文章编号:1007-5321(2021)04-0082-07　　**DOI**:10.13190/j.jbupt.2020-137

数据驱动的城镇智慧水务日用水量预测算法

姚俊良，　薛海涛，　刘　庆
(西安理工大学 自动化与信息工程学院，西安 710048)

摘要:针对国内某中小型自来水公司的实际供水情况,通过对比相关系数分析了天气等因素对日供水量的影响,确定了日用水量预测所需的输入参数;比较了 3 种用传统基于大数据的水量预测方法在该自来水公司中应用的性能,针对用传统方法预测误差较大的问题,提出引入前一日用水量和前 8 h 的用水量作为影响因素的改进方法. 将所提方法在该自来水公司的信息系统中进行了实际测试,验证了所提算法的有效性. 根据算法性能和实现复杂度,给出了适用于城镇水务的水量预测算法和算法执行形式,能够帮助水务企业提高水量预测精度,有效提升水资源的利用率.

关　键　词:智慧水务;水量预测;大数据;神经网络算法
中图分类号:TP183　　**文献标志码**:A

Daily Water Volume Prediction Algorithm of Urban Smart Water Based on Big Data

YAO Jun-liang,　XUE Hai-tao,　LIU Qing
(School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: According to the actual water supply situation of a small and medium-sized water company in China, the influences of weather and other factors on daily water supply are analyzed by comparing the correlation coefficient, so as to determine the input parameters required for daily water consumption prediction. The application performance of three traditional water volume prediction methods is compared using the actual operating data. To solve the severe errors existing in the traditional methods, an improved method is proposed , which takes the water consumption of the previous day and 8 hours into consideration. The efficiency of the proposed algorithm is verified by the tests in the information system of the water supply company. According to the performance and implementation complexity of the algorithm, a water quantity prediction algorithm and its suitable implementation form for urban water affairs are proposed, which can help the water affair system improve the water quantity prediction accuracy, thus effectively improving the utilization rate of water resources.

Key words: smart water platform; water volume prediction; big data; neural network algorithm

国家智慧城市(区、镇)试点指标体系(试行)和《国家新型城镇化规划(2014—2020 年)》中明确指出,智慧水务建设是智慧城市建设的重要组成部分,要利用信息技术手段,对从水源地监测到水龙头管

收稿日期:2020-08-28
基金项目:国家自然科学基金项目(51706180, 61502385);陕西省自然科学基础研究计划项目(2020JM-456)
作者简介:姚俊良(1984—),男,副教授,硕士生导师,E-mail:yaojunliang@xaut.edu.cn.

理的整个供水工程实现实时监测管理，保障居民用水安全. 精确地用水量预测是实现智慧水务建设重要且必要的环节. 吴弯等[1-2]通过分析不同水平年居民的发展指标和水供需发展趋势进行了需水量预测，属于宏观水量预测，无法用于日水量或时段水量等更细粒度的预测. 基于大数据的学习算法能够精细刻画用水量与影响因素之间的关系，可用于更细粒度的水量预测，主要包括多元线性回归（LR，linear regression）方法、支持向量回归（SVR，support vector regression）方法和误差反向传播神经网络（BPNN，back propagation neural network）方法等[3]. Karamaziotis 等[4]以预测当天的天气情况、最高温、最低温以及是否节假日为主要影响因素，分别利用改进的时间序列预测方法和机器学习算法进行短期水量预测，取得了较为理想的效果. 然而传统的回归学习算法通常不会考虑数据之间的时序关系，而时序预测算法在应用于城市用水量预测这类多因素、非线性、时变性的问题时效果不理想. 笔者以国内某中小型自来水公司 2019 全年的供水数据为对象，对 3 种典型算法的性能进行评估. 根据对算法进行的改进，最后给出了仿真结果及对比分析.

1　城镇用水量及其影响因素分析

以国内某中小型自来水公司 2019 年的日用水量数据作为样例对城镇的用水量规律进行分析. 首先，获取 2019 年的水量数据，在对数据进行抽取和异常值剔除之后，将所得的 2 ~ 11 月的日用水量数据作为整体数据集.

天气情况、日最高温、日最低温以及是否节假日等因素对水量有较大影响[4-10].

笔者统计了 2019 年全年日用水量和 4 个影响因素的相关数据并选取气温较为适宜的 4 月份数据进行分析，如图 1 所示. 其中，天气因素是从中国天气网获取的当地气象数据，图 1(c)右纵坐标各个数值对应的天气情况如表 1 所示. 图 1(d)右纵坐标中的 1.0 表示节假日，0 表示非节假日.

从图 1(a) ~ (c)可以看出，除个别日期以外，日用水量与当日最高温和最低温呈现明显的正相关性. 同时，当天气状况良好且未发生较大改变的情况下，日用水量没有大幅波动且呈现逐步上升的趋势，当出现阴雨天气时，日用水量相对较小. 因此，

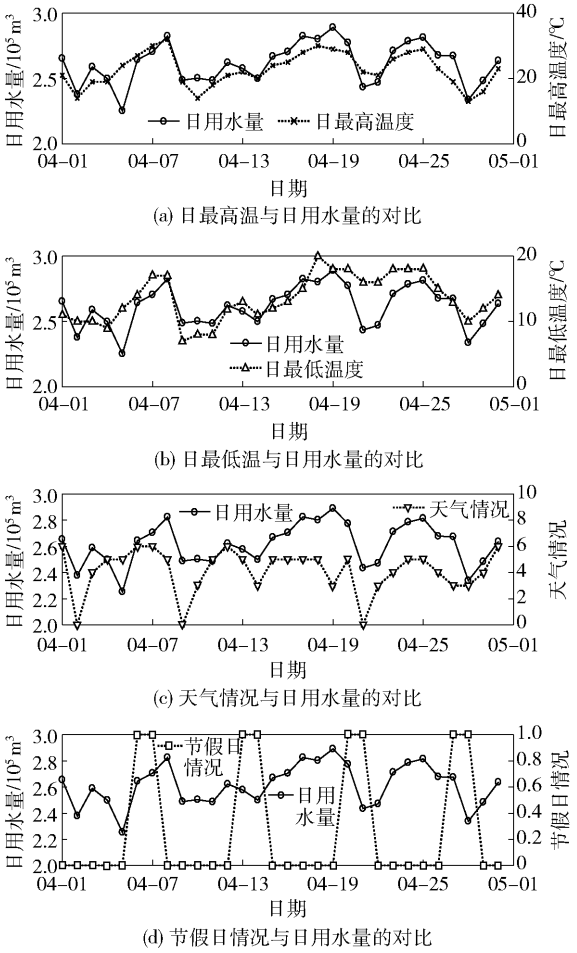

图 1　2019 年 4 月份日用水量与影响因素对比

将日最高温度、日最低温度以及天气情况作为影响日用水量的因素是合理的. 而由于在自来水公司的供水范围内工业企业的分布性质，节假日对日用水量的影响并不大，如图 1(d)所示.

表 1　不同天气情况量化对照表

天气情况	数值
大到暴雪、浮尘、扬沙、强沙尘暴、中到大雪、大雪、暴雪、沙尘暴	0
中到大雨、暴雨、大暴雨、特大暴雨、大到暴雨、暴雨到大暴雨、大暴雨到特大暴雨	1
小到中雨、小到中雪、雨夹雪、中雪、冻雨	2
小雨、小雨转多云、阵雨、雷阵雨、雷阵雨伴有冰雹、阵雪、小雪	3
阴、雾、霾	4
多云、晴转多云	5
晴	6

为了准确对比上述各因素对日用水总量的影响程度，选取 4 月、9 月、11 月以及全年的数据，分别从

单个月份和整体上分析不同气温条件下各因素与日用水量之间的相关性. 通过求取两者之间的相关系数进行具体分析,若影响因素用向量 $\boldsymbol{x}$ 表示,日用水量用向量 $\boldsymbol{y}$ 表示,则两向量的相关系数为

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

其中:$x_i(i=1,2,\cdots,n)$ 为向量 $\boldsymbol{x}$ 的元素,$y_i(i=1,2,\cdots,n)$ 为向量 $\boldsymbol{y}$ 的元素,n 为总的样本数目,$\bar{x}$ 和 $\bar{y}$ 分别为2个向量均值. 利用式(1)计算得到的各相关系数如表2所示.

表2 各影响因素与日用水量相关系数

影响因素	4月	9月	11月	全年
日最高温	0.779 9	0.660 3	0.895 1	0.825 2
日最低温	0.684 9	0.239 6	0.810 1	0.775 3
天气状况	0.439 8	0.736 1	0.724 6	0.326 2
节假日	−0.129 6	−0.133 6	−0.018 7	−0.032 7

由表2可以看出,每日最高温度与日用水量的相关程度最高且在不同气温条件下整体上的相关系数无较大变化. 每日最低温度与日用水量之间的相关系数虽然有较大波动但整体上二者仍具有较高的相关性. 天气状况与日用水量的相关程度次之,在冬夏极端气温条件下对日用水量的影响程度明显提高,但整体上与日用水量的相关程度较低. 节假日与日用水量的相关系数均为负值,日用水量呈下降趋势,但由于相关系数的数值较小,所以两者相关程度最低.

2 常见的日用水量预测算法及性能

2.1 算法原理

线性回归算法作为机器学习中的一种较为基础的算法,其原理简单且算法运行稳定,是水量预测最常用算法之一.

针对水量预测的问题,可以建立多元线性回归模型,有

$$h_\theta(\boldsymbol{x})=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_lx_l=\sum_{i=0}^{l}\theta_ix_i=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x} \tag{2}$$

其中:$h_\theta(x)$ 为算法所需预测日用水量;$x_1,x_2,\cdots,x_l$ 为影响因变量的自变量,分别对应表2的影响因素;$\theta_1,\theta_2,\cdots,\theta_l$ 为各自变量对应的权重系数,θ_0 为回归常数.

支持向量回归算法[11]是在支持向量机算法的基础上建立的,适用于解决水量预测这类回归问题,其基本思想是在样本空间中找到一个超平面,使得训练样本尽可能分布在靠近超平面的两侧,求解出该超平面即可得到SVR建立的模型.

设超平面方程为 $f(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$,利用SVR求解出的最终模型为

$$\left.\begin{aligned}f(\boldsymbol{x})&=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\kappa(\boldsymbol{x},\boldsymbol{x}_i)+b\\ b&=y_i+\varepsilon-\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}\end{aligned}\right\} \tag{3}$$

其中:m 为训练样本数,采用处理后的日用水量数据作为训练样本;α_i 和 $\hat{\alpha}_i$ 为拉格朗日乘子;$\kappa(x_i,x_j)$ 为核函数;y_i 为第 i 个样本的实际日用水量;ε 为SVR所允许的误差范围.

在上述模型中,b 的计算需选择满足 $0<\alpha_i<C$ 的样本,C 为正则化惩罚系数. 若 C 过大,会降低对误差的容忍程度,易出现过拟合现象;若 C 取值过小,则会造成欠拟合,故需确定适当的 C 值. 当核函数选择径向基(RBF, radial basis function)核函数时还需要选择其自带的参数 g,此参数决定映射后特征空间的分布,可使用网格搜索得到最优的核函数选择以及参数 C 和 g.

BPNN算法[12]主要包括输入层、隐含层和输出层,分别包含不同数目的神经元,因而具有学习、联想、记忆和模式识别等信息处理功能,其基本思想是:输入学习样本,使用反向传播算法对网络的权值和偏差进行反复更新,使输出向量与期望向量尽可能接近. 当网络输出层的误差平方和小于设定的门限时,保存网络的权值和偏差.

BPNN算法用于水量预测的具体步骤如下:

1) 根据日用水量的影响因素确定每一层网络的神经元个数、激活函数、预测精度和最大迭代次数;

2) 初始化连接权重,随机生成每层网络之间的连接权重及偏差;

3) 随机选取一个不重复样本,计算隐层与输出层神经元的输出值;

4) 根据当前网络的输出与实际用水量之间的误差,反向更新连接权重;

5）将所有样本输入更新后的网络，计算全局误差，判断预测精度是否达到要求或迭代次数是否达到最大，若是，则保存当前网络连接权重及偏差，训练完成；否则，执行第 3）步.

2.2　仿真结果

利用 2.1 节所述算法对自来水公司 2019 年全年的运行数据进行实验. 在对数据进行归一化处理后选取 85% 的数据作为样本，其余 15% 的数据作为测试集，针对 4 个影响因素，基于 Python 平台对算法进行建模与测试. 其中，SVR 算法参数：C 取值为 3，核函数选用 RBF 核函数，g 取值为 0.016. BPNN 算法参数：输入层神经元设为 4 个，隐层神经元设为 20 个，输出层神经元设为 1 个，激活函数为 identity 函数. 日用水量的预测误差为

$$e = \frac{Q_1 - Q_2}{Q_1} \tag{4}$$

其中：Q_1 为归一化后的真实日用水量，Q_2 为预测日用水量.

经过统计平均，LR 算法、SVR 算法和 BPNN 算法用于该自来水公司实际数据的平均预测误差分别为 8.53%，8.71% 和 8.37%.

3　改进的日用水量预测算法

从实验结果可以看出，使用 LR、SVR、BPNN 这 3 种算法在所分析的数据集上的预测精度不高，无法满足实际应用的要求.

结合自来水公司运行的实际情况，需要在每日上午 8:00 给出当日用水量数据，每日前 8 h 的用水量数据具有很高的价值，获取也比较容易，但在实际中并没有被很好地利用；另外，时间序列规律特别是前一日用水量对于水量预测也非常重要. 因此，考虑将前一日用水量与前 8 h 用水量也作为影响因素进行分析. 统计自来水公司 2019 年 4 月份前一日用水量和前 8 h 用水量的信息如图 2 所示. 可以看出，当日用水总量与前一日用水量、前 8 h 用水量基本同步，因此，考虑将两者作为影响日用水量的因素是合理的.

为了更加精确地进行比较，根据式（1）相关性的分析，计算出在不同季节下，前一日用水量和前 8 h 用水量与日用水量的相关系数，结果如表 3 所示.

通过相关性分析可以看出，城市居民日用水量

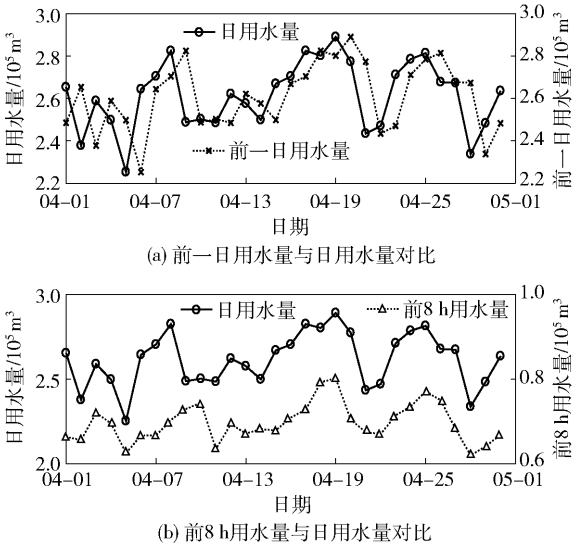

图 2　日用水量与前一日用水量、前 8 h 用水量对比

表 3　两种影响因素与日用水量相关系数

影响因素	4 月	9 月	11 月	全年
前一日用水量	0.364 9	0.626 7	0.816 4	0.969 6
前 8 h 用水量	0.698 8	0.765 0	0.627 1	0.960 7

与前一日用水量以及前 8 h 用水量均具有高相关性且在极端气温下仍具有较大的相关系数. 因此，将上述两因素输入预测算法之中，预计能够达到比较好的预测效果.

结合前文对于各个影响因素的相关性分析，在算法中区分不同因素对预测结果的影响程度，可以利用各个因素与日用水量之间的相关系数，在输入算法时给不同特征分配不同的权重. 例如，可以给日最高温度这一特征分配较大的权重，给节假日这一因素分配以较小的权重，以区分其对日用水量的不同贡献.

对上述改进算法进行建模与测试. 首先在 2019 年的整体数据集上进行测试，设置测试集占总体数据的比例为 15%. SVR 算法参数：C 取值为 68，核函数选用 RBF 核函数，g 取值为 0.01. BPNN 算法参数：输入层神经元设为 6 个.

算法改进前后在测试集上的每日预测误差统计如图 3 所示. 3 种算法改进前后的预测平均误差对比结果如表 4 所示. 可以看出，改进后的算法能够有效减小预测误差且每日预测误差的波动幅度比之前更小，预测结果更加稳定.

在不同的测试集比例下，改进算法的预测效果

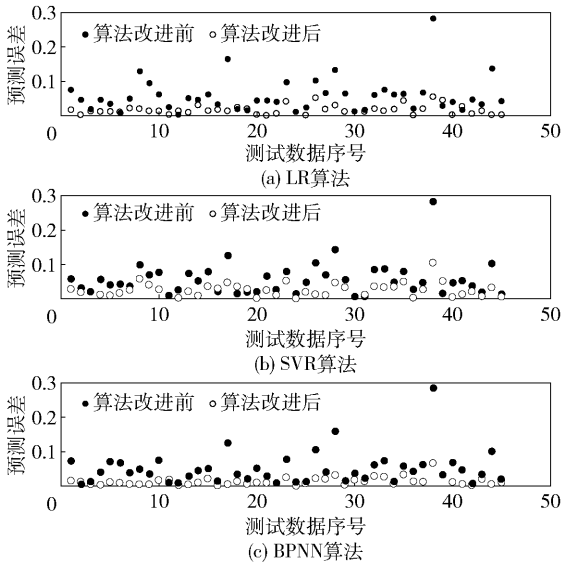

图3 算法改进前后日用水量预测误差

如表5所示.

由表5看出,在不同测试集比例下,线性回归算法均取得了较好预测效果. SVR算法的平均预测误差受测试集比例的影响最大,不够稳定,预测效果最差. BPNN算法受测试集比例的影响较小,在测试集上的预测效果最好.

表4 算法改进前后的平均预测误差 %

算法	改进前预测误差	改进后预测误差
LR	8.53	2.38
SVR	8.71	3.69
BPNN	8.37	2.14

表5 不同测试集比例下算法的平均预测误差 %

测试集比例	LR算法	SVR算法	BPNN算法
10	2.03	4.81	1.53
15	2.38	3.69	2.14
20	2.29	3.60	2.02
25	2.14	3.23	1.94
30	2.03	2.97	1.81

测试改进算法在不同气温条件下的预测精度. 首先对整体数据进行分组测试,将2~6月的数据划分为a组,将7~11月的数据划分为b组. 在a组中,将2~5月的数据作为训练集,6月的数据作为测试集;在b组中,7~10月的数据作为训练集,11月的数据作为测试集. 在此次测试中对于2组训练集和测试集的划分都是固定的,目的是测试算法在极端气温下的性能. 对于a和b两组数据选取相同的SVR参数及神经网络结构,但由于这里改变了训练样本集,故需重新确定SVR的参数选择,利用网格搜索后得到 C 值为1;核函数为线性核函数.

图4所示为改进算法在不同气温条件下的每日预测误差对比.

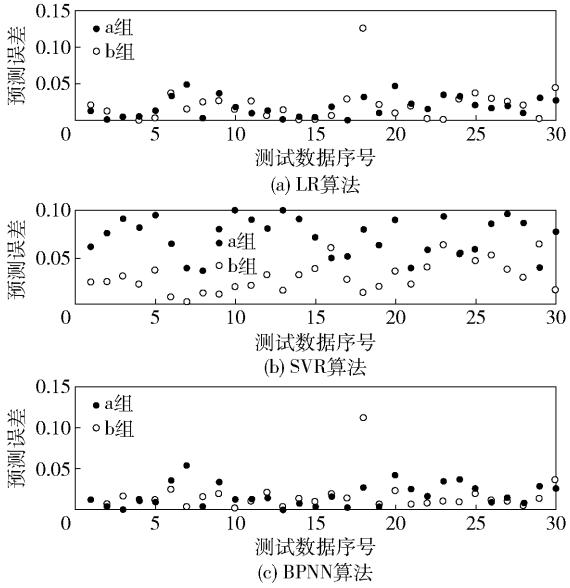

图4 不同气温条件下3种算法的日用水量预测误差

表6所示为3种算法在a和b组中的预测平均误差.

表6 不同气温条件下的预测平均误差 %

算法	a组	b组
LR	2.36	2.70
SVR	9.21	4.01
BPNN	2.22	2.07

从图4和表6可以看出,所提改进方案应用于线性回归算法和BPNN算法时,无论对于极高温天气或极低温天气都能够保持较高的预测精度. 当改进方案应用于SVR算法时,对于极端天气情况下的预测效果会变差. 从表5可以看出,SVR算法在应用于所给日用水量数据集时受训练数据集长度的影响很大,当增加数据集长度时,SVR算法的效果有明显改善. 而在分组测试时每一组的训练数据必然会大幅减少,因此使用SVR算法时会得出较差的预测结果.

下面对所用算法的时间复杂度进行分析. 由于3种算法的复杂度主要存在于训练阶段,预测阶段

可以忽略不计，所以仅分析算法模型训练的复杂度. 设训练数据长度为 k，特征维数为 d，训练数据集支持向量数目为 N，3 层 BPNN 每层的神经元数量分别为 m_1，m_2，m_3，则多元线性回归算法和 SVR 算法的时间复杂度分别为 $O(kd)$ 和 $O(N^2)$. 由于 BPNN 算法输入层和输出层的神经元个数是由数据集的特征决定的，m_1 和 m_3 可以视为常数，所以 BPNN 算法的训练时间复杂度可以简记为 $O(km_2)$.

将改进前后的 3 种算法分别应用于自来水公司 2019 年日用水量数据集并进行不同训练集长度的仿真实验，LR 算法与 SVR 算法各运行 1 万次后的仿真时间，而 BPNN 算法则在训练阶段随机抽取 1 万条数据进行训练后统计仿真时间，结果如图 5 所示.

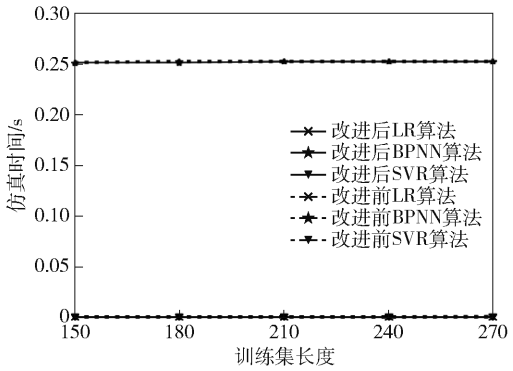

图 5　不同训练集长度下 3 种算法的仿真时间

可以看出，3 种算法在不同长度的训练集上的运行时间整体没有明显变化. 通过对比可以发现，改进前后 3 种算法的仿真时间均没有明显增加. 这是由于算法的改进方案主要是针对数据特征的有效提取，对算法本身没有做较大的变动. 而在自来水公司的实际运行过程中，前 8 h 用水量的获取十分便捷，基本在当天的前 8 h 过后就能够立刻获取到，这也符合国内自来水公司调度周期的划分，因此能够保障水量预测算法的实时性.

综合以上对比可以推断，针对该自来水公司的水量预测问题，所提出的改进算法能够明显改善用水量的预测误差，并且应用于线性回归算法和 BPNN 算法时对于极端天气情况仍能保持令人满意的效果. 而在 3 种算法的对比中，线性回归算法在大多数情况下能够保持稳定且具有较高的预测精度；SVR 算法的整体预测误差较大且对于极端天气情况下的小数据集并不十分适用，但在训练长度足够的情况下预测效果有很大改善；BPNN 算法具有最好的性能，但需要时刻提防出现过拟合和局部极值的问题. 而在算法复杂度方面，经仿真实验结果证明，改进后算法的时间复杂度没有明显增加，基本满足预测实时性的要求. 自来水公司可以根据实际情况选择合适的预测算法.

4　结束语

以国内某自来水公司日用水量作为样本，采用 3 种常见的基于大数据的水量预测算法，通过对天气、温度、节假日等影响因素的分析并结合自来水公司实际的运行情况提出一种日用水量预测改进算法. 从数据的提取与处理、日用水量规律及影响因素的分析、算法模型的建立、不同算法的性能对比等方面进行研究，实验结果证明，改进算法能够明显提高预测精度. 将水量预测算法应用于城镇智慧水务系统的建设中[13]，可以有效地指导生产调度.

参考文献：

[1] 吴弯. 城镇需水量预测方法研究[D]. 广州：华南理工大学，2014.

[2] 孙杰肖. 张家口市水中长期供需预测及平衡分析[D]. 保定：河北农业大学，2013.

[3] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：53-140.

[4] Karamaziotis P I, Raptis A, Nikolopoulos K, et al. An empirical investigation of water consumption forecasting methods[J]. International Journal of Forecasting, 2020, 36(2): 588-606.

[5] 刘志壮，吕谋，周国升. 基于小波组合模型的短期城市用水量预测[J]. 给水排水，2020，56(10)：110-114，131.
Liu Zhizhuang, Lü Mou, Zhou Guosheng. Short term water supply load forecasting based on wavelet combination model[J]. Water & Wastewater Engineering, 2020, 56(10): 110-114, 131.

[6] 罗华毅，王景成，杨丽雯，等. 基于时差系数的城市原水需水量预测应用[J]. 上海交通大学学报，2017，51(10)：1260-1267.
Luo Huayi, Wang Jingcheng, Yang Liwen, et al. Research and application of urban water demand forecasting based on time difference coefficient[J]. Journal of Shanghai Jiao Tong University, 2017, 51(10): 1260-

1267.

[7] 陆维佳, 朱建文, 叶圣炯, 等. 基于多因素长短时神经网络的日用水量预测方法研究[J]. 给水排水, 2020, 56(1): 125-129.
Lu Weijia, Zhu Jianwen, Ye Shengjiong, et al. Research on daily water quantity prediction method based on multi-variable long short term memory neuael network [J]. Water & Wastewater Engineering, 2020, 56(1): 125-129.

[8] 朱慧峰. 基于最小二乘支持向量机的城市供水短期水量预测研究[J]. 电气自动化, 2018, 40(1): 105-107.
Zhu Huifeng. Predictive study on short-term urban water supply based on least squares support vector machines [J]. Electrical Automation, 2018, 40(1): 105-107.

[9] 金冬梅, 荣楠. 基于回归分析的长春市需水量预测研究[J]. 东北水利水电, 2018, 36(10): 19-21, 29, 71.
Jin Dongmei, Rong Nan. Study on the water demand prediction of Changchun city based on regression analysis [J]. Water Resources & Hydropower of Northeast China, 2018, 36(10): 19-21, 29, 71.

[10] 杨丽雯. 城市供水管网短期需水量预测及优化调度研究[D]. 上海: 上海交通大学, 2016.

[11] 刘文, 高立慧. 基于支持向量回归机的城市用水量预测研究[J]. 价值工程, 2017, 36(22): 59-61.
Liu Wen, Gao Lihui. Study on urban water consumption forecast based on support vector regression[J]. Value Engineering, 2017, 36(22): 59-61.

[12] 王圃, 唐鹏飞, 白云, 等. 基于多分辨BP神经网络的城市日供水量预测模型[J]. 中国给水排水, 2018, 34(11): 51-55, 60.
Wang Pu, Tang Pengfei, Bai Yun, et al. Forecasting model of daily urban water supply based on multi-resolution BP neural network[J]. China Water & Wastewater, 2018, 34(11): 51-55, 60.

[13] 罗贤伟. 智慧水务评价指标体系研究[J]. 给水排水, 2020, 56(2): 125-128, 132.
Luo Xianwei. Research on smart water evaluation index system[J]. Water & Wastewater Engineering, 2020, 56(2): 125-128, 132.