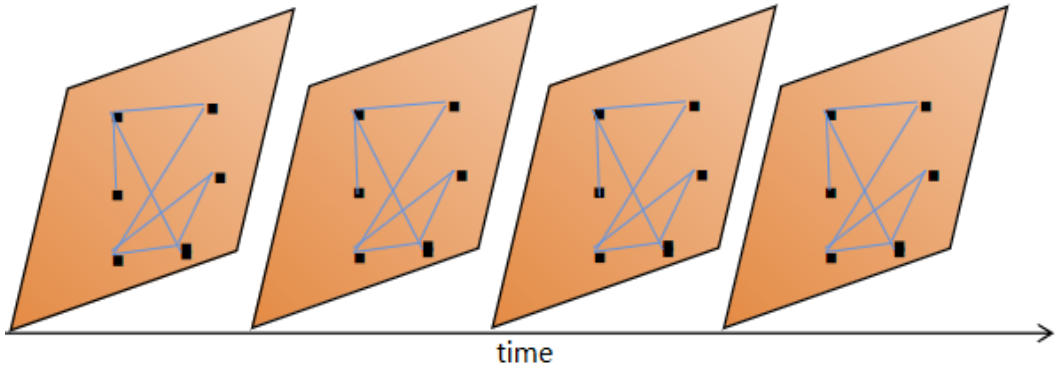

图 2-2 时空图数据实例

Figure 2-2 An example of spatial-temporal graph data

时空轨迹数据通常经过设备进行采样，记录对象各时间点的位置、时间和速度等特征。此数据类型比时空网格数据、时空图数据更加复杂，因为对象的活动灵活且不确定性更大，在区域停留时间也具有不固定性，因此在时间和空间上的特征都呈现出不规则性。因此轨迹预测是根据历史的轨迹预测未来一段时间内对象位置的过程。譬如外卖员的接单路线的数据，属于典型的时空轨迹数据的范围。时空网格数据如图 2-3 所示：

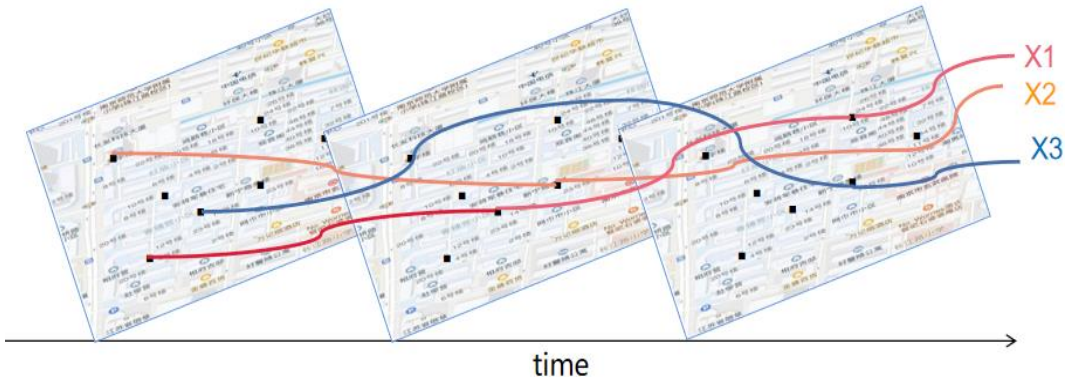

图 2-3 时空轨迹数据示例

Figure 2-3 An example of spatial-temporal trajectory data

2.2 Informer 模型

2017 年，Attention is all you need 为学术界带来了 Transformer 模型[43]，鉴于其在 NLP 领域对时序数据的强大建模能力，Tranformer 也被应用到了时序数据预测上来，并且取得了不错的预测效果。尽管如此，Transformer 在长时间序列预测问题上也存在一些不足，如二次的时间复杂度、编码器-解码器[44]架构的固有限制和高内存使用率。针对这些问题，Informer[45]模型对这些不足采取了如下的改进：①

提出一种稀疏注意力机制，筛选出重要的 query，降低时间复杂度和空间的内存开销；②提出了自注意力蒸馏，通过将层级输入减半来突出主导注意力，减少了维度和网络参数量，并高效处理极端长的输入序列；③提出了生成式解码器，不需要分步操作的方式预测长时序序列，只需一步得到所有预测结果，大大提高预测的推理速度。

Informer 的模型结构如图 2-4 所示，左侧为编码器（Encoder），右侧为解码器（Decoder）。编码器负责长序列数据的输入接收，将传统的自注意力机制替换为稀疏自注意力机制，蓝色梯形部分为自注意力提取操作，可以大幅度地减少网络规模，通过层层的堆叠提高了模型的鲁棒性。解码器（Decoder）接收长序列数据的输入，将目标元素使用零填充，将需要预测的全零序列也作为特征图加权注意力的一部分，接着使用生成式的方式对预测序列进行预测。

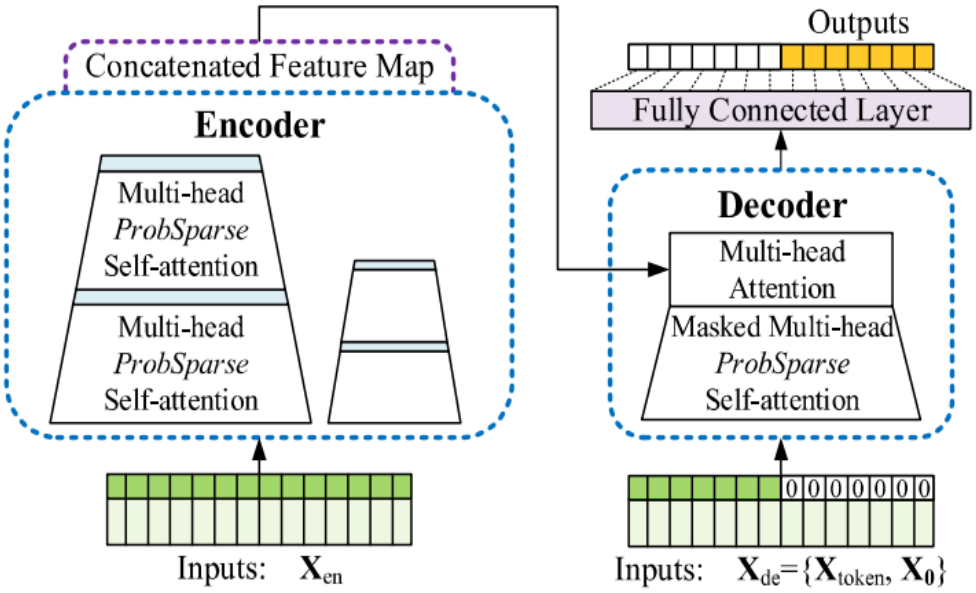

图 2-4 Informer 结构[45]

Figure 2-4 Structure of Informer[45]

2.2.1 Informer 模型的自注意力机制

原始的自注意力（self-attention）机制的输入形式是(Q，K，V)，接着再按比例进行缩放点积（scaled dot-product），即：

$$A(Q,K,V)=Softmax(\frac{(QK^T)}{\sqrt{d}})V \tag{2-1}$$

其中，d 是输入维度，$Q\in R^{LQ\times d}$，$K\in R^{LK\times d}$，$V\in R^{LV\times d}$，分别代表 query，key 和 value。其中 key 和 value 是配对的，query 根据 key 查询 value。第 i 个

query 的注意力被定义为概率形式的核平滑器，即：

$$A(q_i,K,V)=\sum_j \frac{k(q_i,k_i)}{\sum_l k(q_i,k_l)} v_j = E_{p(k_i|\ q_i)}\left[V_j\right] \tag{2-2}$$

其中，$p(l_i|\ q_i)=\frac{k(q_i,k_i)}{\sum_i k(q_i,k_l)}$ ，$p(l_i|\ q_i)=\frac{k(q_i,k_i)}{\sum_i k(q_i,k_l)}$，$k(q_i,k_i)$选择$\exp\left(\frac{q_i k_j^T}{\sqrt{d}}\right)$非对称指数。

自注意力机制使用点积运算来计算概率 $p(q_i,k_j)$，达到二次时间复杂度，且计算需要$O(L_Q,L_K)$的空间复杂度，复杂的时间和空间复杂度是在长序列预测中的主要障碍。基于概率的研究发现，自注意力概率的分布具有稀疏性。首先对学到的自注意力模式做定性评估，发现稀疏性自注意力分数呈现长尾分布，即少数的点积构成了主要的注意力，对主要的注意力贡献比重较大，从而可以忽略其它点积的贡献，所以部分的 query 对于 value 的贡献可以不需计算。为了区分出这部分 query，作者用到了 KL 散度来衡量相似度。第i个 Query 稀疏性的评价公式为：

$$M(q_i,K)=\ln\sum_{K_K}^{j=1} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L_K}\sum_{L_K}^{j=1}\frac{q_i K_j^T}{\sqrt{d}} \tag{2-3}$$

其中第一项是q_i在所有 key 上的 Log-Sum-Exp（LSE），第二项则是算术平均值。如果第 i 个 query 获得较大$M(q_i,K)$，那么它的注意力分布具有更大的多样化，在自注意力长尾分布中的头部中有更大的概率包含占主导地位的点积对。基于此分析，就可以得到 ProbSparse self-attetion 公式，允许每个 key 只处理 u 个占主导地位的 query，即公式：

$$A(Q,K,V)=Softmax\left(\frac{\bar{Q}K^T}{\sqrt{d}}\right)V \tag{2-4}$$

其中，$\bar{Q}$和q有着相同大小的稀疏矩阵，并只包含在稀疏评估$M(q,M)$下 Top-u 个的 query，由一个固定的采样因子c来控制u，设$u=c\cdot\ln L_Q$。这样 ProbSparse self-attention 在每个 query-key 匹配中只需要$O(\ln L_Q)$个点积操作，使得每一层的内存开销降低为$O(L_K \ln L_Q)$。

为了降低时间复杂度，模型提出一种有效获取 query 稀疏度测量的经验近似方法。Informer 提出如下的 max-mean 度量：

$$\bar{M}(q_i,K)=\max_j\left\{\frac{q_i k_j^T}{\sqrt{d}}\right\}-\frac{1}{L_K}\sum_{j=1}L_K\frac{q_i k_j^T}{\sqrt{d}} \tag{2-5}$$

在长尾分布的情况下，随机抽取$U=L_K\ln L_Q$个点击计算$\bar{M}(q_i,K)$，用零填充不重要的点积对。选择 Top-u 个作为$\bar{Q}$，$\bar{M}(q_i,K)$中的最大算子对零不敏感，而且在数值上也很稳定。在实际情况中，query 和 key 的输入长度在自注意力计算中往往是相等，即$L_Q=L_K=L$，这样时间和空间复杂度降低为$O(L\ \ln L\ )$。

2.2.2 Informer 模型的编码器

编码器（Encoder）设计用于提取长序列输入的鲁棒长期依赖。由于编码器的特征映射中包含输入 value 的冗余组合，故使用了上节中的稀疏自注意机制。编码器利用蒸馏操作对具有主导地位的高级特征进行特权化，并在下层生成聚焦的自注意力特征图，这样可以削减输入的长度，从 j 层到 $j+1$层的蒸馏操作如下公式：

$$X_{j+1}^{t} = MaxPool\left(ELU\left(Convld\left(\left[X_{j}^{t}\right]_{AB}\right)\right)\right) \tag{2-6}$$

其中，$\left\lfloor X_{j}^{t}\right\rfloor_{AB}$ 包含了注意力块和稀疏注意力机制的基本操作，$Convld$ 使用 ELU 激活函数在时间维度进行一维卷积滤波器操作，接着使用一个步长为 2 的最大池化层，在每层之后将 X^{t} 下采样到一半长度，这样大大减少了内存的使用。为了增强蒸馏操作的鲁棒性，对编码器结构进行堆叠组合，再将它们的输出进行拼接得到完整的编码器的输出。因此模型提出的自注意力蒸馏机制，可以将每层解码器的输入序列长度进行减半，大大减少了编码器的计算时间和内存开销，从而可以处理更长的时间序列输入。

2.2.3 Informer 模型的解码器

解码器设计目标是通过一个前向过程就可以生成长序列预测。模型使用传统的 Decoder 结构，由两个相同的多头注意力层组成，用生成式预测方式来解决长序列数据预测中时间复杂度较高的问题。解码器的输入向量如下表示：

$$X_{\text{feed\_de}}^{t} = Concat\left(X_{token}^{t}, X_{0}^{t}\right) \in \mathbb{R}^{\left(L_{token}+L_{y}\right)\times d_{\text{model}}} \tag{2-7}$$

其中，X_{token}^{t} 是 start token，从输入序列中选择长度为 L_{token} 的长序列作为 token，也就是需要预测序列前面的一段序列。X_{O}^{t} 则是目标序列的占位符，其设置为 0。将隐藏多头注意力（masked multi-head attention）应用于稀疏自注意力计算中，使用隐藏多头注意力机制的好处是可以避免自回归，原因是不会使每一个位置都注意到下一个位置。全连接层最后获得了最终的输出，可以进行单变量和多变量的预测，这取决于它的输出维度。该方法将生成式结构使用在编码器中，能够一次性的输出预测序列，摆脱原本的动态解码操作，从而预测的解码时间得到了大幅度缩短。

2.3 图神经网络

实际的场景中很多数据都是由非欧式空间产生，传统深度学习在处理非欧式空间数据上变得艰难，故学者们设计了应用图数据的神经网络结构，即图神经网络。

本节将从图的基本概念、图的表示、图神经网络模型展开介绍。

2.3.1 图的基本概念

图（Graph）是现有生活中最常见的数据结构之一，例如用户关系网络、通信网络、交通路网拓扑等领域都可以用图结构来表示。图是由“节点”和“边”组成的结构。“节点”譬如地铁站点、社交网络中的用户、风电场的风机等，一般都会携带一些额外信息。“边”的作用是图中的两个“节点”进行连接，当两个节点存在关联关系的时候，就可以通过边把他们连接起来。通常定义$G = (V, E, A)$来表示成一个图，其中 V 代表的是图 G 的有限节点集合，E 代表的是图 G 的边集合，每条边对应图中的两个节点，A 代表的是图的邻接矩阵。邻接矩阵使用二维数组存放顶点之间的关系。根据是否区分图的方向，又可细分为无向图的邻接矩阵和有向图的邻接矩阵。

2.3.2 图的表示

从信息传递的方向对图分类，可以将图分为两类，即有向图和无向图，根据边之间是否有权值又可分为不加权图和加权图。任意类型的图都可以使用图的节点矩阵和节点之间的拓扑关系来进行唯一性确定。

一般使用特征矩阵用来表示图中各个节点的特征。若 M 代表的是图中节点的个数，N 代表的是节点的特征维度，那么特征矩阵的维度即是 M×N。比如在风电场中，节点代表的是风机，比如有 10 台风机，而每一台风机具有风速、温度、湿度、空气密度特征，比如有 4 个特征，那么特征矩阵的维度可以表示为 50*4。然而如果只有一个特征矩阵表示，则一个图是不能被唯一确定的。若想唯一确定，那么就需要了解每个节点之间的拓扑关系。

邻接矩阵就可以用来存放图的边或弧的信息，借此用来表示节点与节点之间的复杂拓扑关系，邻接矩阵为一个二维数组，维度为 M×M。其中每一行代表对应节点和除它之外其他节点之间的连接关系或权值。若是无权图，则邻接矩阵中的元素值要么为 0，要么为 1 。若为 0 表示不连通，若为 1 表示连通，若是有权图，则邻接矩阵中的元素值则为边的权值。有向图的邻接矩阵不一定为对称矩阵，两个节点之间不一定是完全一致的。但是在无向图中邻接矩阵是对称矩阵的，代表着从节点 a 到节点 b 和从节点 b 到节点是完全一致的。

2.3.3 图神经网络模型

图神经网络[46]最早提于2009年，随着研究的深入，又衍生出图生成网络（Graph Generative Networks）、图时空网络（Graph Spatial-temporal Networks）、图注意网络（Graph Attention Networks）、图卷积神经网络（Graph Convolution Networks）等。图卷积神经网络由于具有捕获非结构化数据的处理能力而受到学者们的大量研究。本文使用到的就是图卷积网络。图卷积神经网络是为了捕捉数据的全局相关性和局部相关性而设计。图卷积神经网络可以分为空间图卷积神经网络和谱图卷积神经网络两大类。空间图卷积中的卷积运算通过直接定义卷积在图上的邻域节点来进行，不断聚合节点的邻居信息。谱卷积通过傅里叶变换将空域信号转化到频域，再将卷积的结果转换到空域，谱域卷积其实是空域卷积的特例。本文主要采用的基于谱域的图卷积。

第一代的谱图卷积（Spectral CNN）[47]是在2014年由Bruna等人提出，使用拉普拉斯矩阵把图数据转换到谱域上进行卷积运算，但是此运算存在三个不足，首先是计算复杂度太大，这是因为使用了拉普拉斯矩阵，导致在特征值分解时，计算的效率受到了影响。其次是在进行特征聚合的时候，有可能会将不相连的节点的信息进行聚合，却没有聚合邻居的节点，并没有满足卷积神经网络设计的初衷，实现局部连接。最后是因为图数据的数据量庞大，导致参数量过大，计算的时间和空间的复杂度过高。

针对SCNN的缺点，2016年Defferrard等人提出ChebNet[48]网络可以捕获局部信息，使用切比雪夫多项式中的K阶截断来计算卷积核，代替谱域的卷积核，同时在不使用拉普拉斯矩阵的情况下使得计算的复杂度得到减少。

2017年Kipf等[49]对ChebNet中的卷积操作做了进一步优化，通过使用一阶邻居仅考虑1阶切比雪夫多项式。并且提出了图卷积GCN（Graph Convolutional Network）的简单公式推导模型，计算公式如下：

$$x_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \frac{1}{C_{ij}} x_j^{(l)} w^{(l)} + b^{(l)} \right) \tag{2-8}$$

其中 $x_i^{(l)}$ 代表的是在第1层节点 i 的特征，c_{ij} 代表归一化因子，N_i 代表节点 i 的所有邻居也包括 i 本身，$w^{(l)}$ 代表的是第 l 层的权重，$b^{(l)}$ 代表的时候第 l 层的截距。

2018年Velickovic等[50]把注意力机制运用到了图卷积中并且提出了GAT（Graph Attention Network）模型，这是一种空域模型。此模型使用注意力机制对邻居节点进行特征加权求和，节点特征决定了邻居节点特征的权重，这也是它相对于谱卷积方法的优点。GAT根据每个节点在其邻居节点上的注意得分更新节点

的表示。它的优点是节点并行计算、计算速度快和同时处理不同程度节点。其计算公式为：

$$\boldsymbol{h}_i^{(l+1)}=\sigma\left(\sum_{j\in\mathcal{N}_i}\frac{\exp\left(a\left(\boldsymbol{w}\vec{\boldsymbol{h}}_i,\boldsymbol{w}\vec{\boldsymbol{h}}_j\right)\right)}{\sum_{k\in\mathcal{N}_{(i)}}\exp\left(a\left(\boldsymbol{w}\vec{\boldsymbol{h}}_i,\boldsymbol{w}\vec{\boldsymbol{h}}_j\right)\right)}\boldsymbol{w}\vec{\boldsymbol{h}}_j\right) \tag{2-9}$$

其中$W\in\mathbb{R}^{F'\times F}$为权重系数矩阵，$F$为节点的特征维度，$\vec{h}$表示 GAT 输入节点的特征值，$a$ 代表的是$\mathbb{R}^{F'}\times\mathbb{R}^{F'}\to\mathbb{R}$的映射。

2.4 组合模型

在预测的研究中，同一个预测问题可以使用不同预测方法，由于考虑的角度等不一致，不同的模型预测结果包含不同的预测信息，将这些预测结果进行有效组合，可以提升预测的信息量，提高预测的精度。组合模型就是将单模型缺陷尽可能地缩小到最小化，从而构建最优模型。组合模型的研究重点将数个模型的预测结果通过选取合适的权重进行组合，其中方差-协方差组合预测法是电力负荷预测中最常用的一种组合预测方法。

2.4.1 基于方差-协方差组合预测法

设有q个单一预测模型，q个模型预测值的组合预测结果如下：

$$f_c=\sum_{q}^{i=1}w_i f_i,\sum_{q}^{i=1}w_i=1 \tag{2-10}$$

式中$f_1,f_2,\ldots,f_q$为q个预测模型的预测值，$w_1,w_2,\cdots,w_q$为相应的权重系数。

由于各模型预测结果之间是互为独立，记各模型预测误差的方差为$\delta_{11},\delta_{22},\cdots,\delta_{qq}$，则组合预测的方差可以表示为

$$D\left(e_c\right)=\sum_{q}^{i=1}w_i^2\delta_{ii} \tag{2-11}$$

为寻各模型的最优权重，即求$D\left(e_c\right)$关于$w_i(i=1,2,\cdots,q)$的极小值，且需满足$\sum_{q}^{i=1}w_i=1$，引入拉格朗日乘子来求最小值，可以得到：

$$w_i=\frac{1}{\delta_{ii}\left(\frac{1}{\delta_{11}}+\frac{1}{\delta_{22}}+\cdots+\frac{1}{\delta_{qq}}\right)}(i=1,2\cdots,q) \tag{2-12}$$

该方法理论上可以获得最佳的组合权系数，预测的稳定性和准确性较高。

2.5 预测评价指标

本小节介绍风功率中长期预测所采用指标评价。风功率中长期预测的好坏通常由真实功率值与预测功率值之间的误差值反应，为了客观评价本文的预测实验结果，采用了如下的评价方法：

（1）均方误差（MSE）：均方误差是反映预测值和真实值之间差异的一种度量。均方误差值越小，说明预测的模型具有更好的精度。如式（2-13）所示，其中 $\hat{y}_l$ 表示预测值，y 为真值。

$$MSE = \frac{1}{n}\sum_{n}^{i=1}\left(\hat{y}_l - y_i\right)^2 \tag{2-13}$$

（2）均方根误差（RMSE）：均方根误差是预测值与真实值偏差的平方与观测次数 n 比值的平方根。如式（2-14）所示，其中 $\hat{y}_l$ 表示预测值，y 为真实值。

$$RMSE = \sqrt{\frac{1}{n}\sum_{n}^{i=1}\left(\hat{y}_l - y_i\right)^2} \tag{2-14}$$

（3）平均绝对误差（MAE）：表示预测值和预测值之间的绝对误差的平均值，是一种线性分数。平均绝对误差越小，说明模型的预测效果更好，预测精度更高。如式（2-15）所示，其中 $\hat{y}_l$ 表示预测值，y 为真实值。

$$MAE = \frac{1}{n}\sum_{n}^{i=1}\left|\hat{y}_l - y_i\right| \tag{2-15}$$

（4）决定系数（R^2 SCORE）：绝对系数是用来表示拟合的曲线的拟合优度。取值范围为[0,1]，结果为 0，说明拟合的效果很差，结果为 1，说明拟合的效果较好。

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{n}^{i=1}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{n}^{i=1}\left(y_i - \bar{y}\right)^2} = 1 - \frac{SSE}{SST} \tag{2-16}$$

2.6 本章小结

本章先对时空数据挖掘的理论基础做了介绍，然后对本文用到的 Informer 模型进行介绍，对比 Transformer 模型，介绍了它的优点，再介绍了图神经网络模型，说明在空间特征提取的优越性，接着介绍了本文所用到的基于方差-协方差的组合预测法。最后介绍了本文实验中用到的评价指标。

3　基于工程先验知识和期望最大的风功率缺失值填充方法

无论是在科学研究还是在现实调查中，大部分的数据集都不可避免出现缺失情况，这极大地影响了模型的训练和预测的准确性。无论是简单的删除还是忽略缺失数据都会使原始数据信息量发生变化，因此缺失数据的填充是数据预处理阶段重要步骤，如何对风功率历史缺失值进行高精度的填充是本节接下来研究重点。

本章基于期望最大算法缺失值填充进行优化。该算法是求最大似然估计的迭代算法，在数据量较大时候，算法执行简单，通过迭代的步骤找到全局的最优解，理想情况下，填充的精度是较高的。但是此算法忽略了数据的局部相似性，需要遍历整个数据集，同时它的收敛速度也较慢，时间复杂度较高。本节结合工程知识，通过缺失时刻的风速对应的理论功率，并且充分利用数据之间的相似性，对期望最大算法初始化参数进行合理的范围限制，提高收敛的速度，避免陷入局部最优解。

3.1 相关工作

风机数据往往存在于时序数据库中，虽然历史数据庞大，但是其存储形式复杂，很难直接应用到时序数据预测工作中去，因此需要对时序数据库中数据做个性化抽取和整理。本节介绍了数据采集和数据整理的相关工作。

3.1.1　数据采集

风电场的数据都是通过采集设备 24 小时不断采集并写入到时序数据库中去。本文需要从实时库中整理出与功率输出相关的特征数据。根据功率公式(3-1)可以看出，p 为风机输出功率，C_p 为风轮的功率系数，固定系数不影响输出，A 是风轮扫掠面积，ρ 和 V 为当前时刻空气密度和风速。通过公式可以看出功率和风速及空气密度相关，且风速影响最大，其中空气密度由风机的海拔和风机当前测得的环境温度计算而来。风机特征数据是每隔 15 分钟采集的风机风速、温度、功率，因此通过获取风机对应的风速、温度、功率对应的测点，再根据测点从实时库获取对应的特征历史值。本文选取了某区域风电场下 2018 年初到 2021 年底的对应数据作为数据集。

$$P=\frac{1}{2}C_P A\rho V^3 \tag{3-1}$$

3.1.2 数据整理

本文将采集到的风速、温度、功率三个风机特征的历史数据进行脚本处理，形成 140208*3 特征维度的数据集。

由于数据集存在部分功率缺失，因此需根据工程先验知识计算每 15 分钟间隔下的理论功率，计算逻辑如下：

Ⅰ：根据风机当前空气密度 ρ 寻找风功率厂家提供的 2 条空气密度相邻风速功率曲线，满足 $\rho_1 < \rho < \rho_2$；

Ⅱ：根据 v_f 判定风速区间，在空气密度 ρ_1 和 ρ_2 的风速功率曲线中得到 v_f 所处风速区间的理论功率值 p_1 与 p_2；

Ⅲ：根据如下公式(3-2)所示线性内插算法，计算当前风速 v_f 和空气密度 ρ 对应的理论功率 p_L：

$$P_L = P_1 + \frac{P_2 - P_1}{\rho_2 - \rho_1}\left(\rho - \rho_1\right) \tag{3-2}$$

3.2 缺失值填充方法

数据集的缺失影响模型的训练精度，会导致不可靠的预测结果输出。因此缺失值填充是数据预处理的一个重要的环节。本节详细介绍本论文中所提出的基于工程先验知识和期望最大算法的风功率缺失值填充方法，包括算法流程及各关键步骤的详细说明。

3.2.1 算法流程

首先采用 ST-DBSCAN 算法检测数据集中的异常点，将异常点与功率缺失值点作为需要填充的对象，将剩下的历史数据集进行聚类操作，接着对功率缺失值进行如下步骤填充。首先根据当前功率缺失时刻遥感器测得的风速计算出理论功率，通过风速和理论功率求出聚类中与当前功率缺失时刻最近的簇心，再根据 K 近邻算法寻找簇心附近最近 K 个邻居，接着根据 K 个邻居的功率值求得均值和方差，最后将计算出的均值和方差作为期望最大算法的初始化参数进行迭代，最终收敛的值作为此次要填充的功率数值。具体流程如算法 3-1 所示

算法 *3-1*：基于工程先验知识和期望最大算法的风功率缺失值填充值方法

1: 预处理：通过 ST-DBSCAN 找到故障点

2: K-Means 聚类：

2.1：通过 Inertia 、Silhouette、 Calinski-Harabase 三种评价指标确定 K 值

2.2: 对不包含故障点和缺失点的数据集进行聚类

3: 期望最大算法初始值获取：

3.1：根据风速求得待填充时刻的理论功率

3.2：根据风速和理论功率求最近的簇心

3.3：K 近邻取簇心的最近 K 个邻居

4: 期望最大算法迭代：

4.1：计算邻居节点的 var 和 mean，初始化模型参数 θ 的初始值 θ_0

4.2：for j from 1 to j:

4.3：E 步，计算联合分布的条件概率期望

4.4：M 步，极大化 $L(\theta,\theta_j)$，得到 θ_{j+1}

4.5：如果 θ_{j+1} 收敛，则算法结束，否则继续进行 E 步迭代，直到收敛

5: 输出：填充值

3.2.2 异常数据检测

由于风电机组一般安装在偏远的山地或沿海地区，长期处于气候恶劣的环境，加上随着风机本身服役时间增长，发生故障的概率也在增大，因此采集到的数据包含一定故障数据。过多的故障数据，会导致模型训练的误差急剧累积，降低模型的可信性与准确性，故对异常数据的检测变得尤为重要。

ST-DBSCAN（Spatial Temporal-DBSCAN）由 DeryaBirant[51]提出，是一种基于相似性度量对时空数据进行聚类的算法。相比较传统的 DBSCAN 多出一个维度的聚类，其异常识别方法的思路是以每个点为中心，设定领域及领域内至少需要的点数，若样本点大于指定要求，则认为该点与领域内的点属于同一类，如果小于指定要求，若该点位于其它点的领域内，则属于边界点，其它未分类的点被称为噪声点，也就是不属于任何集群的数据点，作为需要处理的异常点。

根据风机功率的实际呈现出的函数关系，选择对时间、功率和风速进行分析。通过指定半径 eps 和中心领域内最少点的数量 min_samples，获取聚类之后的标签，标签值小于 0 的点为未分类，将筛选出聚类后少数未分类的样本群体作为异常点，这些异常点最终需通过缺失值算法进行填充处理。

3.2.3　聚类最优 K 值选取

聚类中 K 值的取值直接影响到聚类的效果。本小结通过 Inertia、Silhouette 和 Calinski-Harabase 三个聚类评价指标进行聚类效果评估，以确定最优 K 值。

数据集采用上节中真实点。实验k值范围为5-15，通过观察Inertia、Silhouette和Calinski-Harabase三个评价指标曲线，如图3-1，3-2，3-3。通过观察可以看出，Inertia在聚类中心为10的时候趋向收敛，Silhouette在聚类中心为10左右斜率变化最大，Calinski-Harabase值在聚心9到10左右最大，故本文将聚类最优K值定为10。

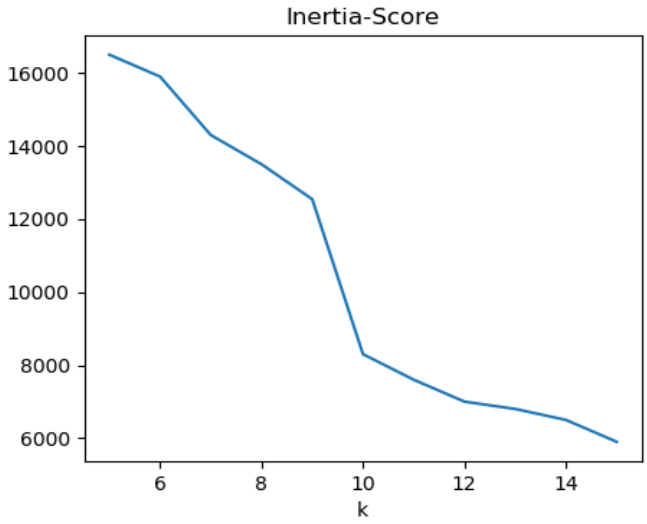

图 3-1 Inertia 评价指标

Figure 3-1 evaluation of Inertia

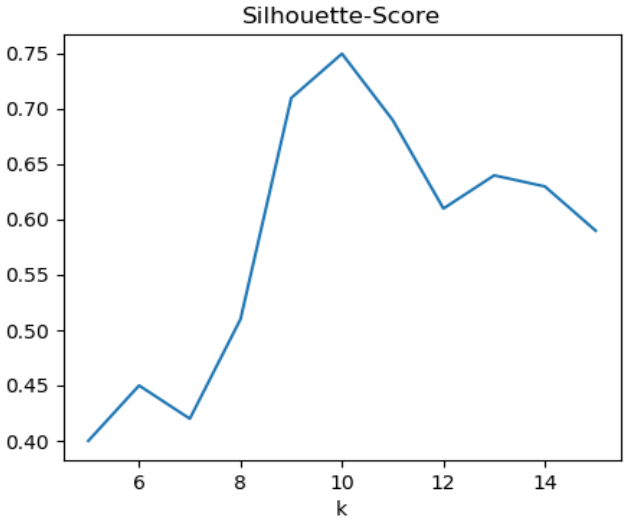

图 3-2 Silhouette 评价指标

Figure 3-2 evaluation of Silhouette

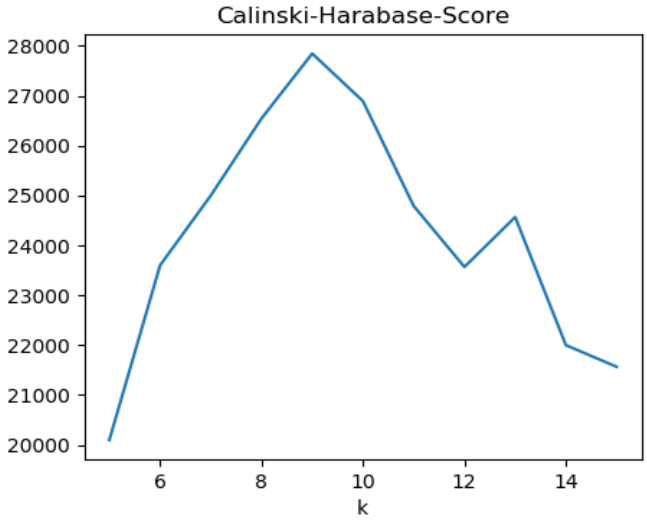

图 3-3 Calinski-Harabase 评价指标图

Figure 3-3 evaluation of Calinski-Harabase

3.2.4 期望最大算法初始化参数计算

在分析出聚类中心数值10为合理的中心数情况下，通过K-Means算法进行聚类，聚类的特征为风速与功率。接着对需要填充的数据依次遍历，由于需要填充的数值是功率，首先计算出当前风速下的理论功率，将理论功率与当前的风速作为距离计算点，寻找聚类中与此点距离最小的簇心，距离采用欧式方法计算，如公式(3-3)所示：

$$|AB|_{\min}=\sqrt{(x_i-x_2)^2+(y_i-y_2)^2}\ (0<i<=k) \tag{3-3}$$

找到最近簇心后，通过K近邻算法，寻得簇心附近最近的K个邻居，求得这些邻居的均值和方差，分别如公式(3-4)和公式(3-5)所示：

$$\mu=\frac{\sum_{N}^{i=1}Xi}{N} \tag{3-4}$$

$$\sigma^2=\frac{\sum_{N}^{i=1}(Xi-\mu)^2}{N} \tag{3-5}$$

将求得的均值和方差作为期望最大算法初始化的参数，进行算法最后的迭代。最终收敛的值作为此次需要填充的功率值。

3.3 实验结果与分析

本节展示了本文所提缺失值算法在数据集上的实验结果，并且以图表的形式

与主流的缺失值算法进行了实验结果对比和分析。

3.3.1 实验环境

本文实验均在 Windows 操作系统下进行，具体硬件环境信息如表 3-1 所示。实验所用语言主要为 Python，实验框架为 PyTorch。后续所有实验均在此环境下进行。

表 3-1 实验配置及环境

Table 3-1 Experimental configuration and environment

名称	版本型号说明
操作系统	Windows10
CPU 型号/内存	Intel i7-9700k/16G
GPU 型号	NVIDIA 2080Ti
CUDA 版本	10.1
Python 版本	3.7
深度学习框架	PyTorch 1.10.1

3.3.2 基准方法

为了验证该模型的有效性，本节使用三种对比方法。基于回归的缺失值填充方法，基于随机森林的缺失值填充方法，基于多重插补的缺失值方法，具体算法填充如下描述：

（1）回归缺失值填充算法

此算法将缺失属性当作因变量，其他关联属性当作自变量，建立属性之间的关系回归模型，通过回归模型进行缺失值插补。

（2）随机森林缺失值填充算法

此算法首先为缺失值预先设定一些估计值，如数值型特征中，选择剩余的平均数或者众数作为当前的预估值，再根据预估值，建立随机森林，将所有数据在随机森林跑一次。在决策树中逐步记录每组数据的分类路径，判断哪组数据与缺失数据的路径最接近，引入相似度矩阵，记录数据间的相似程度，相似度矩阵大小为 N*N。如果是数值型变量，通过加权求平均的方式得到新的预估值，如果缺失值是类别变量，采用权重投票的方式得到新的预估值，依次迭代，最终获得稳定的预估值。

（3）多重插补缺失值填充算法

基于多元回归模型来预测，每个不完整变量由单独的模型估算，进行多重混合评估。mice 算法可以实现对于连续型，二进制，无序分类和有序分类数据的进行混合。此外，mice 包可以处理连续的两级数据，并通过被动插补来保持插补之间的一致性，通过各种统计诊断图，保证插补数据的质量。

3.3.3　对比实验结果与分析

本实验采用 3.1.2 节整理后的数据集中部分数据。将某区域风电场下 1 号风机在 2020 年 7 月 23 日到 7 月 25 日的连续三天完整数据作为本节实验数据，分别在 10%、20%、30%、40%和 50%的随机缺失比例下进行实验，通过 MAE 和 RMSE 评价指标进行实验效果分析，验证基于工程先验知识和期望最大算法的风功率缺失值填充的有效性。实验的评价指标如表 3-2 所示。

表 3-2 各算法在不同缺省值下评价指标

Table 3-2 evaluation of different algorithm under different default values

Percent	**10%**		**20%**		**30%**		**40%**		**50%**	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Regression	20.54	160.53	32.34	190.35	40.23	364.54	45.67	401.34	52.45	413.43
IterForest	18.63	145.42	29.47	184.48	37.65	345.03	41.54	397.28	48.61	389.98
MICE	16.54	133.23	26.31	175.32	32.94	333.23	38.69	374.69	47.75	373.05
Ours	**15.71**	**107.48**	**24.32**	**144.65**	**30.69**	**304.28**	**34.81**	**343.04**	**44.31**	**355.39**

通过分析实验结果，可以看出本文提出的基于工程先验知识和期望最大算法风功率缺失值填充具有明显的优势。与基准模型中较好的多重插补填充算法相比，在 10%缺失比例下，MAE 和 RMSE 分别下降了 5%和 19.3%；在 20%缺失比例下，MAE 和 RMSE 分别下降了 7%和 17.7%；在 30%的缺失比例下，MAE 和 RMSE 分别下降了 9%和 8%；在 40%的缺失比例下，MAE 和 RMSE 分别下降了 10.5%和 8.3%；在 50%的缺失比例下，MAE 和 RMSE 分别下降了 6.4%和 4.9%。综合 MAE 和 RMSE 评价指标，可以证明本文提出的风功率缺失值填充方法的有效性。

通过分析图 3-4 和图 3-5，即使随着缺失比例的增大，本文提出的功率缺失填充的误差也可以在一个工程项目可接受的范围之内，这证明了本文提出的缺失值填充算法的有效性。

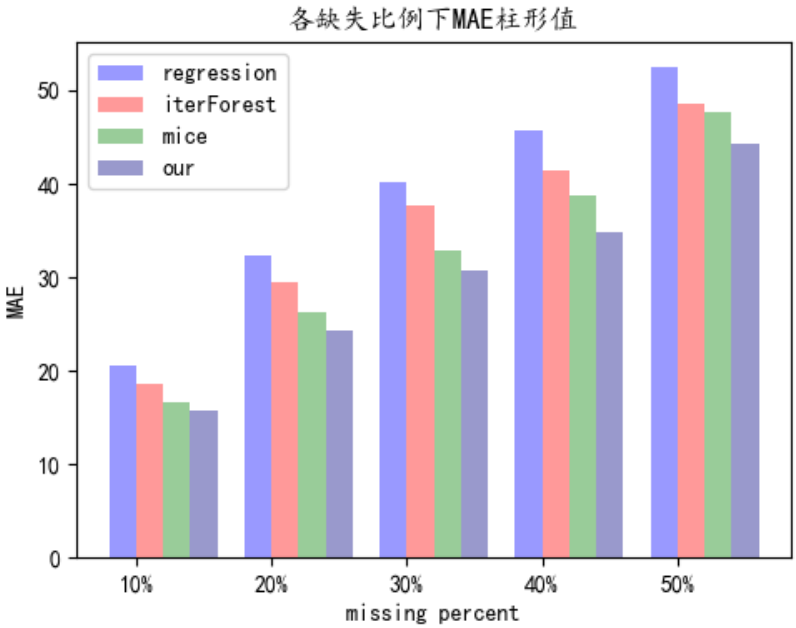

图 3-4 MAE 柱形图

Figure 3-4 MAE histogram

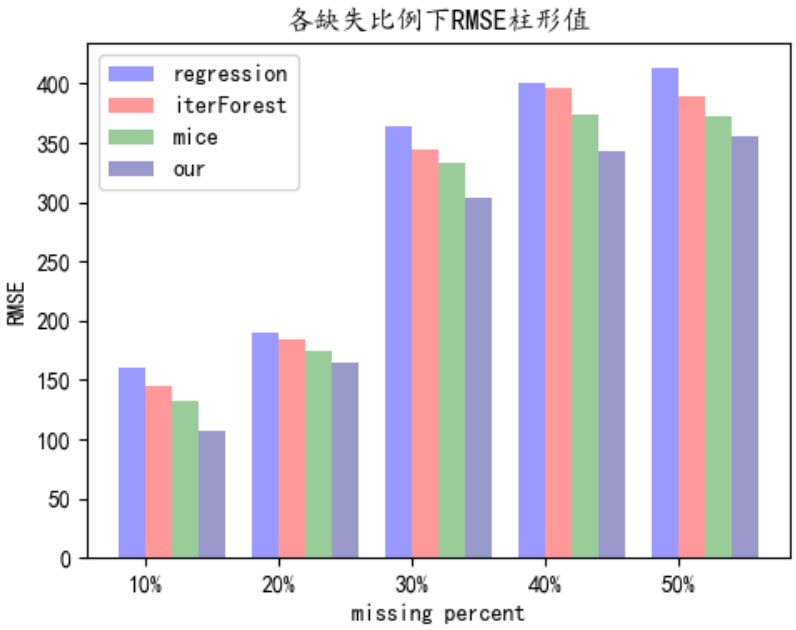

图 3-5 RMSE 柱形图

Figure 3-5 RMSE histogram

3.3.4 消融实验结果与分析

为了进一步验证工程先验知识在填充算法中的有效性，本节进行了充分的消融实验，在无工程先验知识的期望最大算法和结合了工程先验知识的期望最大算

法分别进行了缺失比例为 5%、15%、25%的实验，并且使用 MAE 和 RMSE 分析不同缺失比例下的评价指标。

表 3-3 消融实验评价指标

Table 3-3 evaluation of the ablation experiment

Missing percent	**5%**		**15%**		**25%**	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
期望最大算法	17.27	64.31	23.19	162.43	37.19	307.92
Ours	**14.89**	**56.48**	**20.76**	**131.22**	**31.27**	**275.52**

从表可以看出，在缺失比例 5%、15%和 25%下，结合了工程先验知识的期望最大算法确实在提出的算法中具有提升填充精度的作用。这也充分证明了本文的工程先验知识，即通过理论功率限定期望最大算法初始化参数，确实在提高缺失填充精度上起了一定的积极作用。

3.4 本章小结

本章节介绍了结合工程先验知识和期望最大算法的阐述。首先介绍了前期数据的相关工作，包括数据采集和数据整理；接着详细地介绍了填充方法的各步骤流程，包括通过实验确定聚类中的最优 K 值，从而结合理论功率限制期望最大初始化迭代的参数；之后在数据集上，通过随机缺失比例为 10%、20%、30%、40%和 50%下进行了对比实验，对比的基准方法为回归填充法、随机森林填充法和多重插补填充法，对比评价指标是 MAE 和 RMSE，最终证明了本文提出的基于工程先验知识风功率缺失值填充方法的有效性。同时，为了证明工程先验知识在填充过程中起到了积极的影响，本节进行了消融实验，对比无工程先验知识期望最大填充算法和结合了工程先验知识的期望最大填充算法，证明结合了工程先验知识确实可以提高预测的精度。

4　基于 Informer 和图卷积的组合模型风功率预测算法

第三章对缺失值的填充为本章的模型实验的数据集提供了良好的质量保证。本章针对中长期风功率预测进行研究，此研究在深度学习领域的研究较少，有很大的研究空间。本章使用图卷积神经网络进行风机功率之间的空间特征的提取，使用 Informer 模型进行功率时间特征的提取，再通过组合预测模型将时空特征进行融合，最终进行中长期风电功率的预测。

4.1 相关性分析

风机功率预测是一个典型的时序预测问题，但是根据实际工程知识，风具有延时的性质，较近距离风机风的传播速度更快，风速更加接近，同时风机本身转速产生的风速及对附近气候的影响也会对邻近风机功率产生一定的空间影响。因此理论上风机功率之间是具有空间相关性。

为了验证风功率之间的时空相关性，需要借助时空分析的方法对风机功率进行量化分析。本节从时间、空间、时空三个角度对风机功率进行深入的相关性分析，并且引入了一些量化的指标及方法，用来验证风机功率在时间、空间和时空上的基本规律和特征。

4.1.1 相关性指标

本节采用 Pearson 相关系数、ACF 自相关性系数、MIC 最大互信息和 CCF 互相关系数[52]对风功率进行时空特征分析。使用 Pearson 相关系数和 MIC 系数对风功率的空间相关性进行分析，使用 ACF 和 CCF 对风功率时间序列进行量化分析。

（1） Pearson 相关系数

用于度量两个变量之间的相关性大小，值介于-1 和 1，计算公式如下：

$$\rho(X,Y)=\frac{E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]}{\sigma_X\sigma_Y}=\frac{E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]}{\sqrt{\sum_{n}^{i=1}\left(X_i-\mu_X\right)^2}\sqrt{\sum_{n}^{i=1}\left(Y_i-\mu_Y\right)^2}} \tag{4-1}$$

（2） ACF 自相关性系数

用来描述数据自身不同时期的相关程度，即度量历史数据对现在产生的影响，若某风机功率的时间序列为 $X=[x_1,x_2,\ldots,x_T]$，k 为时延阶数，T 为时间序列长度，ACF 自相关性系数公式如下：

$$C_{ACF}^{k}=\frac{\sum_{T-k}^{t}\left(x_t-\bar{x}\right)\left(x_{t+k}-\bar{x}\right)}{\sqrt{\sum_{T-k}^{t=1}\left(x_t-\bar{x}\right)^2\sum_{T-k}^{t=1}\left(x_{t+k}-\bar{x}\right)^2}} \tag{4-2}$$

（3） MIC 最大互信息

衡量两个变量之间的线性或者非线性程度，具有普适性、公平性和对称性，计算公式如下：

$$I(x;y)=\int p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \tag{4-3}$$

$$mic(x;y)=\max_{a*b<B}\frac{I(x;y)}{\log_2\min(a,b)} \tag{4-4}$$

其中 $I(x;y)$ 为联合概率密度分布， $p(x)$ 和 $p(y)$ 为 X 和 Y 的边缘分布。

（4） CCF 互相关系数

CCF 互相关性系数对风机之间的空间相关性进行验证，假设两个风机的功率时间序列为 $X=[x_1,x_2,\ldots,x_T]$ 和 $\boldsymbol{Y}=[y_1,y_2,\ldots,y_T]$，$\tau$ 代表时延阶数，互相关系数计算公式如下：

$$c=\frac{\sum_{T-\tau}^{t=1}\left(x_t-\bar{x}\right)\left(y_{t+\tau}-\bar{y}\right)}{\sqrt{\sum_{T}^{t=1}\left(x_t-\bar{x}\right)^2\sum_{T}^{t=1}\left(y_t-\bar{y}\right)^2}}\tau\in(0,1,2,\ldots,2T-1) \tag{4-5}$$

4.1.2 时间相关性

如图 4-1，对 1 号风机、2 号风机在 2021 年下不同时延阶数下的 ACF 自相关性系数进行计算，横坐标为时延阶数，纵坐标为自相关系数。

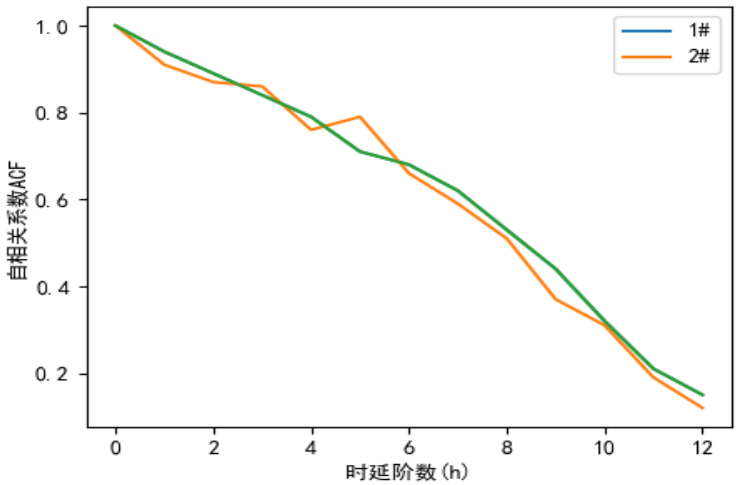

图 4-1 自相关系数图

Figure 4-1 Image of Autocorrelation coefficient

观察图 4-1 可以发现，两台风机的功率自相关曲线趋势接近，在时延阶数小于 6 时，两台风机的时间序列显著相关，相关性较强，而随着时延阶数的增大，两台风机的自相关性逐渐减弱。因此可以得出结论，风功率历史数据具有很强的时间相关性。

4.1.3 空间相关性

以 1 号风机为例，根据 2021 年 1 号风机与 2 号至 6 号风机的功率计算 Pearson 和 MIC 相关性系数，绘制相关性系数柱形图，如图 4-2 所示：

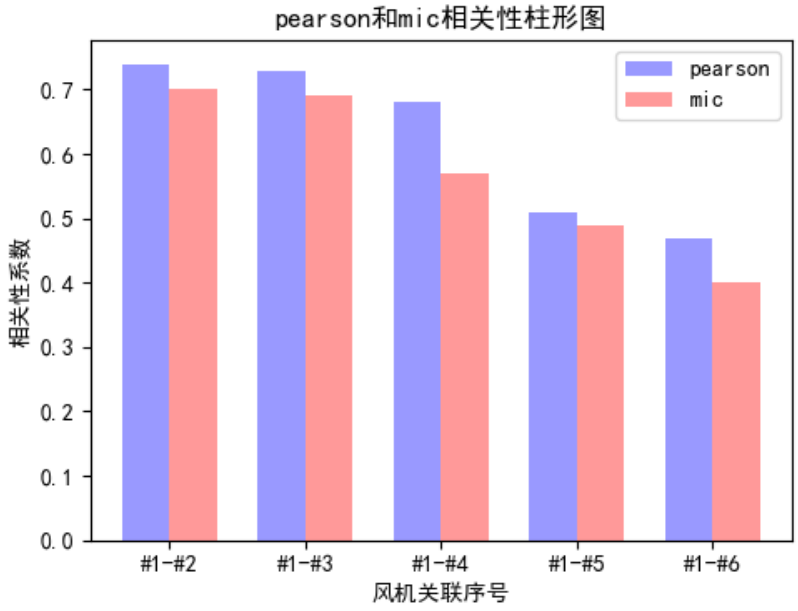

图 4-2 Pearson 和 MIC 自相关系数图

Figure 4-2 Image of Pearson and MIC autocorrelation coefficient

观察图 4-2 可以发现，1 号风机与 2 号、3 号和 4 号风机的 Pearson 和 MIC 相关性系数都大于 0.5，存在一定的空间相关性。而与距离较远的 5 号和 6 号风机，相关性系数小于 0.5，空间相关性系数较弱。

4.1.4 时空相关性

在以上两节对时间和空间相关性分析基础上，本节对风功率的时空相关性进行研究分析，采用 CCF 互相关性系数。数据集使用 2021 年风机功率数据，分别分析 1 号风机与 2 号至 6 号风机之间的时空相关性，并绘制 CCF 对比图并分析。CCF 对比如图 4-3 所示。

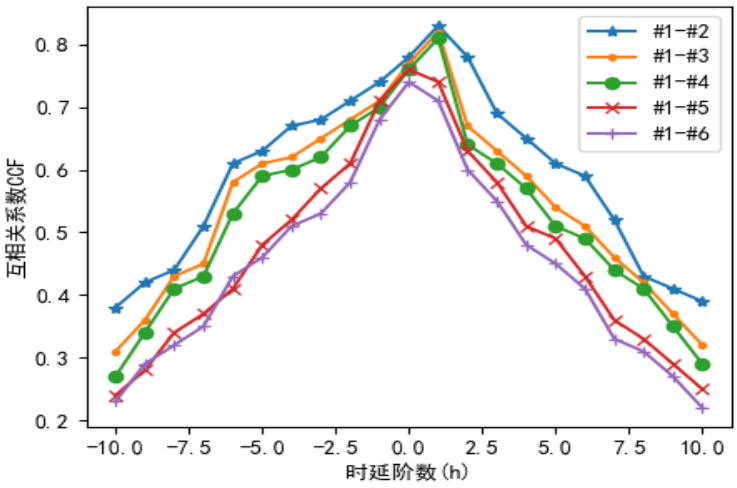

图 4-3 CCF 对比图

Figure 4-3 Image of CCF

通过 CCF 对比图可以看出，1 号与距离较近的 2 号、3 号、4 号风机的互相关性最强，与距离较远的 5 号和 6 号风机互相关性较差。当时延为 0 时，1 号-2 号风机互相关性高达 0.83，1 号-3 号风机互相关性达到 0.81，1 号-4 号风机达到 0.78，而与空间位置较远的 5 号和 6 号风机，互相关性只能达到 0.75 和 0.64。这是由于距离较近的风机有更加相似的地理和气候条件，在空间上有正相关性。

同时从时延阶数的变化来看，随着时延阶数的增大，相关性系数也随之下降，且空间相关性较大的风机之间下降的速度更快，空间相关性较小的风机之间下降的速度相对缓慢。这是因为风具有延时性的特征，距离较近的风机风的传播速度更快，风速更加接近，时空特征的耦合性更加显著。

综合以上时间、空间、时空的特征分析，可以得出结论，整个风功率预测是一个时空预测的问题，如图 4-4 所示。

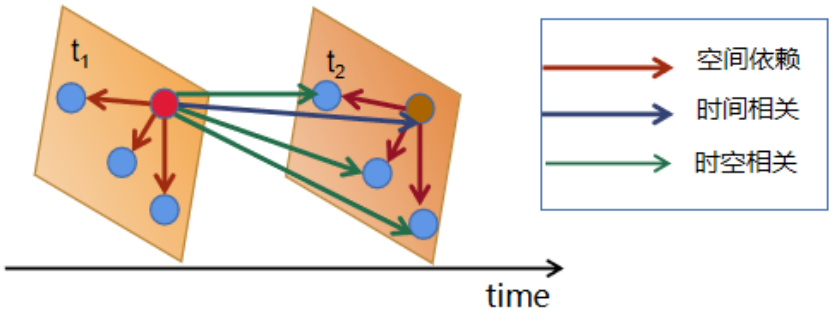

图 4-4 时空相关性图

Figure 4-4 Images of Spatiotemporal correlation graph

4.2 问题分析及模型介绍

本节对风功率预测问题进行详细地分析，指出预测困难点所在，针对难点提出对应的解决方案，最后对本文的预测模型进行详细描述。

4.2.1 问题分析

以往的风功率预测研究，大都在时间序列上进行建模分析，仅仅考虑了风功率时间依赖关系，而少有考虑功率数据之间的空间依赖关系。其它领域中，在时间维度上一些学者使用门控循环单元网络、基于注意力机制的 Transformer 等模型捕获时间特征关系；在空间维度上一些学者使用卷积循环网络来捕获节点的空间关系；在时空维度上一些学者使用时间图卷积等模型来同时捕获节点的时间和空间关系。尽管可以将现有方法运用到风功率预测这项工作中，但是依然有两个重大挑战需要解决。

（1）多步预测的误差损失。在不同的时间步骤会产生不同的时间依赖性，通过在历史的时间序列数据的基础上，考虑使用已有的预测结果再进行多步的预测会产生误差累积的问题，使得之前的预测的误差会传播到下一步，降低预测的性能和准确率。

（2）复杂的空间相关性。风电场各风机之间的功率存在一定的空间相关性，不同的风机之间功率会有一定的影响，且具有很强的动态性，如何有效地捕获空间维度中各节点间的关联关系，是预测过程中具有挑战性的问题。

根据以上问题分析，本文提出了一种基于 Informer 与图卷积的组合模型风功率预测算法。对于时间依赖关系建模，使用 Informer 模型，通过模型的 ProbSparse Self-attention 机制减少时间复杂度和内存上的开销，使用自注意力蒸馏机制，来高效地处理过长的输入序列，通过生成式的解码器，能够一次向前操作而不是分步操作的方式预测长时间序列，极大地提高长序列的预测速度。对于复杂空间依赖关系，从图的角度出发，通过空间相关性分析构建邻接矩阵，通过风机功率特征值构建特征矩阵，输入到两层图卷积模型中从而提取复杂空间特征。最后通过组合模型对学习到的时空特征进行融合，进行某段时间的风功率中长期预测。

4.2.2 模型框架

本文的风功率预测模型是利用 Informer 和图卷积分别学习时间和空间上关系，再通过组合模型进行时空融合进行预测。为了便于描述，本文以 1 号风机为例，预

测 1 号风机未来 p 时刻的功率值。预测模型如图 4-5 所示，整体流程如下：

（1） 根据 4.1.3 的空间相关性分析方法，计算 1 号风机与其他风机的空间关联度$\{R_1,\cdots,R_2,R_t\}$，若空间具有相关性，根据相关性系数值设置连接权值，若无相关性，连接权重值为 0，从而构建带权重的风机连接矩阵 A 。根据风机的功率数据，构建特征矩阵 X，描述了各风机功率随时间的变化，每一行代表的是一台风机，每一列代表不同风机上某一时刻的功率值，时间间隔为 15min。

（2） 通过增加自循环处理后得到的矩阵$\hat{A}$和 1 号风机功率特征矩阵 X 输入两层图卷积网络，学习空间依赖特征信息，进行未来 p 时刻的预测输出 $y_S^t=\left\{y_1^t,\cdots,y_p^t|\ y_i^t\in R^{d_y}\right\}$。

（3） 通过构建风机的长时间序列特征数据输入到 Informer 模型中，学习数据时间依赖特征信息，计算得到预测的相对应未来 p 个时刻序列数据 $y_T^t=\left\{y_1^t,\cdots,y_p^t|\ y_i^t\in R^{d_y}\right\}$。

（4） 通过 2.4.1 节中介绍的组合预测法，分别计算图卷积和 Informer 的组合权重系数。权重系数记为 w_1 和 w_2。

（5） 根据计算好的组合权重，输出未来 p 时刻的预测值 $y^t=w_1y_S^t+w_2y_T^t$。

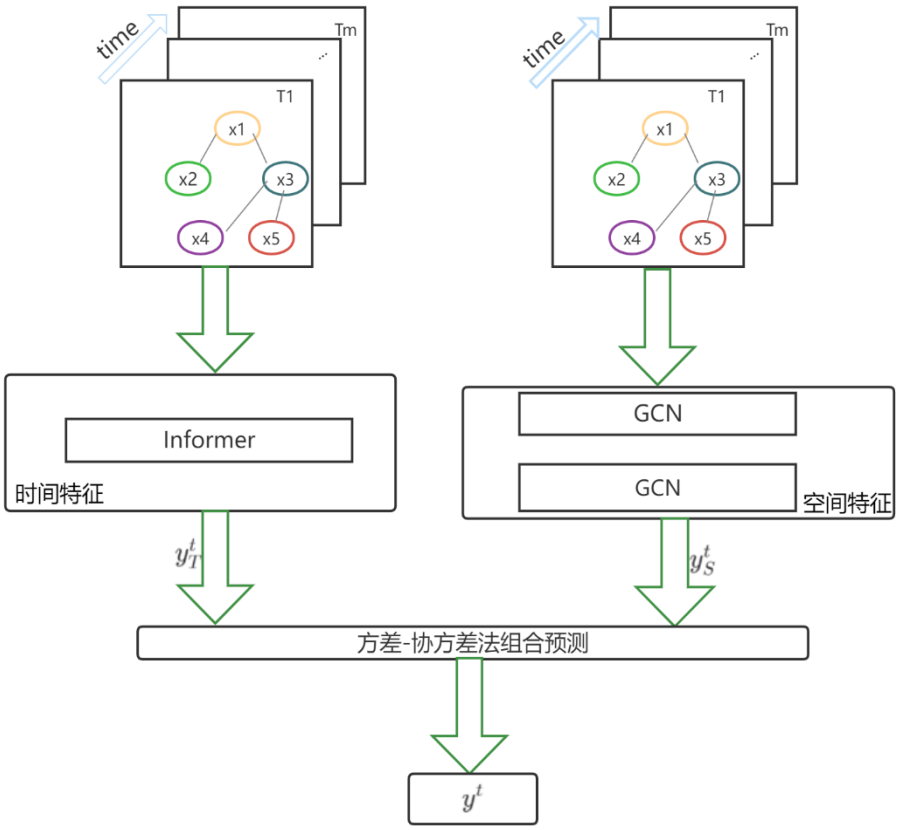

图 4-5 风功率组合预测模型图

Figure 4-5 Images of wind power combination prediction model

4.3 实验设置与分析

本节展示了本文所提风功率预测算法在数据集上的实验结果，并且与主流的算法进行了实验结果对比和分析。

4.3.1 数据集介绍及实验设置

本节数据集采用 3.1.2 节中整理之后的数据集，并对数据集进行了缺失值填充处理。数据集包含 2018-2021 年共 48 个月各风机每隔 15min 时序数据，按照 42:3:3 的比例划分训练集、验证集和测试集，也就是将 2018 年 1 月-2021 年 6 月共 42 个月数据作为训练集，将 2021 年 7 月-2021 年 9 月共 3 个月数据作为验证集，将 2021 年 10 月-2021 年 12 月共 3 个月数据作为测试集。

其中训练集用于模型训练；设置的验证集可以在模型训练的过程中及时发现模型是否存在问题出现异常等，验证模型的泛化能力，检测模型是否出现过拟合；设置测试集用于评估模型的最终的泛化能力。

Informer 在训练过程中，数据集的输入都是采用 Min-Max 归一化方法对数据进行转换，并且最后对预测值进行反归一化操作。Informer 在编码器中包含一个三层堆栈和一个一层的堆栈（1/4 输入），以及包含一个 2 层的解码器，通过 Adam 进行优化，gelu 作为激活函数，初始学习率设置为 1e-4，epoch 设置为 8，batch_size 设置为 64。

图卷积的训练过程中，使用两个叠层的图卷积层作为编解码单元，解码器的输出层使用 tanh 作为激活函数。训练阶段，bathc_size 设置为 64，学习率初始设置为 0.001，学习的参数采用均匀分布初始化，使用 Adma 进行优化。

评价指标采用均方根误差（RMSE）、平均绝对误差（MAE）、拟合优度（R^2）。

4.3.2 基准方法

本文的基准方法包括 Transformer、TCN 和 T-GCN 进行对比分析，通过四组不同的实验验证本文所提出组合模型预测算法的有效性。

（1）Transformer：摒弃了传统的 CNN 和 RNN，整个网络由 self-Attention 和 Feed Forward Neural Network 组成，一个基于 Transformer 的神经网络网络可通过堆叠多个 Transformer 的形式进行构建。在时序数据预测中，可以利用 self-Attention 机制从时间序列数据学习复杂的模式和形态，可以应用于单变量和多变量的时序数据预测。

（2）TCN：该网络模型以 CNN 为基础，是一种利用因果卷积和空洞卷积的神经网络模型，可以适应时序数据的时序性并可以提供视野域用于时序数据建模，具有稳定的梯度和较低的内存消耗的特点。

（3）T-GCN： 首先在交通预测中提出，由图卷积网络和门控单元两部分组成，使用历史时序数据作为模型输入，利用图卷积捕获网络拓扑结果，进行空间特征捕获，再将含有空间特征的时间序列数据输入门控递归单元捕获时间特征，最终，通过全连接层得到预测结果。

4.3.3 对比实验结果与分析

图 4-6 和图 4-7 为 1 号风机和 2 号风机在一段时间的预测结果，使用本文提出的 Informer 和图卷积的组合模型风功率预测算法与基准模型 Transformer、TCN 和 T-GCN 进行预测效果对比。

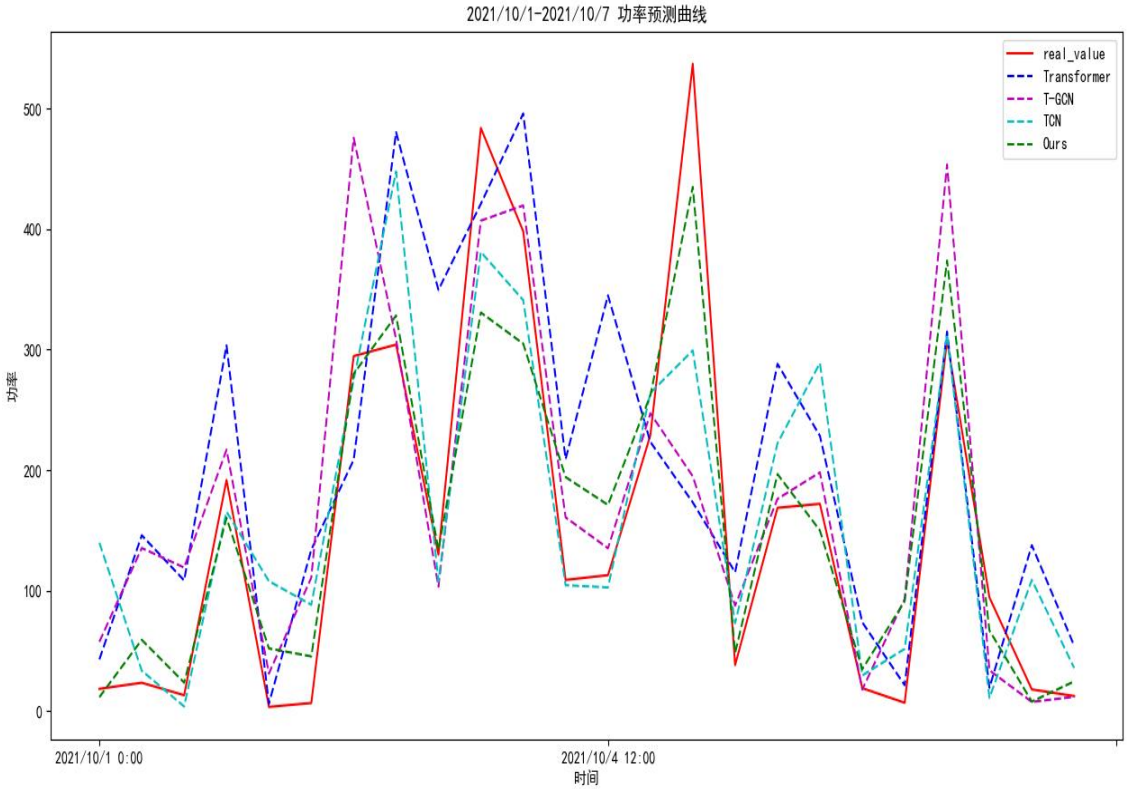

图 4-6 1 号风机预测图

Figure 4-6 Images of prediction of fan one

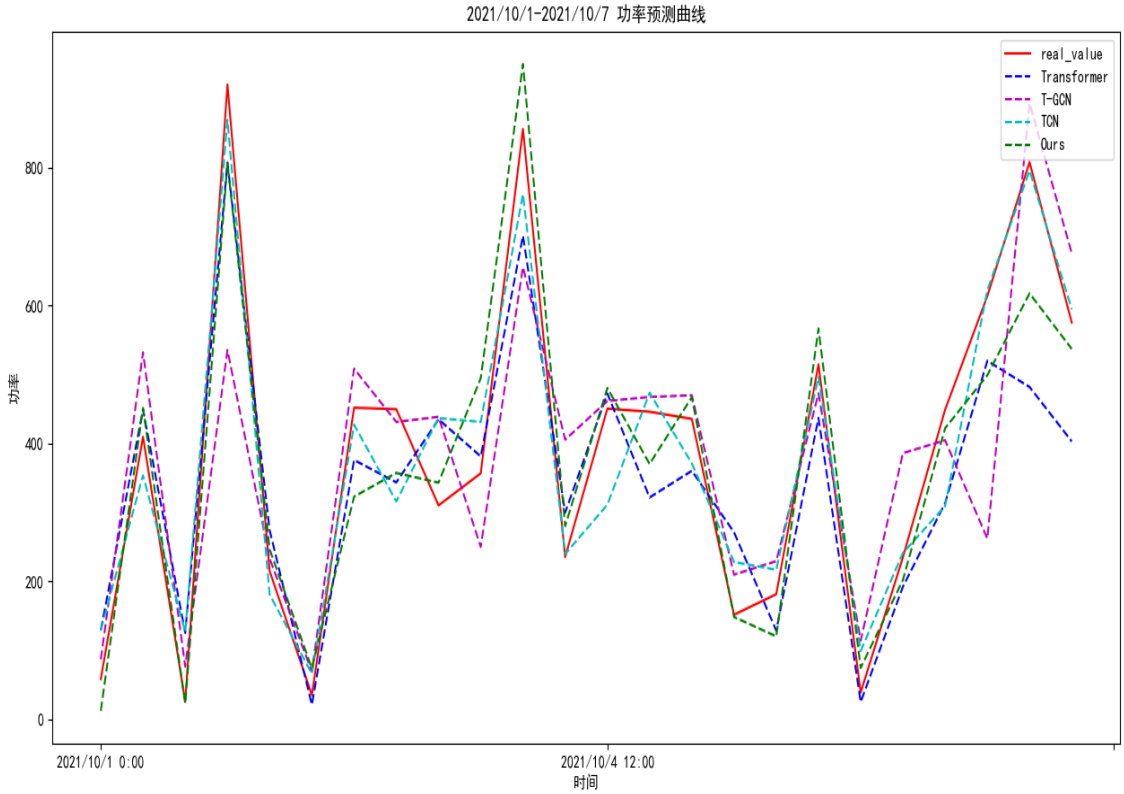

图 4-7 2 号风机预测图

Figure 4-7 Images of prediction of fan two

通过对比曲线可以看出，不管是 1 号风机还是 2 号风机，在对指定时间的一周风功率进行预测时，基于本文提出的模型风机功率曲线图趋势更加接近真实风机功率曲线趋势，同时本模型的预测值和真实值之间误差较其他基准模型也更低。因此本文所提出的模型相比较于基准模型，预测趋势更加符合真实的曲线情况，预测效果更好。

下表 4-1 和表 4-2 分别为 1 号风机和 2 号风机的各模型评价指标。

表 4-1 1 号风机评价指标

Table 4-1 evaluation of fan one

Model	RMSE	MAE	R^2
Transformer	230.1	132.4	0.56
TCN	196.7	128.9	0.63
T-GCN	146.4	117.3	0.67
Ours	132.5	109.5	0.75

表 4-2 2 号风机评价指标

Table 4-2 evaluation of fan two

Model	RMSE	MAE	R^2
Transformer	223.5	129.1	0.57
TCN	181.5	120.5	0.62
T-GCN	143.2	114.7	0.69
Ours	129.1	101.3	0.71

通过分析，本文提出的组合预测模型相对于预测效果较好的 T-GCN 基准模型，在 1 号风机上，RMSE 和 MAE 分别降低了 9.6%和 7.1%；在 2 号风机上，RMSE 和 MAE 分别降低了 9.8%和 11.4%，并且预测模型的拟合效果要优于基准模型。因此本文提出的模型预测算法预测效果较好。

4.3.4 消融实验结果与分析

为了充分验证融合了空间特征的图卷积模块确实在组合预测中起到了一定作用，本节进行了充分的消融实验，实验结果如图 4-8 和图 4-9 所示。

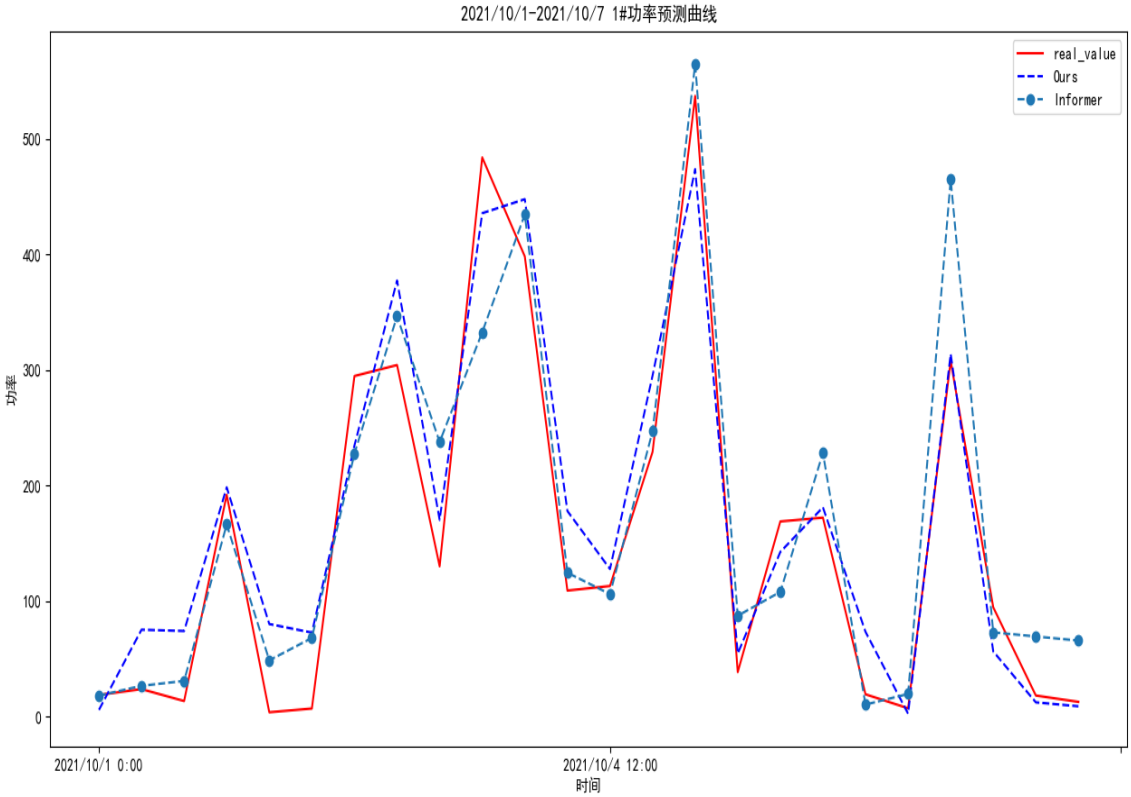

图 4-8 1 号风机消融实验

Figure 4-8 Images of ablation experiment of fan one

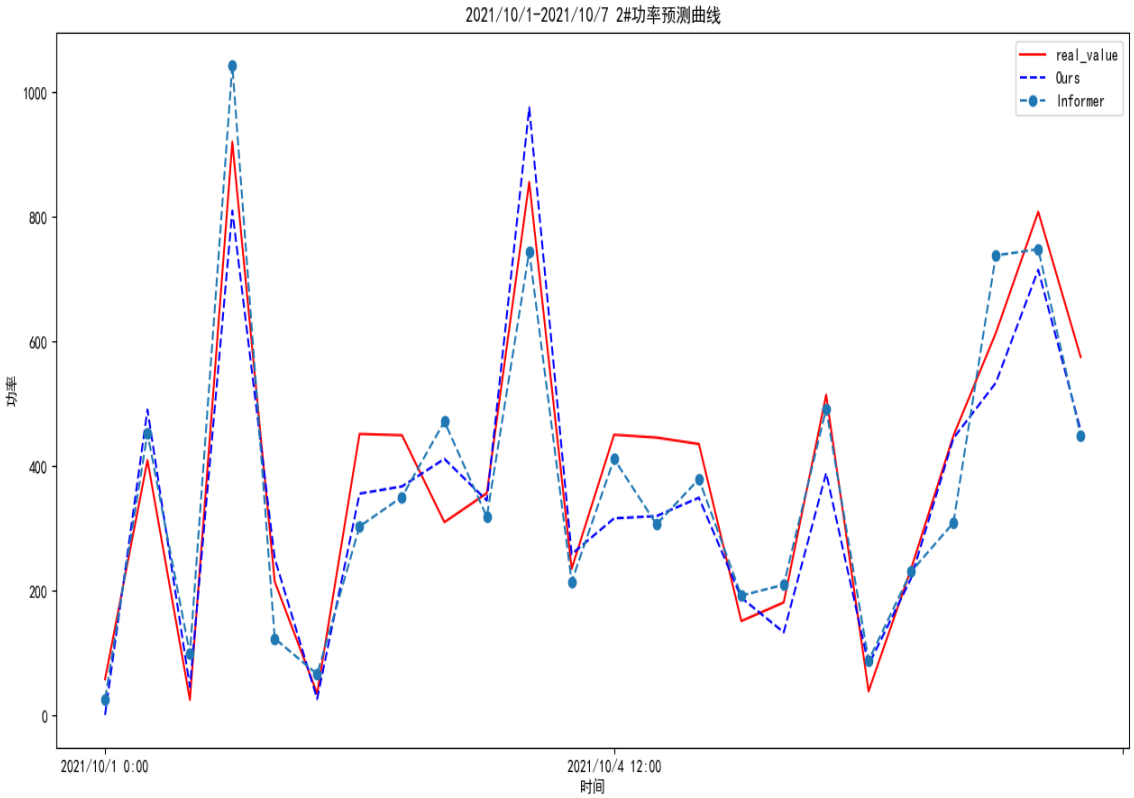

图 4-9 2 号风机消融实验

Figure 4-9 Images of ablation experiment of fan two

下表 4-3 和 4-4 分别为消融实验后的 Informer 和 Informer+图卷积的评价指标。

表 4-3 1 号风机消融实验指标

Table 4-3 evaluation of ablation experiment of fan one

Model	RMSE	MAE	R^2
Informer	147.4	114.7	0.71
Ours	134.3	109.5	0.75

表 4-4 2 号风机消融实验评价指标

Table 4-4 evaluation of ablation experiment of fan two

Model	RMSE	MAE	R^2
Informer	141.3	104.3	0.69
Ours	128.4	98.5	0.72

通过消融实验结果可以发现，在 1 号风机上，RMSE 和 MAE 分别降低了 6.1%和 4.4%；在 2 号风机上，RMSE 和 MAE 分别降低了 9.2%和 5.8%。这表明通过图

卷积层融入了风机功率之间的空间特征关系进行预测确实可以有效提高预测的精度，最终证明了本文提出的 Informer 和图卷积的组合模型风功率预测算法的有效性。

4.4 本章小结

本章基于 3.1.2 节整理后的数据集开展工作。首先对数据集的时空相关性进行了分析，证明了风机之间的功率确实存在空间相关性。接着对时间和空间上预测遇到的问题进行分析，时间维度预测上存在多步预测的误差损失，误差的累积导致精度的急剧下降，空间维度预测上难以有效捕获空间维度中各节点间的关联关系。然后，针对提出的问题，提出了基于 Informer 和图卷积的组合模型，分别对时间和空间上进行特征提取进行预测，再进行组合预测。最后，介绍了此次实验的数据集，基于实际的 1 号风机和 2 号风机的历史数据进行了预测实验和一些基准实验，并在评价指标上进行了对比，证明本文提出的预测模型在评价指标上要优于基准实验模型。同时本章节进行了消融实验，证明本文提出的模型中的图卷积层确实对于模型空间特征的提取起到了一定作用，可以有效地提高预测的精度。

5 风功率预测可视化平台的设计和实现

本章详细介绍了风功率可视化平台的实现。将本文提出的基于工程先验知识和期望最大算法的风功率缺失值填充方法及基于 Informer 和图卷积的组合模型风功率预测算法进行了实际的项目工程应用。本平台提供了风电场运营的重要实时指标和统计指标查询和展示功能，以及风机中长期功率预测，并使用曲线图、表格等形式进行数据可视化展示。同时，通过容器化技术方便平台的部署和运维。此平台方便了运营人员的远程监控和发电计划等工作安排。

5.1 相关工作

本节首先针对风功率预测可视化平台进行了现状分析，然后根据实际的研究内容和业务场景需求，介绍了风功率预测可视化平台的主要技术方案。

5.1.1 风功率预测可视化平台现状分析

风功率预测可视化平台的发展对风力发电的建设起到积极的推进作用，可以进一步提高生产管理水平，对风电企业的健康发展起到良好的保证。结合当下的技术发展和风电场以往的工作经验，对风功率预测可视化平台做如下分析：

（1）目前较多的风功率预测平台部署于 WindowsServer 服务器，受制于操作系统，开源组件较少，开源技术社区活跃度低，技术沉淀相对较弱，平台的可移植性、跨平台性较差，不适用于目前的国产化转型趋势。

（2）传统的部署方式单一，部署步骤繁琐，配置分散各处，缺少统一配置中心管理，给现场实施运维人员带来较多的工作量。且平台大部分基于单体应用研发，后期业务不断迭代会导致系统的复杂性越来越高，后期平台的扩展性和可靠性较差，单个模块异常可能会导致整个平台服务不可用。

（3）风电场的风机数据零散且数据庞大，现场工作人员只能通过数据库客户端管理工具进行实时数据和历史数据的查看，缺少统一的风机指标查询入口，缺少数据的冗余备份和数据恢复功能。

（4）风功率预测大部分还是基于传统的物理方法、统计方法或者专家预测法，且预测有延迟，预测时效性不强。传统的风功率缺失数据填充方法较多采用的是线性插值，填充效果有一定偏差，且较少有将深度学习模型应用到风功率预测工作。

针对风功率预测可视化平台发展的不足，本章旨在采用主流的技术路线开发

一套可跨平台、扩展性强、7*24h 持续稳定运行、部署简易、配置中心化的平台，将本论文提到的风功率缺失值处理方法和风功率预测模型集成到平台中，为风电场的工作人员的决策分析提供可靠的辅助支持。

5.1.2 开发技术方案

针对 5.1.1 小节的风功率预测平台现状，进行技术选型和分析，下面对本平台技术做详细的介绍。

风功率预测可视化平台使用主流的前后端开发技术。前端采用 VUE 框架，页面设计采用 Element-UI 组件库，并且结合 Echats 类库对数据进行可视化展示。后端采用 Java 语言，基于 SpringCloud 微服务架构研发，采用 Docker 容器化对项目进行部署，采用 spark 分布式技术组件进行模型的训练，通过 yarn 实现 spark 的资源调度。下面对 SpringCloud 微服务、Docker 容器化和 spark 技术进行介绍。

（1） SpringCloud 微服务

微服务[53]是一种软件技术架构，是 SOA 架构的变体。将传统单体应用的各模块进行解耦，形成一组小服务，服务之间通过 http 或 rpc 方式进行协调配合，每个服务都运行在单独的进程中。每个服务根据具体的业务进行开发构建，不限开发语言、编译工具、关系数据库等，支持独立部署到生产环境中。微服务架构使得应用程序更加易于扩展和开发迭代。

（2） Docker

Docker[54]是 Go 语言开发的虚拟化应用容器引擎。各容器进程相互隔离，容器性能开销极低。Docker 特点和优点是持续支付和部署、一致的运行环境、更加快速的启动、更加高效的利用系统资源、更便捷的迁移和更轻松的扩展和维护。更重要的是，Docker 的发展完善了微服务架构，使得微服务的组件方便管理、扩容。

（3） Spark

Spark[55]是一个开源的分布式运算框架。它提供了 Java、Python 等高级开发语言 API，该计算引擎支持用于数据分析的通用计算图，并且它还支持一系列丰富的高级工具，包括 SQL 处理结构化数据处理的 Spark SQL、用于机器学习的 MLlib 库、用于图计算处理的 GraphX 等。Spark 是基于内存进行计算，因此处理速度和效率更加高效。Meng 等[56]对 Spark 在机器学习中的应用进行了深入研究。

5.2 风功率预测可视化平台设计

本节首先对风功率预测可视化平台架构进行设计和阐述，接着根据业务需求进行功能模块的设计和业务的开发流程制定。

5.2.1 系统架构设计

风功率预测可视化平台的系统架构设计图如图 5-1 所示。

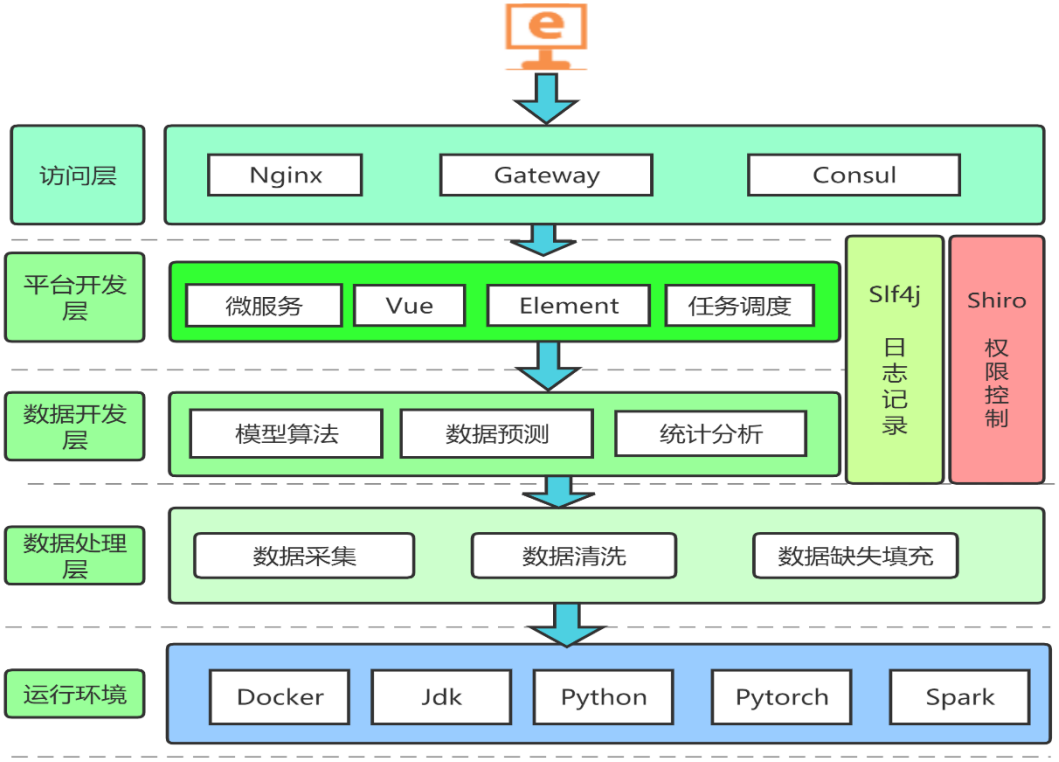

图 5-1 系统架构设计图

Figure 5-1 Image of platform architecture

运行环境：本平台稳定运行在 x86 架构下的 centos7.6 操作系统中。

数据处理层：数据治理层包含数据抽取、数据清洗及基于工程先验知识和期望最大算法的风功率缺失值填充，为后续模型训练提供可靠数据支撑。

数据开发层：数据开发层包含模型算法训练和功率数据预测。模型算法集成 Transformer、TCN、T-GCN 和本文提出的基于 Informer 和图卷积的组合模型风功率预测算法。模型的训练和功率的预测通过任务调度中的定时任务执行，通过 Spark 分布式计算提交到 Yarn 上执行，使用 Yarn 实现资源的调度和管理，提高预测模型的速度和时效性。

平台开发层：平台开发层将现有数据与业务进行结合，提供用户具体的平台服务，将数据开发层的数据进行统计分析，将实时采集数据进行展示，最终将风电场和风机的实时、历史指标，功率预测曲线通过 web 可视化形式展示。

访问层：风功率预测平台必须保证数据的一致性，需满足微服务 CP 定理，故采用 Consul 作为注册中心。由于组件多节点部署，需要统一的网关入口，同时在网关处进行用户的认证和鉴权，故采用原生 GateWay 作为网关。Nginx 作为方向代理服务器，负责请求转发到网关，屏蔽实际后台地址，保证访问安全性。

平台部署方面，通过将数据预处理、用户管理、角色管理和任务调度组件打包成 Docker 镜像，以 Docker 容器进行部署运行，相较传统的风功率预测平台的繁琐部署流程，本平台部署步骤更加简化。

相比较其它风功率预测平台，本平台技术路线更加多样和先进，安全性更强，部署更加便捷，后期开发扩展性更强。同时将本文提出的风功率缺失值填充方法和基于 Informer 和图卷积的组合模型风功率预测算法在平台进行了实际的应用。

5.2.2 平台功能设计

本节根据风功率预测可视化平台需求进行功能模块设计，如图 5-2 所示。

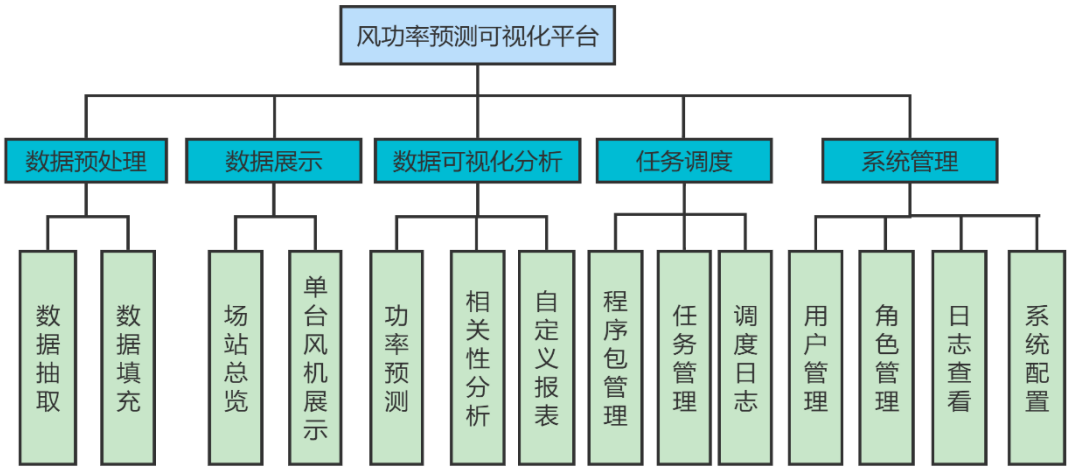

图 5-2 平台功能设计图

Figure 5-2 Images of platform functional design

（1）数据预处理模块：该模块主要由数据抽取、数据填充构成。数据抽取模块为系统提供基础数据支撑，通过任务调度中心调度任务从时序数据库中抽取风电场风机相关指标实时值。数据填充模块首先对抽取的数据通过 3.1.2 节所提的 ST-DBSCAN 算法进行实现异常数据的检测，再将故障值和缺失值采用本文提出的基于工程先验知识和期望最大算法的风功率缺失值填充方法进行数据填充。

（2）数据展示模块：该模块实现包含场站总览和单台风机实时指标和统计指标的展示，同时将本文所提出的风功率预测方法进行预测的值进行可视化展示。

（3）数据可视化分析模块：该模块由功率预测、自定义报表和相关性分析构成。功率预测页面用户可选择展示指定风机的不同时间段的功率预测曲线，自定义报表查询页面可以对指定时间范围下的风机指标进行查询和下载，相关性分析查询不同风机功率之间的时间、空间和时空相关性关系图。

（4）任务调度模块：该模块主要管理本文中所提到的定时任务。通过任务调度中心进行定时任务的调度和管理。

（5）系统管理模块：该模块主基于 RBAC 模型进行认证和鉴权的开发。并且提供数据异常条件配置、机组模型实验服务器节点选择及预测模型选择的功能。

结合如上描述的平台功能设计，进行如图 5-3 的平台核心业务开发。

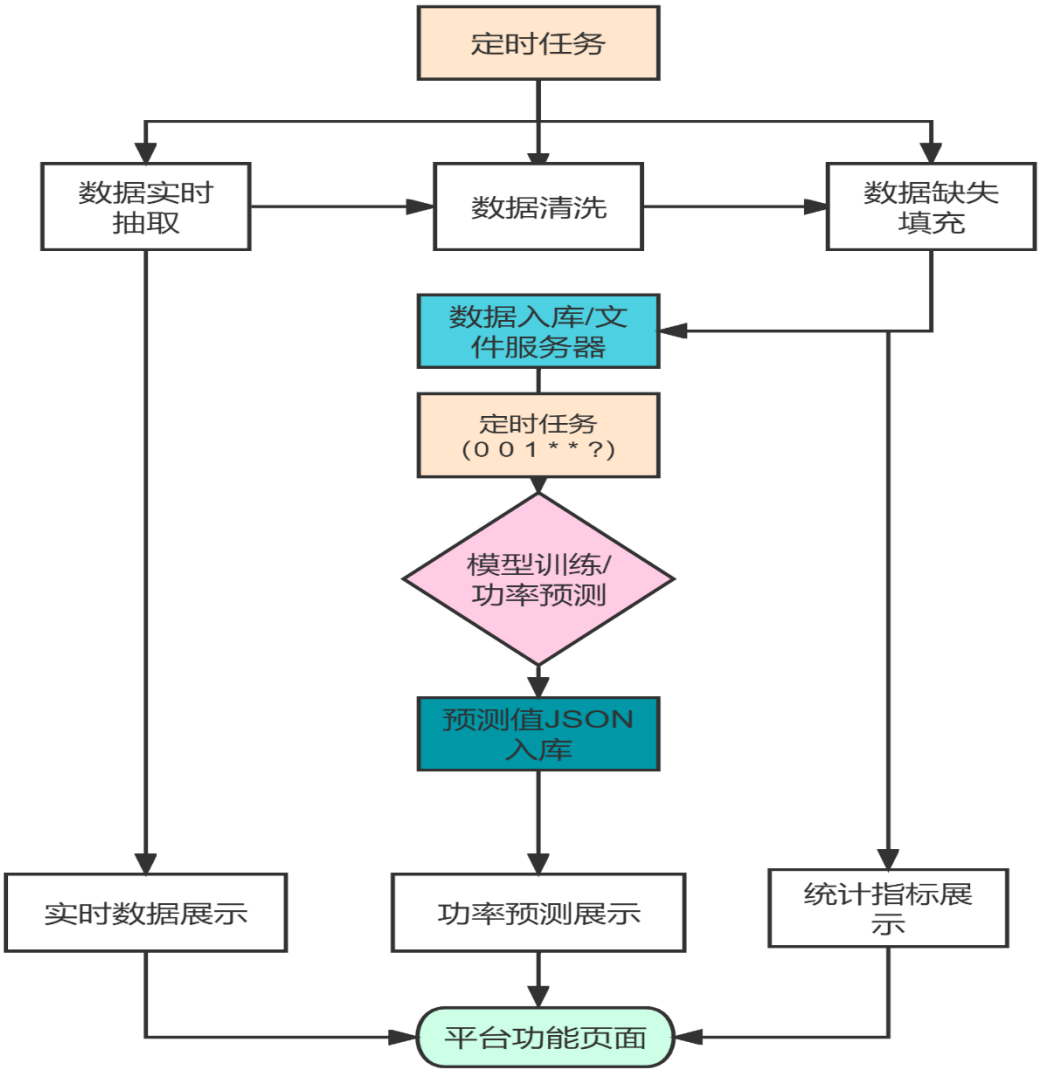

图 5-3 业务开发流程图

Figure 5-3 Images of business development process

通过数据实时抽取任务获取风机相关指标实时值，再传入数据清理模块，对异常数据进行检测，最后将数据传入数据缺失填充模块，进行功率缺失值的填充。填充后数据写入关系库和文件服务器，写入关系库为了方便页面的统计指标分析，存入文件服务器为了方便模型训练中数据集的读取。接着进入模型训练任务，每日凌晨一点进行机组的模型训练，利用预训练好模型对未来时间的风功率进行预测，预测结果直接以 json 格式存储在关系库中，页面直接传入风机编码到后台获取对应未来一段实际功率预测值。实时数据展示部分，直接调用实时数据库服务接口，传入指定的风机指标测点获取实时值。统计指标展示部分，从关系库根据风机测点进行指标值的统计，本平台统计指标是场站发电量、机组发电量和 24 小时故障次数。

5.3 平台功能介绍

本节首先介绍了本平台的微服务功能组件，最后介绍风功率预测可视化平台各功能模块，通过功能描述结合页面截图展示平台功能。

5.3.1 系统首页

系统首页如图 5-4 所示，首页的特点是采用了矢量图 SVG 展示风机实时运行情况，风机图元通过风速测点调用接口获取实时风速值从而控制风机的转速，相比较传统 GIF 展示更加符合实际运行状况，且图片不会出现失真和变形。右下角故障消息采用消息队列推送，后台作为消息生产者，将异常数据推送到页面展示，而非传统轮询的方式，保证故障消息的实时性，提醒运营人员进行故障排除。

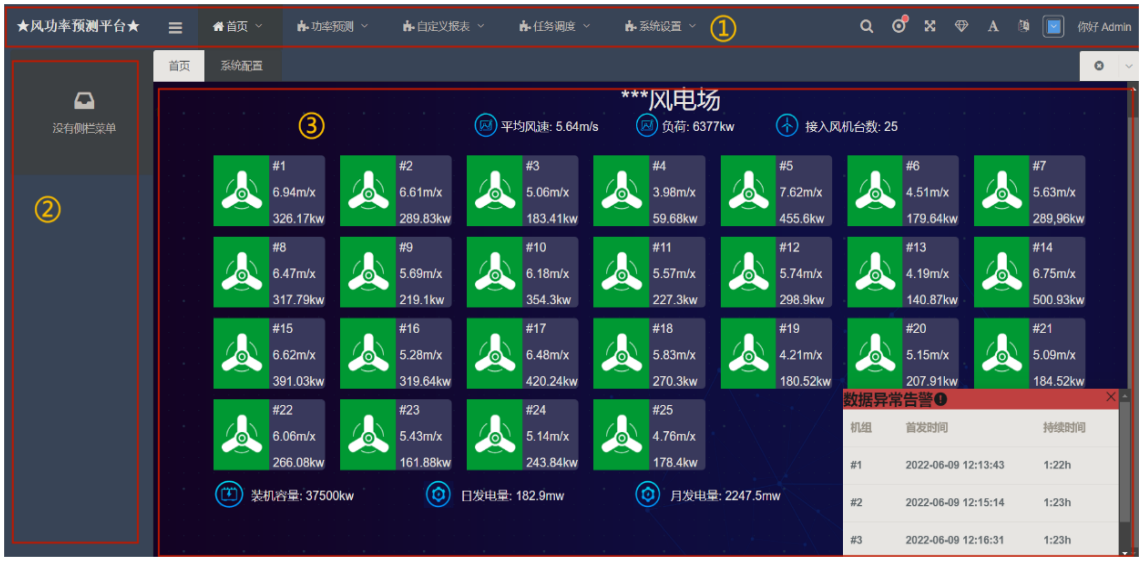

图 5-4 首页风机总览图

Figure 5-4 Images of homepage fan overview

5.3.2 单台风机页面

针对传统的风功率预测平台中单台风机的实时数据分散的问题，本平台合理设计了单台风机页面，展示风机重要的实时指标和统计指标。如图 5-5 所示，为单台风机指标一览页面，左下方展示用户重点关注的风机当日负荷曲线，数据源为通过调用实时库历史插值查询接口，右上方为功率预测模块，默认展示了本页面风机未来一周的功率预测曲线，预测模型默认为本文提出的基于 Informer 和图卷积的组合模型，也可通过系统设置页面进行其它预测模型选择。

图 5-5 单台风机页面图

Figure 5-5 Images of single fan

5.3.3 风功率预测查询页面

如图 5-6 所示，为风功率预测页面，页面提供未来一周、未来半月和未来一月的功率预测查询。页面查询条件包括多台机组的选择，预测模型的选择，预测结果通过曲线形式表现，横坐标为时间，纵坐标为功率预测值，单位为 kw。为了方便对风机未来运行状态进行对比，也可以选择多台风机进行风功率预测曲线。通过风功率预测模块，为风电场的运营和风机的检修等工作开展提供可靠的数据支撑。

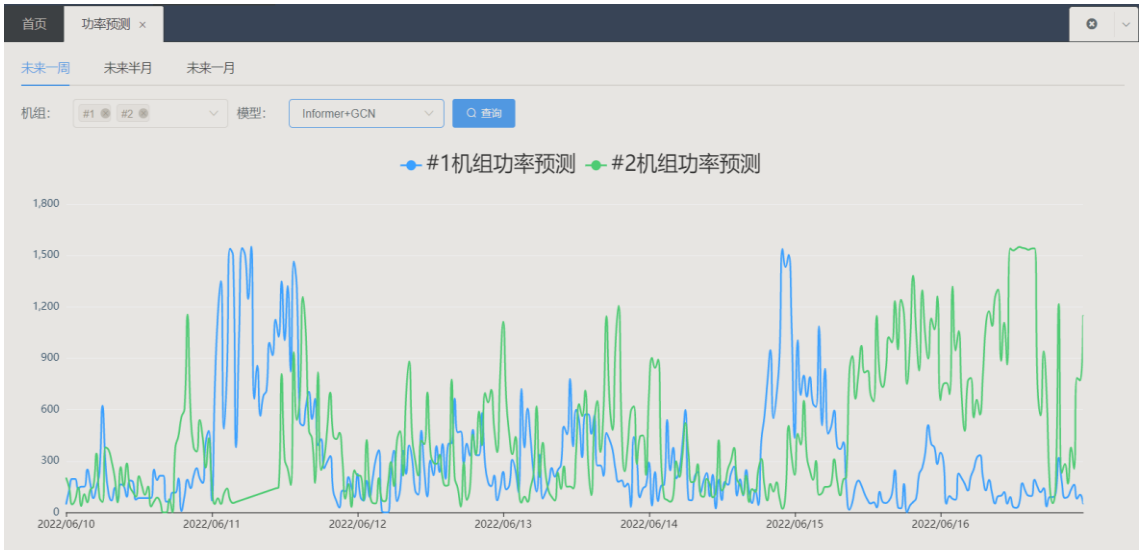

图 5-6 功率预测页面图

Figure 5-6 Images of power prediction

5.3.4 相关性分析

如图 5-7 所示，为风机功率相关性分析页面。时间相关性使用 ACF 相关性系数分析，空间相关性使用 Pearson 和 MIC 相关性系数分析，时空相关性系数使用 CCF 相关性系数分析。后台接收页面请求，调用 python 代码生成相关性分析图，最终后台将相关性分析图返回至页面。

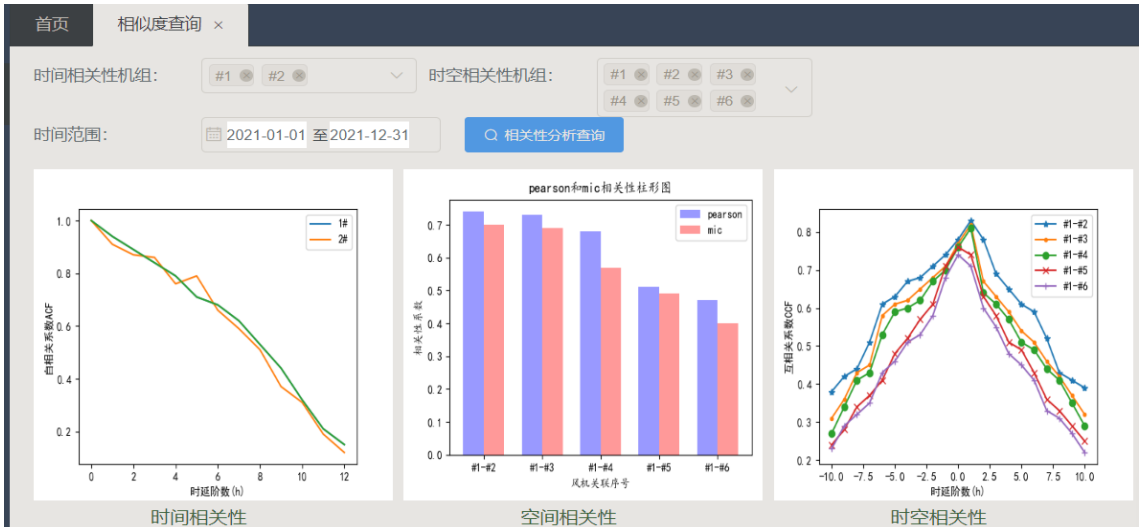

图 5-7 相关性分析图

Figure 5-7 Images of correlation analysis

5.3.5 自定义报表查询

如图 5-8 所示，为平台自定义报表查询模块。页面查询条件包括机组、时间范

围和指标的选择，并且展示本文所提的功率缺失值填充结果，报表支持导出功能。相比较传统的风功率预测平台，本页面将分散的风机数据进行整合，减少繁琐的查询工作，为运营人员提供灵活的指标报表查询。

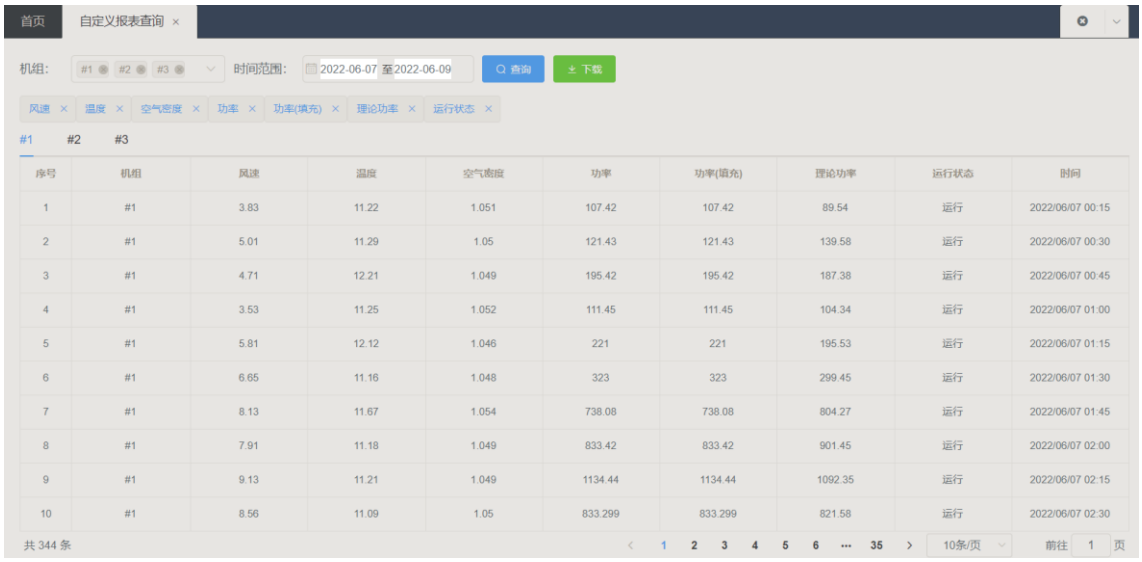

图 5-8 自定义报表查询图

Figure 5-8 Images of custom report query

5.3.6 系统设置

如图 5-9 所示，为平台功能设置页面。由于每台风机的投产运行情况存在差异，本平台可对每台风机进行个性化灵活配置。通过选择一台或多台机组，可对风机的数据异常告警条件进行个性化设置，支持风机模型训练的服务器节点的选择，支持对风机默认的风功率预测训练模型进行灵活选择。

图 5-9 系统设置页面

Figure 5-9 Page of system settings

5.3.7 功能组件

如图 5-10 所示，为本风功率预测可视化平台开发的微服务 Docker 组件包，包括数据预处理组件、用户服务组件、角色服务组件和任务调度组件。本平台通过编写 Dockfile 文件，将 springboot 生成的 jar 包生成到镜像中。平台部署通过编写 Docker Compose 文件实现镜像容器的定义和运行。相比较传统部署方式，容器化的部署方式可以减少繁琐的部署步骤，提高运维效率。

```
[root@VM-16-2-centos ~]# docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
dispatchserver      1.0.0               a957dd7dc662        16 hours ago        473MB
datapreprocessing   1.0.0               7b94cda7ffc7        8 days ago          446MB
basicservice        1.0.0               ae9660359c2a        9 days ago          922MB
userservice         1.0.0               85b121affedd        10 days ago         194MB
roleservice         1.0.0               b692a91e4e15        10 days ago         142MB
[root@VM-16-2-centos ~]# docker ps -a
CONTAINER ID        IMAGE                       COMMAND                  CREATED
70ba40294769        dispatchserver:1.0.0        "catalina.sh run"        7 minutes ago
17163569a772        roleservice:1.0.0           "/docker-entrypoint.…"   2 days ago
07927ee4b858        userservice:1.0.0           "bash"                   2 days ago
0102d9268acd        datapreprocessing:1.0.0     "docker-entrypoint.s…"   2 days ago
9464112a61ff        basicservice:1.0.0          "python3"                2 days ago
```

图 5-10 平台部署包图

Figure 5-10 Images of platform deployment package

5.4 本章小结

本章节详细阐述了风功率预测可视化平台的相关工作。首先分析了风功率预测可视化平台的现状，并且针对现状进行技术分析和技术选型工作；之后介绍了平台的核心技术栈，包括前后端分离技术、SpringCloud 微服务技术、Docker 容器化技术和 Spark 分布式运算框架；接着对平台的系统架构和平台功能的设计进行了全面的阐述，接着根据业务需求分析进行模块和开发流程图的设计，其中包括数据预处理、数据展示、数据可视化分析、任务调度和系统管理；最终对风功率预测可视化平台的功能模块进行逐一介绍。

6 总结与展望

6.1 工作总结

风功率中长期预测有助于我国风电行业的发展，有助于风电场基础设施建设的规划、场站日常检修工作的安排等。但目前风功率预测算法主在深度学习的领域研究不足，具有较大的研究空间。大量的功率数据缺失会对模型的实验造成精度丢失甚至造成工程项目的经济损失。目前应用到行业的预测模型基本都是基于时序数据的关系建模，鲜有融合空间特征。并且目前的模型都是定期更新，时效性不强，没有利用大数据平台分布式计算强大的计算能力提高模型的时效性和预测精度。针对上述问题，本文进行了如下的内容研究：

（1）提出基于工程先验知识和期望最大算法结合的缺失值填充方法。

针对风电功率历史数据缺失的问题，在期望最大算法填充的基础上引入了风功率工程先验知识。根据工程知识计算出风机的理论功率，结合聚类和 K 近邻算法，计算出期望最大算法初始化参数，将初始化参数控制在合理的范围内，提高期望最大算法收敛的速度，减少缺失值填充的误差。

（2）提出了基于 Informer 和图卷积的组合模型风功率预测算法。

针对风功率中长期预测属于时空预测问题，提出了基于 Informer 和图卷积的组合模型风功率预测算法。首先通过 Informer 捕获长序列时序关系，再通过图卷积捕获空间依赖关系，最终通过组合模型融合空间和时间的特征，进行最终的风功率预测。

（3）开发了一套风功率预测可视化平台。

在上述两点研究的基础上，设计开发了一套风功率预测可视化平台。本平台采用 Vue 和 SpringBoot 实现前后端的分离开发，使用 Docker 容器对平台进行部署管理，并且内置主流及本文提出的风功率预测的模型算法，利用 Spark 实时计算能力提高模型的时效性和精度，系统集成风机功率相关指标数据并进行展示，最终将预测结果使用 Echarts 中的曲线图、柱状图和表格等进行可视化，为发电厂工作人员提供可靠稳定的风功率中长期预测平台。

6.2 未来展望

本文提出的基于 Informer 和图卷积的组合模型对风功率预测，有效地提高的预测的精度，虽然在老师的指导下和同学的帮助下取得了一定的研究成果，但是依旧有进一

步的扩展空间，需要继续完善，总结的下一步的工作展望如下：

（1）本文对于功率异常点的检测还可以考虑更多元、更先进的故障检测算法。

（2）将本文提出的基于 Informer 和图卷积的组合预测模型应用到其他时空序列预测问题上，使用更多的数据集进行研究。对于算法模型不断提出改进，以便可以更好地适用于不同领域下的时空预测问题，使本文提出的预测方法更具有泛化能力和普遍性。

（3）本文融合 Informer 和图卷积的组合模型采用基于方差-协方差组合的方式，可以选择其他的组合方式，譬如樽海鞘群算法、粒子群算法等智能优化算法进行参数的优化。

（4）本文的风功率预测可视化平台页面的交互可以进一步优化完善，目前容器的运行还是手动执行命令，未来希望引入 kubernetes 对平台的容器进行管理。

综上分析，本文对风功率中长期预测取得了阶段性的成果，但是依旧存在许多可以完善的地方，后续会进一步对风功率中长期预测进行研究，便于更好的应用到工程中。

参考文献

[1] 刘永奇, 陈龙翔, 韩小琪. 能源转型下我国新能源替代的关键问题分析[J]. 中国电机工程学报, 2022, 42(2): 515-524.

[2] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述[J]. 计算机工程与应用, 2021, 57(14): 27-38.

[3] Zakaria N A, Noor N M. Imputation methods for filling missing data in urban air pollution data formalaysia[J]. Urbanism. Arhitectura. Constructii, 2018, 9(2): 159.

[4] Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data[J]. Statistical Methods in Medical Research, 2016, 25(5): 2021-2035.

[5] Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods[J]. Computational Statistics & Data Analysis, 2015, 90: 84-99.

[6] Wang P, Chen X. Three-way ensemble clustering for incomplete data[J]. IEEE Access, 2020, 8: 91855-91864.

[7] Hastie T, Mazumder R, Lee J D, et al. Matrix completion and low-rank SVD via fast alternating least squares[J]. The Journal of Machine Learning Research, 2015, 16(1): 3367-3402.

[8] Beaulieu-Jones B K, Moore J H, POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM. Missing data imputation in the electronic health record using deeply learned autoencoders[C]//Pacific symposium on biocomputing 2017. 2017: 207-218.

[9] Silva-Ramírez E L, Pino-Mejías R, López-Coello M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns[J]. Applied Soft Computing, 2015, 29: 65-74.

[10] Yoon J, Jordon J, Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets[C]. In International Conference on Machine Learning, 2018: 5675-5684.

[11] Luo Y, Cai X, Zhang Y, et al. Multivariate time series imputation with generative adversarial networks[J]. Advances in neural information processing systems, 2018, 31.

[12] Luo Y, Zhang Y, Cai X, et al. E2gan: End-to-end generative adversarial network for multivariate time series imputation[C]. AAAI Press. 2019: 3094-3100.

[13] Shih S Y, Sun F K, Lee H. Temporal pattern attention for multivariate time series forecasting[J]. Machine Learning, 2019, 108(8): 1421-1441.

[14] 黎维, 陶蔚, 周星宇,等. 时空序列预测方法综述[J]. 计算机应用研究, 2020,37(10):2881-2888.

[15] Farhath Z A, Arputhamary B, Arockiam L. A survey on ARIMA forecasting using time series model[J]. Int. J. Comput. Sci. Mobile Comput, 2016, 5: 104-109.

[16] Duan P, Mao G, Zhang C, et al. STARIMA-based traffic prediction with time-varying lags[C]//2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2016: 1610-1615.

[17] Christiano L J. Christopher A. Sims and vector autoregressions[J]. The Scandinavian Journal of Economics, 2012, 114(4): 1082-1104.

[18] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

[19] Kobayashi K, Kaito K, Lethanh N. A statistical deterioration forecasting method using hidden

Markov model for infrastructure management[J]. Transportation Research Part B: Methodological, 2012, 46(4): 544-561.

[20] 王佳璆、邓敏、程涛. 时空序列数据分析和建模[M]. 北京: 科学出版社,2012.

[21] 柳姣姣, 禹素萍, 吴波, 等. 基于隐马尔科夫模型的时空序列预测方法[J]. 微型机与应用, 2016, 35(1): 74-76.

[22] Salinas D, Flunkert V, Gasthaus J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020, 36(3): 1181-1191.

[23] Zhang J, Zheng Y, Qi D, et al. DNN-based prediction model for spatio-temporal data[C]//Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2016: 1-4.

[24] Guo S, Lin Y, Feng N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 922-929.

[25] Wu Z, Pan S, Long G, et al. Connecting the dots: Multivariate time series forecasting with graph neural networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 753-763.

[26] Zhao L, Song Y, Zhang C, et al. T-gcn: A temporal graph convolutional network for traffic prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(9): 3848-3858.

[27] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. Advances in neural information processing systems, 2015, 28.

[28] Han H, Zhang M, Hou M, et al. STGCN: A Spatial-Temporal Aware Graph Learning Method for POI Recommendation[C]//2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020: 1052-1057.

[29] Saravanan S, Kannan S, Thangaraj C. India's electricity demand forecast using regression analysis and artificial neural networks based on principal components[J]. ICTACT Journal on soft computing, 2012, 2(4): 365-70.

[30] Cui J, Liu S, Zeng B, et al. A novel grey forecasting model and its optimization[J]. Applied Mathematical Modelling, 2013, 37(6): 4399-4406.

[31] 鲁宝春, 赵深, 田盈, 等. 优化系数的NGM(1,1,k)模型在中长期电量预测中的应用[J]. 电力系统保护与控制, 2015,43(12): 98-103.

[32] Ozkan M B, Karagoz P. A novel wind power forecast model: Statistical hybrid wind power forecast technique (SHWIP)[J]. IEEE Transactions on Industrial Informatics, 2015, 11(2): 375-387.

[33] Mana M, Burlando M, Meissner C. Evaluation of two ANN approaches for the wind power forecast in a mountainous site[J]. International Journal of Renewable Energy Research, 2017, 7(4): 1629-1638.

[34] 张健美, 周步祥, 林楠, 等. 灰色 Elman 神经网络的电网中长期负荷预测[J]. 电力系统及其自动化学报, 2013, 25(04): 145-149.

[35] Li Bowen, Zhang Jing, He Yu, et al. Short-term load-forecasting method based on wavelet decomposition with second-order gray neural network model combined with ADF test [J]. IEEE Access, 2019, 5: 16324-16331.

[36] 马星河, 闫炳耀, 唐云峰, 等. 基于优选组合预测技术的中长期负荷预测[J]. 电力系统及其自动化学报, 2015, 27(6): 62-67.

[37] 周滦, 任海军, 李健, 等. 层次结构下的中长期电力负荷变权组合预测方法[J]. 中国电机工程学报, 2010, 30(16): 47-52.

[38] Wang Xiping, Wang Yaqi. A hybrid model of EMD and PSO-SVR for short-term load forecasting in residential quarters [J]. Mathematical Problems in Engineering, 2016.

[39] Wang Jianjun, Li Li, Niu Dongxiao, et al. An annual load forecasting model based on support vector regression with differential evolution algorithm [J]. Applied Energy, 2012, 94: 65-70.

[40] Li Weiqin, Chang Li. A combination model with variable weight optimization for short term electrical load forecasting [J]. Energy, 2018, 164: 575-593.

[41] 赵芝璞, 高超, 沈艳霞, 等. 基于关联模糊神经网络和改进型蜂群算法的负荷预测方法[J]. 中国电力, 2018, 51(02): 54-60.

[42] Guo S, Lin Y, Li S, et al. Deep spatial-temporal 3d convolutional ceural cetworks for traffic data forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2019: 3913-3926.

[43] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[44] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[45] Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of AAAI. 2021.

[46] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[47] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.

[48] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in neural information processing systems, 2016, 29.

[49] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[50] Velickovic P, Cucurull G, Casanova A. Graph attention networks[C]. //International Conference on Learning Representations, Vancouver, Canada. 2018.

[51] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data[J]. Data & knowledge engineering, 2007, 60(1): 208-221.

[52] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis[C]//International conference on machine learning. PMLR, 2013: 1247-1255.

[53] Soldani J, Tamburri D A, Van Den Heuvel W J. The pains and gains of microservices: A systematic grey literature review[J]. Journal of Systems and Software, 2018, 146: 215-232.

[54] Wan X, Guan X, Wang T, et al. Application deployment using Microservice and Docker containers: Framework and optimization[J]. Journal of Network and Computer Applications, 2018, 119: 97-109.

[55] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: Cluster computing with working sets[C]//2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10). 2010.

[56] Meng X, Bradley J, Yavuz B, et al. Mllib: Machine learning in apache spark[J]. The Journal of Machine Learning Research, 2016, 17(1): 1235-1241.

学位论文数据集

表 1.1： 数据集页

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>风功率预测;缺失值研究;时空预测模型;中长期预测</td><td>公开</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">学位授予单位名称*</td><td>学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>电子信息</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名*</td><td colspan="2">并列题名</td><td>论文语种*</td></tr>
<tr><td colspan="2">基于 Informer 的中长期风电功率预测研究</td><td colspan="2"></td><td>中文</td></tr>
<tr><td>作者姓名*</td><td colspan="2">王文贵</td><td>学号*</td><td>20140081</td></tr>
<tr><td colspan="2">培养单位名称*</td><td>培养单位代码*</td><td>培养单位地址</td><td>邮编</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>北京市海淀区西直门外上园村 3 号</td><td>100044</td></tr>
<tr><td colspan="2">专业领域*</td><td>研究方向*</td><td>学制*</td><td>学位授予年*</td></tr>
<tr><td colspan="2">软件工程</td><td>数据挖掘</td><td>2 年</td><td>2022 年</td></tr>
<tr><td>论文提交日期*</td><td colspan="4">2022 年 9 月</td></tr>
<tr><td>导师姓名*</td><td colspan="2">黄华</td><td>职称*</td><td>副教授</td></tr>
<tr><td>评阅人</td><td colspan="2">答辩委员会主席*</td><td colspan="2">答辩委员会成员</td></tr>
<tr><td></td><td colspan="2">常晓林</td><td colspan="2">张英俊
白慧慧</td></tr>
<tr><td colspan="5">电子版论文提交格式 文本（√） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ）
推荐格式：application/msword；application/pdf</td></tr>
<tr><td colspan="2">电子版论文出版（发布）者</td><td colspan="2">电子版论文出版（发布）地</td><td>权限声明</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td><td></td></tr>
<tr><td>论文总页数*</td><td colspan="4">56 页</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 21 项。</td></tr>
</table>