路志英，汪永清，孙晓磊，等. 基于 Focal Loss 改进的 GBDT 模型对天津强对流灾害的预报[J]. 灾害学，2020，35(3)：34－37，50.［LU Zhiying，WANG Yongqing，SUN Xiaolei，etal. Tianjin Strong Convection Disaster Forecast Based on Improved GBDT Model of Focal Loss［J］. Journal of Catastrophology，2020，35(3)：34－37，50. doi：10.3969/j.issn.1000－811X.2020.03.007.］

基于 Focal Loss 改进的 GBDT 模型对天津强对流灾害的预报*

路志英[1]，汪永清[1]，孙晓磊[2]，贾惠珍[3]

（1. 天津大学 电气自动化与信息工程学院，天津 300072；2. 天津市海洋中心气象台，天津 300074；
3. 天津市气象台，天津 300074）

摘　要：强对流灾害是气象研究的重点，不准确的强对流灾害天气预报往往给人们的安全以及社会经济造成影响。该文利用天津 2006－2018 年的地面气象观测站点的地面物理场数据，筛选出强对流灾害性天气过程，对天津强对流灾害天气进行研究。首先对地面站点获取的数据通过主成分分析方法进行降维，然后构建基于 Focal Loss 改进的 GBDT 模型，最后通过五折交叉验证的方式进行训练与测试。结果表明，该模型对强对流灾害天气预报的命中率、误警率和临界成功指数上都有较好的表现，可为天津强对流天气的预报提供有效的依据。

关键词：强对流灾害；主成分分析；Focal Loss；GBDT 模型；交叉验证

中图分类号：X43；P456.7；X915.5　　**文献标志码：**A　　**文章编号：**1000－811X(2020)03－0034－05

doi：10.3969/j.issn.1000－811X.2020.03.007

强对流天气是突发的自然灾害性天气，主要包括冰雹和短时强降水等。这些灾害性天气的出现往往会给人们的生活造成重大的影响。据气象局数据统计，2010－2018 年，我国平均每年的强对流灾害所造成经济损失高达 130 亿元以上。针对强对流天气，国内外学者进行了大量的研究。孙燕等[1]对江淮地区强对流天气的形成原因进行了总结；王峙[2]分析了强降雨、雷暴、大风等强对流天气对高空气象探测造成的影响；郑永光等[3]总结了我国对流初生识别、对流系统强度识别和对流天气类型识别等监测技术的研究进展；Ferran-Fabro 等[4]通过数据分析的方式发现了冰雹出现以及闪电频率之间的联系。

通过实时的雷达信息对强对流天气进行外推预报是气象研究的一种方法，但是，实时雷达数据仅是实况的反映，缺乏足够预报提前量[5]。因此，基于强对流灾害发生的背景，深入挖掘物理场数据的意义，获取一种拥有足够长的预报提前量且更有参考价值的预报方法便是一个研究方向[6]。

天津处于强对流灾害频发的地区，每年的 3－9 月是天津冰雹和短时强降水等强对流灾害频发的时间。本文以天津市的地面气象站物理场数据为基础，首先对数据集进行预处理，然后将预处理后的数据集通过主成分分析(Principle Component Analysis，PCA)的方式进行降维，并使用 Focal Loss 改进的 GBDT 模型对降维后的数据集进行训练，最后使用五折交叉验证的方式验证模型的准确性。强对流灾害的预报命中率(Percent of Doom，POD)为 84.3%，临界成功指数(Critical Success Index，CSI)为 79.4%。模型运行稳定，为天津强对流灾害的预报提供了一种新的思路。

1　数据集预处理

1.1　数据获取

本文所采用的数据为天津 13 个国家级的地面气象站从 2006－2018 年每年的 3－9 月逐小时记录的物理场数据，天津地面气象观测站分布如图 1 所示。记录的数据共计 107 983 个，其中，强对流灾害的过程数据共计 588 个。

天津地面气象观测站采集的物理量有：地平面气压、海平面气压、温度、露点温度、相对湿度、水汽压、2 min 平均风向、2 min 平均风速、10 min 平均风向、10 min 平均风速等 10 个物理量。由于气象数据特征有时间的相关性，在考虑之前时间的气象过程对当前过程的影响后，本文构建的气象数据集使用了每个时刻前 3 h 的物理场数

* 收稿日期：2019－12－14　　修回日期：2020－03－31
基金项目：国家自然科学基金资助项目(51677123)
第一作者简介：路志英(1964－)，女，天津人，教授，博士，主要研究方向为数据挖掘、图像处理方面的研究.
E-mail：luzy@tju.edu.cn

据，因此数据集的特征维数为 10 × 3 = 30 维。

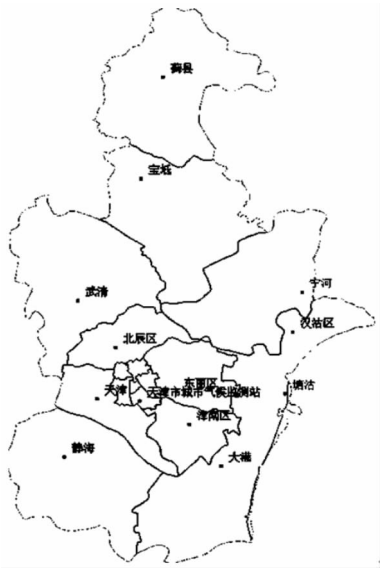

图 1　天津地面气象观测站点分布图

1.2　缺失值处理

在整理获取的天津地面站点气象物理场数据后，发现有许多数据出现缺失的情况，示例如表 1 所示。缺失值问题往往会影响模型的效果。如果对缺失数据不做处理，那么在训练某一个模型，或者整合数据集时，将会使系统丢失很多有效信息；训练后模型的不确定性将会更加的显著，很难抓住数据集的有效成分；如果在训练中某些数值缺失，会造成数据维数的不对齐，模型无法完成训练[7]。

表 1　缺失值数据样例

露点	水气压	2 min 平均风向/(°)	2 min 平均风速/(m/s)
10.2	33	53	1.2
7.5	缺失值	114	2.6
6.9	34	104	2.2
5.5	35	120	2.3

本文所使用的缺失值处理方法为均值插补的方法[8]，即利用研究数据特征的已观测数据的均值替代缺失值。这种方法能有效解决缺失值带来的问题，保证模型训练的准确性。

1.3　样本不均衡处理

由于本文所研究的强对流灾害数据为 588 个，而非强对流天气的数据为 107 395 个，两类样本数据出现了类别不均衡的问题。为了解决不均衡问题，常常使用过采样以及欠采样的方法[9]。本文对非强对流天气数据进行了欠采样，为了保证训练后模型的分类能力与真实样本的分布相似，使用了约 10∶1 的欠采样，欠采样后非强对流天气数据集为 5 822 个。

2　主成分分析

由于本文的建模样本集维数为 30 维，而样本数目少，在训练中，不可避免地使模型复杂度增大，造成模型的过拟合。为了防止过拟合问题的出现，需要对数据集进行降维处理[10]。

主成分分析（Principal Component Analysis，PCA）方法通过线性变换把给定的一组相关量转换成另一组不相关的变量[11]，这些新的变量按照方差依次递减的顺序排列，提取数据的主要特征分量[12]。

对于本文的物理场数据，首先进行数据标准化处理，将输入的各维数据减去各维度平均值；然后将输入数据与标准化后的逆矩阵进行矩阵相乘，通过矩阵的分解获得各个特征值，并进行从大到小的排序，获得各自的特征向量；最后筛选出特征值最大的前几维特征向量，选取的特征向量与占据特征值总和的主成分比重有关，通常会选择占据主成分比重 90% 以上的特征值对应的特征向量。

主成分分析方法对本文模型的评估效果如图 2 所示，可以看出，当数据集的维数降维为 8 维时，模型有最好的评估效果。在图 2 中，横坐标为主成分分析方法降维后的数据集维数（dimension），纵坐标为临界成功指数（CSI）。

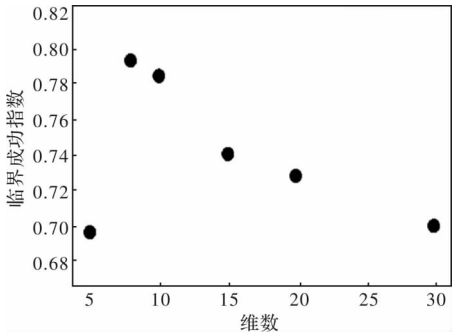

图 2　PCA 降维结果比较

3　强对流灾害预报模型构建

3.1　模型构建流程

图 3 说明了本文模型构建的流程，包括了数据预处理，主成分分析，样本集的划分，模型的训练、改进和比较。

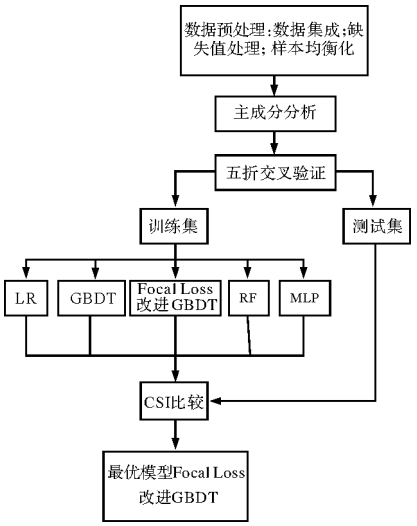

图 3　强对流灾害预报模型构建流程

(1)数据预处理：利用天津地面观测站点从2006－2018年的3－9月的采集的107 983个气象物理场数据，首先对这些气象数据进行缺失值处理；然后并将非强对流天气数据进行欠采样，采样后非强对流天气数据为5 822个，强对流灾害数据为588个。

(2)主成分分析：为了解决样本数据少而维度高的问题，本文经过模型的评估后，选择将原始的30维数据使用主成分分析的方式降维为8维。

(3)样本集的划分：为了获得适用性强、效果稳定的模型，本文使用五折交叉验证的方式划分训练集与测试集，即每次将互斥的20%的数据集作为测试集，剩下的80%的数据集作为训练集。模型评估结果为五个测试集评估结果的平均值。

(4)模型的训练、改进和比较：本文使用的模型为Focal Loss改进的GBDT模型。为了验证模型的效果，本文对比了常用分类模型。评价指标为气象常用指标：命中率，误警率(False Alarm Rate，FAR)和临界成功指数[13]。结果证明Focal Loss改进的GBDT模型在强对流灾害预报上有更好的效果。

3.2 GBDT模型

GBDT模型属于集成学习中的Boosting类，是以CART决策树为基础学习器的Gradient Boosting模型[15]。它利用了前一轮基础学习器的误差来对训练集的权值进行更新，并进行迭代[16]。本文基于GBDT模型对强对流灾害预报问题进行研究。

GBDT模型描述如下：

$$F(x;P) = F(x;\{\beta_n,\alpha_n\}_1^N) = \sum_{n=1}^{N}\beta_n h(x;\alpha_n)。 \quad (1)$$

式中：β为每一个基础学习器的相应权重，α为每一个基础学习器的参数，参数$\{\beta_n, \alpha_n\}$为M个数据(x_i, y_i)损失函数最小最优解P[16]。设损失函数L：

$$L(x,f(x)) = \sum_{i=1}^{M} l(y_i - f(x_i))。 \quad (2)$$

式中：l为每次迭代的基础学习器的损失函数。则

$$\{\beta_n,\alpha_n\}_1^N = \arg\min \sum_{i=1}^{M} L(y_i, \sum_{n=1}^{N}\beta_n h(x_i;\alpha_n))； \quad (3)$$

$$\beta_n,\alpha_n = \arg\min \sum_{i=1}^{M} L(y_i, F_{n-1}(x_i) + \beta h(x;\alpha))。 \quad (4)$$

对于每一个样本x_i，都可以得到一个梯度下降方向，即：

$$\vec{g}_n = -\vec{g}_n(x_i) = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right], F(x) = F_{n-1}(x)。 \quad (5)$$

优化公式(4)得到：

$$\alpha_n = \arg\min \sum_{i=1}^{M} (-g_n(x_i) - \beta h(x;\alpha))^2。 \quad (6)$$

进而求得β_n，

$$\beta_n = \arg\min \sum_{i=1}^{M} L(y_i, F_{n-1}(x_i) + \beta h(x;\alpha_n))。 \quad (7)$$

最终获取GBDT算法模型的迭代描述：

$$F_n(x) = F_{n-1}(x) + \beta_n h(x;\alpha_n)。 \quad (8)$$

本文对GBDT模型调节的特征有：树的最大迭代次数、树的最大深度和学习率。为了获得GBDT最优的参数，对不同的参数进行调整，参数调节的结果如表2、表3和表4所示。因此，本文最终选择最大迭代次数为10，树的最大深度为4，学习率为0.02。

表2 树的最大迭代次数与临界成功指数关系

树的最大迭代次数	临界成功指数
3	0.619
5	0.627
8	0.782
10	**0.794**
12	0.772
15	0.765
20	0.762
50	0.742
100	0.727

表3 树的最大深度与临界成功指数关系

树的最大深度	临界成功指数
3	0.624
4	**0.794**
5	0.782
6	0.765

表4 学习率与临界成功指数关系

学习率	临界成功指数
0.001	0.724
0.01	0.783
0.02	**0.794**
0.03	0.791
0.04	0.786
0.1	0.716
0.2	0.691

3.3 Focal Loss函数的模型改进

在分类问题中，常用的损失函数为交叉熵损失函数。本文所使用的数据集为欠采样后10∶1的非强对流天气与强对流灾害数据集，非强对流天气数据比例大，使用交叉熵损失函数会使训练后的模型向非强对流天气类别偏移。Focal Loss函数则可以有效解决这个问题。其公式为：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)。 \quad (9)$$

式中：α_t，γ代表超参数，p_t代表预测标签概率。

本文使用Focal Loss函数来替代公式(2)中的基础学习器的损失函数l。对于本文的非强对流样

本，Focal Loss 函数将其损失对应的权重变小，而将强对流灾害样本损失对应的权重变大，提升强对流天气样本的重要程度[17]。因此可以提高强对流天气预报的准确率，提升模型的性能。本文 α_t，γ 两个超参数选择经验值，γ 参数选择为 2，α 参数选择为 0.25。

Focal Loss 函数对 GBDT 模型改进的效果如表 5 所示。可见 Focal Loss 改进的 GBDT 模型在强对流灾害预报上有更好的效果。

表 5　Focal Loss 改进模型的强对流灾害评估

	POD	FAR	CSI
Focal Loss	0.843	0.171	0.794
交叉熵	0.802	0.206	0.727

4　强对流灾害预报模型评估

本文所使用的模型评估方式为五折交叉验证。每次将 1 282 个数据作为测试集，将 5 128 个数据作为训练集。将五次的测试集测试结果取平均值作为最终的模型评估结果[18]。

为了说明本文所建立的 Focal Loss 改进的 GBDT 模型对强对流灾害预报方法的有效性，本文对比了逻辑斯特回归(Logistic Regression，LR)、GBDT 模型、随机森林(Random Forest，RF)与多层感知机(Multi－Layer Perceptron，MLP)等常用分类模型。对比评估结果如表 6 所示。

表 6　不同模型对强对流灾害预报评估结果比较

模型	POD	FAR	CSI
LR	0.782	0.363	0.657
GBDT	0.826	0.217	0.769
RF	0.813	0.226	0.754
MLP	0.837	0.189	0782
Focal Loss 改进 GBDT	**0.843**	**0.171**	**0.794**

由此可见，本文所用 Focal Loss 改进的 GBDT 模型相比于逻辑斯特回归、GBDT 模型、随机森林与多层感知机在命中率及临界成功指数都更高，而误警率相对更低。说明 Focal Loss 改进的GBDT 模型在天津强对流灾害预报上有更优秀的性能。

5　结论

本文提出基于 Focal Loss 改进的 GBDT 模型对天津强对流灾害的预报方法，利用天津地面观测站的 2006－ 2018 年每年 3－9 月的逐小时观测的气象数据做训练。通过实验测试，结果表明：强对流灾害预报的命中率为 84.3%，临界成功指数为 79.4%，说明了模型对天津强对流灾害预报的有效性，具体结论如下：

(1)利用前 3 h 的地面观测站点的观测数据，实现了未来 1 h 强对流灾害天气的准确预报。

(2)在采集到的地面物理场数据集中，本文对于缺失值数据使用了均值插补的方法，保证了模型训练的准确性；对非强对流数据使用了欠采样的方法，为模型的有效分类能力提供基础。

(3)本文采用主成分分析方法将数据集进行降维，降维后的数据能有效地解决过拟合问题。

(4)本文所使用的 Focal Loss 改进的 GBDT 模型在对强对流灾害的预报上，命中率、误警率和临界成功指数等指标均相比较其他分类模型表现更好，说明了 Focal Loss 改进的 GBDT 模型在天津强对流灾害预报问题上的优秀性能。

针对本文工作还有待继续扩大数据集，获取更为有效的特征，完善训练模型，增强模型的学习能力和识别能力，为最终实现天津强对流灾害的准确预报奠定基础。

参考文献：

[1] 孙燕，吴海英，蒋义芳，等. 江淮地区致灾强对流天气流型配置及层结特征分析[J]. 灾害学，2019，34(3)：97－102.

[2] 王峙. 强对流天气对高空气象探测的影响及应对处理[J]. 南方农机，2019，50(19)：241.

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

[4] FerranFabro，Joan Montanya，Nicolau Pineda，et al. Analysis of energetic radiation associated with thunderstorms in the Ebro delta region in Spain[J]. Atmospheres，2016：9879－9891.

[5] 张秉祥，李国翠，刘黎平，等. 基于模糊逻辑的冰雹天气雷达识别算法[J]. 应用气象学报，2014，25(4)：415－426.

[6] YOU C H，LEE D I，KANG M Y，et al. Classification of rain types using drop size distributions and polarimetric radar：Case study of a 2014 flooding event in Korea[J]. Atmospheric Research，2016，181：211－219.

[7] 王昀. 新疆农作物生长期雹灾的时空分布及危害性评估[J]. 农业工程学报，2019，35(6)：149－157.

[8] 邓建新，单路宝，贺德强，等. 缺失数据的处理方法及其发展趋势[J]. 统计与决策，2019，23(5)：28－34.

[9] 易辉，宋晓峰，姜斌，等. 样本不均衡条件下基于自调整支持向量机的故障诊断[J]. 北京理工大学学报，2013，33(4)：394－398

[10] HAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over－sampling technique[J]. Journal of Artificial Intelligence Research，2002，16(1)：321－357.

[11] AMITA Nandal，HAMURABI GAMBOA Rosales，NINOSLAV Marina，Modified PCA Transformation with LWT for High-Resolution based Image Fusion[J]. Transactions of Electrical Engineering，2019，43 (1)：141－157.

[12] R Lazcano，Adaptation of an iterative PCA to a manycore architecture for hyperspectral image processing[J]. Journal of Signal Processing Systems，2019，91(7)：759－771.

(下转第 50 页)