摘 要

基于深度学习的海洋温度数据补全及预测方法研究

海洋温度在海洋水文要素中非常重要，是影响气候变化的关键因子，对声速、声波的传播、声呐设备的传播距离均有显著影响，间接影响水声网络连通性。因此，深入挖掘及精准预测海洋温度的时空分布及其变化规律，有助于海洋中不同深度下声速以及声波传播距离的估计，传播路径的分析研究，对水声工程的发展，尤其是水声网络的连通有着重要意义。

然而，现有的海洋观测数据存在数据表征不完备以及时序特征表征不完备问题，给海洋温度的规律性研究带来了挑战。传统的数据补全与温度预测模型主要包括数值模型和基于统计学的模型，他们往往结构简单，无法更好的捕获数据的非线性特征，而新兴的深度学习模型的效果往往依赖于超参数的选择，难以取得理想精度。本文重点从深度学习模型切入，对海洋温度空间和时间上的特征以及分布规律进行充分挖掘和学习，提取空间特征用于缺失数据补全，提取时间特征用于时间序列预测。

从数据表征不完备切入，针对海洋 Argo 数据集存在某些经纬度温度数据缺失的问题，本文将 BiLSTM（Bi-directional Long Short-Term Memory）层与 GRU（Gated Recurrent Unit）层相结合，在双层 GRU 模型可以很好的捕捉海洋温度空间趋势性的基础上，综合了 BiLSTM 能够学习海洋温度空间关联性的能力，提出了基于多层 BiLSTM-GRU 模型的缺失温度数据补全方法，构建了 GRU-BiLSTM-GRU 模型，并对最优神经元进行了选择，获得了理想的回归效果，其精度优于基线模型 LSTM、BiLSTM 和 GRU。然而，超参数的选取往往对精度有着较大影响，现有的模型往往采用手动选取超参数的策略，极大的浪费了时间成本。因此，本文提出了基于贝叶斯优化算法（Bayesian Optimization, BO）的超参数自动寻优方法，对多层 BiLSTM-GRU、LSTM、BiLSTM 和 GRU 进行优化，分别构建了 BO-BiLSTM-GRU、BO-LSTM、BO-BiLSTM 和 BO-GRU 模型，节约了超参数的选择成本，同时提高了模型精度。实验证明，优化后的模型相较于优化前均取得较高的精度提升，且本文提出的 BO-BiLSTM-GRU 模型优于 BO-LSTM、

BO-BiLSTM、BO-GRU、ConvLSTM、ConvGRU 和 M-ConvLSTM。

在数据表征完备的基础上，从时序特征表征不完备切入，对海洋温度在时间分布上的规律性特征进行提取和分析，本文提出了基于时序特征构建与一维卷积神经网络（1D-CNN）的温度预测方法，首先对连续的海洋温度时间序列进行构建，并在选取的 8 个不同的深度样本上对时序特征系数进行了选取，然后利用 1D-CNN 进行时序特征分解与提取，并对温度进行预测，最终将预测精度与基线模型进行比较。实验证明，在 8 个不同深度的样本数据上，其中有 6 个深度样本，1D-CNN 优于模型 BiLSTM、LSTM、GRU、BP、KNN。然而，单个 1D-CNN 模型常常只在某方面学习和预测效果良好，因此，本文将 1D-CNN 与集成学习 Adaboost 模型相结合，提出了 1D-CNN-Adaboost 模型，该方法将 1D-CNN 作为弱学习器训练，通过多个弱学习器，最终构成强学习器，使得模型在整体效果上均表现良好，提升了 1D-CNN 的学习能力和预测效果。实验证明，1D-CNN-Adaboost 模型在 8 个样本深度上均优于 1D-CNN 模型，验证了 Adaboost 模型对提高 1D-CNN 模型预测精度具有有效性。

关键词：

海洋温度，多层 BiLSTM-GRU，贝叶斯优化，1D-CNN，Adaboost

Abstract

Research on Ocean Temperature Data Completion and Prediction Method Based on Deep Learning

Ocean temperature is very important in marine hydrological factors, and is a key factor affecting climate change. It has a significant impact on the speed of sound, the propagation of sound waves, and the propagation distance of sonar equipment, and indirectly affects the connectivity of underwater acoustic network. Therefore, it is of great significance for the development of underwater acoustic engineering, especially for the connectivity of underwater acoustic networks, to deeply excavate and accurately predict the temporal and spatial distribution of ocean temperature and its change rules, which will help to estimate the speed of sound and the propagation distance of sound waves at different depths in the ocean, and analyze the propagation path.

However, the existing ocean observation data have the problems of incomplete data representation and incomplete temporal feature representation, which brings challenges to the study of the regularity of ocean temperature. The traditional data completion and temperature prediction models mainly include numerical models and statistics-based models. They are often simple in structure and cannot better capture the nonlinear characteristics of data. The effect of the emerging deep learning models often depends on the selection of super parameters, and it is difficult to obtain the ideal accuracy. This paper focuses on the deep learning model, fully mining and learning the spatial and temporal characteristics and distribution rules of ocean temperature, extracting spatial features for missing data completion, and extracting temporal features for time series prediction.

From the perspective of incomplete data representation, aiming at the problem that some longitude and latitude temperature data are missing in the ocean Argo data set, this paper combines the Bi-Directional Long Short-Term Memory layer with the GRU (Gated Recurrent Unit) layer, on the basis that the double-layer GRU model can well

capture the spatial trend of ocean temperature, and synthesizes the ability of BiLSTM to learn the spatial correlation of ocean temperature, this paper proposes a method to complete missing temperature data based on multi-layer BiLSTM-GRU model, constructs the GRU-BiLSTM-GRU model, selects the optimal neurons, and obtains the ideal regression effect. Its accuracy is better than the baseline model LSTM, BiLSTM and GRU. However, the selection of hyperparameters often has a great impact on the accuracy. The existing models often use the strategy of manual selection of hyperparameters, which greatly wastes time and cost. Therefore, this paper proposes an automatic optimization method based on Bayesian optimization (BO) for super parameters, which optimizes multi-layer BiLSTM-GRU, LSTM, BiLSTM and GRU, and constructs BO-BiLSTM-GRU, BO-LSTM, BO-BiLSTM and BO-GRU models respectively, saving the cost of selecting super parameters and improving the accuracy of the model. The experiment shows that the optimized model has a higher accuracy improvement than the optimized model, and the BO-BiLSTM-GRU model proposed in this paper is superior to BO-LSTM, BO-BiLSTM, BO-GRU, ConvLSTM, ConvGRU and M-ConvLSTM.

On the basis of complete data representation, starting from the incomplete representation of time series characteristics, the regular characteristics of ocean temperature in time distribution are extracted and analyzed. This paper proposes a temperature prediction method based on the construction of time series characteristics and one-dimensional convolution neural network (1D-CNN). First, the continuous time series of ocean temperature is constructed, and the time series characteristic coefficients are selected on eight different depth samples. Then, 1D-CNN is used to decompose and extract temporal features, and predict the temperature. Finally, the prediction accuracy is compared with the baseline model. The experiment shows that 1D-CNN is better than the model BiLSTM, LSTM, GRU, BP and KNN in 8 different depth sample data, including 6 depth samples. However, a single 1D-CNN model often only has a good learning and prediction effect in some aspects. Therefore, this paper combines 1D-CNN with the integrated learning Adaboost model and proposes a 1D-CNN-Adaboost model. This method trains 1D-CNN as a weak learner, and finally forms a strong learner

through multiple weak learners, making the model perform well in the overall effect, improving the learning ability and prediction effect of 1D-CNN. The experiment shows that the 1D-CNN-Adaboost model is superior to the 1D-CNN model in eight sample depths, which verifies that the Adaboost model is effective in improving the prediction accuracy of the 1D-CNN model.

Key words:

Ocean temperature, Multi-layer BiLSTM-GRU, Bayesian optimization, 1D-CNN, Adaboost

目 录

第 1 章 绪论

1.1 研究背景与意义

在我们赖以生存的地球上，海洋约占 70.8%，在我国，海洋面积约 473 万平方千米，对我国经济、政治、人文科学以及军事均有着不可或缺的作用。海洋与人们的生活息息相关，沿海地区孕育了我国将近 40%的人口，同时其凭借便利的水路交通和发达的进出口贸易，为我国创造了大量的经济价值，是我国市场经济的命脉。海洋中蕴含着丰富的矿产资源，如石油、天然气及其他金属矿物等[1]，同时，海洋也是我国战略安全的天然屏障[2]。因此，探索海洋，开发海洋，研究海洋对我国乃至世界各国均有着深远的经济价值和战略价值。

海洋环境既有巨大的时间和空间差异性，又有十分复杂的周期性和非周期性变化，存在极强的时空不确定性，具有混沌特征。在海洋环境中，海洋温度是极其重要的变量，对声速、声波的传播、声呐设备的传播距离均有显著影响[3]。根据声速随温度的变化，海洋可以在垂直方向被划分为表层（<100m）、温跃层（100~600m）和深海等温层（>600m）[4]。声速在表层随深度的增大而增大，在温跃层随深度的增大而急剧减小，在等温层中随深度逐渐增大[5]。随着深度的变化，声波会发生折射和反射，在一些特定的位置，如表层与温跃层交界处，甚至会发生声波的全反射，使得声波无法穿透温跃层，进而形成声影区，从而影响水声网络的连通性以及通信性能[6]。因此，海洋温度的精准预测对水声网络的路由以及媒体访问控制的设计有很大的帮助，对水声网络的研究具有重要意义。

然而，海洋温度的精准预测往往受限于较为匮乏的海洋观测数据。1998 年，美国的一些科学家推出 Argo 全球海洋观测计划，旨在构建全球范围的海洋观测网，从而能够更加精准快速的收集全球海洋内部温盐度剖面信息[7]。目前，海洋 Argo 数据已经成为天气预报以及海洋时空分布规律性研究的主要数据来源[8]。但是，Argo 浮标在海洋中并不能均匀分布，同时也存在数据传输错误、传感器故障等状况，这使得数据缺失现象较为严重，缺失数据的存在制约了海洋科学的发展，直接影响了海洋实时监测，同时对海洋数据的后续分析和规律性研究有着较大影响[9]。因此，寻找海洋缺失数据的补全方法以及提高其精度是目前亟待解决的问题，同时也是海洋时空规律性探索的支撑。

目前，已经有很多方法应用于海洋温度数据补全，这些方法主要被划分为三类，分别为基于时间策略、空间策略、时空策略的方法。基于时间策略的补全方法仅可用于补全某观测点处的数据缺失情况，由于海洋环境的恶劣，数据缺失往往表现为整块缺失的情况，时间策略补全并不适用；时空策略依赖于高维度的时空数据，当空间比较时大，数据量会非常大，导致模型训练和学习非常复杂；由于本文所研究的空间数据量较大，本文将重点从空间策略补全切入，现有的空间补全策略主要包括传统的数学模型，如空间插值，以及数据驱动的深度学习模型，传统的模型当数据量很大时往往难以计算、有些模型需要数据满足某些特定分布，深度学习模型表现出较好的学习能力，然而简单网络结构的深度学习模型难以取得好的精度，同时其超参数的择优对模型精度有着较大影响，现有的超参数选择方法以人工选择和网格搜索为主，浪费了大量的时间和资源。因此，本文从深度学习模型切入，对模型的网络结构进一步优化，同时，为影响模型精度的超参数提供了良好的择优方法。

海洋温度的时序特征研究是海洋学最基本的研究内容之一，与海洋的探索和开发关系密切，精准的预测海洋温度的时空分布对很多领域均有重要意义，如天气预报、海洋渔业以及海洋军事作战等。然而复杂的海洋环境给海洋温度的精准预测带来了挑战。很多方法已经被用于海洋温度预测，这些方法主要被划分为两类，其中，一类是基于数学或数理统计的数值方法，这一类方法主要是采用多项式、微分方程等数学推理来对海洋温度的变化规律进行拟合，这类方法比较简单，而且海洋变化的非线性特征很难用一个标准的数学方程表示，因此他们往往难以取得理想的拟合效果。另一类是基于数据驱动的方法，随着简单神经网络到深度学习的逐步发展，这一类方法在拟合海洋温度变化的非线性规律方面效果理想。然而单一的深度学习模型往往在某些方面学习效果较好，综合各方面均具有理想效果很难。因此，本文从数据驱动的深度学习方法切入，对海洋温度的时间分布特征进行提取，并对深度学习方法集成以进一步完善优化模型。

准确预测海洋温度有助于海洋开发和利用以及水声网络的研究，本文旨在利用海洋 Argo 数据集对海洋的历史状态及规律进行探索和研究，通过深度学习等方法对海洋温度缺失数据进行补全，同时对海洋温度时间序列进行高精度的预测。

1.2 国内外研究现状

1.2.1 缺失数据补全

数据缺失补全问题根据观测数据、补全需求和数据特点的不同主要分为三种策略，分别为时间补全策略、空间补全策略、时空补全策略。

时间补全策略主要依赖于时间序列数据，根据数据的时间特征来估计缺失数据。其中，插值方法依赖于时间前后数据信息，包括线性插值、样条插值[10]。差分整合移动平均自回归模型（ARIMA）[11]与季节性差分自回归积分滑动平均模型（SARIMA）[12]将时间序列补全问题转化为时间序列预测问题，然而这两种方法仅考虑了缺失值前的信息，并未考虑缺失值后的信息。以上基于数学和统计的方法简单且易于实现，但非线性拟合能力较差。机器学习方法在数据补全中应用广泛，Phan 等人利用基于自适应特征的动态时间规整（DTW）算法寻找相似子序列，并将其下一个子序列用于填充海洋单变量时间序列缺失值[13]。刘等人利用多层 LSTM-GRU 联合模型补全时间序列数据缺失值，并对该模型进行了改进，将缺失数据的前序时间序列正向预测，后序时间序列反向预测，并将预测值结合，取得了理想的效果[14]。

空间补全策略主要依赖于空间数据，根据数据的空间特征来估计缺失数据。空间插值方法同样可以依赖空间位置附近数据信息来对缺失数据补全，例如反距离加权插值[15]、克里金插值[16]等。然而这些方法往往需要已知数据服从特定的分布，对不规则的海洋数据往往难以取得理想的效果，因此，一系列机器学习算法被用于空间补全。张等人利用 SVR 也海洋温度网格数据中缺失数据进行估计，并利用网格搜索对最优超参数进行选取[8]。Chen 等人提出了一种基于节点聚类和遗传编程（NCGP）的新型数据估计方法来估计 WSN 缺失的传感器数据，利用遗传编程来挖掘同一聚类中节点之间的关系[17]。阳等人使用层次聚类对不同特征的数据进行类别划分，然后使用基于网格搜索选取最优超参数的随机森林模型进行海洋温度数据高分辨率构建，从而补全网格数据集，同时为了解决海洋温度数据的不平衡性与稀疏性，将变分自编码器与深度学习回归网络结合，应用到数据补全中，成功缓解了样本的不平衡性问题[18]。由于海洋环境的恶劣，数据采集过程中某些区域的数据大量缺失情况并不少见，Alamoodi 等人提出了一种大

量缺失值补全方法，该方法利用预处理措施，将数据集分解为连续的小部分，然后使用多标准决策分析来选择代表整个破碎数据集的一部分数据，并通过使用不同的机器学习技术进行不同百分比的人工缺失制造过程和插补来扩展缺失数据集，取得了理想的效果[19]。

时空补全策略将时间和空间信息综合考虑，根据时空特征来估计缺失数据。谭等人综合考虑了时间和空间特性，提出一种基于矩阵补全的多视图学习（MC-MVL）方法来填充浮标监测数据中的缺失值，提高了模型的插补能力[20]。关等人提出一种海洋遥感缺失数据智能补全方法，利用图注意力网络（GAT）动态获取海洋遥感数据的时间空间信息，实现了遥感信息的有效聚合，提高了补全的准确性[21]。

时间补全策略主要适用于单个观测点在时间维度上的数据缺失问题，无法解决数据在空间上块状缺失问题，时空补全策略过度依赖于时空数据，当空间比较大时数据量会非常大导致模型训练过于复杂。空间补全策略更适用于空间比较大，存在块状缺失的问题，现有的研究表明深度学习方法在空间补全问题上具有更好的普适性。具有单一网络结构的深度学习方法精度并不高，并且其精度往往依赖于其超参数的设定。然而，现有的方法往往采用手动选择超参数，浪费了人力物力成本。因此，本文从深度学习网络结构及其超参数选择方法入手进行模型优化。

1.2.2 海洋温度预测

海洋温度预测问题可以表述为回归问题。目前常用的回归方法有线性回归、自回归、岭回归、logistic 回归和支持向量机（SVR）回归。Menon 等人使用多元线性回归（MLR）预测温度[22]。Gou 等人应用 KNN 回归算法预测海洋温度和盐度[23]。Jiang 等人利用 SVR 进行高分辨率温度和盐度模型分析，证明了 RBF 核函数适用于复杂的海洋数据[24]。然后，Quan 等人对 SVR 进行改进，利用遗传算法(GA)来优化超参数，提出了一种改进的支持向量机（MGASVR）来预测水温[25]。

但这些传统的回归方法结构简单，非线性拟合能力有限，在准确性方面往往不够。近年来，神经网络和深度学习逐渐成为时间序列预测的主流方法。常见的神经网络主要包括人工神经网络（ANN）、卷积神经网络（CNN）和递归神经网

络（RNN）。

人工神经网络（ANN）包括前馈神经网络（FFNN）和 BP 神经网络。Wei 等人应用人工神经网络模型（ANN）预测海洋温度[26]。Graf 等人将离散小波变换（WT）与人工神经网络（ANN）相结合，并将其应用于水温预测[27]。Ting 等人采用 BP 神经网络预测温度[28]。进而，Xu 等人对 BP 神经网络进行了改进，将动态和静态预测方法结合起来以预测温度[29]。Zhu 等人使用不同版本的前馈神经网络（FFNN）来预测河流的水温[30]。简单人工神经网络是比较基础的神经网络，有时很难获得理想的精度。

卷积神经网络（CNN）在海洋温度预测中也有一些应用。Zoo 等人提出了立体空间和时间四维卷积模型（SST-4D-CNN），该模型充分考虑了时间序列和海洋空间关系的双重特征，提高了预测精度[31]。然而，CNN 被广泛应用于图像处理和图像识别等领域，在海洋温度预测中应用较少。

递归神经网络（RNN）在海洋温度预测中最为广泛，并且随着研究的深入，一系列变种模型应运而生，其中，长短期记忆神经网络（LSTM）被广泛应用于时间序列预测。Zhang 等人使用长短期记忆神经网络（LSTM）预测海面温度（SST）[32]。Liu 等人提出了一种基于时间依赖的 LSTM（TD-LSTM）来预测海面温度[33]。肖等人提出了一种机器学习方法，该方法将 LSTM 与提升决策树算法预测值取均值，以预测中短期每日海面温度，提高了预测精度[34]。Stephen 使用遗传算法（GA）优化 LSTM，然后应用 GALSTM 预测水温[35]。

LSTM 在应用过程中产生了一些变体，并取得了良好的效果，例如 BiLSTM 和 GRU。Jiang 等人提出了一种双向 LSTM（BiLSTM）框架来预测和分析海洋温度和盐度[36]。Jiang 等人使用精英保存遗传算法（EGA）来提高 BiLSTM 的预测精度，并提出 EGA-BiLSTM 来预测温度[37]。Xie 等人提出了基于 SST 编码的门环单元（GRU）编解码器和动态冲击链路（DIL），以及考虑静态和动态影响的 GRU 编解码器（GED），以预测海面温度[38]。覃提出了基于时序分解和 GRU(SGRU)的海洋温度预测模型，提高了连续的预测精度，降低了累计误差[39]。

一些研究者将卷积层与长短期记忆（LSTM）神经网络结合起来，以提高预测精度。Yang 等人将完全连接层 LSTM（CF-LSTM）与卷积层相结合，并提出了预测海洋温度的 CFCC-LSTM 模型[40]。肖等人提出了一种基于时空信息的深

度学习模型，该模型将卷积层与 LSTM 层结合构成卷积长短记忆（ConvLSTM）网络，并以端到端的方式进行训练。该模型可以捕捉温度的时空相关性[41]。张等人将卷积神经网络和循环神经网络扩展算法结合起来，提出了 ConvGRU 用于海表温度预测[42]。Zhang 等人将多层 CNN 和 LSTM 堆叠，提出了一种多层卷积长短记忆（M-convLSTM）模型来预测三维海洋温度[43]。

在已有的研究中，单一的深度学习方法往往难以取得理想的预测效果，本文从深度学习方法集成的角度入手，进一步对海洋温度预测精度进行优化。

1.3 主要研究内容及章节安排

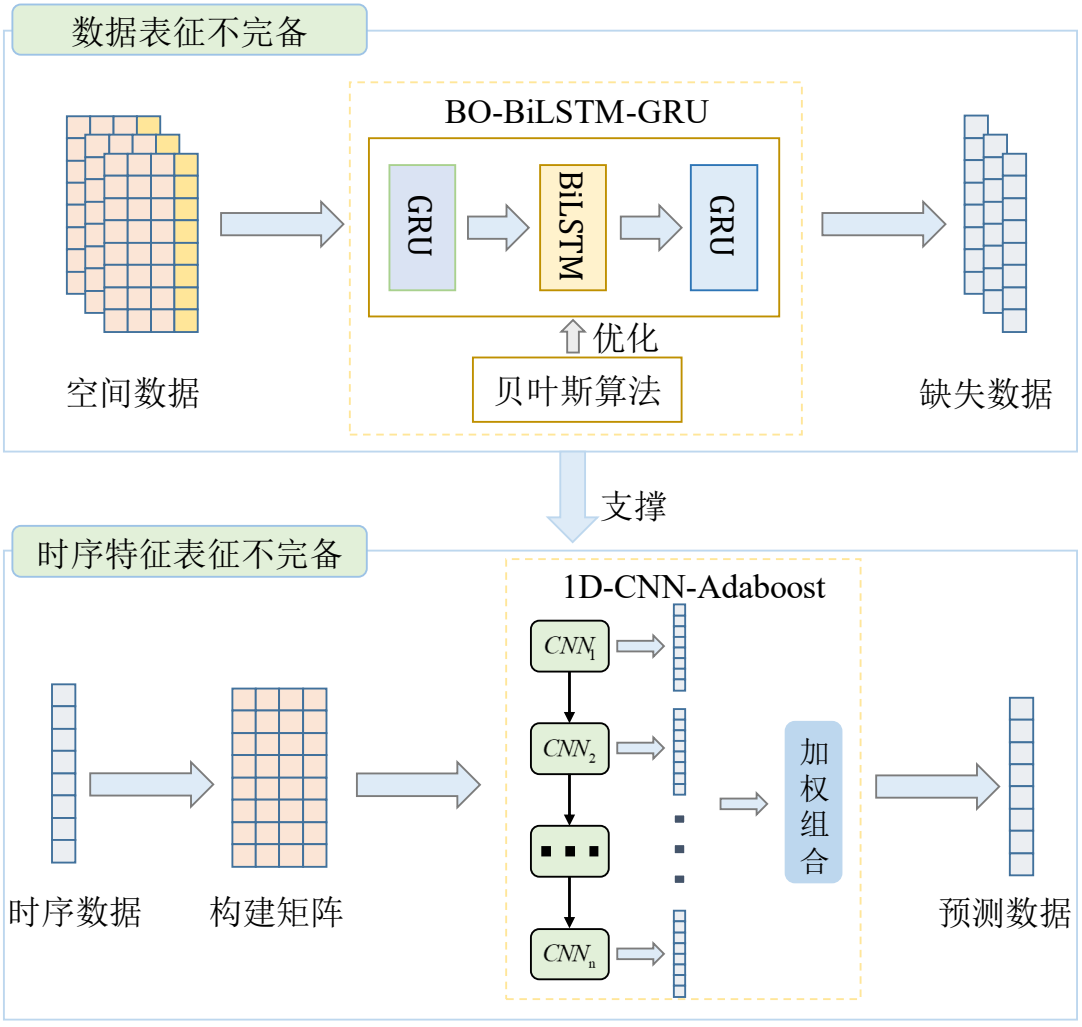

图 1.1 研究整体框架

本文针对海洋观测数据的数据表征不完备和时序特征表征不完备两个问题，从空间和时间的分布规律入手。在空间分布规律性研究上，主要研究内容是对

Argo 数据缺失温度进行补全，在时间分布规律性研究上，对海洋时序温度进行特征构建和预测。如图 1.1 所示，第一部分，首先针对海洋 Argo 原始数据进行预处理，即对缺失值和异常值进行剔除处理，然后提出了多层 BiLSTM-GRU 模型，同时利用贝叶斯算法对超参数进行调优，构建了基于 BO-BiLSTM-GRU 模型的缺失温度数据补全方法。第二部分，首先提出了基于时序特征构建和 1D-CNN 的温度预测模型，该方法针对某观测点处的海洋温度时序数据进行时间序列构建，然后将经时序特征构建的矩阵利用 1D-CNN 进行时序预测。再利用 Adaboost 模型对 1D-CNN 进行集成，提出了基于 1D-CNN-Adaboost 模型的时序温度预测模型。

本论文章节安排如下：

第一章：绪论。描述了海洋温度规律研究、缺失数据补全以及时序预测的重要意义，调研并总结了现有的缺失数据补全方法以及时序温度预测方法的国内外研究现状，阐述了本文的研究目标和研究内容，并组织出本文的整体结构。

第二章：海洋温度缺失数据补全和时序预测相关理论和方法。分别介绍了三种常用的数据补全方法和时间序列预测方法，数据补全方法分别为插值、ARIMA 以及 SVR，时间序列预测方法分别为 KNN 算法、BP 神经网络、循环神经网络及其变种模型。

第三章：基于 BO-BiLSTM-GRU 模型的温度数据补全方法。针对目前海洋 Argo 观测数据中存在的某些经纬度温度数据缺失或者有异常值的问题，本章将缺失数据补全问题表述为多元回归问题，将数据集中的经度、纬度和深度作为回归预测的自变量，温度作为因变量，建立{经度，纬度，深度；温度}结构模型，将 GRU 层与 BiLSTM 层相结合，构建了多层 BiLSTM-GRU 模型，然后介绍了贝叶斯算法相关理论，并利用贝叶斯算法选取模型超参数，提高了回归预测精度。最终，经过实验验证，本章节提出的多层 BiLSTM-GRU 模型优于基线模型 LSTM、BiLSTM 和 GRU，同时，经过贝叶斯算法选取超参数后的 BO-BiLSTM-GRU 模型优于 BO-LSTM、BO-BiLSTM、BO-GRU、ConvLSTM、ConvGRU 和 M-ConvLSTM。

第四章：基于 1D-CNN-Adaboost 模型的海洋时序温度预测。首先介绍了时间序列与 1D-CNN 的基本理论，同时利用 BOA_Argo 数据集构建了不同深度的

海洋温度时间序列数据，提出了基于时序特征构建和 1D-CNN 的温度预测模型，通过实验证明，该模型优于基线模型 KNN、BP、LSTM、GRU 和 BiLSTM。然后介绍了集成学习相关理论，并利用集成学习中的 Adaboost 模型对 1D-CNN 进行优化，实验证明，1D-CNN-Adaboost 模型优于 1D-CNN，提升了预测精度。

第五章：总结与展望。对本文提出的温度数据补全方法以及温度预测方法进行了总结，概括了本文的创新点，同时针对本文的未来工作进行展望。

第 2 章 海洋温度补全与预测相关理论

2.1 缺失数据补全方法

海洋缺失数据补全策略主要包括三类，分别为时间策略、空间策略以及时空策略，其方法主要包括两类，分别为基于数值模型的以及基于数据驱动的，常用数值模型包括插值、差分整合移动平均自回归模型（ARIMA），数据驱动的模型包含 K 近邻（KNN）、支持向量机（SVR）、随机森林等。

2.1.1 基于插值的数据补全

插值方法主要被分为两类，一类是基于时间序列信息的，如线性插值、三次样条插值[10]；一类是基于空间相邻信息的，如反距离加权插值[15]、克里金插值[16]等。

反距离加权插值是基于相近原则的插值方法，即与待补全的点距离相近的点比距离更远的点具有更多的相关性和相似性。每个样本点对待补全样本点均具有一定的影响，距离越近的点权重越大，距离越远的点权重越小，甚至可以忽略。反距离插值对缺失值的估计采用公式[15]：

$$Z=\frac{\sum_{i=1}^{N} z_i \cdot d_i^{-n}}{\sum_{i=1}^{N} d_i^{-n}} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.1)$$

其中 Z 为当前待补全节点的估计值，z_i 为已知点的值，N 为已知点的总数量，n 为基于距离的系数。

克里金插值是基于给定区域内所有相近点的加权均值，又被称为空间局部插值，其要求空间中各个点服从相同的分布，即具有相同的期望和方差。克里金插值的基本原理为利用局部空间内的变量和变异函数来对待补全节点进行最优的、线性无偏的估计，其中最优表示误差的平方和达到最小，无偏表示偏差的数学期望为 0。克里金插值对缺失值的估计采用公式[16]：

$$Z(x_0)=\sum_{i=1}^{n} \lambda_i Z\left(x_i\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2.2)$$

其中 λ_i 为待定权重系数，根据其无偏、方差最小的条件可以得到该方程组：

$$\begin{cases} \sum_{i=1}^{n} \lambda_i C\left(x_i, x_j\right) + \mu = C\left(x_0, x_j\right) (j = 1, 2, \ldots\ldots, \mathrm{n}) \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \quad \cdots\cdots\cdots\cdots \text{（2.3）}$$

其中 μ 为拉格朗日常数，$C\left(x_i, x_j\right)$ 是与原点在 x_i，端点在 x_j 的向量相对应的变异函数值，n 为样本数量。

克里金插值等空间数据补全方法需要数据具有一定的规则和分布，然而已知的海洋数据分布及其缺失数据均具有不规则性，因此简单的空间插值方法往往无法取得理想效果。

2.1.2 基于 ARIMA 的数据补全

时间维度上的海洋数据缺失问题可以表述为时间序列预测问题。差分整合移动平均自回归模型[11]（Autoregressive Integrated Moving Average model，ARIMA）是基于时间策略数据补全的常用算法之一。它使用自回归和移动平均，并结合一个差异顺序来去除趋势或季节性，其基本原理为将随时间变化的非平稳序列转化为平稳序列。

$$y_t^{'} = c + \alpha_1 y_{t-1}^{'} + \cdots + \alpha_p y_{t-p}^{'} + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} + \varepsilon_t \quad \cdots\cdots\cdots \text{（2.4）}$$

其中 $y_t^{'}$ 代表差分级数，它是通过结合差分滞后值与差分滞后误差来计算的。

ARIMA 模型包含 3 个重要参数（p，d，q），参数 p 在模型中代表滞后时间，参数 d 表示为消除趋势和季节性进行的差分转换次数，参数 q 表示误差分量的滞后，误差部分表示在时间序列过程中无法用趋势或季节性进行解释的部分。将三个参数代入方程（2.4）可以得到[11]：

$$\begin{array}{ccccc} \mathrm{AR}(p) & & d \text{ differences} & & \mathrm{MA}(q) \\ \downarrow & & \downarrow & & \downarrow \\ \left(1 - \alpha_1 D - \cdots - \alpha_p D^p\right) & & (1-D)^d y_t & = c + & \left(1 + \beta_1 D + \cdots + \beta_q D^q\right) \varepsilon_t \end{array} \quad \text{（2.5）}$$

其中，L 为滞后算子，AR（Autoregressive）为自回归模块，AR（p）模块代表当前数据与前 p 个数据具有较强的关联性；MA（Moving Average）为移动平均模块，其与自回归模块构成自回归移动平均模块 ARMA（Autoregressive Moving

Average）模块，最终与 d 阶差分模块结合构成 ARIMA 模型。

ARIMA 建模本质上是根据时间序列数据具有自相关或者偏自相关的特性，从中挖掘趋势、随机变化、周期成分、循环模式和序列相关性等信息，进而更加恰当的对时间序列数据的结构进行模拟，该模型具有对数据特征的探索性和对数据的结构及趋势拟合的灵活性，可以很容易的对序列数据的未来值进行一定程度的准确预测。

ARIMA 模型建模步骤为，先获取时间序列数据，然后检验数据平稳性以及进行白噪声检验，平稳性的检验主要通过时序图或者相关图，对于非平稳的时间序列，若存在某种趋势（增长或下降），则需要利用差分的方法处理直到平稳。时序信息随着差分处理的次数增多，平稳信息将提取的更加充分，然而每次差分均会造成信息损失，因此差分阶数通常不超过 2。然后将预处理完毕的数据进行模型识别，同时需要利用 BIC 准则法进行模型定阶，最终验证模型的拟合效果，若不理想，则需要重新拟合模型。

ARIMA 模型的输入序列仅能为单变量序列，仅能用于海洋某观测点处的数据补全，无法对空间块状缺失的数据进行补全。

2.1.3 基于 SVR 的数据补全

支持向量机（SVM）起源于广义肖像算法，是一种分类模型，其利用结构风险最小化的思想对非线性数据进行拟合，具有良好的泛化能力，目前比较常用的支持向量模型主要分为两类，分别为支持向量回归和支持向量分类。而空间数据补全问题往往被表述为回归问题，因此，本文重点讲解支持向量回归模型[8]（support vector regression，SVR）。

SVR 是基于空间策略数据补全比较常用的机器学习方法，是一种有监督的学习的算法，其目标为在给定的 n 维空间中寻找一个最理想的拟合线，该拟合线也被成为超平面，距离超平面很近的点被称为支持向量。SVR 在处理高空间维度上具有明显优势和良好的泛化能力。

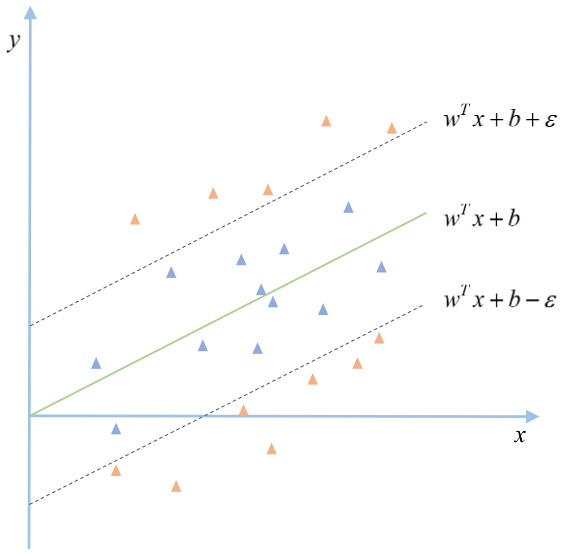

图 2.1 支持向量回归原理示意图

通过图 2.1 可以看出，针对给定的一个 n 维样本 (x_i, y_i)，$i=\{1,2,...,n\}$，目的在于拟合一个最优模型函数 $f(x)=w^T x+b$，其中 w 为权重矩阵，b 为偏差向量。在超平面两侧均建立一个宽度为 ε 的隔离带，构建两个边界函数 $f_1(x)=w^T x+b+\varepsilon$，$f_2(x)=w^T x+b-\varepsilon$，在两个边界内的区域被称为 ε 管，当训练值与真实值的差小于 ε 时，即落入 ε 管内，该值被定义为预测正确，此时不计算误差损失，当训练值与真实值的差大于 ε 时，即落入 ε 管外，该值被定义为预测错误，此时计算误差损失。因此可以得到损失计算公式

$$\xi_i=\max\left(0,\left|w^T\cdot x_i+b-y_i\right|-\varepsilon\right),\xi_i\geq 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (2.6)$$

综上，结合结构风险最小化思想，支持向量回归可以表述为以下优化问题

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{M}l_\varepsilon\left(f\left(x_i\right)-y_i\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (2.7)$$

其中 $l_\varepsilon(z)=\begin{cases}0, & \text{if } |z|\leq\varepsilon\\ |z|-\varepsilon, & \text{otherwise}\end{cases}$，$z=y_i-f(x_i)$，$w=\sum_{i=1}^{M}\left(\beta_i-\beta_i^*\right)\phi\left(x_i\right)$，C 是惩罚因子其目的是平衡训练误差和训练数据与超平面空间之间最大距离，式（2.7）的第一项用于对大权重进行惩罚来保证回归函数平稳性，第二项使用 ε 不敏感损失函数确定置信度与经验风险之间的平衡。线性 SVR 表述为

$$y=\sum_{i=1}^{M}\left(\beta_i-\beta_i^*\right)\cdot\left\langle x_i,x\right\rangle+b \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (2.8)$$

非线性 SVR 表述为

$$y=\sum_{i=1}^{M}\left(\beta_i-\beta_i^*\right)\cdot\left\langle\varphi\left(x_i\right)\cdot\varphi(x)\right\rangle+b \cdots\cdots\cdots\cdots\cdots\cdots \text{（2.9）}$$

通过引入核函数，SVR 统一表述为

$$y=\sum_{i=1}^{M}\left(\beta_i-\beta_i^*\right)\cdot K\left(x_i,x\right)+b \cdots\cdots\cdots\cdots\cdots\cdots \text{（2.10）}$$

其中 β_i 与 β_i^* 为拉格朗日乘子，$K\left(x_i,x\right)$ 为核函数。SVR 利用核函数对其进行线性分离，将数据转化为高维度特征空间。核函数遵循 Mercer 定理，因此，任意半正定对称函数均可作为核函数，核函数包括常用的核函数包括多项式核函数和径向基核函数

$$K\left(x_i,x_j\right)=\left(x_i\cdot x_j\right)^d \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.11)$$

$$K\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.12)$$

核函数的引入使得 SVR 对低维以及高维数据均具有良好适用性，有效的避免了维度灾难，目前 RBF 核函数最为常用。

在海洋温度数据补全的应用中，先对数据进行归一化，然后将数据输入模型，经过训练不断调整 w 和 b 来获得理想的拟合函数，利用训练好的拟合函数及给定的输入值来估计温度的缺失值。

2.2 海洋温度预测方法

目前，机器学习在海洋温度预测中应用广泛，主要包括 K 近邻（KNN），支持向量机（SVR）以及人工神经网络等。目前比较主流的人工神经网络主要有 BP 神经网络、卷积神经网络（CNN）以及循环神经网络（RNN）等。随着研究的深入，他们产生了一些变种模型，如 LSTM、BiLSTM 以及 GRU 等。

2.2.1 基于 KNN 的时序预测

K-近邻(K-Nearest Neighbors，KNN)是一种基本的回归或分类方法[23]，它并

没有显式的学习过程，而是基于实例的学习。KNN 算法将距离作为度量两个样本是否相似的依据，常用的距离计算方法包括欧氏距离、曼哈顿距离等。

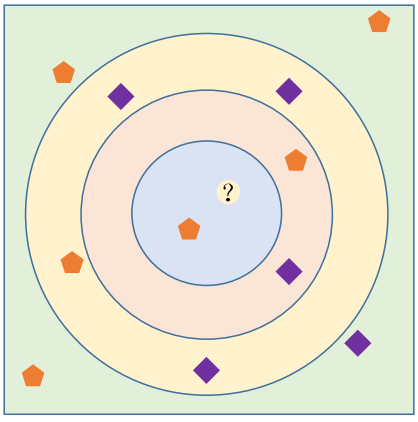

图 2.2 KNN 算法示意图

针对分类问题，如图 2.2 所示，整张图被分为 4 个区域，从里向外分别记为“蓝色区域”，“粉色区域”，“黄色区域”，“绿色区域”。其中样本被分为两类，记为“橙色五边形”和“紫色菱形”。当 K 值取 1 时(蓝色区域)，中心黄色圆形待测样本点将被分为“橙色五边形”类别，当 K 取 3 时(粉色区域及以内)，中心黄色圆形待测样本点将被分为“橙色五边形”类别，当 K 取 7 时(黄色区域及以内)，中心黄色圆形待测样本点将被分为“紫色菱形”类别，当 K 取 11 时(绿色区域及以内)，中心黄色圆形待测样本点将被分为“橙色五边形”类别。针对回归预测问题，待测样本点给会将距离最近的 K 个值的均值作为待测样本点的值。综上可以看出，K 值的选择对分类以及预测结果有着直接的影响，因此无论在预测问题还是回归预测问题上，K 值均扮演着重要角色。

在海洋温度预测中，KNN 模型工作原理如下，首先先对样本数据进行构建并输入，然后计算这几个样本的欧氏距离，对距离进行排序，然后选择最近的 K 个临近点，将最近的 K 个临近点取均值后获得预测值。

2.2.2 基于 BP 神经网络的时序预测

1986 年，Rumelhart 和 McClelland 等人提出了 BP 神经网络[44]，它是一种前馈神经网络，其原理为利用误差反向传播进行训练，是目前应用最广泛的神经网络之一[45]。

BP 神经网络的架构分为 3 层，分别为输入层、隐藏层和输出层，如图 2.3 所

示：

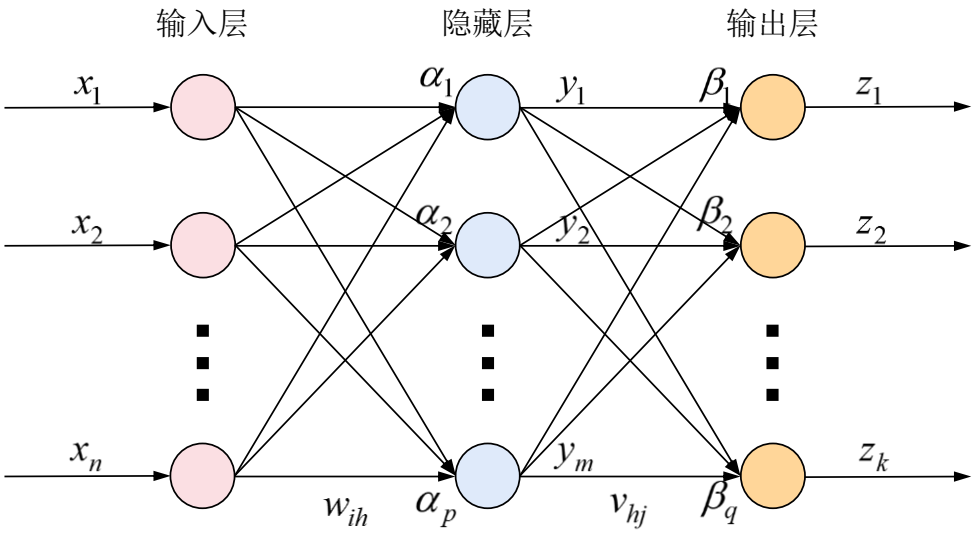

图 2.3 BP 神经网络示意图

其中 $x_1, x_2, ..., x_n$ 为输入， y_i 为隐藏层输出，然后作为输出层的输入， z_i 为输出， w_{ih} 为输入层与隐藏层连接权重， v_{hj} 为隐藏层与输出层连接权重， α_h 为隐藏层第 h 个神经元的输入， β_j 为输出层第 j 个神经元的输入。

$$\alpha_h = \sum_{i=1}^{n} w_{ih} * x_i \cdots\cdots (2.13)$$

$$\beta_j = \sum_{h=1}^{m} v_{hj} * y_h \cdots\cdots (2.14)$$

BP 神经网络的详细工作原理是模拟大脑神经元的信息传递模式，对输入的信息进行非线性变换处理，每一轮训练均会给定输出，根据本次输出与真实数据计算拟合误差，然后该误差将会反向传播，通过隐藏层传播到输入层，最终分别传播到每个神经元上，通过不断的修正神经元之间的连接权重以及阈值，使误差沿着梯度方向下降，最终达到一个拟合度较为理想的效果。

BP 神经网络常用的 3 种激活函数：

Sigmoid 函数是一种 S 型函数，具有连续、光滑以及严格单调等特点，是一种较好的神经元阈值函数。然而 S 型函数具有软饱和(左右两端导数均趋于 0)的缺点，容易在训练过程中产生梯度消失的现象，导致训练效果并不理想。

$$f(x)=\frac{1}{1+e^{x}} \quad \cdots\cdots(2.15)$$

Tanh 函数（双曲正切函数）：Tanh 函数的值在[-1,1]，以[0,0]为中心，可以将正值映射到正值，负值映射到负值，与 sigmiod 函数仅能映射到正值比有一定优势。但 Tanh 函数也有梯度消失现象，这点与 sigmiod 类似。

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \cdots\cdots(2.16)$$

ReLU 函数，有效的弥补了上述两个激活函数存在梯度消失问题的不足，目前在大多数情况下使用。ReLU 函数当输入为正值时，导数恒为 1，因此具有较快的收敛速度，且易于计算，计算速度快。但当输入为负值时，存在梯度失效问题。

$$f(x)=\max(0,x) \quad \cdots\cdots(2.17)$$

在海洋温度预测中，BP 神经网络的详细工作原理如下，先将海洋温度时序数据进行数据集构建，然后将训练集通过输入层输入 BP 神经网络，通过不断迭代逐步修正输入层与隐藏层和隐藏层与输出层之间的权重，将训练完毕的网络用于预测。

2.2.3 基于循环神经网络的时序预测

2.2.3.1 循环神经网络原理介绍

以 BP 神经网络为代表的一些前馈神经网络模型在学习和训练过程中，其输出结果仅由当前时刻的输入决定，过去一段时间的输入无法影响其训练结果。然而，在一些实际的场景中，网络模型的当前输出并不是仅由当前时刻的输入决定，其与前一时间段的输入也有紧密的联系，基于此，循环神经网络(Recurrent Natural Network，RNN)应运而生。RNN 在处理序列数据上具有较强的适用性和相对明显的优势，已经成功应用在众多时间序列数据处理领域中，例如自然语言生成、机器翻译以及语音识别等。

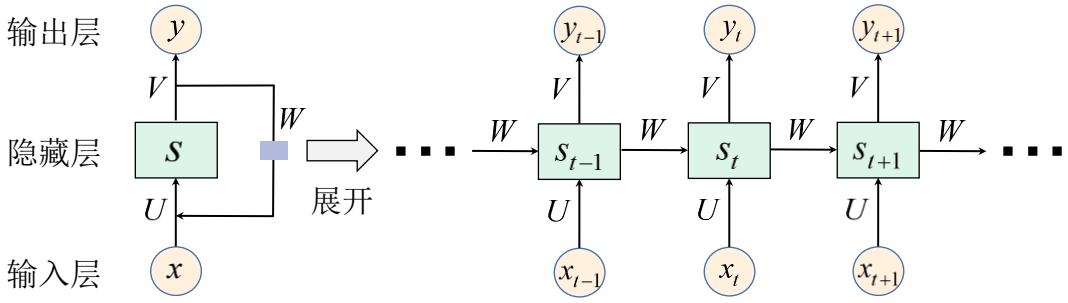

图 2.4 RNN 展开结构

RNN 沿时间展开结构如图 2.4 所示，该图展示了一个典型的 Elman 循环神经网络，其中U 是输入层到隐藏层的权重矩阵，s 为状态，W 为状态到隐藏层的权重矩阵，V 为隐藏层到输出层的权重矩阵。由图 2.4 可以看出，权重W，U，V 是共享参数，可以有效的降低参数数量。

RNN 除了可以沿时间步展开外，也可以增加深度，形成深度循环神经网络，如图 2.5 所示。

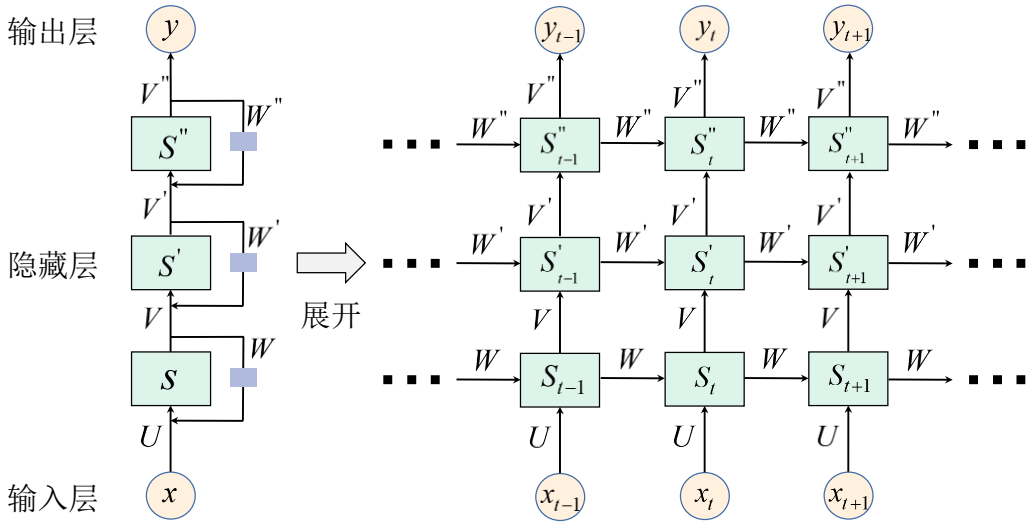

图 2.5 深度 RNN 展开结构

然而，RNN 存在梯度消失问题以及梯度爆炸问题，并且随着 RNN 的深度增加，网络将难以训练，无法实现长时间记忆。

在海洋温度预测中，循环神经网络工作过程如下，海洋温度数据需要先进行构建，获得构建完毕的输入矩阵，然后将其输入循环神经网络，循环神经网络将输入依次在不同时刻输入，RNN 神经元逐渐获得更多的信息并将其记忆，同时将下一时刻的预测值输出。

2.2.3.2 长短时记忆神经网络原理介绍

1997 年，Hochreiter 和 Schmidhuber 将输入门、输出门和遗忘门添加到 RNN 神经元中，提出了长短时记忆神经网络(Long Short-Term Memory，LSTM)[46]，从而有效的解决了 RNN 中所存在的梯度消失问题以及梯度爆炸问题，LSTM 结构如图 2.6 所示。

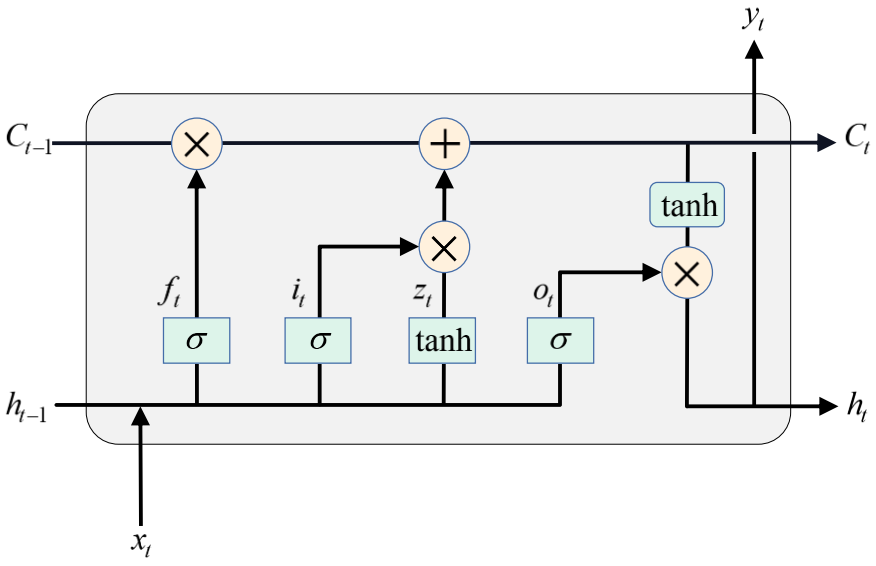

图 2.6 LSTM 单元结构

LSTM 的主要针对其存储单元进行了改进，其本质相当于状态信息转换器。工作原理如下，第一步，由遗忘门 f_t 决定哪些单元状态的信息应当被丢弃，上一时刻的状态输出 h_{t-1} 以及当前时刻的输入 x_t 共同输入到遗忘门中，遗忘门利用一个 sigmoid 激活函数 σ 来进行处理。

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.18)$$

并输出一个介于 0-1 之间的数值，1 代表信息完全保留，0 代表信息完全丢弃。其中 W_f 为权重矩阵，$[h_{t-1}, x_t]$ 表示将两向量进行连接，进而组成更长向量，b_f 为偏置项。

第二步，由输入门来决定应更新哪些状态信息，根据上一时刻的输出和当前时刻的输入，利用 tanh 函数在当前的输入中提取有效信息，然后对有效信息进行筛选。

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.19)$$

$$z_t = \tanh\left(W_z \cdot \left[h_{t-1}, x_t\right] + b_z\right) \cdots\cdots (2.20)$$

第三步，状态由C_{t-1}转换为C_t，将遗忘门和输入门所保留信息进行运算

$$C_t = f_t * C_{t-1} + i_t * z_t \cdots\cdots (2.21)$$

第四步，由输出门决定输出

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \cdots\cdots (2.22)$$

$$h_t = o_t * \tanh\left(c_t\right) \cdots\cdots (2.23)$$

2.2.3.3 双向长短时记忆神经网络原理介绍

与 LSTM 同年，Schuster 和 Paliwal 首次提出双向循环神经网络(Bi-directional Recurrent Neural Network, Bi-RNN)模型[47]，相较于 RNN，Bi-RNN 模型可以充分利用未来信息。其采用两个时序相反的循环神经网络，同时对历史和未来信息进行捕获。

双向长短时记忆神经网络(Bi-directional Long Short-Term Memory，BiLSTM)为 LSTM 的一个变种模型，由两个方向相反的 LSTM 模型组成，其结构如图 2.7。

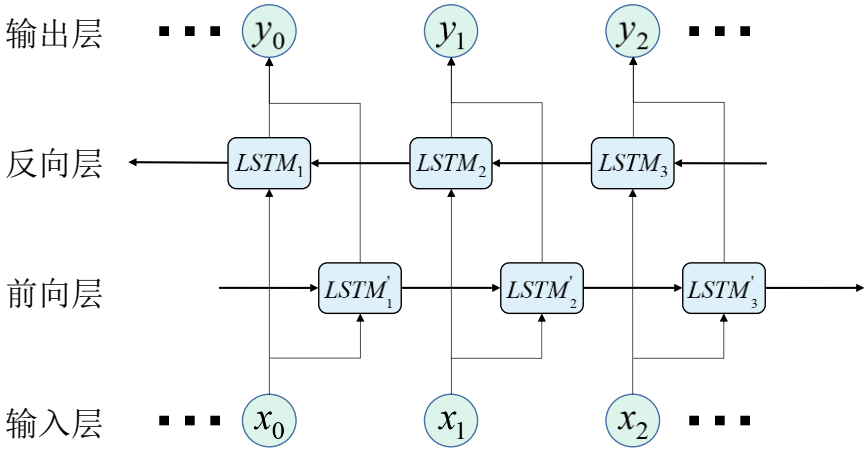

图 2.7 BiLSTM 结构图

2.2.3.4 循环门单元神经网络原理介绍

循环门单元神经网络(Gated Recurrent Unit, GRU)[48]是 LSTM 的一个变体模型，LSTM 有效的解决了长期依赖、梯度消失和梯度爆炸问题，但是其也有一些

缺点，如结构较为复杂，计算复杂度偏高等。GRU 就是基于此提出的，GRU 是 LSTM 的一个简化模型，将输入门、遗忘门和输出门这三个门合并为两个门，分别为重置门和更新门，同时将输出和单元状态这两种状态合并为一种状态，在效果相差不大的同时具有较高的计算效率，降低了内存利用率，提高了训练速度，其结构如图 2.8。

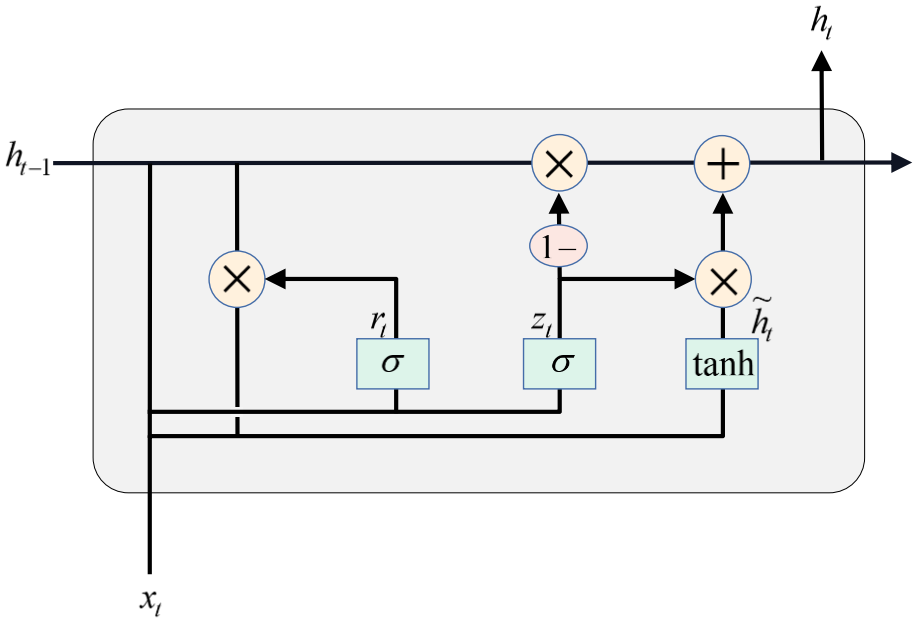

图 2.8 GRU 单元结构

其工作原理，第一步，将前一时刻的输出 h_{t-1} 与当前状态的输入 x_t 合并为新向量 $[h_{t-1}, x_t]$，输入重置门中，由 sigmoid 激活函数 σ 处理并输出为 r_t，然后重置门将上一时刻的状态信息进行筛选输出 $r_t * h_{t-1}$，其中 w_t 是权重矩阵。

$$r_t = \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.24)$$

第二步，将向量 $[h_{t-1}, x_t]$ 由更新门中的 sigmoid 激活函数处理并输出为 z_t。

$$z_t = \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.25)$$

第三步，重置门的输出 $r_t * h_{t-1}$ 与当前输入 x_t 合并为新的向量 $[r_t * h_{t-1}, x_t]$，传入 tanh 激活函数处理并输出候选集 h_t。

$$\tilde{h}_t = \tanh\left(W_{\tilde{h}} \cdot \left[r_t * h_{t-1}, x_t\right]\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.26)$$

第四步，对上一时刻 h_{t-1} 和 h_t 中的重要信息进行选择性保留，由公式 2.27 可

以获得当前层输出状态h_t。

$$h_t = (1-u_t) * h_{t-1} + u_t * \tilde{h}_t \cdots\cdots(2.27)$$

2.3 评价标准

本文使用三种不同的一致性度量来分析算法性能，分别为：平均绝对误差（MAE）、平均绝对百分比误差（MAPE）和均方根误差（RMSE）。MAE、MAPE 和 RMSE 越小，预测精度越好。平均绝对误差表示为

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|r_i - p_i| \cdots\cdots(2.28)$$

其中r_i是实际值，p_i是预测值，N是数据长度。MAE 被用作不确定性度量指标，以评估信任预测的风险。MAE 是绝对误差平均值的量度，其优点是非专业人士更容易理解。平均绝对百分比误差表示为

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|r_i - p_i|}{|r_i|}\times 100 \cdots\cdots(2.29)$$

MAPE 是基于百分比的，使得两个模型之间的结果更易比较。均方根误差表示为

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - p_i)^2} \cdots\cdots(2.30)$$

RMSE 与 MAE 相似，均为绝对误差。

2.4 本章小结

本章主要介绍了目前比较常用的缺失数据补全方法和时间序列数据预测方法。其中缺失数据补全方法主要包括反距离加权插值、克里金插值、ARIMA、SVR，这些基本方法在数据补全过程中往往受限，难以取得理想精度。时间序列预测方法主要包含 KNN 模型、BP 神经网络以及循环神经网络，其中循环神经网络在时序预测中最为常用，包含 LSTM，Bi-RNN，BiLSTM 以及 GRU 等多种变种模型，这些基本模型结构简单，难以适应不同的海洋深度时序数据变化规律。最后，本章节给出了三个模型评价指标，分别为 MAE、MAPE 和 RMSE。

第 3 章 基于 BO-BiLSTM-GRU 的海洋温度数据补全

3.1 模型总体设计

模型总体设计如图 3.1 所示，主要包括四大部分，首先是对样本数据选取并进行预处理，然后构建多层 BiLSTM-GRU 模型，采用贝叶斯算法对模型超参数进行选取，然后和基线模型进行精度对比。

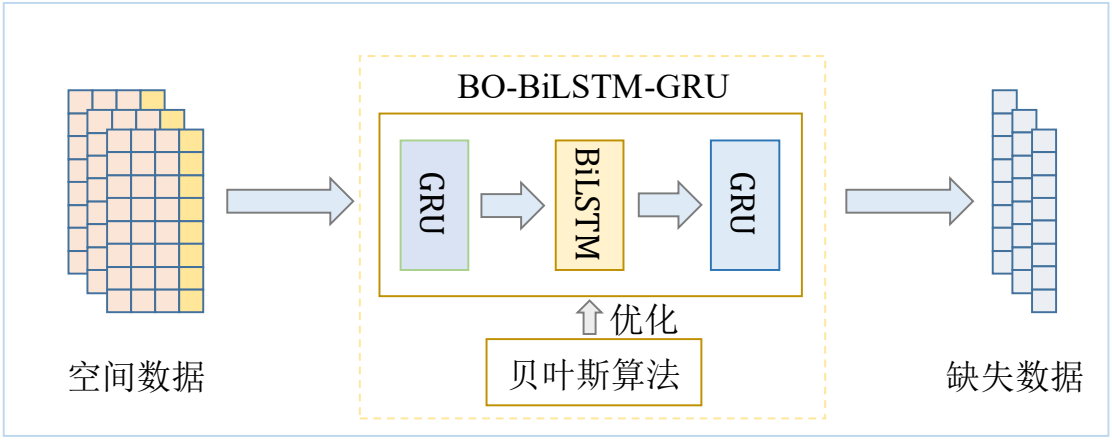

图 3.1 总体设计图

3.2 温度样本选取与预处理

本章数据来自于全球 Argo 资料中心(Argo GDAC)提供的 Argo 数据(https://doi.org/10.17882/42182)，该数据集提供了太平洋、大西洋与印度洋 2000-2022 年逐年逐月逐日的数据，本文将在 2018-2022 年太平洋数据集中选取样本。如图 3.2 所示，温度数据缺失主要分为三类，某观测点在不同深度上少量缺失，大量缺失或全部缺失。本文分别在 2018 年 3 月、2019 年 6 月、2020 年 9 月、2021 年 6 月的逐日观测数据中抽取一日全球观测数据作为实验数据，包括经度、纬度、深度、温度，进而剔除无实际观察值的数据以及异常数据(严重与实际情况不符的数据)。

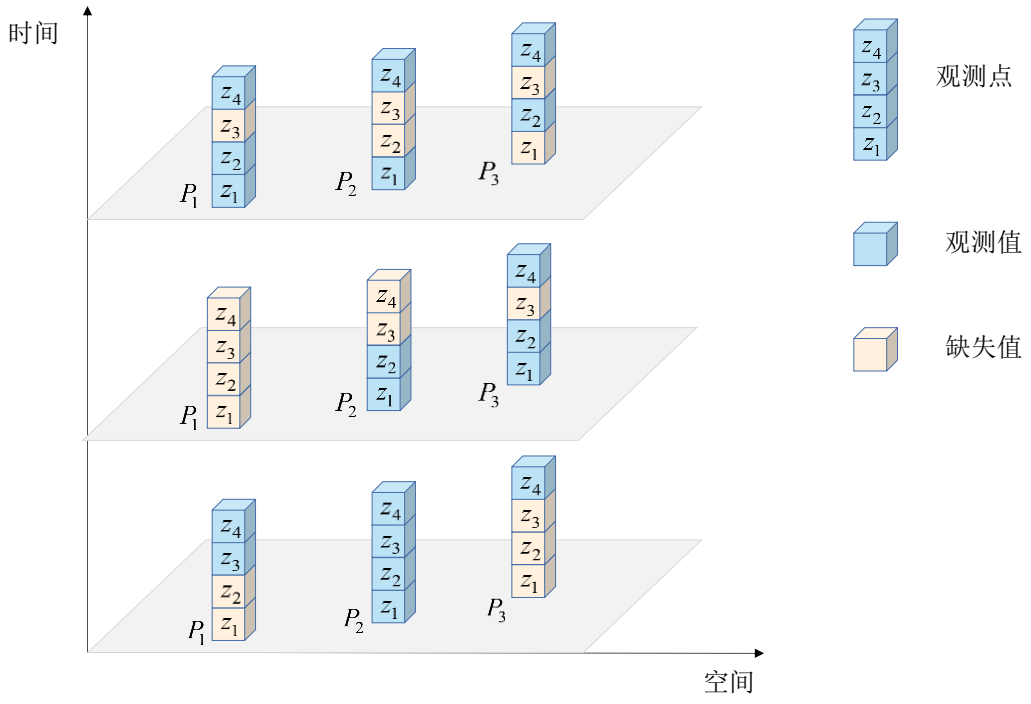

图 3.2 数据缺失状况示意图

在将数据用于模型训练前，需要对数据进行预处理，即归一化。在诸多机器学习算法中，有一部分模型不需要归一化，如树形 model。而对于多个属性特征量纲差别较大，又需要用梯度下降的方法迭代来训练和优化模型时，对数据进行归一化处理可以提高模型精度和收敛速度。如 KNN 模型，其工作原理是基于欧氏距离计算的，如果量纲不同，在计算距离时，具有较大量纲的属性特征将占据主导地位，而在实际上，具有较小量纲的属性应当被作为主要考虑因素，这样将与预期结论有较大差距，采用归一化处理的方式可以很好的解决这一问题，从而大大提高模型正确性或精度。对于需要利用梯度下降来训练的模型，将沿着梯度下降的方法来寻找最优参数。若每个属性特征具有不同的量纲，则其寻找最优参数的空间可以看作椭圆形，量纲较大的属性为长轴，在迭代过程中，可能并不会一直向最小值方向更新。而经过归一化处理后，寻找最优参数的空间可以近似看作圆形，向最小值迭代时路径更短。

本文采用的归一化方法为最大值最小值归一化，将原始数据全部归一化到[0,1]，计算公式为：

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \cdots\cdots (3.1)$$

3.3 多层 BiLSTM-GRU 模型构建

3.3.1 模型原理

所选取的样本中每一条数据均包括经度、纬度、深度、温度，表示为 $x_i=[x_{i1},x_{i2},x_{i3},x_{i4}]$,整个数据集表示为 n*4 矩阵：$\begin{bmatrix} x_{i1} & \cdots & x_{i4} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{n1} \end{bmatrix}$。温度数据补全问题可以表述为一个多元非线性回归问题，即{经度，纬度，深度；温度}，其中经度、纬度和深度为自变量，温度为因变量。本章将数据前 60%作为训练数据集，中间 20%作为验证数据集，后 20%作为测试数据集。

海洋温度数据具有地理空间上的关联性，某观测点与其附近的温度具有强关联性，BiLSTM 具有学习数据前后信息（即某观测点附近信息）的能力，可以实现双向记忆，能够很好的对空间关联性进行学习，同时，温度随着经纬度、深度的不同均具有一定趋势性，例如从赤道向两极随纬度变化而呈现出整体降低的趋势，随深度增大而整体降低的趋势，GRU 模型具有良好的趋势捕捉能力。

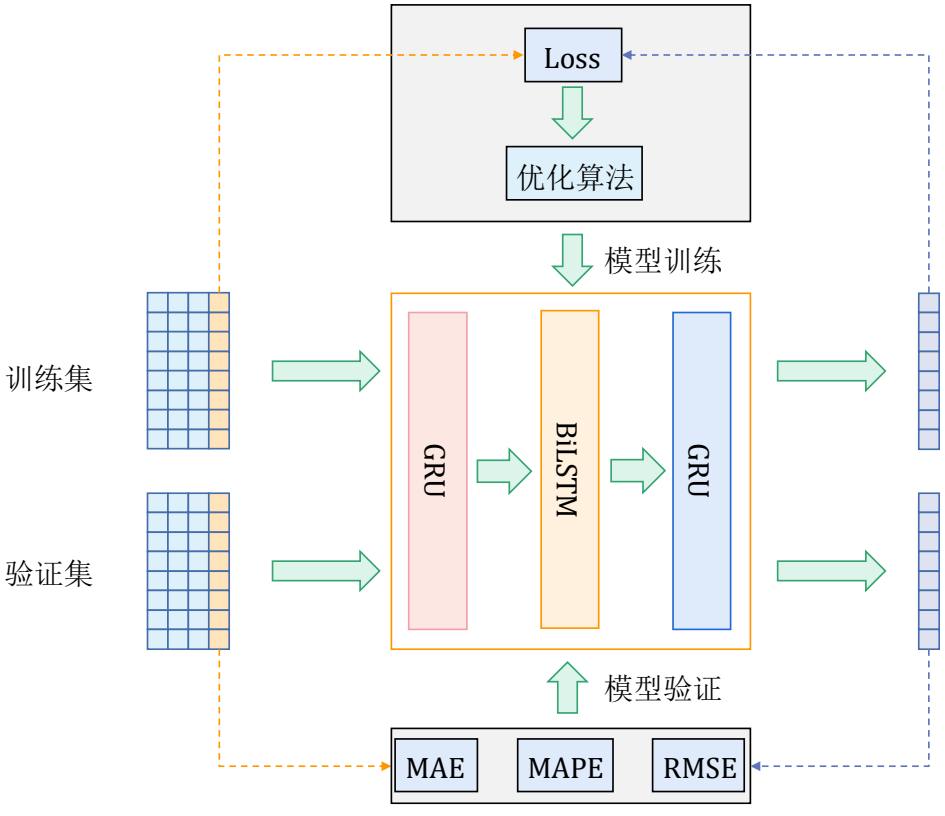

图 3.3 多层 BiLSTM-GRU 模型工作流程图

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/306033130033010054