成振华,周坤论,陶伟,等. 基于三种机器学习算法的降水现象仪和雨量筒数据一致性检验[J]. 气象研究与应用,2022,43(3):115-119.
Cheng Zhenhua,Zhou Kunlun,Tao Wei,et al. Data consistency test of precipitation phenomenometer and rain gauge based on three machine learning algorithms[J]. Journal of Meteorological Research and Application,2022,43(3):115-119.

基于三种机器学习算法的降水现象仪和雨量筒数据一致性检验

成振华，周坤论，陶 伟，黄剑钊，王 玮，景 坤
（广西壮族自治区气象技术装备中心，南宁 530022）

摘要：基于三种机器学习算法，对 2018 年南宁国家气象观测站雨量筒观测数据和降水现象仪的雨滴观测数据进行一致性检验试验。通过降维算法,对降水现象仪数据去除数据冗余,进一步分别采用多元线性回归、决策树回归、最近邻回归等 3 种机器学习算法验证与雨量筒数据的一致性情况。结果表明,综合性能中多元线性回归算法效果最好,在误差范围内的准确率达到 85%以上;最近邻回归算法在小雨量中可以有较好的预测值,综合准确率达到 75%,两种算法均优于决策树算法 70%的准确率。

关键词：多元线性回归;决策树回归;最近邻回归;降维算法;气象数据质量
中图分类号：P412 **文献标识码**：A **doi**：10.19849/j.cnki.CN45-1356/P.2022.3.21

引言

降水现象仪在地面气象观测中逐渐地被普及使用,可以观测降水雨滴谱数据,得到降水过程中的各种参数(如降水粒子的直径、降水粒子的下落速度、不同降水粒子的数量等)。在人工影响天气、天气现象类型判定中具有重要作用[1]。南宁国家气象观测站于 2017 年安装并使用至今,是南宁国家站智能观测天气现象的主要手段之一。降水现象仪是一种采用现代激光技术的光学测量仪器，它主要由激光发生器和激光接收器组成。工作时激光发生器向接收器发射激光束,当降水粒子从激光束中通过时,会遮挡激光束造成接收器端电压的变化，从而确定粒子大小。利用粒子从进入激光束的时间,到完全离开激光束的时间可以判定粒子的下降速度。测量粒子直径范围为 0~24.5mm,速度最大可达 20.8m·s^{-1},分钟降水量最大值为 6mm,可以满足南宁市的降水观测需求[2]。随着自动气象观测的不断发展,“监测精密”对气象探测质量提出了更高要求。周坤论等对比了广西全区的降水现象仪与人工观测天气中存在差距[3]。刘平等指出 81.8%的降水现象仪，得出的降水量绝对偏差和相对偏差都较大[4]。经过分析得出可能的原因，是由于计算过程中对降水现象仪的精度要求较高，在实际中可能会出现降水现象仪在预测过程中,将粒子直径或者速度误判的情况,因此造成较大偏差。

机器学习(machinelearning,ML)算法近年来发展成熟,广泛应用于医学、经济学、生物学、农学、气象学等领域[5]。相关人工智能算法也引入到降水预报业务中[6]。ML 技术在数据挖掘领域有较多应用,在算力充足时可以弥补传统算法建模复杂的问题[7]。按照是否需要训练数据，机器学习分为监督学习和无监督学习两类。其中监督学习主要有线性回归(linearregression)、逻辑回归(logisticregression)、K 近邻算法（K-Nearest neighbor，K-NN）决策树(decisiontrees，DTs）等。无监督学习主要有聚类算法(Clustering Algorithm)以及主成分分析法等(principalcomponentanalysis，PCA)[8]。在数据处理中高维

收稿日期：2022-04-12
基金项目：降水天气现象仪探测数据订正模型本地化研究(桂气科 2022QN09)
作者简介：成振华(1995—),男,山西临汾人,助理工程师,从事地面气象探测工作。E-mail:845869155@qq.com

稀疏数据带来的冗余数据会影响算法模型的准确率，并且提升时间和空间存储的复杂度。通过数据降维可以减少数据冗余，提升算法的表示性能。本文尝试采用机器学习算法对降水现象仪和雨量筒的数据进行拟合试验，从而检验降水数据的一致性情况。提升观测业务的准确性。

1 资料和方法

1.1 资料

本文的试验数据主要是南宁气象观测站 2018 年雨量筒数据和降水现象仪按分钟输出的雨滴观测数据。其中降水资料通过地面气象综合观测业务软件(ISOS)采集，观测粒子速度和直径分为 32 级共 1024 维，每个维度用 4 位数表示该维度雨滴粒子的个数。

1.2 方法

在进行降水现象仪和雨量筒数据一致性检验试验时，本文主要利用 PCA 算法将高维稀疏的降水现象仪输出数据降维，减少数据冗余。进一步分别采用多元线性回归算法、决策树算法、最近邻算法等 3 种机器学习算法对降维后的雨滴谱数据进行拟合计算，并分析各种算法性能差异化的原因。各种降维和机器学习算法介绍如下：

1.2.1 主成分分析法

主成分分析法(PCA)通过去除冗余信息可有效降低数据维度[9]。数据集样本协方差越小，则认为该维度的信息更不容易被区分，该维度携带的信息量越小；反之则认为该维度携带的信息量越大。由于协方差矩阵对称，因此选取的各个特征之间相互正交，保证了分解之后的各个特征之间不相关，使得分解结果的各个分量可以单独讨论。计算公式如(1)所示：

$$\Sigma=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1}$$

其中 Σ 代表样本方差，x_i 为第 i 个特征的样本数据，$\bar{x}$ 为样本均值。

1.2.2 多元线性回归

多元线性回归分析是根据给定的多维自变量数据，计算对应的因变量数据[10]，其公式如(2)所示。本文进行雨量估计时，降水现象仪输出不同粒子直径、不同粒子数量的数据可以看作数据的自变量 x。雨量筒输出的雨量数据，可以看作多元线性回归的因变量 y。在计算中只需找到两者的对应关系，通过给定自变量 x 和因变量 y，可以求得对应的关系 w。

$$y=w\cdot x \tag{2}$$

其中：

$$y=\begin{bmatrix}y_1\\y_2\\\cdots\\y_n\end{bmatrix}\quad w=\begin{bmatrix}w_1\\w_2\\\cdots\\w_n\end{bmatrix}\quad x=\begin{bmatrix}x_{11}&x_{12}&\cdots x_{1m}\\x_{21}&x_{22}&\cdots x_{2m}\\&\cdots\cdots&\\x_{n1}&x_{n2}&\cdots x_{nm}\end{bmatrix} \tag{3}$$

1.2.3 决策树

决策树算法[11]以树形结构建立模型，在模型中，决策节点、叶节点和分支定义了一系列可用于案例分类的决策。通过对树的生长算法做小幅调整，这些树同样可以应用于数值预测。相对于比较常见的回归方法，决策树主要有以下优点：决策树能自动选择特征，允许决策树与大量特征一起使用。具有更多特征的决策树，对某些数据类型的适应能力比线性回归强。分析表明，在降水现象仪统计时，如果出现较多错误数据，则决策树回归模型预测效果相对较好，如果错误数据较少，则决策树模型拟合效果一般。

1.2.4 K 近邻回归

K 近邻回归算法凭借计算量小、可解释强等优点近年来被广泛应用于气象研究中。该算法通过计算目标值周围最近的数值结果，加权平均得到目标值的数据[12]。在计算过程中需要对数据维度进行正确地划分。在雨滴谱拟合降雨量的过程中，雨滴谱输出的每个降雨维度都可以作为降雨量的一个特征，具有相似雨滴谱特征的降雨量，也具有相似的雨量值，因此可以作为拟合雨量的算法。在实际使用中需要求已获得的数据样本足够多，可以在目标值周围产生足够的近邻选择。K 近邻回归算法不用担心雨滴谱中，各个维度之间的参数制约关系，数据即使不为线性数据也可以有较好的回归结果。

k 近邻回归的算法流程：

(1) 找出距离目标值最近的 k 个相似雨滴谱特征的雨量值；

(2) 计算 k 个雨量值的平均数作为预测值的实际雨量；

(3)不断重复筛选 k 值使得结果最优。

2 一致性检验

试验中，首先采用 PCA 算法对数据进行降维处理，其次分别采用多元线性回归算法、决策树算法、最近邻算法等 3 种机器学习算法进行拟合计算。在

进行机器学习算法降水数据集的构建中，筛选出降水现象仪与雨量筒观测的雨量数据记录共有 309 个小时。本文将其中的 80%作为训练集用以训练模型，20%作为测试集验证算法效果。

2.1 PCA 降维

雨滴谱将雨滴数据划分 1024 维雨滴数据，分别对应不同的半径和粒子速度。其中南宁一整年的降雨过程没有雨夹雪和雪，因此在较高维度中数据为 0，因此在计算过程中可以直接将该列数据删除。将空列数据删除后雨滴谱输出维度为 426 维。

删除空维度后保证了每列数据均有不为 0 的雨量记录，但是仍然存在大量为 0 的单列数据。在进行计算时稀疏性的高维数据会使得计算值产生较大偏差，影响计算效果。

通过 PCA 算法将高维数据降维，可以降低计算过程中的计算量，增加每个维度在计算中的占比权重。降维后还可以减少因雨滴谱设备本身性能问题带来的噪声信号。

2.2 多元线性回归

利用小时雨滴谱参数加权求和可以得到小时雨量，经过降维后多元线性回归方程写为：

$$\begin{cases} y_1=w_0+w_1x_{11}+w_2x_{12}+\cdots+w_nx_{1n} \\ y_2=w_0+w_1x_{21}+w_2x_{22}+\cdots+w_nx_{2n} \\ y_m=w_0+w_1x_{m1}+w_2x_{m2}+\cdots+w_mx_{3n} \end{cases} \tag{4}$$

其中 y 表示雨量筒统计的小时降雨量。x 表示对应小时的雨滴谱降维数据。w 为小时雨滴谱各个维度对应的参数。

利用最小化损失函数：

$$s=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 \tag{5}$$

即：

$$s=\sum_{i=1}^{m}(y_i-X_w)^2 \tag{6}$$

其中 y_i 为雨量筒雨量，$\hat{y}_i$ 为回归预测雨量。

利用最小二乘法求参数 w，等式两边对 w 求导并令一阶导数为 0 得到：

$$w=(X^TX)-1X^Ty \tag{7}$$

此时解出参数 w 并可以预测雨量筒雨量 y_i 的值。通过拟合估计得出基于多元线性回归的雨量预测如图 1 所示：

在雨量预测中训练集里的每个样本维度 x 都会被赋予对应的系数 w，因此在拟合中对样本雨强的数据平衡性依赖较低，不会因为某一种雨强较少而降低拟合准确率。可以很好地克服实际中暴雨出现较少的场景。

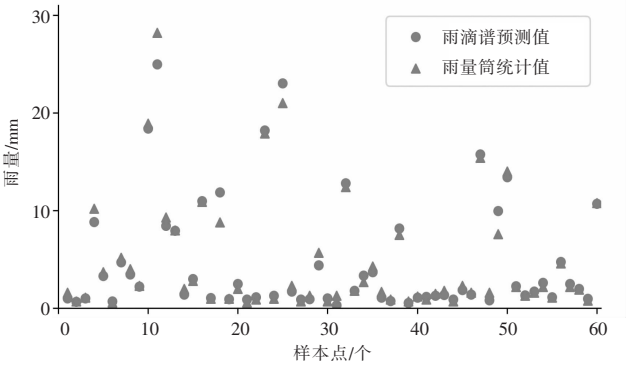

图 1　多元线性回归拟合雨量图

雨滴谱样本中存在 426 维非空的数据，因为降雨过程中雨滴存在相关性(包含线性和非线性相关)，预测雨量值时存在大量的冗余维度作为噪声，干扰多元线性回归系数的选取。通过 PCA 降维算法抑制噪声后进行多元线性回归拟合，会取得较好的效果。从图 1 中可以看出多元线性回归算法在预测值中，大部分的预测值与雨量筒值较为相近。经统计线性回归在降维成 29 维时，预测准确率最高，其中有超过 85%的预测值误差小于 0.7mm。

2.3 决策树拟合

经过 PCA 降维后的雨滴谱数据的每个维度，可以看作计算雨量值的一个特征。因此在计算过程中，可以利用决策树预测实际的雨量值。决策树算法首先构建根节点，将雨滴谱数据放入其中，筛选出根节点中条件最好的分类。然后按照此规则依次将各个分类的子集中的雨滴谱数据继续分类，直到得出叶子节点对应的雨量数据为止。如果重复完该过程后依旧有未成功分类的节点，则根据最优特征分类将雨滴谱数据归为其中。拟合结果如图 2 所示：

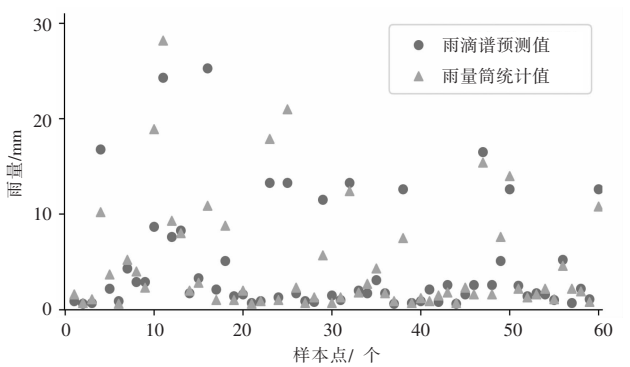

图 2　决策树拟合雨量图

分析图 2 可知，决策树算法在预测值中大部分的预测值与真实值偏差较大。但总体预测稳定性较好。决策树算法通过判断每层逻辑分支进行拟合。当拟合数据维度较少时，会因数据特征重叠导致误判；当拟合数据维度较多时，又会导致分支数较多导致拟合性能下降。经统计，决策树的降维最佳维度为 40 维，其中有超过 70%的预测小时雨量值误差小于 0.7mm。在三种算法中拟合效果最差。

导致这种情况的原因，是由于决策树拟合过程中，将目标训练数据作为离散的分类点。在进行测试数据拟合时，通过平滑离散点形成一条曲线。因此在拟合出的结果中几乎所有点都存在一定量的偏差。

2.4 最近邻拟合

通过找出一个样本的 k 个最近邻居，将这些邻居的某个(些)属性的平均值赋给该样本，就可以得到该样本的值。最近邻算法预测结果如图 3 所示：

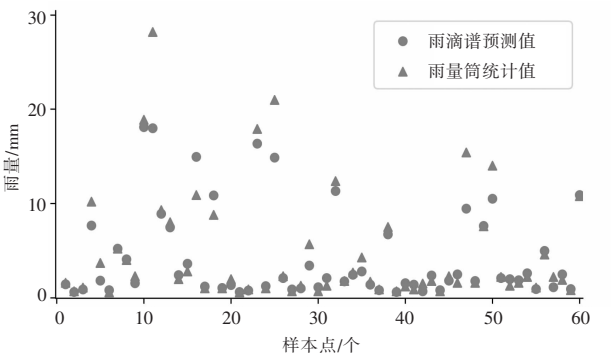

图 3 KNN 拟合雨量图

从图 3 中可以看出 KNN 算法在预测值中大部分的预测值，与真实值偏差总体不大。这是由于 KNN 算法在预测中取周围点的平均值。在降雨过程中小时雨量超过 7mm 的次数较少，测试集目标值周围的点不够均匀，因此 KNN 算法对较大雨量的拟合准确率较低。而小雨量的目标值在样本中较多，可以充分取到周围的近邻点做平均，因此预测准备率较高。

对于近邻值 K 的选择。当 K 值选择较小时，会因选取点不够均匀，导致预测结果产生偏差；当 K 值选择过大时，会导致计算量和空间开销增加，并且会将原本无关的样本点归在一起进行预测。准确率随 K 值变化如图 4 所示，当 K 值取 6 时效果最佳，综合判定超过 75%的预测值误差小于 0.7mm。

2.5 不同机器学习算法估计的对比分析

对上述一致性检验试验的结果进行对比分析发现，3 种机器学习算法中多元线性回归算法的预测成功率最好。这表明降水现象仪输出中的各个粒子直径和速度的雨滴谱数据与雨量之间的确存在线性关系。通过计算各个粒子的权重，累和后可以得到雨量筒数据。计算过程中部分数据出现小时雨量超过 0.7mm 以上情况主要分布在较大量级的雨量附近。分析发现造成这一现象的原因可能是雨量较大时雨量筒在翻转过程中存在一些外溢损耗，因此雨量筒未成功探测到雨量数据。

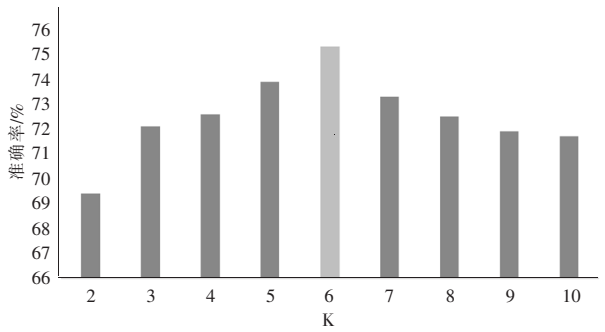

图 4 不同 K 值下的准确率

最近邻回归算法通过计算距离目标值最近的 K 个点的均值，代表目标值数据。通过分析小时雨量数据发现小时雨量超过 50mm 的样本较少。在计算这部分雨量时，没有足够多的近邻数据进行加权平均，因此出现了较大误差。较多的小雨量样本使得在计算小雨量时，有足够多的相似近邻因此预测值较为接近，这也说明降雨过程的雨滴粒子直径和数量较为均匀。

决策树算法在三者当中，误差明显高于另外两种的原因是：决策树在进行回归计算时主要通过分类的方式进行。通过分类得到的决策树具有较好的可理解性，但是在线性拟合中将线性的结果，转化为分类结果会产生过拟合现象。

3 结论与讨论

本文采用三种机器学习算法，对 2018 年南宁国家站降水现象仪观测数据与雨量筒数据，进行了一致性拟合试验。结果表明，不同的机器学习算法在预测中均有较高的拟合度。其中多元线性回归算法在拟合中预测中最多的小时降雨量，达到了 85%；最近邻算法在预测中雨量较小时综合准确率为 75%，在样本较多时具有很好的预测准确性。当设备稳定运行时决策树综合预测值最差只有 70%。

在预测降水过程中并未克服传统机器学习存在

的弊端，无法进一步提升算法的表示性能。在最近邻算法中，距离度量方法有多中，判定出的“近邻”也可能有显著区别，从而影响拟合效果。后续可尝试不仅使用欧式距离，改用马氏距离或其他距离作为选择度量。在最近邻算法中如果解决样本不均衡问题，可以获得更好的预测效果。线性回归中对设备稳定运行率有较高要求，设备出现过大值的原因主要因为设备运行稳定率不高导致。在后续工作中考虑将不同算法的优势结合起来，从而提高算法的效果。

参考文献：

[1] 李力，姜有山，蔡凝昊，等.Parsivel 降水粒子谱仪与观测站雨量计的对比分析[J].气象，2018，44(3)：434-441.

[2] 周坤论，张哲睿，成振华，等.北海一次强降雨过程的雨滴谱特征分析[J].气象研究与应用，2022，43(2)：16-22.

[3] 周坤论，黄剑钊，陶伟，等.降水类天气现象自动与人工观测质量对比分析[J].气象研究与应用，2022，43(1)：112-117.

[4] 刘平，王磊，祁生秀，等.天气现象仪降水观测分析[J].成都信息工程大学学报，2020，35(1)：104-110.

[5] 周冠博，钱奇峰，吕心艳，等.人工智能在台风监测和预报中的探索与展望[J].气象研究与应用，2022，43(2)：1-8.

[6] 何慧，陆虹，覃卫坚，等.人工神经网络在月降水量预测业务中的研究和应用综述[J].气象研究与应用，2021，42(1)：1-6.

[7] 覃卫坚，何莉阳，蔡悦幸.基于两种机器学习方法的广西后汛期降水预测模型[J].气象研究与应用，2022，43(1)：8-13.

[8] 何清，李宁，罗文娟，等.大数据下的机器学习算法综述[J].模式识别与人工智能，2014，27(4)：327-336.

[9] 李蝉娟. 高维数据降维处理关键技术研究[D].桂林：电子科技大学，2017.

[10] 吴玉霜，黄小燕，陈家正，等.机器学习在广西台风极大风速预报中的应用[J].气象研究与应用，2021，42(4)：26-31.

[11] 陆虹，翟盘茂，覃卫坚，等.低温雨雪过程的粒子群-神经网络预报模型[J].应用气象学报，2015，26(5)：513-524.

[12] 黄明明，林润生，黄帅，等.一种新的基于伪最近邻算法的降水预报方法[J].科学技术与工程，2018，18(17)：222-228.

Data consistency test of precipitation phenomenometer and rain gauge based on three machine learning algorithms

Cheng Zhenhua, Zhou Kunlun, Tao Wei, Huang Jianzhao, Wang Wei, Jing Kun

(Guangxi Meteorological Technology Equipment Center, Nanning 530022, China)

Abstract: Based on three machine learning algorithms, a consistency test was conducted on the rain gauge observation data and the raindrop observation data of the precipitation phenomenon instrument of Nanning National Meteorological Observatory in 2018. Firstly, the dimensionality reduction algorithm is used to remove the data redundancy of the precipitation phenomenon meter data. Three machine learning algorithms including multiple linear regression, decision tree regression, and nearest neighbor regression are further used to verify the consistency with the rain gauge data. The results shows that the multiple linear regression algorithm has the best effect in the comprehensive performance, and its comprehensive accuracy rate is more than 85% within the error range, followed by the nearest neighbor regression algorithm with the accuracy rate reaching 75%, which has a better performance in predicting light rainfall. Both of the above algorithms outperformed the decision tree algorithm with 70% accuracy.

Key words: multiple linear regression, decision tree regression, nearest neighbor regression, dimensionality reduction algorithm, quality of meteorological data