Software Engineering and Applications 软件工程与应用, 2021, 10(1), 35-43
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.101006

Hans汉斯

基于机器学习的雷达回波与降雨分析

黄建伟，段　勇，杨　堃，于　霞

沈阳工业大学信息科学与工程学院，辽宁　沈阳
Email: duanyong0607@163.com

收稿日期：2021年1月14日；录用日期：2021年2月13日；发布日期：2021年2月25日

摘　要

多普勒天气雷达产生的雷达回波数据是降雨分析及预测的重要依据，针对如何有效利用雷达回波进行降雨等级分析问题，本文研究了一种基于XGBoost集成学习算法的雷达回波与降雨关系分析模型。本文使用辽宁省气象台提供的历年雷达和降雨气象观测数据，经过数据解码、清洗、匹配后，使用经网格搜索算法优化后的XGBoost方法训练，建立多层雷达回波数据与降雨等级的分类关系。最后通过实验结果表明，基于XGBoost方法得到的结果更接近实际，能够较好地反映云团雷达回波和降雨的关系。

关键词

XGBoost，集成学习，降雨分析，机器学习，雷达回波

Relationship Analysis of Radar Echo and Rainfall Based on Machine Learning

Jianwei Huang, Yong Duan, Kun Yang, Xia Yu

School of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning
Email: duanyong0607@163.com

Received: Jan. 14th, 2021; accepted: Feb. 13th, 2021; published: Feb. 25th, 2021

Abstract

Radar echo data generated by Doppler weather radar is an important basis for rainfall analysis and prediction. Aiming at the problem of how to make effective use of radar echo for rainfall grade analysis, this paper studies an analysis model of the relationship between radar echo and rainfall based on XGBoost ensemble learning algorithm. In this paper, we use the radar and rainfall meteorological observation data provided by Liaoning Meteorological Station over the years. After

文章引用: 黄建伟, 段勇, 杨堃, 于霞. 基于机器学习的雷达回波与降雨分析[J]. 软件工程与应用, 2021, 10(1): 35-43.
DOI: 10.12677/sea.2021.101006

data decoding, cleaning and matching, we use XGBoost method optimized by grid search algorithm to establish the classification relationship between multi-layer radar echo data and rainfall level. Finally, the experimental results show that the results based on XGBoost method are closer to reality and can better reflect the relationship between cloud radar echo and rainfall.

Keywords

XGBoost, Ensemble Learning, Rainfall Analysis, Radar Echo

1. 引言

降水作为全球水循环三个最重要组成部分之一，其对人类生活影响程度不言而喻。多普勒天气雷达是进行气象观测的重要工具，其可产生三个最主要的数据：雷达反射率因子(雷达回波)、平均径向速度和谱宽。利用雷达回波可确定降雨(雪)带，但是如何准确分析二者关系，例如确定某次降雨的等级或者大小，是一个具有挑战性的工作。

目前被广泛研究用于雷达回波与降雨量关系分析的方法主要有两种：Z-I 关系法和雷达外推法。Jones 等发现了反射率因子(雷达回波强度单位)与降水量间的 Z-I 关系。多普勒效应被发现后，雷达数据与雨量站数据开始结合使用，简单的 Z-I 关系也逐渐开始被优化[1]。Suzana R 等人为了获得更准确的区域 Z-I 关系，使用了最优化方法[2]；而汪瑛则提出了动态分级 Z-I 关系估计法[3]。Brandes 引入 Bares 客观分析法来分析雷达回波与降雨关系[4]。目前，与 Z-I 关系结合使用最主要的雷达外推方法是光流法[5]。使用光流法进行雷达外推后，还要使用 Z-I 关系进行反演，二者误差叠加，整体分析精度会降低。

近年来，机器学习相关技术为很多领域解决问题带来了新的途径，使得许多原来棘手的非线性问题迎刃而解。在气象领域，Lee 等人根据当地雷达信息和降水数据通过径向基神经网络，预测了 24 小时后的降雨量[6]；Robert 等在降雨数值分析的基础上，引入了人工神经网络进行局部地区降雨分析[7]；Luk 等人使用多层前反馈神经网络、偏循环神经网络和时间延迟神经网络 3 种方法对帕拉玛塔河流域上游暴雨量进行了预测，表明选用这 3 种模型用来进行分析预测是可行的[8]；Chau 等将极限学习机与马尔科夫蒙特卡洛方法、Copula 和 Bat 算法结合，实现了巴基斯坦 3 个农业带的降雨分析，取得了很好的结果[9]；韩婷婷、时玮域[10]等人使用 SVM 方法对大雾天气进行预测，在大雾预测方面实现了新的突破。

基于此，本文采用 XGBoost 集成学习算法，采用多层雷达数据进行降雨分类问题分析。本文使用辽宁省气象局提供的辽宁省内的雷达回波数据及多个降雨观测站点测得的降雨数据，经数据预处理与匹配后，形成雷达回波与降雨数据集。随后，使用优化后 XGBoost 模型对雷达回波与降雨量之间的关系进行分析。最后，搭建了雷达回波与降雨分析系统，并投入日常气象工作中使用。文中第 2 节描述了数据组织方式以及预处理办法；阐述了如何建立优化模型、搭建降雨分析系统的过程以及系统实现的主要功能。第 3 节则对实验结果进行分析，研究模型内部分析数据规律，以及使用该方法进行分析预测的可行性。论文第 4 节则对全文进行了总结。

2. 多层雷达数据与降雨关系分析

2.1. 雷达和降雨数据集的处理

本模板本文使用历年辽宁省地区的多层雷达回波数据和降雨数据来建立机器学习训练样本数据集，原始数据由辽宁省气象局提供。为了有效利用多层雷达和降雨数据提供的有用信息，构建可靠的雷达回波和降雨关系分析模型，需要将原始的雷达数据和降雨数据进行数据预处理。数据预处理包括去除脏数据、站点观测值提取、强化特征，然后按照时间相近原则匹配雷达观测数据和降雨测量数据。

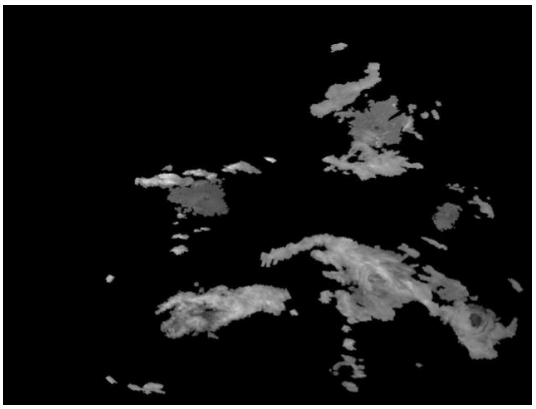

Figure 1. Visualize radar map
图 1. 可视化雷达图

雷达数据每间隔 6 分钟产生一次，每次产生 21 层雷达数据，它们对应间隔 0.5 km 不同高度的雷达回波，其可视化后如图 1 所示。据气象领域经验，通常最上和最下 2 层雷达所得数据存在较强的杂波干扰并且降雨量关系不明确。因此，本文选取 3~19 共 17 层雷达回波值来构成数据集的特征序列，其组织形式如图 2 所示。

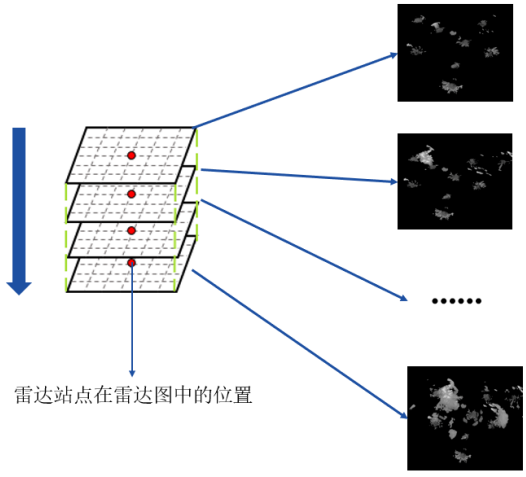

Figure 2. Multilayer radar data organization form
图 2. 多层雷达数据组织形式

首先依据规格说明对原始雷达数据文件进行解析，获得雷达图分辨率、中心经纬度、像素间距对应实际距离等头部信息。然后，将二进制数据流文件转成雷达图对应的数值矩阵。最后，根据气象观测站

点经纬度，在多层雷达数据中找到其位置，并求得该位置八邻域雷达回波数据均值，依据式(1)转换为对应的雷达回波值。其中 dBZ 是雷达回波的单位，Grey 表示灰度值。八邻域雷达回波数据均值是指依据经纬度在雷达图中找到雷达站点对应回波值后，同时取其周围八个回波值，加自身共 9 个数据均值作为站点对应的回波值，其目的是减少雷达回波数据的稳定性，从而更好地反映降雨的变化关系。

$$dBZ = \frac{(Grey - 66)}{2} \tag{1}$$

降雨数据间隔产生，一小时内累计，下一小时重新计算，表示雨量站监测的一小时内降雨量。为了得到固定间隔内准确的降雨值，在数据库中提取有雨数据，然后添加精确位置和时间信息。然后进行数据清洗，删除空数据，得到数据集中的标签列。

进行雷达数据和降雨数据匹配时，由于采样频率不同，进行匹配要求时间接近、站点编号完全一致。匹配完成后，根据气象分析预测需求，将匹配后的数据分为 5 个等级。

2.2. 基于 XGBoost 的雷达回波与降雨分析系统

2.2.1. XGBoost 集成学习算法

本文采用集成学习中的 XGBoost 方法来分析雷达回波与降雨量之间的关系。基于集成学习的雷达回波与降雨分析方法通过集成多个弱分析器的结果来提升分析的准确率[11]。本文所述多层雷达回波数据与降雨量之间存在着复杂的函数映射关系，普通弱学习器很难取得很好的效果。把多个弱学习器组合在一起，形成一个强学习器，这是集成学习的核心思想。

XGBoost 是 GBDT (Gradient Boosting Decision Tree)的改进版本。为了避免过拟合问题，它将树模型的复杂程度纳入到正则项中；损失函数使用泰勒展开式展开，使用了一阶导数与二阶导数。经历t次迭代后样本i的分析结果可以被表示为：前$t-1$棵棵决策树分析结果与使用迭代函数f的第t次迭代结果之和。

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

XGBoost 的目标函数可被表示为式(3)，其中，第一项为损函数差，例如常见的 logistic 或 MSE，其代表了模型的偏差；为了尽可能减小模型方差，控制树的复杂程度，在一定程度上防止发生过拟合现象[12]，在目标函数中加入了第二项正则项Ω，它的更详细表示见式(4)。式(4)中，T 表示每棵树叶子节点的数量，其值越小表示模型越简单，ω 表示这些树叶子节点权重组成的集合，ω 中权重不宜过高，γ、λ 是参数，可依据实际应用自行设置。

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{4}$$

为建立 XGBoost 分析模型，需将前文所述预处理后的数据集划分成 2 部分，一部分用来训练 XGBoost 模型，另一部分用来验证模型效果。XGBoost 模型接受雷达数据的 3~19 层，共 17 层雷达回波作为输入，经内部多棵决策树分析后，输出多层雷达回波数据对应的降雨等级(0~4 级，共 5 个等级)。

2.2.2. 基于 XGBoost 的雷达回波与降雨分析模型优化

虽然 XGBoost 在防止过拟合、泛化能力等方面表现不俗，但只有找到一组最优超参数，该模型表现才能最优。目前，网格搜索、随机搜索、遗传算法以及粒子群算法等常被用来进行参数优化。上述方法除网格搜索方法外，在进行多个超参数优化时，都存在参数间相互影响的问题，往往不能保证结果最优。

网格搜索(Grid Search)是一种穷举搜索调参手段，这种优化方法将尝试范围内的每一种超参数组合，以此确定最优的超参数组合。显然，网格搜索方法可以避免选取局部最优解的问题。为了尽量提高分析模型的准确率，本文选取网格搜索作为超参数优化方法，并且采用K折交叉折验方式来验证选取的超参数组合是否合理。

K 折交叉验证(K-fold cross validation)是一种检验分类器性能的统计学方法[13]。其基本思想是将训练集分为 K 份(均等划分，$K \geq 2$)，其中 $K-1$ 份作为训练集，带入模型训练，剩余 1 份子集作为验证集[14]。上述过程重复 K 次，验证集每次选取第 1，2，3，……，K 份子集，记录每次验证集得分 $E_i(1 < i \leq K)$，并求得均值作为得分[15]。网格搜索与交叉验证关系见图 3。

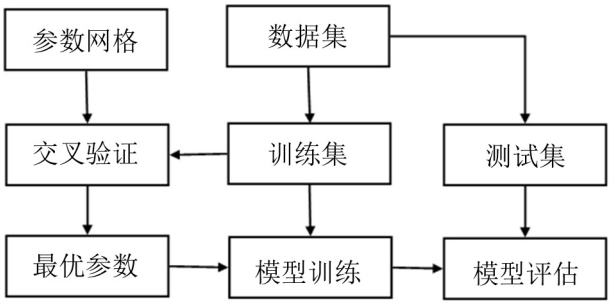

Figure 3. Grid Search and Cross Validation
图 3. 网格搜索与交叉折验确定参数

2.2.3 降水分析系统的设计与实现

为将前文设计实现的分析算法投入日常气象工作中，本文设计并实现了一个功能齐全、用户友好的分析系统。系统功能可以被分为四个部分，分别是：雷达数据自动获取处理模块、雷达数据分析模块、降雨数据分析模块以及结果分析模块，相关模块功能框图见图 4。

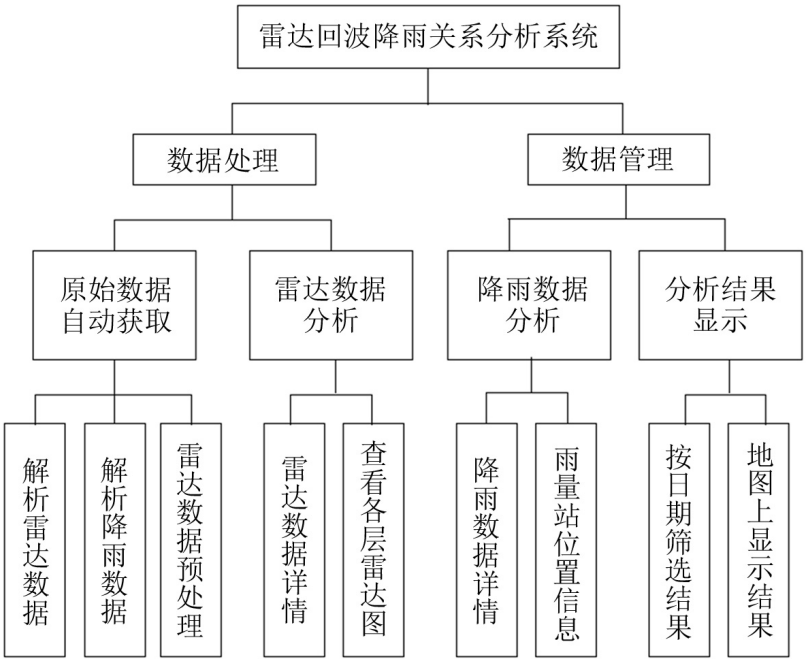

Figure 4. System functional block diagram
图 4. 系统功能框图

雷达数据自动获取处理模块主要完成两项工作：一是在后台自动完成对雷达数据的预处理工作，处理过程如前文所述，处理过后得到数据集；二是对雷达数据进行转码、解码、可视化等一系列操作，得到每间隔 6 分钟的 21 层雷达图，随后使用雷达数据分析模块注册的回调函数告知该模块，雷达数据已准备完成，可供调用。

雷达数据分析模块主要完成雷达数据分析和雷达可视化工作，其界面如图 5 所示。在该模块中可以选择查看指定时间的雷达可视化图片以及雷达数据头部信息，包括：文件说明、纵深数量、网格开始经纬度及中心经纬度、有效站点等信息。

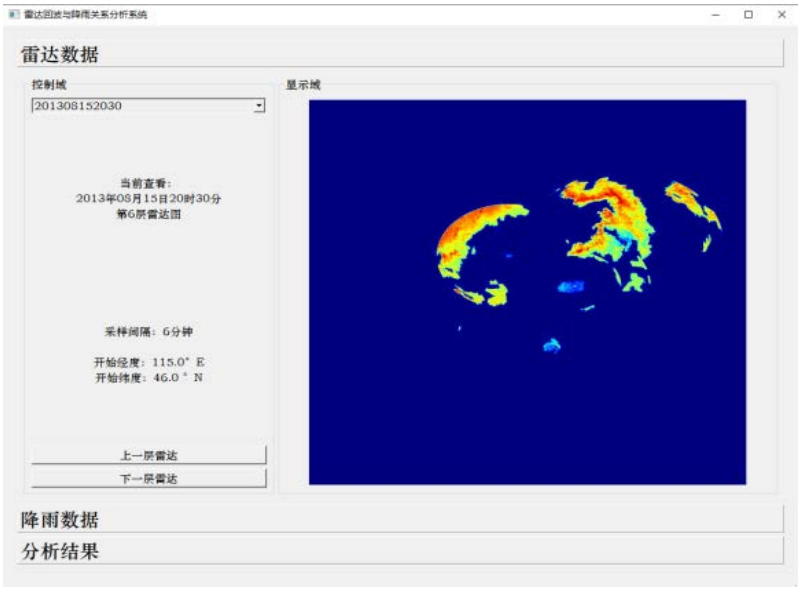

Figure 5. Radar data analysis module
图 5. 雷达数据分析模块

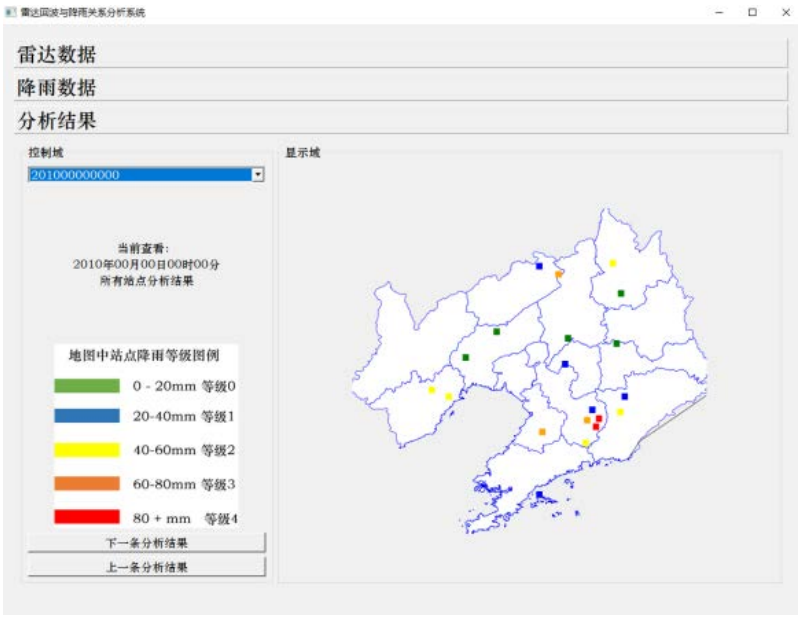

Figure 6. Analysis results display
图 6. 分析结果显示模块

最后一个模块显示 XGBoost 分析模型输出的分析数据。左侧选择当前查看的时刻后，右侧会显示这一时刻 XGBoost 模型的分析结果。分析结果包含一张由 basemap 绘制的辽宁省地图，地图内包含按照经纬度位置信息绘制的降雨雷达监测站点。降雨量从大到小被分为 5 个等级，分别由绿色、蓝色、黄色、橙色以及红色标识，如图 6 所示。

3. 实验结果及分析

文中所述降雨分析属分类过程，其评价指标有准确率(Accuracy，式(5))、精确率(precision，式(6))、召回率(recall，式(7))，F1 得分(F1-Score 式(8)) [16]。上述式中 TP、TN、FP、FN 等含义见表 1，正类使用“1”表示，负类使用“0”表示。

Table 1. Abbreviations used in evaluation method
表 1. 评价方法中涉及缩写含义

缩写	名称	真实类别	预测类别
TP	True Positive	1	1
FP	False Positive	0	1
TN	True Negative	0	0
FN	False Negative	1	0

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 score = \frac{precision \times recall \times 2}{precision + recall} \tag{8}$$

准确率主要评价模型正确预测分析样本能力，精确率即查准率，被用来评价模型的精确性，指类别真正为“1”的数据占所有预测为“1”的数据比例；召回率又称查全率，被用来评价模型对于类别为“1”的数据识别能力；而 F1 得分是折中进行模型评估。

在使用本文所述 XGBoost 方法建立基于多层雷达回波数据的降雨强度等级分析模型时发现，在一定范围内当训练迭代次数(即模型中树的数量)增加时，模型分析效果有一定提升，但超过某个值时，即使训练集误差减小，测试集误差率仍不变甚至轻微上升，这说明发生了过拟合现象，详见图 7。在图 7 中，迭代次数小于 600 时，测试集与训练集误差均在减小，说明此时为欠拟合状态；而迭代次数大于 1000 以后，训练集误差近似一条直线，表明此时发生了过拟合现象，这证明理想的模型弱分析其数量为 700~900 个。

各分类等级指标详细统计数据见表 2，不同降雨等级对应不同的精确度、召回率和 F1-SCORE 得分指标数据，表明对应的多层雷达回波数据的特征显著程度不同。

为了进一步验证研究工作的有效性，使用本文的 XGBoost 方法和当前常用的机器学习算法进行对比实验。比较的方法为 Adaboosting，GBDT，Bagging 三种集成学习算法。实验结果如图 8 所示。本文的 XGBoost 方法建立的基于多层雷达回波数据的降雨强度等级模型准确率为 75.66%，高于其他对比集成学习模型，包括 Adaboosting (72.59%)、GBDT (70.35%)以及 Bagging (71.28%)。由此可见本文使用的 XGboost 模型较其他机器学习模型取得了较高精度，更适合进行气象分析，特别是降雨相关的分析预测。

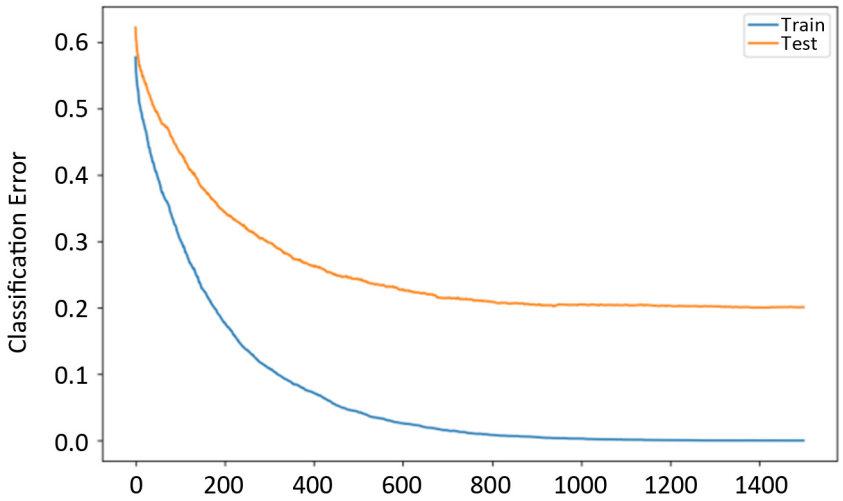

Figure 7. Relationship of iteration and error rate system result of standard experiment
图 7. 迭代次数与错误率关系

Table 2. Each classification level index data
表 2. 各分类等级指标数据

类别	PRECISION	RECALL	F1-SCORE
0	0.69	0.65	0.67
1	0.63	0.61	0.62
2	0.81	0.82	0.81
3	0.92	0.96	0.94
4	0.91	0.95	0.93

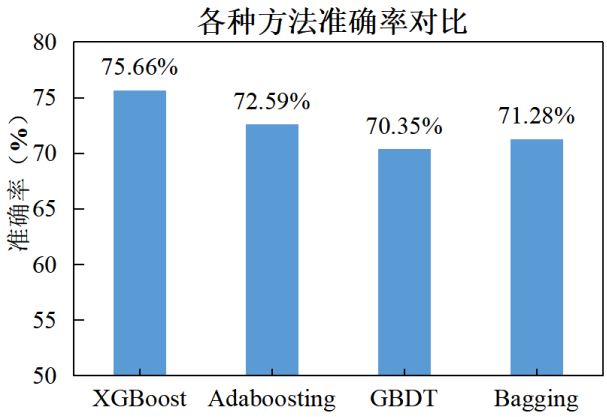

Figure 8. The accuracy of various methods
图 8. 各种方法分析的准确率

4. 结论

使用多层雷达回波数据进行降雨分析预测时，本文尝试采用 XGBoost 方法进行分析，并构建了可用的分析系统。经过数据预处理、模型建立、超参数选取、模型优化等过程后，得到了较好的降雨值与降雨等级分类结果。此外，还比较了 XGBoost 方法与其他机器学习方法的分类结果，结果表明，XGBoost

较主流的 Bagging、Adaboosting、GBDT 等方法，结果更贴近实际，这说明使用 XGBoost 方法进行降雨相关分析预测是可行的。

参考文献

[1] 王慧媛. 基于深度学习的短时定量降水预测研究[D]: [学位论文]. 金华: 浙江师范大学, 2020.

[2] Suzana, R. and Wardah, T. (2011) Radar Hydrology: New Z/R Relationships for Klang River Basin, Malaysia. *Proceedings of* 2011 *International Conference on Environment Science and Engineering*, Bali Island, 1-3 April 2011.

[3] 汪瑛, 冯业荣, 蔡锦辉, 胡胜. 雷达定量降水动态分级 Z-I 关系估算方法[J]. 热带气象学报, 2011, 27(4): 601-608. http://dx.chinadoi.cn/10.3969/j.issn.1004-4965.2011.04.018

[4] Brandes, E.A. (1975) Optimizing Rainfall Estimates with the Aid of Radar. *Journal of Applied Meteorology and Climatology*, **14**, 1339-1345. https://doi.org/10.1175/1520-0450(1975)014%3C1339:OREWTA%3E2.0.CO;2

[5] Sakaino, H. (2013) Spatio-Temporal Image Pattern Prediction Method Based on a Physical Model with Time-Varying Optical Flow. *IEEE Transactions on Geoscience and Remote Sensing*, **51**, 3023-3036. https://doi.org/10.1109/TGRS.2012.2212201

[6] Lee, S., Cho, S. and Wong, M. (1998) Rainfall Prediction Using Artificial Neural Networks. *Journal of Geographic Information and Decision Analysis*, **2**, 233-242.

[7] Kufigowski, R.J. and Barros, A.P. (1998) Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks. *Weather and Forecasting*, **13**, 1194-1204. https://doi.org/10.1175/1520-0434(1998)013%3C1194:LPFFAN%3E2.0.CO;2

[8] Luk, K.C., Ball, J.E. and Sharma, A. (2001) An Application of Artificial Neural Networks for Rainfall Forecasting. *Mathematical and Computer Modelling*, **33**, 683-693. https://doi.org/10.1016/S0895-7177(00)00272-7

[9] Chau, K.W. and Wu, C.L. (2010) A Hybrid Model Coupled with Singular Spectrum Analysis for Daily Rainfall Prediction. *Journal of Hydroinformatics*, **12**, 458-473. https://doi.org/10.2166/hydro.2010.032

[10] 时玮域. 基于机器学习方法的雾天气预测研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2020.

[11] 王训师. XGBoost 机器学习模型在缺血性卒中后早期认知损害诊断中的应用研究[D]: [博士学位论文]. 杭州: 浙江大学, 2018.

[12] 王晓晖, 张亮, 李俊清, 孙玉翠, 田捷, 韩睿毅. 基于遗传算法与随机森林的 XGBoost 改进方法研究[J]. 计算机科学, 2020, 47(Z2): 454-458+463.

[13] Zhang, X.M., Yan, C., Gao, C., Malin Bradley, A. and Chen, Y. (2020) Predicting Missing Values in Medical Data via XGBoost Regression. *Journal of Healthcare Informatics Research*, **4**, 383-394. https://doi.org/10.1007/s41666-020-00077-1

[14] 陈逸伦. 基于多源卫星数据的云团和雨团识别及其特征研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2019.

[15] Kato, S., Rose, F.G., Rutan, D.A., Thorsen, T.J., Loeb, N.G., Doelling, D.R., *et al*. (2018) Surface Irradiances of Edition 4.0 Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Data Product. *Journal of Climate*, **31**, 4501-4527. https://doi.org/10.1175/JCLI-D-17-0523.1

[16] Guo, J.Q., Dai, Y.Z., Wang, C.X., Wu, H., Xu, T.Y. and Lin, K. (2020) A Physiological Data-Driven Model for Learners' Cognitive Load Detection using HRV-PRV Feature Fusion and Optimized XGBoost Classification. *Software: Practice and Experience*, **50**, 2046-2064. https://doi.org/10.1002/spe.2730