李文娟，赵放，郦敏杰，等，2018. 基于数值预报和随机森林算法的强对流天气分类预报技术[J]. 气象，44(12)：1555-1564.

基于数值预报和随机森林算法的强对流天气分类预报技术*

李文娟[1] 赵 放[1] 郦敏杰[2] 陈 列[1] 彭霞云[1]

1 浙江省气象台，杭州 310017

2 浙江省杭州市气象台，杭州 310057

提 要： 随机森林算法是当前得到较为广泛应用的机器学习方法之一，有着很高的预测精度，训练结果稳定，泛化能力强，解决多分类问题有明显优势。本文将随机森林算法应用于强对流的潜势预测和分类，分短时强降水、雷暴大风、冰雹和无强对流四种类别，基于 2005—2016 年 NCEP 1°×1°再分析资料计算的对流指数和物理量，开展强对流天气的分类训练、0～12 h 预报和检验，经 2015—2016 年独立测试样本检验表明，针对强对流发生站点的点对点检验，整体误判率为 21.9%，85 次强对流过程基本无漏报，模型尤其适用于较大范围强对流天气。随机森林算法筛选的因子物理意义较为明确，和主观预报经验基本相符，模型准确率高，可用于日常业务。

关键词： 强对流分类，对流指数，物理量，随机森林

中图分类号： P456 **文献标志码：** A **DOI：** 10.7519/j.issn.1000-0526.2018.12.005

Forecasting and Classification of Severe Convective Weather Based on Numerical Forecast and Random Forest Algorithm

LI Wenjuan[1] ZHAO Fang[1] LI Minjie[2] CHEN Lie[1] PENG Xiayun[1]

1 Zhejiang Meteorological Observatory, Hangzhou 310017

2 Hangzhou Meteorological Observatory of Zhejiang Province, Hangzhou 310057

Abstract: The random forest algorithm is currently one of the more widely used machine learning methods, featuring high prediction accuracy, stable training results and generalization ability, and has obvious advantages in solving the problem of multi-classification. This paper applies the random forest algorithm to the prediction and classification of severe convective weather, which is divided into four categories: short-time heavy rainfall, thunderstorm gale, hail and no severe convection. Then, based on the data of convection index and physics calculated from the NCEP data of 2005－2016, the training, 0－12 h forecasting and testing of classified severe convection are carried out. The results show that the whole misjudgment rate is 21.9% that is calculated out of the independent data of 2015－2016. It has almost no omission in 85 examples of severe convective weather and the model is especially suitable for larger range of severe convective weather. The physical meaning of the factors used in random forest algorithm is relatively clear, and basically consistent with the subjective forecasting experience. It can be used in daily forecasting operation.

Key words: severe convection classification, convective index, physical quantity parameter, random forest (RF)

* 国家气象中心关键技术项目[YBGJXM(2018)02-13]和浙江省科技厅重点项目(2017C03035)共同资助

2017 年 9 月 15 日收稿； 2018 年 3 月 14 日收修定稿

第一作者：李文娟，主要从事强对流天气预报及其研究. Email：liwenjuan1998@163.com

通信作者：赵放，主要从事短时临近天气预报与雷达资料开发. Email：e-zhaofang@163.com

引 言

气象学中，对流指的是大气中由浮力产生的垂直运动所导致的热力输送，强对流天气通常指的是由深厚湿对流(DMC)产生的包括冰雹、大风、龙卷、强降水等各种灾害性天气，具有突发性、生命史短、局地性强、易致灾等特点。强对流天气预报尤其是分类强对流天气一直是业务天气预报的难点之一，热动力物理参数敏感性分析及利用“配料法”、统计分析方法以及高分辨率数值模式进行强对流客观预报方法的研究逐渐成为预报强天气潜势的基础(郑永光等，2015；2017；田付友等，2015；漆梁波，2015；雷蕾等，2011)。Doswell Ⅲ(2001)、俞小鼎等(2012)、孙继松等(2014)系统总结了DMC和不同类型强对流天气(冰雹、雷暴大风、短时强降水和龙卷)发生发展的环境条件、中尺度结构和特征，这些条件和结构特征是目前进行强对流天气分类预报的物理基础。近几年国内一些学者基于数值模式计算的对流参数利用配料法和模糊逻辑法开展了分类强对流潜势预报的业务化试验。曾明剑等(2015)基于中尺度数值模式预报的对流参数，综合历史频率分布和权重分配，构建了分类强对流天气预报概率，并以优势概率作为分类判据，做出强对流分类预报。雷蕾等(2012)将统计的强对流天气判别指标应用到数值模式(快速更新同化系统)，计算模式格点上的强对流发生概率，并针对冰雹、雷暴大风和短时暴雨天气下不同物理量的阈值范围，实现了对强对流的分类概率预报。机器学习等人工智能的方法多应用在强对流临近识别和概率预报中，Mecikalski et al(2015)使用Logistic回归和人工智能随机森林(random forest，RF)等方法发展了基于卫星资料和数值模式资料的对流初生(CI)临近概率预报技术。李国翠等(2014)和张秉祥等(2014)基于雷达三维组网数据利用模糊逻辑方法分别开发了雷暴大风和冰雹的自动识别算法；周康辉等(2017)将模糊逻辑算法用于雷暴大风的监测识别，实现了雷暴大风和非雷暴大风的有效区分；修媛媛等(2016)用机器学习中有监督学习模型支持向量机(support vector machines，SVM)来进行强对流天气的识别和预报。

RF算法在近几年实际应用中得到了广泛关注，已经成为数据挖掘、模式识别等领域的研究热点，在生态学、水文学、经济学、医学等领域得到了广泛应用(张雷等，2014；李欣海，2013；石玉立和宋蕾，2015；侯俊雄等，2017；Belgiu and Drăguţ，2016；Chen et al，2017)。RF是一种基于分类回归树的数据挖掘方法，是由Breiman和Cutler在2001年提出的一种较新的机器学习技术(方匡南等，2011)。RF算法通过聚集大量分类树来提高模型预测精度，与决策树一样，可用来解决分类和回归问题，预测精度很高，在异常值和噪声方面有很高的容忍度，且不易出现过度拟合现象(Breiman，2001)。国内外学者将RF算法与传统的神经网络、SVM、Logistic等机器学习方法做了一些对比，黄衍和查伟雄(2012)证明RF泛化能力在多分类问题上优于SVM；梁慧玲等(2016)在基于气象因子的塔河地区林火发生预测模型研究中，得出RF模型的预测准确率高于传统Logistic模型10%左右；余胜男等(2016)研究表明RF模型预测精度较高、稳定性好、泛化能力强，能有效预测年、月降水量，与BP神经网络模型和SVM模型相比，RF模型效率更高、性能更优，尤其适用于大样本的逐月降水量预测；白琳等(2017)和Zhang et al(2017)研究均证明RF算法比传统的多元线性回归的结果更为理想，处理非线性和分级关系更具优势；Naghibi et al(2017)应用RF、RFGA(random forest genetic algorithm)、SVM三种模型评估地下水资料的潜势，发现RF和RFGA比SVM更高效且更准确；Peters et al(2007)基于RF和Logistic模型建立了生态水文分布模型，对比得出RF的预测误差小于Logistic模型；Kampichler et al(2010)通过5种机器学习方法对比，发现RF明显优于神经网络、SVM等方法；Coussement and Van den Poel(2008)比较了SVM、Logistic模型和RF的客户流失预测能力，RF始终优于SVM和Logistic。由此可见，大量的研究表明RF算法在不同领域已取得较好的应用效果。

RF算法应运而生，给解决很多问题带来了新的方向，但将RF应用于强对流的分类预测，相关研究为数不多。传统的配料法等通过挑选对不同类型强对流天气具有指示意义的物理量，根据历史个例的统计结果挑选预报因子，预测结果完全取决于天气学要素和物理量对强对流天气发生发展物理条件的代表性，而人工智能等机器学习算法可以建立在大数据集的应用基础上，通过智能化的筛选、组合多种因子进行预测分类，尤其在多分类预测方面有一定的优势，能够处理很高维度的数据，在训练完后，

能够给出特征量的重要性排序,可以很好地预测多达几千个解释变量的作用。因此,本文将 RF 算法尝试性地应用于分类强对流的潜势预测,构建反映强对流发生发展环境条件的大数据集,通过训练学习达到预测分类的目的。

1　RF 算法

1.1　RF 算法原理

RF 是由加州大学伯克利分校 Breiman(2001)提出来的一种统计学习理论。RF 的基本组成单元是决策树,又称为分类回归树。基本思想是一种二分递归分割方法,在计算过程中充分利用二叉树,在一定的分割规则下将当前样本集分割为两个子样本集,使得生成的决策树的每个非叶节点都有两个分枝,这个过程又在子样本集上重复进行,直至不可再分成为叶节点为止。由于单棵决策树模型往往精度不高,且容易出现过拟合问题,为此需要通过聚集多个模型来提高预测精度,RF 中采用的是 Bagging 方法来组合决策树,其核心是重抽样自举法,第一步,对样本量为 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(称自举样本),第二步,将自举样本视为训练样本,建立分类树 T_1,重复上述两步 M 次,最终得到 M 个自举样本 $S_1,S_2,\cdots,S_M$ 以及 M 个预测模型 T_1,$T_2,\cdots,T_M$。然后组合 M 个决策树的预测模型,通过投票得出最终预测结果。RF 的思路就是训练出在某一个方面有决策能力的决策树,这个决策树几乎不存在过度复杂和过分拟合数据的问题,相对而言它是一个弱决策树,但是多个方面的弱分类器集成能够形成一个强大的分类器。

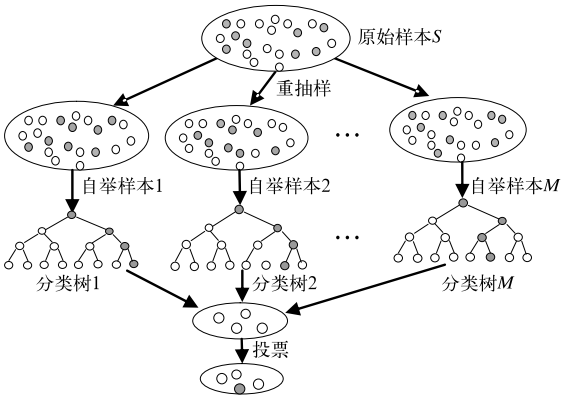

图 1　RF 分类结构图

Fig. 1　Random forest classification structure

1.2　泛化误差与重要性因子评价原理

RF 用 Bagging 方法生成训练集,样本容量为 N 的总训练集 S 中每个样本未被抽取的概率为 $(1-1/N)^N$,当 N 足够大时,$(1-1/N)^N \to 1/e=0.368$,这表明原始样本集中接近 37% 的样本不会出现在训练集中,这些数据称为袋外(out-of-bag, OOB)数据,使用这些数据来估计模型的性能称为 OOB 估计。OOB 数据可以用来估计决策树的泛化误差,或用来计算单个特征的重要性。泛化误差是指分类器对训练集之外数据的误分率,泛化误差越小表示分类器性能越好,相反则表明分类器性能较差。每一棵树都可以得到一个 OOB 误差估计,将森林中所有树的 OOB 误差估计取平均,即可得到 RF 的泛化误差估计。Breiman 通过试验已经证明,OOB 误差是无偏估计,并且相对于交叉验证,OOB 估计是高效的,且其结果近似于交叉验证的结果(杨柳和王钰,2015)。

RF 测度输入变量重要性的基本思路是:对于解释变量重要性,一个直观的评价标准是,该变量越重要,其对预报结果的影响也越大。RF 算法的解释变量重要性评价采用类似标准:对所有检验样本,随机打乱某一解释变量取值,采用原 RF 算法对检验样本进行再次预报,袋外拟合误差增加愈多,该解释变量愈重要,表现为各类别的预测置信度变化明显,总体预测精度变化明显,袋外拟合误差增加量可用于定量评价解释变量重要性。因此,本文对于输入变量重要性的评判指标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。

2　模型建立过程

本文将 RF 算法应用于强对流的环境场分类,基于 NCEP 1°×1° 08 时的分析场资料计算的若干对流指数和物理量指标作为输入变量,输出变量为短时强降水、雷暴大风和冰雹三种类别的强对流天气和无上述强对流天气。从理论和经验角度,不同的环境场有利于不同灾种的强对流发生,因此,采用多种物理量全面描述强对流发生的环境场,再应用机器学习算法,对强对流天气进行预测及分类,预报模型的建立过程如图 2,具体步骤为:

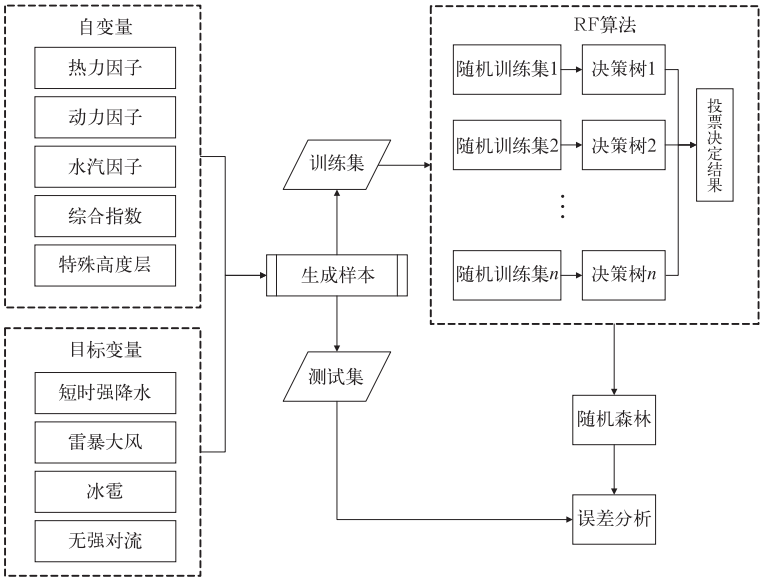

图 2 RF 预报模型的建立过程

Fig. 2 Flow chart of RF forecasting model

2.1 选取预报因子

资料选用 2005—2016 年的强对流监测资料，基于 08 时 NCEP 1°×1°资料计算若干对流指数和 500、700、850、925 hPa 各层物理量场，这些要素涵盖了强对流天气的构成要素包括静力稳定度、水汽、能量及垂直风切变等动力热力因子。代表要素见表 1，表中物理量要素根据其计算公式和物理意义简单归为五类，其分类方法和条件参考文献(刘建文等，2015)。文后附录中列出了四种综合指数和条件-对流稳定度指数的计算公式和物理意义。因此，基于数值模式分析场资料计算的若干对流指数和物理量场共 68 类组成预报因子数据集，构建强对流分类预报模型。

2.2 选取目标变量

目标变量分为四类，分别是短时强降水、雷暴大风和冰雹等强对流天气以及无强对流天气。以 2005—2016 年浙江省 69 个基准站点的监测实况为标准，实况选取时间段为 00—20 时，任一站点观测到冰雹记为一次过程，共监测到 75 次冰雹过程；为了区分强降水拖曳产生的局地性大风，雷暴大风样本选取影响范围相对较大的过程，至少 3 个站点出现 8 级或以上雷暴大风记为一次过程；短时强降水过程历史样本量较多，考虑到数据平衡性问题，仅选取全省11个地、市代表站点的短时强降水过程，任

表 1 应用于 RF 算法的主要预报因子类型和要素

Table 1 Main types and elements of main predictors applied to RF algorithm

水汽因子	整层可降水量(P_w)
	比湿(q)
	水汽通量(Q_{flux})
	相对湿度(R_h)
	水汽通量散度(Q_{fdiv})
	温度露点差($T-T_d$)
	925 hPa 露点温度(T_{d925})
动力因子	散度(D_{iv})
	涡度(V_{or})
	垂直速度(ω)
	垂直风切变(S_{hr})
高度层因子	0℃层高度(Z_{ht})
	−10℃层高度(M_{ht})
	−20℃层高度(F_{ht})
热力因子	假相当位温(θ_{se})
	K 指数(K_i)
	沙氏指数(S_i)
	最佳抬升指数(B_{li})
	最佳不稳定能量(B_{CAPE})
	不稳定能量($CAPE$)
	总指数(TT)
	850 与 500 hPa 假相当位温差($\theta_{se\,850-500}$)
	850 hPa 温度(T_{850})
	条件-对流稳定度指数(I_{lc})
	下沉有效位能(D_{CAPE})
	温度差($T_{850\sim500}$)
综合指数	强天气威胁指数(S_{weat})
	瑞士雷暴指数(S_{wiss00})
	修正深对流指数(M_{dci})
	风暴强度指数(S_{si})

一站点出现 20 mm·h^{-1}以上降水,记为一次过程。三种强对流天气往往是相伴产生的,在选取的强对流样本中,同时观测到短时强降水、雷暴大风或冰雹的有 13 次,观测到冰雹和雷暴大风相伴产生的有 10 次,因此,在分类过程中遵循一定的原则,根据灾害影响程度对强天气进行侧重分类,对于雷暴大风和短时强降水均出现的情况,一般记为雷暴大风过程,而雨强在 50 mm·h^{-1}及以上的极端降水则同时记为短时强降水和雷暴大风过程;对于冰雹和短时强降水均出现的情况,一般记为一次冰雹过程;对于雷暴大风和冰雹均出现的情况,则记为冰雹过程;对于无强对流样本,以 2010—2015 年全省 69 个基准站无雷暴日和 10 mm·h^{-1}以下的弱降水样本为主。因此,模型训练期为 2005—2015 年共 1026 个样本,模型验证期为 2015—2016 年共 406 个独立测试样本(表 2)。

表 2　模型训练和测试样本集

Table 2　Samples of model training and testing

强对流分类	短时强降水	雷暴大风	冰雹	无强对流
训练样本集	255	181	73	516
测试样本集	163	82	2	159

2.3　预报模型构建过程

设 RF 包括 M 颗分类树,在第 i 颗决策树的建立过程中,首先通过随机方式选取 k 个输入变量构成候选变量子集 X_i,依据变量子集 X_i 将建立一颗充分生长的决策树且无需剪枝。确定 k 的依据是:第一,决策树对袋外观测的预测精度,也称决策树的强度;第二,各决策树间的相互依赖程度,也称决策树的相关性。森林中包含的众多决策树形成一个组合预测模型,利用投票原则确定最后的预测结果。

本研究基于 R 语言 RF 程序包进行强对流分类预报研究。RF 算法包含 2 个参数,即 M 颗决策树和每棵树的输入变量 k,M 越大,RF 算法过拟合效应越小;k 越大,子预报模型间差异性越小。对于分类树,变量子集的大小 k 默认为$\sqrt{P}$,P 为预报因子个数。M 取值为 500,k 取值为 8,以选取的 68 个预报因子作为解释变量(自变量),1026 个分类强对流作为目标变量,构建 RF 模型对解释变量进行重要性评价。

2.4　误差分析

利用 OOB 数据估计模型的泛化误差,为了更好地检验模型的预报性能,再利用检验期独立数据集进行验证,采取泛化误差和独立样本测试两种方式可以更全面地说明模型的预报效果。基于 RF 算法对全部观测做预测,计算混淆矩阵和整体的误判率。整体误判率=分类错误的样本数/总的样本数。

3　模型训练与预报结果分析

3.1　评判精度分析

3.1.1　泛化误差

建立模型后需对训练模型与测试结果进行评估,其评估精度满足要求后模型才能被应用。以 2005—2015 年训练期的强对流样本基于 RF 预测模型构建的 OOB 误差见表 3。由表 3 可见,RF 对全部观测进行预测,预测误差很小,仅为0.39%,而单棵树的预测误差约 20%,说明由 RF 构建的预报分类模型效果比较理想。

表 3　2005—2015 年模型训练期 OOB 预测误差表

Table 3　Errors of OOB prediction in the model training period during 2005－2015

实况	预报				
	无强对流	冰雹	强降水	雷暴大风	整体误判率/%
无强对流	516				
冰雹		73			0.39
强降水			255		
雷暴大风		1	2	178	

3.1.2　独立样本测试

根据建立的 RF 模型,对 2015—2016 年检验期的 406 次独立数据进行预测。由于冰雹样本较少,选取了影响浙江省较为严重的三次过程进行预测检验,结果见表 4。检验期的独立测试样本均是点对点的验证,即针对站点监测到短时强降水、雷暴大风的实况进行预测,整体误判率为 21.9%。由于 2016 年基准站点没有观测到冰雹,不能准确判断冰雹过程,因此对于预报出现冰雹必然会增加一定的错误率。三次冰雹过程中一次判断为雷暴大风过程,实际情况既出现了冰雹又伴随大范围的雷暴大风,另外两次过程是 2014 年 3 月 19 日和 2015 年 4 月 5 日均发生了影响较严重的大冰雹天气,模型均准确判断出;无强对流过程判断准确率高。由于短时强降水和雷暴大风的实况很难明确客观的分类,导致强降水和雷暴大风的误判率相对较高;短时强降水

和雷暴大风站点预测存在部分漏报的情况，但是从预报过程的检验来看，共有 85 次过程，包括局地强降水过程和较大范围雷暴大风过程，基本无漏报，预报落区偏差是导致站点漏报的主要原因。对于 5 个基准站点及以上出现强对流天气的较大范围过程，共有 12 次，仅一次大风过程误判为短时强降水，其余都预报正确；40 次无强对流过程，4 次为空报。总体来说，基于 RF 的预报分类模型效果比较理想，强对流过程基本能准确预报，尤其适用于较大范围的强对流天气。但是，由于强对流天气观测的原因，尽管我们采用了 11 年的观测数据，但还是存在样本不足的问题，使得 RF 模型存在一些缺陷，主要是存在训练不充分的情况，且模型的训练期样本在分类的过程中，会存在混淆的情况，因此导致了一定的错误率的增加。

表 4 2015—2016 年模型检验期预测误差表

Table 4 Errors in the model testing period during 2015－2016

实况	预报				
	无强对流	冰雹	强降水	雷暴大风	整体误判率/%
无强对流	148		8	3	21.9
冰雹		2		1	
强降水	18	2	133	10	
雷暴大风	19	4	25	34	

图 3 列出了 2016 年的两次预测个例，简单说明模型的预报效果，2016 年两次过程均出现了较大范围的雷暴大风和短时强降水过程，5 月 5 日过程据了解在浙南出现了局地小冰雹，6 月 1 日过程浙南出现了较大范围的雷暴大风，预测效果来看，预报模型对出现的灾害性天气都有所反映，包括冰雹和雷暴大风的落区，不足的是，预报落区比实况范围大，落区也存在一定的偏差。

3.2 强对流分类指标的重要性分析

RF 在计算过程中能根据预测精度的平均下降量计算各指标重要度。图 4 是 RF 算法对影响强对流分类因子的重要性排序，值越大表示越重要，从中可见，沙氏指数（S_i）在分类过程中的重要性高于其他指标，表明其对强对流分类的贡献程度最大。预测精度平均减少量筛选的前几位的因子分别是 850 和 500 hPa 的温度差（$T_{850\sim500}$）、低层相对湿度（$R_{h\,850}$ 和 $R_{h\,925}$）、整层可降水量（P_w）、总指数（TT）、0～6 km 风垂直切变（$S_{hr\,0\sim6\,km}$）、最大抬升指数（B_{li}）、强天气威胁指数（S_{weat}）、低层假相当位温（θ_{se850}）及风暴强度指数（S_{si}）。

从输入变量对分类强对流的重要性指标排序来看（图 5），区分有无强对流的指标，能量、水汽条件是发生强对流天气的必要条件，其中稳定度因子贡献较为显著，其中 S_i 指数和 B_{li} 等稳定度指数表现较好，这两个指标是强对流主观预报的优选指标，其次综合指数有较好的表现，如 S_{weat}，S_{si}，M_{dci} 可以综合反映中低层热力稳定度特性及适宜风暴发生动力环境对风暴发生所产生的共同作用。从短时强降水的重要性因子排序可见，短时强降水更倾向于表征水汽条件的因子，如 P_w、低层相对湿度、比湿、各层 θ_{se} 表征整层高温高湿的环境场，即深厚的湿对流有利于短时强降水的发生。雷暴大风的环境场特征除了层结不稳定，代表环境温度直减率的因子 $T_{850\sim500}$ 的贡献较为显著，环境大气有较大的温度递减率，既有利于强上升气流，也有利于强下沉气流。此外，下沉有效位能（D_{CAPE}）代表干下沉气流的作用以及低层相对湿度条件（$R_{h\,850}$ 和 $R_{h\,925}$），对雷暴大风的贡献也较显著（图略）。冰雹天气的环境场，贡献最为显著的是不同高度层的风垂直切变，尤其是 $S_{hr\,0\sim6\,km}$，其次－20℃高度层（F_{ht}）、－10℃高度层（M_{ht}）、0℃高度层（Z_{ht}）等特性层的高度对于冰雹的形成起重要作用。综合指数中 S_{si} 对冰雹天气的贡献较为显著，S_{si} 计算方法和垂直风切密切相关，由 0～3600 m 的环境风垂直切变和 CAPE 决定（刘建文等，2005）；此外，温度直减率（$T_{850\sim500}$）较大同样有利于冰雹天气的发生。由此可见，RF 算法筛选的因子的物理意义较为明确，和主观预报经验基本相符，因此，RF 建立的强对流分类模型可信度较高，可以应用于日常业务。

3.3 预报因子特征分析

针对 RF 算法筛选的重要物理量绘制核密度估计图（图 6），为了更加精确刻画变量的分布特点，可在变量的频率分布图上添加核密度估计曲线，将频率转化为概率密度，可直观对比不同组数值的分布形状以及不同组之间的重叠程度（李文娟等，2017）。从不同灾害性天气的对流指数分布可以直观地看出，有无强对流天气表现在环境场的物理量指标存在较明显的差别，尤其是稳定度指标 S_i，K_i，B_{li} 和 S_{weat}，黑色曲线（无强对流）和其他三条曲线分离度

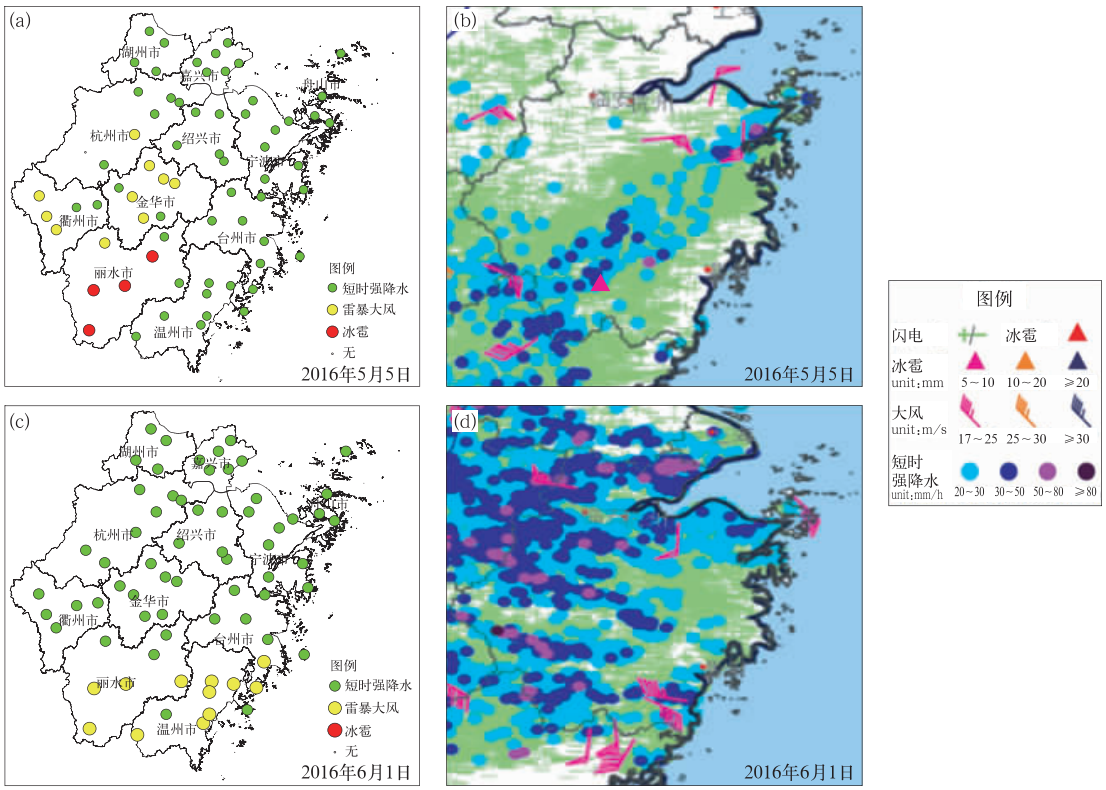

图 3　2016 年 5 月 5 日(a,b)和 6 月 1 日(c,d)强对流个例预测(a,c)和实况(b,d)的对比

Fig. 3　The prediction (a, c) and observation (b, d) of severe convection cases on 5 May (a, b) and 1 June (c, d) 2016

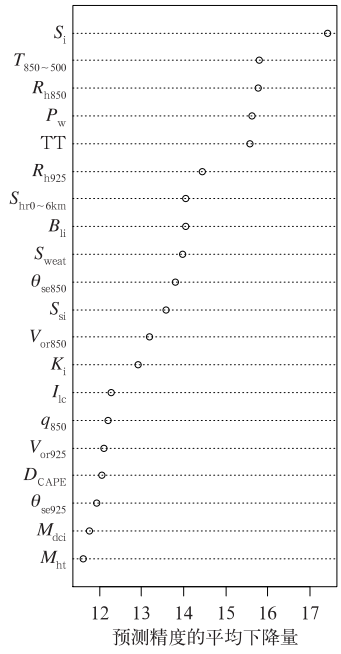

图 4　RF 对强对流分类前 20 项因子的重要性排序

Fig. 4　The importance ranking of the first 20 factors of severe convection classification by RF

较高,峰值处于不同的阈值区间,例如,B_{li}发生强对流的峰值在－5℃左右,无强对流时一般在 0℃以上;S_{weat}发生强对流的峰值一般集中在 250～300,而这个区间对应无强对流的低概率密度区。S_{si}在 50～70 易发生强对流天气,冰雹天气在 60～70 具有高概率密度,而短时强降水集中在 50 左右,雷暴大风的分布较不集中,说明 S_{si}对冰雹天气的指示性较好。$T_{850\sim500}$可以较好地区分短时强降水和风雹类强对流,25～27℃是风雹类强对流的集中区,而短时强降水分布在 22～24℃。和水汽条件密切相关的因子如 P_{w},$R_{\mathrm{h}\,925}$,K_{i},可以较好地指示短时强降水的发生条件,如短时强降水 P_{w} 峰值在 60 mm,而风雹类强对流在 50 mm,此外,雷暴大风的 $R_{\mathrm{h}\,925}$峰值在 60%～80%,明显低于短时强降水 90%的相对湿度;D_{CAPE}可以较好地表征雷暴大风类强对流天气,和短时强降水的曲线存在一定的分离度。同时发现,冰雹的环境指标有明显的双峰特征,如 P_{w},$R_{\mathrm{h}\,925}$,D_{CAPE},分析环境指标不同季节地演变特征可以解释双峰特征,春季是浙江省冰雹天气的高发季,指标和夏季相比有明显的季节性特征,这里不做详

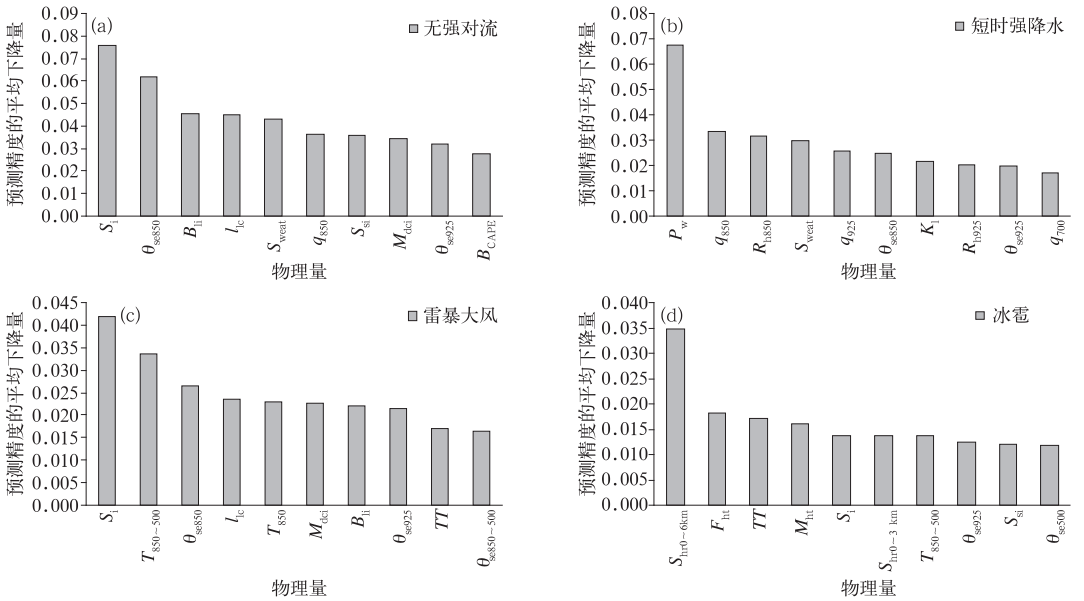

图 5 RF 对各类别前 10 项因子的重要性测度排序

(a)无强对流,(b)短时强降水,(c)雷暴大风,(d)冰雹

Fig. 5 The importance ranking of the first 10 factors of each category by RF

(a) no severe convection, (b) short-time heavy rainfall,

(c) thunderstorm gale, (d) hail

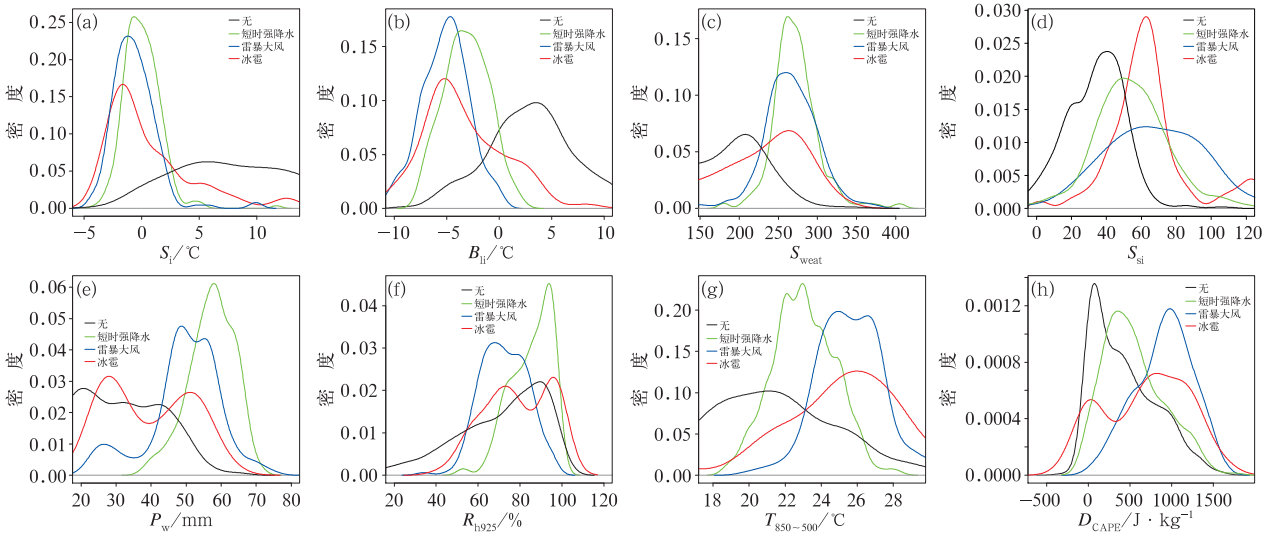

图 6 物理量核密度估计分布:

(a)沙氏指数(S_i),(b)最大抬升指数(B_{li}),(c)强天气威胁指数(S_{weat}),

(d)风暴强度指数(S_{si}),(e)整层可降水量(P_w),(f)925 hPa 相对湿度($R_{h\,925}$),

(g)温度差 850～500 hPa ($T_{850\sim500}$),(h)下沉有效位能(D_{CAPE})

Fig. 6 The kernel density estimation charts of predictors (a) SI index (S_i), (b) max lifted index (B_{li}),

(c) threat index (S_{weat}), (d) storm intensity index (S_{si}), (e) whole atmospheric precipitable water (P_w),

(f) 925 hPa relative humidity ($R_{h\,925}$), (g) temperature difference ($T_{850-500}$),

(h) descending effective potential energy (D_{CAPE})

细讨论。因此，RF 算法可以自动筛选物理量的重要性，再结合核密度估计分布可以直观地反映物理量在分类过程中的作用及阈值区间，为主观预报提供参考。

4　结论与讨论

随着大数据时代的到来，计算机辅助预测的方法日益丰富。一般来说，机器学习算法的性能会随着数据量的增多而提高，但是随着数据量的增大模型也容易出现过拟合的现象，从而影响模型性能，例如 SVM、人工神经网络等机器学习模型都有着类似的特点，而 RF 算法具备训练结果稳定、泛化能力强的特点。因此，将 RF 算法运用于浙江省强对流天气的潜势分类，针对 2005—2015 年 NCEP 再分析资料计算的对流指数和物理量进行分类训练，建立模型预测强对流的潜势和类别。

误差分析结果表明，RF 算法建立的模型准确率高，基于袋外观测 OOB 的泛化误差仅为0.39%，基于 2015—2016 年独立测试样本的整体误判率为 21.9%，85 次强对流过程基本无漏报，模型尤其适用于较大范围的强对流天气，但是预报落区和范围仍存在一定的偏差。强对流分类因子的重要性分析表明，S_{i}，$T_{850\sim500}$，$R_{\mathrm{h}\,850}$，$R_{\mathrm{h}\,925}$，P_{w}，TT，$S_{\mathrm{hr}\,0\sim6\ \mathrm{km}}$，$B_{\mathrm{li}}$，$S_{\mathrm{weat}}$，$\theta_{\mathrm{se}850}$ 及 S_{si} 等指标对 RF 强对流分类模型的贡献较为显著；$S_{\mathrm{hr}\,0\sim6\ \mathrm{km}}$、$T_{850\sim500}$ 及 −20℃层高度对冰雹等强对流天气贡献显著。根据核密度估计分析，强对流天气 S_{weat} 集中在 250～300；S_{si} 对冰雹的指示性较强，在 60～70 具有高概率密度。$T_{850\sim500}$ 可以较好地区分短时强降水和风雹类强对流，25～27℃是风雹类强对流的集中区，而短时强降水分布在 22～24℃。短时强降水易发生在整层高温高湿的环境场，短时强降水 P_{w} 的高概率密度区在 60 mm 左右；此外，雷暴大风的 $R_{\mathrm{h}\,925}$ 峰值在 60%～80%，明显低于短时强降水 90%的相对湿度；D_{CAPE} 也可以较好地表征雷暴大风类强对流天气。

该模型也存在不足，受历史强对流样本的数量限制，训练不够充分，在业务应用的过程中，需要不断动态训练模型，加入新的训练样本以及综合更多因子才能更好地发挥强对流分类模型的作用。

参考文献

白琳，徐永明，何苗，等，2017. 基于随机森林算法的近地表气温遥感反演研究[J]. 地球信息科学学报，19(3)：390-397.

方匡南，吴见彬，朱建平，等，2011. 随机森林方法研究综述[J]. 统计与信息论坛，26(3)：32-38.

侯俊雄，李琦，朱亚杰，等，2017. 基于随机森林的 $PM_{2.5}$ 实时预报系统[J]. 测绘科学，42(1)：1-6.

黄衍，查伟雄，2012. 随机森林与支持向量机分类性能比较[J]. 软件，33(6)：107-110.

雷蕾，孙继松，王国荣，等，2012. 基于中尺度数值模式快速循环系统的强对流天气分类概率预报试验[J]. 气象学报，70(4)：752-765.

雷蕾，孙继松，魏东，2011. 利用探空资料判别北京地区夏季强对流的天气类别[J]. 气象，37(2)：136-141.

李国翠，刘黎平，连志鸾，等，2014. 利用雷达回波三维拼图资料识别雷暴大风统计研究[J]. 气象学报，72(1)：168-181.

李文娟，赵放，赵璐，等，2017. 基于单站探空资料的不同强度短时强降水预报指标研究[J]. 暴雨灾害，36(2)：132-138.

李欣海，2013. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报，50(4)：1190-1197.

梁慧玲，林玉蕊，杨光，2016. 基于气象因子的随机森林算法在塔河地区林火预测中的应用[J]. 林业科学，52(1)：89-98.

刘建文，郭虎，李耀东，等，2005. 天气分析预报物理量计算基础[M]. 北京：气象出版社.

漆梁波，2015. 高分辨率数值模式在强对流天气预警中的业务应用进展[J]. 气象，41(6)：661-673.

石玉立，宋蕾，2015. 1998—2012 年青藏高原 TRMM 3B43 降水数据的校准[J]. 干旱区地理，38(5)：900-911.

孙继松，戴建华，何立富，等，2014. 强对流天气预报的基本原理与技术方法——中国强对流天气预报手册[M]. 北京：气象出版社.

田付友，郑永光，张涛，等，2015. 短时强降水诊断物理量敏感性的点对面检验[J]. 应用气象学报，26(4)：385-396.

修媛媛，韩雷，冯海磊，2016. 基于机器学习方法的强对流天气识别研究[J]. 电子设计工程，24(9)：4-7，11.

杨柳，王钰，2015. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究，32(5)：1287-1290，1297.

余胜男，陈元芳，顾圣华，等，2016. 随机森林在降水量长期预报中的应用[J]. 南水北调与水利科技，14(1)：78-83.

俞小鼎，周小刚，王秀明，2012. 雷暴与强对流临近天气预报技术进展[J]. 气象学报，70(3)：311-337.

曾明剑，王桂臣，吴海英，等，2015. 基于中尺度数值模式的分类强对流天气预报方法研究[J]. 气象学报，73(5)：868-882.

张秉祥，李国翠，刘黎平，等，2014. 基于模糊逻辑的冰雹天气雷达识别算法[J]. 应用气象学报，25(4)：414-426.

张雷，王琳琳，张旭东，等，2014. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报，34(3)：650-659.

郑永光，陶祖钰，俞小鼎，2017. 强对流天气预报的一些基本问题[J]. 气象，43(6)：641-652.

郑永光，周康辉，盛杰，等，2015. 强对流天气监测预报预警技术进展[J]. 应用气象学报，26(6)：641-657.

周康辉，郑永光，王婷波，等，2017. 基于模糊逻辑的雷暴大风和非雷暴大风区分方法[J]. 气象，43(7)：781-791.

Belgiu M, Drăguţ L, 2016. Random forest in remote sensing: a review of applications and future directions[J]. ISPRS J Photogramm Remote Sens, 114: 24-31.

Breiman L, 2001. Random forests[J]. Mach Learn, 45(1): 5-32.

Chen Tao, Trinder J C, Niu Ruiqing, 2017. Object-oriented landslide mapping using ZY-3 satellite imagery, random forest and mathematical morphology, for the three-gorges reservoir, China[J]. Remote Sens, 9(4): 333.

Coussement K, Van den Poel D, 2008. Chum prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques[J]. Exp Syst Appl, 34(1): 313-327.

Doswell III C A, 2001. Severe convective storms[M]// Meteorological Monographs. Boston: American Meteorological Society: 1-525.

Kampichler C, Wieland R, Calmé S, et al, 2010. Classification in conservation biology: a comparison of five machine-learning methods[J]. Ecol Inform, 5(6): 441-450.

Mecikalski J R, Williams J K, Jewett C P, et al, 2015. Probabilistic 0-l-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data[J]. J Appl Meteor Climatol, 54: 1039-1059. DOI: 10.1175/JAMC-D-14-0129.1.

Naghibi S A, Ahmadi K, Daneshi A, 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping[J]. Water Res Manag, 31(9): 2761-2775.

Peters J, De Baets B, Verhoest N E C, et al, 2007. Random forests as a tool for ecohydrological distribution modelling[J]. Ecolog Modell, 207(2/4): 304-318.

Zhang Huan, Wu Pengbao, Yin Aijing, et al, 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model[J]. Sci Total Environ, 592: 704-713.

附录：

对流指数	计算公式	物理意义
条件-对流稳定度指数(I_{lc})	$I_{lc}=(\theta^*_{se\,500}-\theta_{se\,0})+(\theta_{se\,500}-\theta_{se\,0})$ $\theta_{se\,0}$表示地面假相当位温，θ^*_{se}表示饱和假相当位温	条件性稳定度是考虑一小块空气上升得到的，而对流性稳定度是考虑厚度相当大的某一层空气抬升得到的，常把 I_L 与 I_C 相加称为条件-对流稳定度指数
修正的深对流指数(M_{dci})	$M_{dci}=(T_{850}+T_{d850})/2+(T_s+T_{ds})/2-LI$ T_s 与 T_{ds}分别表示地面温度和地面露点温度，LI 表示抬升指数	综合反映低层(地面～850 hPa)温湿特性及中低层条件稳定度的参数。
强天气威胁指数(S_{weat})	$S_{weat}=12T_{d850}+(TT-49)+2f_{850}+f_{500}+125(s+0.2)$ $S=\sin(\alpha_{500}-\alpha_{850})$ α_{500}与 α_{850}分别代表 500 和 850 hPa 风向	综合反映了中低层热力稳定度特性及适宜风暴发生动力环境对风暴发生所产生的共同作用。
瑞士雷暴指数(S_{wiss00})	$S_{wiss00}=SI_{850}+0.4S_{hr_{3-6}}+0.1(T-T_d)_{600}$ $S_{hr_{3-6}}$为 3～6 km 垂直风切变	当 $S_{wiss00}<5.1$ 时预报有雷暴，否则无雷暴。
风暴强度指数(S_{si})	$S_{si}=100\{2+[0.276\ln(S_{hr})+(2.011\times10^{-4}CAPE)]\}$ S_{hr}代表 0～3600 m 的环境风垂直切变；$CAPE$ 代表对流有效位能	一个由 $CAPE$ 和 S_{hr}组合的函数，可将强雷暴与非强雷暴分开。