水电水利
Hydropower and Water Resources
第 8 卷◆第 10 期◆版本 1.0◆2024 年
文章类型：论文|刊号（ISSN）：2529-7821 /（中图刊号）：868GL002

基于气象驱动的中长期水文预报初探

郑翰　孙克雷
安徽理工大学
DOI:10.12238/hwr.v8i10.5808

[摘　要] 鉴于近年以来全球变暖趋势日益明显,极端天气频现,防汛抗旱形势日益严峻。本文根据民间谚语中的“冬行春令,夏必干旱”和“春行冬令,夏必水涝”的思想启发,与日益火爆的人工智能机器学习预测算法——随机森林算法,使用长江安徽段多年的气象水文资料作训练集,验证探索冬行春令和春行冬令的现象对当地的水旱影响情况,为中长期水文预报工作提供一个创新的思路与方法,也证明了古圣先贤对自然社会的观察与规律总结的智慧。
[关键词] 水文预报；冬行春令夏必旱；随机森林算法
中图分类号：P336　文献标识码：A

Study on medium and long term hydrological forecast based on meteorological drive

Han Zheng　Kelei Sun
Anhui University of Science and Technology

[Abstract] In view of the increasingly obvious trend of global warming in recent years, the frequent occurrence of extreme weather, and the increasingly severe situation of flood control and drought relief, this paper is inspired by the ideas of "winter travel in spring, summer will be drought" and "spring travel in winter, summer will be flood", and the increasingly popular artificial intelligence machine learning prediction algorithm – random forest algorithm. Using the meteorological and hydrological data of Anhui section of the Yangtze River for many years as a training set, to verify and explore the effects of the phenomenon of winter and spring on local floods and droughts, to provide an innovative idea and method for the medium and long term hydrological forecast, and to prove the wisdom of ancient sages in observing and summarizing the laws of natural society.
[Key words] hydrological forecast; Winter line spring when summer will be drought; Random forest algorithm

引言

水文预报是一种重要的防洪非工程措施,在我国防汛、抗旱、水资源开发利用、国民经济建设和国防领域有广泛应用,经济效益巨大。水文预报主要研究如何根据水文规律,预报未来水情、旱情和冰情,以短期的水文预报方法为主,而缺少中长期的水文预报方式。目前,水文气象预报技术主要是以基于统计学的致灾阈值模型和分布式水文模型等为主,结合大数据分析与人工智能的气象-水文-地质耦合预报模式将在水文气象预报中发挥重要作用。而气候系统中的水循环也处于不断运转演化和更新中,民间谚语中的“冬行春令”和“春行冬令”便是影响水循环的重要气象因素。

1 理论与方法

1.1民谚“冬行春令夏必旱”的相关理论

“冬行春令”指如果冬天至雨水节气前气温较高,湿度较大,像春天一样,到了夏季一定干旱。这句话也是说冬天提前过了春天,把天上的云即水汽等都提前降了下来,而且春季本身多雨,这样从冬行春令开始的漫长雨季便把云中水汽减少很多。虽然也有地表水的蒸发作用,但春季的气温正常情况下蒸发量较少。所以到了夏天云中没有了多少水汽,便一定干旱。

“春行冬令”指春天过得像冬天一样干冷少雨,夏必水涝。这句话的原理是冬行春令的补充,就是说春天若是干冷少雨,那么云中汽未降,地表水总是会蒸发部分,这样就会导致夏天多雨,而且夏天温度高蒸发量也大,更加速了天地间的水循环,必然导致洪涝灾害。

而如果冬行春令和春行冬令同时存在,譬如2024年上半年长江安徽段芜裕河段的情况。春节前后如暖春气候,气温高达10~20摄氏度,按民谚的理论,夏必干旱。但春季却偏寒冷,雨水甚少,所以到了夏季,便洪涝严重。但是洪涝结束后,有干旱且大地龟裂。所以,如果暖冬和寒春同时存在,便是夏天旱涝也是同时发生。

文章类型：论文|刊号（ISSN）：2529-7821 /（中图刊号）：868GL002

1.2随机森林算法

为验证民谚的“冬行春令,夏必干旱”和“春行冬令,夏必水涝”的理论的准确性,本文通过水文预报模型来检验实现其理论。

水文预报模型可分为水文模拟模型、传统统计模型和现代智能算法模型三大类。水文模拟模型主要基于产汇理论和河道演进理论对洪水过程进行模拟[1],需要较为完整的水文实测数据和下垫面资料；统计模型是采用分析预报对象与前期预报的特征因子之间的关联关系,然后基于训练样本拟合变量间的数值统计关系进行预报洪水[2]；现代智能算法大多是属于数据挖掘技术,基于智能算法的预报模型是直接通过训练样本确定模型结构、参数、对变量的依赖关系进行拟合[3],而无需明确的假设条件,具有强大的功能[3]。

随机森林采用Bootsrap重抽样方法从原始样本中随机抽取n组样本,并对每组样本进行决策树建模,得到序列分类模型;预测时有建立的分类模型序列可得到n种结果,最后采用投票方法或取平均值作为最终分类或预测的结果。

随机森林算法与其他常见的集成学习算法,如Adaboost、GBDT等存在一些区别。在样本权重处理方式上,Adaboost会根据前一轮基学习器的预测结果,为样本赋予不同的权重,错误分类的样本权重增加,正确分类的样本权重降低；随机森林在构建每棵树时,对样本进行有放回的抽样,样本权重相对较为均衡。

在弱学习器的构建方式上,GBDT中的弱学习器是回归树,通过不断拟合前一轮学习器的残差来构建新的学习器；而随机森林的弱学习机是决策树,并且在构建决策树时,对特征的选择也是随机的。在并行性方面,随机森林中的树可以并行生成,计算效率较高;ADaboost和GBDT则是按顺序依次构建弱学习机,难以并行计算。在对异常值的敏感度方面：随机森林对异常值不敏感,因为它是通过多个树综合结果进行预测的;而GBDT相对更容易受到异常值的影响。在模型复杂度上,随机森林的模型相对简单,不需要复杂的调参;GBDT等算法在调参方面可能需要更多的技巧和经验。在可解释性上,随机森林中单个决策树具有一定的可解释性,但整体模型的解释性相对较弱;GBDT由于是基于残差的拟合,其可解释性相对较准。

随机森林算法在众多领域都有广泛的应用,常见领域有金融领域的信用风险评估和股票市场预测、医疗领域、市场营销领域、环境领域、农业领域、工业领域、图像识别和计算机视觉、自然语音处理等方向。

综上所述,随机森林算法具有的优越性比较适合中长期水文预报实际采用。

1.3选取特征因子

在水旱灾害防御中,依据历史洪涝实勘水文资料,分析上下游站点的情况、前后时间段流量的相关关系,就能获得影响控制断面洪水流量的特征因子。例如2024年汛前的长江安徽段实际流量情况,春季少雨,但长江上游多雨,洪涝灾害严重,湖南暴雨连绵不断,洞庭湖漫灌决堤。于是洪峰过境,上游水流全排到了长江中下游也就是安徽段,所以长江安徽段的水位居高不下。以至于到了汛期暴雨来临之时,芜裕河段的水位即超警戒线。所以,选取特征因子需要考虑的主要因素固然是借鉴民间谚语的“冬行春令”和“春行冬令”的重要影响,还要联系上下游的气候影响和下游的堤坝拦截情况,如大型水电站、拦水坝等的建筑影响情况。

2 实例应用

2.1流域概况

皖江即专指长江安徽段干流,是长江的重要组成部分。而芜裕河段为长江安徽段的芜湖至裕溪口河段。芜湖自古以来便有万里长江的咽喉重镇之说。又因长江在芜湖段拐了一个弯,俗称大拐,水流至此湍急而过,水下地形更显复杂多变。如此复杂多样的水文情境,是否还符合“冬行春令,夏必干旱”、“春行冬令,夏必水涝”的规则,可以通过历年的冬季和春季气象和水旱灾害预报验证,同样也可通过人工智能算法神经网络方法等来验证说明。同时,也可依据验证结果来决定是否能通过此规则来做中长期水文预报的预警方式。

2.2算法应用

随机森林算法作为集成学习算法,它结合了多个决策树的预测结果,以提高预测的准确性和稳定性。其实验步骤和主要特点原理包括：数据抽样：通过自助抽样(bootstrap sampling)从原始数据集中有放回地抽取多个样本集；特征随机选择：在构建每个决策树时,不是使用全部特征,而是随机选择部分特征来进行分裂；构建决策树：基于抽样得到的样本集和随机选择的特征,构建多个决策树；集成预测：综合多个决策树的预测结果,通过投票(分类任务)或取平均值(回归任务)来得到最终的预测结果。

使用随机森林算法时,可以通过调整参数来优化模型性能：n_estimators参数决定了森林中树的数量,树越多模型效果通常越好,但训练和预测的时间复杂度也会增加；max_features参数是控制每棵树分裂时考虑的最大特征数,较小的值通常能增加树的差异性,防止过拟合的出现；max_depth参数能限制树的最大深度,防止单棵树过于复杂导致过拟合,降低算法的泛化能力；min_samples_split和min_samples_leaf参数是控制节点分裂和叶子结点的最小样本数,防止模型出现过拟合。Bootstrap参数决定是否使用Bootstrap抽样法；oob_score参数使用袋外样本(Out-of-Bag)来评估模型性能,而无需额外的验证集。n_jobs参数指定并行运行的任务数,可以加速训练过程;random_state参数用于设置随机种子,确保参数的可重复性。作为一个强大灵活的机器学习算法,随机森林算法适用于各种需要高准确性和稳定性的任务。通过合理选择参数和调优模型,可以进一步提升其预测性能。

2.3实验步骤

使用随机森林算法进行预测涉及到以下实验步骤,这些步骤可以通过各种编程语言和机器学习库来实现,其中Python和

文章类型：论文|刊号（ISSN）：2529-7821 /（中图刊号）：868GL002

scikit-learn库是常用的组合。

2.3.1安装必要的库

首先,安装Python和scikit-learn库。可以通过pip命令进行安装：

```
pip install scikit-learn pandas
```

2.3.2导入所需的库

在Python脚本中导入必要的库：

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier #对于分类任务
from sklearn.metrics import accuracy_score, classification_report
```

2.3.3加载数据

加载数据集,即气象水文资料数据集。

2.3.4数据预处理

根据数据集的特点进行预处理,如缺失值处理、异常值处理、特征编码、特征放缩等。比如日期数据,需要处理成二维矩阵的形式表示。

2.3.5划分训练集和测试集

将数据集划分为训练集和测试集,以便评估模型的性能：

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=42)
```

2.3.6创建随机森林模型

创建随机森林分类器的实例并设置相应的参数：

```
rf_classifier=RandomForestClassifier(n_estimators=100,random_state=42)
```

2.3.7训练模型

使用训练集数据训练随机森林模型：

```
rf_classifier.fit(X_train,y_train)
```

2.3.8进行预测

使用训练好的模型对测试集进行预测：

```
y_pred=rf_classifier.predict(X_test)
```

2.3.9评估模型

计算并输出模型的性能指标,如准确率、分类报告等：

```
accuracy=accuracy_score(y_test,y_pred)
print(f'Accuracy:{accuracy:2f}')
print(classification_report(y_test,y_pred)
```

2.3.10特征重要性评估

还可以评估每个特征对模型预测的重要性,这是随机森林算法的又一大功能：

```
importances=rf_classifier.feature_importances_feature_nammes = X.columns
indices = np.argsort(importances)[::-1]
plt.figure(figsize=(10,6))
plt.title("Feature Importances")
plt.bar(range(X.shape[1]),importances[indices],align="center")
plt.xticks(range(X.shape[1]),feature_names[indices],rotation=90)
plt.show()
```

2.3.11参数调优

随机森林有多个超参数可以调节,如n_estimators(树的数量)、max_depth(树的最大深度)、min_samples_split(内部节点再划分所需最小样本数)等。可以通过网格搜索(Grid Search)或随机搜索(Random Search)进行超参数调优,以进一步提升模型性能。

3 实验结果

依据历史气温和降水数据、对应于夏天的旱涝与否,可以观测到冬行春令与夏必旱、春行冬令与夏必涝的强相关性。同时,根据随机森林算法的实验结果,我们也验证了这一俗语的准确性,并能进一步强化这一特征因子。

但由于实验条件的局限性和数据采集处理的差异和异常数据的影响干扰因素,可能对实验结果产生影响,对结论的建立可能还欠缺严谨性与全面性。

4 结语

通过历史数据的直观观测和机器学习算法的数据验证,本文丰富了中长期水文预报的理论与方法,对实际经济社会与生产生活提供了气象水文方面的中长期预测保障,对防汛抗旱工作的预测调度和堤防修坝等水利工程设施建设都有一些经验等贡献。虽然实验固有一些局限性,但对实际水文工作的建议和未来气象水文研究方面的启示探索的新方法、新领域、新问题的产生都是有一定贡献意义的。

[参考文献]

[1]芮孝芳.洪水预报理论的新进展及现行方法的适用性[J].水利水电科技进展,2001,21(5):1-4.

[2]XU K Q,BROWN C,KWON H-H,et al.Climate teleconnections to Yangtze river seasonal streamflow at the Three Gorges Dam,China[J].International Journal of Climatology,2007,27(6):771-780.

[3]BREIMAN L.Statistical modeling:The two cultures[J].Statistical Science,2010,16(3):199-215.

作者简介：

郑翰(1992--),女,安庆桐城人,硕士研究生,工程师,从事测绘制图有关工作。