各位在座的评委老师们，大家好！我是2018级计算机基地班的蒋嵩林。我的毕业论文题目是《基于ICEEMDAN多特征分解和Prophet-GRU-NN组合模型多步预测短期风速》。下面我将从研究背景、建模过程、总结展望这三个方面来介绍我的毕业论文工作。

首先是研究背景。这部分我将主要围绕使用多特征组合模型进行风速预测的意义来进行介绍。

使用风能是出于能源变革的需要。对于短期近地面风速的预测能够很好的帮助并促进风能的应用。然而，现有的研究成果表明，单一模型并不能很好地进行风速预测。同时，在研究数值预报方法的过程中，借鉴大气动力学方程的思想，我注意到不同于风速，其它气象要素随着季节变化，通常具有很强的周期性，一般较容易预测，因而产生了使用多种气象要素辅助风速预测的想法。

接下来我将介绍本次论文的模型构建过程。

本次模型数据来源于国家青藏高原科学数据中心提供的中国区域地面气象要素驱动数据集，时间分辨率为3个小时，包含了传统数值天气预报使用的7个气象要素，可以为中国区域陆面大气动力学过程提供模拟。

选取的地点位于甘肃玉门的中电酒泉第四风力发电有限公司附近，从图2所示的局部范围内卫星地图我们可以看到清晰的风力发电机组。

表1是数据集中7个气象要素指标的具体介绍。此次建模过程选取该地点2017至2018年共计5832条数据。

根据斯皮尔曼相关性分析，去除掉相关性不显著的地面降水率特征，保留其余5种辅助气象要素特征。

使用Python语言的pyEMD实现ICEEMDAN分解。下面展示6种气象要素的ICEEMDAN分解结果。

随后使用Prophet模型自带的方法，交叉验证搜寻Prophet模型对于风速数据的最优超参数，从而使得MAPE值最小化。最终得到如表3所示的最优值。

训练时根据5828条数据点，以8个连续时间序列数据点，即1天，为一个时间窗口制作为一个训练样本。每次向后滑动1个数据点，即3个小时，制作为5820个训练样本。神经网络部分的实现使用Tensorflow Keras，训练参数如PPT第二段所示。使用多步预测的方法。每个时间窗口的输入为现在已知的数据点值，混合上一个时间窗口的预测输出值，从而进行四步预测12小时的结果。

为了验证本文提出组合模型的优越性，本文还与只使用单个或多个本文提出的组合模型中的成分进行组合的其它简单模型进行对比。首先是只基于风速的ICEEMDAN-GRU模型。该模型首先对近地面全风速进行ICEEMDAN分解，去除高频信号所在的IMF分量，将低频信号的IMF分量相加，构建ICEEMDAN分解处理后的风速序列输入模型。

对于模型内部部分，训练集样本窗口首先通过标准化后，进行RBF核主成分分析升维，降低复杂度，再将其输入进GRU模型。

这里使用的GRU模型如图11所示，模型四层之间使用序贯模型的形式进行连接。第一层为模型的输入层；第二层使用标准GRU模型单元，即激活函数为sigmoid，循环激活函数为tanh的隐藏层，其一共有1024个GRU单元。第三层使用将激活函数和循环激活函数都改为GELU的GRU单元，其共计512个GRU单元。第四层为全连接输出层，为输出的最终风速预测值。

接下来基于刚刚的风速ICEEMDAN-GRU模型，将每个窗口的输入值同时使用Prophet框架进行统计分析处理，得到该窗口数据对应的拟合值、趋势分量、累加式季节性分量以及上述分量所对应的预测最大边界和最小边界共计9个特征，并与原始数据进行拼接整合。随后将最后两个维度展平并进行数据标准化缩放，再通过使用RBF核函数的核主成分分析，将训练集升维后再输入进同上一种模型GRU结构的GRU模型，得到最终预测结果，形成风速ICEEMDAN-Prophet-GRU模型。

最后，将其余辅助的5种气象要素的原始历史数据的每个气象要素通过和刚刚对近地面全风速进行ICEEMDAN分解一致的方式得到6个气象要素ICEEMDAN分解处理后的序列输入模型。

随后，同风速ICEEMDAN-Prophet-GRU模型一致，对每个气象要素训练独立预测的ICEEMDAN-Prophet-GRU模型，得到每个GRU模型预测值，再将这些预测值通过使用RBF核函数的核主成分分析升维，随后将这些预测值输入NN模型，对风速的预测值进行修正，最终得到对风速的预测结果。

神经网络模型NN六层之间使用序贯模型的形式进行连接。第一层为模型的输入。第二层为一个展平层。第三、四层均为全连接层，其激活函数均为GELU，各有128、64个神经元。第五层为丢弃层，其丢弃比率为20%，用于防止神经网络的过拟合。最后一层为输出层，将前一层的输出值通过全连接的方式，输出最终预测值。

模型的评价指标选用了MSE、MAE、MAPE、RMSE、决定系数、15%准确率。

最终模型四步预测结果的评价指标如表4所示，另外从下面的几张预测结果对比图也可以看到提出模型的优越性。

最后一部分是总结展望。

本文有如下创新：第一，从数值预报、大气动力学获得灵感，多气象要素联合预测。第二，使用最新信号分解技术ICEEMDAN来特征分解。采用Facebook推出的新型时间序列框架 Prophet与最新循环神经网络 GRU 模型组合预测。第三，从支持向量回归机获得灵感，使用了RBF核PCA升维。第四，将GELU激活函数以及Nadam优化器和Huber损失函数应用到短期风速预测领域。第五，选择甘肃酒泉一个真实风力发电厂附近气象数据验证模型。

当然，因为论文时间紧迫，展望未来，本次论文写作过程也存在着许多遗憾的点，如果有机会的话可以在后续进一步研究过程中完成。本次建模过程只选择了两年跨度的时间序列数据。只对风速原始时间序列数据这一个模型进行了搜索最优超参数的过程。本文只选择了甘肃酒泉的一个真实风力发电厂附近的地点进行短期风速的预测。没有探究模型中所用神经网络层数进一步加深的情况。也可尝试更多各种不同新型模型，用各种不同方法组合起来。

以上就是我毕业论文答辩的全部内容，敬请各位老师批评指正，谢谢！

**答疑素材：**

ICEEMDAN分解图中第一行为原始数据，最后一行为ICEEMDAN分解后得到的残差，中间的行为ICEEMDAN分解后得到的本征模函数（IMF）分量。每行左侧为原始得到的时域数据，右侧蓝色部分为经过傅里叶变换后得到的频域数据。可以看到对于每种气象要素数据，高频信号主要都集中在IMF1和IMF2分量上。

图16是基于风速历史的Prophet模型的可视化图，该模型图中黑点部分为实际值，实蓝色曲线部分为预测值，透明蓝色部分的上下边界分别为预测结果可能的最大值与最小值。从图中我们可以看到其实际值与预测值的偏差很大，结合该图与表4中对未来4步预测数据评价指标值可知，Prophet模型的预测结果很差。从表4中我们还可以看到，对风速数据进行ICEEMDAN分解处理，剔除高频噪声后，其准确度相较于未经分解的基于风速历史Prophet模型而言各项指标都有比较好的提升效果。

图17为基于神经网络的三种模型预测可视化图。由图和表4可见，使用GRU模型之后，模型的各项指标相较于只使用Prophet而言都得到了大幅度地显著提高。进一步地，加入Prophet的统计分析数据之后交由GRU进行训练，模型的准确度和各项指标进一步提升，但是在进行第四步预测时都出现了预测值的较大偏离问题。最终，使用多气象要素特征训练GRU模型，并将这些气象要素模型的预测结果输入进NN，从而对风速的预测值进行校正，产生的预测结果MAPE值达到了个位数，准确率有了比较大的提升，且第四步预测时准确度也较好，体现了本文最终提出模型的优越性和实用价值。

创新点。本文综合使用目前最新的科研成果，从数值预报、大气动力学获得灵感，使用多气象要素联合预测，同时预测其他具有强周期性、易预测的气象要素，而不仅仅基于风速的历史时间序列，达到了对风速预测进行修正的目的。另外，模型基于最新信号分解技术ICEEMDAN，还采用Facebook推出的新型时间序列框架 Prophet与最新循环神经网络 GRU 模型进行组合。从支持向量回归机获得灵感，将GRU模型的输入数据使用RBF核PCA升维，从而将非线性的特征历史数据映射到高维线性可分，降低复杂度。同时对相关模型进行优化，突破模型的传统，将GELU激活函数以及Nadam优化器和Huber损失函数应用到短期风速预测领域。通过选择甘肃酒泉的一个真实风力发电厂附近进行短期风速的预测从而验证模型，而并非像同类研究一样为了预测风速选择了任意的非风力发电场所的气象数据集进行验证，因而本论文提出的模型通过了实用性的验证。最终预测结果显示了该模型的优势和极大的实用价值。

但是，因为论文时间紧迫，展望未来，本次论文写作过程也存在着许多遗憾的点，如果有机会的话可以在后续进一步研究过程中完成。一、本次建模过程出于性能的考虑，在源数据集1979-2018年份中只选择了最后两年跨度的时间序列数据，没有将更大尺度的数据投入训练。对于更大尺度数据而言，模型能进一步强化对年季节性和长期变化关系的学习，因而能够取得更好的效果。二、由于Prophet搜寻最优超参数所耗费时间过长，本文只对风速原始时间序列数据这一个模型进行了搜索最优超参数的过程。因为对于一个模型的超参数搜寻在作者电脑中运行完成大概就要花上三、四天的时间，耗费时间太长，因而对于其他特征时间序列本文则直接沿用了风速原始时间序列数据的最优超参数，因为这些参数都比默认的要取得更好的效果。如果对每个特征的时间序列构建的Prophet模型都加以优化，预计会产生更好的效果。三、本文只选择了甘肃酒泉的一个真实风力发电厂附近的地点进行短期风速的预测。当然本文作者坚信在其他风力发电场所本模型也会产生很好的效果，但后续有待验证。四、受本文作者使用的电脑用于训练的显卡和CPU以及内存的限制，没有进一步将本文提出模型中所用神经网络层数进一步加深，因而可以进一步探究加深后模型的表现。五、随着数据科学的发展，新模型正在不断地涌现出来。由于单一的模型各有优缺点，尝试更多各种不同的新型模型，用各种不同的方法组合起来，会产生更好的效果。

本次论文模型构建过程中使用到的技术：

这里使用基于EMD的时间序列信号分解技术而非小波分解，是因为EMD技术依据信号自身的时间尺度特征对信号没有任何先验主观标准选择地自动分解，因而无需像傅里叶变换与小波分解一样设定谐波基函数或小波基函数。EMD将时间序列分解为包含不同频率的本征模函数，也即IMF分量，从而可以较好地将高频噪声和低频信号分离开来。文章使用目前最前沿的于2014年提出的改进自适应噪声完备集合经验模态分解，即ICEEMDAN分解。与传统基于EMD的系列算法而言，ICEEMDAN具有许多显著的优势。

Prophet是由Facebook公司于2018年推出的一款全自动时间序列预测框架。对于气象数据而言，其将一个时间序列拆解为趋势分量，周、日、年的季节性分量以及服从高斯白噪声分布的分量。

主成分分析可以降低时间序列的复杂度，但是其仅适用于线性数据。因而这里引入借鉴了SVR思想的核主成分分析。通过高斯径向基函数，即RBF，将非线性数据映射到高维线性可分后，再使用主成分分析进行降维。

如图2可见经典BP神经网络和经典循环神经网络的示意图。经典的神经网络，由于其时序无关，采用了层状传递结构，模型的内部前一层神经元直接传递给后一层神经元。这样做模型由于无记忆功能，无法学习时序数据中的历史相关性，因而需要循环层，通过内部传递时序数据来学习记忆内容。

神经网络激活函数将神经网络模型中神经单元的输出映射到一个范围，从而起到控制输出数据值域的作用。常见的激活函数有sigmoid，tanh，ReLU三种。由于sigmoid，tanh在无穷小和无穷大处总是趋于常数，因而很容易引起RNN的梯度消失现象。然而尽管ReLU由于其性质可以克服梯度消失问题，但是由于ReLU是线性激活函数，而风速数据具有较强的非线性、非平稳特征，因而用于风速数据效果不是很理想。高斯误差线性单元，也即GELU，是谷歌于2016年推出的一种新型激活函数。Phi(x)是标准高斯累积分布函数，通过高斯加权使得其非线性化，在许多循环神经网络任务中都取得了最好的效果。

神经网络模型优化器直接参与神经网络的训练，通过对应算法搜索神经网络的最优权重参数。目前主流的深度学习模型优化器为Adam，几乎成为了所有神经网络的标配。Nadam是Adam的一种变体算法，于2016年推出，其融合了能够提高RNN任务性能的传统涅斯捷罗夫加速梯度下降，即NAG算法的精髓，因而论文中采用Nadam优化器。

神经网络模型损失函数度量当前权重参数下输出值和真实值之间的误差大小。这里使用将MAE和MSE组合起来的损失函数Huber，从而取长补短。

本次论文采用GRU模型，是因为经典RNN模型存在缺陷。同时，相较于LSTM存在三个门控单元，于2014年推出的GRU模型只有两个门控单元，因而GRU的参数更少，复杂度更低，收敛速度更快，并且许多研究都显示GRU和LSTM所能达到的训练以及预测效果同样出色。

因为传统的皮尔逊相关系数衡量的只是线性相关关系，且对数据的要求较高，因而这里采用可以度量单调关系的斯皮尔曼相关系数。