DOI:10.3969/j.issn.1004-6933.2014.01.007

基于随机森林模型的需水预测模型及其应用

王 盼,陆宝宏,张瀚文,张 巍,孙银凤,季 妤

(河海大学水文水资源学院,江苏 南京 210098)

摘要:为解决需水预测模型精度问题,尝试基于随机森林模型的分类和回归功能构建需水预测模型。以苏州市需水量预测为研究实例,首先应用随机森林模型的分类功能将需水预测因子分类,经计算发现第一产业比例、人口、灌溉面积、万元产值用水量和国民经济生产总值为最重要的解释变量。在此基础上,用随机森林模型的回归功能对需水进行预测,同时采用相同的训练数据建立基于 BP 神经网络和 RBF 神经网络的需水预测模型,通过对比 3 个模型的预测结果,发现随机森林模型能有效预测需水量,且精度较高。

关键词:需水预测;随机森林模型;神经网络模型;解释变量;OOB 交叉验证

中图分类号:P333.9 **文献标志码**:A **文章编号**:1004-6933(2014)01-0034-04

Water demand prediction model based on random forests model and its application

WANG Pan, LU Baohong, ZHANG Hanwen, ZHANG Wei, SUN Yinfeng, JI Yu

(*College of Hydrology and Water Resources, Hohai University, Nanjing* 210098, *China*)

Abstract: In order to improve the accuracy of a water demand prediction model, we attempted to use the classification and regression functions of the random forests model to construct a water demand prediction model. Taking the water demand forecast in Suzhou City as a case study, we used the classification function to classify the water demand prediction factors, and found that the most significant explaining variables were the primary industrial structure, population, irrigation area, water use per 10 000 yuan, and GDP. On this basis, we used the regression function to predict the water demand, and used the same training data to construct the water demand prediction model based on the BP neural network and RBF neural network models. Through comparison of the prediction results of the three models, we drew the conclusions that the random forests model can effectively forecast the water demand, and it has higher precision than the other two models.

Key words: water demand prediction; random forests model; neural network model; explaining variable; OOB cross validation

随着人口的增加和经济的快速增长,水资源短缺已成为制约国民经济和社会可持续发展的瓶颈。研究需水问题,科学预测需水量,对环境和社会经济的协调发展有重要的意义。

传统的需水预测方法有时间序列法和相关分析法。时间序列法主要包括趋势外推法、随机模型等,相关分析法主要包括回归分析法、弹性系数法、灰色预测以及用水定额法。影响需水量的因素较多,且我国需水量数据时间系列短,因此用常规的预测方法来保证需水预测精度是比较困难的。近年来有学者利用人工神经网络[1-2]、小波神经网络模型[3]和支持向量机[4]等方法对区域需水量进行预测。这

基金项目:国家自然科学基金(NSFC-50979023);水利部公益性行业科研专项(201201026)

作者简介:王盼(1989—),女,硕士研究生,研究方向为水资源规划与管理。E-mail:hhu_wp@163.com

通信作者:陆宝宏,副教授。E-mail:lubaohong@126.com

些方法能以较高精度逼近任何非线性连续函数[5]，从而通过训练样本数据得到精确的预测结果。单分类器的计算精度往往差强人意，且容易引发过度拟合现象，因此，越来越多的学者通过集成的方法来提高预测精度。集成的方法即分类器组合方法。2001年，Breiman[6]将他在1996年提出的Bagging集成学习理论和Ho[7]在1998年提出的随机子空间理论相结合而提出随机模型；2007年，有学者[8]通过对Random Forests进行分析，提出了Random Multi-Nomial Logit（RMNL）和Random Naive Bayes（RNB）两种算法，并进行了实验与SVM进行了比较；2008年，该学者又将MultiNomial Logit和Random Forests这两种方法结合起来，实现了一种新的多分类器的模型Random Multi-Nomial Logit（RMNL），通过使用Random Forests弥补了Multi-Nomial Logit方法不能处理大的特征空间的缺点。国内有关随机森林模型的应用集中在生物信息学[9-10]、医学诊断[11]以及文档检索[12]上，对理论的探讨还比较少。笔者考虑到影响需水量的因素众多，采用随机森林算法构建需水预测模型，并进行实例研究。

1 随机森林需水预测模型

1.1 随机森林理论

随机森林（random forests）模型是一种比较新的机器学习模型。经典的机器学习模型是神经网络，有半个多世纪的历史了。神经网络预测结果精确，但是计算量很大。20世纪80年代Leo Breiman等[6]发明了分类树的算法，通过反复二分数据进行分类或回归，使计算量大大降低。2001年Breiman把分类树组合成随机森林，在每棵决策树的构建过程中使用了两次随机，一是构建决策树时使用的训练样本是通过bootstrap法在原始数据集中随机获取的；二是每棵决策树所使用的解释变量也是在原有特征集的基础上随机获取，生成很多分类树，再汇总分类树的结果。随机森林模型在运算量没有显著提高的前提下提高了预测精度。由于随机森林模型对多线性不敏感，预测结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用[13]，被誉为当前最好的算法之一[14]。

决策树是随机森林模型的基础分类器，其构造由一个独立同分布的随机向量决定。如果将决策树看成分类任务中的一个专家，那么随机森林模型就是许多专家在一起对某种任务进行分类。随机森林模型是由多个决策树$\{h(X,\theta_k)\}$组成的分类器，其中θ_k是独立同分布的随机变量。输入向量$\boldsymbol{X}$的最终所属类别由森林中所有决策树投票决定。

在随机森林模型中，当决策树的数目很大时，遵循大数定律。随机森林模型不会随着分类树的增加而过度拟合，但有一个有限的泛化误差值，这表明随机森林模型可以为未知实例预测提供很好的参考思路。随机森林模型也可以用来处理回归问题。有关随机森林模型的基本理论的介绍详见文献[8]。

1.2 模型的构建和参数选取

文中使用R语言中的Random Forests包构建随机森林分类及回归模型，使用的函数主要有Biplot、partialPlot、Random Forests、importance、varImpPlot、predict等。

随机森林模型的构建过程中，预测性能很大程度上依赖于参数的选取。为了保证参数选取完全独立于测试数据，降低泛化性，在模型训练的过程中，随机森林模型使用的是OOB交叉验证算法。在随机森林的试验中，装袋法（Bagging）和随机特征选择并行应用。装袋法中每一个新训练集，都是由原始训练集使用Bootstrapping随机重复采样得到的。应用这种方法生成的训练集只能包含原训练集约67%的样本，其余的样本作为袋外数据，这样就可以用袋外数据来评价模型的性能。随机森林模型对所有袋外样本的拟合误差是其预报误差的无偏估计，这样可以保证训练数据集和测试数据集完全独立，从而提高预测精度。

笔者应用随机森林模型的分类功能来评价各解释变量的重要性。文中的需水量预测因子是解释变量，需水量是被解释变量，构建的随机森林模型可解释各变量的重要性。对所有检验样本来说，随机打乱某一解释变量取值，采用原模型对测试样本进行再次预测，袋外拟合误差越大，说明该解释变量越重要。在解释变量贡献度的基础上，应用随机森林模型的回归功能进行需水预测。

2 实例应用

2.1 苏州市用水现状

苏州市位于长江三角洲、太湖流域腹部，多年平均降水量为1 100 mm，年径流量的多年平均值为240 000万m^3。随着城市化的发展，苏州市工业用水和生活用水量增长明显，用水总量从2000年的516 060万m^3增至2010年的783 140万m^3。尽管水利部门加大从长江及太湖的引水力度，但水资源短缺仍是制约苏州市持续快速发展的重要因素之一。

2.2 需水预测因子的选取与解释

影响需水预测的因子有很多，综合考虑社会经济因素，笔者选取人口（a）、灌溉面积（b）、国内生产总值（c）、第一产业比例（d）、第二产业比例（e）、第三产

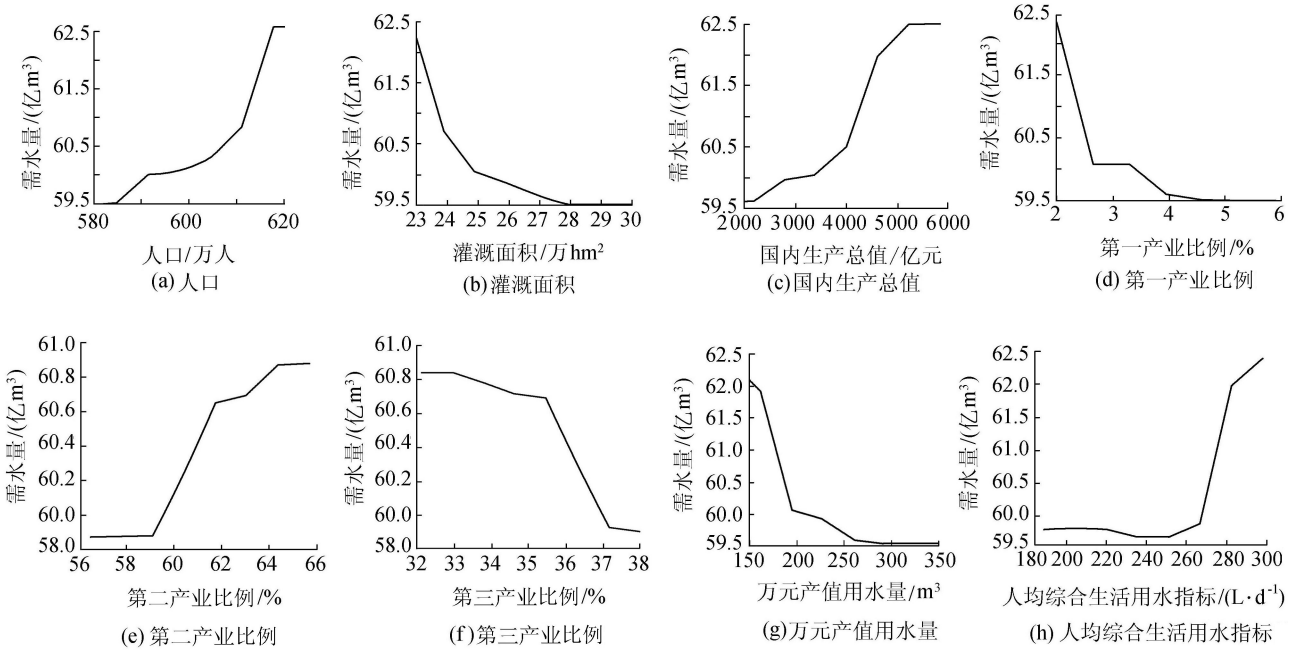

图1　各解释变量对需水量的影响

业比例(f)、万元产值用水量(g)以及人均综合生活用水量(h)8个指标。这些指标能全面反映苏州市的需水现状。由于数据预处理对随机森林模型的计算结果基本没有影响,所以可直接使用指标值进行计算,但是在使用BP模型和RBF神经网络模型进行需水预测时,需要对原始数据进行归一化处理。这里选取《苏州市水资源公报》2000—2010年的需水量数据,其中,将2000—2007年的数据作为训练数据,将2008—2010年的数据作为测试数据。

使用R语言中的程序包来分析解释变量的重要性。图1为各解释变量的部分依赖图,表明每个解释变量对需水量的影响,是对每个变量的边际效应的图形描述。图2为主成分得分及荷载在主成分空间的散点图,是变量和记录在第一、第二主成分上的值。图3中,精度平均下降率是用来度量去除一个解释变量后随机森林分类的准确度降低的程度,该参数值越大,则该变量越重要。对于文中的需水预测模型来讲,最重要的解释变量为第一产业比例、

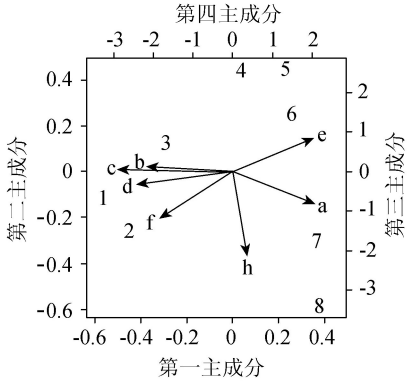

图2　主成分分析散点图

人口、灌溉面积、万元产值用水量和国内生产总值,在进行需水预测时,应注意对这5个变量的预测精度进行控制。

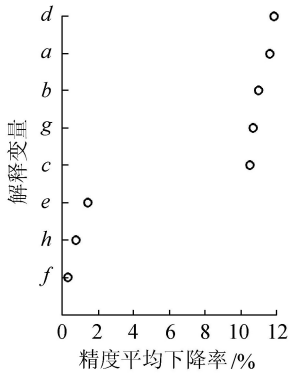

图3　各解释变量重要性度量

2.3　预测结果及误差分析

考虑到随机森林模型是一种可以有效避免过度拟合的新型机器学习语言,适用于计算长系列数据,笔者尝试将其应用到短系列数据的计算中。为了验证预测结果的可靠性和预测精度,笔者用相同的训练样本分别建立了基于BP神经网络和RBF神经网络的需水预测模型,以便进行预测结果的比较。其中,BP神经网络模型的网络隐含层和输出层均采用S型函数,而RBF神经网络的作用函数为高斯函数。这两种模型的预测过程均是在Matlab中实现的。

利用OOB交叉验证法得到的随机森林模型预测结果为:决策树的数量在取120的时候,模型误差最小。随机森林模型计算得到的预测值和实际需水量的比较结果见表1。使用随机森林模型和BP神经网络模型和RBF神经网络模型训练数据所得到

的对比结果见图 4,3 种模型平均相对误差(绝对值)分别为 2.009%、4.588%、3.603%。使用 3 种模型对 2008—2010 年的需水预测结果的比较见表 2。由表 2 可见,使用随机森林模型、BP 神经网络模型和 RBF 神经网络模型,预测结果的平均相对误差(绝对值)分别为 2.318%、4.834% 和 3.920%。由此可见,随机森林模型的预测精度较高,用于需水预测效果很好,是一种有效的需水预测方法。不难看出,随机森林模型也可以应用于短系列数据的计算中,且预测精度高于其他两种模型。

表 1　随机森林(RF)模型的预测值与实际需水量的比较

年份	实际需水量/万 m^3	预测需水量/万 m^3	绝对误差/万 m^3	相对误差/%
2000	516060	523855.9	−7795.9	−1.511
2001	503501	518124.1	−14623.1	−2.904
2002	540727	536304.5	4422.5	0.818
2003	570836	581131.5	−10295.5	−1.804
2004	606928	595574.4	11353.6	1.874
2005	614585	599247.3	15337.7	2.496
2006	745479	731026.6	14452.4	1.939
2007	747678	727276.5	20401.5	2.729

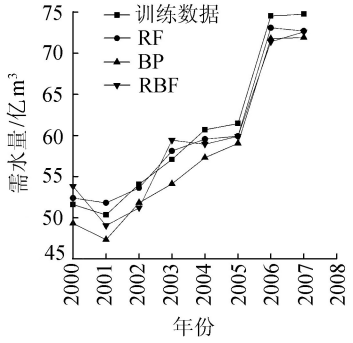

图 4　3 种需水预测模型预测结果对比

表 2　3 种模型的预测结果与实际需水量的相对误差

年份	实际需水量/万 m^3	RF 预测		BP 预测		RBF 预测	
		需水量/万 m^3	相对误差/%	需水量/万 m^3	相对误差/%	需水量/万 m^3	相对误差/%
2008	755238	763848.4	−1.141	726714	3.777	733901	2.823
2009	767234	742151.6	3.269	735581	4.123	739973	3.553
2010	783140	763220.0	2.544	731450	6.600	741003	5.381

随机森林模型适用于计算大规模数据,而文中由于资料有限,用于计算的样本偏少,在计算中发现,随着选取解释变量个数的增加,单棵决策树的强度不会产生很大变化,但树与树之间的相关性会增加;同时,泛化误差与 OBB 误差随着特征个数的增加而增大。而根据 Breiman 的理论,强度越大且相关性越小的随机森林模型预测结果越好。样本过少是导致误差存在的主要原因,增大样本容量,增加解释变量的数量,可以很大程度提高预测精度。随机森林模型是一种有效的估计缺失数据的方法,当集中有大比例的数据缺失时仍然可以保持预测精度不变。

3　结　论

利用随机森林模型的特征及其在回归和分类上的重要应用,笔者使用短系列数据建立了基于随机森林模型的需水预测模型,用于解释变量的重要性和进行需水量的预测。经过独立数据集的验证,表明所选影响因子能够较好反映需水量特征,该模型能有效预测需水量。此外,文中还利用 BP 神经网络和 RBF 神经网络算法构建了预测模型,3 种模型的预测结果表明,使用随机森林模型得到的预测结果要优于 2 种神经网络算法的结果,同时反映出解释变量选取的合理性和有效性。

对随机森林模型的泛化性有严格的数学证明,该算法不会过度拟合。比起广义线性模型,随机森林模型对数据前提条件的要求等要宽松得多,因此其应用前景非常广阔。

参考文献:

[1] 雷雨,郑旭荣,李玉芳,等. 神经网络方法在城市需水量预测中的应用[J]. 石河子大学学报:自然科学版,2006(1):22-25. (LEI Yu, ZHENG Xurong, LI Yufang, et al. Application of neural network methods on the urban water demand forecasting [J]. Journal of Shihezi University: Natural Science, 2006(1):22-25. (in Chinese))

[2] 龙训建,钱鞠,梁川. 基于主成分分析的 BP 神经网络及其在需水预测中的应用[J]. 成都理工大学学报:自然科学版,2010,37(2):206-210. (LONG Xunjian, QIAN Ju, LIANG Chuan. Water demand forecast model of BP neutral networks based on principle component analysis [J]. Journal of Chengdu University of Technology: Science & Technology Edition, 2010, 37(2): 206-210. (in Chinese))

[3] 罗利民,方浩,仲跃,等. 小波神经网络算法在区域需水预测中的应用[J]. 计算机工程与应用,2006(3):200-201. (LUO Limin, FANG Hao, ZHONG Yue, et al. The algorithm of wavelet neural network and its application to area water demand prediction [J]. Computer Engineering and Applications, 2006(3):200-201. (in Chinese))

[4] 刘俊萍,畅明琦. 基于支持向量机的需水预测研究[J]. 太原理工大学学报,2008(3):299-302. (LIU Junping, CHANG Mingqi. Water demand prediction model based on support vector machine [J]. Journal of Taiyuan University of Technology, 2008(3):299-302. (in Chinese))

[5] LU B, GU H, XIE Z, et al. Stochastic simulation for determining the design flood of cascade reservoir systems [J]. Hydrology Research, 2012, 43(1/2):54-63.

[6] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2):123-140.

(下转第 89 页)