分类号 TP399 学号 14063030

U D C 000 密级 公 开

工学硕士学位论文

带约束的多尺度风速插值和预测方法研究

硕士生姓名 朱祥茹

学 科 专 业 计算机科学与技术

研 究 方 向 大规模科学与工程计算

指 导 教 师 张卫民 研究员

国防科学技术大学研究生院

二〇〇五年一月

Interpolation and Prediction of Wind Speed with Multi-scale Process and Multivariable Constraint

Candidate：Zhu Xiangru

Advisor：Zhang Weimin

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha，Hunan，P.R.China

（January，2005）

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目：带约束的多度风速预测方法研究

学位论文作者签名：朱祥茹　　日期：2017 年 2 月 13 日

学位论文版权使用授权书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目：带约束的多尺度风速预测方法研究

学位论文作者签名：朱祥茹　　日期：2017 年 2 月 13 日

作者指导教师签名：　　日期：2017 年 2 月 13 日

目 录

表 目 录

图 目 录

摘 要

大气再分析数据在应用过程中需要根据需求将低分辨率数据插值成高分辨率数据。传统方法多采用线性插值和指数插值。这些方法误差较大，因此在二维空间序列计算中，通常采用单一尺度高斯过程回归进行插值；在一维时间序列计算中，通常采用循环神经网络对风速值进行预测。但是上述方法都建立在被插值变量组分单一且与其他变量相互独立的假设上，这些方法都忽略了气象数据的物理意义和物理约束。风速信号本身是多个物理过程导致的结果，与温度、气压、密度等气象变量有复杂的非线性关系，因而风速序列的插值计算需要考虑多尺度特征和多变量约束，否则可能会丢失序列中隐藏的重要信息。本文将在不同方面对风速空间插值和时间预测方法进行改进，以提高再分析数据应用计算的准确率。

本文的主要工作和贡献是以下两个方面：

（1）本文提出了一种风速多尺度插值核函数，改进了用于风速空间序列插值的高斯过程回归方法。传统的核函数是由一个单一尺度的风速协方差函数和独立高斯噪声项组成。本文根据风速多尺度特点，将核函数构造为由大尺度风速协方差函数、中小尺度风速协方差函数、空间相关噪声协方差函数和独立高斯噪声组成的组合核函数。实验选取东经 16°不同高度和东经 17°~东经 27°50km 高度纬向风和经向风再分析数据为样本数据，以 2°为间隔建模，以 1°为间隔插值验证。实验结果表明，相对于线性插值、指数插值和单一平方指数协方差核函数插值，本文提出的风速多尺度插值方法纬向风和经向风插值均方根误差（RMSE）更低，使用纬向风和经向风分量计算的误差远小于使用风速向量之模和风速向量角度计算的误差。风速多尺度插值方法的优化效果随垂直高度增加而增大。由于数据是沿着经线选取，因此多尺度核函数经向风分量插值的均方根误差比单一尺度核函数插值、线性插值的均方根误差减小了 1/2 以上，这表明本文提出的风速多尺度插值核函数是有效的。

（2）本文提出了一种采用风速状态向量进行多变量约束短期风速预测方法。本文将风速时间序列元素由单一纬向风（或经向风）标量改进为纬向风（或经向风）与温度、位势高度、相对湿度组成的状态向量，分别应用到基于长短期记忆网络（LSTM）的序列到序列（Seq2Seq）预测方法中。实验选取东经 115°和 130°北半球区域的 1948 年-1972 年再分析数据进行训练，选取 1973-1975 年数据进行验证。在相同时间步长情况下，由风速、温度、位势高度组合构成的三种状态向量明显改善了向量整体预测结果和纬向风分量（或经向风分量）预测结果。其中，由风速 u 、温度 t 、位势高度 h 组成的 3 维状态向量 (u,t,h) 的两种预测均方根误差

（RMSE）都达到最低。而包含相对湿度 rh 的状态向量 (u,t,h,rh) 不仅未改善预测结果反而使均方根误差（RMSE）更大。上述实验结果表明，基于长短期记忆网络（LSTM）的多变量约束风速预测方法能够有效地优化风速时间序列预测结果，但是优化程度取决于风速状态向量各个分量的相关程度。相关度高的变量构成的风速状态向量优化效果明显，而不合适的状态向量可能会让结果变差。在中高纬度，状态向量 (u,t)、(u,h)、(u,t,h) 会整体提高向量的预测准确性，也会单独提高风速变量的预测准确性。其中，状态向量 (u,t,h) 在实验区域所有范围中（东经 115°~130° 北纬 10°~80°）预测的准确性最高，且适用范围最广。本文方法在西雅图地表风预测中的预测结果比 NOAA 的高空风预报的预测结果准确性更高。

主题词：风速插值 风速预测 多尺度过程 多变量约束 高斯过程回归 LSTM

ABSTRACT

Atmosphere reanalysis data will be interpolated from high resolution to low resolution according to the requirements of the application. Linear interpolation and exponential interpolation is usually used, of which the bias is big. Thus, we usually interpolate by single-scale Gaussian process regression in 2-D space series prediction, and we usually calculate wind values by recurrent neural network in 1-D time series prediction. However, these methods all base on the hypothesis that the evaluated values are single signals and are independent to other variables, which ignore the physical significance and physical constraints of meteorological data. Wind speed results from multiple physical processes, having a complex nonlinear relationship with temperature, pressure, density and other meteorological variables. Wind speed sequence is a signal with multi-scale process and multivariable restrictions in which there is lots of important hidden information. This thesis will improve the interpolation and extrapolation of wind speed sequence in different aspects to increase the accuracy.

The thesis emphasized on several aspects as follows:

(1) In this thesis we improve the Gaussian process regression method for wind speed space interpolation, and propose a multi-scale interpolation kernel function. Usually the kernel function consists of a single-scale covariance function and an independent Gaussian noise. Based on the characteristics of the multi-scale process of wind speed sequence, the new combined kernel function consists of a large-scale wind speed covariance function, a middle-scale and small-scale wind speed covariance function, a spatial correlated periodic covariance function, a spatial correlated noise covariance function and an independent Gaussian noise. The kernel function is modeled with interval of 2 degree and the target output is reanalysis data with interval of 1 degree. The results show that the interpolation prediction of the kernel function with multi-scale is effective. The RMSE of the approach in this thesis is smaller than the one of linear interpolation, exponential interpolation and the interpolation with single square exponential covariance kernel function. Better results have been achieved in zonal wind component and meridional wind component prediction. In interpolation calculation of wind speed, the error generated by the interpolation of zonal wind component and meridional wind component is much smaller than the one generated by the norm of wind vector and the angle of wind vector.

(2) We improve the data format in the prediction of wind speed time series, and change the sequence elements from single zonal wind (or meridional wind) scalar to the wind state vector with multivariable constrains. The wind speed state vector made up by some of zonal wind (or meridional wind), geopotential height, relative humidity and temperature will be applied to the LSTM based Sequence to Sequence prediction

approach. The reanalysis data from 1948 to 1972 are chosen as the training set and cross-validation set, and the reanalysis data from 1973 to 1975 are chosen as the testing set. Note that the prediction from state vector to state vector shows a good result, in which the state vector made by wind speed component, temperature and geopotential height generates the smallest RMSE, while the state vector containing relative humidity doesn't work well. In conclusion, if the elements of wind speed state vector have strong relationship with each other, the RMSE of prediction will be lower than the RMSE of prediction without wind state vector. Otherwise, the results of prediction will not be improved.

Key Words： wind speed prediction, multi-scale process, multi-variable constrains, Gaussian Process Regression, LSTM

第一章 引言

1.1 课题研究背景及意义

风速数据广泛应用于风力发电、航天航空等领域。例如需要高空风场统计火箭预期轨道上危险风的出现概率[1]。而飞机起飞、航路飞行、降落阶段轨迹的预测和燃油率的经验计算也基于高空风历史数值[2]。航天器系统的设计和任务规划需要的气象要素不仅仅局限于风分量，也包括大气温度、密度、气压、露点温度和其他化学成分浓度值。在给定飞行轨迹上，根据参考大气模型模拟轨迹上的气象要素，以便于对飞行器进行姿态控制系统设计、制导精度分析和载荷分析[3]。

天气气候变化预测、风力发电、航空航天研究对于大气要素的准确预报提出了更高的要求。大气预测方法基于空气动力学原理，运用统计学手段，对未来大气状态进行推演，或建立物理数学模型研究异常天气气候或者未来天气气候趋势。研制高精度的气象数据对规避自然灾害破坏、促进经济健康快速发展、推动科技成果造福于民具有重要意义。

但是，目前一些统计预报方法得到的气象数据精度和分辨率还达不到应用的需求。例如，全球参考大气模型系统（GRAM）中 0-27km 全球高空气候图集数据库的数据分辨率是 2.5°×2.5°，20-120km 的纬向月平均数据集的分辨率是 10°[4]，远低于应用所需分辨率（1°×1°）。为了得到合适的高分辨率数据，需要对原有数据进行插值。在水平方向上，温度、气压、密度插值是根据理想气体定律和流体静力学假设推导的，垂直方向上则采用线性插值、二维均匀插值计算。这种方法假设大气在水平和垂直方向是均匀变化的，而实际数据往往并非如此。由于我们对大气真实运动状态的太多还处于不完善阶段，为了简化，在气象预测中采用了多种近似方法，因此得到的大气数据具有一定的误差。

影响大气要素变化的因素有很多。以风速为例，地理形貌、非地形重力波、地表植被、海表的水汽热交换等等都会引起风速、风向的变化，导致风的运动十分复杂。因此在风的预测中需要考虑温度、气压、密度、相对湿度等多变量约束；此外，从空间角度来看，风的运动尺度既可以小到米至百米级别，如地表湍流，也可以是上千公里的大尺度平流，如平流层上风的萍乡运动，因此在风的预测中还要综合考虑多个空间尺度。

1.2 国内外研究现状及分析

1.2.1 风速预测技术

气象预报作为一门预测学科，存在一定的不确定性。根据使用模型的不同，通常将预测方法分为两种：一种是数值预报方法，一种是统计预报方法。不同的预测方法目的都是以各种方式减少未知的不确定性因素。数值预报方法是先建立起描述天气变化的满足流体力学和热力学关系偏微分方程组，通过超级计算机，在恰当初值的条件下，用数值方法解方程组求得变量值进行预测。统计预报方法是利用概率统计，分析历史数据的特点，建立回归模型对风速数据进行拟合和预测，推测在未来某个环境下出现的可能性。两类方法也不是互斥孤立的，两类模型各自有各自的特点，在某些应用场合中，也有结合二者特点的混合预报方法。气象预报的分类[5]如图 1-1。

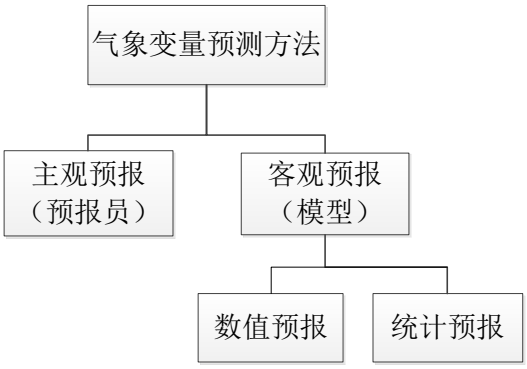


图 1-1 气象预报预测方法分类

1.2.1.1 数值预报方法

数值预报方法是以流体力学和热力学偏微分方程组为基础，通过大规模的数据处理，利用高性能计算机的并行计算优势求数值解，预测未来天气的方法。数值预报方法是一种客观的方法，可以克服统计计算方法中依赖于统计人员经验、预报时效短、难以充分融合多传感器数据的缺陷。数值天气预报所使用的方程组是由牛顿运动定律、质量守恒定律、能量守恒定律、热力学方程、水汽方程、状态方程等组成。得到方程组后，通过资料同化技术求解预定时刻的大气状态的初值，从而得到模式的预测信息，预测结果是速度的水平分量u、v、垂直分量w、温度T、气压p等[6-8]。目前气象部门、军事单位、政府机构的天气预报以数值预报为基础。

1.2.1.2 统计预报方法

统计预报方法是建立预测回归模型，将历史数据作为模型输入，训练得到特定模型参数，得到未来预测值的方法[9]。这种方法主要是基于统计学习理论，包括概率统计、机器学习、深度学习等方法。相比于数值预报方法，历史数据的质量和数量对统计预测方法的准确性至关重要。此时的模型并不追求物理学上的规律，更多的规则是由历史数据和具体训练方法协同确定的。统计预报方法常见的

方法种类有：

（1）Kalman 滤波法。

Kalman 滤波假设动态系统是线性的，并且后验概率分布是高斯分布，系统噪声和测量噪声也是高斯分布。Kalman 滤波需要在已知初始状态统计特性的基础上，建立一套观测方程与状态方程，不断进行时间更新和状态更新，自回归地递推运行下去以求得最小误差的最优解[10]。Kalman 滤波的条件比较苛刻，必须知道系统噪声和测量噪声的协方差矩阵，而在实际应用中，噪声的协方差矩阵很难求得[11]。

（2）时间序列分析法。

时间序列分析法就是利用已经获得的观测数据，通过曲线拟合和参数估计（例如最小二乘法）建立数学模型预测未来时次数据的方法。时间序列分析法需要求自相关函数来分析相关性，建立合适的序列模型来进行拟合。以常用的自回归滑动平均模型（ARIMA）和向量自回归模型（VAR）模型为例，它们都是处理时间序列。Arash 根据近地面风场的特点，设计了同时考虑时间和空间的 GSTAR 模型，效果明显优于 ARMA 和 VAR[12]。

（3）人工神经网络法[6,13-16]。

人工神经网络是模拟人脑神经元传递信息的方式，由大量输入层、隐藏层、输出层组成的层与层之间相互连接的一种拓扑结构。人工神经网络层与层之间的神经元以一定权值连接，每个神经元的输出都有激活函数。回归方法与分类方法的不同就在于输出层的激活函数。理论上，三层人工神经网络可以以任意精度拟合任意函数。但是实际应用中，人工神经网络训练过程中容易陷入局部极小，并且神经网络对权重的初值敏感[17]。

（4）支持向量机法[18,19]。

SVM 的最大特点是对于非线性可分问题，采用核函数将低维特征空间线性不可分的样本转化为高维特征空间线性可分的数据，从而在高维特征空间采用线性分析方法对样本分析。SVM 学习的问题可以转化为凸优化问题，即寻找全局最小值。而其他方法如多层感知机，都是采用贪心测量来搜索目标函数极小值，非常容易陷入局部极小而非全局最小。风速序列预测问题上，SVM 与多层感知机方法对比，结果表明 SVM 的预测效果更好，均方根误差更小[20]。

（5）模糊逻辑法。

模糊逻辑法需要专业人员建立模糊规则库，根据专业人员的经验选择线性模型不断逼近非线性系统。模糊逻辑法可以很好地解决不确定性大的问题，但是这种方法的学习能力较弱，目前还需要结合其他方法一起使用[21]。

（6）组合方法。

应对复杂的气象业务背景，单一的预测方法不能满足有着复杂规律的风速序

列处理，因此更多的组合方法被挖掘出来[22]。刘鸣凤等人将 Kalman 滤波与 ARMA 方法结合起来，预测结果的延迟比传统 Kalman 方法少，且在风速序列极值点的误差原小于单一的 BP 神经网络方法[23]。Zhenghai Guo 等人使用经验模态分析（EMD）分解风速信号，在得到的多个分解信号上分别使用一个神经网络，得到了较好的预测效果[24,25]。席剑辉等人将主成分分析和神经网络结合应用在多元变量非线性时间序列预测中，在降雨、气温序列的验证中证明了有效性[26]。

1.2.2 空间插值方法

空间插值算法目的在于利用离散的观测点测量数据估算同一区域中未采样点的估计值，获得高精度的数据和高分辨率空间表面模型。

传统空间插值有反距离加权方法（inverse distance weighted，IDW）、克里金（Kriging）等，但是函数本身并未考虑地形、空间位置、多影响因子的影响。有些方法模型本身就可以模拟复杂过程，例如张升堂等人发现椭圆指数函数模型能在降水数据插值中指明降水中心位置以及中心降水量，与反距离加权方法、空间线性插值方法、修正反距离方法相比，具有更高精度[27]。

许多方法也在传统插值方法基础上对参数和变量值进行修改以引入地形因素。樊子德等人提出了一种考虑多因素影响的自适应反距离加权插值算法（ACGIDW）。以气象数据为例，考虑复杂地形因素、经纬度和高程的影响，根据采样点空间分布对反距离加权算法中的距离衰减参数α进行自适应调整，这种方法与其他反距离加权方法、普通克里金方法对比具有较高的精度和自适应性[28]。杨成生等人在对 GPS 可将水汽观测值进行空间插值时，改进了 Kriging 插值方法，不仅考虑了插值点与样本点之间的一维距离，还考虑了插值点与样本点的高程差，改进的 Kriging 方法表现出了明显的优越性[29]。

除了地形、空间位置等地理因素，某些变量也要考虑其他影响因子。多元线性回归插值方法在影响因素较多的空间插值计算中应用广泛。黄安等人利用多元线性回归分析继承所有成土因子对土壤养分进行空间分布预测，结果表明虽然多元线性回归预测结果与 Kriging 预测结果在宏观上是一致的，但是多元线性回归预测的土壤有机质空间分布考虑了多种成土因子的变化特征，克服了传统插值方法存在的斑块状分布，预测结果更加精确[30]。吴小芳等人根据农作物病虫害空间插值的特殊性，提出了一种基于空间方位关系、拓扑关系、距离关系以及自然气候条件影响的多因子插值模型。这种该方法可更好地表现农作物病虫害时空变化规律[31]。

除了基于地理因素的插值外，在一些考虑时空相关的序列预测中会同时考虑时间和空间插值。风力发电行业，通常通过预测发电场中每个风力发电机风速大小从而预测发电量。Arash Pourhabib 等人同时考虑风速数据的时空相关性，改进

了自回归滑动平均模型（ARMA）和向量回归模型（VAR）中的单一时间变量，每个空间位置的时间项都增加了一个由时间衰减参数和空间协方差参数构成的综合系数。与前二种时间序列预测相比，加入空间相关关系的序列预测方法表现出了更优异的性能[32]。

模拟包含许多不确定因子的序列模型也会考虑高斯过程回归（Gaussian Process Regression, GPR）方法，该方法特点是可以表示不同尺度上的变化。Christopher J. Moore 等人用 GPR 模型模拟两个黑洞之间未知尺度和参数的重力波，这种方法提高了重力波参数估计的准确性[33]。Aditya Grover 等人在基于深度学习的天气预测系统中，风速的空间插值使用了 GPR+DBN（深度置信网络，Deep Belief Network）的方式，其中 GPR 的核函数采用多个变量单一尺度核函数乘积的方式，该方法的预测结果优于数值天气预报结果[34]。GPR 方法在空间序列插值方面的应用不多，在时间序列预测中应用得更多一些，这与其多尺度变化的描述能力有关。通常使用 GPR 对大气变量时间序列预测时，会采用长期、中短期、短期多个时间尺度的核函数的组合模型，例如大气、水流、地表温度时间序列[35-37]、长期、短期风速时间序列等[38,39]。

1.2.3 多变量时间序列预测方法

多变量时间序列数据广泛存在于多媒体[40,41]、气象[42,43]、农业[44]、金融[45]、环境[46]等领域，如何有效地管理和应用这些历史数据，使之为变量预测、内在规律提供方向，是一个具有重要意义的论题。序列分析也是数据挖掘的一项重要内容。

有的方法主要思想是采用信号分解，将单一信号分解成多束信号时间序列进行分析。谭忠富等人提出了一种多因素小波变换和多变量时间序列模型的电价预测方法，利用小波变换解构历史电价序列和负荷序列，建立多源序列模型，这种方法比单一的电价预测模型能提供更准确的预测[47]。

有的方法主要思想是将多变量组成一个特征向量，预测单变量的过程就转变为预测特征向量的出现概率。李权等人提出一种了采用 KPCA 技术获取多变量时间序列数据高维特征空间的主成方向矢量，使用主成方向矢量内积作为异常度量，检测过程中通过计算实际数据主成方向矢量出现的概率大小来判断异常的发生，这种方法与传统方法比较，不依赖于先验的专家知识，具有较高的有效性[48]。

有的方法首先要对多变量的数据进行聚类，再将高维数据转变为低维的时间序列进行预测。一般的聚类方法多采用 K-means 方法对低维数据进行聚类，但是该方法不能有效应用高维多变量时间序列（MTS）。周大镯等人提出的多变量时间序列聚类算法（PCA-CLUSTER）利用主成分分析对 MTS 降维，选取主成分序列再进行 KNN 聚类分析，实验结果效果显著[41]。Yang K 等人基于主成分

分析（PCA）提出了一种 MTS 数据集相似度度量的方法，将 MTS 数据集用矩阵表示来产生主成分和相关特征值，然后用这些主成分特征值来比较 MTS 之间的相似性，实验表明这种方法优于传统相似度度量方法，如欧氏距离（ED）、动态时间规整（DTW）、加权和 SVD（WSSVD）、PCA 相似系数（SPCA）等[49]。

有的方法使用神经网络对多变量序列信号进行控制。Hwang CL 等人应用循环神经网络（RNN）学习非线性自回归移动平均模型（NARMA）的动态特征，为了克服系统向量函数变化波动大的缺点，用一个简单网络来补偿 RNN 线性化近似误差的残差上界来修正学习率，从而提出了基于 RNN 的多变量自适应控制系统[50]。

作为 RNN 的升级版，长短期记忆网络（LSTM）在克服了 RNN 误差消失的问题，可以连接起输出目标和输出目标之间长时间的滞后关系，从而组合整个上下文结构。LSTM 可以用来重构可预测、不可预测、周期性、非周期性、准周期性的时间序列[51]，在语言建模[52,53]、语言理解[54,55]、机器翻译[56]等自然语言处理领域，LSTM 已经被证实具有非常大的优势。在机器翻译领域，将一个词通过 Word Embedding 技术从输入域词维度转换成向量表示，那么一句话就变成了一个高维向量的时间序列，输出结果为输出域所对应的高维向量时间序列，将这些输出向量映射到输出域，就得到了翻译结果[56]。这种序列到序列（Sequence to Sequence，Seq2Seq）方法应用到数值计算领域，就可省略 Word Embedding 映射过程，输入和输出结果直接由数值表示。Mohamed Akram Zaytar 等人利用 LSTM 进行 24 小时和 72 小时序列到序列天气预报，这两个时间步长预测的 RMSE 都非常小，这说明这种时间序列预测方法在纯数值领域也有发挥空间[57]。

以上是比较常见的多变量时序序列方面的研究，还有许多其他方法进行多变量和时间序列结合的预测，尤其是与领域特点和变量特点结合，这样的方法对于特定问题更有适用性。

1.3 研究内容及结构

1.3.1 研究内容

如图 1-2，传统统计预报方法有许多是基于机器学习或者深度学习。先对空间中的点进行单点、单变量的时间序列预测，再根据其空间位置进行插值。这种方法的缺陷：首先，单个点上，预测过程缺乏了经向风、纬向风、温度、气压、密度、湿度、露点温度、位势高度等变量的约束，体现不出成因复杂、影响因子众多的气象变量的特点。其次，除了垂直方向的温度、密度、气压插值会考虑理想气体定律和流体静力学关系，风速、水平方向的温度、密度、气压等其他变量

插值往往都是基于单一的简单模式，如线性插值、二维均匀插值、应用简单核函数的高斯过程回归等。

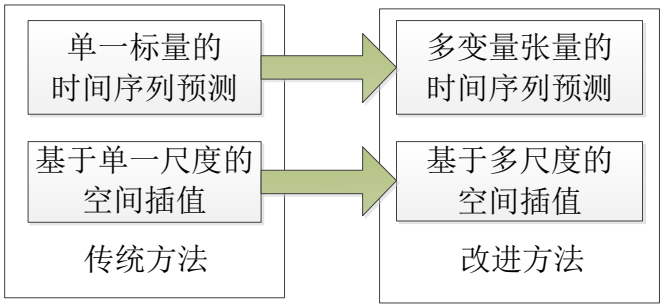


图 1-2 本文对气象序列预测改进思路

针对这个问题，本文主要从两方面进行考虑：

（1）将单一标量预测变为多变量张量的预测。

考虑到多变量的约束关系，一种有效方法是神经网络。神经网络对于非线性关系的拟合具有非常好的效果。由于我们需要对时间序列进行预测，因此可以采用具有记忆功能的循环神经网络，能够处理前后关联的问题。本文将多个变量组成的状态向量加入到循环神经网络中，讨论不同状态向量对预测结果的影响。

（2）将单一尺度空间插值改进为多尺度空间插值。

气象变量序列之间是复杂的非线性关系。我们需要将复杂的非线性关系变为易于处理的计算方式。高斯过程回归多数都应用于时间序列预测，但是对于空间序列，高斯过程回归也有很好的拟合与预测效果。难点是需要根据气象变量的特点，找到合适的核函数结构和合适的超参数。考虑到大气运动和大气科学、流体力学、热力学等方面的约束关系，建立新的核函数，并研究优化结构对预测结果的影响。

1.3.2 组织结构

本文内容共分为五章。具体内容安排如下：

第一章，引言。首先介绍了课题研究背景及意义，并总结了国内外风速预测研究、空间序列研究、时间序列研究发展现状，最后给出了研究内容和文章结构。

第二章，相关序列预测方法介绍。首先介绍了高斯过程回归的理论依据、核函数的分类和超参数的确定方法，然后介绍了长短期记忆网络的各种门结构和前向传播、后向传播。

第三章，基于高斯过程回归的多尺度的空间序列预测方法。首先，介绍了物理学上的多尺度的概念，然后，讨论了核函数的选择和超参数的选择对插值结果的影响，并针对这些特点分别建立了多尺度和多变量的核函数。

第四章，基于 LSTM 的多变量约束时间序列预测方法。首先介绍了基于

LSTM 的 Seq2Seq 的时间序列预测方法，之后再此基础上对风速状态向量进行时间序列预测，分析不同时间步长和风速状态向量对预测结果的影响。

第五章，结论与展望。总结了全文所做的研究，指出了论文的不足，讨论了下一步工作的方向。

第二章 相关序列预测方法

本章主要介绍两大部分：高斯过程回归理论（GPR）和基于长短期记忆网络（LSTM）的序列到序列（Sequence to Sequence）方法。第一大部分中，首先介绍高斯过程回归不同的表述方式，然后介绍核函数的定义和类型，最后介绍利用极大似然估计计算超参数的方法。第二大部分中，首先沿着历史发展介绍人工神经网络、循环神经网络、长短期记忆网络，随后引出基于长短期记忆网络的序列到序列预测方法，最后介绍神经网络中损失函数最小值计算中常用的优化方法。

2.1 高斯过程回归（GPR）理论

高斯过程是一类监督学习理论，应用于从经验数据集（训练集）中学习训练集与测试集的某种映射关系。高斯过程是一个随机变量的集合，集合中任意有限个随机变量都服从联合高斯分布。本质上，高斯过程就是一个多元高斯分布，形式如下：

$$f \sim GP\left(m,k\right), \tag{2-1}$$

其中，f 是均值函数为 m，协方差函数为 k 的高斯过程分布函数。

高斯过程可以分为回归和分类两类问题。分类问题的输出是离散分类标签，回归问题的输出是预测的连续值。例如，在一个经济学应用中，我们可能想知道某种商品价格受利率、货币汇率、供需关系变化的影响。这一章，我们主要讨论高斯过程回归理论。

高斯过程（GP）回归模型有几种表示方式：一种高斯过程回归可以看做分布函数，在空间函数中直接推理，成为函数空间论（function-space view）；还有一种更熟悉的、更易理解的表示方式是权重空间论（weight-space view）。高斯过程的超参数和核函数是其特点之一，通常随着其超参数和核函数的形式不同，其分布特点会发生变化。高斯模型的预测可以提供一种全覆盖的预测方式。

2.1.1 权值空间论定义

本节用权值空间[58]的角度解释高斯过程回归理论。简单线性回归模型的输出是输入数据的线性组合，其主要优势在于易于实现和解释。主要缺点是复杂度要求太严格，如果输入输出的关系不能近似用一个线性函数来表示，预测结果就会非常差。

假设训练集 D 有 n 个观测， $D=\{(x_i,y_i)\,|\,i=1,...,n\}$，其中 $\boldsymbol{x}$ 是 d 维输入向量（自变量）， $\boldsymbol{y}$ 是一个标量输出（因变量）。n 个实例的列向量输入组成了 d×n 维的矩阵 $\boldsymbol{X}$，目标结果用向量 $\boldsymbol{y}$ 表示，所以我们写作 $D=(\boldsymbol{X},\boldsymbol{y})$。回归中的目标值都是实数。我们并不关心输入的分布，但是我们关心在已知输入条件下，目标值的条件分布，这可以用贝叶斯理论来表示。

等式（2-2）是一个基于贝叶斯理论的带高斯噪声的标准线性回归模型[58]

$$f(\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{w},\ y=f(\boldsymbol{x})+\varepsilon, \tag{2-2}$$

其中 $\boldsymbol{x}$ 是输入向量， $\boldsymbol{w}=\{w_1,\ldots,w_n\}^T$ 是线性模型的权重（参数）向量， f 是函数值， $\boldsymbol{y}$ 是观测目标值。假设观测值 $\boldsymbol{y}$ 和函数值之间相差一个噪声值，再假设该噪声是均值为 0、方差为 σ_n^2 的独立同一高斯分布 $\varepsilon\sim N\left(0,\sigma_n^2\right)$ 。

这类问题的实质是要通过观测值 $\boldsymbol{y}$ 来求权重向量 $\boldsymbol{w}$。假设变量 $\boldsymbol{X}=\{x_1,\ldots,x_n\}^T$ ， $\boldsymbol{y}=\{y_1,\ldots,y_n\}^T$ ，分别是独立的，结合模型已知的概率密度的参数，可得到目标观测值的先验概率似然函数[58]

$$\begin{aligned} p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}\right)&=\prod_{i=1}^{n}p\left(y_i\,|\,\boldsymbol{x}_i,\boldsymbol{w}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{\left(y_i-x_i^T\boldsymbol{w}\right)^2}{2\sigma_n^2}\right)\\ &=\frac{1}{\left(2\pi\sigma_n^2\right)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma_n^2}\left|\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}\right|^2\right)=N\left(\boldsymbol{X}^T\boldsymbol{w},\sigma_n^2I\right)\end{aligned} \tag{2-3}$$

在贝叶斯形式中，我们需要确定 $\boldsymbol{w}$ 的先验概率分布的参数。我们假设 $\boldsymbol{w}$ 服从均值为 0、协方差矩阵为 Σ_p 的高斯分布 $\boldsymbol{w}\sim N\left(0,\Sigma_p\right)$[58]。

$$p(w)=\frac{1}{(2\pi)^{n/2}}\frac{1}{\left|\Sigma_p\right|^{1/2}}exp\left(-\frac{1}{2}\boldsymbol{w}\Sigma_p^{-1}\boldsymbol{w}\right) \tag{2-4}$$

根据贝叶斯理论，权重的后验分布概率密度为

$$p\left(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}\right)=\frac{p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}\right)p\left(\boldsymbol{w}\right)}{p(\boldsymbol{y}\,|\,\boldsymbol{X})} \tag{2-5}$$

其中的分母的常数项 $p(\boldsymbol{y}\,|\,\boldsymbol{X})$ 又称作边缘似然，与权重 $\boldsymbol{w}$ 无关。而 $p(\boldsymbol{y}\,|\,\boldsymbol{X})$ 又可以用如下形式表示

$$\mathrm{p}\left(\boldsymbol{y}|\boldsymbol{X}\right)=\int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\,p(\boldsymbol{w})dw \tag{2-6}$$

因此我们可以得到这样的表示：

$$p\left(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}\right)\propto p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\,p(\boldsymbol{w}) \tag{2-7}$$

带入（2-3）和（2-4）我们可以得到

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) &\propto exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}))exp(-\frac{1}{2}\boldsymbol{w}^T\sum\nolimits_p^{-1}\boldsymbol{w}) \\ &\propto exp(-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^T(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T+\sum\nolimits_p^{-1})(\boldsymbol{w}-\bar{\boldsymbol{w}})) \end{aligned} \tag{2-8}$$

其中，$\bar{\boldsymbol{w}}=\sigma_n^{-2}\left(\sigma_n^{-2}\boldsymbol{X}\boldsymbol{X}^T+\sum\nolimits_p^{-1}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$。我们认为，这种后验概率分布服从均值为$\bar{\boldsymbol{w}}$协方差矩阵为$\boldsymbol{A}^{-1}$的高斯分布[58]：

$$p\left(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}\right)\sim N\left(\bar{\boldsymbol{w}}=\sigma_n^{-2}\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{y},\boldsymbol{A}^{-1}\right) \tag{2-9}$$

其中，$\boldsymbol{A}=\sigma_n^{-2}\boldsymbol{X}\boldsymbol{X}^T+\sum\nolimits_p^{-1}$。等式（2-9）中，$p\left(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}\right)$的均值又称为$\boldsymbol{w}$的最大后验概率估计。因此，已知输入$\boldsymbol{X}$和对应的目标观测$\boldsymbol{y}$，我们可以得到$\boldsymbol{w}$的后验概率分布。

我们最终的目的是要通过输入$\boldsymbol{X}$和目标观测$\boldsymbol{y}$，得到新位置的预测值。假设新的输入集为$\boldsymbol{x}_*\in\boldsymbol{R}^n$，我们令预测值的分布$\boldsymbol{f}_*\triangleq f\left(\boldsymbol{x}_*\right)$，则预测值$\boldsymbol{f}_*$的概率分布为[58]

$$p\left(\boldsymbol{f}_*|\boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}\right)=\int p\left(\boldsymbol{f}_*|\boldsymbol{x}_*,\boldsymbol{w}\right)p\left(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}\right)d\boldsymbol{w}=N\left(\sigma_n^{-2}\boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{y},\boldsymbol{x}_*^T\boldsymbol{A}^{-1}\boldsymbol{x}_*\right) \tag{2-10}$$

等式（2-10）可以看出，预测值的分布还是一个高斯分布，均值是（2-9）中权重的后验均值乘以测试集输入，方差是后验协方差矩阵作为系数的测试集输入向量的二次型形式。因此，预测的均值、方差随着测试输入规模增大而呈现线性、平方规模增大，因此这个模型只适应于线性函数。

在非线性的问题中，我们通常把低维的非线性数据投射到高维空间中，使得非线性的数据在高维空间线性可分，此时，我们就可以运用线性贝叶斯模型来处理非线性函数关系的问题。例如，标量x投射到x的幂空间$\varphi\left(x\right)=\left(1,x,x^2,x^3,\ldots.\right)^T$来实现多项式回归。只要这些映射是固定的，该模型的参数就依然是线性可解的。

假设$\varphi(x)$是向量D维向量$\boldsymbol{x}$到N维特征空间的一个映射，且$N>D$。我们用

$\Phi(\boldsymbol{x})$表示$\varphi(\boldsymbol{x})$的长度为 n 的集合，即 $\Phi(\boldsymbol{x})=\left[\varphi(x_1),\ldots\varphi(x_n)\right]^T$ 。则现在模型为

$$f(\boldsymbol{x})=\varphi(\boldsymbol{x})^T\boldsymbol{w} \tag{2-11}$$

其中，参数向量长度为N。这个模型与线性模型相似，只是把$\boldsymbol{X}$替换成了$\Phi(\boldsymbol{x})$。从而预测分布满足

$$\boldsymbol{f}_*\mid \boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}\sim N\left(\sigma_n^{-2}\varphi(\boldsymbol{x}_*)^T\boldsymbol{A}^{-1}\boldsymbol{\Phi}\boldsymbol{y},\varphi(\boldsymbol{x}_*)^T\boldsymbol{A}^{-1}\varphi(\boldsymbol{x}_*)\right) \tag{2-12}$$

其中，$\boldsymbol{\Phi}=\Phi(\boldsymbol{x})$，$\boldsymbol{A}=\sigma_n^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\sum_p^{-1}$。由于$\boldsymbol{A}$的矩阵大小是 N×N 的，如果特征空间的维度 N 非常大，计算过程会非常不便。根据矩阵求逆法则，我们以如下方式重写等式（2-12）得到

$$\boldsymbol{f}_*\mid \boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}\sim N(\boldsymbol{\varphi}_*^T\sum_p\boldsymbol{\Phi}\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y},\boldsymbol{\varphi}_*^T\sum_p\boldsymbol{\varphi}_*-\boldsymbol{\varphi}_*^T\sum_p\boldsymbol{\Phi}\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^T\sum_p\boldsymbol{\varphi}_*) \tag{2-13}$$

其中$\varphi(\boldsymbol{x}_*)=\boldsymbol{\varphi}_*$，$\boldsymbol{K}=\boldsymbol{\Phi}^T\sum_p\boldsymbol{\Phi}$。

现在，特征空间只与$\boldsymbol{\Phi}^T\sum_p\boldsymbol{\Phi}$、$\boldsymbol{\varphi}_*^T\sum_p\boldsymbol{\Phi}$和$\boldsymbol{\varphi}_*^T\sum_p\boldsymbol{\varphi}_*$有关。而$\varphi(\boldsymbol{x})^T\sum_p\varphi(\boldsymbol{x}')$中的$\boldsymbol{x}$和$\boldsymbol{x}'$无论在训练集还是测试集中也都是常数。由于$\sum_p$是正定的，定义$\left(\sum_p^{1/2}\right)^2=\sum_p$。令$k(\boldsymbol{x},\boldsymbol{x}')=\varphi(\boldsymbol{x})^T\sum_p\varphi(\boldsymbol{x}')$，$k(\cdot,\cdot)$称为协方差函数（核函数）。经过奇异值分解$\sum_p=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$，$D$是对角阵，令$\Psi(\boldsymbol{x})=\sum_p^{1/2}\varphi(\boldsymbol{x})$，则$k(\boldsymbol{x},\boldsymbol{x}')=\Psi(\boldsymbol{x})^T\Psi(\boldsymbol{x}')$，$k$是高维映射函数的内积形式。

令$k(\boldsymbol{x}_*,\boldsymbol{x})=\varphi(\boldsymbol{x}_*)^T\sum_p\Phi(\boldsymbol{x})=\boldsymbol{k}_*$，$k(\boldsymbol{x}_*,\boldsymbol{x}_*)=\varphi(\boldsymbol{x}_*)^T\sum_p\varphi(\boldsymbol{x}_*)=\boldsymbol{k}_{**}$，即得[58]

$$\boldsymbol{f}_*\mid \boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y}\sim N(\boldsymbol{k}_*\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y},\boldsymbol{k}_{**}-\boldsymbol{k}_*\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{k}_*^T) \tag{2-14}$$

其中，$\boldsymbol{k}_*\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}$为预测值，$\boldsymbol{k}_{**}-\boldsymbol{k}_*\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{k}_*^T$为预测方差。

2.1.2 函数空间论定义

这一节，我们用函数空间[58]的角度解释前一章的另一种等价的表述。这一节用一个高斯过程近似表示函数的分布。通常定义：高斯过程是任何有限个服从联合高斯分布的随机变量的集合。

高斯过程完全是由其均值函数和协方差函数定义的。假设一个高斯过程 $f(x)$，均值是 $m(x)=E\left[f(x)\right]$，协方差函数是 $k(x,x')=E\left[\left(f(x)-m(x)\right)\left(f(x')-m(x')\right)\right]$，则该高斯过程 $f(x)$ 可记作：

$$f(x)=GP\left(m(x),k(x,x')\right) \tag{2-15}$$

通常为了计数方便会把均值函数设置为 0，不过这也并不是必须的。

我们假设一个贝叶斯线性回归模型 $f(\boldsymbol{x})=\varphi(\boldsymbol{x})^T\boldsymbol{w}$，其中先验 $\boldsymbol{w}$ 满足 $\boldsymbol{w}\sim N(0,\Sigma_p)$，即 $E(\boldsymbol{w})=0$，$E(\boldsymbol{w}\boldsymbol{w}^T)=\Sigma_p$。则我们得到均值和方差[58]：

$$\begin{aligned} E\left[f(\boldsymbol{x})\right]&=\varphi(\boldsymbol{x})^T E(\boldsymbol{w})=0 \\ E\left[f(\boldsymbol{x})f(\boldsymbol{x}')\right]&=\varphi(\boldsymbol{x})^T E(\boldsymbol{w}\boldsymbol{w}^T)\varphi(\boldsymbol{x}')=\varphi(\boldsymbol{x})^T\Sigma_p\varphi(\boldsymbol{x}') \end{aligned} \tag{2-16}$$

因此 $f(\boldsymbol{x})$ 与 $f(\boldsymbol{x}')$ 都服从均值为 0、协方差为 $\varphi(\boldsymbol{x})^T\Sigma_p\varphi(\boldsymbol{x}')$ 的联合高斯分布。也就是说，任意 n 个输入点对应的函数值 $f(x_1),\ldots,f(x_k)$ 都服从联合高斯分布 $N(0,\varphi(\boldsymbol{x})^T\Sigma_p\varphi(\boldsymbol{x}'))$，虽然当联合协方差矩阵的阶数 $N<n$ 时，这时的高斯过程是奇异的。

本章我们采用平方指数协方差函数（SE）为协方差函数。协方差函数明确了两个随机变量之间的协方差[58]：

$$\operatorname{cov}\left(f(x_p),f(x_q)\right)=k(x_p,x_q)=\exp\left(-\frac{1}{2}\left|x_p-x_q\right|^2\right) \tag{2-17}$$

我们可以将输出值的协方差视作是输入值的一个函数，即由前者的相关性是由后者的相关性表示的。对于具体的协方差函数，两个非常相近的相关输入变量之间的协方差几乎是一致的，并且随着输入距离的增大而不断减小。

我们主要关注的并不是根据先验概率描述的随机函数，而是由训练数据得到的函数的结构。首先，我们先假设观测没有噪声的简单情况，即 $y=f(x)$。已知输入集 $\boldsymbol{X}$ 和给出的对应高斯分布函数 $f(\boldsymbol{x})$ 为训练集，$f\sim N\left(0,k(\boldsymbol{x},\boldsymbol{x})\right)$。需要估计的输入集 $\boldsymbol{X}_*$ 和对应的高斯分布函数 $f_*(\boldsymbol{x}_*)$ 为测试集，$f_*\sim N\left(0,k(\boldsymbol{x}_*,\boldsymbol{x}_*)\right)$。训练集输出 $\boldsymbol{f}$ 和测试集输出 $\boldsymbol{f}_*$ 服从联合高斯分布：

$$\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{bmatrix}\sim N\left(0,\begin{bmatrix}K(\boldsymbol{X},\boldsymbol{X}) & K(\boldsymbol{X},\boldsymbol{X}_*)\\ K(\boldsymbol{X}_*,\boldsymbol{X}) & K(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right) \tag{2-18}$$

假设训练集中有n个点，测试集中有n_*个点，上式中的协方差$K(\boldsymbol{X},\boldsymbol{X}_*)$为$n\times n_*$矩阵，协方差$K(\boldsymbol{X},\boldsymbol{X})$为$n\times n$矩阵，协方差$K(\boldsymbol{X}_*,\boldsymbol{X}_*)$为$n_*\times n_*$矩阵。根据已知的观测点，我们得到满足的先验概率，由此我们得到高斯分布函数的后验概率：

$$\boldsymbol{f}_*\mid \boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{f}\sim N(K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X}_*)^{-1}\boldsymbol{f},K(\boldsymbol{X}_*,\boldsymbol{X}_*)-K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X}_*)^{-1}K(\boldsymbol{X},\boldsymbol{X}_*)) \tag{2-19}$$

根据(2-19)，已知$\boldsymbol{X}$和$\boldsymbol{X}_*$的情况下，我们可得$\boldsymbol{f}_*$概率分布的均值和协方差矩阵。

实际情况中，我们很难得到观测中函数值$\boldsymbol{f}$的值本身，我们得到的观测值y是一个带有噪声的项。考虑引入噪声，观测值就为$\boldsymbol{y}=f(\boldsymbol{x})+\varepsilon$。假设添加独立同分布的高斯噪声$\varepsilon\sim N(0,\sigma_n^2)$，则带有噪声的观测值的：

$$cov(y_p,y_q)=k(x_p,x_q)+\sigma_n^2\delta_{pq}?\ cov(\boldsymbol{y})=K(\boldsymbol{X}\ \boldsymbol{X})+\sigma_n^2\boldsymbol{I} \tag{2-20}$$

其中，当$p=q$时，$\delta_{pq}=0$；$\boldsymbol{I}$为单位矩阵。如果$\boldsymbol{y}$满足了包含独立噪声的假设，协方差项就要在无噪声的矩阵上再加上一个对角矩阵。所以，观测目标值和我们测试集需要求的值都服从联合高斯分布，（2-18）的等式我们就可以改写作：

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{f}_*\end{bmatrix}\sim N(0,\begin{bmatrix}K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I} & K(\boldsymbol{X},\boldsymbol{X}_*)\\ K(\boldsymbol{X}_*,\boldsymbol{X}) & K(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}) \tag{2-21}$$

根据等式（2-19），我们可以得到高斯过程回归的关键预测结果，即有噪声情况下，$\boldsymbol{f}_*$的后验概率分布为：

$$\begin{aligned}&\boldsymbol{f}_*\mid \boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{y}\sim N(\overline{\boldsymbol{f}_*},\mathrm{cov}(\boldsymbol{f}_*))\\ &\overline{\boldsymbol{f}_*}=E[\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{y}]=K(\boldsymbol{X}_*,\boldsymbol{X})\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}\\ &\mathrm{cov}(\boldsymbol{f}_*)=K(\boldsymbol{X}_*,\boldsymbol{X}_*)-K(\boldsymbol{X}_*,\boldsymbol{X})\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}K(\boldsymbol{X},\boldsymbol{X}_*)\end{aligned} \tag{2-22}$$

如果测试集的输入为标量，即$\boldsymbol{X}_*=\boldsymbol{x}_*$，$\boldsymbol{f}_*=f(\boldsymbol{x}_*)$，则：

$$\begin{aligned}&\overline{\boldsymbol{f}_*}=E[\boldsymbol{f}_*|\boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{y}]=K(\boldsymbol{x}_*,\boldsymbol{X})\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}\\ &\mathrm{cov}(\boldsymbol{f}_*)=K(\boldsymbol{x}_*,\boldsymbol{x}_*)-K(\boldsymbol{x}_*,\boldsymbol{X})\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}K(\boldsymbol{X},\boldsymbol{x}_*)\end{aligned} \tag{2-23}$$

均值$\overline{\boldsymbol{f}_*}$是观测值$\boldsymbol{y}$的线性组合，这一类预测子被称为线性预测子。另一种解

释是 $\overline{\boldsymbol{f}_*}$ 可看作 n 个核函数的线性组合：假设训练集的目标观测值 $\boldsymbol{y}=[y_1,\ldots,y_n]^T$，则 $\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}=\left[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}\right]^{-1}[y_1,\ldots,y_n]^T=[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n]^T=\boldsymbol{\alpha}^T$，等式（2-23）可改写为[58]：

$$\overline{\boldsymbol{f}_*}=K(\boldsymbol{x}_*,\boldsymbol{X})\boldsymbol{\alpha}^T=\left[k(\boldsymbol{x}_*,\boldsymbol{x}_1),\ldots,k(\boldsymbol{x}_*,\boldsymbol{x}_n)\right]\boldsymbol{\alpha}^T=\sum_{i=1}^{n}k(\boldsymbol{x}_*,\boldsymbol{x}_i)\boldsymbol{\alpha}_i \qquad (2\text{-}24)$$

其中，$\boldsymbol{\alpha}=\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}$。根据（2-24），均值函数可以表示成有限个核函数的和，非线性空间的数据根据一定组合的映射关系映射到线性空间，将非线性问题转化为一个线性问题，从而以线性问题的思路解决问题。

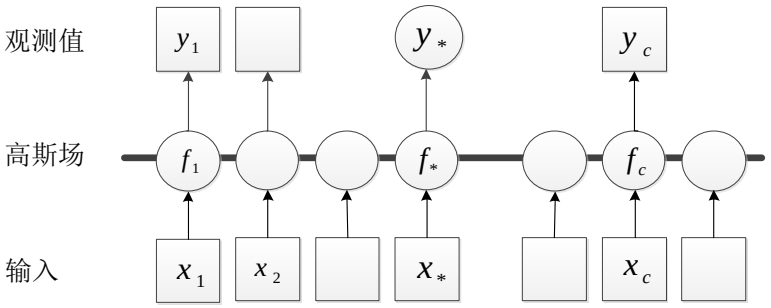


图 2-1 高斯过程回归的链式图模型

总的来说，高斯过程回归如图 2-1 的结构，正方形代表已知观测变量，圆形代表未知值。粗实水平线代表全连接结点的集合。除了 f_1 观测值 y_1 完全独立于其他所有的 $f_2\ldots f_*$。无论测试集的输入 x_* 和未知目标值 y_* 是多少，都不会改变隐藏的 f 的概率分布。不论是权重空间论还是函数空间论，都可以推导出相同的目标预测值的概率分布，我们假设两种情况都已知作为基底的核函数的结构。

2.1.3 核函数的定义和类型

如果一个函数 $k(x,x')$ 可以使低维线性空间不可分的模型 $x\in\boldsymbol{X}\subset R^n$ 通过非线性映射到高维特征空间 $x'\in\boldsymbol{X}'\subset R^m$ 线性可分（$n\ll m$），则这个函数就是核函数。核函数的形式是：$k(x,z)=<\Phi(x),\Phi(z)>$，其中 $<,>$ 为内积。核函数将高维空间的内积运算转化为低维空间的核函数计算，大大减小了计算量，从

而巧妙地解决了高维特征空间计算中的“维度灾难”的问题。核函数矩阵必须是对称矩阵，$k\left(x,x'\right)=k\left(x',x\right)$。协方差函数$k(x_1,x_2;\theta)$是空间随机场中两个随机变量点对应的函数值的二阶混合中心距，是用于衡量两个数据之间的相似程度和自相关关系。协方差矩阵的每个元素都是协方差函数，协方差矩阵一定是对称矩阵，且必须是半正定矩阵。

我们介绍一些常用的各项同性协方差函数。协方差函数的标准形式是$k\left(0\right)=1$。在k项前乘以一个正项常数σ_f^2来表示任意表示的过程方法。下面的r表示两点距离$r=\left|x_p-x_q\right|$。

1. 平方指数（SE）协方差函数：

$$k_{SE}\left(r\right)=\sigma^2\exp\left(-\frac{r^2}{2l^2}\right) \tag{2-25}$$

其中，σ是输出规模参数，l是特征长度尺度参数。如果$\mathrm{r}\to 0$，两个位置非常接近，那么k取到最大值。如果$\mathrm{r}\to\infty$，此时距离远近对核函数影响多大取决于l的取值：如果l取值很小，k趋向于0；如果l取值非常大，k可能会得到一个常数。

2. Matern 协方差函数

$$k_{matern}\left(r\right)=\sigma^2\frac{2^{1-\nu}}{\Gamma\left(\nu\right)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu}K_{\nu}\left(\frac{\sqrt{2\nu}r}{l}\right) \tag{2-26}$$

其中，ν、l是正数，K_{ν}是个修正的 Bessel 函数。特别地，当$\nu=\mathrm{p}+1/2,\mathrm{p}\geq 0$，则：

$$k_{\nu=p+1/2}\left(r\right)=\sigma^2\exp\left(-\frac{\sqrt{2\nu}r}{l}\right)\frac{\Gamma\left(p+1\right)}{\Gamma\left(2p+1\right)}\sum_{i=0}^{p}\frac{\left(p+i\right)!}{i!\left(p-i\right)!}\left(\frac{\sqrt{8\nu}r}{l}\right)^{p-i} \tag{2-27}$$

3. 有理二次协方差函数（RQ）：

$$k_{RQ}\left(r\right)=\sigma^2\left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{2-28}$$

其中，$\alpha,l>0$，有理二次协方差核函数可以看作有限个不同尺度的平方指数协方差核函数之和。我们令$\tau=l^{-2}$，假设$p\left(\tau\mid\alpha,\beta\right)$满足伽玛分布，即$p\left(\tau\mid\alpha,\beta\right)\propto\tau^{\alpha-1}\exp\left(-\alpha\tau/\beta\right)$，得到：

$$k_{RQ}(r)=\int p(\tau\mid\alpha,\beta)k_{SE}(r\mid\tau)d\tau\propto\int\tau^{\alpha-1}\exp\left(-\frac{\alpha\tau}{\beta}\right)\exp\left(-\frac{\tau r^2}{2}\right)d\tau\propto\left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{2-29}$$

当$\alpha\to\infty$，RQ 协方差函数就变成了 SE 协方差函数。

4. 周期协方差函数（PER）：

$$k_{PER}(r)=\sigma^2\exp\left[-\frac{2}{l^2}sin^2(\pi fr)\right] \tag{2-30}$$

其中，f 为频率相关参数。

其余的常用协方差核函数还有：常数核函数、线性核函数、神经网络核函数等。如表 2-1 分别列出了各种常用核函数的稳定性和非退化性[58]。

表 2-1 常用核函数的稳定性和非退化性

协方差函数	表述	稳定性	非退化性
常数	σ_0^2	√	
线性	$\sum_{d=1}^{D}\sigma_d^2\boldsymbol{x}_d\boldsymbol{x}'_d$		
多项式	$(\boldsymbol{x}\cdot\boldsymbol{x}'+\sigma_0^2)^p$		
平方指数	$\exp(-\frac{r^2}{2l^2})$	√	√
Matern	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu}K_{\nu}\left(\frac{\sqrt{2\nu}r}{l}\right)$	√	√
指数	$\exp(-\frac{r}{l})$	√	√
γ-指数	$\exp(-(\frac{r}{l})^{\gamma})$	√	√
有理二次	$\left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha}$	√	√
神经网络	$\sin^{-1}(\frac{2\boldsymbol{x}^T\sum\boldsymbol{x}'}{\sqrt{(1+2\boldsymbol{x}^T\sum\boldsymbol{x})(1+2\boldsymbol{x}'^T\sum\boldsymbol{x}')}})$		√

2.1.4 组合核函数

1. 两个核函数之和也是一个核函数。

假设有随机过程 $f(x)=f_1(x)+f_2(x)$，其中 $f_1(x)$ 和 $f_2(x)$ 相互独立。那么 $k(x,x')=k_1(x,x')+k_2(x,x')$。这种结构可以用来描述多个不同特征长度尺度的组合的场景。

2. 两个核函数之积也是一个核函数。

假设有随机过程 $f(x)=f_1(x)\times f_2(x)$，其中 $f_1(x)$ 和 $f_2(x)$ 相互独立。那么 $k(x,x')=k_1(x,x')\times k_2(x,x')$。这种结构可以用来描述多个不同特征长度尺度的组合的场景。同理，对于 $p\in N$， $k^p(x,x')$ 也是一个有效的核函数。

3. 一个尺度变化的核函数也是一个核函数。

假设 $a(x)$ 是已知的确定函数， $f(x)$ 是一个随机过程， $g(x)=a(x)f(x)$，那么 $\mathrm{cov}(g(x),g(x'))=a(x)k(x,x')a(x')$。令 $a(x)=k^{-\frac{1}{2}}(x,x)$，我们可得到标准化的核函数： $\tilde{k}(x,x')=\dfrac{k(x,x')}{\sqrt{k(x,x)}\sqrt{k(x',x')}}$。这确保了对于所有 x， $\tilde{k}(x,x)=1$。

4. 核函数的卷积依然是一个核函数。

假设一个任意已知的核函数 $h(x,z)$，映射 $g(x)=\int h(x,z)f(z)dz$，则很明显 $cov(g(x),g(x'))=\int h(x,z)k(z,z')h(x',z')dzdz'$ 。

5. 不同空间上核函数的简单加法乘法组合依然是有效核函数。

如果 $k(x_1,x'_1)$ 与 $k(x_2,x'_2)$ 是不同空间域 χ_1 和 χ_2 下的核函数，那么直和 $k(x,x')=k_1(x_1,x'_1)+k_2(x_2,x'_2)$ 和张量积 $k(x,x')=k_1(x_1,x'_1)k_2(x_2,x'_2)$ 都是定义 $\chi_1\times\chi_2$ 空间下的核函数。

上述核函数的组合构建方法只是一些基本方法，还有其他组合方式可以见[58]。不同场景下，针对目标预测值的特点，多种组合核函数的灵活应用会给预测的准确性带来提升的空间。

2.1.5 超参数

核函数的最终确定不仅需要选择核函数的类型，也要选择核函数里的参数。所谓超参数，就是机器学习模型里面控制模型框架的参数，比如聚类方法里面类的个数，或者话题模型里面话题的个数。简言之，就是参数的参数，因为参数本

身符合一个分布，而这个分布中的参数就是超参数。如果选择适当的超参数，理论上高斯过程回归可以逼近任意非线性系统。以常用的平方指数协方差为例

$$k\left(x,x'\right)=\sigma_f^2\exp\left(-\frac{\left(x-x'\right)^2}{2l^2}\right)$$

其中，l为特征长度尺度参数，σ_f^2为输出规模参数。参数集合$\boldsymbol{\theta}=\left\{l,\sigma_f,\sigma_n\right\}$即为超参数，一般通过极大似然法求得，即在已知训练集输出结果的情况下，通过极大化似然函数来获得满足这种情况下的参数估计值，把可能性最大的那个参数$\boldsymbol{\theta}$作为真实$\boldsymbol{\theta}_*$的参数估计。

假设$\boldsymbol{X}$为训练集输入，$\boldsymbol{y}$为训练集输出，$\boldsymbol{\theta}$是所有待求超参数组成的向量，$\boldsymbol{\theta}=\left\{\theta_1,\theta_2,\ldots,\theta_n\right\}$。根据贝叶斯原理：

$$p\left(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X}\right)=\frac{p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{2-31}$$

其中$p(\boldsymbol{y}|\boldsymbol{X})=\sum p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)$。$p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}\right)$称为边缘似然函数，服从高斯分布。

首先建立训练样本条件概率的福对数似然函数，并令其对超参数$\boldsymbol{\theta}$求偏导；然后采用牛顿法、共轭梯度法等优化方法求偏导最小值以得到超参数最优解；最后，将求得的超参数带入（2-23），得到测试集点$\boldsymbol{x}_*$对应的预测值均值$\overline{\boldsymbol{f}_*}$及其方差$\sigma_{f_*}^2$。

负对数似然函数$L\left(\boldsymbol{\theta}\right)$的形式如下[59]：

$$L\left(\boldsymbol{\theta}\right)=-\log\left(p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}\right)\right)=\frac{1}{2}\boldsymbol{y}^T\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}+\frac{1}{2}log\left|\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right|+\frac{n}{2}log2\pi \tag{2-32}$$

其中，只有第一项包含训练集输出$\boldsymbol{y}$，第二项是一个惩罚项，第三项是一个归一化常数，是数据拟合过程中的重要成分。

超参数θ的偏导数形式具体如下[59]：

$$\frac{\partial}{\partial\theta_j}L\left(\boldsymbol{\theta}\right)=\frac{1}{2}tr\left[\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^T-\boldsymbol{C}^{-1}\right)\frac{\partial\boldsymbol{C}}{\partial\theta_j}\right] \tag{2-33}$$

其中，$\boldsymbol{C}=\boldsymbol{K}+\sigma_n^2\boldsymbol{I}$，$\boldsymbol{\alpha}=\left(\boldsymbol{K}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}=\boldsymbol{C}^{-1}\boldsymbol{y}$。

如果输入的训练集样本足够大，那么如上方式学习到的参数会非常逼近实际的后验概率模型。如果输入的训练集样本非常小，参数的训练过程中很容易出现过拟合的情况。因此训练数据样本数量不能太少。

2.2 长短期记忆网络（LSTM）时间序列预测方法

为了讲清楚长短期记忆网络（LSTM）的前向后向传播，需要先介绍基本的多层感知机前向后向传播。之后，引入时间步的概念，介绍处理时间序列的循环神经网络的前向后向传播。针对循环神经网络的缺陷和改进方法，介绍长短期记忆网络的结构和特点。最后，介绍几种常用的最优化方法。

2.2.1 人工神经网络的向前向后传播

人工神经网络（ANN）最初是研究生物学神经元处理信号的过程中产生的数学模型。人工神经网络的基本结构是由小处理单元、节点组成的网络，这些神经元、节点都是按权重相互连接在一起。在最初的生物模型中，节点代表着神经元，连接权重代表着神经元之间的突触的强度。网络的输入经过权重连接扩散到整个网络。ANN 节点之间的活动就是模仿神经元之间的信息传输方式。

ANN 有许多有着不同性质、适用于不同场景的变种。其中的一个重要区别就是是否是有向无环图。有环 ANN 又叫循环神经网络，其反馈、循环机制具体会在 2.2.2 节讲述。无环的 ANN 被称为前馈神经网络（FNNs）。被大众熟知的 FNN 包括感知机、径向基函数网络、自组织映射等。FNN 应用最广泛的就是多层感知机（MLP）[60-62]。

如图 2-2 所示，多层感知机的神经元是通过层间的前馈链接组合在一起的。输入内容由输入层进入网络，通过隐藏层传播到输出层。这个过程就叫做网络的前向传播过程。

由于 MLP 的输出只与当前输入有关，与过去或者未来的输入无关，MLP 更适合当前模式的分类或者标记，而不适合序列的标记或预测。这一点我们在 2.2.2 节继续介绍。

从本质上说，建立 MLP 的目的是拟合一个从输入向量到输出向量的非线性函数，而最终效果就是，建立的网络就是这个函数。其中，MLP 各层权重就是函数的参数，我们训练网络的目的就是训练这些权重。改变权重，单个 MLP 可以适用于许多不同函数实例。理论上已经证明，如果 MLP 只含有一个隐藏层，而这个隐藏层包括足够多数目的非线性神经元，那么这个 MLP 可以近似表示输入为任意精度的任何连续函数[63]。因此我们可以说，MLP 是通用的函数近似方法。图 2-2 中的圆圈表示神经元，里面的“S”表示包含 sigmoid 激活函数，除了输入层，隐藏层和输出层的每一层都会用到激活函数。

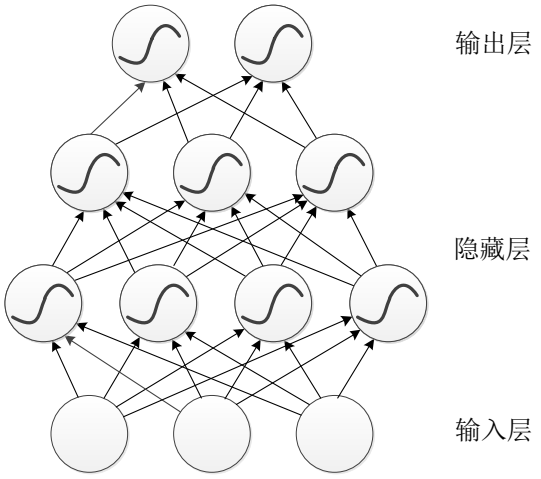


图 2-2 多层感知机模型

1. 前向传播

假设 I 代表输入神经元的个数， x 代表输入向量（即 $|x|=I$ ）。第一个隐藏层的每个神经元计算全部输入神经元的一个权重和。对某个隐藏神经元 h ，前面得到的这个和就作为该层神经元 h 的输入，记作 a_h 。 θ_h 表示每个神经元最后的激活函数。 w_{ij} 表示从神经元 i 到神经元 j 的权重。假设网络一共有 L 个隐藏层， H_{l-1} 表示隐藏层的某层， H_L 表示隐藏层的最后一层。等式（2-34）（2-35）表示的是输入层和隐藏层之间的信息传递。等式（2-36）（2-37）表示的是多层隐藏层的逐层之间的信息传递。即有

$$a_h = \sum_{i=1}^{I} w_{ih} x_i \tag{2-34}$$

$$b_h = \theta_h\left(a_h\right) \tag{2-35}$$

$$a_h = \sum_{h'=1}^{H_{l-1}} w_{h'h} b_{h'} \tag{2-36}$$

$$b_h = \theta_h(a_h) \tag{2-37}$$

如图 2-2 是几种激活函数，激活函数最主要的特征就是“S”形。

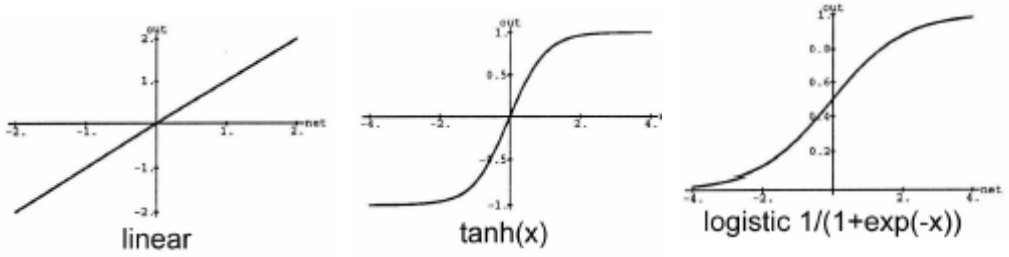


图 2-3 几种常见激活函数

其中（b）tanh 双曲正切函数，（c）是 logistic sigmoid 函数：

$$\tanh(x)=\frac{e^{2x}-1}{e^{2x}+1} \tag{2-38}$$

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{2-39}$$

$$\tanh'(a)=1-\tanh(a)^2 \tag{2-40}$$

$$\sigma'(a)=\sigma(a)(1-\sigma(a)) \tag{2-41}$$

这些不同的激活函数本质是相同的，可以经过一些运算关系相互转化。也就是说，某个隐藏层用tanh作为激活的任意网络都可以用另一个隐藏层用logistic sigmoid作为激活的网络表示。而区分这些激活函数的原因是，对于相同的输入，在不同的应用场景可能需要不同的输出，例如如果输出的内容是一个概率，那么激活函数就需要用logistic sigmoid。

tanh 和 logistic sigmoid 的一个重要特点是——非线性。在非线性分类界限划定或者模拟非线性等式上，非线性网络比线性网络更加强大。此外，线性运算的任意组合都是一个现行运算，也就是说，多个线性运算隐藏层的 MLP 与只有一个线性运算隐藏层的 MLP 是完全一样的。这与非线性网络相比，后者可以通过连续多个隐藏层实现输入数据的复杂表示，可以更好地“记忆”输入数据的特征，从而获得相当大的信息量[64,65]。

另一个关键性质就是，这两个激活函数都是可导的，这一点使得网络可以用梯度下降方法训练。由于我们把无限的输入域转化为了一个有限的输出域，有时候神经网络的激活函数有叫挤压函数（squashing functions）。

MLP 的输出向量 y 是由输出层的激活神经元得出的。每个输出层神经元 k 的输入 a_k 都是由最后一个隐藏层输出的 b_h 求和得到的，即

$$a_k=\sum_{h\in H_L} w_{hk}b_h \tag{2-42}$$

输出层的神经元个数和输出激活函数的选择是根据实际任务需求确定的。例如二元分类为题，激活函数就采用 logistic sigmoid 函数；多元分类问题，激活函数采用 softmax 函数；回归问题，激活函数采用 linear 函数。以分类问题为例，假如最后一个是 softmax 层，分类结果为 C_k 的概率值是

$$p(C_k|x)=y_k=\frac{e^{a_k}}{\sum_{k'=1}^{K}e^{a_{k'}}} \tag{2-43}$$

如果输出的分类结果是完全由 0 和 1 组成的和为 1 向量，如$(0,1,0,0,0)$，1 代表分类结果所在的类。z 代表标签，或者说输出值的真值，y 代表输出为 z 的概率，则最终的模型预测准确率是

$$p(z|x)=\prod_{k=1}^{K}y_k^{z_k} \tag{2-44}$$

2. 目标函数

目标函数是我们需要最小化的目标值。等式（2-45）的形式是交叉熵[66]：

$$L=-\ln\prod_{(x,z)}p(z|x,w)=-\sum_{(x,z)}\ln p(z|x,w) \tag{2-45}$$

假设神经网络最后一层采用的是多分类问题，即采用（2-44），则目标函数形式如(2-46)。交叉熵越小，说明模型的准确率越高。

$$L=-\sum_{(x,z)\in S}\sum_{k=1}^{K}z_k\ln y_k \tag{2-46}$$

最小化目标函数就需要用梯度下降法，每训练一个 mini-batch 的数据更新一下权重。

3. 后向传播

这个过程实际上是从后向前推导误差传播积累的过程。我们根据最后一层的误差递推地计算出前一层直至第一层误差，根据误差值修正权重系数的值，使之误差最小。前向传播和后向传播不断循环进行，直至找到目标函数的最小值。目标函数对 a_k 的导数[66]：

$$\frac{\partial L}{\partial a_k}=\sum_{k'=1}^{K}\frac{\partial L}{\partial y_{k'}}\frac{\partial y_{k'}}{\partial a_k} \tag{2-47}$$

假设最后一层是 softmax 层，则

$$\frac{\partial L}{\partial a_k}=y_k-z_k \tag{2-48}$$

从最后一层往前一层推导。我们令 $\delta_j=\frac{\partial L}{\partial a_j}$，

$$\delta_h=\frac{\partial L}{\partial b_h}\frac{\partial b_h}{\partial a_h}=\frac{\partial b_h}{\partial a_h}\sum_{k=1}^{K}\frac{\partial L}{\partial a_k}\frac{\partial a_k}{\partial b_h} \tag{2-49}$$

由于（2-42），我们可得

$$\delta_h = \theta'(a_j)\sum_{k=1}^{K}\delta_k w_{hk} \tag{2-50}$$

由此，我们可推广到除了最后的 softmax 层的其它层的递归关系

$$\delta_h = \theta'(a_h)\sum_{h'=1}^{H_{l+1}}\delta_{h'} w_{hh'} \tag{2-51}$$

那么，最终目标函数对权重系数的导数变得十分简洁：

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} = \delta_j b_i \tag{2-52}$$

2.2.2 循环神经网络（RNN）前向后向传播

循环神经网络的目的是处理序列数据。传统的神经网络中，数据沿着输入层→隐藏层→输出层的顺序流动，层与层之间是全连接的，但是一层内的神经元之间是相互隔离的。这种神经网络对于需要前后关系的序列数据无能为力。比如，我想预测句子里下一个单词是什么，我需要知道句子里前面的单词，因为一句话里的每个单位都不是孤立的，这是我就需要记录之前时间步的信息，同时作为下一个时间步的输入。循环神经网络就是如此，网络会记录前面的信息并将输出输入到当前计算的神经元上。循环神经网络的结构特点就是，在每一个隐藏层上的输入值包括上一个时刻该隐藏层的输出值。理论上，循环神经网络的序列计算长度是无限的，但是实践中过长的序列表现并不好。因此，我们可以根据实际应用假设当前状态只与之前的几个状态相关。网络原理图如下：

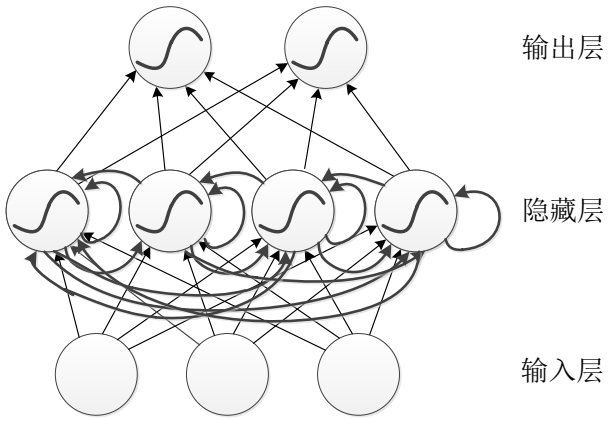


图 2-4 循环神经网络模型

1. 前向传播

RNN 的前向传播和只有一个隐藏层的多层感知机类似，不同之处在于激活函数的输入来源于该时刻的外部输入和前一个时刻的该隐藏层激活函数的输出。假设输入序列 x 的时间长度是 T，I 代表输入神经元 x 的个数，H 个隐藏神经元，

K 个输出神经元。x_i^t 表示时刻 t 时第 i 个输入，a_j^t 和 b_j^t 分别是时刻 t 的第 j 个神经元的输入和激活函数。隐藏层神经元有

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1} \tag{2-53}$$

$$b_h^t = \theta_h(a_h^t) \tag{2-54}$$

a_h^t 是非线性可微的，a_h^t 和 b_h^t 可从 $t=1$ 时刻递推逐时次计算，通常 b_i^0 的初始值设置为 0。而网络的输出层：

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^k \tag{2-55}$$

2. 后向传播

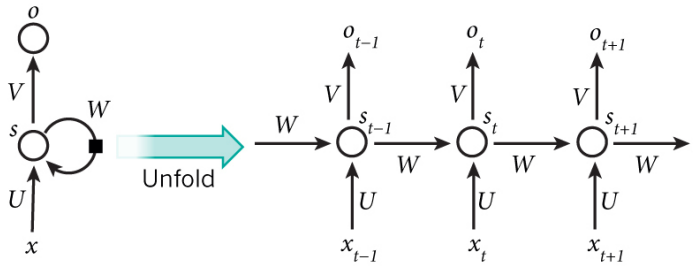


图 2-5 按时间步展开的循环网络

如图 2-5 是展开的循环神经网络，每个节点代表单一时间步的神经网络的一层。输入和隐藏层之间的权重是 U，隐藏层之间的权重是 W，隐藏层和输出层之间的权重是 V。需要注意的是，每个时间步会用相同的权重。

循环神经网络最后一个隐藏层激活函数值不仅仅影响到输出层，还受到隐藏层，也影响到了下一个时间步的隐藏层的值。令 $\delta_j^t = \dfrac{\partial L}{\partial a_j^t}$，可得[66]：

$$\delta_h^t = \theta'(a_h^t)\left(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'}\right) \tag{2-56}$$

该项可从 $t=T$ 开始倒序递推运算每一个时间步，其中我们设 $\delta_j^{T+1} = 0 \forall j$，因为超出序列时间范围是没有误差的。最终，目标函数对权重参数的导数

$$\frac{\partial L}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial L}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t \tag{2-57}$$

2.2.3 长短期记忆网络（LSTM）前向后向传播

在 RNN 的 BPTT 方法中，误差消失比较常见。为了克服该问题，加上了两道控制门：输入门（input gate）和输出门（output gate）。由于门控的限制，cell 在训练时保持内部梯度不受不利变化的干扰。CEC（Constant Error Carousel）是指 Cell 自己对自己，权重为 1 的连接。在计算 CEC 的输入之前，要乘以输入门的输入，计算 CEC 的输出时，将结果乘以输出门的输出。但是这样带来的缺点就是 CEC 的状态值可能会一直增大，我们需要对 CEC 的状态进行控制，因此添加忘记门（forget gate），控制 CEC 的状态值：在需要时使之为 1，意味着“所有数据通过”；否则为 0，意味着“无数据通过”。。系统的这种形式可以记住过去很远距离的值。但是这样也有一个缺点，就是当前的 CEC 状态不能影响到输入门、忘记门在下一时刻的输出，所以增加了窥视孔连接（peephole connections），此时忘记门、输入门的输入来源就增加了 CEC 前一时刻的输出，输出门的输入来源增加了 CEC 当前时刻的输出。LSTM 神经元内部门结构如图：

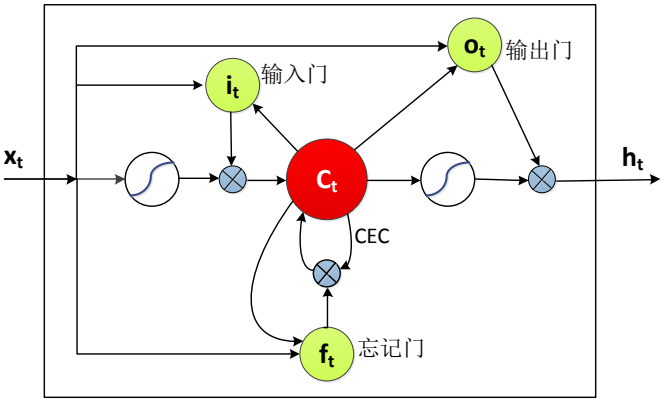


图 2-6 仅包含输入门、输出门、忘记门的简单 LSTM 门

w_{ij} 表示从神经元 i 到 j 的连接权重，a 表示每个神经元输入，b 表示每个神经元的输出。下标 ι、ϕ、ω 分别表示输入门（input gate）、忘记门（forget gate）、输出门（output gate），下标 c 表示 cell，从 cell 到输入、忘记、输出门的 peephole 权重分别是 $w_{c\iota}$、$w_{c\phi}$、$w_{c\omega}$。s_c^t 表示 t 时刻 cell 的状态。控制门的激活函数用 f 表示，cell 的输入、输出激活函数分别用 g、h 表示。I 表示的是输入层神经元的个数，K 是输出层神经元个数，H 是隐藏层 cell 的个数。

1. 前向传播[66]

输入门：

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} \omega_{c\iota} s_c^{t-1} \tag{2-58}$$

$$b_t^t = f(a_t^t) \tag{2-59}$$

忘记门：

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} \omega_{c\phi} s_c^{t-1} \tag{2-60}$$

$$b_\phi^t = f(a_\phi^t) \tag{2-61}$$

Cell:

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1} \tag{2-62}$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_t^t g(a_c^t) \tag{2-63}$$

输出门：

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} \omega_{c\omega} s_c^{t-1} \tag{2-64}$$

$$b_\omega^t = f(a_\omega^t) \tag{2-65}$$

Cell 输出：

$$b_c^t = b_\omega^t h(s_c^t) \tag{2-66}$$

2. 后向传播[66]

令 $\delta_c^t = \frac{\partial L}{\partial b_c^t}$， $\delta_s^t = \frac{\partial L}{\partial s_c^t}$，则

Cell 输出：

$$\delta_c^t = \sum_{k=1}^{K} w_{ck} \delta_k^t + \sum_{h=1}^{H} w_{ch} \delta_h^{t+1} \tag{2-67}$$

输出门：

$$\delta_\omega^t = f'(a_w^a) \sum_{c=1}^{C} h(s_c^t) \delta_c^t \tag{2-68}$$

Cell 状态：

$$\delta_s^t = b_\omega^t h'(s_c^t) \delta_c^t + b_\phi^{t+1} \delta_s^{t+1} + w_{ct} \delta_t^{t+1} + w_{c\phi} \delta_\phi^{t+!} + w_{c\omega} \delta_\omega^{t+1} \tag{2-69}$$

Cell：

$$\delta_c^t = b_t^t g'(a_c^t) \delta_s^t \tag{2-70}$$

忘记门：

$$\delta_{\phi}^{t} = f'(a_{\phi}^{t})\sum_{c=1}^{C} s_{c}^{t-1}\delta_{s}^{t} \tag{2-71}$$

输入门：

$$\delta_{\iota}^{t} = f'(a_{\iota}^{t})\sum_{c=1}^{C} g(a_{c}^{t})\delta_{s}^{t} \tag{2-72}$$

2.2.4 序列到序列方法

序列到序列（Sequence to Sequence）方法，简称 Seq2Seq 方法。Seq2Seq 模型就是一个翻译模型，把一个序列翻译成另一个序列。基本思想是两个 RNN（LSTM 可视作添加了门结构的更复杂的 RNN），一个作为编码器（encoder），一个作为解码器（decoder），称之为 RNN Encoder-Decoder。我们是基于 LSTM 进行序列到序列的预测，因此我们的两个 RNN 都是 LSTM。作为编码器的 LSTM，目的是把一个输入序列压缩表示成一个固定长度的向量。作为解码器的 LSTM，根据编码器的向量，生成一个 token 序列，这个 token 序列就是生成的要翻译的另一个输出序列。我们采用极大似然估计优化损失函数，使输入序列通过编码器编码再经过解码器得到输出序列的概率最大。输入序列和输出序列的长度可以不同。这种方法的优势在于，以往的 DNN 要求输入和输出序列都已经处理成一个固定长度的向量，但是实际中很多问题都是用一些位置长度的序列。而基于 RNN 的 Seq2Seq 方法里编码器将输入序列自动转化成固定长度向量[67]。

根据 Sutskever 等人的论文介绍的序列到序列方法，用到的两个 LSTM 都是 4 层，而且深层的 LSTM 表现明显比浅层好很多[67]。但是为了方便表示，下面把图 2-7 当成是 1 层来分析。

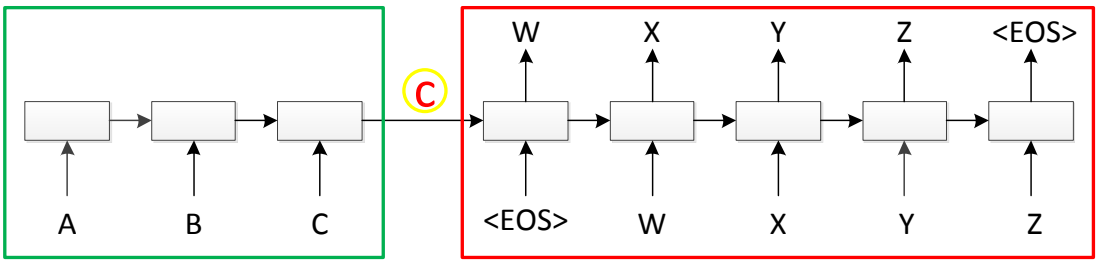


图 2-7 Seq2Seq 方法模型示意图

左边绿色框部分是 LSTM1（encoder），右边红色框部分是 LSTM2（decoder）。模型中，如果输入序列是“ABC”，输出的翻译结果是“WXYZ”。这里是假设输入输出向量的每个元素都是 1 维元素。可以看出，输入和输出的向量长度是不同的，模型会不断根据前面的状态和结果来预测下一个输出的符号。当模型预测到要输出一个<EOS>符号时，表示句子结束（end-of-sentence），翻译过程才会停止

输出。模型展开图如图 2-8。$x_i \in R^{n\times 1}(i=1,2,...,T)$ 表示输入向量，如果每个 x_i 是个 10 维的向量，那么 $n=10$。$y_i \in R^{n'\times 1}(i=1,2,...,T')$ 表示输出向量，x 和 y 的维度可以不同。$h_i, h_j \in R^{k\times 1}(i=1,2,...,T; j=1,2,...,T')$ 图中每个圆圈表示的是隐藏层的状态，每个箭头表示一个函数操作。图 2-8 是按时间步展开模型进行讨论，绿色的圆圈表示的是编码器 encoder 同一层按时间步排列的函数操作，红色的圆圈表示的是解码器 decoder 同一层按时间步排列的函数操作，从左到右表示该层状态随输入的变化。

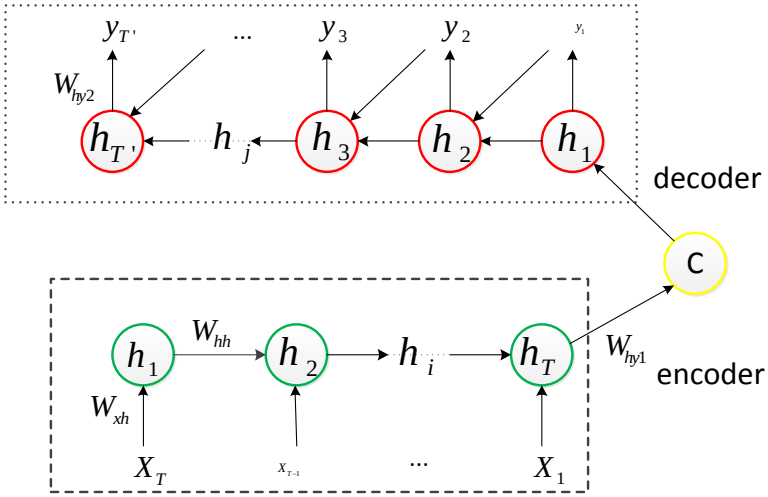


图 2-8 Seq2Seq 模型细节展开图

Seq2Seq 最初是应用于机器翻译和对话机器人。不过由于端到端的模型，减少了很多人工处理和规则制定的步骤。如果把序列中每个时次的值设置为多变量组成的向量，序列到序列的预测同时考虑了向量多个维度的信息，要考虑多个维度多时次的相似度，即可在时间序列预测中考虑隐藏的多变量的相互关系。

2.2.5 优化方法

传统的批量梯度下降算法（batch gradient descent）中，如等式（2-73），每一步迭代都要使用训练集的所有数据，使损失函数达到最小值。但是由于每一步都要用数据集的所有数据，因此运行速度会比较慢。但是优点是学习速率一经设定可以不必改变。

$$\begin{aligned} \hat{g} &= \frac{1}{n}\nabla_\theta \sum_i L(f(x_i,\theta), y_i) \\ \theta &= \theta - \varepsilon \hat{g} \end{aligned} \tag{2-73}$$

下面介绍几种加快梯度下降、改变学习速率的方法。

1. 随机梯度下降（SGD）

随机梯度下降方法在每次迭代的过程中，随机抽取一批固定大小 m 的样本以及相关输出，以此为根据来更新参数。这种方法的优点是对于很大的数据集，训练的收敛速度较快。缺点是由于迭代是依据随机抽取的样本子集，因此下降的梯度和全局梯度存在误差，梯度含的噪声比较大，如果学习速率不变的情况下，可能会导致模型无法收敛。因此训练过程中需要不断减小学习速率。

$$\begin{aligned}\hat{g} &= \frac{1}{m}\nabla_\theta \sum_i L(f(x_i,\theta),y_i)\\ \theta &= \theta - \varepsilon\hat{g}\end{aligned} \tag{2-74}$$

2. 动量法

动量法可以一定程度上解决迭代梯度噪声大的问题，最适用于梯度连续且变化率不大但含有很大噪声的情况。动量法里引入了一个表示动量的新变量 υ，是前面的梯度的累加项。每次迭代动量值都要以一定的衰减程度 α 加上新的梯度，直至动量项趋于稳定。这种方法的优点是抑制震荡，梯度方向变化小的情况下也可以加速学习。

$$\begin{aligned}\hat{g} &= \frac{1}{m}\nabla_\theta \sum_i L(f(x_i,\theta),y_i)\\ \upsilon &= \alpha\upsilon - \varepsilon\hat{g}\\ \theta &= \theta + \upsilon\end{aligned} \tag{2-75}$$

3. 自适应梯度下降方法（AdaGrad）

自适应梯度下降方法最大的特点是可以自动修正学习速率。我们需要设置一个初始全局学习速率 ε 和一个中间变量 r，这个中间变量 r 内容是累计的梯度模之和，每次迭代的学习速率都是全局学习速率与梯度模之和的一个比值（δ 是一个很小的值，防止分母为 0），目的是控制学习速率合理范围内自动更新。如果某次的梯度过小，实际学习速率衰减速度就会变慢；如果某次的梯度过大，实际学习速率衰减速度就会变快。这种方法的缺点是需要设置一个全局学习速率，而且如果网络的深度过深，优化效果会不好。

$$\begin{aligned}\hat{g} &= \frac{1}{m}\nabla_\theta \sum_i L(f(x_i,\theta),y_i)\\ r &= r + \hat{g}\odot\hat{g}\\ \Delta\theta &= -\frac{\varepsilon}{\delta+\sqrt{r}}\odot\hat{g}\\ \theta &= \theta + \Delta\theta\end{aligned} \tag{2-76}$$

4. RMSProp

RMSProp 是 Geoff Hinton 在 Coursera 课程的第 6 章[68]提出的未发表的一种自适应学习速率法。RMSProp 引入了一个衰减系数 ρ，每次迭代都会使梯度累积

量 r 衰减一定的比例 ρ，方法与动量法类似，都是为了控制学习速率。这种方法适用于处理非平稳变化的序列，在 RNN 中效果很好。这种方法的缺点是需要设置一个全局学习速率和一个梯度累积量衰减系数。

$$\begin{aligned}
&\hat{g}=\frac{1}{m}\nabla_{\theta}\sum_{i}L(f(x_i,\theta),y_i)\\
&r=\rho r+(1-\rho)\hat{g}\odot\hat{g}\\
&\Delta\theta=-\frac{\varepsilon}{\delta+\sqrt{r}}\odot\hat{g}\\
&\theta=\theta+\Delta\theta
\end{aligned}\tag{2-77}$$

5. 自适应动量估计法（Adam）

这种方法就是带有动量项的 RMSprop。这种方法比较复杂，需要引入一阶动量 s 和二阶动量 r，一阶动量衰减系数 ρ_1，二阶动量衰减系数 ρ_2。目的是平稳地调节学习速率。这种方法的优点是学习速率范围稳定，使得梯度变化平稳。

$$\begin{aligned}
&\hat{g}=\frac{1}{m}\nabla_{\theta}\sum_{i}L(f(x_i,\theta),y_i)\\
&s=\rho_1 s+(1-\rho_2)g\\
&r=\rho_2 r+(1-\rho_2)g\odot g\\
&\hat{s}=\frac{s}{1-\rho_1}\\
&\hat{r}=\frac{r}{1-\rho_2}\\
&\Delta\theta=-\varepsilon\frac{\hat{s}}{\delta+\sqrt{\hat{r}}}\\
&\theta=\theta+\Delta\theta
\end{aligned}\tag{2-78}$$

总之，我们可以网络特点选取合适的优化方法，加快数据收敛速度。

2.3 本章小结

本章首先从权值空间论和函数空间论两方面介绍了高斯过程回归的定义，之后介绍了核函数的类型和超参数的计算方法，为第三章核函数的设计和超参数求解打下基础。随后，本章又介绍了基于长短期记忆网络（LSTM）的时间序列预测方法，从人工神经网络、循环神经网络到长短期记忆网络渐进推导了前向和后向传播的原理，最后介绍了序列到序列（Seq2Seq）方法和几种损失函数优化方法，为第四章风速时间序列预测方法提供理论基础。

第三章 基于多尺度核函数的高斯过程回归空间风速插值算法

大气再分析数据在应用过程中需要根据需求将低分辨率数据插值成高分辨率数据。传统方法多采用线性插值和指数插值。在二维空间序列计算中，通常也采用单一尺度高斯过程回归插值。高斯过程回归方法不仅考虑了临近点的相关性，对于某些核函数，插值过程中也可考虑到较远区域点的相关性。但是上述方法都建立在被插值变量组分单一的假设上，这些方法都忽略了风速数据多尺度非平稳的特点。影响风速的因素众多，作用机理十分复杂，风速信号本身就表现出很强的多尺度特性。从信号分解的角度来看，不同作用源产生的信号波长和振幅都不同，最终的风速序列可看作是多个信号耦合在一起的结果，因此风速序列表现出了较强的非平稳性。近些年来，针对风速的此特点，广泛采用的方法是利用小波分析或者经验模态分解将信号分解为不同波长的子序列，再对每个子序列进行回归建模，最终将所有子序列的插值结果直接求和或者加权求和。本文改进高斯过程回归的核函数也是此思路，假设对风速空间序列进行分解，那么每个序列高斯过程回归插值都有对应的核函数，根据核函数的定义，核函数之和也是核函数，因此总的核函数就是多个尺度核函数之和。

3.1 物理学上的尺度约束

小于月时间尺度的大气变化是由几种物理现象产生的。行星尺度罗斯贝波有几天的时间，更长的波长可能会产生准静止波模式。罗斯贝波的斜压不稳定性会产生对流天气常见的锋面、气旋、反气旋模式。大气潮汐，主要由对流层的水蒸气和平流层的臭氧被太阳能加热而产生，其波长是行星尺度波长，且主要是一天一次和一天两次。大气潮汐随海拔高度增加而扩大，尤其在 90km 以上高空。地表增温产生对流循环，对流循环会产生雷暴。不稳定性或者其他物理过程会产生有序的雷暴漩涡线。大气重力波可能由地形流作用产生，或者由雷暴、热带风暴或其他扰动引发。虽然像潮汐一样，重力波随高度增加而扩大，但是由于在自然中它们都是无规律的，重力波不能准确模拟。大气湍流是一个小尺度过程，由地表增温、地形造成的影响，或者由重力波、潮汐、罗斯贝波相关的急流切变产生的不稳定过程都可以引发大气湍流[4]。

在实际的气象模拟中，可能会采用多尺度扰动的模型模拟。美国 NASA 的全球参考大气模型系统 2010 版（GRAM2010）[69]在模拟高空大气变量时，就采

用了多尺度扰动模型来模拟各种大气过程所引起的温度、气压、密度、露点温度等的变化。较小尺度的参数用来描述小尺度过程，如扰动、中尺度暴风雨和重力波，而大尺度的参数用来描述大尺度过程，如罗斯贝波、潮汐、气旋和反气旋。从光谱积分尺度上说，这两种尺度参数中的任何一种都能描述一段横跨有效波数范围光谱的特征。假设这两种尺度参数只依赖于海拔高度和纬度。同时，还要考虑一些微小扰动，这些小尺度参数值的随机变化也会影响变量值的最后结果。比如间歇效应，也就是扰动出现在小块区域或整层的趋势。

我们假设风速向量$\boldsymbol{w}=\{u,\ v\}$，风速大小是$w=\|\boldsymbol{w}\|$，风速方向是α，单位是弧度，其中风向量的两个分量$u=-w\sin\alpha$，表示的是纬向风，即沿着纬度方向的风；$v=-w\cos\alpha$，表示的是经向风，即沿着经线方向的风。实际风速分析中，经常把风速分为纬向风u和经向风v两个变量。这两个变量不仅与气压p、密度ρ、温度T有关，也与风速大小w和方向α有关。由此，本章的目标就是完成对风速两个分量的定量分析。

3.2 核函数对插值的影响

高斯过程回归建模的关键就在于核函数的选择。核函数的选择大多基于经验方法，根据数据特点，选择合适的局部核函数或全局核函数。本节会结合实际数据，讨论不同单一核函数的结构特征，讨论不同的超参数对插值结果的影响，为下一节核函数的选择提供理论基础。

3.2.1 核函数的结构特点

1. 单一核函数

核函数由于结构的不同，对数据结构关系的描述也不同。我们先分析几个常用单一核函数的结构特点和适用情况，以便能更好地选择核函数。图 3-1 是几种常用单一核函数的图像，图中表示不同结构的$k(x,0)$，横坐标是距离大小，表示$r(r=|x_i-x_j|)$与k的关系。，也就是k对r范围的数据还保持着怎样的关注度。从可以看出：

对于平方指数协方差核函数，两个点越靠近，协方差越大，当两个点越来越远，协方差函数的描述能力逐渐变弱，一定长度之后相关性变为 0。平方指数协方差函数是一种描述局部特征的函数，如果数据变化不是特别剧烈，会取得不错的效果。

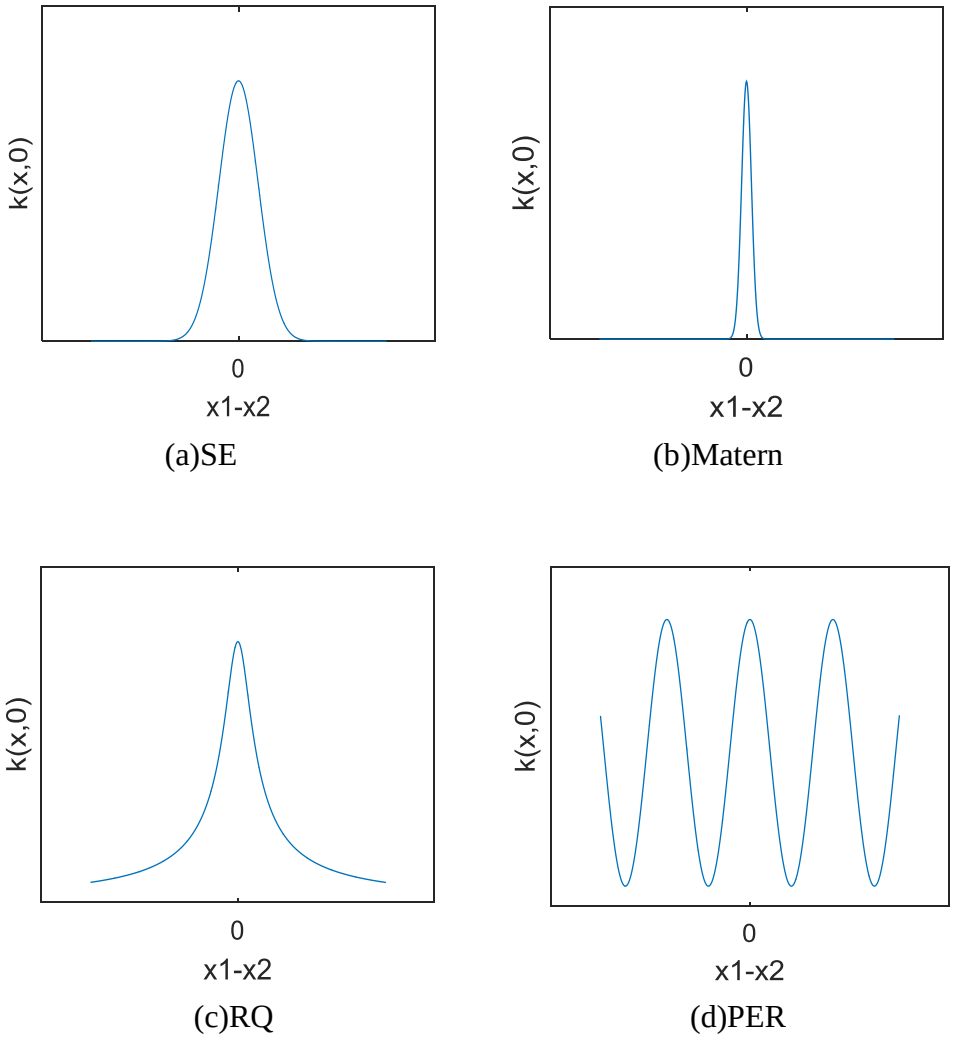


图 3-1 几个常用单个协方差函数结构

Matern 协方差核函数，也是一种描述局部特征的核函数。不过，它的形状比平方指数协方差核函数和有理二次协方差核函数更尖锐，描述局部特征的能力更强，但是 Matern 的局部区域长度比平方指数协方差的局域区域长度小很多，大范围相关性很弱。

有理二次协方差核函数可以看做是多个不同特征尺度、量级的平方指数协方差函数之和。虽然随着 r 减小，核函数的描述能力也会逐渐下降，但是 k 的值变化为 0 的过程非常缓慢，局部特征的区域长度也比平方指数协方差函数大很多。有理二次协方差函数可以描述局部特征，在两点相距不是特别远的情况下也可以描述部分全局特征。

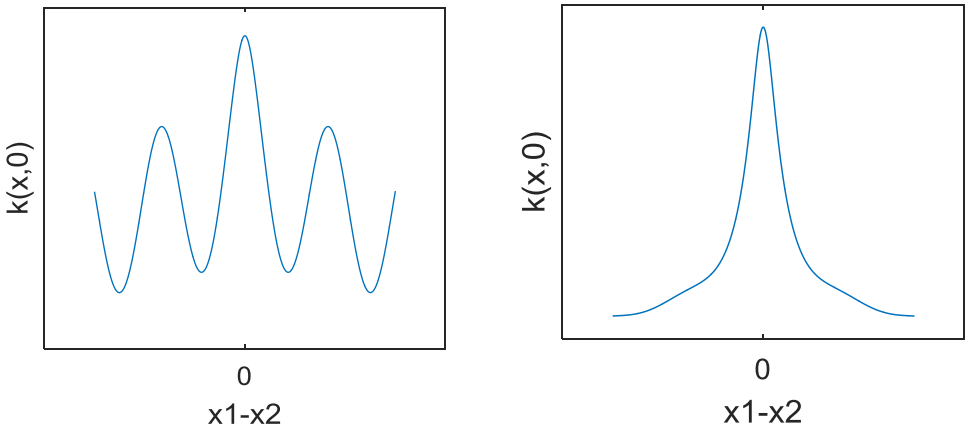
周期协方差核函数适用于描述周期性序列的核函数，核函数 k 不会随着 r 增大而变小，而是呈现周期性的相关关系，适用于描述周期现象明显的数据，但是对距离很近的数据相关性描述得并不好。周期协方差核函数是描述全局相关关系的核函数，不适合描述局部关系。

总之，周期性协方差核函数适用于描述全局特征，有比较好的泛化能力；平

方指数协方差核函数、Matern 协方差核函数、有理二次协方差核函数适用于描述局部特征，其中有理二次协方差核函数也可描述某些距离不是太远的全局特征。

2. 组合核函数

为了获得更好的描述能力，根据组合核函数的定义，得到了一系列组合核函数。以有理二次协方差核函数和周期核函数为例如图 3-2。横坐标表示距离，纵坐标是核函数 $k(x,0)$ 的值。图（a）中，相加组成的核函数同时具有两个核函数的特征，既有周期性特征，又在 $r=0$ 处显示出了比较强的局部相关性，随着 r 的增大，这种局部特征逐渐消失。（b）中，由于局部特征明显、全局特征较弱的平方指数协方差核函数值在因变量远离 0 处逐渐趋于 0，因此相乘组成的核函数在远离 0 处值也为 0， 只在 $r=0$ 的周围表现出了很弱的周期性。



(a)相加　　(b)相乘

图 3-2　RQ、PER 核函数组成的组合核函数结构图

实际建模中，除了如上两个不同核函数简单相加和简单相乘这种方式，还有如表 3-1 的有数学意义的多个核函数组合[58]。

表 3-1 多个重复结构组合核函数的意义

数学意义	组合方式
贝叶斯多项式回归	$\prod LIN$
广义傅里叶变换	$\sum PER$
广义相加模型	$\sum SE$
自相关计算	$\prod SE$

根据以上分析，核函数的选择具有非常大的灵活性。根据变量特点将其扩展，选择更适应历史数据特征的核函数来提高插值准确性。

3.2.2 超参数的影响

这一节讨论核函数中不同的超参数选择对于插值结果的影响。以有理二次核函数和周期核函数为例，在其他参数采用最优值的基础上，分析特征长度尺度l、扩散形状参数α、变化频率f取值的影响。数据选取 1948 年某地月平均的温度序列，如图 3-3。

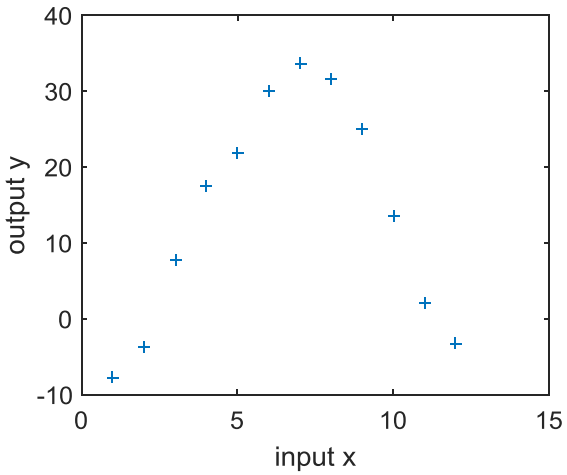


图 3-3 示例序列真值示意图

1. 特征长度尺度l取值的影响。

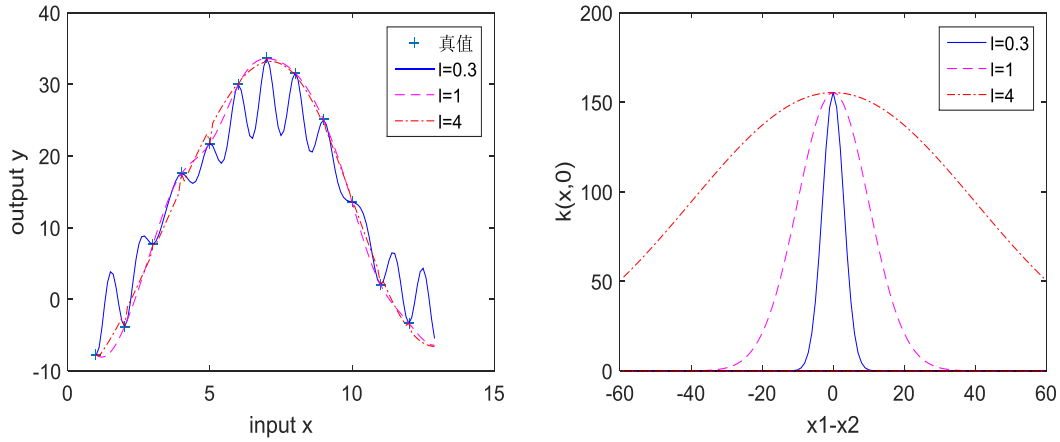
分别对l取不同的值，l=0.3,1,4。当l取值非常小时，预测的函数趋势非常震荡，而l=1,4 时，预测的函数趋势比较平缓。当l取值非常小时，核函数的影响范围会变得很小，局部性非常强，稍微远一点位置的点就不会被考虑。

2. 扩散形状参数α的影响。

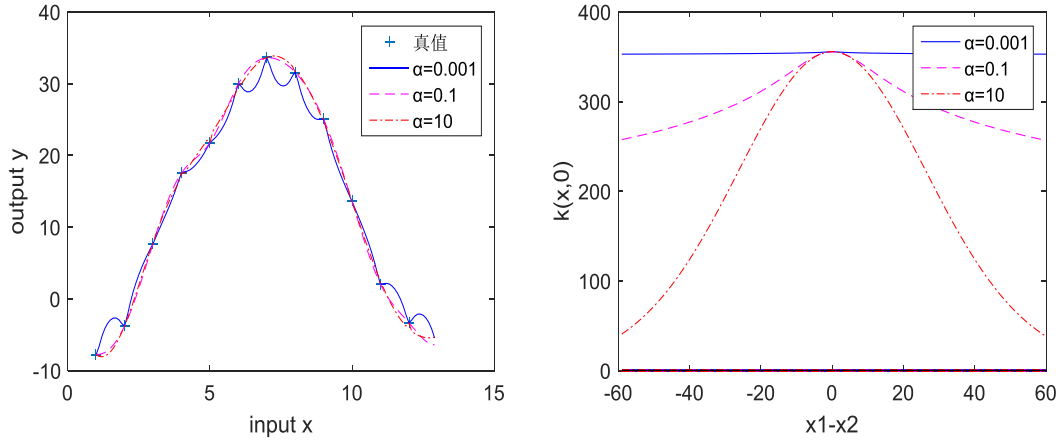
分别对α取值 0.001，0.1，10。当α取值非常小时，数据的预测也出现了震荡，但是相同量级α变化对结果的影响比l要小。α取值越大，核函数的局部特征越明显，越像平方指数核函数（SE）；α取值越小，全局的特征越明显，泛化能力越强。

3. 变化频率f的影响。

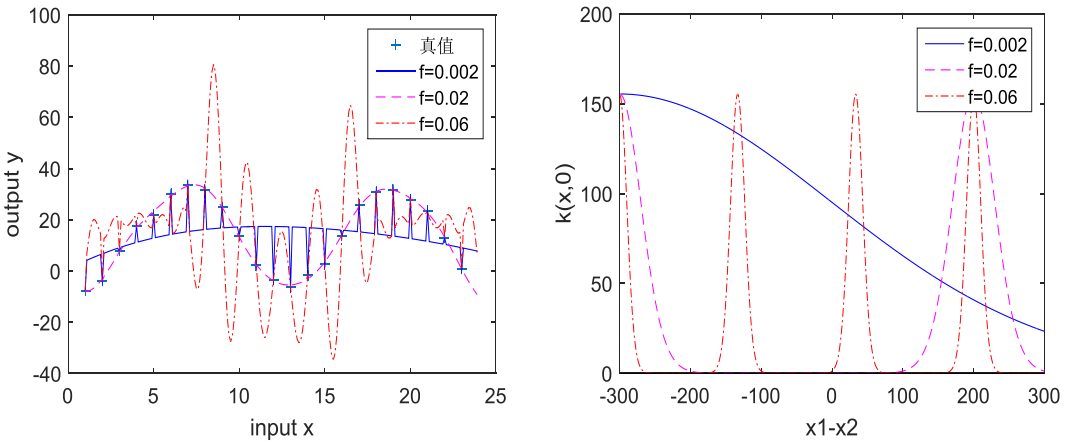
分别对f取值 0.02,0.2,0.6。f越大，预测的函数趋势波动性越大；f越小，预测函数趋势变化越平缓。右图中，无论r取何值，周期核函数始终保持着一定周期的波动性。



（a）l



(b) α



(c) f

图 3-4 l、α、f 对插值结果和核函数结构的影响

不同种类的核函数描述能力不同，超参数取值不同也会使核函数描述能力产

生差异。我们需要选取最优的超参数来拟合数据，同时我们也可应用组合核函数来增加核函数的描述内容。

3.3 基于风速多尺度核函数高斯过程回归的风速插值

本章选取国防科技大学的第二代军用数值天气预报系统中的全球中期数值天气预报模式（YHGSM）的再分析数据作为源数据，选取格林威治时间（UTC）2001 年 1 月 1 日 0 时第 50 个模式层（大约 0.95km）、东经 16°南纬 90°~东经 16°北纬 90°范围的风速数据作为实验数据集，主要用到的数据变量有：

表 3-2 风速插值选取的气象变量

编号	气象变量	表示符号	单位
1	风速	w	m/s
2	角度	α	rad
3	经向风分量	u	m/s
4	纬向风分量	v	m/s

许多场合需要用到经向风分量u、纬向风分量v作为设计依据，有的场合这两个分量也可以通过风速w和角度α通过等式（3-1）求得。但许多情况气象数据的精度达不到实际应用的精度要求，比如气象数据集是 2.5°×2.5°的网格数据，而实际应用需要 1°×1°网格的风速数据；或者气象数据集提供的是不均匀的模式层数据而实际应用需要均匀的风速数据，这就需要进行插值计算。通常情况下，假设大气在水平或者垂直方向是均质大气，利用均匀插值、线性插值、指数插值、三次样条插值、克里金等其他插值方法计算内插点的数值，而这些方法往往要求数据平缓变化，对于有起伏变化的数据插值效果并不好。尤其是大气变量，由于在地球表面 1°的距离实际上是非常远的（经线上，纬度每差 1 度，实地距离大约为 111 千米；在纬线上，经度每差 1 度,实际距离为$111\times\cos\theta$千米，其中θ表示该纬线的纬度），因此可能风速的变化并不平缓。风速的变化是由附近位置的气压差导致的，气压的变化又与温度、大气密度的变化息息相关，因此风速的变化在空间上也不是孤立的，与不同尺度大气运动的临近位置风速也是有关系的。

风速w（m/s）、角度α（rad）与经向风分量u（m/s）、纬向风分量v（m/s）之间的转换关系：

$$u=-w\sin\alpha, v=-w\cos\alpha \tag{3-1}$$

其中，u、v分别是水平风的两个分量，$w=\sqrt{u^2+v^2}$，$\alpha=arctan\frac{u}{v}$。通常，只要提供了w和α，就可以计算得到u'和v'。下面我们探究一下对w和α插

值间接计算u'、v'和对u和v插值直接计算u、v的准确性。

通常使用平方指数协方差函数做为高斯过程回归的核函数进行拟合。即一般核函数的形式为

$$k\left(x,x'\right)=\sigma_f^2\exp\left(-\frac{\left(x-x'\right)^2}{2l^2}\right)+\sigma_n^2 \tag{3-2}$$

其中，σ_f 表示数量级大小，l表示尺度长度，σ_n表示模型误差。这是最简单、最普遍的基于距离的核函数。在上一节知，平方指数协方差函数对拟合局部特征有优势，不同大小的σ_f和l对拟合结果有一定的影响。这个核函数只是表示一个单一尺度过程的变量之间相关关系。

3.3.1 基于高斯过程回归风速插值算法的建模方法

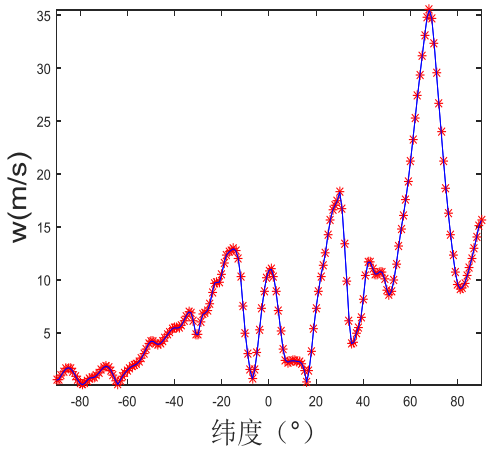
GPR 算法的建模原理如图 3-6。首先根据训练样本选取合适的核函数，设置合适的超参数初始值。核函数需要考虑到多个尺度的风速变化，应该是多种简单核函数组合的结果。超参数初始值的设置需要根据样本数据特点选取，本文采用的是经验方法，即设置一个较大范围，得到若干组该范围的随机初始值组合，通过后续的模型不断训练修改超参数。此时，确定了初始先验概率模型的形式。其次，利用最大似然估计从训练样本学习得到超参数的最优值，得到后验概率模型，本质上这就是利用梯度下降法寻找使得损失函数最小的超参数θ。最后，将测试集放入 GPR 后验模型中，得到测试集的插值结果估计。

图 3-5 高斯过程回归风速插值方法原理图

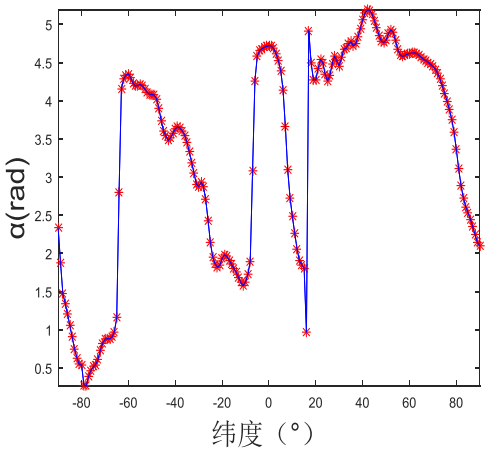
本文的主要工作主要在“选择多尺度风速插值核函数”之后的过程，包括确定多尺度核函数、确定 GPR 先验模型、最大似然估计确定超参数最优解、使用 GPR 后验模型进行插值验证。

3.3.2 风速多尺度插值核函数的构建

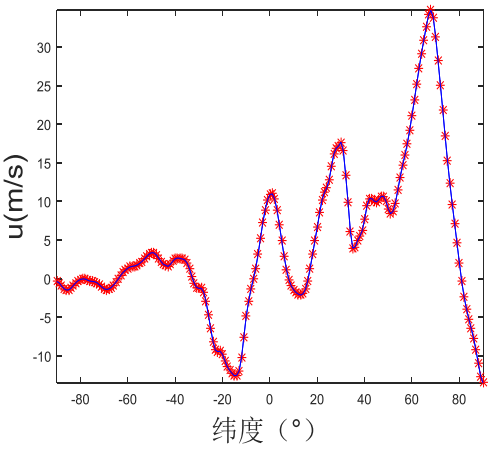
为了进一步提高数据拟合的精确度，更好地描述数据特征，本节对风速多个尺度的核函数进行组合。在单一核函数的基础上，根据数据特点和物理规律，本文构建了包含大中小尺度和空间重复性变化的风速多尺度插值核函数。虽然图 3-5 数据是一维的，但考虑到数据的大尺度、小尺度、空间重复性变化过程和随机扰动的复杂情况，本节采用了 12 个超参数来拟合这个过程，分别用 $\boldsymbol{\theta}=\{\theta_1,\ldots,\theta_{12}\}$ 表示。具体含义会在下面进行解释。



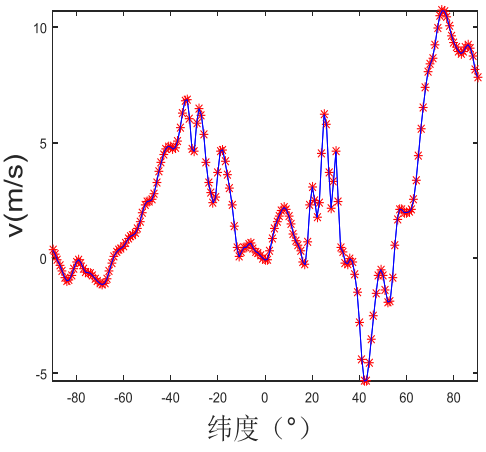
(a)w



(b) α



(c)u



(d)v

图 3-6 2001.1.1 东经 16° 南纬 90° ~北纬 90° 20km 高空风速数据

数据是东经 16°南纬 90°~北纬 90°、高度 20km 的风速数据，单位是 m/s，风速随纬度的分布如图 3-5，其中北纬为正，南纬为负。目标是建立风速与纬度x的一个函数。根据相关气象知识，从图中我们也可以看出风速的几个特点：一个大尺度的趋势、一个中尺度的空间重复性过程变化趋势和一些微小的扰动。风速数据可以看作一个信号，信号一般通过小波分解或者 EMD 等其他分解方式，都可分解为若干个振幅与波长不同的信号叠加的方式，其中有的波长很长，可以看作大尺度过程，有的波长很短，可以看作中小尺度过程，更小者可以看作随机扰动或者噪声[70]。由此，核函数需要考虑到至少这三方面的性质，构建复杂度与之匹配的高斯过程回归核函数。

Carl Edward Rasmussen 在研究 20 年夏威夷莫纳罗亚天文台采集的上空 CO_2 浓度的案例中，采用了多尺度过程的组合核函数[58]，拟合数据是随时间变化的浓度信息。在这个案例中，组合核函数比较准确地模拟了各种尺度的过程，并且以低方差预测了连续 60 个月的 CO_2 浓度。Carl Edward Rasmussen 的多尺度组合核函数是考虑时间过程的预测核函数，考虑了大中小尺度、季节性变化过程和非独立高斯噪声。受到这种多尺度模型构建方法的启发，本文也构建包含大尺度项、纬度相关空间重复波长项、中小尺度项、多依赖噪声项的空间插值组合核函数。下面分别对这几个尺度内容进行介绍。

为了描述一个大尺度平滑变化的趋势，我们用平方指数（SE）协方差项，其中有两个参数：振幅θ_1和特征长度θ_2。该项形式为

$$k_1\left(x,x'\right)=\theta_1^2\exp\left(-\frac{\left(x-x'\right)^2}{2\theta_2^2}\right) \tag{3-3}$$

这只是一个平滑的大尺度的过程，不论θ_1和θ_2取值大小，这个模型过于简单，效果可能并不理想。对于纬度相关的空间重复性过程，例如空间中由风带、气压带引发的风速变化，可以用空间重复波长项表示，即周期协方差核函数，如 2.1.3 的等式（2-30）。但是我们并不清楚空间变化的具体波长，可以修改（2-30），加入一个平方指数项，来模拟一个没有确切波长的空间重复性过程。该项形式为

$$k_2\left(x,x'\right)=\theta_3^2\exp\left(-\frac{\left(x-x'\right)^2}{2\theta_4^2}-\frac{2\sin^2(\pi\theta_{12}(x-x'))}{\theta_5^2}\right) \tag{3-4}$$

其中，θ_3表示数量级，θ_4表示空间重复波长的影响范围，θ_5表示平滑程度，θ_{12}是与波长相关的项。数据中的空间重复波长项主要是由赤道到两极低压高压交

替出现的不同风带气压带变化产生的，因此$k_2(x,x')$这一项的存在是有意义的。如果训练得到的θ_4很小，那么这一项的影响就会比较大，如果训练得到的θ_4很大，那么空间重复性过程项的影响就会被大大削弱，由此这一项的设置更具有泛化意义。 中小尺度的扰动可以用一个如等式（2-29）的有理二次项来模拟

$$k_3\left(x,x'\right)=\theta_6^2\left(1+\frac{\left(x-x'\right)^2}{2\theta_8\theta_7^2}\right)^{-\theta_8} \tag{3-5}$$

其中θ_6表示数量级，θ_7是一个标准长度尺度，θ_8是确定长度尺度扩散形状的参数。这个过程我们也可以用平方指数协方差项（SE）的形式表示，如k_1的形式，采用与之不同大小的参数。但是通常情况下理性二次项（RQ）形式的效果更好（给出了更高的边缘似然），因为有理二次项形式可以包含几种不同的长度尺度。等式$k_3\left(x,x'\right)$并不是一个过程的表示，而是可看作多个平方指数协方差核函数项按不同量级、不同特征长度的比例组合。当时，$\theta_8\to\infty$时，其结构与相同参数的平方指数协方差项相同。

最后，我们需要确定一个噪声模型。我们用平方指数分布再加一个独立项的形式

$$k_4\left(x_p,x_q\right)=\theta_9^2\exp\left(-\frac{\left(x_p-x_q\right)^2}{2\theta_{10}^2}\right)+\theta_{11}^2\delta_{pq} \tag{3-6}$$

其中θ_9表示相关噪声项的量级，θ_{10}是相关噪声的长度尺度，θ_{11}是独立噪声项的量级。序列中的噪声可能是受测量误差或者本地短期天气现象的影响，所以我们可以假设噪声与空间也有某种适度的相关性。等式（3-6）中的相关噪声项与等式（3-3）的大尺度项的表达方式是一样的。在优化超参数时，这些项中的某些尺度项相关的一些项会比较大，但是同时其他尺度相关的一些项会很小。这里的“噪声”实际上并不一定是真正意义上的噪声。本文根据大、中尺度、空间重复性过程这种方法来划分风速核函数的尺度，但是可能存在某些没有关注到的其他尺度过程的项，而实际上它们并不是噪声，只是本文不关心这些过程。如果在其他的问题中，需要更关注另外一些尺度的问题，那么这些尺度的问题就依然是信号，而非“噪声”。。

最终的风速多尺度协方差核函数表示为

$$
\begin{aligned}
k(x,x') &= k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x') \\
&= \theta_1^2 \exp\left(-\frac{(x-x')^2}{2\theta_2^2}\right) \\
&+ \theta_3^2 \exp\left(-\frac{(x-x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi\theta_{12}(x-x'))}{\theta_5^2}\right) \\
&+ \theta_6^2\left(1+\frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8} + \theta_9^2 \exp\left(-\frac{(x-x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2\delta
\end{aligned}
\tag{3-7}
$$

其中，用到的超参数 $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_{12})^T$。数据再训练的过程中，先要减去经验平均值，然后通过共轭梯度法优化边缘似然函数来找到合适的超参数。为了避免陷入不好的局部极小值，需要设定若干个随机起点重复实验，挑选出最佳的边缘似然值。

3.3.3 实验结果分析

根据以上理论分析，用上述风速多尺度协方差核函数对东经 16°南纬 90°~北纬 90°、高度 20km 的风速数据，包括风速 w（m/s）、角度 α（rad）、经向风分量 u（m/s）、纬向风分量 v（m/s）。风速随时间的分布如图 3-5。从图中可以看出，w、u 数据都比较圆滑，α、v 数据只在局部有较明显的跳跃。用平方指数协方差函数进行拟合，检验并解释优化得到的超参数结果。这里引入均方根误差（RMSE）来定量地分析模型插值结果的准确性。RMSE 是衡量平均误差的常用方法，RMSE 越小，说明模型对已有数据拟合得越好，精确度越高。它的等式：

$$
\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^t - y_i^e\right)^2}
$$

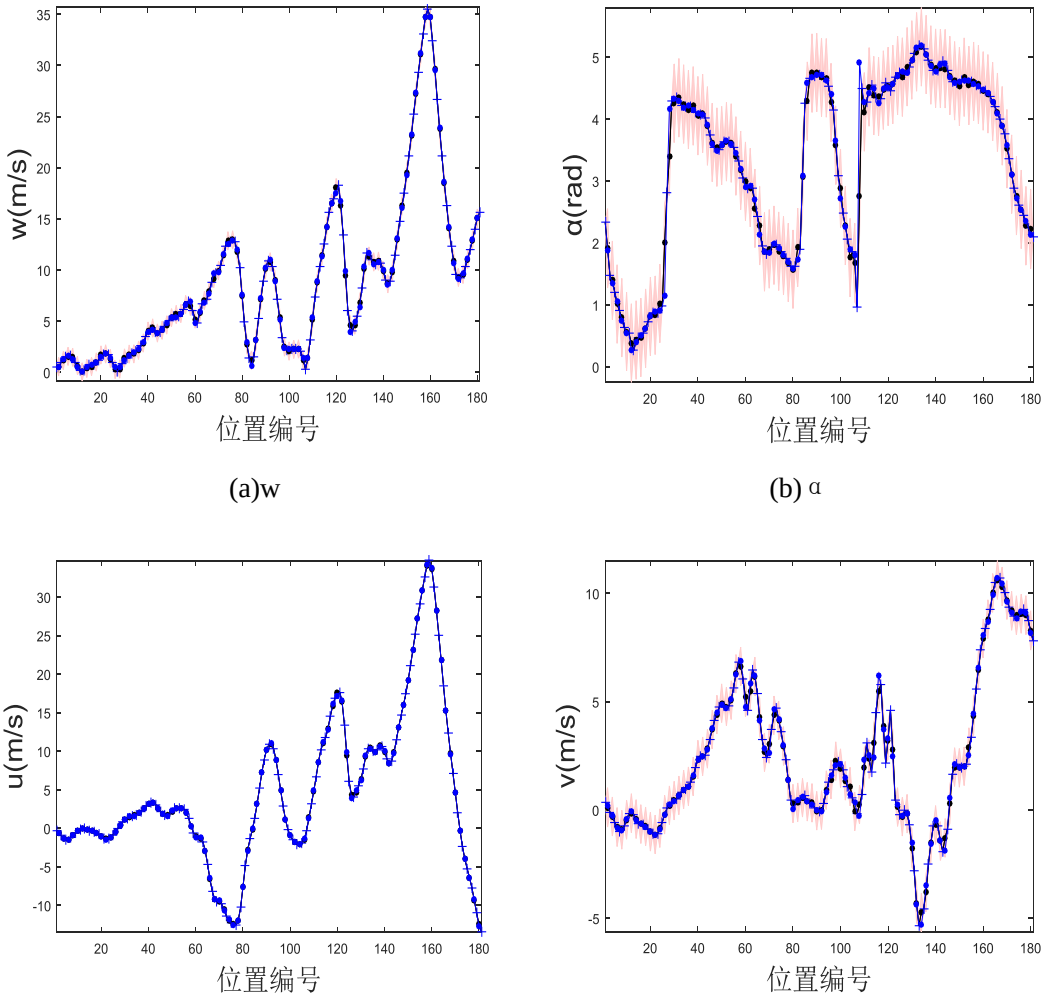
其中，y^e 为插值结果，y^t 为真实值，n 为待插值点个数。

以经向风 v 为例：大尺度过程的振幅量级 $\theta_1 = 0.7345$，长度尺度 $\theta_2 = 1.7035$（单位 m/s）。中等尺度过程的振幅量级 $\theta_3 = 7.7877$，标准尺度长度 $\theta_4 = 6.7599$，形状参数 $\theta_5 = 1.3532$。这说明中等尺度过程项的影响较大尺度过程的影响更大，拟合模型的主要影响因素还是中等尺度过程的核函数。

空间重复性过程的超参数：尺度大小 θ_3=0.067，空间影响范围 θ_4=2.2205，平滑参数 θ_5=0.4785，空间重复波长参数 θ_{12}=0.03236。相对较小的尺度大小说明

数据空间重复性过程的影响相比大尺度和小尺度过程更小。

对于噪声项，我们得到了相关噪声项 $\theta_9 = 0.0055$，长度尺度项 $\theta_{10} = 0.0969$，独立噪声项 $\theta_{11} = 8.9965 \times 10^{-7}$。因此，噪声项的相关长度很短，总的噪声量级仅仅 $\sqrt{\theta_9^2 + \theta_{11}^2} = 0.0055$，说明数据可以用这个模型模拟很好地解释。除了外插值区间，内插值区间插值结果的 95%的置信区间的宽度都小于 1.0389m/s。



(a)w　　(b) α

(c)u　　(d)v

图 3-7 20km 高度风速多尺度核函数高斯过程回归插值结果图

图 3-7 中，蓝色“+”表示已知点的数据，黑色“ •”表示中间点利用插值算法计算的数据，蓝色“ •”表示中间点位置真实值的数据，粉色区域是插值结果的 95%置信区间的值，即 $\varepsilon \pm 1.96\sigma$ 范围。w 和 u 的 95%的置信区间非常窄，几乎看不出较大的粉色区域，黑色和蓝色“ •”不重合的区域也比较少。α 的数据不平稳区域置信区域比较宽，在数据跳跃非常强烈的区间数据模型插值和真实值不重合的比较明显，数据误差比较大的区间也出现在这里。

从图3-7中可以看出，风速多尺度插值核函数插值结果与真实数据非常吻合，可以很好地对数据进行回归拟合。基于单一核函数与多尺度核函数的插值结果在图像上相差不明显，因此直观上难以看出二者拟合的差别。为了验证这种模型的泛化效果，我们将 90°S~90°N 编号 1-181，数据以 1°为间隔，只取编号为单数的点进行建模，用双数点进行内插值，我们得到的 90 个内插值点的均方根误差如表 3-3，3-4,3-5，其中u'和v'是利用等式（3-1）求得的。

表 3-3 空间间隔为 1°1km 高度四种插值结果的 RMSE

	w	α	u’	v’	u	v
线性插值	0.9611	1.1432	1.2776	4.3901	0.9254	0.5391
指数插值	1.0453	1.2088	1.7994	3.9912	1.0442	0.6030
单一平方指数协方差核函数	0.7875	1.1934	1.0622	4.3256	0.4796	0.4444
风速多尺度协方差核函数	0.9332	1.1672	1.3318	4.5139	0.4675	0.4421

表 3-4 空间间隔为 1°20km 高度四种插值结果的 RMSE

	w	α	u’	v’	u	v
线性插值	0.2816	0.2577	0.3324	0.3348	0.2607	0.2241
指数插值	0.3858	0.2311	0.4224	0.3727	0.3920	0.2501
单一平方指数协方差核函数	0.1705	0.2663	0.2894	0.4509	0.1889	0.1474
风速多尺度协方差核函数	0.1396	0.2634	0.3238	0.5116	0.1023	0.1339

表 3-5 空间间隔为 1°50km 高度四种插值结果的 RMSE

	w	α	u’	v’	u	v
线性插值	0.2956	0.3298	0.3204	1.1545	0.3045	0.1234
指数插值	0.7791	0.3975	0.4542	1.3015	0.7902	0.1826
单一平方指数协方差核函数	0.1742	0.3366	0.2899	0.9534	0.1891	0.1129
风速多尺度协方差核函数	0.1478	0.3361	0.2556	0.7355	0.1154	0.0476

表 3-3 为不同模型对风速数据插值结果的定量比较。线性插值的形式为

$$\frac{y_x - y_a}{y_b - y_a} = \frac{h_x - h_a}{h_b - h_a}, h_a \le h_x \le h_b$$

指数插值的形式为

$$\frac{y_x - y_a}{y_b - y_a} = e^{-\frac{h_x - h_a}{h_b - h_a}}, h_a \le h_x \le h_b$$

其中，y_x是待求点的值。y_a、y_b分别是左临点和右临点的函数值，h_a、h_x、h_b分别是三个位置的因变量值。这两种插值方法在气象风速、密度、温度、气压插值中也经常应用。从表 3-3、3-4、3-5 中可看出，单一平方指数协方差核函数和风速多尺度协方差核函数的高斯过程回归插值方法在w、u、v三方面的插值都明显优于线性插值和指数插值，并且，风速多尺度协方差核函数插值比单一平方指数协方差核函数有明显的优化。这说明对于风速数据w、α、u、v来说，多尺度组合核函数的学习能力最强，其次是单一尺度高斯过程回归插值和线性插值，最后是指数插值。从四个值的真实值的图像中也可以看出，w、u、v虽然变化也比较大，但是边缘都比较平缓。而α的变化呈现区域性现象，即部分区域内变化平缓，而在区域边缘出现跳跃，进入下一个区域。对于这种跳跃剧烈的变化，单一平方指数协方差核函数核和风速多尺度协方差核函数对比线性插值和指数插值都没有表现出明显的优越性。

表中数据横向比较，u'和v'的 RMSE 明显大于直接计算u和v插值的 RMSE。纵向比较，单一平方指数协方差核函数和风速多尺度协方差核函数的高斯过程回归插值方法虽然对w的拟合要优于线性插值和指数插值，但是α的拟合要明显劣于后者，因此导致了u'和v'的误差较大，从而导致了较大的均方根误差。总之，在实际应用中，u和v直接插值比其他计算的精确度更高一些。

表中数据纵向比较，1km、20km、50km 不同高度中多尺度协方差核函数插值结果表现出来的优化效果是不同的。高度越高，该方法对比线性插值和单一尺度核函数插值优化效果越好。其中，50km 高度的u和v的均方根误差分别减小了 1/3 和 1/2。为了探究风速多尺度核函数对高空风速区域插值的效果，本文选取了东经 17°~东经 27°50km 高空风作为样本，分别采用四种方法实验，实验结果对比散点图如图 3-8 和图 3-9。从图 3-8 和 3-9 可以看出，多尺度核函数 GPR 方法的 RMSE 比线性插值和单一尺度 GPR 小非常多，优化效果非常明显。从图中，还可以注意到一个有趣的地方：单一尺度 GPR 和线性插值u和v的表现上非常不同，在 50km$u$插值中，单一尺度 GPR 始终明显优于线性插值结果，而在 50km$v$插值中，单一尺度 GPR 与线性插值结果相近，有时候劣于线性插值。这也许与样本点的选取有关，因为样本点是沿着同一经线选取的，所以多尺度核函数对经向风的空间隐藏关系挖掘得更多，而沿着同一经线的数据无法体现纬向风的水平变化，空间纬向风序列中隐藏的多尺度空间相关关系也少很多。不过，无论是哪种风分量，风速多尺度核函数的插值结果都远优于其余几种常见方法。

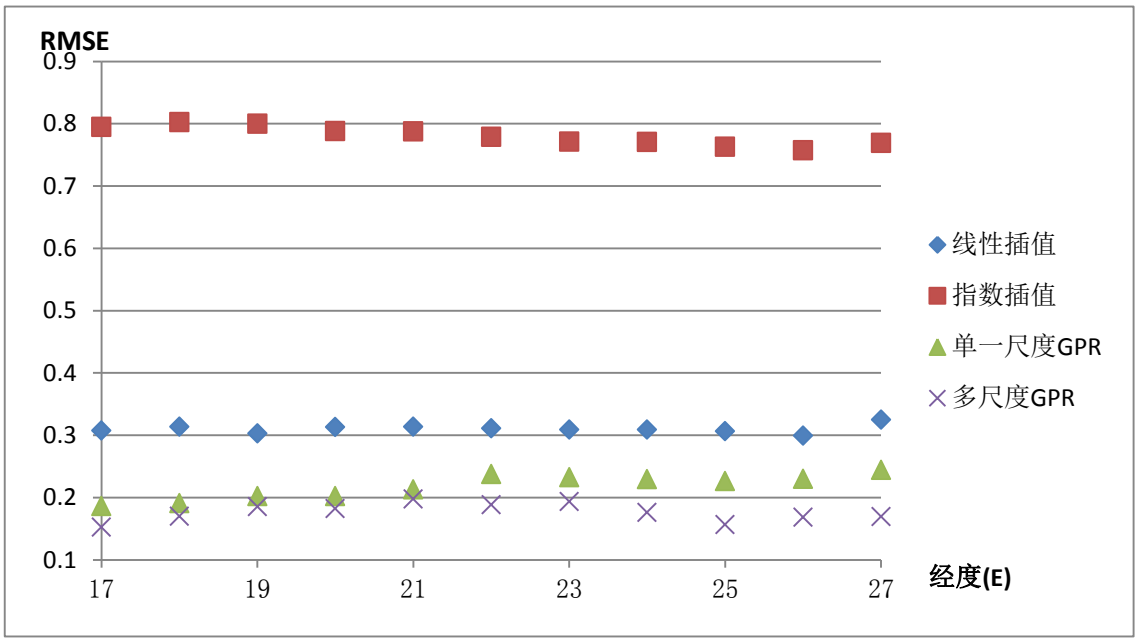


图 3-8 东经 17°~27° 50km 高度纬向风 u 四种插值方法 RMSE

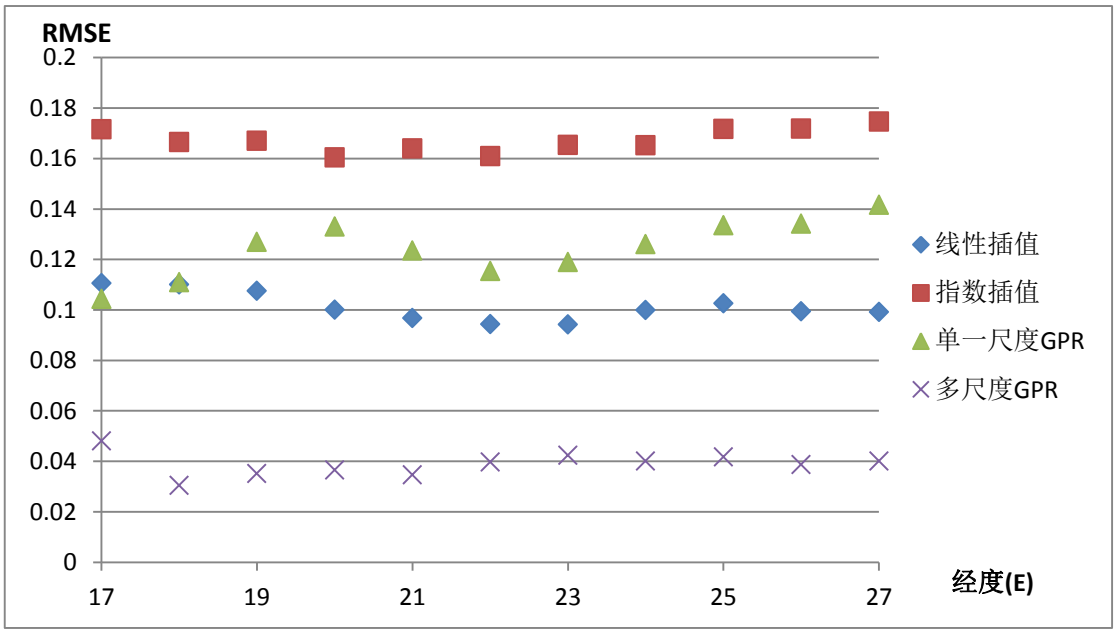


图 3-9 东经 17°~27° 50km 高度经向风 v 四种插值方法 RMSE

总之，复杂结构的高空风速可以用风速多尺度协方差核函数模拟。单一核函数的描述能力比较有限，只能学习到待插值样本邻域的训练数据，学习得到的数据趋势并不是全局的变化趋势。而风速多尺度核函数综合了泛化能力比较强的全局核函数和学习能力较好的局部核函数，前者增强了序列全局特征的模拟，后者增强了高斯过程回归模型的局部非线性逼近能力。将局部和全局核函数的根据物

理现象的特点进行多种空间尺度的划分，通过整合不同尺度的空间插值核函数来获得更好的数据描述能力。

3.4 本章小结

针对气象领域远距离大气变量插值计算的需要，本章系统地讨论了使用高斯过程回归进行插值不同核函数和超参数选取对结果的影响。组合核函数相对于单一核函数而言，对复杂序列的描述能力比较强。大气变量在空间中的空间关系不仅有局部的平滑关系，也遵循一定物理定律会有大、中、小不同尺度的全局和区域影响。

本文提出了一种风速多尺度插值核函数，将组合核函数高斯过程回归算法应用到高空风速数据建模中。根据气象变量纬向风和经向风的特点，将核函数构造为由大尺度风速协方差函数、中小尺度风速协方差函数、空间相关噪声协方差函数和独立高斯噪声组成的组合核函数。本文分别选取东经 16°经线所有区域 1km、20km、50km 高度和东经 17°~东经 27°50km 高度进行风速插值实验，实验结果表明，相对于线性插值、指数插值和单一平方指数协方差核函数插值，本文提出的风速多尺度插值方法在纬向风和经向风插值上都取得了更好的效果，并且垂直高度越高优化效果越明显，均方根误差越小。其中，由于数据是沿着经线选取，因此多尺度核函数经向风分量插值的均方根误差比单一尺度核函数插值、线性插值的均方根误差减小了 1/2 以上，这表明本文提出的风速多尺度插值核函数是有效的。

第四章 多变量相关的长短期记忆网络时间序列预测算法

风速信号本身是多个物理过程导致的结果，与温度、气压、密度等气象变量有复杂的非线性关系，因而风速序列的插值计算需要考虑多变量约束关系，否则可能会丢失序列中隐藏的重要信息。在数据量充足的情况下，长短期记忆网络因其时间记忆功能成为了时间序列预测非常有效的方法。本章引入机器翻译领域 Seq2Seq 方法，假设每个字由一个向量表示，则一句话就由一个向量序列表示。该方法的功能就是输入某个长度的句子（向量序列），输出后续的词或句子（向量序列）。本章的主要思想就是将多个变量组成一个风速状态向量，预测单变量的过程就转变为预测风速状态向量的出现概率。因此，风速状态向量时间序列预测的过程就变成了根据历史风速状态向量时间序列建模，从而推测出后续风速状态向量序列的值。以纬向风（或经向风）、温度、位势高度、相对湿度为约束变量进行实验，验证方法的有效性。

4.1 物理学中的变量约束

在物理学约束中，风的产生是由于水平气压梯度力的作用下产生的空气运动，大气运动就是由此产生，而气压梯度力的出现是由于各地存在的气压差产生的由高气压区流向低气压区的力。单位质量的气体，气压梯度力的方向与等压面垂直，且从高压指向低压，其大小与气压梯度成正比，与空气密度成反比。如果在高气压中心，水平气压梯度力垂直于等压线且由中心指向外；如果在低气压中心，水平气压梯度力垂直于等压线且由外指向中心。由于地球公转和自转的原因，地表温度不均，在高低纬度之间产生了热量差异，带来的气压差使得气团出现了上升或下降运动，水平方向的气压差导致了水平风的存在。

也就是说，风速的大小和方向与气压梯度大小和方向有关。以日本气象厅（JMA）和美国联合台风警报中心（JTWC）为首的各大气象研究机构也有自己的一套风速-气压梯度对应规则，尤其针对台风预报，台风中心的风速和气压也有一定的对应关系。

上面讨论了风速和气压梯度、气压值的关系。下面我们讨论风速和温度、密度的关系。假设气压 p，温度 T 和密度 ρ。而根据理想气体定律

$$p = \rho RT \tag{4-1}$$

其中 R 是理想气体常数。空间中两个位置之间气压 p 会存在气压差，如果我们认为这个差值是一个扰动，我们称之为 p'，则气压、温度、密度扰动值之间必须

存在的关系，即：

$$p'/p_0 = \rho'/\rho_0 + T'/T_0 \tag{4-2}$$

其中，p'、ρ'、T'分别是气压、密度、温度扰动值，定义为平均气压p_0、密度ρ_0、温度值T_0的一个偏差。因此，气压差的变化也与密度差、温度差的变化有关。从而，我们可以得知，风速的变化直观上是受气压差的影响，其间接也受到密度差、温度差的影响。

4.2 基于高维度张量的长短期记忆网络算法的风速预测方法

在 Mohamed Akram Zaytar 的论文中[57]，作者已经将 LSTM 方法引入到风速时间序列中来，利用 LSTM 强大的记忆功能，对风速进行预测，结果证明该方法进行逐小时的气象变量预测是可行的。但是作者没有考虑到温度、气压等变量与风速以及风速分量的约束关系，没有讨论单一风速标量和多种气象变量组合的风速状态向量对风速预测结果的不同影响，也并未讨论使用何种气象变量组合的风速状态向量可以能真正有效地提高预测结果。本文在 LSTM 方法基础上，考虑多变量约束关系，采用不同的风速状态向量取代单一风速标量，建立多变量约束的短期风速预测模型，进一步讨论风速状态向量对预测结果的影响，并且讨论气象变量何种组合对预测结果的优化作用最为明显。

4.2.1 基于长短期记忆网络风速预测算法的建模

本文设计的风速预测算法是基于 Seq2Seq 方法，原理如图 4-1。基于 LSTM 的 Seq2Seq 方法是机器翻译领域一种已有的方法，本文将词向量的概念引入风速预测中，设计一种采用风速状态向量进行多变量约束的短期风速预测方法，通过风速状态向量加入其他变量因子，通过“状态-状态”的“翻译”实现预测功能。首先选取一定区域的训练样本。之后选取不同组合的风速状态向量用以对比预测结果，设计 Seq2Seq 方法的各层网络结构，并设置网络参数的初始值。此时，确定了风速预测的数据格式和网络结构。其次，将样本区域里每个点的多变量数据组合成风速状态向量，每个点每个时刻对应一个风速状态向量，之后将向量序列送入网络中训练参数。最后，我们在训练过程中交叉验证得到最适合训练集的网络结构，将这个网络应用到测试集，得到测试集的预测结果，并对比预测结果与单变量预测的差异。

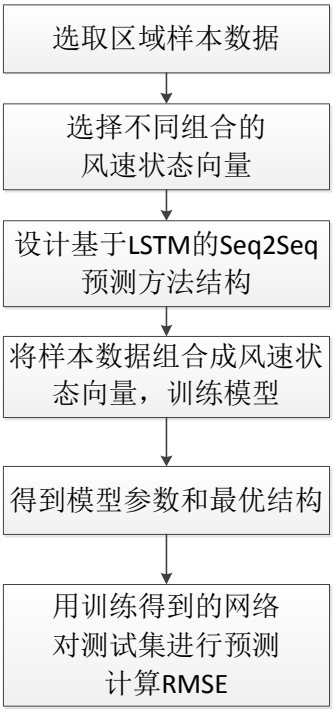


图 4-1 基于 LSTM 的多变量约束风速预测方法原理图

4.2.2 Seq2Seq 模型设计

根据历史信息来预测未来信息，历史信息可能会帮助我们建立能够蕴含常规趋势和运动特点的模型。对于这种长期依赖问题，LSTM 能够较好地解决此类时间序列问题。LSTM 是一类能够学习长期依赖的循环神经网络。从数学含义上来量，LSTM 的目的是预测条件概率 $\mathrm{p}(y_1,\ldots,y_N \mid x_1,\ldots,x_N)$，其中 $(x_1,\ldots,x_N)$ 是输入序列，$(y_1,\ldots,y_N)$ 是相同长度的输出序列。LSTM 计算最后一个隐藏状态经过输入序列 $(x_1,\ldots,x_N)$ 得到的固定维度的序列 v，得到该状态的条件概率。计算 $(y_1,\ldots,y_N)$ 有个标准 LSTM-LM 等式，其初始隐藏状态是 $(x_1,\ldots,x_N)$ 的代表 v：

$$\mathrm{p}(y_1,\ldots,y_N \mid x_1,\ldots,x_N) = \prod_{t=1}^{N} p(y_t \mid v, y_1,\ldots,y_{t-1}) \tag{4-3}$$

本章根据基于 LSTM 的 Seq2Seq 方法[73]，建立了一个包括两个 LSTM 层、一个包含 100 个神经元的全连接层的多层模型。从 LSTM 的外层看，模型的结构非常简单。输入是历史数据序列，输出是接下来 timestep 步预测结果。假设输入的是某个点的风速相关状态序列，每组 (u)、(u,t)、(u,h)、(u,t,h)、(u,t,h,rh) 状态数据的维度分别是 1、2、3、4 维的。假设输入数据是 (u,t,h) 序列，输入的时间步长度是 100，输出的时间步长度是 4，那么输入时间序列组成了一个长度是 100 的三元组，输出序列组成了一个长度是 4 的风速状态三元组。根据 Seq2Seq 方法，输入序列和输出序列的长度可以不同。输入输出和模型的数据形式如图

4-2。

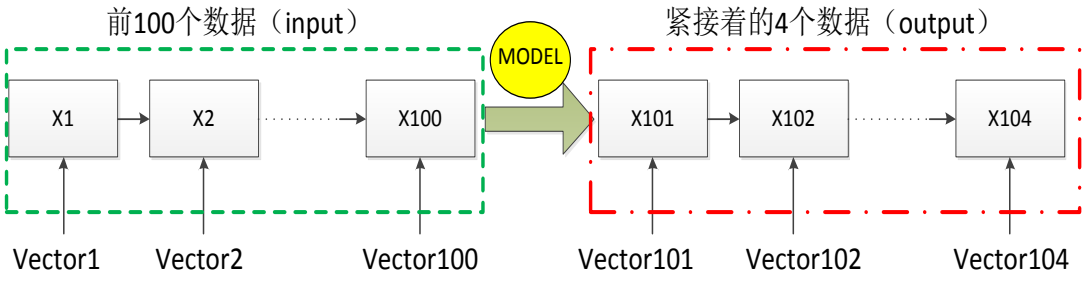


图 4-2 输入输出数据流

假设每个风速状态向量的维度是 r 维，状态向量序列的历史输出长度为 $len1$，预测时间步长度为 $len2$ 其总的层结构和 I/O 维度如下图所示：

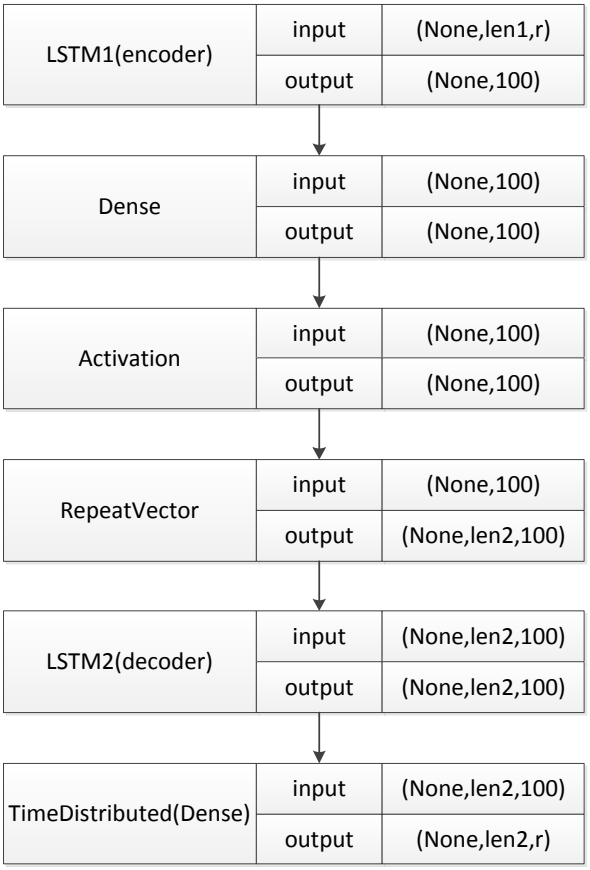


图 4-3 包含输入输出格式的 Seq2Seq 模型结构

其中，各层的功能介绍如下：

1）LSTM1 层：作为编码器的 LSTM，目的是把一个输入序列压缩表示成一个固定长度的向量。

2）Dense 层：一个神经元的全连接层，根据多次试验选择得到的 100 个神经元的值

3）Activation 层：选定的整流函数（Rectifier function）用作激活函数。

4）Repeat Vector 层：该层重复最终 encoding 层的输出向量作为 decoder 的

每一个时间步的一个恒定输入。

5）LSTM2 层：作为解码器的 LSTM，根据编码器的向量，生成一个 token 序列，这个 token 序列就是生成的要翻译的另一个输出序列。

6）Time Distributed(Dense)层：把一个相同的 Dense 层（全连接层）操作应用到一个 3 维张量的每一个时间步上。

我们用 RMSprop 作为网络的优化器，它每个权重都保持上一个时刻平方梯度的一个平均移动量：

$$E\left[g^2\right]_t = 0.9E\left[g^2\right]_{t-1} + 0.1g_t^2 \tag{4-4}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E\left[g^2\right]_t + \varepsilon}} g_t \tag{4-5}$$

RMSprop 方法用一个平方梯度的指数衰减平均项除以学习率。这种方法推荐γ设置为 0.9，推荐学习率η设置为 0.01。神经网络选择均方误差预测量作为损失函数，形式如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^p - Y_i^r\right)^2} \tag{4-6}$$

如上建立了多变量风速预测的核心模块。然后，将建立好的风速状态向量序列作为历史数据加入网络模型中，采用极大似然估计优化损失函数，使输入序列通过编码器编码再经过解码器得到输出序列的概率最大。如此，我们就确立了模型的参数和结构。

4.2.3 数据预处理及参数的确定

训练数据选取纬向风风速u、经向风风速v、温度t、位势高度h、相对湿度rh五个变量，目的在于研究风速数据u、v预测过程中，与u相关的风速相关状态向量为(u)、(u,t)、(u,h)、(u,t,h)、(u,t,h,rh)，与v相关的风速相关状态向量为(v)、(v,t)、(v,h)、(v,t,h)、(v,t,h,rh)。

表 4-1 构建风速状态向量的变量

编号	气象变量	符号	单位
1	纬向风风速	u	m/s
2	经向风风速	v	m/s
3	温度	t	K
4	位势高度	h	gpm
5	相对湿度	rh	$\%RH$

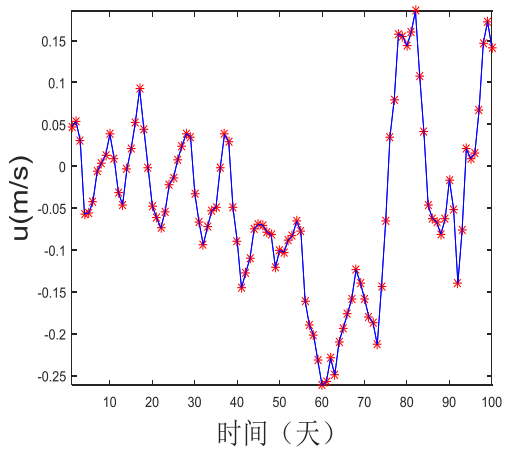
数据来源是美国环境预报中心(NCEP)和国家大气研究中心(NCAR)联合推

出的日平均再分析资料集，其中提供经向风风速、纬向风风速、垂直风速、温度、位置高度、相对湿度、绝对湿度等变量。其中，模式层是根据气压分层，在同一个等压面上，海拔高度和位势高度各不相同。我们选取固定点的三维坐标，最佳选择是经度、纬度、海拔高度，这样在固定点可以得到与风速密切相关的气压信息。不过由于源数据的限制，实际得到的固定点是根据经度、纬度、模式层选取的，因此选取点的气压都是一致的，但是海拔高度不同。

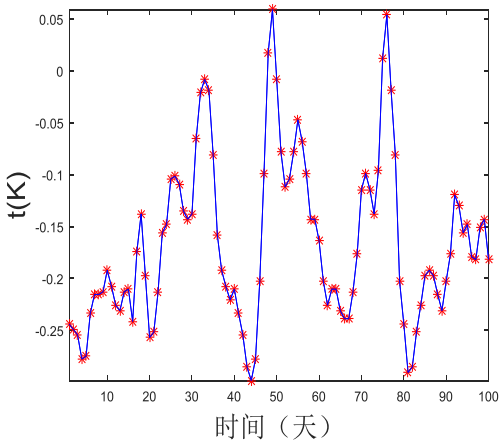
数据清洗后，非常重要的一点，所有的值都需要归一化，使其值收缩到$[-1,1]$的范围内来避免训练中出现的陷入局部最小值的问题，优化加速损失函数的衰减过程。同时，归一化方法也使得在标准输入下权重参数和贝叶斯估计能够更加快速。常用的方法如下：

$$X_i = \frac{X_i - Mean(X)}{MAX(X) - MIN(X)} \tag{4-7}$$

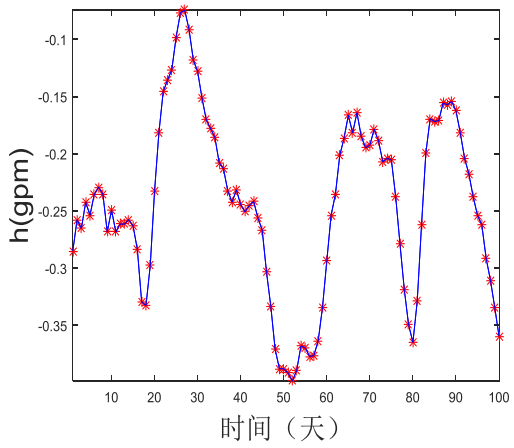
归一化处理后的 1948-1975 年东经 115°北纬 40°1000hpa 高度日平均纬向风、温度、位势高度、相对湿度的部分真值如图 4-5。

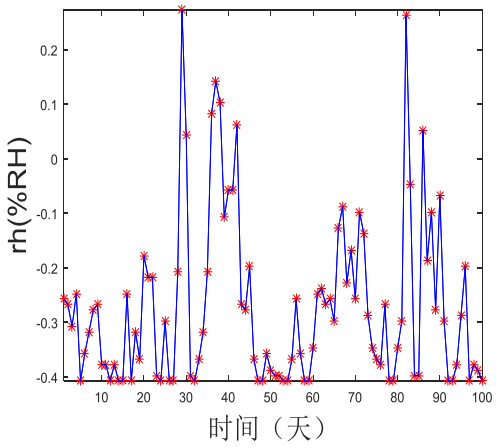


（a）u



（b）t





（c）h　　（d）rh

图 4-4 1948-1975 年东经 115°北纬 40°1000hpa 高度日平均变量分布

所有的数据都要改造成状态向量序列的形式，以(u,t,h)为例，每一个输入包括 100 个 3 元组(u,t,h)。选定的天气数据都要分到三个选定的序列里。训练集包括序列的 80%的数据，验证集包括序列的 10%的数据，验证方法采用交叉验证，测试集包括序列 10%的数据。网络衡量误差的标准是均方根误差。

LSTM 模型的训练过程相对就比较简单。模型里一共包括两个 LSTM 层和一个全连接层，每个 LSTM 都只有一个层，都分别包含 100 个神经元。训练细节如下：

1）所有的 LSTM 参数初始化都服从[-0.05,0.05]的均匀分布。

2）批量梯度下降法的学习率用一个固定值 0.001。

3）梯度下降算法一次训练的序列长度是 512，循环 50 次。

4.2.4 实验结果分析

本节引入均方根误差 RMSE 来判别预测的质量。我们记录下测试集输出结果的均方根误差进行对比分析。表 4-2 是 115°40°1000hpa 风速分量值（归一化）预测的 RMSE。由于进行了公式 4-7 的归一化处理，因此所有值 RMSE 都小于 1。

表 4-2 115°40°1000hpa 风速u v预测的 RMSE（归一化）

输入维度	输出维度	RMSE					
		Timestep=1（6 小时）		Timestep=2 （ 12 小时）		Timestep=4 （ 24 小时）	
		u (或v)	状态向量	u (或v)	状态向量	u (或v)	状态向量
(u,t,h,rh)	(u,t,h,rh)	0.0444	0.1094	0.0686	0.1228	0.1052	0.1355
(u,t,h)	(u,t,h)	0.0412	0.0318	0.0664	0.0415	0.1036	0.0617
(u,t)	(u,t)	0.0418	0.0334	0.0668	0.0443	0.1005	0.0648
(u,h)	(u,h)	0.0417	0.0366	0.0665	0.0479	0.1004	0.0675
(u)	(u)	0.0454	0.0454	0.0683	0.0572	0.1014	0.0814
(v,t,h,rh)	(v,t,h,rh)	0.0592	0.1095	0.0855	0.1274	0.1162	0.1475
(v,t,h)	(v,t,h)	0.0568	0.0389	0.0834	0.0484	0.1114	0.0641

(v,t)	(v,t)	0.0601	0.0482	0.0931	0.0624	0.1215	0.0789
(v,h)	(v,h)	0.0561	0.0417	0.0828	0.0531	0.1133	0.0686
(v)	(v)	0.0609	0.0609	0.0958	0.0808	0.1260	0.1030

从变量类型角度看，(u,t)、(u,h)、(u,t,h) 预测结果要明显优于 u 的结果，由于 u、t、p 有比较强的相关性，而在 NCEP/NCAR 的数据中，虽然选定点是由气压模式层确定的，即气压是个定值，但是等压面上的位势高度是不同的，该值间接地体现了气压的影响，因此三种状态向量的序列预测准确度较好，其他两个变量的引入提高了风速预测的准确性。而加入了 rh 的预测结果并没有显示出向量预测的优势，误差比较大，在更长时间步中甚至比单一 u 变量的误差还要打，这说明 rh 与其他三个量的相关性比较远，该值的引入对风速 u 时间序列规律的探索造成了干扰。

从序列预测长度看，预测步长越长，误差越大。横向看，(u)、(u,t)、(u,h)、(u,t,h)、(u,t,h,rh) 中 u 的预测结果随着步长的增长而增大。纵向看，大多序列都保持了“状态向量序列预测结果优于单一纬向风分量预测结果”，但是 (u,t,h,rh) 在长时间步长的结果却比单一 u 变量的结果差。由于 rh 和风速、温度、位势高度等的相关性低，随时间变化的趋势和周期与其他变量的趋势和周期不同，由于该值只占向量的一项，其他相关性强的变量占向量的三项，为了使目标函数值最小，模型参数会更偏向于遵循温度 t、位势高度 h 等的规律，因此 rh 的拟合程度较差。经向风分量 v 的结果与纬向风分量 u 类似。

为了进一步验证这种规律是否适用于更广泛的区域，我们增大实验范围，采用东经 115° 和东经 130° 两条经线的所有纬度范围进行实验。实验在北纬 10° ~北纬 80° 之间，以 10° 为间隔，得到如表 4-3 和表 4-4 的纬向风状态向量预测结果。上文为了对比单变量和状态向量整体的预测准确性，表 4-2 的采用了归一化处理的结果。而下文单独讨论变量 u 和变量 v，我们将采用由逆归一化处理恢复得到的值。从表 4-3 和表 4-4 中可看出，从整体趋势上看，(u,t)、(u,h)、(u,t,h) 整体上的 RMSE 是小于 (u) 的 RMSE 的，但是在在赤道和靠近北极点附近，(u,t) 的表现不佳，会出现较大 RMSE 的情况，(u,t) 在中纬度地区表现较好。(u,h) 在所有纬度表现都较好，RMSE 都低于 (u)，但是效果不如 (u,t,h) 明显。(u,t,h) 在绝大多数纬度的 RMSE 最低，且整体低于 (u,h) 的结果。(u,t,h,rh) 在整个区域的 RMSE 不稳定，没有明显优化纬向风 u 的预测结果，在北纬 60° ~80° 范围出现了比 (u) 差的结果，这与表 4-2 单点预测中 (u,t,h,rh) 状态向量出现较大误差是一致的。(u,t,h,rh) 在高纬度地区会降低风速预测的准确性。

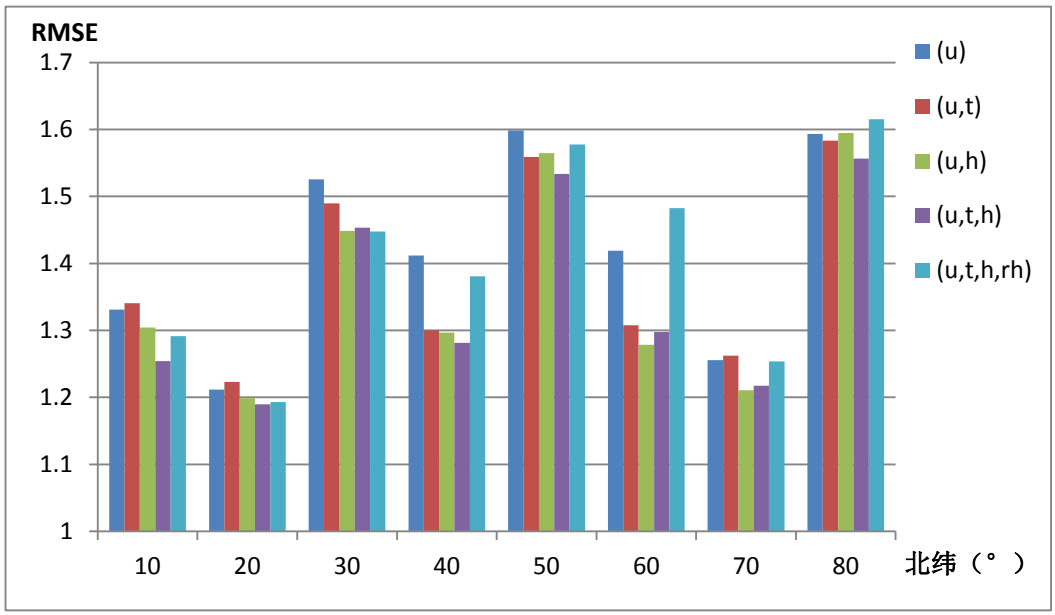


图 4-5 115°E10°~80°N 1000hpa 纬向风u预测结果 RMSE(timestep=1)

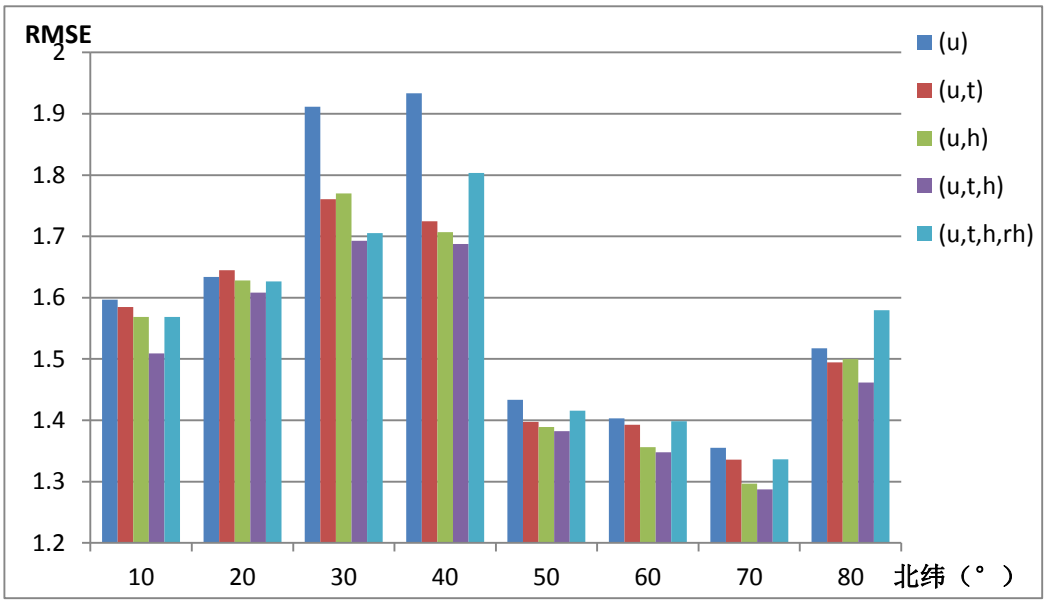


图 4-6 130°E10°~80°N 1000hpa 纬向风u预测结果 RMSE(timestep=1)

至此，我们得到了适合用于多变量约束的风速状态向量(u,t,h)。但是这种方法没有与其他预报的结果进行对比。本文选取两个预报系统模型的风速预测结果进行对比：第一个是 Aditya Grover 在 KDD2015 提出的梯度增强决策树（Gradient Boosted Decision Tree，简称 GBDT）和深度置信网络（DBN）结合的统计预报方法（后文简称深度混合方法），思路是用 GBDT 对单独的风速序列进行时间上的预测，然后将待预测点的其他变量（如温度t，气压p，密度ρ，露点温度d）和由 GBDT 得到的风速预测加入 DBN 中利用约束关系修正预测值[34]；第二个是 NOAA 提出的每 6 小时高空风预报（Winds Aloft Forecast）。第一项和第二项预测使用的源数据是地表风速观测，预测结果 RMSE 是由 Aditya 的文章[34]提供，

而本文使用的风速状态向量 LSTM 方法使用的源数据是再分析风速资料。本文提取三个时间步长用于比较：6 小时、12 小时、24 小时。表 4-3 是几种方法在西雅图站（47.6°N， 122.3°W，本文采用附近 47.5°N，122.5°W 的数据近似）的 RMSE 比较，表中上述几个系统的数据历史数据长度为 1 个月的后续预测。表中可知，在西雅图站风速状态向量 LSTM 方法的表现明显比 NOAA 的 Winds Aloft Forecast 好，风速状态向量(u,t,h)LSTM 方法在 12 小时、24 小时v预测上与深度混合表现相当，在u的预测要明显好于深度混合方法。

表 4-3 风速状态向量 LSTM 方法与当前风速预测表现较好的方法比较

时间步长 timestep	模型	RMSE	
		u	v
6 小时	风速状态向量 LSTM 方法	1.94	1.93
	深度混合方法	2.29	1.33
	NOAA	3.18	3.44
12 小时	风速状态向量 LSTM 方法	2.55	2.63
	深度混合方法	4.44	2.59
	NOAA	5.13	4.34
24 小时	风速状态向量 LSTM 方法	3.40	3.86
	深度混合方法	6.57	3.82
	NOAA	8.79	6.37

4.3 本章小结

常见大气变量统计预测方法中，大都是基于单一变量时间域上的预测，鲜有基于多变量向量时间域上的预测。LSTM 在处理时间序列问题上有很好的效果，基于 LSTM 的序列到序列（Seq2Seq）可使输入序列数据和预测输出序列数据的长度不一样。得益于这两种方法，我们将风速数据和与之相关的温度、气压、位势高度、相对湿度等量组成风速状态向量，作为输入来预测后续时间步的状态向量。实验结果表明，与温度、位势高度等与气压有关的量组成的风速状态向量的预测效果最好，向量中包含的元素越多，预测值越精确。但是，相对湿度对序列预测准确性的负面影响较大，不适合加入状态向量。因此，我们需要根据物理学关系，寻找与风速相关性大的变量组成状态向量。在中高纬度，状态向量(u,t)、(u,h)、(u,t,h)会整体提高向量的预测准确性，也会单独提高风速变量的预测准确性。其中，状态向量(u,t,h)在实验区域所有范围中（东经 115°~130°北纬 10°）预测的准确性最高，且适用范围最广。最后，本文提出的方法在西雅图地表风预测中得到了比 NOAA 的高空风预报更好的效果。

第五章 结论与展望

5.1 总结

针对高空风速具有多尺度混合、多变量影响的特点，本文将高斯过程回归算法和基于 LSTM 的序列到序列预测算法引入到气象变量的研究中，对高空风的时间序列预测和空间插值展开了研究。

论文的主要工作与结论如下：

（1）粗网格的大气资料在实际应用中需要进行经度、纬度、垂直方向的多重插值，大多数系统采用的是线性插值、指数插值等简单方法。在计算二维平面上的一些插值问题时也会用到经过贝叶斯优化的高斯过程回归。大气变量在应用广泛适用的插值方法时，需要结合变量自身的特点，考虑多尺度运动的插值计算。本文提出了一种风速多尺度插值核函数，改进了用于风速空间序列插值的高斯过程回归方法。传统的核函数是由一个单一尺度的风速协方差函数和独立高斯噪声项组成。本文根据风速多尺度特点，将核函数构造为由大尺度风速协方差函数、中小尺度风速协方差函数、空间相关噪声协方差函数和独立高斯噪声组成的组合核函数。实验选取东经 16°不同高度和东经 17°~东经 27°50km 高度纬向风和经向风再分析数据为样本数据，以 2°为间隔建模，以 1°为间隔插值验证。实验结果表明，相对于线性插值、指数插值和单一平方指数协方差核函数插值，本文提出的风速多尺度插值方法纬向风和经向风插值均方根误差（RMSE）更低，使用纬向风和经向风分量计算的误差远小于使用风速向量之模和风速向量角度计算的误差。风速多尺度插值方法的优化效果随垂直高度增加而增大。由于数据是沿着经线选取，因此多尺度核函数经向风分量插值的均方根误差比单一尺度核函数插值、线性插值的均方根误差减小了 1/2 以上，这表明本文提出的风速多尺度插值核函数是有效的。

（2）深度学习理论因为其强大的非线性映射能力而在众多领域应用。风速变量的变化不是孤立的，而是受到温度、气压等其他变量影响的，一个时刻的其他气象变量共同组成的状态与风速变量是有某种非线性关系的。因此本文提出了一种采用风速状态向量进行多变量约束短期风速预测方法。本文将风速时间序列元素由单一纬向风（或经向风）标量改进为纬向风（或经向风）与温度、位势高度、相对湿度组成的状态向量，分别应用到基于长短期记忆网络（LSTM）的序列到序列（Seq2Seq）预测方法中。实验选取东经 115°和 130°北半球区域的 1948 年-1972 年再分析数据进行训练，选取 1973-1975 年数据进行验证。在相同时间步长情况下，由风速、温度、位势高度组合构成的三种状态向量明显改善了

向量整体预测结果和纬向风分量（或经向风分量）预测结果。其中，由风速u、温度t、位势高度h组成的3维状态向量(u,t,h)的两种预测均方根误差（RMSE）都达到最低。而包含相对湿度rh的状态向量(u,t,h,rh)不仅未改善预测结果反而使均方根误差（RMSE）更大。实验结果表明，基于长短期记忆网络（LSTM）的多变量约束风速预测方法能够有效地优化风速时间序列预测结果，但是优化程度取决于风速状态向量各个分量的相关程度。相关度高的变量构成的风速状态向量优化效果明显，而不合适的状态向量可能会让结果变差。在中高纬度，状态向量(u,t)、(u,h)、(u,t,h)会整体提高向量的预测准确性，也会单独提高风速变量的预测准确性。其中，状态向量(u,t,h)在实验区域所有范围中（东经115°~130°北纬10°~80°）预测的准确性最高，且适用范围最广。本文提出的方法在西雅图地表风预测中得到了比NOAA的高空风预报更好的预测结果。

5.2 未来工作

目前本研究对考虑多变量约束的风速序列预测和插值研究尚处于起步阶段，在未来研究工作中，还有如下几个方面的问题需要进一步研究：

（1）目前只是提出了对风速多尺度运动核函数，没有讨论风速多变量约束核函数。实际问题中可以考虑将两个方面的内容组合起来，建立多尺度和多变量统一的核函数，以更好地适用于本文所提到的应用背景。

（2）本文进行水平面上变量的高斯过程回归中也考虑了距离因素，从减少计算复杂性的角度只考虑了线性距离，并未考虑二维空间距离。下一步研究可以考虑将一维空间关系的插值扩展到二维平面上的多尺度插值。

（3）由于大气垂直方向上的数据比水平方向上的少，因此时间和空间序列预测只建立了水平方向的模型。虽然垂直方向上的数据相对少，但是垂直方向上大气变量值变化非常平缓，跳跃情况少。对于实际应用中非常广泛的垂直变量预测，可以寻找合适的模型进行垂直方向的研究。

（4）在Sutskever的文章中，他发现Seq2Seq方法使用深层的LSTM比浅层的优化效果要好很多[67]。本文从时间复杂度的角度考虑仅使用了一个隐藏层的LSTM，没有验证过多个隐藏层的LSTM情况下的效果。下一步可以逐步添加多个隐藏层来验证深层网络对风速状态向量预测效果有多大程度上的提高。

（5）本文进行的所有模型方法都没用进行并行化，特别是基于深度学习的方法在时间复杂度上比较高，导致实用性不佳。从现实角度，需要利用GPU等其他资源进行并行优化，这一点的在其他领域也非常有研究价值。

总之，本文的工作是为了多变量序列预测提供了一定的研究基础，在未来的工作中还需要许多方面的努力。

致　谢

我衷心地感谢导师张卫民研究员在我硕士三年学习过程中对我学术上的帮助和指导，引领我入门资料同化领域，出差之余不忘给我开发上的建议，鼓励我参与自己感兴趣的活动，我都记在心里。导师严谨的治学态度、敬业的精神、对我生活和工作上的关心和帮助是我受益匪浅。同时，也感谢师母，师母的热情使我感受到实验室是一个大家庭，导师和师母是大家庭的家长，使我们有机会相聚团结在一起，借此机会，对他们表示最诚挚的感谢！

感谢刘柏年师兄对我学术上无微不至的关心，我没能给刘师兄的工程上贡献多少帮助，但是刘师兄一直指导我小论文、帮助我大论文定题，给我的论文提出了很多宝贵建议和意见，在方向上给予指导，细心认真地帮我解决困扰。感谢方民权师兄对我学术和生活上的建议，方师兄的学习态度给我做出了榜样。感谢孙敬哲师兄，虽然我们的主攻方向相差甚远，但是坐在身后的他让我在实验室找到了信心和勇气，许多细致的气象业务系统的问题都要多亏孙师兄的指导。

感谢朱孟斌、余意、张泽、段博恒、陈妍等师兄师姐的建议和帮助，感谢林士伟、邢德、邢翔、赵盼盼在生活学习上的帮助，有你们在，实验室始终是一个欢乐、温暖、相互关心的大家庭。

感谢谢婷婷、高畅等八队的同学和孙友佳队长、朱涛政委，感谢你们陪伴我走过快乐或失意的时光，你们认真帮我分析、为我考虑，正式你们的陪伴，我的硕士生活才如此充实。

感谢实习实验室的赵建辉老师和薛崇，点亮了我未来的方向，让我更有信心迎接新阶段的工作。

感谢我的父母和家人，感谢他们对我的无私支持和理解，他们是我坚强的后盾，我心灵的港湾，我奋斗的动力。

最后，衷心感谢各位在百忙之中评阅的各位老师、专家。

参考文献

[1] Smith O E. Vector wind and vector wind shear models 0 to 27 km altitude for Cape Kennedy, Florida, and Vandenberg AFB, California[J]. 1976.

[2] 张荣, 祁伟, 许坚, 等. 高空风 GRIB 报文解析及精度分析[J]. 空中交通管理, 2010 (4): 17-20.

[3] 赵人濂, 陈振官. 风切变与运载火箭设计[J]. 宇航学报, 1998, 19(2):105-108.

[4] Justus C G, Jeffries III W R, Yung S P, et al. The NASA/MSFC global reference atmospheric model-1995 Version (GRAM-95)[J]. 1995.

[5] 苏鹏宇. 考虑风速变化模式的风速预报方法研究[D]. 哈尔滨工业大学, 2013.

[6] Krasnopolsky V M, Fox-Rabinovitz M S. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction[J]. Neural Networks, 2006, 19(2):122-134.

[7] Richardson L, withaforwordbyPeterLynch. Weather prediction by numerical process[M]. New York, 2007.

[8] Marchuk G. Numerical Methods in Weather Prediction[J]. Numerical Methods in Weather Prediction, 1974:259–273.[9] K. Hanjalić, R. Van de Krol, A. Lekić. Sustainable Energy Technologies[M]. Springer Netherlands, 2008.

[10] 潘迪夫, 刘辉, 李燕飞. 基于时间序列分析和卡尔曼滤波算法的风电场风速预测优化模型[J]. 电网技术, 2008, 32(7):82-86.

[11] Huang Z, Chalabi Z S. Use of time-series analysis to model and forecast wind speed[J]. Journal of Wind Engineering & Industrial Aerodynamics, 1995, 56(2–3):311-322.

[12] Pourhabib A, Huang J Z, Ding Y. Short-term Wind Speed Forecast Using Measurements from Multiple Turbines in a Wind Farm[J]. Technometrics, 2016, 58(1):00-00.

[13] Kuligowski R J, Barros A P. Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks[J]. Weather & Forecasting, 2010, 13(4):1194-1204.

[14] Horenko I, Klein R, Dolaptchiev S, et al. Automated Generation Of Reduced Stochastic Weather Models I: Simultaneous Dimension And Model Reduction For Time Series Analysis[J]. Siam Journal on Multiscale Modeling &

Simulation, 2008, 6(4):14241-9.

[15] Chen L, Lai X. Comparison between ARIMA and ANN Models Used in Short-Term Wind Speed Forecasting[C]// Asia-Pacific Power and Energy Engineering Conference. IEEE Computer Society, 2011:1-4.

[16] Voyant C, Muselli M, Paoli C, et al. Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation[J]. Energy, 2012, 39(1):341–355.

[17] Giorgi M G D, Ficarella A, Tarantino M. Error analysis of short term wind power prediction models[J]. Applied Energy, 2011, 88(4):1298-1311..

[18] Radhika Y, Shashi M. Atmospheric Temperature Prediction using Support Vector Machines[J]. 2009, 1(1):55-58.

[19] Sapankevych N I, Sankar R. Time Series Prediction Using Support Vector Machines: A Survey[J]. Computational Intelligence Magazine IEEE, 2009, 4(2):24-38.

[20] Mohandes M A,Halavani T O, Rehman S,et al. Support vector machines for wind speed prediction[J]. Renewable Energy,2004,29(6):939-947.

[21] 吴栋梁, 王扬, 郭创新,等. 基于改进 GMDH 网络的风电场短期风速预测[J]. 电力系统保护与控制, 2011, 39(2):88-93.

[22] Gneiting T, Raftery A E. Weather Forecasting with Ensemble Methods[J]. Science, 2005, 310(5746):248.

[23] 刘明凤, 修春波. 基于 ARMA 与神经网络的风速序列混合预测方法[C]// 中国智能自动化会议. 2013.

[24] Guo Z, Zhao W, Lu H, et al. Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model[J]. Renewable Energy, 2012, 37(1):241–249.

[25] Liu H, Chen C, Tian H Q, et al. A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks[J]. Renewable Energy, 2012, 48(6):545–556.

[26] 席剑辉, 韩敏. 主成分分析与神经网络的结合在多变量序列预测中的应用[J]. 控制理论与应用, 2007, 24(5):719-724.

[27] 张升堂, 张楷. 椭圆指数函数降水空间插值模型[J]. 南水北调与水利科技, 2015, 13(3):530-533.

[28] 樊子德, 李佳霖, 邓敏. 顾及多因素影响的自适应反距离加权插值方法[J]. 武汉大学学报(信息科学版), 2016, 41(6):842-847.

[29] 杨成生, 张勤, 张双成,等. 改进的 Kriging 算法用于 GPS 水汽插值研究

[J]. 国土资源遥感, 2013, 25(1):39-43.

[30] 黄安, 杨联安, 杜挺,等. 基于多元成土因素的土壤有机质空间分布分析[J]. 干旱区地理(汉文版), 2015, 38(5):994-1003.

[31] 吴小芳, 包世泰, 胡月明,等. 多因子空间插值模型在农作物病虫害监测预警系统中的构建及应用[J]. 农业工程学报, 2007, 23(10):162-166.

[32] Pourhabib A, Huang J Z, Ding Y. Short-term Wind Speed Forecast Using Measurements from Multiple Turbines in a Wind Farm[J]. Technometrics, 2016, 58(1):00-00.

[33] Moore C J, Berry C P L, Chua A J K, et al. Improving gravitational-wave parameter estimation using Gaussian process regression[J]. Physics, 2015, 93(6).

[34] Grover A, Kapoor A, Horvitz E. A Deep Hybrid Model for Weather Forecasting[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015:379-386.

[35] Wang Y, Chaib-Draa B. An Online Bayesian Filtering Framework for Gaussian Process Regression: Application to Global Surface Temperature Analysis[J]. Expert Systems with Applications, 2016, 67:285-295.

[36] Grbić R, Kurtagić D, Slišković D. Stream water temperature prediction based on Gaussian process regression[J]. Expert Systems with Applications, 2013, 40(18):7407-7414.

[37] Mihoub R, Chabour N, Guermoui M. Modeling soil temperature based on Gaussian process regression in a semi-arid-climate, case study Ghardaia, Algeria[J]. Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 2016, 2(4):397-403.

[38] Yu J, Chen K, Mori J, et al. A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction[J]. Energy, 2013, 61(6):673-686.

[39] Zhang C, Wei H, Zhao X, et al. A Gaussian process regression based hybrid approach for short-term wind speed prediction[J]. Energy Conversion & Management, 2016, 126:1084-1092.

[40] 翁小清, 沈钧毅. 多变量时间序列异常样本的识别[J]. 模式识别与人工智能, 2007, 20(4):463-468.

[41] 周大镯, 姜文波, 李敏强. 一个高效的多变量时间序列聚类算法[J]. 计算机工程与应用, 2010, 46(1):137-139.

[42] 刘志红, Ti m R.McVicar, Tom G.Van Nie,等. 基于 ANUSPLIN 的时间序列气象要素空间插值[J]. 西北农林科技大学学报自然科学版, 2008,

36(10):227-234.

[43] 张玉虎, 刘凯利, 陈秋华,等. 区域气象干旱特征多变量 Copula 分析——以阿克苏河流域为例[J]. 地理科学, 2014, 34(12):1480-1487.

[44] 管孝艳, 王少丽, 高占义,等. 基于多变量时间序列 CAR 模型的地下水埋深预测[J]. 农业工程学报, 2011, 27(7):64-69.

[45] 刘立霞. 多变量金融时间序列的非线性检验及重构研究[M]. 南开大学出版社, 2011.

[46] 曾耀, 李春峰. 基于 RBF 多变量时间序列的滑坡位移预测研究[J]. 长江科学院院报, 2012, 29(4):30-34.

[47] 谭忠富, 张金良. 利用多因素小波变换和多变量时间序列模型的日前电价预测[J]. 中国电机工程学报, 2010, 30(1):103-110.

[48] 李权, 周兴社. 基于 KPCA 的多变量时间序列数据异常检测方法研究[J]. 计算机测量与控制, 2011, 19(4):822-825.

[49] Yang K, Shahabi C. A PCA-based similarity measure for multivariate time series[C]// ACM International Workshop on Multimedia Databases, Acm-Mmdb 2004, Washington, Dc, Usa, November. DBLP, 2004:65-74.

[50] Hwang C L, Jan C. Recurrent-Neural-Network-Based Multivariable Adaptive Control for a Class of Nonlinear Dynamic Systems With Time-Varying Delay[J]. IEEE Transactions on Neural Networks & Learning Systems, 2015, 27(2):388-401.

[51] Malhotra P, Ramakrishnan A, Anand G, et al. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection[J]. 2016.

[52] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[C]// Interspeech. 2012:601-608.

[53] Sundermeyer M, Ney H, Schluter R. From Feedforward to Recurrent LSTM Neural Networks for Language Modeling[J]. IEEE Transactions on Audio Speech & Language Processing, 2015, 23(3):517-529.

[54] Yao K, Zweig G, Hwang M Y, et al. Recurrent Neural Networks for Language Understanding[C]// Interspeech. 2013.

[55] Mesnil G, He X, Deng L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[J]. Interspeech, 2013.

[56] Sundermeyer M, Alkhouli T, Wuebker J, et al. Translation Modeling with Bidirectional Recurrent Neural Networks[C]// Conference on Empirical Methods in Natural Language Processing. 2014.

[57] Mohamed Akram Zaytar, Chaker El Amrani. Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks[J]. International Journal of Computer Applications (0975 - 8887), 2016,Volume 143 - No.11, June 2016,143(11).

[58] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)[M]// Gaussian processes for machine learning /. MIT Press, 2006:69-106.

[59] 何志昆, 刘光斌, 赵曦晶,等. 高斯过程回归方法综述[J]. 控制与决策, 2013(8):1121-1129.

[60] Rumelhart D, Mcclelland J. Learning Internal Representations by Error Propagation[M]// Parallel Distributed Processing. Exploration of the Microstructure of Cognition. 1986:318-362.

[61] Werbos P J. Generalization of backpropagation with application to a recurrent gas market model[J]. Neural Networks, 1988, 1(4):339-356.

[62] Bishop C M. Neural Networks for Pattern Recognition[J]. Agricultural Engineering International the Cigr Journal of Scientific Research & Development Manuscript Pm, 1995, 12(5):1235 - 1242.

[63] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5):359-366.

[64] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets.[J]. Neural Computation, 1989, 18(7):1527-1554.

[65] Bengio Y, LeCun Y. Scaling learning algorithms towards AI[J]. Large-scale kernel machines, 2007, 34(5): 1-41.

[66] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks[M]. Springer Berlin Heidelberg, 2012.

[67] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[68] Geoffrey Hinton. Overview of mini-batch gradient descent. http://goo.gl/A9lKPi, 2014. [Online; accessed 09-June-2016].

[69] Leslie F W, Justus C G. The NASA Marshall Space Flight Center Earth Global Reference Atmospheric Model—2010 Version The NASA STI Program…in Profile[J]. 2011.

[70] 张宇, 郭振海, 张文煜,等. 中尺度模式不同分辨率下大气多尺度特征模拟能力分析[J]. 大气科学, 2010, 34(3):653-660.

作者在学期间取得的学术成果

[1] 朱祥茹, 刘柏年, 张卫民. 全球大气再分析产品应用平台的通用框架及组件设计[C]// 中国气象学会年会 s20 气象信息化. 2016.