基于探空资料和随机森林算法的河北省强对流天气分类预报技术研究

张玉婷[1]，刘炳杰[2]，宋　灏[1]，孙　卓[3]，李月英[1]

（1.衡水市气象局，河北　衡水　053000；2.沧州市气象局，河北　沧州　061000；3.河北省气象局，河北　石家庄　050000）

摘　要：文章利用2000年～2019年河北省邢台探空站、张家口探空站每日8时、20时探空资料及地面、高空气象观测资料，结合人工智能随机森林算法，构建河北省强对流天气分类预报模型，并不断训练及检验模型，得出以下结论：经2015年～2019年模型测试机测试模型，模型整体误判率2%～3%，构建预报分类模型效果较为理想。虽然每个时次各类强对流天气对应的影响性较大的物理量不尽相同，但其表征的环境场基本一致，这与笔者的主观预报经验基本相符。因此，由随机森林算法筛选出的影响度较高的物理量较为准确，模型可信度较高，可以应用于日常业务。从影响性较高的各物理量的核密度估计曲线看出，有无强对流天气对应的环境场在各物理量的数值上存在明显差别，这些物理量可以辅助笔者判断是否出现强对流天气。对于出现强对流天气对应的核密度估计曲线，其分离度越高，在判断强对流天气类型时越好用，不同的强对流天气对应物理量的阈值均不同。

关键词：探空；随机森林；强对流；分类预报

中图分类号：P456　　**文献标识码：**A　　**文章编号：**1674-1064（2022）01-001-03

DOI：10.12310/j.issn.1674-1064.2022.01.001

强对流天气一般是指，由深层湿对流（DMC）产生的各种灾害性天气，如雷暴大风、短期强降水、冰雹、龙卷风等。主要发生在中小尺度天气系统，水平范围一般约是十公里到两三百公里，甚至有的水平范围只有几十米到十几公里。其生命周期很短，而且有明显的突发性，大约一小时到十小时，最短的只有几分钟到一小时[1-5]。强对流天气预报，特别是强对流天气预报分类，一直是业务天气预报的难点之一。

河北省夏季强对流天气频发，由强对流天气造成的生命和财产损失也非常巨大。目前，利用探测资料结合人工智能方法开展对河北省强对流天气分类预报的研究较少，因此开展此项工作对研究河北省强对流天气发生发展机制，提高河北省强对流天气预报预警准确率意义重大。

1　资料与方法

1.1　主要资料

文章所用资料为2000年～2019年的地面、高空等气象观测资料。探空站点选取河北省北部的张家口站和河北省南部的邢台站。

1.2　主要方法

1.2.1　随机森林算法

随机森林算法是由美国加州大学伯克利分校的Breiman提出的一种统计学习理论，其是指一种使用多棵树对样本进行训练和预测的分类器，输出类别是个体生成的类别中的大多数树木，取决于数量。其基本单元是决策树，也称为分类回归树。随机森林使用bagging方法组合决策树，其核心是重抽样自举法。

首先，对样本为N的原始样本集S进行有放回的随机抽样，得到一个容量为N的随机样本S_1（称为自举样本）。

其次，将自举样本视为训练样本，建立分类树T_1，重复以上两个步骤M次，最终得到M个自举样本S_1，S_2，…，S_M以及M个预测模型T_1，T_2，…，T_M。

最后，组合M个决策树的预测模型，通过投票得出最终预测结果，如图1所示。

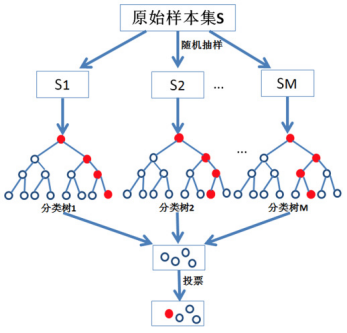

图1　随机森林分类结构图

作者简介：张玉婷（1990—），女，河北衡水人，本科，工程师，研究方向：天气预报及环境气象。

1.2.2　重要性因子评价原理与核密度估计曲线

文章采用预测精度的平均下降量测度输入变量对输出变量的重要性，即该解释变量越重要，对预报结果的影响越大。

对于随机森林算法筛选出的对于预报结果影响较大的物理量，为了更加精确地描述其分布特点，文章采用在各个物理量的频率分布图上添加核密度估计曲线的方法，将频率转化为概率密度，可更加直观地对比不同类型的强对流天气对应的物理量的分布情况及相应数值。

2　模型训练与结果分析

2.1　误差分析

模型建立后，要不断训练模型以达到最优，通过测试集的测试，使其精度满足要求后才能被应用。笔者使用2000年～2014年的数据训练模型，使用2015年～2019年的数据测试模型，整体误判率为邢台探空站8时2%、20时3%，张家口探空站8时2%、20时3%，整体误判率较小，说明构建的预报分类模型效果比较理想。

表1　2015年～2019年模型训练集预测误差表

站号	时次	误判率
53798	8时	雷暴大风1.5%
		短时强降水1.7%
		冰雹2.4%
	20时	雷暴大风2.7%
		短时强降水2.9%
		冰雹3.3%
54401	8时	雷暴大风2.1%
		短时强降水2.2%
		冰雹2.2%
	20时	雷暴大风2.8%
		短时强降水2.9%
		冰雹3.3%

为说明模型的预测效果，笔者随机抽取2011年6月7日及2012年7月26日发生在邢台的两次强对流天气，如图2所示。

2011年6月7日，邢台出现一次小范围雷暴大风天气，此次过程模型预报出现偏差，未能报出。2012年7月26日，邢台中西部出现一次范围较大的强对流天气，临城、内丘、邢台、任县、沙河均出现雷暴大风和短时强降水，此次过程模型预报较为准确，对这两种强对流天气类型均有体现。

2.2　强对流分类预报影响因子重要性排序

在模型计算过程中，可根据预测精度的平均下降量计算各物理量的重要程度，值越大表示越重要。如图3所示，邢台探空站8时对出现冰雹较为重要的物理量为沙氏指数、瑞士第一、第二雷暴指数、K指数、强天气威胁指数等，可见热力因子和综合指数是对其影响较大的因素，前几项与环境温度、空气质量二者的温度差有关，强天气威胁指数则反映了不稳定能量与风速、风向切变对风暴强度的综合作用。这表明，不稳定层结及大的垂直风切变与冰雹产生密切相关。对于短时强降水，其影响较大的因素为热力因子和水气因子，例如各高度处温度、整层比湿积分、云层厚度等，这表明，高温高湿的环境产生的深厚湿对流更有利于产生短时强降水。对产生雷暴大风影响较大的物理量，为瑞士第一、第二雷暴指数、静力能条件稳定度、沙氏指数、对流稳定度指数等。这表明，上干冷下暖湿的不稳定层结更有利于出现雷暴大风。即使影响冰雹和大风的前十项重要因子较为相似，也可以很好地将其区分。对于冰雹，要求沙氏指数为0左右，对流稳定度指数为－4左右。对于雷暴大风，沙氏指数在2左右，对流稳定度指数为－5最为合适。分析邢台站20时及张家口站8时、20时的模型计算结果可知，虽然每个时次各类强对流天气对应的影响性较大的物理量不尽相同，但其表征的环境场基本一致，这与笔者的主观预报经验基本相符。因此，由随机森林算法筛选出的影响度较高的物理量较为准确，模型可信度较高，可以应用于日常业务。

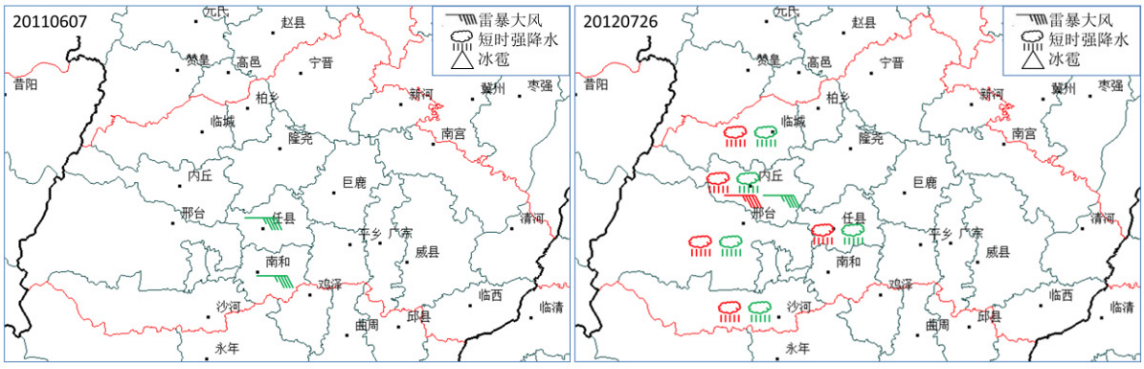

图2　2011年6月7日和2012年7月26日邢台市强对流天气实况（绿）与预测（红）

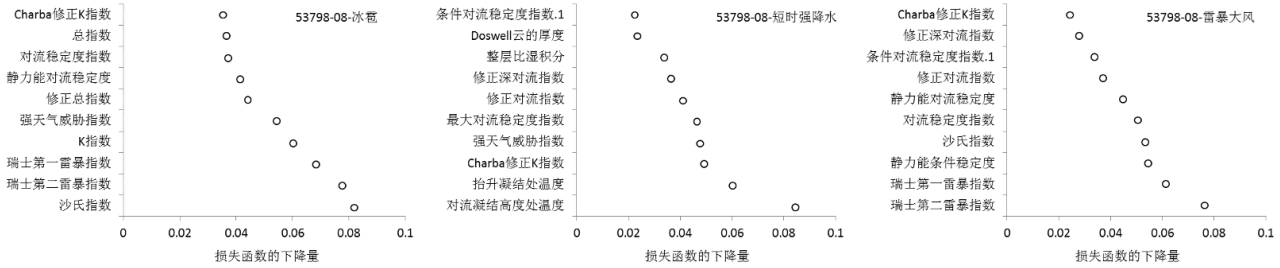

图3 邢台探空站8时随机森林算法对强对流天气分类预报前10项因子的重要性排序

2.3 预报因子特征分析

随机森林算法可以自动筛选物理量的重要性，再结合核密度估计曲线可以直观反映物理量在分类中的作用和阈值，为强对流预报提供参考。从影响性较高的各物理量的核密度估计曲线可以看出，有无强对流天气对应的环境场，在各物理量的数值上存在明显差别，这些物理量可以辅助笔者判断是否出现强对流天气。尤其是对流稳定度指数、沙氏指数、条件对流稳定度指数、整层比湿积分，无强对流时（红色线）与其他曲线出现了明显的分离度。如无强对流时，对流稳定度指数约为13～14，而有强对流时为－3～－5。出现强对流时，沙氏指数对应数值为－1～－5，而这个区间对应无强对流的低概率密度区。表2为各高影响物理量的核密度估计，可为日常预报业务提供参考。

表2 各高影响物理量的核密度估计

	雷暴大风	短时强降水	冰雹	无强对流
Charba 修正 K 指数	39	40	36	36
K 指数	32	30	26	28
对流凝结高度	870	900	760	740
对流凝结高度处温度	18	22	15	17
对流稳定度指数	−5	0	−4	12
沙氏指数	−1	−1	−2	8
抬升指数	−1	−2	−1	1
条件对流稳定度指数	−2	−5	−2	3
修正深对流指数	37	37	26	35
整层比湿积分	3 900	4 000	2 500	500

强对流天气对应的三条曲线分离度越高，判断强对流天气类型时越好用。如－20 ℃高度，对于冰雹，邢台探空站为8 500 m～8 600 m，张家口探空站约7 500 m左右，均高于雷暴大风和短时强降水。冰雹对应的修正深对流指数为27～28，短时强降水和雷暴大风约为36～37，说明这两个物理量对于判断冰雹天气比较好用。同理，整层比湿积分、对流凝结高度、对流凝结高度处温度，则可将短时强降水与雷暴大风、冰雹区别开来。

通过模型计算可知，某些物理量对应的核密度估计曲线出现了明显的双峰分布，如雷暴大风的对流凝结高度核密度估计曲线、短时强降水的整层比湿积分核密度估计曲线，这可能与季节变化相关，文章不作赘述。所以，要求研究者不能只关注其中一个阈值，而是要综合分析季节、天气实况等要素。

3 结语

2000年～2019年，邢台共出现雷暴大风111次、短时强降水76次、冰雹19次，张家口共出现雷暴大风245次、短时强降水60次、冰雹122次。

笔者使用2000年～2014年的数据训练模型，并使用2015年～2019年的数据测试模型，整体误判率为邢台探空站8时2%、20时3%，张家口探空站8时2%、20时3%，整体误判率较小，模型效果较为理想。

虽然每个时次各类强对流天气对应的影响性较大的物理量不尽相同，但其表征的环境场基本一致，这与笔者的主观预报经验也基本相符。因此，由随机森林算法筛选出的影响度较高的物理量较为准确，模型可信度较高，可以应用于日常业务。

有无强对流天气对应的环境场，在各物理量的数值上存在明显差别，在日常业务工作中，这些物理量可以辅助研究者判断是否出现强对流天气。强对流天气对应的核密度估计曲线分离度越高，在判断强对流天气类型时越好用，不同的强对流天气对应物理量的阈值均不同。相对于大量的计算机运算数据，文章的数据量稍显不足。因此，在以后的日常工作中，要对此模型不断进行动态训练，促使结果达到更优。

参考文献

[1] 曹艳察,田付友,郑永光,等.中国两级阶梯地势区域冰雹天气的环境物理量统计特征[J].高原气象,2018(1):185-196.

[2] 樊李苗,俞小鼎.中国短时强对流天气的若干环境参数特征分析[J].高原气象,2013(1):156-165.

[3] 方翀,王西贵,盛杰,等.华北地区雷暴大风的时空分布及物理量统计特征分析[J].高原气象,2017(5):1368-1385.

[4] 费海燕,王秀明,周小刚,等.中国强雷暴大风的气候特征和环境参数分析[J].气象,2016(12):1513-1521.

[5] 高晓梅,俞小鼎,王令军,等.鲁中地区分类强对流天气环境参量特征分析[J].气象学报,2018(2):196-212.