Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(1), 90-97
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131011

基于自动机器学习的辽宁地区雷雨大风天气预测

宋红凯[1,2]，段　勇[1,2]，赵婷婷[1]

[1]沈阳工业大学信息科学与工程学院，辽宁　沈阳
[2]沈阳市先进计算与信创技术重点实验室，辽宁　沈阳

收稿日期：2023年12月11日；录用日期：2024年2月23日；发布日期：2024年2月29日

摘　要

针对辽宁地区雷雨大风天气的不确定性和时空差异性的特点，本文提出了一种基于自动机器学习的雷雨大风天气预测方法。首先由历史再分析数据集和地面实况数据集构建了需要的雷雨大风数据集；其次对经过预处理后的数据进行特征工程；然后使用基于多层堆栈集成、重复k-折交叉装袋策略的AutoGluon自动机器学习方法建立雷雨大风预测模型。最后，通过实验结果表明，使用AutoGluon方法构建的最佳模型在多项评估指标中，命中率为96.72%，漏报率为0.46%，误报率为1.62%。

关键词

灾害天气，自动机器学习，雷雨大风预测，AutoGluon

Thunderstorm Gale Weather Prediction in the Liaoning Area Based on Automatic Machine Learning

Hongkai Song[1,2], Yong Duan[1,2], Tingting Zhao[1]

[1]School of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning
[2]Shenyang Key Laboratory of Advanced Computing and Application Innovation, Shenyang Liaoning

Received: Dec. 11th, 2023; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

Abstract

A method for predicting thunderstorm and gale weather in Liaoning area using automatic ma-

文章引用: 宋红凯, 段勇, 赵婷婷. 基于自动机器学习的辽宁地区雷雨大风天气预测[J]. 人工智能与机器人研究, 2024, 13(1): 90-97. DOI: 10.12677/airr.2024.131011

chine learning to address uncertainties and spatial-temporal differences is proposed in this paper. To begin with, we construct the necessary dataset for thunderstorms and gales using both the historical reanalysis data set and the ground live data set. Then we perform feature engineering on the preprocessed data. After that, we establish the thunderstorm gale prediction model using the AutoGluon automatic machine learning method, which is based on multi-layer stack integration and a repeated k-fold cross-bagging strategy. Finally, the experimental results show that the best model constructed by the AutoGluon method has a hit rate of 96.72%, a false negative rate of 0.46%, and a false positive rate of 1.62% in several evaluation indicators.

Keywords

Disaster Weather, Automatic Machine Learning, Thunderstorm Gale Prediction, AutoGluon

1. 引言

雷雨大风是指平均风力值大于 17.2 m/s，且降雨量大于 20 mm 的一种强对流天气[1]，具有突发性强、发生范围小、持续时间短、破坏力强等特点。由于雷雨大风天气的发生极为局促和突然，常常给农业、林业、畜牧业造成巨大的经济损失，严重影响人民群众的正常生活。

目前，国内外已有一些研究工作使用机器学习方法针对雷雨大风进行短临预报。王兴等[2]基于深度神经网络天气识别算法，以表征回波移动路径的光流图像和雷达回波图像作为输入，对雷雨大风天气进行识别，降低了误报率。Jiang Y 等[3]提出了一种基于多源卷积神经网络的预测方法，提升了预测准确率。路志英等[4]提出基于物理量参数和深度学习模型 DBNs 的短时强降水天气识别模型，对于短时强降水的命中率、误警率和临界成功指数，都有着较好的表现。

虽然机器学习方法相较于传统计量模型，能够处理高维度、非线性和复杂的数据集，但由于气象问题具有不确定性和时空差异性，单一的机器学习方法难以适应错综复杂的变化。基于此，本文提出一种基于 AutoGluon 自动机器学习的雷雨大风预测方法，该方法可以通过一定的策略建立方法集成，不仅能够得到拟合效果更好、精度更高的预测模型，而且无需反复调整参数[5]。本文将气象天气中常见的降雨和大风作为研究对象，基于真实的气象数据集，使用 AutoGluon 方法，完成辽宁地区未来 3 小时内的多个站点的雷雨大风天气预测，及时地进行预报预警。

2. 气象数据集的构建

本文以辽宁地区的降雨和大风作为研究对象，主要使用以下两类资料作为基础数据：一类是来自美国气象环境预报中心 NCEP (National Centers for Environmental Prediction)的历史再分析资料中的数据，其中特征属性包含了与研究对象降雨和大风相关的属性，有位势高度、纬向风速等预报数据；另一类是来自辽宁省地面观测站获取的 ECMWF 数值预报资料中的历史实况数据，主要包括降水量和大风值等数据资料。

研究历史再分析资料中与雷雨大风有关的特征要素，选取 46 个相关的属性，接着对其进行解析和预处理等操作；同样，地面实况观测数据资料也要经过数据处理，得到有效的降雨量和大风值数据。这样就生成了两类数据集，因为雷雨大风天气大多集中在夏季，故两类数据集时间均选择 2015~2018 年中 5~9 月的数据进行导出构建。除此之外，因两个数据集的时区不同，还需要经过时间的转化。最后，通过经

纬度和时间这两个匹配条件(如图 1 所示)，将两类数据集进行匹配合并。

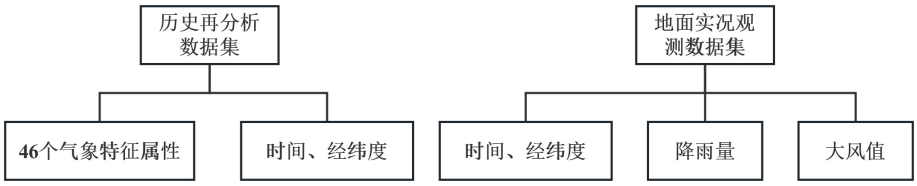

Figure 1. Construction of two types of datasets
图 1. 两类数据集的构建

本文将当前时刻地面观测数据与 3 小时前的 NECP 历史再分析数据进行匹配，建立了实验所需要的完整的训练样本数据集，用于预测未来 3 小时辽宁地区的雷雨大风情况。其中，每一条数据包含的 50 项属性从左至右分别为站点编号(StationNum)、观测时间(ObservTimes)、温度(MaxTemp)、风向(WindDirect)、风速(WindVelocity)、湿度(RelHumidity)、气压(StationPress)、降雨(Precipitation)、大风(MaxWindV)以及其他历史再分析数据属性特征，如 HGT1000、RH500、TMP200、UGRD1000、VGRD500、VVEL850 等。

3. 雷雨大风预测模型自动化建立

3.1. AutoGluon 方法概述

AutoGluon 是一种自动机器学习方法，能自动实现数据特征选择并进行模型训练，依赖于融合多个不需要超参数搜索的模型，支持图像、表格、文本等多种格式数据的处理，适用于文本分类、图像分类、对象检测等多种任务类型[6]。本文使用 AutoGluon 中的表格预测功能(AutoGluon-Tabular)实现雷雨大风的预测，完成自动化模型处理、训练。

3.2. AutoGluon 模型准备

为了解决前馈神经网络或卷积神经网络架构在处理表格数据方面表现不佳的问题，AutoGluon 使用如图 2 所示的新型神经网络架构[7]。

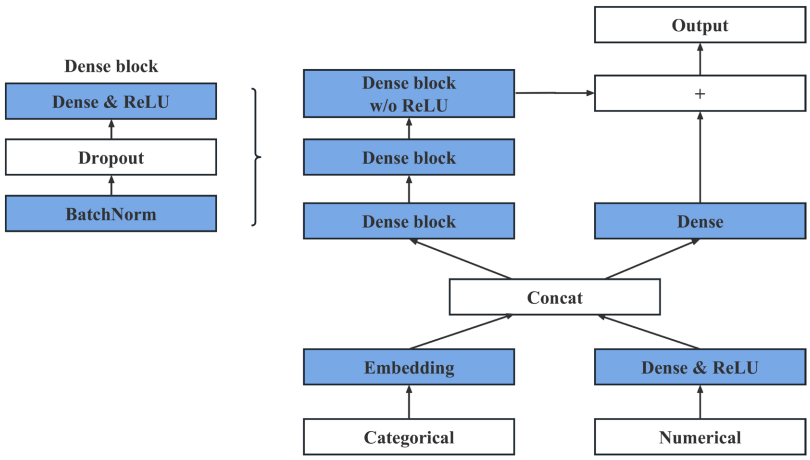

Figure 2. New network architecture of AutoGluon
图 2. AutoGluon 的新型网络架构

在新型网络架构中，网络为每个类别特征都引入一个嵌入层，其中嵌入维数根据特征中唯一类别的数量按比例选择。类别特征的嵌入与数值特征连接成一个矢量，该矢量既被馈送到 3 层前馈网络，也通

过线性的跳跃连接(skip-connection)直接连接到输出预测。神经网络中的每个密集块包括几个卷积层、池化层、ReLU 激活函数层、批归一化(batch normalization)层和随机失活正则化(dropout)。

此外，本文研究的自动机器学习预测模型中，将 LightGBM 提升树算法、CatBoost、随机森林算法(Random Forest)、极端随机树算法(Extremely Randomized Trees)、k 最近邻算法(KNN)、XGBoost 以及加权集成模型(WeightedEnsemble_L2)等多种模型作为基础模型。其中 XGBoost 是对原始版本的 GBDT 算法的改进，而 LightGBM 和 CatBoost 则是在 XGBoost 基础上做了进一步的优化，在精度和速度上都有各自的优点。

3.3. AutoGluon 模型训练

算法采用两大运算策略：多层堆栈(stack)集成、重复 k-折交叉装袋(bagging)，以此来进一步提高预测准确度和减少过拟合。

(1) 模型构建。堆栈集成，将堆栈器模型输出的预测作为输入，提供给其他更高层堆栈器模型。将其所有基础层模型类型重新用作堆栈器，堆栈模型的最后一层使用集成选择以加权的方式，聚合堆栈模型的预测。

(2) 模型训练。将数据随机划分为 k 个不相交的块，随机选取一块为验证集，其余部分为测试集。然后使用堆栈模型，训练和测试之前准备的基础模型，在堆栈的所有层对所有模型进行 k-折交叉装袋。

(3) 通过重复上述过程，执行多次 k-折交叉装袋，得到多层堆栈模型。

(4) 模型测试，确定最佳模型。

本文采取的模型结构如图 3 所示，图中显示了 AutoGluon 的多层堆栈策略，这里使用两个堆栈层和 n 种类型的基础学习器。

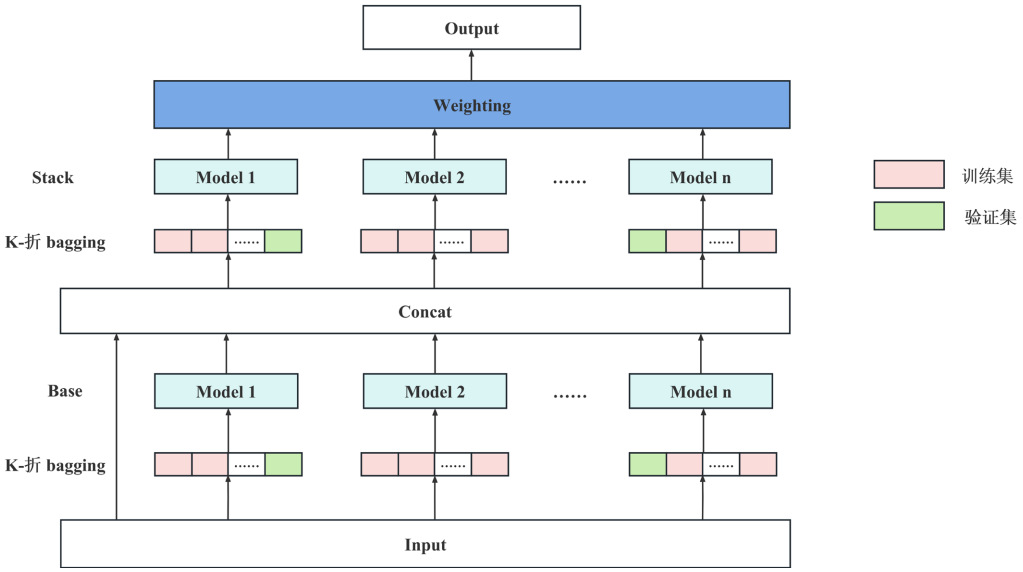

Figure 3. AutoGluon's multi-layer stack strategy
图 3. AutoGluon 的多层堆栈策略

图 3 中，第一层有多个基础模型，其输出被级联，然后被馈送到下一层，而下一层本身由多个堆栈模型组成。然后，这些堆栈器充当附加层的基础模型。

3.4. 评价指标

本文对基于自动机器学习的雷雨大风预测方法的评价指标包括 Pearson 相关系数(PCCs)、均方根误差

(RMSE)、平均绝对误差(MAE)和 R2_score 等统计误差评估雷雨大风预测情况，具体公式见式(1)~式(4)：

$$PCCs=\frac{\sum_{i=1}^{m}\left(y\_true_i-\overline{y\_true}\right)\left(y\_pre_i-\overline{y\_pre}\right)}{\sqrt{\sum_{i=1}^{m}\left(y\_true_i-\overline{y\_true}\right)^2}\sqrt{\sum_{i=1}^{m}\left(y\_pre_i-\overline{y\_pre}\right)^2}} \tag{1}$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y\_true_i-y\_pre_i\right)^2} \tag{2}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|y\_true_i-y\_pre_i\right| \tag{3}$$

$$R2\_score=1-\frac{\sqrt{\sum_{i=1}^{m}\left(y\_pre_i-y\_true_i\right)^2}}{\sqrt{\sum_{i=1}^{m}\left(y\_true_i-\overline{y\_true}\right)^2}} \tag{4}$$

其中，y 表示需要预测的降水值或大风值，在式(1)~式(4)中，m 表示雷雨大风数据集总条数，y_true 表示真实的 y 值，y_pre 表示模型预测的 y 值。PCCs 取值范围[−1, 1]，越接近 1，效果越好。其余指标范围参见文献[8]。

从气象局提供的参考资料可知，雷雨大风是指平均风力值大于 17.2 m/s 且降雨量大于 20 mm 的一种强对流天气。具体判断有无雷雨大风现象的方法如图 4。

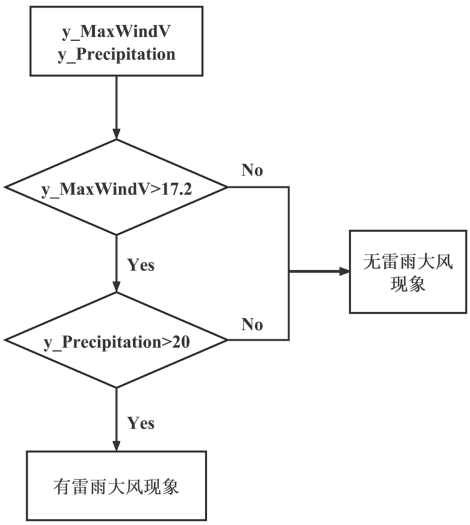

Figure 4. Judging whether there is a thunderstorm or strong wind phenomenon
图 4. 判断是否有雷雨大风现象

TS 评分(True Skill Statistic)是一种用于评估气象模型预报性能的方法，其取值范围在 0 到 1 之间，TS 评分越高，表示模型预报的准确性越高[9]。此外，权威的检验预测效果的统计量还包括：命中率(POD)、准确率(ACC)、漏报率(FNR)、误报率(FAR)。其中，本文研究使用的各指标代表含义如下：

(1) TP 为实际有雷雨大风且预测有雷雨大风；

(2) FP 为实际有雷雨大风而预测无雷雨大风；

(3) FN 为实际无雷雨大风而预测有雷雨大风；

(4) TN 为实际无雷雨大风且预测也无雷雨大风。

各指标的公式如式(5)~式(9)：

$$TS = TP/(TP + FP + FN) \tag{5}$$

$$POD = TP/(TP + FP) \tag{6}$$

$$FAR = FN/(TP + FN + FP + TN) \tag{7}$$

$$FNR = FP/(TP + FN + FP + TN) \tag{8}$$

$$ACC = (TP + TN)/(TP + FN + FP + TN) \tag{9}$$

4. 实验结果与分析

本研究针对于辽宁地区的雷雨大风天气问题，建立了雷雨大风数据集，使用基于多层堆栈集成、重复 k-折交叉装袋策略的 AutoGluon 自动机器学习方法建立雷雨大风预测模型。

AutoGluon 以特定选择的顺序(fit_order)训练各个模型。在训练过程中，AutoGluon 根据每个模型的性能和复杂度等因素进行评估和比较，以便选择出最优的模型。其主要过程首先是训练性能可靠的模型，如 RandomForest 模型，然后逐步训练计算成本更高但可靠性较低的模型，如 KNN 模型，这样可以在特定成本或时间预算下获得最佳精度。同时，在本实验中，堆栈层数没有超过 3 层，保证了训练效率。

为了验证 AutoGluon 方法在预测雷雨大风方面的效果，将 AutoGluon 方法预测的结果与决策树(Decision Tree)、线性模型(Linear)、K-近邻(KNN)、随机森林(RandomForest)、LGBM、XGBoost、支持向量机(SVM)这些常见的机器学习模型进行对比实验，具体结果见表 1。

Table 1. Comparison of statistical error indicators between autogluon and other machine learning models
表 1. AutoGluon 与其他机器学习模型统计误差指标对比

模型	MaxWindV (大风)				Precipitation (降水)			
	RMSE	MAE	R2	PCCs	RMSE	MAE	R2	PCCs
DecisionTree	5.18	1.43	0.98	0.95	9.21	2.05	0.90	0.93
Linear	90.29	55.01	0.18	0.42	29.72	42.61	0.41	0.65
KNN	57.88	28.67	0.66	0.76	13.75	29.38	0.72	0.72
RandomForest	17.12	7.69	0.97	0.88	12.70	17.26	0.85	0.89
LGBM	18.29	6.79	0.96	0.96	11.68	14.59	0.89	0.85
XGBoost	4.18	1.45	0.94	0.95	10.22	12.05	0.89	0.90
SVM	107.42	48.90	0.15	0.42	21.46	59.55	0.14	0.34
AutoGluon	**4.19**	**0.93**	**0.99**	**0.99**	**6.32**	**1.30**	**0.98**	**0.99**

从表 1 可知，无论是大风值预测还是降水值预测，RMSE 和 MAE 指标结果数值均为最低，R2 和 PCCs 评价指标数值均最接近于 1。表明 AutoGluon 算法构建的雷雨大风预测模型的统计误差指标明显好于其他模型的效果。

接着，使用气象学中常使用的 TS 评分评价模型。其中，部分实验样本的分析结果如表 2 所示。

Table 2. Analysis results of part of the experimental samples
表 2. 部分实验样本分析结果

实际大风值	预测大风值	实际降水量	预测降水量	实际	预测	结果
0	18.34	24	30.89	无	有	空报
23	16.17	78	10.74	有	无	漏报
332	340.33	226	223.33	有	有	命中

续表

10	12.5	0	0	无	无	不计数
139	63.28	26	20.84	有	无	漏报
255	269.54	130	100.22	有	有	命中

在多项评估指标中，除了本文的 AutoGluon 方法以外，预测效果最好的机器学习模型是决策树模型，其中 R2_score 和 PCCs 已经很接近 AutoGluon 方法，其次是 XGBoost 方法，该方法在大风预测方面效果良好，但其在降水方面的预测效果不佳。其余方法(如 SVM、KNN)均表现效果一般。使用 TS 评分、POD、ACC、FNR、FAR 等描述准确率的指标，将这三种方法作对比实验，具体结果如表 3。

Table 3. Comparison of accuracy metrics between AutoGluon and other machine learning models
表 3. AutoGluon 与其他机器学习模型准确率指标对比

	TS	POD	ACC	FNR	FAR
XGBoost	80.44%	83.45%	89.12%	2.01%	8.87%
Decision Tree	84.98%	91.46%	91.82%	3.85%	2.52%
AutoGluon	**95.97%**	**96.72%**	**97.99%**	**0.46%**	**1.62%**

从上表可知，使用 AutoGluon 方法构建雷雨大风预测模型具有很好的效果。在 TS 评分方面，该方法比 Decision Tree 和 XGBoost 分别高出 10.99%和 15.53%。同时，AutoGluon 方法构建雷雨大风预测模型从各种指标来看，均属于最佳指标。另外，通过多次的 AutoGluon 方法构建模型，并进行同类型的指标评估，评估效果相差不到 1%，说明使用 AutoGluon 自动机器学习方法构建雷雨大风预测模型，具有很强的鲁棒性。

5. 结论

灾害性天气中的雷雨大风天气，持续困扰和影响人们的正常生活并对社会造成不同程度的危害，但对其进行准确预测亦面临巨大挑战。本文旨在深入探讨提高对此类天气的预测精度。基于真实的气象数据集，研究使用自动机器学习方法构建雷雨大风预测模型，并通过实验验证了研究工作的有效性。实验结果表明，基于 AutoGluon 自动机器学习的预测方法对辽宁地区未来 3 个小时的雷雨大风预测具有较好的效果，高于部分主要机器学习模型的预测精度。

基金项目

辽宁省高等学校优秀科技人才支持计划(LR15045)；辽宁省教育厅科学研究经费面上项目(LJKZ0139)；辽宁省气象台科技项目(201903256, 201803276)。

参考文献

[1] 沈平, 余小嘉, 郑晓志, 等. 广东省雷雨大风预警信号的智能监控[J]. 广东气象, 2022, 44(3): 85-88.

[2] 王兴, 吕晶晶, 王璐瑶, 等. 基于深度神经网络的强对流天气识别算法[J]. 科学技术与工程, 2021, 21(7): 2737-2746.

[3] Jiang, Y., Yao, J. and Qian, Z. (2019) A Method of Forecasting Thunderstorms and Gale Weather Based on Multi-source Convolution Neural Network. *IEEE Access*, **7**, 107695-107698. https://doi.org/10.1109/ACCESS.2019.2932027

[4] 路志英, 任一墨, 孙晓磊, 等. 基于深度学习的短时强降水天气识别[J]. 天津大学学报(自然科学与工程技术版), 2018, 51(2): 111-119.

[5] Wen, Z., Yang, G. and Cai, Q. (2021) An Improved Calibration Method for the IMU Biases Utilizing KF-Based Ada-Grad Algorithm. *Sensors*, **21**, 5055-5075. https://doi.org/10.3390/s21155055

[6] Qi, W., Xu, C. and Xu, X. (2021) AutoGluon: A Revolutionary Framework for Landslide Hazard Analysis. *Natural Hazards Research*, **1**, 103-108. https://doi.org/10.1016/j.nhres.2021.07.002

[7] Erickson, N., Mueller, J., Shirkov, A., *et al*. (2020) Autogluon-Tabular: Robust and Accurate Automl for Structured Data. arXiv: 2003.06505.

[8] Xu, B. (2021) Assessment of a Gauge-Radar-Satellite Merged Hourly Precipitation Product for Accurately Monitoring the Characteristics of the Super-Strong Meiyu Precipitation over the Yangtze River Basin in 2020. *Remote Sensing*, **13**, 3850-3863. https://doi.org/10.3390/rs13193850

[9] 郑超昊, 尹志伟, 曾钢锋, 等. 基于时空深度学习模型的数值降水预报后处理[J]. 浙江大学学报(工学版), 2023, 57(9): 1756-1765.