随机森林方法在机场温度预测中的应用

朱国栋　孙建杰　陈阳权　王照刚
（民航新疆空中交通管理局气象中心，乌鲁木齐 830016）

摘要：利用2015—2017年欧洲中心细网格数值预报产品，使用随机森林方法，结合不同数量的决策树进行模型训练，研究建立基于随机森林方法的乌鲁木齐机场逐时温度回归预报模型。通过对模型预测结果的检验可以看到，模型预测乌鲁木齐机场温度平均绝对误差≤2 ℃的占样本总数的94%，机场温度−10～30 ℃，平均绝对误差为1.2 ℃，该方法预测效果较好，因此可以尝试使用本方法制作民航机场客观要素指导产品。

关键词：随机森林方法，机场温度，预报

DOI：10.3969/j.issn.2095−1973.2021.04.008

A Random Forest Application for Predicting Airport Temperature

Zhu Guodong, Sun Jianjie, Chen Yangquan, Wang Zhaogang
(Xinjiang Civil Aviation Meteorological Center, Urumqi 830016)

Abstract: Using the 2015−2017 European Center numerical forecasting product, combined with different numbers of decision trees for model training, establishes regression forecasting model based on the random forest method at Urumqi airport. Through the inspection of the prediction results of the model, we can see the average absolute temperature error is less than or equal 2 ℃, accounting for 94% of the total number of samples, the airport temperature is between −10 and 30 ℃, and the average absolute error is 1.2 ℃. The effect is good, so you can try to use this method to produce objective product guidance products for civil aviation airports.

Keywords: random forest method, airport temperature, forecast

0　引言

随着近年民用航空产业的不断发展，航空公司、机场、空管等民航气象用户单位对机场预报的准确性和时效性要求不断提高。其中机场温度作为民用航空器配载的重要指标，准确的预报将会对飞机旅客、货物、油料的装载数量提供科学的参考，同时为确保航班起降安全提供帮助。

目前对地面温度的预测主要依托数值预报产品，但是模式直接输出的温度预测产品与实况存在一定的偏差[1-2]，为了解决模式直接输出产品的误差，通过对不同数值模式产品的检验和误差订正[3]，并应用机器学习方法开展模式解释应用[4-6]，较好地改善了温度预测的效果。同时参考不同的机器学习方法的特性和在气象领域的预测效果[7-8]，本文选取能够较好地处理非线性问题的随机森林方法，结合欧洲中心细网格数值预报产品，实现对乌鲁木齐地窝堡国际机场的逐小时地面2 m温度的预测，为民航运行单位提供科学、可靠的温度预报产品，进而更好地为民航安全、效益服务。

收稿日期：2019 年 9 月 17 日；修改日期：2021 年 1 月 28 日
第一作者：朱国栋，Email：blueet@163.com

1　随机森林方法

随机森林是基于决策树的集成学习算法[9]，决策树是一种广泛应用的树状分类器，在树的所有节点上，通过选择最优的特征不断进行分类，直到达到建树的停止条件。决策树是无参数有监督的机器学习方法，不需要先验知识，相比神经网络等方法更容易解释，但是单个决策树对问题预测性能有限，为了改善单个分类器的预测性能，将单个分类器聚集起来，通过对每个基本分类器的分类结果进行组合，也就是形成多个决策树组成的随机森林，提升模型的预测精度和泛化能力，避免出现过拟合现象。

2　数据预处理

在机器学习方法中，虽然算法的选型很重要，但是良好的数据才是算法的基本。然而在实际的应用中，产生的气象数据并不一定符合算法的要求，总会由于一些客观因素影响数据的收集，例如观测设备故障、数值模式传输错误等。

本文整理乌鲁木齐地窝堡国际机场（以下简称机场）逐小时地面观测资料，将机场温度作为预测对象，筛选气温对应时刻的前24 h地面风、气温等要素。同时，利用2015—2017年逐日20时起报的欧洲中

心细网格数值预报产品，包括2T、2D、高空温度、湿度、高度场、UV风场等要素，预报有效时间12～36 h的预测产品，由于不同的预测要素产品网格距离不同，本文采用查找距离机场最近网格点上的数据，与机场温度构建训练样本序列。通过对收集到的数据进行数据筛查、清洗等预处理，剔除数值预报产品中的缺测记录后，共得到22056条样本记录。

由于不同的物理量组成的因子存在着量级差异，在进行模型训练和参数寻优前，需要归一化处理所有的因子，将其限定在0～1，具体处理方法如式（1）所示：

$$X_n = \frac{X_n - \min(X)}{\max(X) - \min(X)} 。 \quad （1）$$

为了评估不同算法模型的预测能力，同时避免模型出现过拟合，本文利用开源工具包scikit-learn对归一化后的样本进行随机切分，确保检验样本的独立性，将样本数据中随机抽取33%作为检验样本，67%作为训练样本，进行模型训练和参数寻优。

3 温度预测模型构建

3.1 模型构建

利用Python的开源机器学习库scikit-learn构建预测模型，为了评估不同的方法和模型参数下的温度预测效果，本文选取决策树回归方法和随机森林回归方法进行建模，并通过设置不同决策树数量来评估随机森林方法的预测能力，具体模型实现代码如下：

1） 决策树方法

```
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor()
model.fit(trainX,trainY)
```

2） 随机森林方法

```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(n_estimators=200)
model.fit(trainX,trainY)
```

3.2 模型寻优

通过分析预测模型可以发现，决策树回归方法对测试样本的机场温度预报平均绝对误差为1.01 ℃，而采用5个决策树的随机森林方法较好地改善了平均绝对误差，达到0.82 ℃，随着增加决策树数量，模型对测试样本预测结果的平均绝对误差逐渐减小，当决策树数量达到200个以上后，模型预测误差逐渐趋于稳定，达到0.70左右。由此可见随机森林用于温度回归预测效果明显好于单个决策树，同时随着决策树数量的增加，对测试样本的平均绝对误差逐渐减小。具体模型预测结果如表1所示。

表1 不同模型及参数构建的预测模型效果评价
Table 1 Evaluation of prediction models for different models and parameters

序号	模型类型	决策树数量	测试样本平均绝对误差/℃	模型训练耗时/s
1	决策树	-	1.01	2
2	随机森林	5	0.82	4
3		10	0.76	9
4		20	0.74	17
5		50	0.71	41
6		100	0.70	83
7		200	0.696	106
8		500	0.694	418
9		1000	0.695	915

3.3 温度回归预测

利用模型寻优得到的预测模型，对样本中随机抽取的7279条测试样本进行预测，其中预测结果平均绝对误差≤1 ℃的占样本总数78%，平均绝对误差≤2 ℃的占样本总数94%，模型预测结果的平均绝对误差能够控制在2 ℃以内，对于温度业务预报有较好的指导作用。为了充分评估模型预测能力，本文将预测对象机场温度按照10 ℃为一个量级（表2），划分出8个量级。乌鲁木齐机场温度主要分布在−10～30 ℃，模型的平均绝对误差主要在1.2 ℃左右，其中−10～0 ℃平均绝对误差最小，为0.939 ℃。而−30～−20 ℃考虑到样本数量占比较少，仅为31个，此范围内的预测误差不单独进行统计。

表2 不同区间段内的温度预测误差分析
Table 2 Analysis of temperature prediction error in different interval segments

温度区间/℃	样本数量	平均绝对误差/℃
[30, −20)	31	0.583
[−20, −10)	751	1.129
[−10, 0)	1434	0.939
[0, 10)	1128	1.121
[10, 20)	1467	1.330
[20, 30)	1920	1.185
[30, 40)	541	1.209
[40, 50)	7	1.129

通过对四个季节的样本建立独立的随机森林预测模型，分析模型对训练样本的预测误差可以看到，春季气温预测模型的平均绝对误差为0.956 ℃，夏季气温预测模型的平均绝对误差为1.100 ℃，秋季气温预测模型的平均绝对误差为0.935 ℃，冬季气温预测模型的平均绝对误差为1.067 ℃。对比全年样本数据构建的预测模型，按季节构建的预测模型，在不同温度量级下的春季预测效果更优，具体见表3。

表3 不同季节的温度预测误差分析
Table 3 Analysis of temperature prediction error in different season

温度区间/℃	春季		夏季		秋季		冬季	
	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃
[30, −20)	-	-	-	-	-	-	34	1.172
[−20, −10)	7	0.829	-	-	27	0.906	725	0.937
[−10, 0)	363	0.807	-	-	234	0.702	771	0.920
[0, 10)	419	0.906	-	-	732	0.885	38	1.241
[10, 20)	829	1.105	221	1.240	399	1.037	-	-
[20, 30)	448	0.916	1237	1.016	265	0.980	-	-
[30, 40)	33	1.174	469	0.945	22	1.105	-	-
[40, 50)	-	-	6	1.200	-	-	-	-

4 数值预报对乌鲁木齐机场温度预测误差分析

利用欧洲中心细网格数值预报输出72 h的0.125°×0.125°2 m气温产品资料，结合乌鲁木齐机场本地特点，采用最近经纬网格点的数据做为乌鲁木齐机场的气温预报结果，通过对不同预报有效时间下的预报数据进行筛选，每个预报有效时间大约获得3100个样本，平均绝对误差为2.151 ℃，误差最小的为预报有效时间21 h，为1.932 ℃，误差最大的为预报有效时间72 h，为2.357 ℃，通过分析不同温度区间内的平均绝对误差可以看到，其中0～10 ℃平均绝对误差为1.839 ℃。具体见表4。

结合乌鲁木齐机场季节划分以及模式不同预报有效时间下的预测效果，选取预报有效时间为24 h的结果进行分析，春季气温预测的平均绝对误差为2.043 ℃，夏季气温预测的平均绝对误差为1.982 ℃，秋季气温预测的平均绝对误差为2.238 ℃，冬季气温预测的平均绝对误差为2.288 ℃。对比全年预测误差结果可以看到，夏季预测效果更优，具体见表5。

通过对比分析可以看到，欧洲中心细网格数值预报直接输出的温度预测结果较为稳定，平均绝对误差在2 ℃左右，利用随机森林方法的温度回归预测结果，平均绝对误差在1 ℃左右，对模式直接输出的温度结果有了较大的提升，其预测效果明显优于模式直接输出的结果。

表4 数值预报温度预测的平均绝对误差分析
Table 4 Analysis of the numerical forecast temperature prediction error

预报有效时间/h	总样本数量	平均绝对误差/℃	不同温度区间的平均绝对误差/℃							
			[30, −20)	[−20, −10)	[−10, 0)	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)
0	3174	2.027	2.572	2.23	1.704	1.839	2.181	2.207	1.631	2.965
3	3095	1.971	2.646	2.275	1.863	1.667	2.039	1.948	2.007	
6	3093	2.055	2.268	2.053	2.022	1.828	2.172	2.225	1.585	2.075
9	3116	1.970	2.947	2.052	1.948	1.734	2.01	2.194	1.481	1.823
12	3131	2.102	2.866	2.224	1.687	1.933	2.29	2.336	1.741	2.97
15	3125	1.964	3.093	2.375	1.978	1.554	1.951	1.931	1.89	
18	3107	2.052	2.943	2.183	2.209	1.792	2.109	2.147	1.417	1.59
21	3113	1.932	2.377	2.124	2.005	1.726	1.969	2.062	1.399	1.627
24	3122	2.082	2.914	2.312	1.702	1.828	2.28	2.281	1.671	2.63
27	3119	2.118	3.383	2.508	2.062	1.752	2.216	2.045	1.963	
30	3111	2.191	3.4	2.295	2.219	2.026	2.283	2.29	1.555	1.715
33	3116	2.076	3.453	2.358	2.005	1.812	2.101	2.251	1.589	1.627
36	3107	2.204	3.937	2.477	1.785	1.985	2.402	2.347	1.802	2.715
39	3109	2.156	3.73	2.835	2.162	1.638	2.143	2.065	1.851	
42	3102	2.209	3.411	2.524	2.398	1.926	2.246	2.266	1.438	1.295
45	3104	2.085	2.461	2.583	2.143	1.745	2.037	2.243	1.495	1.347
48	3115	2.233	3.753	2.673	1.83	1.913	2.381	2.407	1.746	2.665
51	3098	2.292	4.039	3.021	2.177	1.782	2.359	2.182	2.011	
54	3089	2.317	3.153	2.702	2.315	2.067	2.461	2.347	1.548	1.035
57	3091	2.203	3.374	2.634	2.177	1.931	2.209	2.329	1.606	0.92
60	3100	2.316	4.21	2.834	1.967	2.03	2.426	2.41	1.759	2.39
63	3097	2.298	3.883	3.231	2.251	1.745	2.256	2.16	1.922	
66	3074	2.334	3.639	2.906	2.522	1.97	2.365	2.314	1.569	0.855
69	3079	2.220	2.999	2.812	2.267	1.807	2.23	2.323	1.625	1.18
72	3081	2.357	3.965	3.011	1.869	1.946	2.513	2.499	1.797	2.295
平均	3107	2.151	3.257	2.529	2.051	1.839	2.225	2.232	1.684	1.880

表5 数值预报温度预测（24 h）在不同季节的平均绝对误差分析
Table 5 Analysis of the numerical forecast temperature prediction error in different season

温度区间/℃	春季		夏季		秋季		冬季	
	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃	样本数量	平均绝对误差/℃
[30, −20)	-	-	-	-			24	2.914
[−20, −10)	2	1.655	-	-	35	2.38	401	2.309
[−10, 0)	126	1.66	-	-	108	1.705	319	1.717
[0, 10)	169	2.14	-	-	252	1.609	7	2.21
[10, 20)	349	2.42	81	1.447	220	2.364	-	-
[20, 30)	169	2.156	488	2.219	126	2.692	-	-
[30, 40)	12	2.23	211	1.63	2	2.68	-	-
[40, 50)	0		2	2.63	-	-	-	-

参考文献

[1] 苟杨. 欧洲中心细网格两米温度短期数值预报检验. 农技服务, 2017, 34(5): 93.
[2] 王满, 边小勇, 张旭东. 基于WRF模式的局地短期气温预测研究. 现代计算机(专业版), 2017, (30): 6-11.
[3] 王焕毅. 三种数值模式气温预报产品的检验及误差订正方法研究//中国气象学会. 第35届中国气象学会年会 S1灾害天气监测、分析与预报. 北京: 中国气象学会, 2018: 12.
[4] 谭江红, 陈伟亮, 王珊珊. 一种机器学习方法在湖北定时气温预报中的应用试验. 气象科技进展, 2018, 8(5): 46-50.
[5] 陶晔, 杜景林. 基于随机森林的长短期记忆网络气温预测. 计算机工程与设计, 2019, 40(3): 737-743.
[6] 蔡凝昊. 江苏省温度精细化客观释用方法研究与讨论//中国气象学会. 第33届中国气象学会年会 S1灾害天气监测、分析与预报. 北京: 中国气象学会, 2016: 5.
[7] 余胜男, 陈元芳, 顾圣华, 等. 随机森林在降水量长期预报中的应用. 南水北调与水利科技, 2016, 14(1): 78-83.
[8] 黄衍, 查伟雄. 随机森林与支持向量机分类性能比较. 软件, 2012, 33(6): 107-110.
[9] 王奕森, 夏树涛. 集成学习之随机森林算法综述. 信息通信技术, 2018, 12(1): 49-55.