分类号 TP391 学号 18020038

U D C 004 密级 公 开

工学硕士学位论文

机器学习方法在处理风云四号红外高光谱资料中的应用

硕士生姓名 张琪

学 科 专 业 计算机科学与技术

研 究 方 向 海洋信息工程

指 导 教 师 周恩强 研究员

协助指导教师 张卫民 研究员

国防科技大学研究生院

二〇二〇年十月

机器学习方法在处理风云四号红外高光谱资料中的应用

国防科技大学研究生院

Application of Machine Learning Methods in Processing Fengyun Infrared Hyperspectral Data

Candidate：Zhang Qi

Supervisor：Associate Prof.　Zhou Enqiang

Associate Supervisor：Prof.　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

October, 2020

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技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：《机器学习方法在处理风云四号红外高光谱资料中的应用》

学位论文作者签名：张琪　　日期：2020 年 10 月 26 日

学位论文版权使用授权书

本人完全了解国防科技大学有关保留、使用学位论文的规定。本人授权国防科技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：《机器学习方法在处理风云四号红外高光谱资料中的应用》

学位论文作者签名：张琪　　日期：2020 年 10 月 26 日

作者指导教师签名：　　日期：2020 年 10 月 26 日

目 录

摘 要 iv
ABSTRACT vi
第一章 绪论 1
1.1 研究背景 1
1.1.1 卫星遥感与数值天气预报 1
1.1.2 大气垂直探测技术发展 1
1.1.3 GIIRS 资料应用前景 3
1.1.4 GIIRS 资料应用面临的问题 4
1.1.5 机器学习与遥感和数值天气预报 8
1.2 论文选题及研究现状 9
1.2.1 论文选题 9
1.2.2 问题研究现状 10
1.3 主要研究内容和论文结构 12
1.3.1 主要研究内容 12
1.3.2 论文结构 13
第二章 机器学习算法 13
2.1 机器学习基础理论 14
2.2 模型泛化能力 15
2.3 模型性能评估 16
2.3.1 分类模型 16
2.3.2 回归模型 17
2.4 机器学习算法简介 18
2.4.1 Logistic 回归 18
2.4.2 人工神经网络 19
2.4.3 高斯过程回归 21
2.4.4 随机森林 23
2.5 本章小结 24
第三章 基于 L1 正则化 Logistic 回归的 GIIRS 云检测算法 25
3.1 数据集和研究区域概况 25
3.1.1 GIIRS 数据 25

3.1.2 AGRI 成像仪云掩膜产品 28
3.1.3 葵花 8 号云图 28
3.1.4 研究区域概况 29
3.2 成像仪-高光谱仪器匹配云检测算法 29
3.3 机器学习云检测算法 31
3.3.1 代价函数定义 32
3.3.2 数据预处理 32
3.3.3 模型输入特征选择 32
3.3.4 模型参数调节 33
3.4 机器学习云检测算法试验 33
3.4.1 训练样本数 33
3.4.2 超参数调节 35
3.4.3 分类概率阈值调节 36
3.5 机器学习云检测结果 36
3.5.1 云检测误差统计结果 37
3.5.2 云检测可视化结果 37
3.6 试验结果讨论 38
3.6.1 算法适用性 38
3.6.2 特征通道贡献 39
3.7 本章小结 40
第四章 大气温度廓线反演 41
4.1 数据集说明 41
4.1.1 欧洲中期数值预报中心第五代再分析资料 41
4.1.2 AIRS Level 2 级标准反演产品-大气温度廓线 41
4.2 构建样本集 42
4.2.1 训练样本集 42
4.2.2 均衡数据集采样 43
4.2.3 温度探测特征通道选取 45
4.3 机器学习算法设置 47
4.3.1 人工神经网络 47
4.3.2 高斯过程回归 48
4.3.3 随机森林 48
4.4 机器学习反演大气温度廓线结果 49
4.4.1 非台风过程 49

4.4.2 台风过程 50
4.5 试验结果讨论 52
4.5.1 反演误差 52
4.5.2 训练时间 54
4.6 本章小结 55
第五章 总结与展望 57
5.1 论文总结 57
5.2 研究展望 57
致 谢 59
参考文献 60
作者在学期间取得的学术成果 67
附录 A 模型特征通道索引号 68

表 目 录

表 1.1 长波红外光谱波段的作用 3
表 3.1 正负样本数目 34
表 3.2 两个模型在测试集上云检测统计结果 37
表 3.3 三个云图场景的信息 37
表 4.1 非台风样本集和台风样本集中的样本数量 42
表 4.2 $D_1 \sim D_4$ 的概率分布和香农熵 44
表 4.3 RF 的最优参数组合 49
表 4.4 6 个模型的十次平均训练时间 55
表 A.1 云检测模型 38 个特征光谱通道索引号 68
表 A.2 反演大气温度廓线模型的 50 个特征光谱通道索引号 68

图 目 录

图 1.1 红外高光谱仪器的光谱探测范围及不同波段的探测目标.....3
图 1.2 红外高光谱资料在同化系统中的处理流程.....7
图 2.1 建立机器学习预测模型的流程.....15
图 2.2 混淆矩阵.....17
图 3.1 （a） GIIRS 扫描范围；（b）单个驻留点 128 个像元排列方式；（c）单个驻留点 128 个像元扫描顺序.....26
图 3.2 （a）相邻扫描带像元重叠示意图；（b）GIIRS 长波第 516 个光谱通道上相邻扫描带重叠区域量温差。.....26
图 3.3 GIIRS 的观测辐射率。（a）切趾前；（b）切趾后.....27
图 3.4 GIIRS689 个长波红外通道光谱信息。蓝色线为光谱辐射率，红色点为 NEdR 大于 0.15 的通道，绿色点为 NEdR 的平均值，绿色折线为 NEdR 的标准差。.....28
图 3.5 （a）研究区域；（b）1990-2019 年生成于南海和西北太平洋的热带气旋的路径汇总.....29
图 3.6 GIIRS 星下点像元和 AGRI 星下点像元匹配的空间示意图.....30
图 3.7 2019 年 5 月 15 日 03 时吕宋岛附近海域上空。（a） Himawari-8 可见光云图；（b）CLM；（c）AGRI-GIIRS 云检测算法得到的 GIIRS 云标签分布，其中红色点为 GIIRS 晴空像元，蓝色点为 GIIRS 有云像元，绿色点为 GIIRS 部分有云像元。.....31
图 3.8 LR_689 和 LR_38 的学习曲线；第一行采用 L_2 正则化，第二行采用 L_1 正则化；红色曲线为训练集，绿色曲线为验证集.....35
图 3.9 LR_689 和 LR_38 的 AUC 随超参数 C 的变化.....36
图 3.10 LR_689 的混淆矩阵随分类概率阈值 Ω 的变化.....36
图 3.11 云检测可视化结果.....38
图 3.12 三个模型的运行时间随 GIIRS 像元数的变化.....39
图 3.13 LR_689 的判别函数中特征通道前的系数.....40
图 4.1 研究区域对应的 GIIRS 观测起始和终止时间.....43
图 4.2 GIIRS 光谱通道的温度雅克比曲线；（a）689 个长波红外通道；（b）保留的 50 个长波红外通道.....47
图 4.3 非台风过程模型误差统计结果。第一行的三个机器学习反演模型由 S 训练得到，第二行的三个模型由 B 训练得到。.....50
图 4.4 台风过程中 6 个模型的误差统计结果;图片布局同图 5.3。.....52

图 4.5　Temp_AIRS 同 Temp_ERA5 的偏差统计结果。(a)非台风过程; (b)台风过程。
..... 54

摘 要

强对流发生前的天气预警、临近和短期天气预报都需要描述大气温度、湿度精细结构的观测资料的支持，搭载在静止卫星上的红外高光谱探测仪器是这种资料的主要来源之一。干涉式大气垂直探测仪（Geosynchronous Interferometric Infrared Sounder，GIIRS）是搭载在我国静止气象卫星风云四号 A 星上的红外高光谱仪器，相对于极轨卫星上的高光谱仪器，它具有高频次连续观测局地区域的优势，能够监测快速变化的中尺度大气系统。为了在数值天气预报系统中充分发挥 GIIRS 数据的特性和优势，本文在充分掌握 GIIRS 数据的空间分布和光谱特征后，采用机器学习方法针对其在数值天气预报中亟待解决的两个问题展开了研究：

（1）高效、准确的云检测：快速大气辐射传输模式很难精准地模拟出云中的辐射传输过程；当探测像元视场直径大于 10KM 时，红外高光谱仪器约有 90%的视场都被云污染；目前的数值预报系统无法高效处理 GIIRS 每日巨大的数据量。基于以上三点的考虑和实时预报的需求，如何在 GIIRS 资料进入同化系统前对其进行高效、准确的云检测是一个重要的问题。本文将云检测视为一个分类问题，通过训练鲁棒性强的 L_1 正则化 Logistic 回归算法达到了对 GIIRS 视场进行高效云检测的目的。与经典的成像仪匹配云检测法相比，本文提出的云检测模型在具有高准确率（>95%）的同时检测效率显著提升；对比目前三维变分同化系统中依赖于背景场的云检测方法，本文提出的方法只需将 GIIRS 观测辐射作为输入，避免了由于背景场不准确引起的误差。综上，本文提出的云检测方法能够用于资料同化系统中对 GIIRS 资料进行实时云检测。

（2）高精度大气温度廓线的反演：由于具有高光谱分辨率，相对于传统多光谱传感器，红外高光谱仪器能够探测更加精细的大气垂直结构，对天气过程预警具有重要作用。然而由卫星接收到的辐射反演大气状态这个逆问题是不适定的；此外，卫星接收辐射和大气参数之间的非线性关系、通道噪声和通道高相关性给反演大气温度廓线进一步增加了难度，使其成为一个具有挑战性的问题。为了反演出高精度的大气温度廓线，以光谱通道的温度雅可比和噪声水平作为依据，从 GIIRS 的上百个长波通道中选出了对不同高度温度敏感的 50 个低噪声通道作为输入特征，采用了三种非线性拟合能力强的机器学习算法（高斯过程回归、随机森林、人工神经网络）作为基础模型。试验中创新性地使用香农熵采样法构造出了有代表性的样本集，在反演出高精度大气温度廓线的同时显著提升了训练效率。三种模型中，基于随机森林方法的模型反演效果最为突出：非台风过程时，随机森林反演模型在 10hPa 以下的气压层上均方根误差在 1.5K 以内，偏差在±1K 以内；台风过程中，10hPa 以下均方根误差在 2K 以内，偏差在±2 K 以内。综上所述，本

文提出的大气温度廓线反演模型能够高效地获得高精度的大气温度廓线。

关键词：静止卫星；红外高光谱；GIIRS；机器学习；云检测；温度廓线反演

ABSTRACT

Weather warning before heavy convection, now-casting and short-term weather forecast rely on the support from continuous observations of the structure of atmospheric temperature and humidity. The infrared hyperspectral detection instrument carried on the geostationary satellite is one of the main sources of this kind of data. Geosynchronous Interferometric Infrared Sounder (GIIRS) is an infrared hyperspectral instrument onboard Fengyun-4A, a geostationary meteorological satellite of China. Compared with the hyperspectral instrument onboard polar-orbiting satellites, GIIRS has an unprecedented advantage to observe the fast-changing water vapor and temperature structure related to severe weather events. In order to give full play to the characteristics and advantages of GIIRS data in numerical weather prediction (NWP) system, after fully understanding the spatial distribution and spectral characteristics of GIIRS data, this paper uses machine learning method to study two problems to be solved urgently in NWP:

(1)Efficient and accurate cloud detection: the current atmospheric radiative transfer model is difficult to accurately simulate the radiative transfer process under cloudy situation; due to technical limitations, the horizontal spatial resolution of hyperspectral instruments is low, so about 90% of the field of view is contaminated by clouds; the current NWP system can not deal with the massive daily GIIRS data. Based on the above three considerations and the needs of real-time forecast, how to carry out efficient and accurate cloud detection of GIIRS data before entering the data assimilation system is an important problem. In this paper, cloud detection is regarded as a classification problem, and efficient cloud detection is achieved by training robust Logistic regression algorithm based on L_1 regularization. Compared with the classical imager assisted cloud detection method, the detection model proposed in this paper can significantly improve the detection efficiency while having high accuracy (> 0.95). Compared with the current cloud detection methods which depend on the background field in the three-dimensional variational assimilation system, the method proposed only needs to take the GIIRS observation radiation as the input to avoid the error caused by the inaccuracy of the background field. Based on the above discussion, the cloud detection method proposed in this paper can be used for real-time cloud detection of GIIRS data in data assimilation system.

(2)Retrieval of high-precision atmospheric temperature profile: because of its high spectral resolution, compared with the traditional multispectral sensor, infrared hyperspectral instrument can detect more elaborate atmospheric vertical structure, which plays an important role in the early warning of important weather processes. However, the problem of retrieving the atmospheric state from the radiation received by

the satellite is ill-posed, and the noise, channel correlation and the nonlinear relationship between the radiation received by the satellite and the atmospheric parameters make the retrieval of the atmospheric temperature profile a challenging problem. In order to retrieve the high-precision atmospheric temperature profile, based on the temperature Jacobian and noise level of spectral channels, 50 low-noise channels sensitive to temperature at different heights are selected from hundreds of long-wave channels as input features. Three machine learning algorithms with strong nonlinear fitting capability (Gaussian process regression, random forest and artificial neural network) are used as the basic model. In the experiment, Shannon entropy sampling method is creatively used to construct a representative sample set, which not only helps retrieve the high-precision atmospheric temperature profile, but also significantly improves the training efficiency. Among three algorithms, random forest bassed retrieval model has higher training efficiency and outstanding retrieval effect: In the non-typhoon process, below 10hPa, the Root Mean Squared Error (RMSE) of the random forest is less than 1.5K and the Bias is less than ±1K below 10hPa; in the typhoon process, the RMSE is less than 2K and the Bias is less than ± 2K. In summary, the atmospheric temperature profile retrieval model proposed in this paper can obtain high-precision atmospheric temperature profile efficiently.

Key Words：Geostationary satellite，Hyperspectral infrared data，GIIRS，machine learning，cloud detection，temperature profile retrieval

第一章　绪论

1.1　研究背景

1.1.1　卫星遥感与数值天气预报

数值天气预报的本质是在给定的大气初始和边界条件下，对一组描述大气运动的非线性偏微分方程组向前积分实现对未来大气状态的预报[1]。在过去的 20 年间，数值预报技巧的提高主要归因于数值模式的发展、星载遥感仪器探测能力的提升以及能够得到初始分析场的同化系统的改进。由于地球上 70%的面积都被水体（海洋、内陆水域）、山体、沙漠和极地冰体覆盖，因此采用传统的地面观测手段探测全球大气是不可能的；即使一些区域的传统地面探测相对比较密集，但这些观测也很难满足中尺度预报的时间和空间需求。此外，大气成分信息，如 O_3、沙尘、CH_4、SO_2 和 CO 通常是无法通过地面观测探测到的[2]。在卫星遥感高速发展之前，海洋上空的观测数据异常稀少，因此模式对于海洋面积广阔的南半球的预报能力差，远远比不上陆地面积更大、气象观测资料更丰富的北半球，星载遥感仪器的发展深刻提高了海面上空的预报水平。除此之外，气象卫星自大气上层向下探测，不受地形和下垫面类型的影响，因而可以获得空间覆盖更加均匀、时间连续且稳定有效的观测资料。目前为止，卫星遥感探测已经成为数值天气预报中不可缺少的观测资料来源。

1.1.2　大气垂直探测技术发展

1969 年美国“云雨 3 号”试验天气卫星是全球天气试验项目的先驱，其上搭载的红外辐射分光仪（ Satellite Infrared Radiation Spectrometer，SIRS-A）是世界上第一台对地观测的探测类仪器。SIRS-A 在 15μm 的 CO_2 吸收波段设置了 8 个长波红外通道用于探测大气温度的垂直分布，其扫描方式为星下点单点观测。虽然 SIRS-A 获得的观测资料的星下点空间分辨率和光谱分辨率都很低，但是仍然给当时的天气预报带来了正向改进效果。1978 年搭载在泰罗斯-N （TIROS-N）上的泰罗斯业务垂直探测器 （TIROS Operational Vertical Sounder，TOVS） 实现了从星下点单点观测到跨轨扫描观测的进步，标志着探测技术的一次重大改进；同时该仪器增加了对大气湿度的探测，并且开始为天气预报业务提供温湿度反演产品。90 年代末，先进大气探测系统 ATOVS（Advanced TIROS Operational Vertical

Sounder，ATOVS）再次升级，其上增加了微波湿度探测器，进一步提升了对大气水汽垂直分布的探测能力，使卫星实现了从以晴空探测为主到全天候观测的转变。但以上仪器都被认为是传统卫星探测器，由于受限于仪器的分光技术，该类仪器的光谱分辨率普遍较低，导致它们对地球大气的垂直探测能力不强。1987 年，世界气象组织对全球大气温度、湿度探测提出要求，卫星探测必须达到在对流层大气温度均方根误差（Root Mean Squared Error）小于 1K，湿度均方根误差小于 10%，大气探测的垂直分辨率达到 1KM，即达到无线电探空水平，才能为数值天气预报的改进做出重要贡献[3]。传统的滤光片式大气垂直探测无法满足达到改进天气预上述要求，因此人们开始研究发展具有高光谱分辨率的红外高光谱探测技术。

红外高光谱仪器主要有两种实现分光的技术：（1）光栅分光技术；（2）干涉分光技术。与干涉分光技术相比，光栅分光的成本高，且会降低入射光的光通量，因此对红外探测器的探测敏感度要求更高，实现技术更难，还容易对后续仪器引入随机的辐射定标误差。搭载在 Aqua 上的大气红外探测器 （Atmospheric Infrared Sounder，AIRS）是世界上首个真正意义上的红外高光谱仪器，该仪器采用光栅分光技术，具有 2378 个红外探测通道，光谱分辨率大于 1200，实现了遥感领域对大气温度、湿度廓线和臭氧总量探测能力的显著提升[4]。继 AIRS 之后，有多个采用干涉分光器技术的红外高光谱仪器成功在轨运行，分别是 2006 年发射并载于欧洲气象业务 A 星（Meteorological Operational Satellite-A，MetOp-A）的干涉式超高光谱分辨率红外大气探测仪（Infrared Atmospheric Sounding Interferometer，IASI）[5]，它具有 8461 个通道，最大的特点是实现了对光谱范围 3.62μm~15.5μm 的连续探测，可以用于提取诸多大气成分，如 CO，N_2O，CH_4，SO_2；2011 年发射并载于索米国际极地伙伴卫星（Suomi National Polar-orbiting Partnership，SNPP）的跨轨道红外探测仪（Cross-Track Infrared Sounder，CrIS）[6]在长、中、短波波段具有不同的光谱分辨率，分别为 0.625cm^{-1}、1.25cm^{-1} 和 2.5cm^{-1}。2017 年我国发射了风云-3D 极轨气象卫星，其上搭载的红外高光谱大气垂直探测仪（High-resolution Infrared Atmospheric Sounder， HIRAS）是风云三号极轨卫星系列上的首个红外高光谱探测仪。

以上红外高光谱仪器的观测资料已经应用于全球或区域数值预报系统并且产生了正向效果，但它们均搭载在极轨卫星上。极轨卫星的轨道高度在几百到上千公里，对于同一地点的观测时间间隔长，每天只能对同一地点观测两次，无法满足监测快速变化的天气系统的需求。搭载在静止卫星上的仪器可以对某一固定区域进行连续观测，因此可以获得高时间频次的观测资料，对于追踪大气中快速变

化的重要天气过程具有重要意义，因此静止卫星上的红外高光谱仪器已经成为世界各个数值天气预报（Numerical Weather Prediction，NWP）中心现在以及未来的研究重点之一。我国于 2016 年发射的风云四号上搭载的干涉式垂直探测仪（Geosynchronous Interferometric Infrared Sounder，GIIRS）是世界上第一台搭载于静止卫星上的红外高光谱探测仪，它对于数值天气预报模式同化高频次红外高光谱资料的研究以及发展红外高光谱仪器技术具有重要意义。

以上 5 个在轨运行的红外高光谱仪器探测的光谱范围[7]如图 1.1 所示；表 1.1 给出了长波红外不同波段对应的主要探测目标， GIIRS 的长波光谱范围在 700cm^{-1}~1130cm^{-1}，对应的探测目标包含大气温度、云、地表和臭氧。

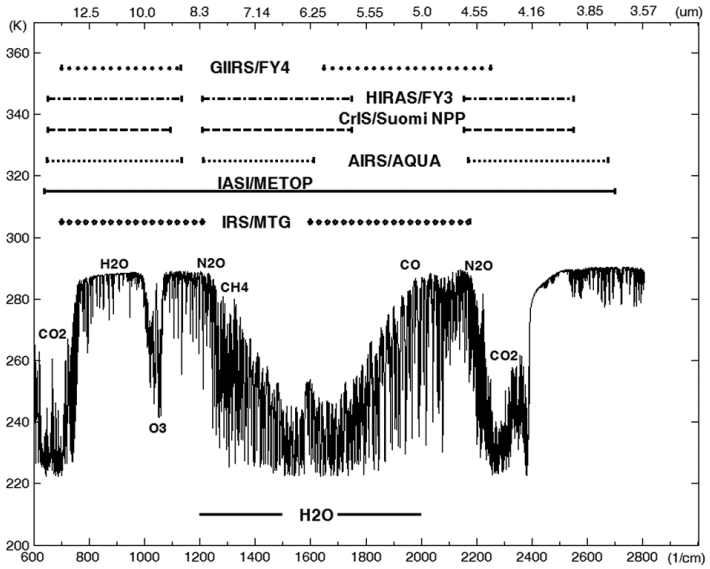


图 1.1　红外高光谱仪器的光谱探测范围及不同波段的探测目标

表 1.1 长波红外光谱波段的作用

光谱范围	主要作用
$650cm^{-1}$~$790cm^{-1}$	大气温度廓线和云产品反演
$790cm^{-1}$~$1180cm^{-1}$	探测地表、云物理属性和臭氧
$1210cm^{-1}$~$1650cm^{-1}$	探测大气水汽、温度、N_2O、CH_4 和 SO_2
$2100cm^{-1}$~$2150cm^{-1}$	探测大气柱总 CO 含量
$2150cm^{-1}$~$2250cm^{-1}$	探测大气温度和大气柱总 N_2O 含量
$2350cm^{-1}$~$2420cm^{-1}$	探测大气温度
$2420cm^{-1}$~$2700cm^{-1}$	探测地表/云物理属性

1.1.3　GIIRS 资料应用前景

（1） 数值天气预报中的应用

由于高光谱仪器具有高光谱分辨率和低噪声水平，探测到的辐射率中包含比以前的垂直探测仪器多一个量级以上的大气结构信息。根据全球多个 NWP 中心的报告，红外高光谱仪器是单个仪器中对当前业务 NWP 系统的同化和预报结果有最大正效果的仪器[2]。一些区域预报模型中同化入了来自高光谱仪器的观测发现其对降水预报产生了显著的正向效果。

（2） 天气预警应用

红外高光谱的光谱区域涵盖大气窗区许多相对较弱的水汽和 CO_2 的吸收线，这些吸收线上的观测可以为监测对流层低层热力状态的演变提供关键信息[8]。此外，从红外高光谱资料中反演得到的大气参数廓线可以为中尺度环境下的强风暴短时预报和短期预报提供更丰富的信息，例如与暴雨、大气稳定性和边界层结构有关的指数[9]。Sieglaff 等人[10]提出 800hPa 和 600hPa 之间的等效位温差可以作为雷电潜势的预警指标。Li 等人[11]发现 AIRS 的探测资料能够在对流发生前几个小时刻画出晴空下的不稳定区域，这对于降低强对流风暴的虚警率尤为重要。尽管目前极轨卫星上的红外高光谱探测仪的垂直分辨率与以往相比有了质的飞跃，但由于水平空间分辨率和时间分辨率的限制，其捕捉对流势能的能力仍然有限。通常需要将红外高光谱资料和其他卫星资料，或地面观测资料以及数值预报中的分析场结合，才能提高对强对流天气的临近预报能力。同极轨卫星上的红外高光谱探测仪器相比，搭载在静止卫星上的红外高光谱仪器兼具高光谱分辨率和高时间分辨率这两个优势，能够监测一些变化快、强度大的重要天气过程中水汽场和温度场的连续变化信息，在恶劣天气事件的早期预警方面具有显著的优势。

（3） 大气成分和云属性反演应用

云在地气系统的能量平衡和水文循环中起着重要作用，然而由于气候模式中云类型的多样性、云结构的复杂性、辐射分布的非均匀性，云的反馈机制存在很大的不确定性。红外波段的 12.5μm ~15μm 光谱区域对云顶高度和有效云量特别敏感，9μm 波段包含了有关云层光学厚度和粒子大小的信息。红外高光谱仪器在这些波段的高光谱探测辐射已经成为联合反演大气和云参数的基础。红外高光谱数据的另外一个重要目标是大气成分反演和监测。由于可以反演痕量气体和气溶胶的垂直分布廓线（如 O_3、CO、CH_4 等）[2]，红外高光谱资料在空气质量监测中具备极大的潜力，而且红外高光谱资料可以提供长期连续的大气成分记录用于研究其年际变化。

1.1.4 GIIRS 资料应用面临的问题

（1） 如何处理高相关性的海量数据

目前搭载在风云四号上的 GIIRS 探测仪具有 1650 个光谱通道，能够提供四维的大气信息。由于其高频率的探测能力，其每日的观测数据量巨大，在目前的探测器规模下的数据传输量已达到 66Mb/s。GIIRS 每日可提供约10^9个亮温观测值，这个数目和欧洲中期天气预报中心（European Centre for Medium-range Weather Forecasts，ECMWF）同化入集成数值预报系统（Integrated Forecast System，IFS）中的观测总量相当[12]。NWP 系统利用 GIIRS 观测存在以下问题：一方面，由于计算资源的限制，这些海量的数据不能全部进入同化系统；另一方面，变分同化理论中假设观测之间是不相关的，而 GIIRS 资料间存在高相关性。GIIRS 资料的高相关性体现在两方面：高空间相关性和高光谱相关性。高光谱相关性是在同一个观测像元下，很多光谱通道对应的探测高度非常相近，因此不同通道探测的信息存在相关性；高空间相关性是对于相同的光谱通道，某一像元与相邻像元的探测对象属性相同或者相似，所以探测到的信息具有高空间相关性。面对巨大的数据量以及数据间存在的高相关性，如何从红外高光谱资料中高效提取出能够被同化系统吸收的有效信息是值得研究的问题。

（2）GIIRS 资料进入同化系统前的资料预处理问题

利用卫星观测结合先进的同化系统生成可靠的分析场是改进数值预报模式精度的主要方向之一。目前的同化系统框架中假设背景场和观测场满足高斯正态的无偏分布[13]，但实际上由于观测视场中云的影响、探测仪器本身的误差、数据传输中的偶然误差、数值天气预报系统并非完美等原因导致背景场和观测场并不满足这一假设，所以在观测资料进入同化系统前，需要对观测资料进行一系列严格的处理使其达到上述要求。图 1.2 为卫星红外高光谱观测资料在同化系统中的处理流程[14,15]，在如今的数值预报系统的资料同化模块中，该流程中的每一步都是红外高光谱资料应用过程中亟待解决或是需要进一步改进的问题，下面对该流程做出介绍：

- 通道选择

在变分同化系统中，为了在节省计算资源的同时保证得到的分析场的质量，假设观测协方差矩阵为对角矩阵，即观测变量在空间和光谱上是相互无关的。由于红外高光谱资料的成百上千个通道间存在明显的空间相关性和光谱相似性[14]，这些通道全部进入数值天气预报系统将违背观测变量独立的假设，还会造成同化和反演问题的不适定、消耗大量的计算资源，因此需要对红外高光谱资料进行合理的通道选择。

- 异常值剔除

该步骤通过剔除离群点资料，使得经过质量控制后的通道亮温偏差服从高斯分布。考虑到模式空间中地表温度、下垫面发射率的不确定性，观测空间中卫星观测的临边扫描变暗等因素[15]，需要对观测值和模拟值之间的偏差进行检验。离群值的定义将影响经过该步骤后是否在真正保留下有效信息的同时剔除了“错误”的资料。需要注意的是，一些离群值代表了天气过程，因此过度的剔除会导致很多有用的信息的丢失。如何度量离群值以及对经典的变分代价函数进行改写，使其能同化部分误差为非高斯分布的观测资料是资料质量控制和同化的方向和难点。

- 偏差订正

资料同化理论中假设观测误差满足随机、无偏分布，但实际上辐射观测值和模拟值之间的误差包含许多系统性的偏差的混合，而不是随机的，且这些系统性偏差常常超过了观测资料本身的噪声水平。例如，卫星观测包含依赖于仪器的偏差；辐射传输模式中的近似可能会在同化中造成复杂的、依赖于数值预报得到的大气状态的系统偏差等[16]。来自于不同源头的偏差常常表现出一些时间和空间上的分布特征，例如偏差具有仪器扫描角依赖、地理位置或气团特征依赖等。消除观测和模拟值之间的系统性偏差是观测资料进入变分系统前的重要步骤。当观测偏差具有很强的时空变异性时，对其建模是非常具有挑战性的。

- 云检测

在红外波段，云吸收红外辐射同时以自身的温度发射辐射。目前的研究对于云内的物理过程尚不十分清楚；快速辐射传输模式对云内的物理过程的模拟能力不强；此外，数值天气预报模式也无法提供准确的云水廓线信息。因此红外资料在进入到同化系统前需要进行云检测，把晴空视场和有云视场分开，晴空视场可以进入下一步，有云视场则需要做单独处理。

- 数据稀疏化

为了保证观测误差协方差矩阵为对角矩阵的假设，要求进入同化系统的观测资料在光谱和空间上均无关。观测的误差相关是未知的，如果对观测误差相关做出估计会极大地增加计算复杂度。此外，观测资料的密度和模式网格分辨率之间存在一种联系：当观测资料的密度太大并且忽略观测误差相关时，同化得到的分析场质量将会下降。以上问题是研究数据稀疏化方法的动机，其目的在于减小观测之间的空间误差相关性，同时从观测资料中提取出能够在资料同化系统中最优使用的必要信息。

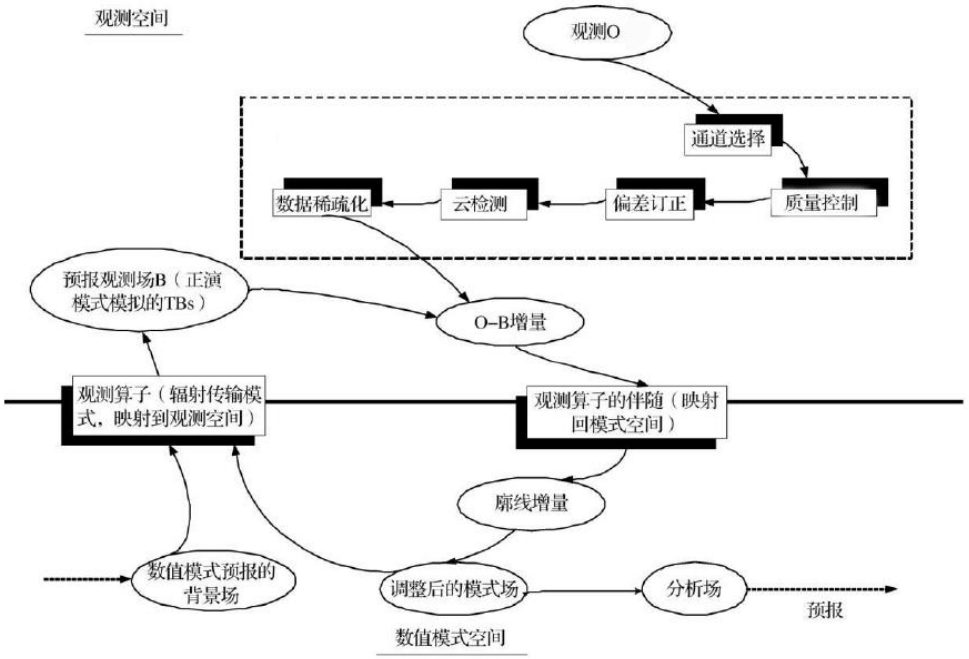


图 1.2　红外高光谱资料在同化系统中的处理流程

（3）非线性不适定的大气参数反演问题

卫星探测辐射的原理可以用大气辐射传输方程来表示，以晴空大气辐射传输方程为例，在只考虑大气吸收，不考虑散射，并且满足局地热力平衡的条件下，红外波段接收到的晴空大气辐射[17]为：

$$I=\varepsilon_s\tau(p_s,\mu)B(T_s)+\int_{p_s}^{0}B(T_p)\frac{\partial\tau(p,\mu)}{\partial(\ln p)}d(\ln p)+(1-\varepsilon_s)\tau(p_s,\mu)\int_{0}^{p_s}B(T_s)\frac{\partial\tau(p,\mu)}{\partial(\ln p)}d(\ln p) \quad (1.1)$$

式（1.1）为卫星接收到的辐射，右边的三项分别为地表发射辐射，大气每一层向上发射辐射，大气下行辐射被地表反射后的辐射。（1.1）中 μ 为辐射传输方向天顶角的余弦，ε_s 为地表发射率；B 为普朗克函数，与发射源的波长和温度有关；p 为气压；τ 为大气透过率，$\tau(p_s,\mu)$ 为整层大气透过率，$\tau(p,\mu)$ 为从气压层 p 到大气层顶的大气透过率；τ 由多种大气吸收成分（如氧气、臭氧、水汽、气溶胶等）对辐射的削减程度决定。

由（1.1）可知卫星在某个波段/通道接收到的辐射 I 与大气和地表的温度、大气和地表的发射率、大气中的吸收气体有关，所以高光谱仪器探测的辐射携带了大气和地表热力状态以及大气中的物理参数信息。根据大气辐射传输理论，可以从观测辐射中得到以下信息：（1）普朗克函数中的大气温度和地表温度；（2）

地表发射率；（3）通过透过率得到各种大气成分（如水汽、O_3）和大气气溶胶（如SO_2、N_2O）的含量。然而，通过大气和地表状态基于大气辐射传输方程计算卫星接收到的辐射是一个正向过程，对于一种给定的大气和地表状态计算出的辐射值唯一，该过程在数学上是一个适定问题；但是由卫星的多个通道对于同一个像元观测到的一组辐射值可能对应多种不同的大气状态，所以通过卫星接收到的辐射反演大气和地表的状态这个问题是不适定的[2]。此外，还有以下多种因素增加了反演问题的难度：（1）由于电磁波和大气的复杂作用（比如折射、透射、吸收），以及探测的目标区域受到临近地物发射反射的电磁波的干扰等因素，红外高光谱数据和大气参数之间的关系是非线性的，（2）由于通道权重函数存在重叠，所以不同光谱通道探测到的信息具有高垂直相关性；（3）遥感观测的噪声限制了由观测辐射率反演大气参数廓线所能达到的垂直精度。以上因素使得用红外高光谱资料反演高精度大气参数廓线成为一个具有挑战性的问题。

1.1.5 机器学习与遥感和数值天气预报

卫星遥感和数值预报面临的许多挑战可以归结到以下三个方面：大数据、先进的预报模型的需求和应用以及增加的用户需求[18]。

首先，目前的 NWP 系统不能充分利用日益增长的类型各异的观测资料和巨大的观测量。由于传统处理方法十分耗时，全球 NWP 系统真正使用的卫星数据仅占其总量的 1%~3%。然而由于传感器设计技术的发展，未来的观测数据将朝更高的空间、时间和光谱分辨率发展，因此观测数据量和观测多样性快速增长的趋势将持续下去。

其次，NWP 在改进计算资源、模式初始化、次网格物理过程描述和模式输出后处理等方面的需求还未得到满足。

此外，气象预报数据的不同用户群体为满足不断扩大的应用需求，对气象产品的精确度和分辨率有越来越高的要求。

总之，面对以上挑战，需要解决的问题可以被描述为：需要在允许进行资料处理的时间窗口不断缩短的情况下，采用新的方法来充分利用越来越多的观测，使得越来越准确、对计算能力要求越来越高的资料同化和数值预报系统吸收更多来源的观测，从而改善 NWP 的产品精度。

近年来，机器学习（Machine Learning，ML）技术在多个领域的应用取得了重大突破，这些领域与卫星遥感和数值预报领域需要解决的问题具有很大相似性，例如信号和图像处理、质量控制、模式识别、数据融合、正向及反演、预测等问

题。机器学习若用于解决上述 NWP 面临的挑战，它的优势和可行性体现在以下方面：

（1）计算效率。ML 模型通常比原始的基于物理方案的模型快几个量级；

（2）精确度。对于具有代表性并且非常精确的数据集，ML 模型可以比传统的参数化方案更加精确；

（3）可迁移性。处理各种领域的先进 ML 方法可以用于尝试用于解决地球物理问题；

（4）可协同性。由于 ML 方法和传统方法的优劣不同，可以通过优化它们的组合达到最佳实践效果。

（5）灵活性。ML 方法可以适应非线性过程、包含物理约束（如平衡方程或守恒定律）、非高斯观测误差以及物理过程尚不清楚的经验数据；

（6）易于使用；ML 技术已经和现代编程语言相结合，例如基于 Python 的 Keras 和 Tensorflow 的机器学习框架，这使得机器学习方法易于在具体问题中得到实践。

1.2 论文选题及研究现状

1.2.1 论文选题

根据 GIIRS 红外高光谱数据应用时面临的诸多问题，本文选取了对红外高光谱资料进行云检测以及利用红外高光谱资料反演大气温度廓线两个方面作为研究内容；考虑到 GIIRS 每日巨大的数据量以及 1.1.5 节中机器学习方法在面临大数据挑战时的优势，本文采用了多种机器学习算法解决两个研究内容中的具体问题。下面对研究问题做出介绍：

（1） 如何高效、准确地对 GIIRS 视场进行云检测

在红外波段，当遥感仪器的视场中存在云时，由于云顶对于到达云顶部的辐射的反射，以及云底部对地表和下层大气发射辐射的吸收、散射，有云视场接收到的辐射和晴空视场接收到的辐射差异巨大。目前数值天气预报系统中的大气辐射传输模式很难精准地模拟出云中的辐射传输过程[19]，而使用晴空条件下的辐射传输模式对有云视场的辐射进行模拟在物理上是不合理的，会给数值天气预报系统引入错误信息。由于技术限制，红外高光谱仪器的水平空间分辨率往往大于 10KM，在这个水平空间分辨率下，几乎有 90%的高光谱资料视场都被云污染[20]。如果将这些有云的资料全部舍弃，一方面将会对观测资料造成巨大的浪费，另一

方面也会丢失与云相关的重要天气过程信息。为了有效利用大部分观测资料，需要首先判断像元是否被云污染，再对晴空像元和有云像元分别做处理：晴空像元可以直接进入资料同化模块进行下一步处理，有云像元则需要经过如晴空通道云检测[21]、清云[22]等处理。GIIRS 资料具有局地连续观测的特征，局地观测密度大，因此高效、准确地辨别视场是否被云污染更能满足 GIIRS 资料实时同化的需求。

（2） 如何从高光谱分辨率的 GIIRS 数据中反演得到高精度的大气温度廓线

对大气温度和湿度垂直变化的探测是气象卫星的重要贡献之一。利用卫星红外观测辐射反演大气温度的概念最早是由 King[23]提出的。Kaplan[24]指出，卫星探测到的不同光谱区域的辐射信息来自于不同高度的大气层，因此可以利用红外探测辐射反演大气中不同高度的大气状态信息。后期学者研究发现，卫星垂直探测仪探测大气结构的能力取决于光谱通道权重函数的垂直宽度、噪声水平以及独立光谱通道的数量。权重函数被定义为大气透过率在垂直方向上的导数，表示卫星在某个光谱通道接收到的辐射主要来自于大气哪个高度层的贡献，因此，权重函数分布可以确定每个光谱通道对应的探测敏感高度。传统多光谱传感器的光谱通道分辨率低，光谱通道宽度是单一大气吸收线的百倍，这些传感器探测到的辐射是很多吸收强度不同的吸收线上观测辐射的平均值，所以这些传感器探测到的宽波段辐射来自于相对较厚的大气层，反演得到的大气参数廓线精度较低，只能提供比较粗糙的大气结构信息。红外高光谱仪器以极高的光谱分辨率测量地球系统，使用成百上千个红外通道探测从近地表到平流层的辐射。由于它们的光谱分辨率小于大气吸收线的光谱间距，所以能够得到更加精密的大气结构。GIIRS 资料对局地区域的高频次观测资料能够连续反演得到连续的不同高度上的大气温度场，这对于重要天气预警具有指示作用。因此解决 GIIRS 资料反演大气温度廓线面临地诸多挑战（1.1.4 节），获取高精度大气温度廓线是一个值得研究的问题。

1.2.2 问题研究现状

（1） 云检测研究现状

根据云在可见光和近红外波段相对于陆/海面具有更高的反射率，其在红外波段的温度相对于周围晴空环境更低的物理特性，一些学者提出了基于多通道阈值判定的成像仪云检测方法[25-29]。多通道阈值云检测法具有坚实的物理背景支持，这种方法对空间分辨率较高的成像仪的云检测效果较好。但是由于需要特定的可见光/近红外光谱波段以及较高的空间分辨率的支持，多通道阈值法不能直接被红外高光谱仪器所使用。天气预报三维变分同化系统（Weather Research Forecasting

model’s 3-Dimensional Variational Data Assimilation System，WRF-3DVar）中将背景场和观测场的差值和所设固定阈值进行比较，将不满足阈值条件的视场视为有云视场，但是这种方法依赖于背景场的给定，背景场的好坏以及阈值的设定直接决定了云检测的效果。由 NWP 系统产生的背景场必然存在误差，此外也不会存在一个固定阈值适用于所有的天气场景，所以这种阈值比较法存在较大的不确定性。如何摆脱背景场的束缚，只利用观测数据进行准确而高效的云检测对于同化红外高光谱资料具有积极的意义。

在气象目标检测领域，很多学者利用机器学习方法做出了探究。这些检测方法根据输入大致可以分为两类：（1）将高分辨率的卫星图像作为输入。通过对一些经典的神经网络架构做出调整，很多学者在使用这些网络对遥感图像进行云检测的课题上取得了显著的效果[30-33]；（2）将成像仪的多个通道的观测值或者它们的组合输入到分类机器学习模型中，实现对成像仪探测像元的云检测。一些学者发现了 Logistic 回归、随机森林、极端随机树等模型在此类问题上的表现较好[34-39]；这些方法所需的训练时间短，并且各个通道或特征输入在分类过程中的重要程度可以清晰呈现出来。对于红外高光谱图像来说，若采用第一类方法可能会存在两个问题:（1）目前，经典网络架构中的输入通常是 RGB 三色图像或是灰度图像，而高光谱资料包含上百条通道，这些通道对不同的高度层次敏感，每一个通道都可以形成一幅图像。我们并不知道云出现在在哪些高度上，所以选取哪些通道作为网络的输入是难以确定的。（2）红外高光谱资料的低水平空间分辨率也会为从红外高光谱图像中检测云增加难度。

目前，对于红外高光谱资料来说，由成像仪辅助的红外高光谱云检测方法被广泛使用。AIRS 云检测结果由落在 AIRS 视场中的 1KM 的中分辨率光谱成像仪（Moderate Resolution Imaging Spectroradiometer， MODIS）云检测产品来确定[40]。Eressmaa 使用了三个标准来评估所有落入 IASI 视场中的高级甚高分辨率辐射仪（The Advanced Very High Resolution Radiometer，AVHRR）的辐射特征，只有当三个判定标准都通过时才认为 IASI 视场是无云的[41]。将由成像仪辅助的云检测方法应用于 GIIRS 时，即是使用同时搭载于 FY-4A 上的多通道扫描成像辐射计（Advanced Geosynchronous Radiation Imager， AGRI）对其进行云检测。但由于与 GIIRS 相匹配的 AGRI 的经纬度数据存储规律不明显，所以两者精确匹配的过程中会耗费大量的时间和内存空间。为了满足实时天气预报的需求，需要研究针对于 GIIRS 的更加快速、准确的云检测方法。

（2） 红外高光谱资料反演大气温度廓线研究现状

到目前为止，一些学者对于从红外高光谱观测中反演大气温度廓线这一问题已经做出了从统计学方法到物理方法的各种尝试[42]，这些方法可以分为统计回归反演法、基于变分理论的物理反演法以及神经网络反演法三类。统计回归算法计算简便且反演过程稳定，但是它不考虑大气辐射传输过程且难以处理非线性问题，反演精度较低；基于变分理论的物理反演法是对大气辐射传输过程建模后求逆，该方法建模复杂，且需要进行大量的迭代，所以计算开销大；神经网络反演法具有很强的非线性表达能力以及良好的容错能力，已经有研究表明其在多种大气参数反演试验中表现良好；但是这种方法的中间过程目前无法从物理上得到解释。近年来基于统计学的机器学习算法在研究多个领域的科学问题中取得了成功。在卫星遥感资料反演大气参数领域，相比于物理反演法，机器学习算法不需要对数据分布做出任何假设，且反演精度较好。辐射在传输过程中受到大气中、地面上多种物理化学过程的干扰，所以反演参数和卫星辐射观测之间是非线性的，有三种适用于非线性关系的机器学习方法已经在物理参数反演问题的研究中取得一些进展，它们分别是高斯过程回归[43-46]、；随机森林[47-49]）和神经网络[50-55]。对于GIIRS 仪器，有学者使用大气廓线数据集通过 RTTOV 模拟 GIIRS 观测辐射[50]，并利用神经网络建立从模拟辐射到大气温度廓线的映射模型，但是这个模型在 GIIRS 真实观测上的表现尚无法得知。

1.3 主要研究内容和论文结构

1.3.1 主要研究内容

目前对红外高光谱资料的处理和应用研究主要针对于极轨卫星平台上的红外高光谱仪器，为了充分发挥搭载于风云四号静止卫星上的 GIIRS 对固定观测区域连续探测的优势，需要充分考虑 GIIRS 的数据特点和存储方式，重新开发适用于 GIIRS 在数值天气预报中应用的关键技术。基于这一研究背景，本文的研究内容和主要创新点如下：

（1）高效、快速的云检测：考虑到资料同化系统对红外高光谱资料的预处理需求、传统方法无法高效处理海量的高维数据这两点，本文提出了基于 L_1 正则化的 Logistic 回归云检测模型。与传统的成像仪-红外高光谱匹配云检测法相比，本文提出的云检测模型在具备高准确率（>95%）的同时检测效率显著提升；对比目前三维变分同化系统中依赖于背景场的云检测方法，本文提出的方法只需将 GIIRS 观测辐射作为输入，避免了背景场不准确引起的误差。综上，本文提出的云检测

方法能够用于对 GIIRS 资料进行实时云检测。

（2）高精度大气温度廓线的反演：红外高光谱资料反演得到的大气温度廓线对于重要天气过程具有预警作用。反演问题本身的不适定性，通道噪声、通道相关性以及观测辐射和大气参数之间的非线性关系这些因素使得用红外高光谱资料反演大气温度廓线成为一个具有挑战性的问题。本文在分析光谱通道的温度雅可比和噪声水平后，选取了对不同高度温度敏感的 50 个低噪声通道作为输入特征，使用了三种非线性拟合能力强的机器学习算法作为基础模型。试验中创新性地使用香农熵采样法构造出了有代表性的样本集，反演出高精度大气温度廓线的同时显著提升了训练效率。三种机器学习模型中，基于随机森林方法的反演模型效果突出：非台风过程中，随机森林模型在 10hPa 以下的压层上均方根误差（Root Mean Squared Error，RMSE）在 1.5K 以内，偏差（Bias）在 ±1K 以内；台风过程中，10hPa 以下 RMSE 在 2K 以内，Bias 在 ±2 K 以内。

1.3.2 论文结构

本文一共分为六章，具体章节安排如下：

第一章，绪论。首先介绍了课题研究背景，总结了大气垂直探测技术的发展、GIIRS 资料的应用前景和面临的诸多问题、机器学习方法解决海量遥感资料应用的优势和可行性，基于以上背景介绍了论文选题的研究现状和研究内容。

第二章，相关技术介绍。依次介绍了机器学习基础理论和本文中使用的机器学习算法，它们分别为 Logistic 回归、人工神经网络、高斯过程回归和随机森林算法。

第三章，基于 L_1 正则化的 Logistic 回归云检测试验。依次介绍了 AGRI-GIIRS 云检测方法、机器学习云检测模型的构建过程、模型效果评估和对模型适用性的讨论。

第四章，基于机器学习方法反演大气温度廓线试验。依次介绍了样本集构建方法和香农熵采样法、机器学习反演大气温度廓线模型的构建过程、模型效果评估和对试验结果的讨论。

第五章，结论与展望。总结全文所做的研究，指出论文的不足，并讨论下一步工作的方向。

第二章　机器学习算法

机器学习方法通常可以分为有监督学习算法和无监督学习算法两类。有监督算法通过学习已有输入数据和输出数据之间的对应关系，生成映射函数，将输入映射到合适的输出。利用有监督学习算法可以解决分类和回归等问题。在本文中，把对 GIIRS 像元的云检测问题视为分类问题，把利用 GIIRS 观测反演大气温度廓线视为回归问题。下面对机器学习算法的基础理论和本文采用的几种机器学习算法原理作出介绍。

2.1 机器学习基础理论

机器学习研究的是在计算机上从数据中产生模型的算法，即学习算法[56]。我们把大量经验数据提供给学习算法，算法就能基于这些数据产生预测模型，在面对新的数据时，模型会提供给我们相应的判断。大量经验数据构成的数据集称为样本集，它是机器学习算法最重要的基础支持。在样本集中，每一个数据样本是对一个对象或者一个事件的描述，包含对象的相关特征或同时包含其相应的学习目标。机器学习的目标是从由大量经验数据构成的样本集中学习到一个映射函数$h:X \to Y$，X是样本输入特征，Y是学习目标，该映射函数h也被称为模型，确定模型的思路一般为：

（1）制作一个由大量有效数据构成的样本集；

（2）根据学习目标，设定模型的假设形式；

（3）定义衡量模型预测值和真实值差异的指标，一般将这个指标定义为损失函数 loss，记作$l(\theta)$，θ即为模型中我们需要求解的参数；

（4）指定迭代求解算法，通过迭代求解算法得到$l(\theta)$取到极小值时θ的取值；

（5）由 4）得到θ即可确定预测模型，通过分析模型在新数据集上的效果，对模型做出评估。

一般将样本集划分为训练集、验证集和测试集。模型在训练集上学习参数θ，根据模型在验证集上的效果对模型进行调整，测试集用于评判最终模型的泛化能力，不能作为调节参数或者选择特征等过程的依据。最终的目标是建立的模型不仅要在训练集和验证集上表现良好，还要在未见过的测试集上表现良好。结合样本的使用和建立模型的一般思路，图 2.1 给出了建立基于机器学习算法的预测模型的流程。当学习目标是离散变量时，学习得到分类模型；当学习目标是连续变量时，学习得到回归模型。

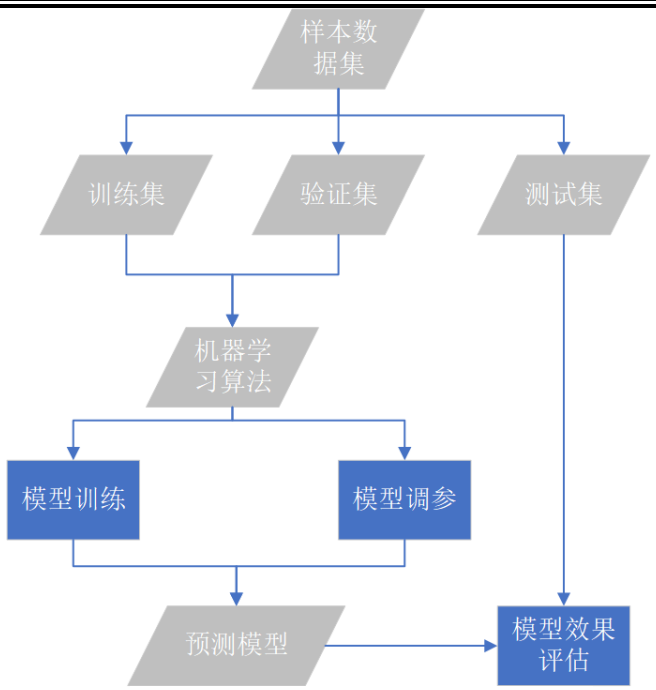


图 2.1　建立机器学习预测模型的流程

2.2　模型泛化能力

机器学习中，泛化能力是指学习到的模型对于未知数据的预测能力[57]。通常我们假设样本空间中的全体样本来自于同一个未知分布，并且每一个样本都服从独立同分布假设[58]。具有良好泛化能力的模型能够对从真实分布中任意抽取出的样本数据做出较好的预测。

过拟合或欠拟合是泛化能力差的模型中常出现的两种现象[59]。过拟合是指模型在训练集上表现很好，而在验证集上表现很差，这是因为模型过分贴合训练集，将训练集中的噪声也作为规律学习进来；欠拟合是指模型在训练集和验证集上都表现较差，这是因为模型相对于真实映射函数的复杂度较低，并没有学习到数据中的规律。导致过拟合和欠拟合的原因通常源于样本数据集的特征选取不合适、样本数量不足、或模型的复杂度过高或过低。

在训练机器学习模型时，使用过多的输入特征可能会增加训练和预测时间，有时甚至会导致过拟合，所以通常在建立模型前需要进行特征选择。特征选择是从高维的输入特征中选取对分类目标最有效的特征，在精简数据集的同时，能够提高模型的分类能力，而其过程本身并不改变特征的物理属性。特征选择具备的优点是保留下的特征一般对分类目标具有物理上的解释意义，因此在后续对分类目标的研究可以将保留的特征作为线索展开。

在机器学习理论中，对模型的选择通常会参考奥卡姆剃刀原则[60]：在模型能够较好地拟合训练数据的前提下，尽量减小模型的复杂度。即希望模型的准确度较高的同时模型不要太复杂。基于这两点的考虑，在训练模型时需要进行方差和偏差的权衡。假设参数的真实向量为θ，参数的估计向量为$\hat{\theta}$，二者的均方误差由方差和偏差两部分构成[59]:

$$\begin{aligned} MSE&=E[(\hat{\theta}-\theta)^2] \\ &=E(\hat{\theta}^2)+E(\theta^2)-2E(\theta\hat{\theta}) \\ &=[E(\hat{\theta})^2-2E(\theta\hat{\theta})+E(\theta^2)]+[E(\hat{\theta}^2)-E(\hat{\theta})^2] \\ &=[E(\hat{\theta})-E(\theta)]^2+Var(\hat{\theta}) \\ &=Bias(\hat{\theta})^2+Var(\hat{\theta}). \end{aligned} \tag{2.1}$$

模型的复杂度增大时，虽然模型的偏差会减小，但是方差会增加。为了尽可能提升模型的泛化性能，需要了解模型当前的状态，通过对模型做出调整来平衡偏差和方差，选择出最优模型。交叉验证[61]和正则化[62]是模型验证中常采用的模型选择方法。交叉验证避免了数据随机划分时的偶然误差引起的模型整体偏离，通常采用的交叉验证方法有将数据集按比例划分的简单交叉验证、K 折交叉验证和留一折交叉验证，一般取在交叉验证中准确度较高的模型。正则化方法通过在代价函数中对参数加以限制改变模型的复杂度。常见的正则化方法有 L_1 正则化和 L_2 正则化。正则化为取参数的平方，L_1 正则化为取参数的绝对值。L_1 正则化可以通过稀疏特征系数达到特征选择的目的。

2.3 模型性能评估

2.3.1 分类模型

分类模型的评估指标通常使用混淆矩阵（图 2.2）计算。通过混淆矩阵可以计算得到 5 项评估模型分类性能的指标：（1）准确率（ACC）；（2）检测率（Probability Of Detection，POD）；（3）虚警率 （False Alarm Rate， FAR）；（4）海德克技巧评分（Heidke Skill Score，HSS）；（5）ROC 曲线下方面积大小（Area Under Roc Curve，AUC）。ROC 曲线的横坐标为假阳率（False Positive Rate， FPR），纵坐标为真阳率（True Positive Rate，TPR），AUC 被定义为 ROC 曲线和 X 轴间的面积。AUC 取值在（0，1）之间，论文中用于在模型参数调节过程中评估模型的分类性能；AUC 值具有当正负样本个数不均衡时仍然保持不变的优点。当真实样本为正样本时，AUC 值越大，模型将该样本预测为正样本的概率相对于预测为

负样本的概率越大。POD，FAR 和 HSS 在论文中被用来评估最终模型的分类效果，其中 HSS 得分消除了由于随机情况而产生的正确预测的情况。

Confusion Matrix		True Label		Margin Total
		Clear	Cloud	
Predicted Label	Clear	TP	FP	TP+FP
	Cloud	FN	TN	FN+TN
Margin Total		TP+FN	FP+TN	Total

图 2.2　混淆矩阵

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)}. \tag{2.2}$$

$$POD = \frac{TN}{TN + FP}. \tag{2.3}$$

$$FAR = \frac{FP}{FP + TP}. \tag{2.4}$$

$$HSS = \frac{2(TP \times TN - FP \times FN)}{((TP + FN) \times (TN + FP) + (TP + FP) \times (TN + FN))}. \tag{2.5}$$

$$AUC = \int ROC. \tag{2.6}$$

$$FPR = \frac{FP}{FP + TN}. \tag{2.7}$$

$$TPR = \frac{TP}{TP + FN}. \tag{2.8}$$

2.3.2　回归模型

对于回归模型通常采用偏差（Bias）、均方根误差（Root Mean Squared Error，RMSE）来评估模型的预测效果。假设共有 m 个样本，真实观测值为 y_i，模型预测值为 $\hat{y}_i$，则：

$$\text{Bias} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i). \tag{2.9}$$

$$\text{RMSE}=\sqrt{\frac{1}{\text{m}}(\hat{\text{y}}_{\text{i}}\text{-y}_{\text{i}})^{2}}\ . \qquad (2.10)$$

Bias 表示预测值相对于观测偏大或偏小的程度。RMSE 表示预测值和观测值偏差的离散程度。

2.4 机器学习算法简介

论文中采用 Logistic 回归[63]模型解决云检测问题，采用人工神经网络（Artificial Neural Network，ANN）[64]、高斯过程回归（Gaussian Process Regression，GPR）[65]和随机森林（Random Forest，RF）反演大气温度廓线。本章将详细介绍以上机器学习算法的原理以及它们的优缺点和应用。

2.4.1 Logistic 回归

Logistic 回归是广义线性模型的一种，用于解决线性可分数据的分类问题，包括二分类和多分类。

以二分类为例，如果数据线性可分，则存在一条直线使得两种类别的数据分别分布在该直线两侧，该直线的表达式为：

$$y=\theta_1 x_1+\theta_2 x_2+...+\theta_i x_i+b=\theta^T X\ . \qquad (2.11)$$

其中 b 为直线的截距 $X=[1,x_1,x_2,...,x_m]$ 为模型的输入特征，$\theta^T=[b,\theta_1,\theta_2,...,\theta_m]=[\theta_0,\theta_1,\theta_2,...,\theta_m]$ 为特征系数，y 为分类结果。Logistic 回归的目标就是找到这个分类的判别函数，即求解出判别函数的特征系数。在二分类中，假设分类的可能结果为有云（$y=0$）或者晴空（$y=1$），对于一组数据中的 m 个样本，每个样本满足独立同分布假设，因此 Logistic 回归满足伯努利分布，假设伯努利分布的均值为 ϕ：

$$p(y=1)=\phi\ . \qquad (2.12)$$

$$p(y=0)=1-\phi\ . \qquad (2.13)$$

将上式改写为：

$$\begin{aligned} p(y)&=\phi^{y}(1-\phi)^{1-y}\\ &=e^{y\log(\phi)+(1-y)\log(1-\phi)}\\ &=e^{\log(1-\phi)+y\log(\frac{\phi}{1-\phi})} \end{aligned}\ . \qquad (2.14)$$

令 $\eta = \log(\frac{\phi}{1-\phi})$，则 $e^{\eta} = \frac{\phi}{1-\phi}$

$$\phi = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}. \tag{2.15}$$

令 $\eta = \theta^T X$

$$\phi = \frac{1}{1+e^{-\theta^T X}}. \tag{2.16}$$

该式表示了在给定一个样本的特征输入后，该样本为某个类别的概率。
将上式带入到 p 的伯努利分布中得到：

$$p(y \mid X, \theta) = (\frac{1}{1+e^{-\theta^T X}})^{y} \times (1 - \frac{1}{1+e^{-\theta^T X}})^{(1-y)}. \tag{2.17}$$

我们的目标是求解出参数 θ^T，对于数据集样本分布中参数的点估计一般采用最大似然法，记似然函数为 $L(\theta) = p(y \mid X, \theta)$，对其两边同时取对数得：

$$InL(\theta) = yIn(\frac{1}{1+e^{-\theta^T X}}) + (1-y)In(1 - \frac{1}{1+e^{-\theta^T X}}). \tag{2.18}$$

对上式取负数，则得到损失函数：

$$\ell(\theta) = -InL(\theta). \tag{2.19}$$

特征系数通过最小化损失函数得到：

$$\theta = \text{argmin}(\ell(\theta)). \tag{2.20}$$

Logistic 回归被广泛用于解决各个领域中的分类问题，此外 logistic 回归能通过赋予与分类目标相关性高的输入特征大的特征系数，达到筛选出重要输入特征的目的。Logisitc 回归在识别数据中的异常值[66]、选取部分气象要素预测雷暴强度[67]、检测与土地重金属污染关联性最强的人类活动[68]等研究中表现出良好的识别、检测能力。此外，与决策树和 K 临近算法相比，Logistic 回归具有更低的泛化误差，同时比支持向量机更容易建立。

2.4.2 人工神经网络

通常的人工神经网络（Artificial Neural Network，ANN）是一种按照误差从后向前反向传播算法训练的多层前馈网络，它能够在无需提供任何先验知识的情况下，学习到输入和输出之间的映射关系。ANN 通常由输入层、隐藏层和输出层构成，每一层具有多个神经元，网络中的信息按照输入层-隐藏层-输出层的顺序流动，

即上一层的输出作为下一层的输入。第$l-1$层神经元的输出y_{l-1}为第l层的输入，得到第l层神经元的输出y_l，即：

$$y_l = b_l + \sum_{i=1}^{n_l} w_i^l \times g(y_{l-1}) \tag{2.21}$$

上式中b_l为第l层输入的偏置项，w_l为第l层输入的权重；$g()$为激活函数，激活函数是非线性且可导的，正是激活函数造就了神经网络非线性拟合的能力。常用的激活函数有 sigmoid（$g_1(x)$）、tanh（$g_2(x)$）和 ReLu（$g_3(x)$）三种。Sigmoid 和 tanh 的函数图像非常相似，不同点在于 sigmoid 函数的值域在（0，1）之间，图像关于 y=0.5 对称；tanh 函数的值域在（-1，1）之间，图像关于原点对称。两个函数在 x 较大或者较小时都有梯度趋向于 0 的现象，这种现象在神经网络的反向传播过程中会引起梯度消失的问题，导致网络无法继续学习；此外，这二者由于需要进行指数运算，计算量较大。Relu 是分段线性函数，Relu 的导数清晰简单，从而避免了计算量大和梯度消失的问题。当 x>0 时，Relu 的导数始终为 1，而当 x<0 时导数为 0，在网络中意味着神经元将不再起作用，这个特点使得 Relu 具有稀疏化的能力。

$$g_1(x) = \frac{1}{1+e^{-x}} \tag{2.22}$$

$$g_2(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.23}$$

$$g_3(x) = \max(0, x) \tag{2.24}$$

神经网络的每一层相当于一个函数，多层网络连接相当于一个复合函数，构成了从输入到输出之间的非线性/线性映射。网络的目标是学习得到每一层神经元的权重和偏置。网络的学习规则采用梯度下降方法，通过误差的反向传播不断调整各层网络的各个神经元的权重和阈值，使得代价函数最小。代价函数$l(w,b)$由预测值和真实值间的差异和网络参数的正则化两部分组成，在本文中使用 RMSE 衡量预测值和真实值的差异，所以代价函数写为：

$$l(w,b) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} + \| w \|_p + \| b \|_p \tag{2.25}$$

由于激活函数的使用，神经网络具对于非线性映射函数具有较强的逼近能力。现有的研究结果显示，ANN 已在非线性强的印度夏季风降水时间序列预测[69]、城

市用水需求预测[70]、日均 $PM_{2.5}$ 浓度的预测等问题中得到良好的应用。Dorling[71]对人工神经网络在大气科学中的应用作出了综述，他指出了人工神经网络在处理非线性问题时，特别是在理论模型难以建立的情况下具有优势。ANN 的建立没有确定的规则[72]，对于不同的问题，一个较好的 ANN 是通过经验和不断试错构建得到的，但是其并不一定就是最优的；此外，ANN 的两个重要的缺点分别为“黑匣子”性质和相比于传统机器学习算法需要更大的计算代价。

2.4.3 高斯过程回归

高斯过程是有监督学习过程，可以用于解决分类和回归问题，当输入和输出均为连续变量时称为高斯过程回归。高斯过程回归（Gaussian Process Regression，GPR）通过训练数据生成数据的先验分布，从而实现对未知数据的后验分布估计，后验分布通过贝叶斯原理计算得到。高斯过程回归中采用不同核函数的组合来逼近真实数据，是一种适用于非线性回归的概率方法。下面将介绍使用高斯回归做预测的思路。

有标签已知数据集 $D=\{X,Y\mid(x_i\in\in X,y_i\in Y),i=1,2,...,m\}$，将该数据集称为训练集，其中 x_i 为 N 维输入特征向量，y_i 为该特征向量对应的观测标签。标签未知数据集 $D^*=\{X^*\mid x_j^*\in X^*,i=1,2,...,n\}$，将该数据集称为测试集。目标是对 D^* 中样本的标签做出预测。

由 D 和 D^* 的联合概率分布的计算，得到的对 D^* 中样本的估计值为：

$$Y^*=K_*K^{-1}Y. \tag{2.26}$$

Y_* 服从高斯分布：

$$P(Y^*\mid X^*,X,Y)=\mathcal{N}(K_*K^{-1}Y,K_{**}-K_*K^{-1}K_*). \tag{2.27}$$

Y 和 Y^* 的联合分布为：

$$\begin{bmatrix}Y\\Y^*\end{bmatrix}=\mathcal{N}(0,\begin{bmatrix}K & K_*\\K_* & K_{**}\end{bmatrix}). \tag{2.28}$$

其中，K、K_*、K_{**} 都是由核函数生成的协方差矩阵，K 为训练集的协方差矩阵，K_* 为训练集和测试集间的协方差矩阵，K_{**} 为测试集上的协方差矩阵。

实际上，由于实际的观测中存在噪声，会对 Y^* 的预测产生影响，假设真实标签为 $\dot{Y}$：

$$\dot{Y}=Y^*+\varepsilon. \tag{2.29}$$

ε 为高斯噪声，满足独立同分布假设，即 $\varepsilon = \mathcal{N}(0,\sigma^2)$

样本的估计值为：

$$Y^* = K_*(K+\sigma^2 I)^{-1}Y. \tag{2.30}$$

$\dot{Y}^*$ 服从高斯分布：

$$P(\dot{Y}^* \mid X^*, X, Y, \varepsilon) = \mathcal{N}(K^*(K+\sigma^2 I)^{-1}Y, K^{**} - K^*(K+\sigma^2 I)^{-1}K^*). \tag{2.31}$$

Y 和 $\dot{Y}*$ 的联合高斯分布为：

$$\begin{bmatrix} Y \\ \dot{Y}^* \end{bmatrix} = N(0, \begin{bmatrix} K+\sigma^2 I & K^* \\ K^* & K^{**} \end{bmatrix}) \quad . \tag{2.32}$$

GPR 不需要知道输入特征向量和输出之间的具体映射函数，由Y的联合分布可以看出，在 GPR 中通过输入向量的相关性来描述输出值之间的相关性。在对标签未知的数据做预测的计算中采用了协方差核函数，协方差核函数用于衡量不同样本之间的相似度或者相关程度。核函数可以将输入向量投影到高维特征空间中使得在低维空间线性不可分的数据转化为线性可分的数据，所以 GPR 可用于解决非线性问题[73]。不同的核函数适用于不同的问题，为了之后 GPR 建模的需要，在此介绍几种常用的核函数。

（1）平方指数协方差核函数（SE）

平方指数协方差核函数可表示为：

$$K_{SE}(x,x') = \sigma^2 e^{[-(\frac{d(x,x')^2}{2l^2})]}. \tag{2.33}$$

其中 $d(x,x')$ 为欧几里得距离，l 为特征尺度。SE 用来描述天气过程是过于平滑的。

（2）Matern 协方差核函数

$$K_{Matern}(x,x') = \frac{2^{1-\upsilon}}{\Gamma(\upsilon)}(\frac{\sqrt{2\nu d(x,x')}}{l})^{\upsilon} K_{\upsilon}(\frac{\sqrt{2\nu d(x,x')}}{l}). \tag{2.34}$$

其中 υ 和 l 是超参数，K_{υ} 是修正的贝塞尔函数。Matern 核函数避免了 SE 核函数过于平滑的问题，因此更适合于天气过程。

（3） 周期协方差函数

$$K_{PER}(x,x') = e^{[-\frac{2}{l^2}\sin^2(\frac{\pi d(x,x')}{p})]}. \tag{2.35}$$

其中 p 为周期性参数。

单一的核函数不能够对输入数据的多方面特征进行全面的描述，所以通常选

择多个核函数的组合构成高斯过程回归中的协方差核函数。由核函数的性质，不同核函数通过以下组合方法仍然可得到一个核函数。假设有 $K_1(x,x')$ 和 $K_2(x,x')$ 两个核函数，经过式（2.36）~（2.38）的组合和变换后得到的 $K(x,x')$ 依然为核函数。

$$K(x,x') = K_1(x,x') + K_2(x,x') . \quad (2.36)$$

$$K(x,x') = K_1(x,x') \times K_2(x,x') . \quad (2.37)$$

$$K(x,x') = \beta K_1(x,x') . \quad (2.38)$$

GPR 中不同核函数的结合使用可以刻画数据中的不同的特征，这一点相对于 ANN 来说使算法更具有解释力。但是由于需要对矩阵求逆，所以 GPR 的计算开销和内存开销较大，适合于解决小样本问题。通过结合不同尺度信息的核函数，冯淼等[74]利用 GPR 实现了低分辨率风场插值得到高分辨率的风场；使用双基地雷达资料，GPR 从稀疏采样的传播因子中预测出了海洋大气边界层内的电磁管道高度[75]。

2.4.4 随机森林

随机森林（Random Forest，RF）既可以用于回归问题也可以用于分类问题。RF 是一种集成方法，由多棵不同的决策树构成，对于一个输入向量，每棵树都会产生一个预测值；对于回归问题，采用所有树的预测值的平均作为最终的预测输出值；对于分类问题，采用投票法决定样本的最终类别。无论是应用于回归问题还是分类问题，构建 RF 的思路是相同的：

（1）利用 bootstrapping 抽样法[76]有放回地随机抽取得到 K 个样本集，由这 K 个样本集构建 K 棵决策树，每一次没有被抽到的样本组成测试集（Out Of Bag，OOB）；

（2）RF 中的任意一棵决策树是这样形成的：从输入向量的所有特征中随机选出 m 个特征，分别按照 m 个特征划分数据集，每次按照节点分裂指标（节点分裂指标一般有信息增益和基尼指数两种）的取值最小原则从 m 个特征中选取一个特征进行节点分裂，直到满足停止规则时停止分裂生长；

（3）由 1）和 2）得到 K 棵决策树，每个决策树对于输入向量得到预测值，最终一起决定最终的预测值。

RF 算法有以下两个优点：a.构建 RF 模型的过程中由于抽取样本集和特征挑选的随机性使得该算法不容易陷入过拟合，具有较好的泛化性能力；b.通过（1）中的 OOB 可以计算得到袋外误差（OOB_Error），袋外误差被证明是对模型误差的

无偏估计。对于回归模型，袋外误差可以用真实值和预测值的 RMSE（式（2.10））计算；对于分类模型，OOB_Error 用准确率（式（2.2））来计算。由于 OOB_Error 的使用，随机森林算法中不需要再单独划分数据集验证模型效果。但是当用于解决回归问题时，随机森林无法预测出训练数据范围之外的结果。

2.5 本章小结

本章对机器学习基础理论、模型泛化能力、模型性能评估以及本文中使用的几种机器学习方法的原理和优劣势进行了介绍，主要内容如下：

（1）介绍了建立机器学习的概念以及建立机器学习模型的一般流程；

（2）介绍了模型泛化能力不强的两种表现：过拟合和欠拟合，以及解决这两种问题的方法；

（3）介绍了评估分类模型和回归模型性能的多种指标；

（4）云检测问题实际为分类问题，介绍了被广泛用于解决分类问题的 Logistic 回归算法的原理。大气温度廓线反演为非线性回归问题，介绍了三种非线性逼近能力较强的机器学习算法，分别为人工神经网络、高斯过程回归、随机森林。由于这三种算法各有优劣，并没有哪一种算法在回归问题上表现出绝对的优势，所以在反演大气温度廓线的试验中分别采用三种算法各自建立了温度廓线反演模型。

第三章　基于 L1 正则化 Logistic 回归的 GIIRS 云检测算法

本章将对基于L_1正则化 Logistic 回归的 GIIRS 云检测算法进行介绍。云检测的目的是识别任意卫星探测像元是否被云污染，本章将其视为一个二分类问题，即将 GIIRS 像元分为被云污染的像元（标签值为 0）和晴空像元（标签值为 1）两类。试验中采用成像仪-高光谱仪器匹配方法（后称为 AGRI-GIIRS 匹配云检测法）制作标签样本集；在 Logistic 回归模型的代价函数中分别加入L_1和L_2正则化项进行试验，由于这两种正则化方法得到的模型在验证集上的云检测结果十分相近，最终选取了能够用于特征选择的L_1正则化Logistic回归云检测模型实现了对GIIRS高效、准确的云检测。下面将依次介绍试验中使用的数据集、AGRI-GIIRS 云检测法，Logistic 回归云检测模型的训练和优化，模型效果评估和试验结果讨论。

3.1　数据集和研究区域概况

3.1.1　GIIRS 数据

迄今为止，国家卫星气象中心公布了从 2019 年 1 月 24 日至今的 GIIRS Level 1 级辐射率资料。在试验期间，关于 GIIRS 数据的文献较少，作者通过下载和分析大量的 GIIRS 辐射率资料，较清楚地掌握了 GIIRS 数据的空间分布和光谱特征。

3.1.1.1　GIIRS 数据的空间结构

根据 GIIRS Level 1 级辐射率资料，GIIRS 数据在两个小时的观测周期内的扫描区域如图 3.1（a）所示。一个观测周期内包含 7 条扫描带（T1-T7），若非加密观测，GIIRS 每日大致在固定的时间段探测每条扫描带。每一条扫描带包含 60 个驻留点，每个驻留点包含 32×4 排列的 128 个像元（图 3.1（b）），每一列 32 个像元的扫描顺序为自北向南（图 3.1（c）），每个像元的观测视场大小为 120μm/448μrad，对应星下点地面瞬时视场大小为 16km，南北方向像元之间间隔 34μm，对应地面距离为 4.44km，东西方向像元间隔为 120μrad，对应地面距离为 16km。由图 3.2（a）所示，对于相邻的两条扫描带，上层扫描带的底层两行像元和下层扫描带的顶层两行像元视场重叠，ECMWF 关于 GIIRS 资料的报告[77]指出这些相互重叠像元的观测不连续。图 3.2（b）为长波红外第 516 个通道像元重叠处观测亮温的差值，可以看出很多重叠像元的亮温差绝对值都大于 3K，而且差值的水平空间分布没有明显的规律，因此目前无法进行有效订正。在本文的试验中，这些相互重叠的像元均被剔除。

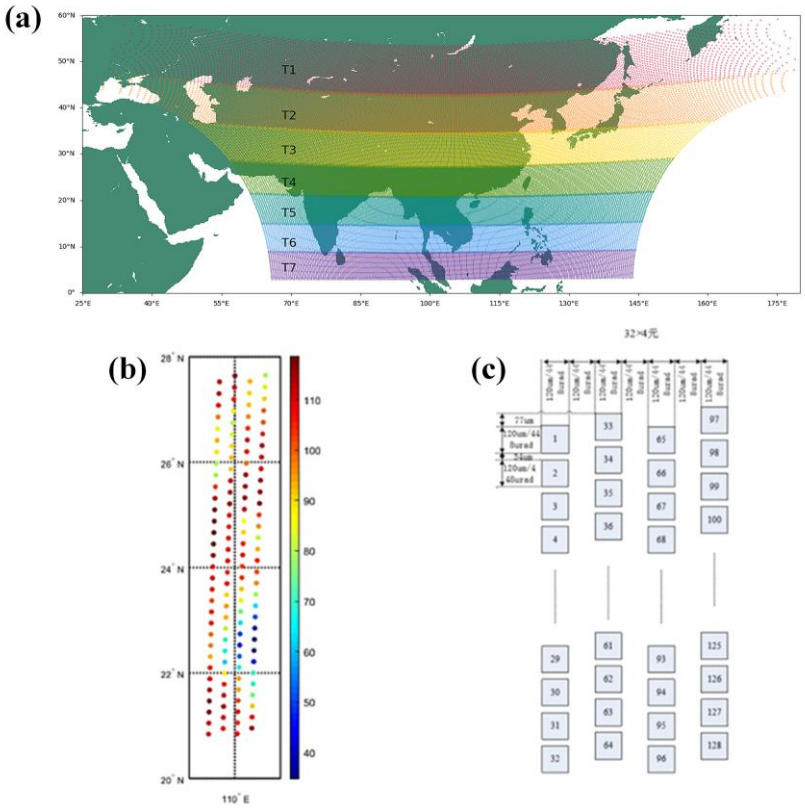
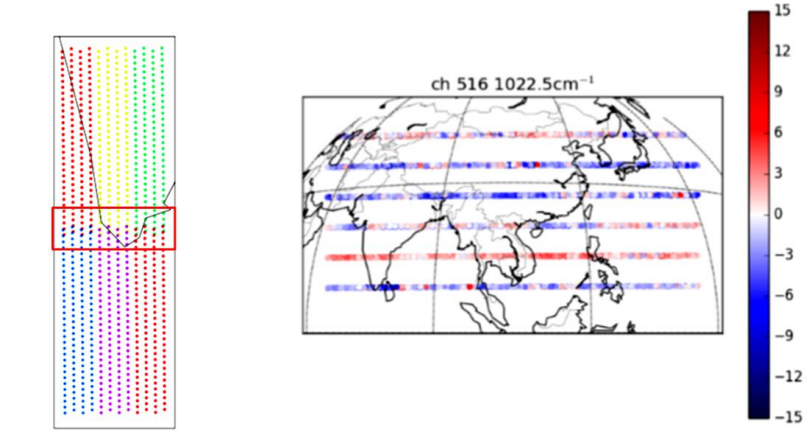


图 3.1 （a） GIIRS 扫描范围；（b）单个驻留点 128 个像元排列方式；（c）单个驻留点 128 个像元扫描顺序



(a) Co-location overlap (b) Co-located differences [K]

图 3.2 （a）相邻扫描带像元重叠示意图；（b）GIIRS 长波第 516 个光谱通道上相邻扫描带重叠区域量温差。

3.1.1.2 GIIRS 数据的光谱特性

GIIRS 具有两个红外探测波段，分别为：长波红外波段，覆盖 700cm^{-1}~1130cm^{-1}，包含 689 个通道；中波红外波段，覆盖 1650cm^{-1}~2250cm^{-1}，包含 961 个通道。二者的光谱分辨率均为 0.625cm^{-1}。其中长波红外波段包含 15μm 附近的 CO_2 吸收带，8μm~12μm 的大气红外窗区以及 9.6μm 附近的的 O_3 吸收带。中波红外波段包含 6.3μm 附近的水汽强吸收带以及 4.3μm 附近的 CO_2 吸收带。

GIIRS 采用干涉分光技术，接收到的是两束干涉光叠加的信号经傅里叶变换后得到的光谱分布，干涉信号与两束光的光程差相关，所以实际测量到的光谱分布需要考虑因光程差移动范围有限以及由仪器自身局限性引入的窗函数带来的影响。窗函数在经傅里叶变换后变为 Sinc 函数，Sinc 函数在主峰两翼有很强的旁瓣（又称为“趾”）效应，这会影响光谱分解的精度，所以通常需要对干涉得到的光谱分布做切趾处理。国家卫星气象中心发布的 GIIRS Level 1 级辐射率资料在 2019 年 8 月 13 日前未做切趾处理，称为 V1 版数据，在 2019 年 8 月 13 日后采用了 Hamming 切趾函数进行切趾处理。图 3.3（a）为切趾前的 GIIRS 的 Level 1 光谱辐射率，图 3.3（b）为切趾后的光谱辐射率，可以看出切趾后 GIIRS 辐射率随波数的变化更加平滑。

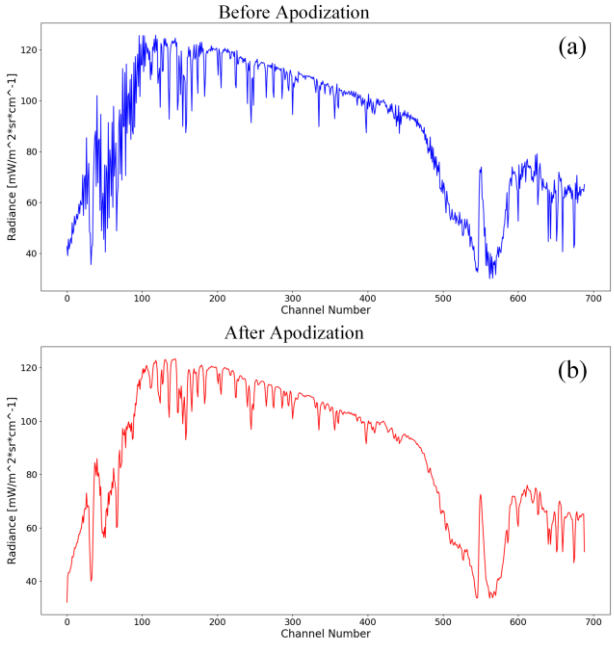


图 3.3　GIIRS 的观测辐射率。（a）切趾前；（b）切趾后

本文的试验针对 GIIRS 长波红外波段展开，图 3.4 为 689 个长波红外通道的光谱信息；其中蓝线为观测亮温，绿色点为每个通道等效噪声辐射（Noise Equivalent delta Radiance，NEdR）的平均值，浅绿色折线为 NEdR 的标准差，红色点为 NEdR

平均值大于 0.15 的通道。

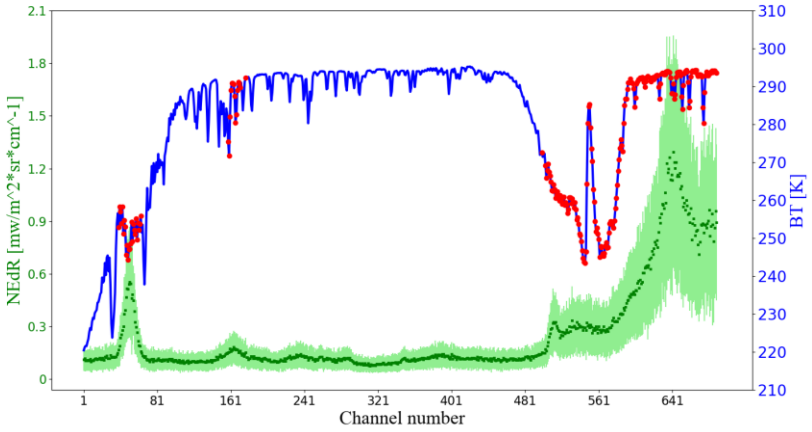


图 3.4　GIIRS689 个长波红外通道光谱信息。蓝色线为光谱辐射率，红色点为 NEdR 大于 0.15 的通道，绿色点为 NEdR 的平均值，绿色折线为 NEdR 的标准差。

3.1.2　AGRI 成像仪云掩膜产品

AGRI 是和 GIIRS 一同搭载在 FY-4A 上的成像仪，它配备了 14 个通道，包括可见光波段、近红外波段、短波红外波段、中波红外波段和长波红外波段；星下点的观测像元直径为 4KM。AGRI 的全圆盘图像不仅可以观测到大尺度天气系统的全景，还可以观测到中小尺度系统的快速变化过程。国家气象卫星中心发布了 AGRI 的 Level 2 级云掩膜产品（CloudMask，CLM）。在 CLM 中，AGRI 视场的云标签被分为四类：有云（标签值为 0），可能有云（标签值为 1），可能晴空（标签值为 2），晴空（标签值为 3）。MODIS 的云掩膜产品的精度通过和主动遥感仪器和辐射模拟值的比较得到了很好的验证，所以 MODIS 的第六版（Collection 6，C6）云检测产品通常被用来作为评估新的云检测算法准确度的基准或者是真值。闵敏等将 MODIS C6 云掩膜产品作为基准值，研究发现 AGRI 的四类云标签相对于 MODIS C6 的误差均小于 2%，说明 AGRI 的云掩膜产品的精度较高，因此本试验中使用 CLM 为机器学习云检测算法中的 GIIRS 像元定义云标签。

3.1.3　葵花 8 号云图

葵花 8 号相比于 AGRI 具有绿光通道，所以本文中选用葵花 8 号云图作为可视化验证机器学习云检测算法的辅助数据。本文使用的葵花 8 号云图分为可见光云图和红外云图两类。可见光云图使用了可见光波段中的红光（0.47μm）、绿光（0.51μm）和蓝光（0.64μm），以及近红外波段中的通道 4（0.86μm）、通道 5

（1.6μm）和通道 6（2.3μm）。红外云图使用通道 11（8.6μm）、通道 12（9.6μm）、通道 13（10.4μm）、通道 14（11.2μm）、通道 15（12.4μm）和通道 16（13.3μm）。

3.1.4 研究区域概况

试验中选取 5°N~30°N，110°E~145°E 作为研究区域（图 3.5（a））， 该区域包含在南海和西北太平洋范围内。西北太平洋是全球热带气旋（Tropical Cyclone，TC） 活动最为频繁的区域，也是唯一全年各月都有 TC 生成的海域；南海是西北太平洋热带气旋的重要生成源地之一。图 3.5（b）为日本气象厅发布的 1990 年至 2019 年在南海及西北太平洋生成的台风的路径汇总图，可以看出在经过研究区域的大多数台风具有先西北后东北的路径走向。这些走向的 TC 经常在我国沿海引起暴雨、大风和风暴潮等灾害，以及山体滑坡、泥石流等次生灾害，造成了巨大的经济损失和人员伤亡，因此有必要在同化系统中利用时间频次高的红外高光谱资料改善数值模式对该区域内重要天气过程的预报，减小灾害性天气给人民和国家带来的损失。

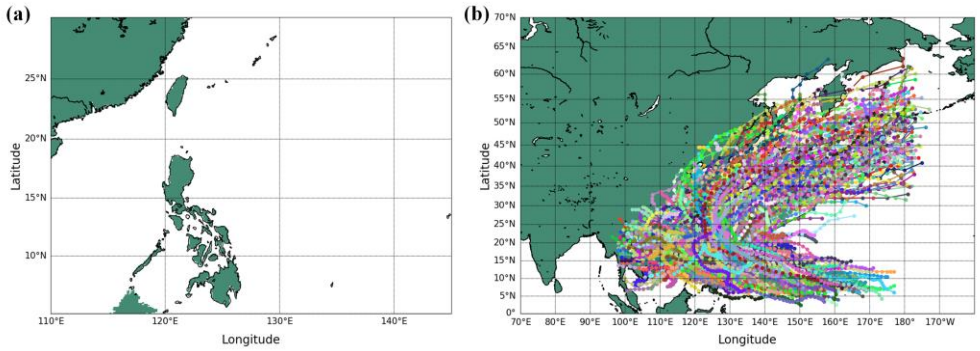


图 3.5 （a）研究区域；（b）1990-2019 年生成于南海和西北太平洋的热带气旋的路径汇总

3.2 成像仪-高光谱仪器匹配云检测算法

与 GIIRS 同时搭载在 FY-4A 上的的成像仪为 AGRI，相应的成像仪-红外高光谱匹配云检测算法为 AGRI-GIIRS 匹配云检测算法。GIIRS 的星下点分辨率为 16KM，AGRI 的星下点分辨率为 4KM，GIIRS 的云标签由落在 GIIRS 视场中的 AGRI 像元的晴空云标签的比例决定。具体流程分为以下几步：

（1）时间匹配

$$t_{GIIRS}-t_{AGRI} \le t_{max\_sec}\,. \qquad (3.1)$$

其中，t_{GIIRS} 为 GIIRS 仪器的观测时间，t_{AGRI} 为 AGRI 像元的观测时间，δ_{max_sec} 为最大观测时间间隔，设置为 600s。

（2） 空间匹配

如图 3.6 所示，GIIRS 像元的中心经纬度为（x1，y1），AGRI 像元的中心经纬度为（x2，y2），当 GIIRS 和 AGRI 的中心经纬度的距离（式（3.2））小于设定的阈值时认为 AGRI 视场落入到 GIIRS 视场中，即二者满足空间匹配。式（3.2）中，R 为地球的半径（6371KM）；若不考虑像元随着卫星观测角度变化引起的形变，那么阈值应为 GIIRS 像元的半径（8KM）。然而实际上，随着卫星扫描角增大，观测像元的视场逐渐变为椭圆，甚至趋向于难以用数学公式刻画的鸡蛋形，考虑到这一点，将阈值设置为 9KM。

$$d = 2R\sin^{-1}\sqrt{(\sin\frac{x2-x1}{2})^2+\cos x1\times\cos x2\times(\sin\frac{y2-y1}{2})^2}\,. \qquad (3.2)$$

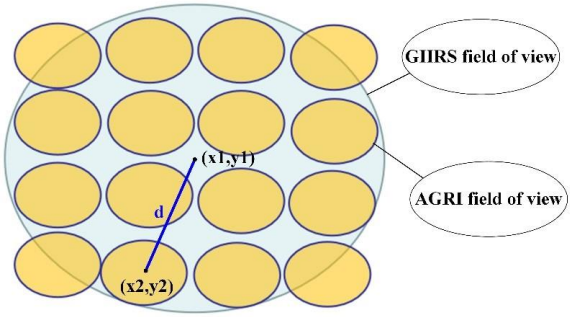


图 3.6　GIIRS 星下点像元和 AGRI 星下点像元匹配的空间示意图

（3）定义 GIIRS 像元云标签

在时间匹配和空间匹配后，从匹配结果来看，一般有 13-17 个 AGRI 像元落入到 GIIRS 像元视场内。根据 AGRI 像元中晴空像元的比例，将 GIIRS 像元的云标签定义为三类：（1）如果所有匹配上的 AGRI 像元的云标签均为晴空，则给该 GIIRS 像元赋予晴空标签（label=1）；（2）如果所有匹配上的 AGRI 像元的云标签均为有云，则给该 GIIRS 像元赋予有云标签（label=0）；（3）如果匹配上的 AGRI 像元中，既包括有云标签又包括晴空标签，则给该 GIIRS 像元赋予部分有云标签（label=2），在制作训练样本集时只使用了 label=0 和 label=1 两类数据。非以上情况的 GIIRS 像元均被剔除。

以 2019 年 5 月 15 日 03 时吕宋岛附近海域上空云图为例，根据 AGRI-GIIRS

匹配云检测算法得到的 GIIRS 云检测结果如图 3.7 所示。其中（a）为葵花 8 号的可见光云图，（b）为 AGRI 的云掩膜产品（CloudMask，CLM），（c）为 AGRI-GIIRS 匹配云检测算法定义的 GIIRS 视场的云标签，其中红色点为 GIIRS 晴空像元，蓝色点为 GIIRS 有云像元，绿色点为 GIIRS 部分有云像元。

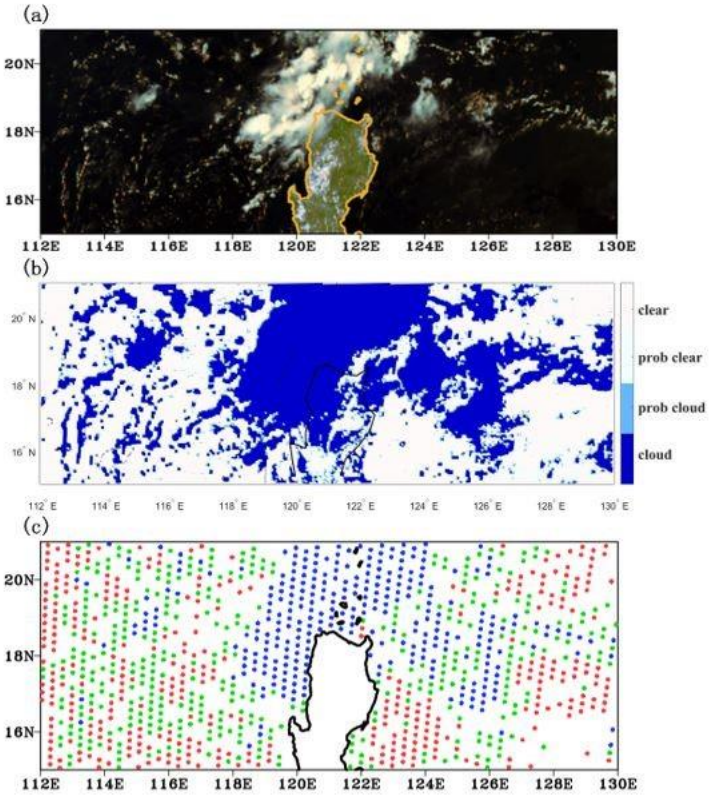


图 3.7　2019 年 5 月 15 日 03 时吕宋岛附近海域上空。（a） Himawari-8 可见光云图；（b）CLM；（c）AGRI-GIIRS 云检测算法得到的 GIIRS 云标签分布，其中红色点为 GIIRS 晴空像元，蓝色点为 GIIRS 有云像元，绿色点为 GIIRS 部分有云像元。

3.3　机器学习云检测算法

本文中将对 GIIRS 的云检测看作一个二分类问题，GIIRS 像元为有云（label=0）或 GIIRS 像元为晴空（label=1）。机器学习算法是一种受数据驱动的自学习算法，它通过对大量样本中“规律”的学习，找到 GIIRS 的观测辐射值和 GIIRS 的云标签之间的映射关系。本节利用由 3.2 节得到的带云标签的 GIIRS 观测数据训练机器学习模型，达到自动对 GIIRS 像元进行高、准确的云检测的目的。主要分为以下几个步骤：（1）Logistic 回归的代价函数定义；（2）数据预处理；（3）模型输入特征选择；（4）模型参数调节；（5）模型效果评估。

3.3.1 代价函数定义

p_i为 Logistic 回归算法将样本 i 预测为正类的概率，其中θ为判别函数系数，即机器学习算法需要学到的参数。$J(\theta)$为 Logistic 回归算法的代价函数，N 为训练样本数，代价函数取到最小值时的θ是模型学习的目标。代价方程包含两项，第一项为损失函数，第二项为正则项。在式（3.4）中 y 为样本类别标签， p 为样本被预测为 y 类的概率； C 为正则系数， C 与正则化强度成反比， C 越小正则性越强，模型越简单。对于正则项， $p=1$时为 L_1 正则化，$||\theta||_1 = |\theta|$； $p=2$ 时为 L_2 正则化，$||\theta||_2 = \sum_{k=1}^{n} \theta_k^2$，n 为输入特征的数目。$L_1$ 正则化可以通过稀疏特征达到特征选择的目的。这两种正则化方法都能尽量减小过拟合的程度。对于小样本集，L_1 正则化对应的代价函数可以通过 liblinear 方法迭代收敛，L_2 正则化的代价函数可通过 Newton-CG， lbfgs 和 liblinear 迭代方法收敛。

$$p_i = \frac{1}{1+e^{-\theta^T X_i}}. \tag{3.3}$$

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \times \log(p_i) + (1-y_i) \times \log(1-p_i)] + C\|\theta\|_p. \tag{3.4}$$

3.3.2 数据预处理

对于 Logistic 回归模型，如果采用正则化，那么必须要对输入数据做归一化处理。如果输入特征没有做归一化处理，那么在数值大小上大的观测值对应的特征会取得更大的特征系数，包含重要信息但是观测值较小的特征就会被赋予较小的特征系数，导致这些特征在判别函数中被忽视。在本研究中对每个特征通道下的观测值采取式（3.5）将观测值 x 全部变换到（0，1）区间内，其中 $x_{\max}$ 为特征通道最大值， $x_{\min}$ 为特征通道最小值。

$$\dot{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{3.5}$$

3.3.3 模型输入特征选择

GIIRS 长波红外包含 689 个通道，通道数目众多且通道间相关性高，所以韩威[82]等学者在分析 689 个光谱通道的权重函数和通道噪声后挑选出了 35 个光谱通

道，分布在前 113 个光谱通道（ 701.25cm^{-1}~770cm^{-1} ）中，这个区间的光谱通道通常被用于探测大气温度和反演云产品。目前与云相关的物理化学过程尚不完全清楚，所以所有与云相关的光谱通道是哪些没有定论。因此，本文采用了两组试验：第一组试验使用全部 698 个光谱通道作为特征输入，以下内容中将其称为 LR_689；第二组试验采用了上述 35 个通道和 3 个大气窗区通道，即共 38 个光谱通道作为输入特征，以下称为 LR_38，38 个通道的通道号和波数见附表。

3.3.4 模型参数调节

模型分类效果随训练样本数的变化能够反映出在当前训练样本数量下模型的状态（欠拟合/过拟合），为分类问题需要多少样本才算足够做出指导。只有在了解模型当前的状态后，才能够根据模型实际的情况向正确的方向对模型做出调整。试验中采用学习曲线来判断模型状态。学习曲线是模型在训练集和验证集上的 AUC 随样本数量变化的曲线。

在机器学习模型中，超参数的选取非常重要，不同的超参数可能会导致分类结果结果截然不同。在 Logistic 回归模型中，正则系数 C 是需要进行调节的超参数，采用设置不同量级的 C 值分别训练模型来选取 C 的最佳取值。

Logistic 回归分类模型通过对比预测概率（ p ）和设定的分类概率阈值（ Ω ）对每个样本进行分类。如果 $p \geq \Omega$，则视场的云标签为 1，代表晴空视场；如果 $p < \Omega$，则视场的云标签为 0，代表有云视场。尽管一般 Ω 的取值为 0.5，但它并不总是每个分类问题的最佳分类概率阈值。在本研究中，将采用混淆矩阵来指导云检测机器学习模型的分类概率阈值的选取。由混淆矩阵的组成可知，改变分类阈值将导致 TP、FN、FP 和 TN 的变化。当 TP 与 TN 越接近 1，FN 与 FP 越接近 0 时，分类阈值越合适。

3.4 机器学习云检测算法试验

3.4.1 训练样本数

通常训练样本数目应该大于输入特征数目，但是训练一个模型具体需要多少训练样本数事先是未知的。为了避免由于训练样本数不足导致的模型欠拟合，需要首先确定不再影响模型效果的最小训练样本数目。由 3.2 节 AGRI-GIIRS 匹配云检测法，我们得到了 6270 个有云样本和 6740 个晴空样本，为了避免样本不均衡对建立模型的影响，划分给训练集和验证集的有云和晴空样本数一样，剩余样本

划分到测试集。三个样本集中的样本数量如表 3.1 所示。此外，验证集上的 AUC 值由五折交叉验证得到。

表 3.1　正负样本数目

类型	训练集	验证集	测试集	总计
有云	3500	1300	1470	6270
晴空	3500	1300	1940	6740

图 3.3 为 LR_689 和 LR_38 分别采用L_1和L_2正则化时的学习曲线，红色折线为模型在训练集上的 AUC 值，绿色折线为模型在验证集上的 AUC；第一列是 LR_689 的学习曲线，第二列是 LR_38 的学习曲线；第一行使用L_2正则化，第二行使用L_1正则化。其中超参数即正则化系数 C 均设置为 1。折线为五折交叉验证的 AUC 平均值，阴影部分为 AUC 的标准差。根据图 3.8 我们可以得到以下结论：

（1）当训练样本数目大于等于 4000 时，4 个模型的 AUC 值均趋于稳定，说明训练这四个模型至少需要 4000 个样本。

（2）训练集上 LR_689（b，d）的 AUC 值总是高于验证集，说明 LR_689 模型存在过拟合。一般可以通过增加数据量、降低模型复杂度（正则性较强）或减少模型特征数来降低过拟合程度。与 LR_689 相比，当训练样本数超过 4000 个时，LR_38 模型（a，c）在训练集上的得分与在测试集上的得分基本相同，这表明与 LR_689 相比，一些输入特征被剔除后，过拟合现象消失了。另外，当训练样本从 4000 个增加到 7000 个时，LR_689 仍然存在过拟合现象，说明对于 LR_689，增加训练样本数量并不能改善该问题中的过拟合问题。在下一节中通过调整超参数“C”进一步改善过拟合问题。

（3）对于具有相同输入特征的模型，当训练集和测试集的 AUC 不随样本大小变化时，在测试集上L_1和L_2正则化的 AUC 值几乎相同。

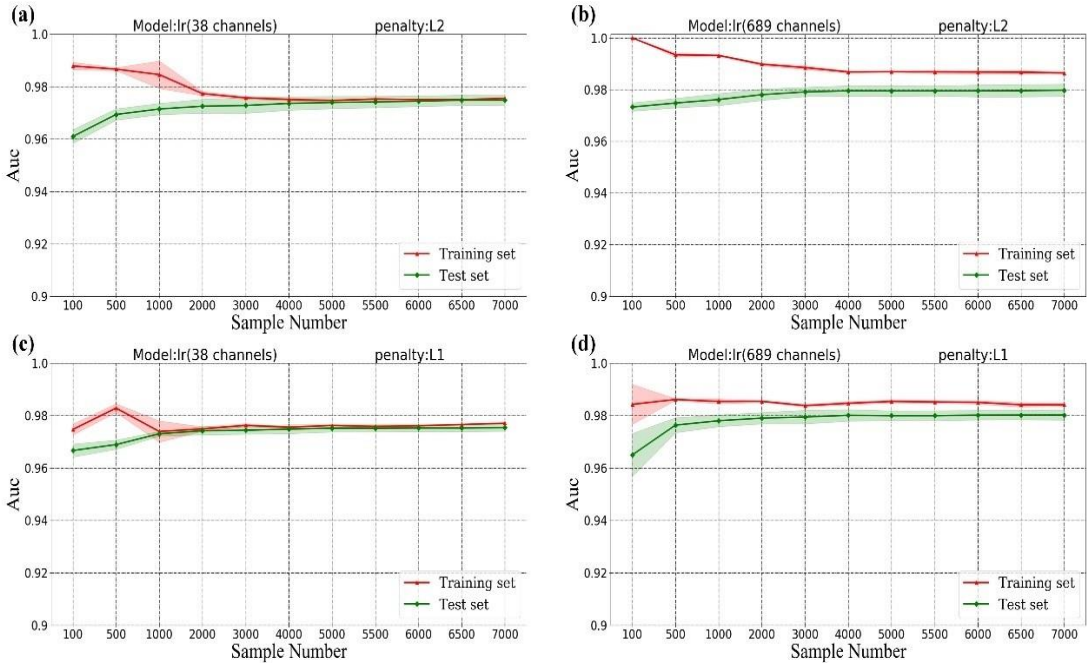


图 3.8　LR_689 和 LR_38 的学习曲线；第一行采用 L_2 正则化，第二行采用 L_1 正则化；红色曲线为训练集，绿色曲线为验证集

3.4.2　超参数调节

由 3.4.1 节知，C=1 时，LR_689 存在过拟合；LR_38 未表现出过拟合，但该模型是否存在欠拟合需要进一步讨论。模型分类效果已经不再随训练样本数目增加而变化时，则需要通过改变超参数来进一步调节模型。在本节中，观察到当超参数“C”以十倍的间隔从 0.0001 变化到 1000 时，4 个模型的 AUC 的变化（图 3.4）。

在图 3.9（a，c）中，当 C 大于 1 时，AUC 值相对于 3.8（a，c）明显提高，C 的增加表明正则化的强度降低，模型的复杂性增加，由此可以推断图 3.8（a，c）中 LR_38 存在欠拟合。虽然当 C 大于或等于 10 时，两个数据集上的 AUC 保持不变，但将“C”的最佳值选取为 10，这是因为 C 越大，正则化能力越弱，所以模型的复杂度越高，根据奥卡姆剃刀原则选择 C 值较小、复杂度低的模型。考虑到两种正则化方法在两个训练集和测试集上的得分相近，并且 L_1 正则化可以用于特征选择，在后续的实验中对于 LR_689 和 LR_38 均选择了 L_1 正则化（C=10）。

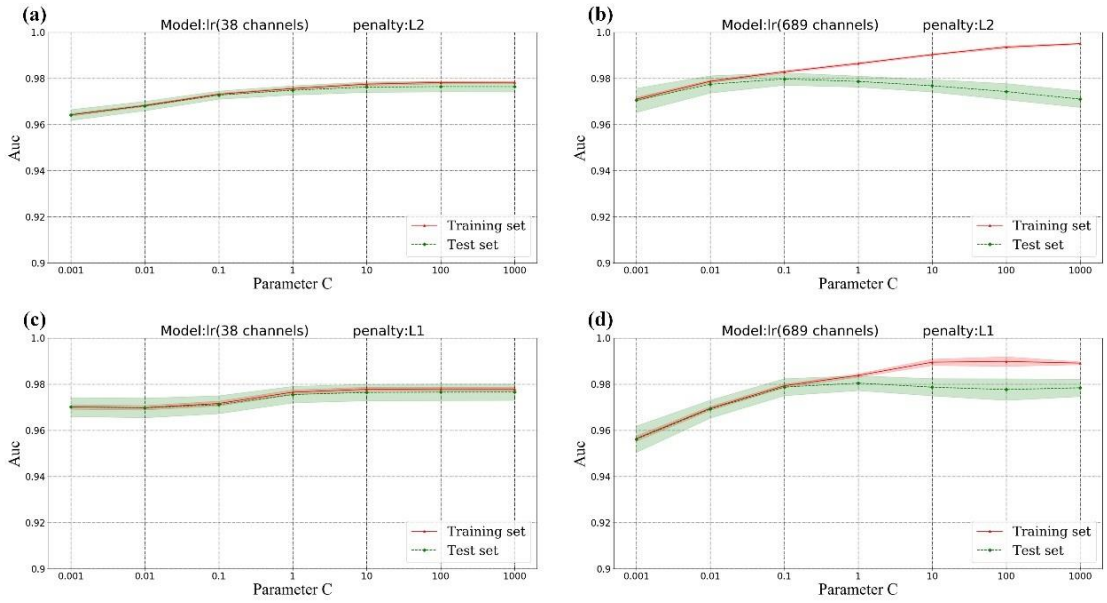


图 3.9　LR_689 和 LR_38 的 AUC 随超参数 C 的变化

3.4.3　分类概率阈值调节

以 LR_689 模型为例，图 3.10 显示了当概率阈值（Ω）从 0.1 变化到 0.9 时混淆矩阵是如何变化的。结果表明，对于 LR_689，最佳分类阈值在 0.4~0.6 之间，缩小 0.4~0.6 之间的区间可以得到更精细的分类阈值。LR_689 的最佳分类概率阈值为 0.55，LR_38 的最佳分类概率阈值为 0.5。

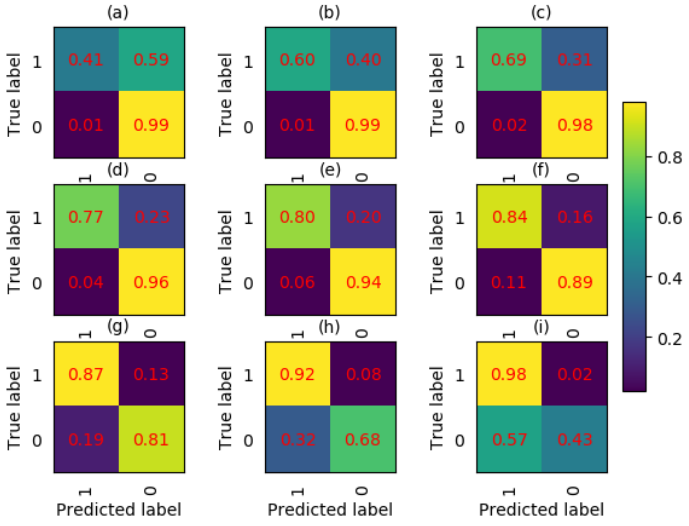


图 3.10　LR_689 的混淆矩阵随分类概率阈值Ω的变化

3.5　机器学习云检测结果

3.5.1 云检测误差统计结果

表 3.2 为 LR_689 和 LR_38 在测试集上的误差统计结果。两个云检测模型在测试集上均表现出高准确率、高检测率、低虚警率的预测结果。两个模型的准确率均大于 95%，POD 大于 98%，HSS 评分大于 0.91，虚警率小于 1%，LR_689 的预测结果略好于 LR_38。

表 3.2 两个模型在测试集上云检测统计结果

模型	POD	FAR	ACC	HSS
LR_38	0.986	0.071	0.956	0.912
LR_689	0.993	0.047	0.973	0.942

3.5.2 云检测可视化结果

3.5.1 节中的测试集是多天各个时刻的样本的混合，为了更加形象地展示 LR_689 和 LR_38 的云检测效果，本节另外选择了三个时刻的真实天空场景，把真实天空云图和云检测模型的可视化结果进行对比从而进一步验证模型的分类效果。表 3.3 列出了这三个真实场景的时间信息和云分布特征。图 3.11 为可视化结果对比，第一列为葵花 8 号云图，第二列为 LR_38 的检测结果，第三列为 LR_689 的检测结果，红色点为晴空像元、蓝色点为有云像元。以葵花 8 号云图为真实参考，两种特征输入的模型的检测结果和真实场景基本一致，LR_38 和 LR_689 的检测结果差异很小。以真实云图黄色圈内的云分布为参考，第二和第三列中红色圈内为与真实场景一致的检测结果，蓝色圈内为与真实场景相比错误的检测结果，对比结果显示模型整体检测较好，对于云片较小、分布散乱的情况分类效果相对较差。

表 3.3 三个云图场景的信息

视场号	时间	视场特点
1	2019-08-14-02:30	台风“KROSA”
2	2019-08-21-04:30	云片较小、分布较分散
3	2019-08-08-12:45	台风“LIKEMA”

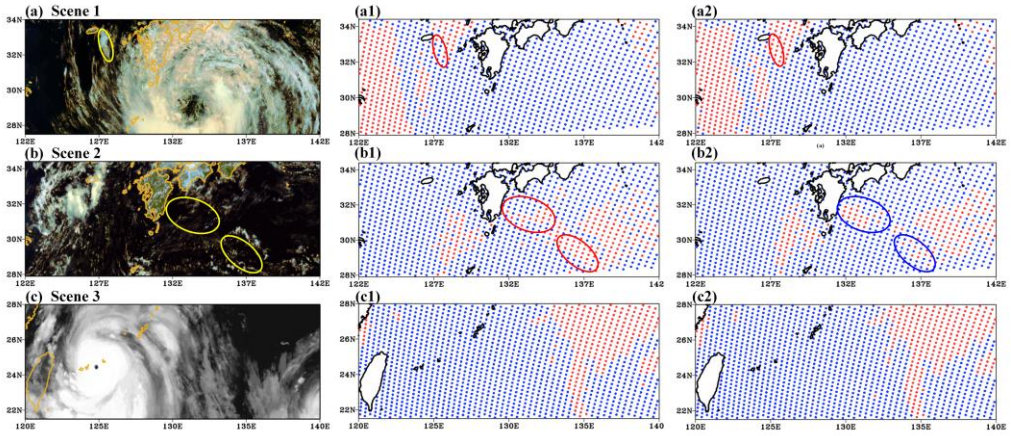


图 3.11　云检测可视化结果

3.6　试验结果讨论

3.6.1　算法适用性

AGRI 的经度和纬度数据单独存放，均为 2748×2748 的矩阵，纬度和经度并非按照在 x 轴上自西向东，在 y 轴上自南向北的顺序规律存储，导致在 AGRI-GIIRS 匹配云检测算法中，AGRI 像元和 GIIRS 像元的空间匹配过程需要遍历经纬度矩阵而消耗了大量的时间。机器学习云检测算法对于每个 GIIRS 像元的云检测结果只依赖于其自身的观测，且本文选取的 Logistic 回归算法的本质是一个线性判别式（式（2.10））， 当模型建立完成后，式（2.10）中的特征系数即被确立，因此对于每一个 GIIRS 像元的云检测的时间复杂度为$O(p)$，其中 p 为判别式中特征个数。图 3.7 为 AGRI-GIIRS 匹配云检测法、LR_689、LR_38 输入的 GIIRS 像元数目分别为 1280、1960、2560 时进行云检测所需要的时间，纵轴对运行时间取了以 10 为底的对数。从图 3.12 可以看出，在输入数据量级在10^3时，LR_689 和 LR_38 的运行时间在 1 秒以内，且运行时间随数据量变化微小；而 AGRI-GIIRS 匹配云检测法的运行时间随输入数据量的增加呈现出明显增长的态势。因此，相较于 AGRI-GIIRS 匹配云检测法，本试验中的两个模型更能满足对 GIIRS 实时观测数据进行高效准确云检测的需求。

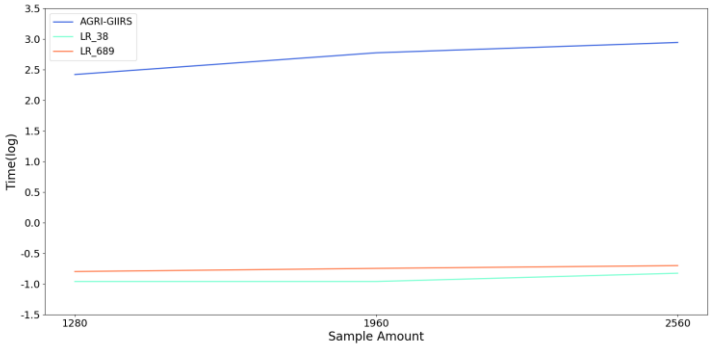


图 3.12　三个模型的运行时间随 GIIRS 像元数的变化

3.6.2　特征通道贡献

试验中选择的模型是采用了 L_1 正则化的 Logistic 回归模型。对于 Logistic 回归模型，模型建立的本质是得到式（2.10）中特征输入的系数参数，即θ^T。L_1 正则化是在代价函数中对θ^T取 1 范数，即所有系数的绝对值之和，通过 L_1 正则化使部分对于预测目标没有贡献的特征的系数趋向或等于 0，从而达到特征选择的目的。图 3.13 中蓝线为长波波段 689 个光谱通道的亮温，红色点代表 L_1 正则化中保留下的特征通道的系数，圆圈的大小代表了系数参数的绝对值。可以看出，保留的系数参数对应在探测对流层中层及以上各个高度温度的光谱通道、大气窗区波段和 O_3 强吸收波段。在温度探测波段，云的存在使得红外探测得到的亮温比晴空下的亮温低，可以为大部分温度探测通道被保留做出解释。在大气窗区，大气分子的吸收较弱，所以被用来探测地表属性和云属性特征。而在臭氧的强吸收线附近的许多通道被保留，说明云的存在和臭氧含量存在联系，以前的研究认为云的存在会影响形成臭氧的光化学反应，且云中液态水对臭氧和前提物具有吸收作用。从臭氧吸收波段对于云检测结果具有贡献研究作用来看，该波段可用于研究云对臭氧的影响。

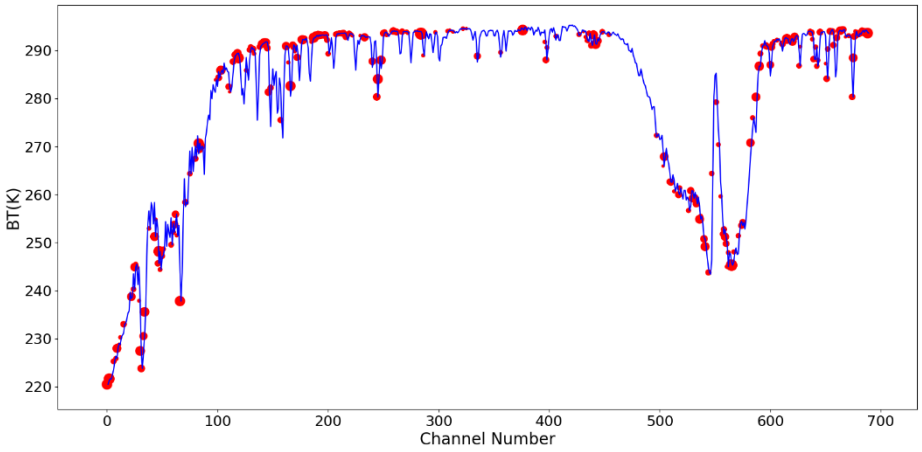


图 3.13　LR_689 的判别函数中特征通道前的系数

3.7　本章小结

本章首先介绍了 GIIRS 数据的空间和光谱特征，然后实现了传统的成像仪-高光谱云检测方法，将该方法得到的带云标签的 GIIRS 数据作为样本集，分别建立了不同特征输入的 Logistic 云检测模型，在试验过程中得到如下结论：

（1）当训练数据足够时，在学习曲线上，LR_689 出现了明显的过拟合-训练集上的准确率始终大于验证集，而 LR_38 在训练集和测试集上的得分相同，说明选取部分通道特征相对于选取全部通道特征更容易避免过拟合；

（2）训练过程中，根据学习曲线判断出模型存在的问题，进而调节模型中的超参数正则系数达到了降低 LR_689 模型复杂度减弱其过拟合的问题，增大 LR_38 模型复杂度解决了其欠拟合的问题；

（3）LR_689 和 LR_38 在测试集上均取得了良好的云检测结果，说明只依赖于 GIIRS 观测的机器学习云检测算法能够取得高效准确的云检测结果，同时避免了传统的阈值检测法由于分类阈值和背景场的不确定性引入的误差。

（4）L_1 正则化可以使对预测目标没有贡献的特征的系数参数趋近或等于 0 达到特征选择的目的。分析保留下的光谱通道的物理探测特性和预测目标的关系，可以为以后的研究工作提供指导。

第四章　大气温度廓线反演

大气温度廓线为某一经纬度点上大气温度在垂直方向上的分布。温度和卫星辐射观测值均为连续变量，因此利用 GIIRS 观测资料反演大气温度廓线为回归问题。使用第三章建立的云检测模型筛选出晴空 GIIRS 像元，将 ERA5 的大气温度廓线（Temp_ERA5）视作真值，将晴空 ERA5 格点和晴空 GIIRS 长波像元（LW_GIIRS）进行时空匹配后可得到 GIIRS 像元和大气温度廓线样本对集。将该样本集分为训练集、验证集和测试集，训练集和验证集用于训练机器学习模型以及调节模型参数，测试集用于评估最终反演大气温度廓线的模型（Temp_Retrieval_Model）的效果。此外，还将 Temp_Retrrieval_Model 预测的大气温度廓线（Temp_Pred）和 AIRS 的 Level 2 级大气廓线产品（Temp_AIRS）进行对比对模型效果做出了分析评估。

4.1　数据集说明

4.1.1　欧洲中期数值预报中心第五代再分析资料

2019 年 3 月，ECMWF 发布了其第五代再分析资料（ERA5）[78]，提供了自 1979 年以来描述全球天气和气候变化的相关产品。ERA5 是由 IFS-Cy41r2 的先进地球系统模式和同化系统通过同化大量历史常规和非常规观测产生的，它提供了水平方向上 30KM 分辨率的逐小时大气、陆地和海洋变量产品，分为垂直方向上 37 层和 137 层两种类型。本试验采用了 ERA5 的 37 层大气温度廓线产品（下面称为 Temp_ERA5），其气压层范围在 1100hPa~1hPa 内。本试验中，将 Temp_ERA5 作为采用机器学习方法反演大气温度廓线的标签数据。

4.1.2　AIRS Level 2 级标准反演产品-大气温度廓线

AIRS 的标准反演产品是全天候的，即包括了晴空和有云的情况。其大气温度廓线产品（下面称为 Temp_AIRS）包含了从 1000hPa 到 0.1hPa 的 27 个气压层上的温度。在 AIRS 的大气温度廓线反演算法文档中指出，虽然 $4.3\mu m$ 和 $15\mu m$ 附近为 CO2 的强吸收带，探测辐射中携带了从对流层到平流层的温度信息，但是考虑到部分位于 $15\mu m$ 附近的长波通道容易受到云的影响，在大气温度反演的过程中，最终只保留了 $15\mu m$ 附近对云不敏感的 53 个平流层探测通道（$662cm^{-1} \sim 713cm^{-1}$），此外还选取了探测对流层和平流层温度的 30 个通道

（$2358cm^{-1} \sim 2418cm^{-1}$）。本试验中，采用 Temp_AIRS 评估了 ERA5 的温度廓线产品的精度。

4.2 构建样本集

4.2.1 训练样本集

本研究中将 37 层的 Temp_ERA5 作为真值，通过 GIIRS 像元和 Temp_ERA5 的时空匹配得到（LW_GIIRS，Tenp_ERA5）样本对集。从空间上来看，Temp_ERA5 为网格点数据，格点分辨率为0.25°，约 30KM；GIIRS 视场在星下点的分辨率为 16KM，当 GIIRS 中心经纬度和 Temp_ERA5 格点在球面上的距离（d）最近且小于 8KM 时认为二者在空间上匹配。从时间上来看，Temp_ERA5 提供全球逐小时数据；我们研究的区域涵盖在 GIIRS 的三条扫描带中（对应图 3.1（a）的 T4-T6），三条扫描带对应的扫描时间如图 4.1 所示，其中 H 为每日的奇数时。在本研究中假设对于晴空视场，半小时内大气的状态变化极小，则认为 H 时的 Temp_ERA5 和 T4-T5 内 GIIRS 的观测在时间上是匹配的。为了尽量满足上述假设，在实际匹配过程中要求空间上匹配的 GIIRS 视场和 ERA5 格点需要同时满足晴空条件。采用第三章的 LR_689 模型对 GIIRS 像元进行云检测；ERA5 的大气产品中包含格点云占比产品（CLD_FRAC），本试验中将 CLD_FRAC<0.2 的 ERA5 格点定义为晴空。

采用 2019 年 5-8 月的 GIIRS 数据和 ERA5 数据构造样本对集，在此期间共有两个台风经过研究区域，10 号台风“利奇马”于 2019 年 8 月 4 日被定级为强热带风暴，8 月 12 日衰减为热带低压。11 号台风“罗莎”于 2019 年 8 月 6 日被定级为强热带风暴，8 月 16 日衰减为温带气旋。本试验对于台风过程和非台风过程分别建模。使用 5~7 月的样本作为非台风数据集，训练集和验证集来自于 5-7 月的 1，5、9、13、21，25 日的每日 3、7、11、15、19、23 时，测试集来自于 5-7 月 15 日的 3、11、19 时。使用 8 月的样本作为台风数据集，将 8 月份中 8 月 4 日前和 8 月 16 日后研究区域内的晴空样本作为训练样本和验证样本，将 8 月 4 日-8 月 16 日的样本作为测试数据。表 4.1 列出了非台风和台风过程中的样本数量。

表 4.1 非台风样本集和台风样本集中的样本数量

类型	训练集	验证集	测试集
非台风样本集	6690	3295	1820
台风样本集	3159	1579	481

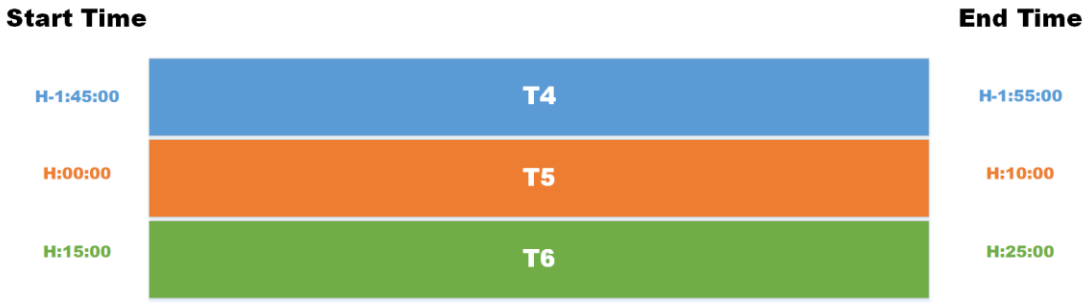


图 4.1　研究区域对应的 GIIRS 观测起始和终止时间

4.2.2　均衡数据集采样

本节将对 4.2.1 节生成的（LW_GIIRS，Temp_ERA5）样本对集进行挑选，获取样本间具有差异性、样本具有代表性的均衡数据集，这样做是出于以下考虑：

（1）由于大多数情况下，在某一时刻的天空中，晴空区域和有云区域总是成片出现，所以经空间匹配得到的 GIIRS 晴空像元往往是相邻的像元，它们携带的大气探测信息非常相似；另还考虑到这样的情况，尽管为了保证样本集能够包含大气的日变化信息，在 AGRI-GIIRS 时间匹配时选取了一日多个时刻进行采样，但是对于有些时刻研究区域对应的天空云覆盖面积很大，因此时空匹配后保留的样本点寥寥无几，而对于另外一些时刻研究区域对应的天空晴空面积很大，保留下的样本点非常多。以上两种情况造成了很多样本相似性很大，以及从时间上来看样本不均衡的问题。机器学习是受训练数据驱动的算法，如果训练数据所代表的情况过于局限，就会造成训练得到的模型在训练数据集上表现非常好，而在未见过的数据上表现很差的问题。

（2）一些机器学习算法具有牢靠的数学理论支持，因此鲁棒性很强，但是计算的时间和空间开销随样本数量增加剧烈。例如通过计算样本间距离衡量样本相似度的 K-临近算法、需要对协方差矩阵求逆的高斯过程回归算法等。对于这些机器学习算法，除了通过对算法的改进减小计算代价，还可以通过构建具有代表性的小样本集训练模型，在减少样本量同时达到学习的目标。

（3）GIIRS 具有局地高频次周期性观测的特点，每日在固定观测区域的观测起止时间变化很小。这意味着 GIIRS 数据包含很强的局地特征和日变化特征，因此可以从中提取出局地代表性信息。

将 4.2.1 节得到的样本集称为原始样本集 S，如何从原始样本集采样得到样本间具有差异，同时样本分布比较均匀的代表性样本集是接下来要解决的问题。

4.2.2.1　最大香农熵采样原理

香农熵被用来衡量系统的无序程度，也是对系统所包含的信息的一种度量，香农熵越大系统的无序程度越高，系统所包含的信息越多。假设有长度为 m 的随机离散变量$X=(x_1,x_2,\dots,x_m)$和一个区间变量$I=(b_1,b_2,\dots,b_{n+1})$，这个长为$(n+1)$的区间变量可以构成$n$个区间（称为$bin$），$BIN=(bin_1,bin_2,\dots,bin_n)$，其中$bin_i=[b_i,b_{i+1})$。通过变量值落在每个$bin$的占比计算得到$X$的香农熵，其中$Number_{x_i\in bin_j}$指$X$中落入区间$bin_j$中的变量值的个数。

$$p_j=\frac{Number_{x_i\in bin_j}}{m}, i=1,2\dots m; j=1,2\dots n \quad . \tag{4.1}$$

$$E(X)=\sum_{j=1}^{n}-[p_j\times\log(p_j)] \quad . \tag{4.2}$$

下面以一个简单的例子介绍最大化香农熵如何帮助我们构造出有代表性的数据集。

假设有 4 个温度数据集D_1~D_4，每个数据集中包含 10 个温度样本，假设温度的最大值为 300K，最小值为 275K，将最大值和最小值等区间分成 5 个bin，$BIN=([275,280),[280,285),[285,290),[290,295),[295,300])$则根据式（4.1）和式（4.2）可以计算出$D_1$~$D_4$每个数据集在每个$bin$的样本占比和最终的香农熵，如表 4.2 所示。在四个数据集中，D_1中所有样本值相同，香农熵为 0；从D_2~D_4，随着数据集中样本间的差异加大以及样本在最大最小值内分布得越来越均匀，计算得到的香农熵也越来越大。D_4中的样本差异最大，并且样本均匀分布在BIN中，计算得到的香农熵最大。

$$D_1=\{275;275;275;275;275;275;275;275;275;275\}$$
$$D_2=\{275;277;276;279;278;297;295;299;296;300\}$$
$$D_3=\{279;287;299;294;300;299;282;277;282;275\}$$
$$D_4=\{282;287;296;282;276;289;297;275;294;292\}$$

表 4.2　D_1 ~ D_4的概率分布和香农熵

数据集	p_1	p_2	p_3	p_4	p_5	香农熵
D_1	1.0	0	0	0	0	0

D_2	0.5	0	0	0	0.5	0.35
D_3	0.3	0.2	0.1	0.1	0.3	0.65
D_4	0.2	0.2	0.2	0.2	0.2	0.70

4.2.2.2 基于最大香农熵对训练数据集采样

由于温度廓线包含 37 层，所以对一条廓线进行采样时需要同时考虑 37 层上的温度，因此将对温度廓线样本采样的香农熵计算公式定义为：

$$E(B) = -\sum_{l=1}^{L_T}\sum_{n=1}^{N} p_n^{\,l} \times \log(p_n^{\,l}) \quad . \tag{4.3}$$

$$p_n^l = \frac{Number_{\mathbf{x}_l \in bin_{l,n}}}{m}, \quad l = 1,2,\ldots L_T;\ n = 1,2,\ldots,\mathrm{N}. \tag{4.4}$$

其中B为最终采样得到的新廓线集；$L_T = 37$，为温度廓线层数；$N = 25$，为BIN的长度，每一个大气层上的BIN由该层的所有温度样本的最大值和最小值等区间划分成 25 个bin得到；p_n^l为第l层上落在第 n 个bin的样本占比。

假设目标廓线集包含的样本数量为 M，香农熵取得最大值时的廓线集即为我们所求的代表性样本集 B，B 可以通过迭代优化得到，迭代过程如下：将 S 中的样本顺序打乱，从中随机选取 M 个样本作为初始廓线集 U_0 并计算熵$E(U_0)$，剩下的廓线集称为 R。在第 t 次迭代过程中，从 R 中无放回抽取一条廓线和 U_t 中的每一条廓线做交换，若当前与 U_t 中第 k 条廓线做交换，计算交换后的熵$E(U_t^k)$，若存在一次交换使得$\Delta_k = E(U_t^k) - E(U_{t-1}) > 0$，并且$\Delta_k = (\max(\Delta_i), i = 1,2,\ldots,2000)$，则保留交换后的数据集，并记$E(U_t) = E(U_t^k)$。遍历 R 中每一条廓线循环进行以上过程，最终得到新廓线集 B。

基于香农熵采样方法对非台风过程样本集采样 2000 个样本（M=2000），对台风过程样本集采样 1500 个样本（M=1500），采样得到的小样本集称为 B，原始大样本集称为 S。对于非台风过程， S 的香农熵为 90.08， B 的香农熵为 97.39；对于台风过程，S 的香农熵为 96.31，B 的香农熵为 101.19。说明相对于 S，B 中样本分布地更加均匀，样本间的平均差异程度更大。

4.2.3 温度探测特征通道选取

卫星对三维大气温度和湿度的探测依靠吸收特性不同的大气分子吸收光谱通

道。强吸收通道探测来自大气高层的辐射，而弱吸收通道探测来自大气低层和地球表面的辐射。由于CO_2几乎均匀地混合在大气中，所以CO_2在红外光谱波段的吸收带可以提供全球大气温度垂直分布的信息；H_2O红外吸收带可以提供有关大气湿度分布的信息。红外光谱的振动吸收中心分别为：4.3μm（2310 cm^{-1}）15.5μm（660 cm^{-1}）的CO_2吸收中心，6.7μm 和 12.7μm 的水汽吸收中心以及 9.7μm（1040 cm^{-1}） 的 O_3 吸收中心。此外，红外波段中存在一些波段，大气分子在这些波段对红外辐射并不敏感，即这些波段的光谱辐射几乎不会被大气分子吸收，除非遇到云、雨等的作用。由于这种特性，这些通道被称作大气窗区通道，在无云的时候，大气窗区的观测值最能确切地反映当前视场的近地表情况。

GIIRS 具有 689 个长波红外通道，覆盖 700cm^{-1}~1130cm^{-1} 波段，其中 700cm^{-1}~790cm^{-1} 位于探测大气温度的 CO_2 强吸收波段；790cm^{-1}~1130cm^{-1} 波段的辐射观测资料在晴空条件下携带了近地表大气的信息，在该波段内，同时也包含了 O_3 的强吸收波段（1041cm^{-1} 和 1111cm^{-1} 附近）。

考虑到 GIIRS 长波 689 个通道的高相关性，本文将从 689 个通道中选取部分通道作为机器学习反演大气温度廓线模型的输入特征。选取哪些光谱通道作为机器算法的特征输入主要从通道噪声和通道的温度雅克比曲线两方面考虑，其中温度雅可比定义为，给予每一个气压层相同的气温扰动时光谱通道辐射相应的变化值，它反映了各个探测通道对于哪个高度层的大气温度敏感，本文的温度雅克比曲线使用 RTTOV12.1 对 6 条标准大气廓线模拟光谱通道辐射后计算平均值得到。具体的通道挑选方案如下：（1）剔除光谱通道等效噪声辐射大于1.5mW / m^2•sr•cm^{-1}的通道；（2）剔除温度雅克比曲线为双峰的光谱通道；（3） 剔除温度雅克比存在负值的光谱通道；（4） 所选的所有光谱通道在垂直方向的探测特征和原始所有光谱通道相似。

基于以上筛选条件最终保留了 50 个光谱通道用于反演大气温度廓线。图 4.2（a）为长波 689 个通道的温度雅可比曲线，（b）所选 50 个通道的温度雅可比曲线。可以看出（a）中有很多通道的温度雅克比曲线存在大全面积重叠，说明它们探测到的信息高度相似。（b）中本文所选通道包含了对各个高度敏感的光谱通道，可以看出（b）中保留下的通道的温度雅克比曲线峰值覆盖了对流层低层到平流层中层，大多数通道的探测高度不同，且保留下的通道的峰值整体分布和（a）相似，达到了降低输入特征相似度且保留所需要信息的目的。

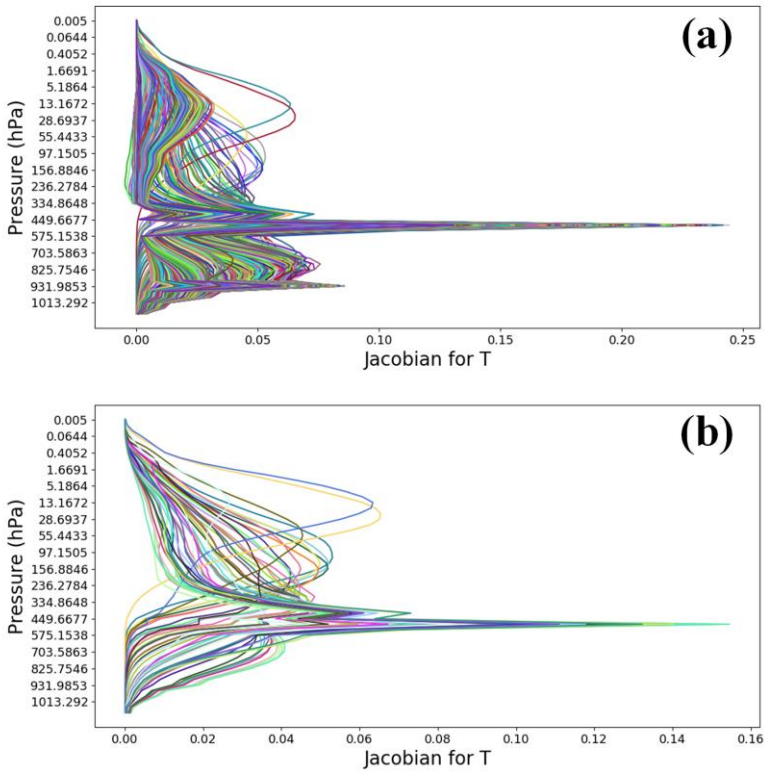


图 4.2　GIIRS 光谱通道的温度雅克比曲线；（a）689 个长波红外通道；（b）保留的 50 个长波红外通道

4.3　机器学习算法设置

选取 ANN、GPR 和 RF 对台风过程和非台风过程分别建立模型，本节对于三个模型的构建的细节分别做出说明。

4.3.1　人工神经网络

（1）模型架构

ANN 由四层构成，第一层为输入层，中间两层为隐含层，最后一层为输出层，隐含层中的激活函数采用 Relu 函数，第二个隐含层的输出和最后的输出层之间的映射采用线性函数。

（2）代价函数

在 ANN 中，将代价函数定义为验证集上 37 层的平均均方根误差（MRMSE），即$MRMSE = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{N}\sum_{k=1}^{K}(y_k^l - \widehat{y_k^l})}$，其中 L=37，K 为样本数，$y$为真值，$\hat{y}$为预

测值，在 GPR 和 RF 也采用了 MRMSE 作为代价函数。

（3）迭代设置

在神经网络中，梯度更新一次称为一次迭代，一次迭代采用 500 个样本；当神经网络遍历完所有样本，称为一个 epoch，当连续的 1500 个 epoch 中 MRMSE 的增量小于 0.1 时训练停止。

4.3.2 高斯过程回归

GPR 中采用了 Matern 核、周期核以及一个白噪声核，具体如下：

$$k = k_1 + k_2 + k_3 \tag{4.5}$$

$$k_1(x,x') = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2vd(x,x')}}{l}\right)^{v} K_v\left(\frac{\sqrt{2vd(x,x')}}{l}\right) \tag{4.6}$$

$$k_2(x,x') = e^{\left[-\frac{2}{l^2}sin^2\left(\frac{\pi d(x,x')}{p}\right)\right]} \tag{4.7}$$

$$k_3(x,x') = noise_{level}, if\ x = x^1\ else\ 0 \tag{4.8}$$

其中，$v = \frac{3}{2}$，$d(x,x')$为欧几里得距离，$\Gamma(v)$为 gamma 函数，K_v为 Bessel 函数，p为周期性参数，l为特征尺度参数。

Matern 核用于描述由于天气过程引起的气温的异常变化，l取 0.1；周期核用于气温日变化特征，p取 24，l取 1；白噪声核用于描述观测辐射中的噪声水平，其中$noise_level$为噪声方差，这里取 0.15。

4.3.3 随机森林

RF 中有多个超参数需要确定，分别为决策树的个数（n_estimators），最大搜索深度（max_depth），单个树允许采用的最大特征数（max_features），单个树最大深度（max_depth），叶子节点最小样本数（min_samples_leaf），内部节点再分裂所需的最少样本数（min_samples_split）。本文采用网格搜索方法（GridSearch）寻找以上几个参数的最佳组合。网格搜索法即给定每个参数的区间，遍历所有参数组合，搜索得到最佳参数组合，表 4.3 为以模型在验证集上的 MRMSE 最小为评

估指标得到的最佳参数组合。

表 4.3　RF 的最优参数组合

n_estimators	max_features	max_depth	min_sample_leaf	min_samples_split
130	30	10	1	2

4.4　机器学习反演大气温度廓线结果

训练集和验证集用于训练和调整模型，测试集用于评估模型最终的效果，所以本节所有的结果均为机器学习模型在测试集上的表现。以下将分别介绍台风过程和非台风过程下大气温度廓线反演结果。一共有 6 个模型，把用原始数据集 S 训练出的三个模型分别称为 GPR_S、RF_S 和 ANN_S，把使用香农熵采样后得到的新数据 B 训练出的三个模型分别称为 GPR_B、RF_B 和 ANN_B。

4.4.1　非台风过程

本节将 Temp_ERA5 作为基准值，评估了非台风过程时反演温度（Temp_Pred）的精度。图 4.3 第一行是用 S 训练的三个模型在测试集上反演结果的 Bias 和 RMSE，其中 Bias=Temp_Pred-Temp_ERA5；第二行是用 B 训练的模型的误差统计结果。图 4.3 所反映的结果和可得的结论如下：

（1）GPR 和 RF 的反演结果相比 ANN 更好。除了 70hPa 上反演误差突增外，10hPa 以下三个模型的 RMSE 和 Bias 均较小且比较稳定，在 10hPa 以上 RMSE 和 Bias 变动幅度较大。

（2）GPR 和 RF 在 RMSE 和 Bias 上的表现十分一致，且由 S 得到的模型和由 B 得到的模型效果相近。在 70hPa 上 Bias 和 RMSE 都出现明显的突增，70hPa 以下，GPR_S、GPR_B、RF_B，RF_S 的 Bias 在 ±1K，RMSE 在 1.3K 以内。说明基于 RF 和 GPR 的反演模型能够在具有代表性的小样本集上获得和不经过筛选的大样本集上十分相近的反演效果。

（3）ANN_S 的 Bias 和 RMSE 和 GPR_S 以及 RF_S 基本一致，但在 775hPa 以下 ANN_S 的 RMSE 和 GPR_S 以及 RF_S 相差较大。与之相对的 ANN_B 相对于 Temp_ERA5 差异很大。

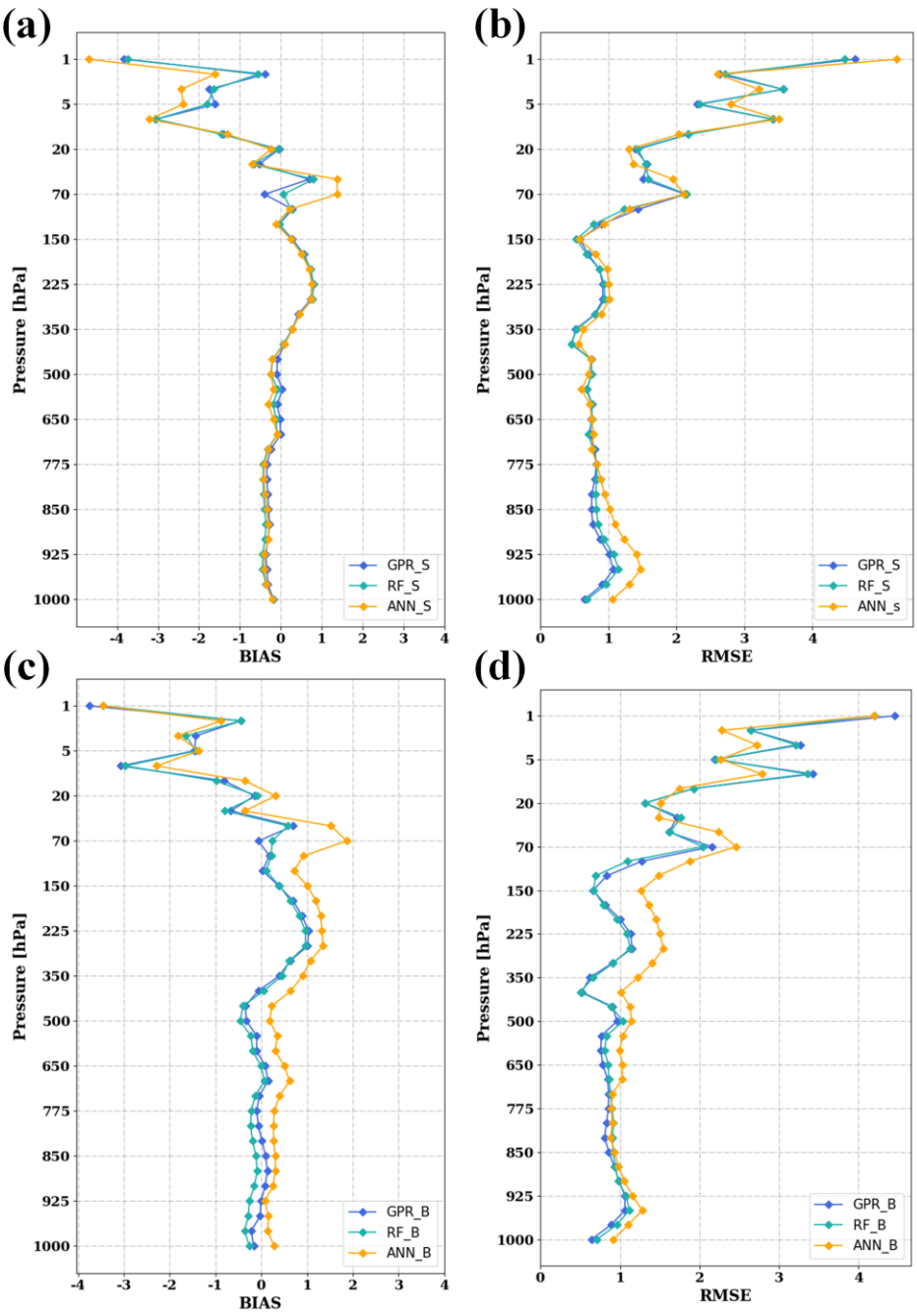


图 4.3 非台风过程模型误差统计结果。第一行的三个机器学习反演模型由 S 训练得到，第二行的三个模型由 B 训练得到。

4.4.2 台风过程

图 4.4 为台风过程的六个模型的反演结果，图片布局和图 4.3 相同。图 4.4 所反映的结果和可得的结论如下：

（1）相对于非台风过程，台风过程中的六个模型的反演效果均有所降低，Bias 和 RMSE 从 1000hPa 到 1hPa 波动均较大。尤其需要指出的是非台风过程中，500hPa~100hPa 这个高度区间的 Bias 为 0.8K 以内的正偏差；而在台风过程中，这个高度区间 Bias 随着高度先出现明显的负偏差后出现明显的正偏差。此外，这个高度区间内的 RMSE 在 1K~2K 之间，较非台风过程有明显增大。台风过程反演精度明显降低主要原因可能是台风过程中研究区域内绝大多数 GIIRS 视场都被云污染，晴空视场十分稀少，导致台风过程的训练样本较少。

（2）GPR_S、GPR_B、RF_S 和 RF_B 统计结果相近，在 125hPa 以下的 RMSE 均为 1.5K 及以下。在对流层中低层 Bias 呈现出负偏差特征，在对流层中高层呈现出正偏差特征，偏差在 ±1.5K 以内。

（3）与非台风过程不同，从台风过程的统计结果来看，ANN_B 相对于 ANN_S 更加接近于 Temp_ERA5。

（4）在非台风过程中，70hPa 上 Bias 和 RMSE 出现明显的增大，而在台风过程中这样的突增出现在 100hPa。

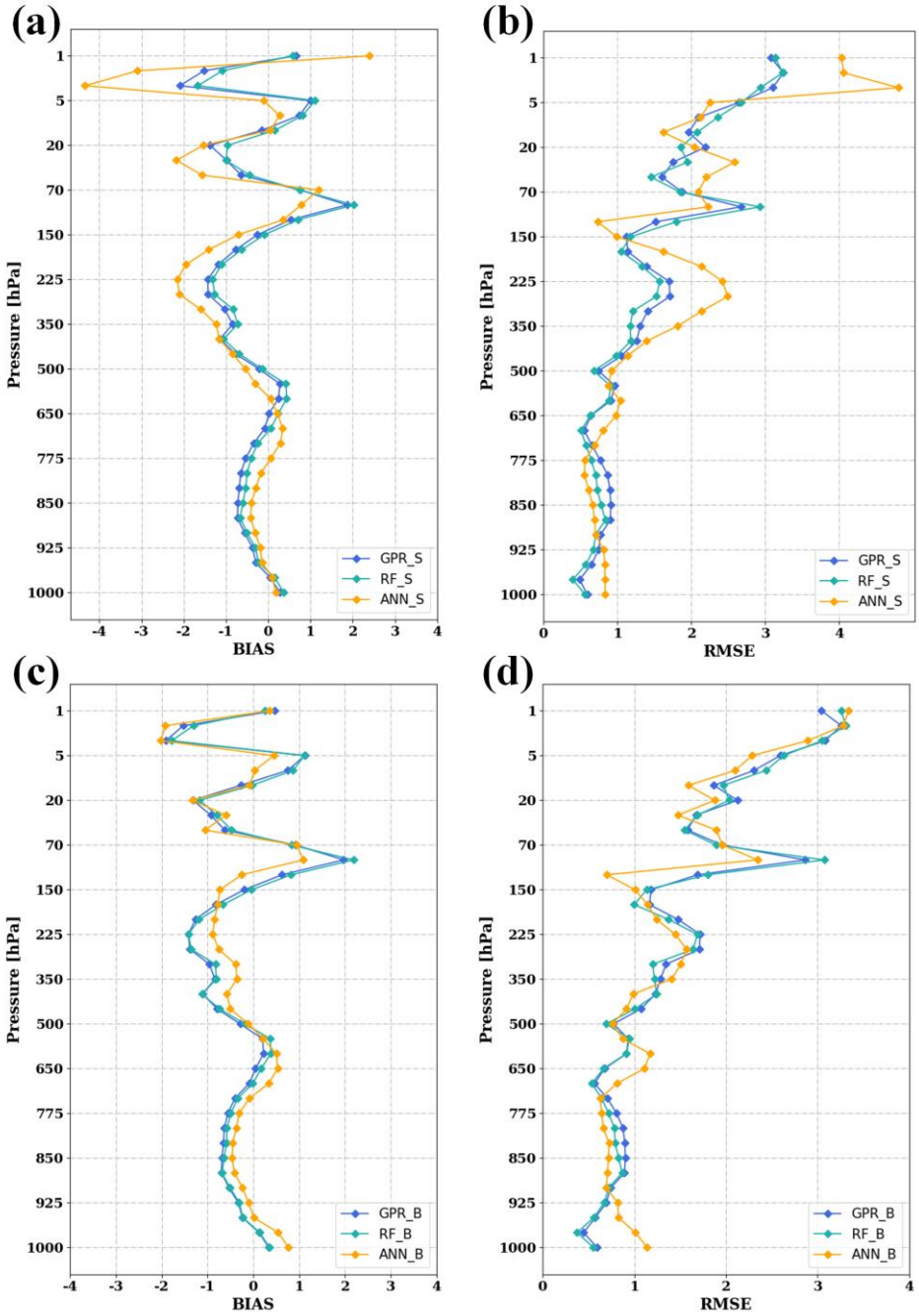


图 4.4　台风过程中 6 个模型的误差统计结果;图片布局同图 5.3。

4.5　试验结果讨论

4.5.1　反演误差

10hPa~1hPa 为平流层中层到平流层高层，700cm^{-1} 附近的光谱通道通常被用于

探测这个高度区间的温度。4.4.1~4.5.1 节的误差统计结果表明，以 Temp_ERA5 作为参考值，6 个温度反演模型的反演结果在 10hPa 以上的误差显著增大，且随高度增加出现较大的波动。本节将从温度雅克比和 10hPa 以上 ERA5 温度廓线的准确度两方面解释出现这种现象的可能原因。

4.5.1.1 温度雅克比

温度雅克比能够反映每个探测通道对于哪个高度层的大气温度敏感。本文将 6 条 54 层的标准大气廓线输入进快速辐射传输模式，计算得到每个通道在各个气压层上的平均雅克比值构成了每个通道的温度雅克比曲线。图 4.2（a）为长波 689 个光谱通道的温度雅克比曲线，可以看出峰值在 28.6937hPa 以上的雅克比曲线对应的探测高度均在 13.1672hPa 附近，而在 13.1672hPa 以上的气压层对应的都是温度雅克比曲线的尾部，意味着在 GIIRS 的光谱通道中不存在温度雅克比曲线峰值在 13.1672hPa 以上的光谱通道，即 GIIRS 的 689 个长波红外通道对于这个高度以上的温度不敏感，这可能是 4.4.1 和 4.4.2 中所有模型在 10hPa 以上的温度反演结果均较差的原因之一。

4.5.1.2 对 Temp_ERA5 的评估

ERA5 在东亚和热带地区的对流层温度廓线同无线电探空仪[79]、全球无线电掩星资料[80]、多种再分析资料[79-80]间的相互对比结果表明，ERA5 在对流层的温度垂直分布最接近东亚和热带地区的真实大气。Graeme 等学者[81]将平流层激光雷达探测温度（Temp_Lidar）同 ERA5 的平流层温度进行了对比，发现在 10hPa 以上 Temp_ERA5-Temp_Lidar 主要表现为负偏差，且偏差具有随着高度显著增大的趋势，最大负偏差约在-5K 左右。

为了进一步调查是否是 ERA5 在 10hPa 以上温度廓线精度相对于对流层有所降低导致 4.4.1~4.4.2 节中 10hPa 以上反演误差显著增大，下面以 AIRS 的温度廓线产品作为参照，统计出了 AIRS 的 L2 级大气温度廓线产品（Temp_AIRS）和 Temp_ERA5 间的偏差（图 4.5）。Temp_AIRS 由对云不敏感的 53 个探测平流层温度的通道（$662cm^{-1}\sim713cm^{-1}$）和 30 个探测对流层和平流层温度的通道（$2358cm^{-1}\sim2418cm^{-1}$）经物理反演法得到，包含了从 1000hPa 到 0.1hPa 的 27 个气压层上的温度。本节选取了台风过程和非台风过程中 AIRS 晴空像元的温度廓线（Temp_AIRS），计算了 Temp_AIRS 与 Temp_ERA5 的偏差统计结果（其中 Bias=Temp_AIRS-Temp_ERA5），并将该结果同 4.4.1 节和 4.4.2 节的统计结果进行了对比。结合图 4.3~图 4.5 可以看出：

（1）在非台风过程中，20hPa 以下 RF 和 GPR 的反演结果和 Temp_AIRS 相近，表现为 Bias 均在 ±1K 以内，RMSE 随高度的走向较为一致，且均在 2K 以内。10hPa 及以上，Temp_AIRS 和本文的温度反演结果均较差，Bias 和 RMSE 最大值达到 5.5K 左右；

（2）在台风过程中，Temp_AIRS 在 400hPa 以下的 RMSE 相对于非台风过程明显增大，其他层次 RMSE 变化不大。与 Temp_AIRS 不同，RF 和 GPR 的 RMSE 在台风过程中相对于非台风过程在 500hPa 到 150hPa 明显增大。10hPa 及以上，Temp_AIRS 和本文的温度反演结果均较差，Bias 和 RMSE 最大值达到 6.5K 左右。

上述对比结果表明，Temp_ERA5 和 Temp_AIRS 在 10hPa 以上的差异显著增大。由于 AIRS 的温度反演产品在平流层也存在误差，因此以上对比结果无法有效地定量地说明 Temp_ERA5 相对于真实状态的差异。但是通过以往的研究[79-81]和本节的对比，但是可以发现平流层的温度反演比对流层更难，平流层温度产品精度比对流层低，这可能也是导致 4.4.1~4.4.2 节中 10hPa 以上反演误差结果显著增大的可能原因之一。

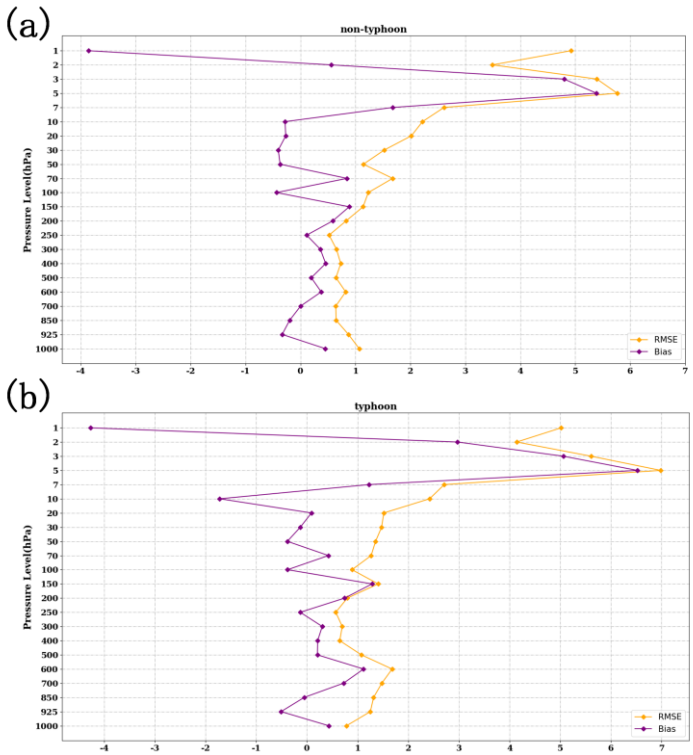


图 4.5　Temp_AIRS 同 Temp_ERA5 的偏差统计结果。(a)非台风过程; (b)台风过程。

4.5.2　训练时间

本文采用了 S 和 B 两个样本集训练模型，其中 GPR_S、RF_S 与 GPR_B、RF_B 在非台风过程和台风过程的反演统计结果非常相近，这说明采用具有一定代表性的小样本集能够达到和使用包含冗余的大样本集十分相近的反演结果。此外，采用 B 训练模型相对于采用 S 训练模型在时间开销上具有较大优势，表 4.4 为非台风过程时，采用 S 和 B 两个样本集训练 GPR、RF 和 ANN 三个模型的十次平均训练时间。结合反演结果和训练时间来看，RF_B 模型的表现较为突出，它的训练时间显著减小的同时具有较高的反演精度。GPR_S 和 GPR_B 虽然也具有较高的反演精度，但是由于需要计算矩阵的逆，所以计算代价高，训练时间长。

表 4.4　6 个模型的十次平均训练时间

GPR_S	GPR_B	RF_S	RF_B	ANN_S	ANN_B
35min	15min	3min	1min	16min	4min

4.6　本章小结

本章以 Temp_ERA5 作为基准值，根据 GIIRS 长波红外光谱通道的温度雅克比选择了 50 个观测值作为输入特征，采用高斯过程回归、随机森林和人工神经网络三种机器学习算法从 GIIRS 观测辐射率中反演得到了 37 层大气温度廓线，得出了以下结论：

（1）在大气温度廓线反演问题上，GPR 和 RF 表现出比 ANN 更好的反演能力，对于非台风过程，在 10hPa 以下基本能达到 RMSE<1.3K，Bias 在 ±1K 以内的反演效果，和 Temp_AIRS 的产品误差统计结果相近。

（2）台风过程中，三种模型在两个训练集上的效果均有所降低，在 10hPa 以下 RMSE<2K，Bias 在 ±2 K 以内，模型效果降低可能是由台风过程中的训练样本数减小导致。

（3）基于最大香农熵采样得到的具有代表性的小样本集 B 训练出的 GPR_B 和 RF_B 模型与带有冗余信息的原始数据集 S 训练出的 GPR_S 和 RF_S 的温度反演效果十分相近。对于非台风过程，ANN_S 比 ANN_B 温度反演能力更强，而对于台风过程结论相反。此外，相对于采用 S 训练，三种模型在 B 上的训练时间均大幅减小。综上所述，随机森林模型在训练时间和温度反演精度上表现更为优异。

（4）在台风和非台风过程中，根据 689 个光谱通道的温度雅克比曲线反映出的信息，三种模型在 10hPa 以上的反演精度很低可能与 GIIRS 的 689 个长波红外通道对于这个高度以上的温度不敏感，以及 ERA5 的温度产品在平流层的精度相对于

对流层降低有关。

第五章　总结与展望

5.1　论文总结

本文针对风云四号上的干涉式垂直探测仪（GIIRS）在数值天气预报系统中应用时面临的挑战，结合具体问题，在 GIIRS 的云检测、反演大气温度廓线两方面采用机器学习算法展开了相关工作，完成了红外高光谱资料进入同化系统前的高效预处理、得到了高精度温度廓线产品。本文主要的研究内容和结论如下：

（1） 针对 GIIRS 资料的云检测

选取了能够进行特征选择的L_1正则化 Logistic 回归算法对 GIIRS 观测像元进行云检测。构建了两种特征输入数量的云检测模型，两个模型在测试集上的 HSS 得分和准确率均在 95%以上；从可视化角度来看，两个模型均能够在台风、云片散布等情况下达到较好的云检测效果。本文提出的机器学习云检测算法相对于传统的成像仪-高光谱匹配云检测算法在检测效率上有明显优势；基于L_1正则化方法在全通道输入模型中保留得到的通道对于未来研究与云物理过程相关的物理参数具有一定意义。

（2） 反演大气温度廓线

分别采用了 GPR、RF 和 ANN 三种机器学习算法从 GIIRS 观测辐射中反演大气温度廓线。考虑到光谱通道的温度雅克比和噪声水平，从 689 个长波红外通道中选取了 50 个通道作为输入特征。试验中创新性地使用香农熵采样法构造出了有代表性的样本集，反演出高精度大气温度廓线的同时显著提升了训练效率。三种机器学习模型中，基于随机森林方法的反演模型效果突出：非台风过程中，随机森林模型在 10hPa 以下的压层上均方根误差（Root Mean Squared Error，RMSE）在 1.5K 以内，偏差（Bias）在±1K 以内；台风过程中，10hPa 以下 RMSE 在 2K 以内，Bias 在±2 K 以内。

。

5.2　研究展望

基于本文已有的研究结果以及红外高光谱资料在数值天气预报系统中的应用中存在的诸多挑战，以下方面仍需改进或是做进一步研究：

（1） 红外高光谱部分有云像元检测

由于红外高光谱资料的水平空间分辨率低，被云污染的像元几乎占了所有像

元的 90%，如果直接剔除所有被云污染的像元将造成对观测资料的极大浪费。因此，通常要对有云像元做进一步处理使其能够最大化地被数值天气预报系统所利用。其中被广泛使用的一种处理方法称为云清除方法，这种方法将部分有云像元的观测辐射率修正成晴空条件下的辐射率，已经使 50%的部分有云像元能够被数值天气预报系统使用。受该研究启发，未来的云检测应该被视为三分类问题：晴空、部分有云、完全有云，其中区分部分有云和完全有云的晴空无疑是这个三分类问题的难点。

（2） GIIRS 资料偏差订正

资料同化理论中假设观测误差是随机、无偏分布的，然而实际上，辐射观测值和模拟值之间的误差包含许多系统性的偏差的混合，而且这些系统性偏差常常超过了观测资料本身的噪声水平，并不符合随机无偏假设。例如，卫星观测包含依赖于仪器的偏差、辐射传输模式中的近似可能会在同化中造成复杂的、依赖于数值预报得到的大气状态的系统偏差等。来自于不同源头的偏差常常表现出一些时间和空间上的分布特征，例如表现为偏差具有仪器扫描角依赖、地理位置或气团特征依赖等。消除观测和模拟值之间的系统性偏差是观测资料进入变分迭代过程前的重要步骤，当观测偏差具有很强的时空变异性时，对其建模是非常具有挑战性的。

致　谢

在硕士毕业论文即将完成之际，回首两年半的硕士生活，一些场景浮上心头。

在科研课题的研究中，张老师和周老师给予了我支持、指导，在我遇到瓶颈时给予我宝贵的建议。张老师为学生创造了舒适的科研环境，使我们可以全身心投入学业；他拥有深厚的专业知识，在每次组会报告中，都会对报告人的课题情况做出一针见血的评价并给于切实的指导和建议。我由衷地感谢在硕士期间两位老师的帮助与引导。

余意师姐和罗藤灵师兄在我选题和进行试验时给予了无私的帮助和指导，在此感谢他们的付出。此外，同门中兄弟姐妹给予了我很多帮助和包容，大家一起营造的学术严肃、生活活泼的氛围，给我的硕士生活留下了难忘的回忆。冯淼师姐如今求学在外，不知何时能再见。师姐真诚可爱，在生活中给予我温暖，在学术中给予我帮助，能够认识冯淼师姐是特别幸运的事情。希望她在外求学顺利，身体健康。在两年半的硕士生涯中，有幸能与戴俊、张永顺、范茂廷同学一起经历科研的毒打，感谢他们对我的各种疑问的耐心解答。

今年 8 月我们的实验室从天河楼搬到科技楼。非常幸运的是，两个实验楼的保安叔叔都非常亲切随和，很多时候，每天早上我见到的第一个微笑都是他们给予的，感谢这些温柔可爱的人们。

接着就是我们 305 聊天室，我和聊天室的其他两名成员吕翾和赵茜行走在学校的每一条道路上、飞驰在开福区夜晚的马路上的画面历历在目。笑颜、焦躁、活蹦乱跳被收藏在心中的宝箱里，成为我的某种力量。

刘畅和梁振，我的好友。感谢我们彼此间存在的相互支持、倾听、看望、拥抱、发呆、自在。

我的爸爸妈妈，一切尽在彼此相通的心间。

最后是关于对自己的期望，我的成长很重要的一部分是认识平凡和接受平凡，这说起来奇怪但也不奇怪。如果未来某一天突然翻到自己的学位论文，希望接下来的话，能够让那时的我想起真正的自己。性格并不分好坏，希望自己能永远真诚善良。于自己，追逐勇敢，成为自己的荧光和炬火。于国家，永远忠诚。

参考文献

[1] 朱国富.数值天气预报中分析同化基本方法的历史发展脉络和评述[J].气象, 2015（8）:68~78.

[2] Menzel W P, Schmit T J, Zhang P, et al. Satellite Based Atmospheric Infrared Sounder Development And Applications[J]. Bulletin of the American Meteorological Society [J], 2018, 99（3）:583~603.

[3] 刘辉,董超华,张文建.AIRS 晴空大气温度廓线反演试验.气象学报，2008（66）:513~519.

[4] Chahine M T, Pagano T S, Aumann H H, et al. AIRS: Improving Weather Forecasting and Providing New Data on Greenhouse Gases [J]. Bulletin of the American Meteorological Society, 2006, 87（7）: 238.

[5] Clerbaux C， Hadi-Lazaro J， Turquety S，et al. The IASI/MetOp 1 mission: First Observations And Highlights of Its Potential Contribution to GMES 2. Space Research Today [J], 2007（168）: 19~24.

[6] Smith A, Atkinson N, Bell W, Doherty A. An Initial Assessment of Observations from The Suomi-NPP Satellite: Data from the Cross-track Infrared Sounder (CrIS). Atmospheric Science Letters [J], 2015（16）: 260~266.

[7] 董超华,李俊,张鹏.卫星高光谱红外大气遥感原理和应用[M].科学出版社, 2013.

[8] Li, J., F. X. Zhou, and Q. C. Zeng. Simultaneous Non-linear Retrieval of Atmospheric Temperature And Absorbing Constituent Profiles from Satellite Infrared Sounder Radiances. Advances in Atmospheric Sciences [J], 1994（11）: 128 ~138.

[9] Zheng J, Li J, Schmit T J, et al. The Impact of AIRS Atmospheric Temperature And Moisture Profiles on Hurricane Forecasts: Ike (2008) And Irene (2011). Advances in Atmospheric Sciences [J], 2005（32）: 319~335.

[10] Sieglaff J M, Schmit T J, Menzel, W P, et al. Inferring Convective Weather Characteristics With Geostationary High Spectral Resolution IR Window Measurements: A Look Into The Future. Journal of Atmospheric and Oceanic Technology [J], 2019（26）: 1527~1541.

[11] Wu X, Li J, Menzel W P, et al. Evaluation of AIRS Cloud Properties Using MPACE Data. Geophysical Research Letters [J], 2005（32）.

[12] Chirs B. Assimilation of Radiance Observations from Geostationary Satellites:Third Year Report. ECMWF, 2020

[13] 官元红,周广庆,陆维松.资料同化方法的理论发展及应用综述.气象与减灾研究, 2007, 1~8.

[14] Collard A D. Selection of IASI Channels for Use in Numerical Weather Prediction [J]. Quarterly Journal of the Royal Meteorological Society, 2010, 133（629）.

[15] 王根,张华,杨寅.高光谱大气红外探测器 AIRS 资料质量控制研究进展[J]. 地球科学进展, 2017（2）: 139~150.

[16] Yin R, Han W, Gao Z, et al. The Evaluation of FY4A's Geostationary Interferometric Infrared Sounder (GIIRS) Long-wave Temperature Sounding Channels Using the GRAPES Global 4D‐Var [J]. Quarterly Journal of the Royal Meteorological Society, 2020, 146.

[17] 盛裴轩.气物理学[M].北京大学出版社, 2013.

[18] Lary D J, Zewdie G K, Liu X, et al. Machine Learning Applications for Earth Observation[M].Earth Observation Open Science and Innovation, 2018.

[19] 陈渭民.卫星气象学[M].气象出版社, 2003.

[20] Wylie D P, Menzel W P, Woolf H M, et al. Four Years of Global Cirrus Cloud Statistics Using HIRS [J]. Journal of Climate. 1994（7）: 1972~1986.

[21] McNally A P, Watts P D. A Cloud Detection Algorithm for High-spectral-resolution Infrared Sounders [J]. Quarterly Journal of The Royal Meteorological Society, 2003（129）: 3411~3423

[22] Li J, Liu C, Huang H, et al. Optimal Cloud-clearing for AIRS Radiances Using MODIS. IEEE Transactions on Electron Devices. 2005（43）: 1266~1278.

[23] King J I. The Radiative Heat Transfer of Planet Earth. Scientific Uses of Earth Satellites [D]. University of Michigan Press, Ann Arbor, 1996: 133~136

[24] Kaplan L D. Inference of Atmospheric Structure from Remote Radiation Measurements [J]. Journal of The Optical Society of America A-optics Image Science And Vision, 1959（49）: 1004~1007.

[25] Stowe L L, Davis P A, Mcclain E P. Scientific Basis And Initial Evaluation of The CLVAR-1 Global Clear/cloud Classification Algorithm for The Advanced Very High Resolution Radiometer [J]. Journal of Atmospheric And Oceanic Technology, 1999（16）: 656~681.

[26] Baum B A, Arduini R F, Wielicki B A, et al. Multilevel Cloud Retrieval Using Multispectral HIRS and AVHRR Data: Nighttime Oceanic Analysis [J]. Journal of Geophysical Research Atmosphere. 1994（99）: 5499~5514.

[27] Ackerman S A, Strabala K I, Menzel W P, et al. Discriminating Clear Sky from Clouds with MODIS [J]. Journal of Geophysical Research. 1998（103）: 141~157.

[28] Ackerman S A, Holz R E, Frey R, et al. Cloud Detection with MODIS. Part II: Validation [J]. Journal of Atmospheric and Oceanic Technology. 2010（25）: 1073~1086.

[29] Yang J, Guo J, Yue H, et al. CNN-based Cloud Detection for Remote Sensing Imagery [J]. IEEE Transactions on Geoscience and Remote Sensing. 2019（99）: 1~17.

[30] Mohajerani S, Saeedi P. Cloud-Net: An End-To-End Cloud Detection Algorithm for Landsat 8 Imagery [Z]. IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019.

[31] Li P, Dong L, Xiao H, et al. A Cloud Image Detection Method Based on SVM Vector Machine [J]. Neurocomputing 2015（169）: 34~42.

[32] Xie F, Shi M, Shi Z, et al. Multilevel Cloud Detection in Remote Sensing Images Based on Deep Learning [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2017（10）: 3631~3640.

[33] Mohajerani S, Krammer T A, Saeedi P . A Cloud Detection Algorithm for Remote Sensing Images Using Fully Convolutional Neural Networks [Z]. The 20th IEEE International Workshop on Multimedia Signal Processing.

[34] Li P, Dong L, Xiao H, et al. A Cloud Image Detection Method Based on Support Vector Machine [J]. Neurocomputing, 2015（169）: 34~42

[35] Bai T, Li D, Sun K, et al. Cloud Detection for High-resolution Satellite Imagery Using Machine Learning And Multi-feature Fusion [J]. Remote Sensing. 2016（8）: 715.

[36] Han B, Kang L, Song H. A Fast Cloud Detection Approach by Integration of Image Segmentation And Support Vector Machine [R]. the International Symposium on Neural Networks, 2006.

[37] Latry C, Panem C, Dejean, P. Cloud Detection with SVM Technique. the IEEE Geoscience and Remote Sensing Symposium.

[38] Xu L, Wong A, Clausi D A. A Novel Bayesian Spatial-temporal Random Field Model Applied to Cloud Detection from Remotely Sensed Imagery [J]. IEEE Transactions on Geoscience and Remote Sensing. 2017（55）: 4913~4924.

[39] Li Q, Lu W, Yang J, et al. Thin Cloud Detection of All-sky Images Using Markov Random Fields [J]. IEEE Geoscience and Remote Sensing Letters, 2012（9）: 417~421.

[40] Li J, Menzel P W, Sun F, et al. AIRS Subpixel Cloud Characterization Using MODIS Cloud Products [J]. Journal of Applied Meteorology, 2004（43）: 1083~1094.

[41] Eresmaa R. Imager-assisted Cloud Detection for Assimilation of Infrared Atmospheric Sounding Interferometer Radiances [J]. Quarterly Journal of the Royal Meteorological Society. 2014（140）: 2342~2352

[42] 蒋德明.高光谱分辨率红外遥感大气温湿度廓线反演方法研究[D].南京信息工程大学. 2007.

[43] Campos T M, Garcia H, Francisco J, et al. Multitemporal And Multiresolution Leaf Area Index Retrieval for Operational Local Rice Crop Monitoring [J]. Remote Sensing of Environment, 2016（187）: 102~118.

[44] Valls G, Svendsen D, Martino L, et al. Advances in Gaussian Processes for Earth Sciences: Physics-aware, interpretability and consistency [Z], EGU General Assembly 2020, Online.

[45] Jose E, Vicent J, Rivera J P, et al. Gaussian Processes Retrieval of LAI from Sentinel-2 Top-of-atmosphere Radiance Data [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020（167）: 289~304.

[46] Yin G, Verger A, Qu Y, et al. Retrieval of High Spatiotemporal Resolution Leaf Area Index with Gaussian Processes, Wireless Sensor Network, and Satellite Data Fusion[J]. Remote Sensing, 2019（11）: 244.

[47] Paola F D, Cersosimo A. Retrieval of Temperature and Water Vapor Vertical Profile from ATMS Measurements with Random Forests Technique [Z], IEEE International Geoscience and Remote Sensing Symposium, 2018: 6014~6017.

[48] Huttunen J, Kokkola H, Mielonen T, et al. Retrieval of Aerosol Optical Depth from Surface Solar Radiation Measurements using Machine Learning algorithms, Non-linear Regression And A Radiative Transfer-based Look-up Table [J]. Atmospheric Chemistry and Physics, 2016（30）: 1~23.

[49] Bi J, Belle J H, Wang Y, et al. Impacts of Snow And Cloud Covers on Satellite-derived PM2.5 Levels [J]. Remote Sensing of Environment. 2019（221）: 665~674.

[50] 周爱明.基于风云四号高光谱红外模拟资料反演大气温湿廓线试验研究[D]. 2017.

[51] Luchetta A, Serio C, Viggiano M. A Neural Network to Retrieve Atmospheric Parameters from Infrared High Resolution Sensor Spectra [R]. The 2003 International Symposium on Circuits and Systems, 2003.

[52] Polyakov A V. The Method of Artificial Neural Networks in Retrieving Vertical Profiles of Atmospheric Parameters [J]. Atmosphere Ocean . 2014（27）: 247~252.

[53] Aires F, Alain C, Noelle A S. A Regularized Neural Net Approach for Retrieval of Atmospheric And Surface Temperatures with the IASI Instrument [J]. Journal of Applied Meteorology. 2002, 144~159.

[54] Kolassa J, Reichle R H, Liu Q, et al. Estimating surface soil moisture from SMAP observations using a Neural Network technique [J]. Remote Sensing of Environment, 2018（204）: 43~59.

[55] Kim H S, Park I, Song C H, et al. Development of a daily PM 10 and PM 2.5 prediction system using a deep long short-term memory neural network model[J]. Atmospheric Chemistry and Physics, 2019（19）: 12935~12951.

[56] Mitchell T M. Machine Learning [M]. McGraw-Hill, 2003.

[57] 朱劲夫,刘明哲,赵成强.正则化在逻辑回归与神经网络中的应用研究[J]. 信息技术, 2016（7）:1-5.

[58] Chen L, Li Z S, Yu Z H, et al. Classifier-Guided Topical Crawler: A Novel Method of Automatically Labeling the Positive URLs [M]. IEEE Computer Society, 2009.

[59] Czajkowski M, Kretowski M. Decision Tree Underfitting in Mining of Gene Expression Data-An Evolutionary Multi-test Tree Approach [J]. Expert Systems with Applications, 2019（137）: 392-404.

[60] Domingos P. The Role of Occam's Razor in Knowledge Discovery [J]. Data Mining & Knowledge Discovery, 1999,（4）: 409-425.

[61] Kohavi R. A Study of Cross-validation And Bootstrap for Accuracy Estimation And Model Selection [A] . The 14th Joint Int Conf Artificial Intelligence [C], IEEE, 1995: 1137~1145.

[62] Andrew Y N. Feature Selection, L_1 VS. L_2 Regularization, And Rotational Invariance. The International Conference on Machine learning, Louisville, KY, USA, 16–18 December 2004.

[63] Menard S. Logistic Regression[J]. American Statistician, 2004, 58（4）: 364.

[64] Judith E D, Deleo J M .Artificial Neural Networks [J]. Encyclopedia of Microfluidics & Nanofluidics, 2001（S8）: 23-33.

[65] Eric S A. A Tutorial on Gaussian Process Regression: Modelling, Exploring, And Exploiting Functions [J]. Journal of Mathematical Psychology, 2018（85）: 1-16.

[66] Dreiseitl A S. Logistic regression and Artificial Neural Network Classification Models: A Methodology Review [J]. Journal of Biomedical Informatics, 2002（35）: 352-359.

[67] 田琨.南京地区雷暴活动强度潜势预报[J].气象科技, 2013（1）:177-183.

[68] Lin Y P, Cheng B Y, Chu H J, et al. Assessing How Heavy Metal Pollution And Human Activity Are Related by Using Logistic Regression And Kriging Methods [J]. Geoderma, 2011（163）: 275-282.

[69] Sahai A K, Soman M K, Satyan V. All India Summer Monsoon Rainfall Prediction Using An Artificial Neural Network [J]. Climate Dynamics, 2000（16）: 291-302.

[70] Adamowski J, Chan H F, Prasher S O, et al. Comparison of Multiple Linear and Nonlinear Regression, Autoregressive Integrated Moving average, Artificial Neural Network, And Wavelet Artificial Neural Network Methods for Urban Water Demand Forecasting in Montreal, Canada [J]. Water Resources Research, 201（48）: 273-279.

[71] Gardner M W. Artificial Neural etworks (The Multilayer Perceptron)—A Review of Applications in The Atmospheric Sciences [J]. Atmospheric Environment, 1998（32）.

[72] Feng X, Li Q, Zhu Y, et al. Artificial Neural Networks Forecasting of $PM_{2.5}$ Pollution Using Air Mass Trajectory Based Geographic Model And Wavelet Transformation [J]. Atmospheric Environment, 2015（107）: 118-128.

[73] Yi G, Shi J Q, Choi T. Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data [J]. Biometrics, 2015（67）: 1285-1294.

[74] Feng M, Zhang W M, Zhu X R , et al. Multivariate Interpolation of Wind Field Based on Gaussian Process Regression [J]. Atmosphere, 2018（9）: 194.

[75] Sit H, Earls C J. Gaussian Process Regression for Estimating EM Ducting Within the Marine Atmospheric Boundary Layer [J]. Radio ence, 2020（55）.

[76] Mooney, Christopher Z. Bootstrapping:A Nonparametric Approach to Statistical Inference [M]. Sage Publications, 1994.

[77] Chris B. Assimilation of Radiance Observations from Geostationary Satellites: Second Year Report [R]. ECMWF, 2019: 1~48.

[78] Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyzes of the global climate [Z] . Copernicus Climate Change Service Climate Data Store (CDS), 2020

[79] Alghamdi A S. Evaluation of Four Reanalysis Datasets Against Radiosonde over Southeast Asia [J]. Atmosphere. 2020（11）: 402.

[80] Tegtmeier S, Anstey J, Davis S M, et al. Temperature And Tropopause Characteristics From Reanalyses Data in The Tropical Tropopause Layer [J]. Atmospheric Chemistry and Physics. 2020（20）: 1~28.

[81] Graeme M, Andrew C P, Giles H, et al. Using A Global Network of Temperature Lidars to Identify Temperature Biases in The Upper Stratosphere in ECMWF Reanalyses [J]. Atmospheric Chemistry and Physics. 2020.

[82] Han, W. Assimilation of GIIRS radiances in GRAPES [R]. The 35th Chinese Meteorological Society Conference.

作者在学期间取得的学术成果

Zhang Q, Yu Y, Zhang W, et al. Cloud Detection from FY-4A's Geostationary Interferometric Infrared Sounder Using Machine Learning Approaches [J]. Remote Sensing, 2019（24）: 3035.

附录 A　模型特征通道索引号

表 A.1 云检测模型 38 个特征光谱通道索引号

3	4	6	9	11	12	15	27	33	34
38	41	63	65	70	72	73	75	77	78
79	80	82	83	84	85	86	87	88	89
90	91	111	112	113	281	425	449		

表 A.2 反演大气温度廓线模型的 50 个特征光谱通道索引号

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	59	60	61	62	63	64	65	66