竞赛介绍

2021 “AI Earth” 人工智能创新挑战赛，以 “AI 助力精准气象和海洋预测” 为主题，旨在探索人工智能技术在气象和海洋领域的应用。

本赛题的背景是厄尔尼诺 - 南方涛动(ENSO)现象。ENSO现象是厄尔尼诺(EN)现象和南方涛动(SO)现象的合称，其中厄尔尼诺现象是指赤道中东太平洋附近的海表面温度持续异常增暖的现象，南方涛动现象是指热带东太平洋与热带西太平洋气压场存在的气压变化相反的跷跷板现象。厄尔尼诺现象和南方涛动现象实际是反常气候分别在海洋和大气中的表现，二者密切相关，因此合称为厄尔尼诺 - 南方涛动现象。

ENSO现象会在世界大部分地区引起极端天气，对全球的天气、气候以及粮食产量具有重要的影响，准确预测ENSO，是提高东亚和全球气候预测水平和防灾减灾的关键。Nino3.4指数是ENSO现象监测的一个重要指标，它是指Nino3.4区(170°W - 120°W，5°S - 5°N)的平均海温距平指数，用于反应海表温度异常，若Nino3.4指数连续5个月超过0.5℃就判定为一次ENSO事件。本赛题的目标，就是基于历史气候观测和模式模拟数据，利用T时刻过去12个月(包含T时刻)的时空序列，预测未来1 - 24个月的Nino3.4指数。

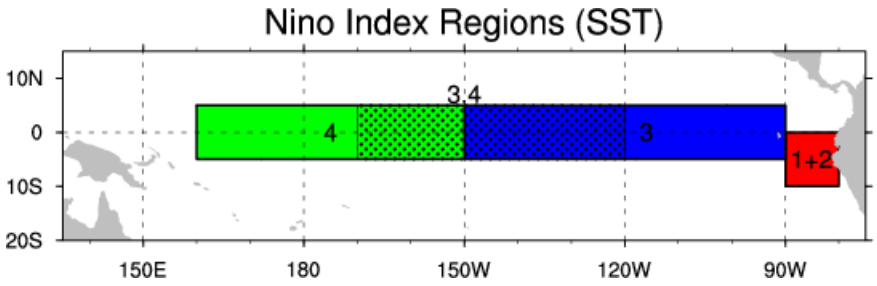

图1 Nino3.4区域

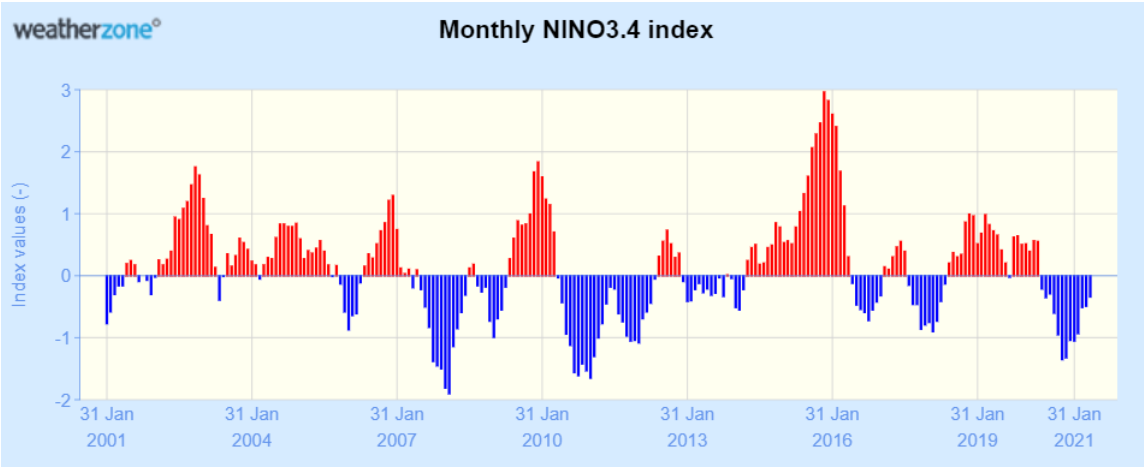

图2 Nino3.4指数（图片来源于weatherzone.com.au）

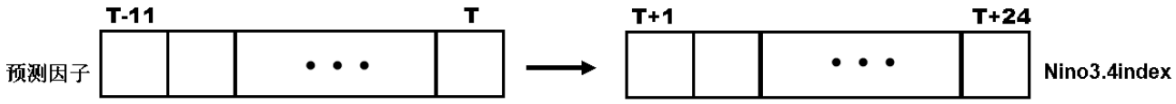

图3 赛题示意图

基于以上信息可以看出，我们本期的组队学习要完成的是一个时空序列的预测任务。

竞赛题目

数据简介

本赛题使用的训练数据包括CMIP5中17个模式提供的140年的历史模拟数据、CMIP6中15个模式提供的151年的历史模拟数据和美国SODA模式重建的100年的历史观测同化数据，采用nc格式保存，其中CMIP5和CMIP6分别是世界气候研究计划(WCRP)的第5次和第6次耦合模式比较计划，这二者都提供了多种不同的气候模式对于多种气候变量的模拟数据。这些数据包含四种气候变量：海表温度异常(SST)、热含量异常(T300)、纬向风异常(Ua)、经向风异常(Va)，数据维度为(year, month, lat, lon)，对于训练数据提供对应月份的Nino3.4指数标签数据。简而言之，提供的训练数据中的每个样本为某年、某月、某个维度、某个经度的SST、T300、Ua、Va数值，标签为对应年、对应月的Nino3.4指数。

需要注意的是，样本的第二维度month的长度不是12个月，而是36个月，对应从当前year开始连续三年的数据，例如SODA训练数据中year为0时包含的是从第1 - 第3年逐月的历史观测数据，year为1时包含的是从第2年 - 第4年逐月的历史观测数据，也就是说，样本在时间上是有交叉的。

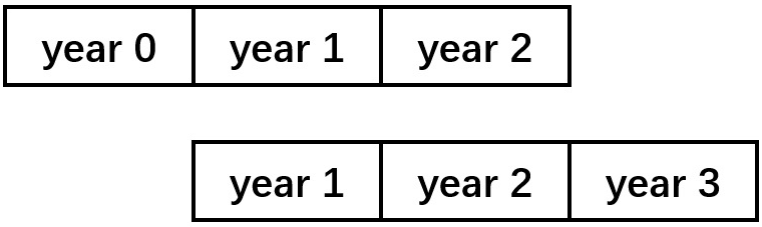

图4 样本时间跨度示意图

另外一点需要注意的是，Nino3.4指数是Nino3.4区域从当前月开始连续三个月的SST平均值，也就是说，我们也可以不直接预测Nino3.4指数，而是以SST为预测目标，间接求得Nino3.4指数。

测试数据为国际多个海洋资料同化结果提供的随机抽取的N段长度为12个月的时间序列，数据采用npy格式保存，维度为(12, lat, lon, 4)，第一维度为连续的12个月份，第四维度为4个气候变量，按SST、T300、Ua、Va的顺序存放。测试集文件序列的命名如test_00001_01_12.npy中00001表示编号，01表示起始月份，12表示终止月份。

评估指标

本赛题的评估指标如下：

$$Score = \frac{2}{3} \times accskill - RMSE$$

其中$accskill$为相关性技巧评分，计算方式如下：

$$accskill = \sum_{i=1}^{24} a \times ln(i) \times cor_i$$

$$(i \leq 4, a = 1.5; 5 \leq i \leq 11, a = 2; 12 \leq i \leq 18, a = 3; 19 \leq i, a = 4)$$

可以看出，月份i增加时系数a也增大，也就是说，模型能准确预测的时间越长，评分就越高。

cor_i是对于N个测试集样本在时刻i的预测值与实际值的相关系数，计算公式如下：

$$cor_i = \frac{\sum_{j=1}^{N}(y_{truej} - \bar{y}_{true})(y_{predj} - \bar{y}_{pred})}{\sqrt{\sum(y_{truej} - \bar{y}_{true})^2 \sum(y_{predj} - \bar{y}_{pred})^2}}$$

其中y_{truej}为时刻i样本j的实际Nino3.4指数，$\bar{y}_{true}$为该时刻N个测试集样本的Nino3.4指数的均值，y_{predj}为时刻i样本j的预测Nino3.4指数，$\bar{y}_{pred}$为该时刻N个测试集样本的预测Nino3.4指数的均值。

$RMSE$为24个月份的累计均方根误差，计算公式为：

$$RMSE = \sum_{i=1}^{24} rmse_i$$

$$rmse = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_{truej} - y_{predj})^2}$$

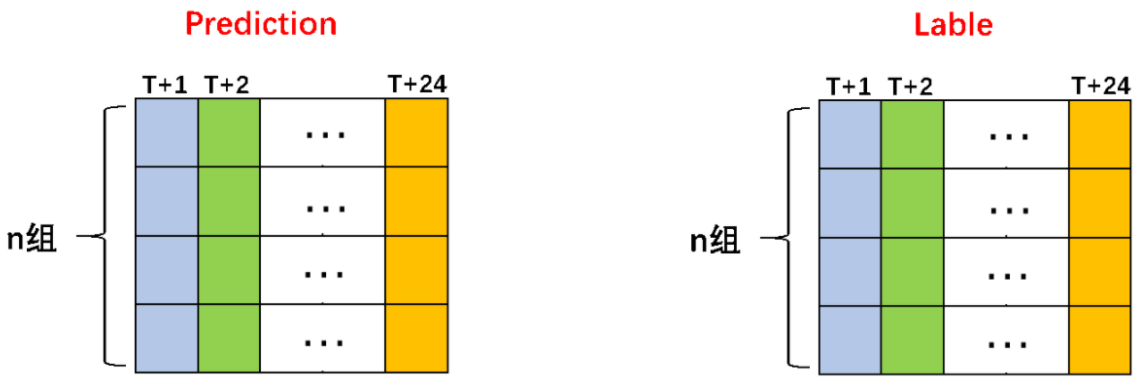

图5 评估指标计算示意图

赛题分析

分析上述赛题信息可以发现，我们需要解决的是以下问题：

- 对于一个时空序列预测问题，要如何挖掘时间信息？如何挖掘空间信息？
- 数据中给出的特征是四个气象领域公认的、通用的气候变量，我们很难再由此构造新的特征。如果不构造新的特征，要如何从给出的特征中挖掘出更多的信息？
- 训练集的数据量不大，总共只有$140 \times 17 + 151 \times 15 + 100 = 4745$个训练样本，对于数据量小的预测问题，我们通常需要从以下两方面考虑：
  - 如何增加数据量？
  - 如何构造小（参数量小，减小过拟合风险）而深（能提取出足够丰富的信息）的模型？

学习目标

我们对比赛top选手的方案进行了梳理和整合，形成了本次组队学习的五个小目标，希望你能够带着以上问题进行学习，在学习过程中找到答案。

我们希望你在本次组队学习中能有以下收获：

1. 掌握气象数据分析的常用工具。
2. 掌握时空数据的分析能力。
3. 掌握在本次组队学习中用到的模型。
4. 学会在时空序列预测问题中进行模型选择和模型构造的一些思路和方法。

同时，期待你在本次组队学习中不止局限于给出的任务，能够有更多的思考和拓展。

组队学习安排

Task01：气象数据分析的常用工具（2天）

Task02：数据分析（2天）

Task03：模型建立之 CNN + LSTM（3天）

Task04：模型建立之 TCNN + RNN（4天）

Task05：模型建立之 SA-ConvLSTM（4天）

相关资料

一、比赛官网

https://tianchi.aliyun.com/competition/entrance/531871/information

二、比赛开源方案

1. https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.15.561d53309Kn9hK&postId=210391（swg-lhl，Rank1）
2. https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.9.561d53309Kn9hK&postId=210734（ailab）
3. https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.6.561d53309Kn9hK&postId=210836（有源码，神之一手YueTan，Rank5）
4. https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.18.561d5330HKwYOW&postId=196536（有源码，学习AI的打工人）
5. https://github.com/jerrywn121/TianChi_AIEarth?spm=5176.21852664.0.0.6b612aedW6oyIQ （有源码，吴先生的队伍）

项目贡献情况

- 项目构建与整合：曾海如