计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500128

基于 XGBoost 算法的多元水文时间序列趋势相似性挖掘

丁　武[1,3]　马　媛[2]　杜诗蕾[2]　李海辰[3]　丁公博[3]　王　超[3]
1 华中科技大学水电与数字化工程学院　武汉 430074
2 太湖流域管理局水文局(信息中心)　上海 200434
3 中国水利水电科学研究院　北京 100038
(M201873779@hust.edu.cn)

摘　要　针对传统的利用神经网络等工具进行水文趋势预测得出结果不具备解释性等不足,文中提出一种基于机器学习算法的水文趋势预测方法,该方法旨在利用 XGBOOST 机器学习算法建立参照期与水文预见期之间各水文特征的相似度映射模型,从而在历史水文时间序列中匹配出与预见期水文趋势最相似的序列,从而达到水文趋势预测的目的。为了证明所提方法的高效性和可行性,以太湖水文时间序列数据为对象进行了验证。分析结果表明,基于机器学习的多元水文时间序列趋势相似性分析可以满足调度人员对未来水文趋势预测效果的要求。

关键词:机器学习;多元时间序列;水文趋势预测;时间序列数据挖掘;相似性度量
中图法分类号　TV121;P333

Mining Trend Similarity of Multivariate Hydrological Time Series Based on XGBoost Algorithm

DING Wu[1,3],MA Yuan[2],DU Shi-lei[2],LI Hai-chen[3],DING Gong-bo[3] and WANG Chao[3]
1 School of Hydropower and Information Engineering,Hua Zhong University of Science and Technology,Wuhan 430074,China
2 Taihu Basin Authority of Ministry of Water Resources(Information Center),Shanghai 200434,China
3 China Institute of Water Resources and Hydropower Research,Beijing 100038,China

Abstract　In view of the shortcomings of the traditional hydrological trend prediction using neural networks and other tools,the results are not interpretable and so on. This paper proposes a method of hydrological trend prediction based on machine learning algorithms,which aims to use the XGBOOST machine learning algorithm to establish a similarity mapping model for each hydrological feature between the reference period and the hydrological prediction period,thus,the most similar sequence to the hydrological trend in the foreseeing period is matched in the historical hydrological time series,so as to achieve the purpose of hydrological trend prediction. In order to prove the efficiency and feasibility of the proposed method,it was verified with the Taihu hydrological time series data as the object. The analysis results show that the multi-variable hydrological time series trend similarity analysis based on machine learning can meet therequirements of dispatchers for the prediction effect of future hydrological trends.

Keywords　Machine learning,Multivariate time series,Hydrological trend prediction,Time series data mining,Similarity measure

1　研究背景

水文学是一个依托大数据驱动的研究领域,由于水文数据量大、维度高、类型复杂,使得仅用传统统计方法结合水文学知识加以分析的方案受到了计算能力、数据存储能力以及维度灾难等的制约,从而造成计算时间过长、精度过低。如何利用人工智能和数据挖掘技术从这些大数据中挖掘出水文时间序列中的信息是将计算机技术和水文学有机结合的一个重要突破点。本文主要研究基于机器学习的多元水文时间序列(Multivariate Time Series,MTS)相似性挖掘,其目的是现时刻相似水文过程发现以及未来水文趋势预测。

时间序列相似性度量是时间序列数据挖掘的重要分支,其中时间序列相似性度量又分为单特征的一元时间序列相似性度量和多特征的多元时间序列相似性度量。在一元时间序列相似性度量方面,比较典型的算法有欧氏距离法、基于斜率表示方法[1]、基于形态的度量方法[2]以及增量动态时间弯曲法[3]等,Bagnall 等[4]提出了 3 种算法用于寻找时间序列的相似过程,其中两种算法是基于两两比较的贪心算法,第三种是利用集质量的启发式测度直接找到相似序列集。在多元时间序列相似性度量方面,Li 等[5]提出了一种基于动态时间规整的相似度度量方法,该方法能在较低的计算成本下有效地度量多变量时间序列的相似性。相比传统的度量距离方法,Duchêne 等[6]针对不满足长度匹配或轴向拉伸的时间序列相似度量问题,提出了一种基于最长公共子序列的度量法。

基金项目:青年人才托举工程(2019QNRC001);中国水利水电科学研究院基本科研业务费专项(WR0145B012020)
This work was supported by the Young Elite Scientists Sponsorship Program by the CAST (2019QNRC001)and Fundamental Research Funds of China Institute of Water Resources and Hydropower Research (WR0145B012020).
通信作者:王超(wangchao@iwhr.com)

Shen 等[7]提出一种结合最大边际近邻分析(LMNN)与动态时间扭曲(DTW)的多元时间序列相似性度量模型,首先对多元时间序列采用基于马氏距离的 DTW 测量方法,通过马氏距离矩阵来衡量变量之间的关系,其次向 LMNN 算法模型输入马氏矩阵,通过迭代最小化损失函数训练模型。Li 等[8]在对多元水文时间序列相似性度量的研究中,首先采用一元时间序列度量算法计算出多元时间序列中各单特征的相似度,再利用 BORDA 计数法综合各单序列相似度量的结果得到多元水文时间序列相似性度量。该模型能够有效地将高维度时间序列相似度量转化为计算一元时间序列相似性度量,有效地降低了数据计算的复杂度。

对于时间序列相似性数据挖掘,相同的数据应用到不同的算法模型时,挖掘效果可能会有较大的差别,能否找到与数据匹配最佳的挖掘算法,对相似性挖掘模型有着关键的影响。本文针对多元水文时间序列的相似性度量,提出一种基于机器学习的多元水文时间序列相似性数据挖掘方法,并以太湖流域的水位、降雨、流量等多特征水文时间序列数据相似性分析为例,验证了该方法的适用性与正确性。

2 基于机器学习的多元水文时间序列数据挖掘

2.1 多元时间序列相似性度量

2.1.1 一元时间序列相似性度量

给定两个一元时间序列 S_1 和 S_2 的相似性度量函数 $Dist(S_1,S_2)$,相似程度与相似性度量函数成反比关系,相似性度量函数越小,两时间序列就越相似。经典的相似性度量函数方法总体上被分为两种:在两个长度一致的时间序列之间进行相似性度量的方法称为锁步度量法(lock-step measures);反而言之,可将两个长度不一致的时间序列进行拉伸比较的方法为弹性度量法(elastic measures)。

欧氏距离(ED)是锁步度量法最经典的代表。采用欧氏距离作为相似性度量的优点是计算简单、适应性强以及具有趋势衡量效果等,缺点是当序列数据存在噪音的情况下,相似性度量受影响较大,以及不允许序列在时间轴上进行长度伸缩以及平移。

弹性度量法经典的代表是动态时间规整(Dynamic Time Warping,DTW)。两序列相似度的衡量在距离测度的基础上进行,DTW 还将数据的时间维度纳入计算范围,因此 DTW 相似度衡量能够支持序列在时间轴上的伸缩,对于相似性呈现出按照时间轴平移、伸缩或不连续的时间序列效果很好,其缺点是计算成本较高,计算结果打破了时间维度上的依次顺序,不利于在水文场景上的应用。

考虑到本文分析的时间序列数据在序列长度上是相等的,且水文预测要求数据不为变相时间序列等因素,本文采用欧氏距离作为单一序列的相似性度量法。

2.1.2 多元时间序列相似性度量

多元时间序列在水文领域广泛存在,对于水文大数据分析来说,如何从高维度的时间序列中挖掘出潜在的水文信息,是一个具有重要意义的方向,然而相似性挖掘在多元时间序列中的应用还有较大的空缺。多元时间序列相似性度量算法除了研究背景中介绍的几种方法外,主要常用的算法有多元时间序列降维算法,如 Karamitopoulos[9]利用主成分分析法(Principal Components Analysis,PCA)将高维数据降为低维数据,以及拆分为单特征的一元时间序列相似性度量两种。

利用数据降维的思想,把高维的时间序列数据降维成一维时间序列数据,再利用一维时间序列数据的相似性度量法计算序列之间的相似性,常用的线性降维算法除了 PCA 还有线性判别分析法(Linear Discriminant Analysis,LDA)和多维尺度变换法(Classical Multidimensional Scaling)等,考虑到水文特征之间常具有非线性性,常用的非线性降维法有 t-SNE[10]与 LargeVis[11],它们的算法思路基本相似,即认为在高维空间中相似的点,映射到低维空间中也是相似的,这一思想非常符合相似性度量的思想。

利用 BORDA 计数法可将多元时间序列相似性度量拆分为单列时间序列相似性度量和 BORDA 计数法两个步骤。BORDA 是一种排序投票表决法,投票人根据自己的想法给各位候选者排序,假如候选总人数为 N,如果候选者 C 在某张选票上排第 n 位,它就得 $N-n+1$ 分,将所有投票人给候选者 C 打的分数求和,结果即为 C 的累计总得分,总分最高的候选者即为获胜者。在拆分多元时间序列数据为单特征时间序列做相似性度量阶段,首先在多元时间序列数据中按照单个数据记录时间进行取样得到样本集,依次取样本集中的样本,再分别取样本中的单列序列,通过相似度量函数计算样本列与查询序列对应列的相似性。对于 BORDA 计数法选取综合相似度量阶段,样本集中每一个样本类似于一个候选者,多元时间序列中的每列类似于一个投票人,而每个投票人则是依据单列时间序列相似性度量阶段计算得出的相似度集进行排列,最后按照 BORDA 计数法计算每个样本的 BORDA 分数,总分最高的样本序列即是与查询序列最相似的序列。

2.2 基于相似度的未来水文趋势预测模型

水文趋势预测挖掘是水文时间序列数据挖掘 HYDM(Hydrological Data Mining)研究工作的一部分,水文时间序列趋势预测挖掘的主要目的是在历史水文时间序列数据中,挖掘出时间序列变化的趋势,并做出预测。例如可以通过现时刻的水情来预测未来时间段的水情,从而辅助防汛调度部门做出相应的防汛调度决策。

2.2.1 机器学习训练数据获取

本文引入机器学习算法对时间序列趋势预测进行数据挖掘。首先在多元时间序列数据中按照一定抽样间距时间长度进行取样得到样本集,每个时间序列样本由参照期和预见期组成,在样本集中依次取出一个样本作为目标样本,其余样本集作为参照样本集,按照单列时间序列相似性度量算法计算出目标样本和每个参照样本在参照期与预见期各列的相似度,以参照期各列相似度作为模型的输入数据、预见期各列相似度作为模型的输出数据,组成机器学习算法的训练数据集。

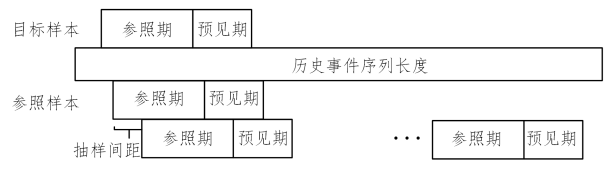


图 1 时间序列样本选择

Fig. 1 Time series sample selection

2.2.2 机器学习算法选择

机器学习中最突出、最常用的回归模型有:线性回归模型、基于深度学习的神经网络回归模型、支持向量机回归模型

和树模型。对于线性回归模型，其原理简单、计算时间短，但是其缺点是模型容易过拟合，即虽然可以完全适应训练数据，但是难以推广到新数据，因此为了防止过拟合，常需要在数据上多下功夫。对于神经网络模型，其在处理复杂的任务时，例如人脸识别等，表现效果很好，但是面对庞大的训练数据时，模型训练非常耗时，且其对数据预处理的要求高。对于支持向量机回归模型，其难以实现多元输出回归，且难以胜任大规模的训练样本，计算成本高。对于决策树模型，它是一种简单易用的非参数模型，适合作为数据量大的回归模型，计算速度较快，结果具有解释性，而且稳健性强，其对数据的要求低，是一种有着独特优势的机器学习算法。

树模型的典型代表算法有随机森林(Random Forest)、Gradient Boosting Decision Tree(GBDT)、eXtreme Gradient Boosting(XGBoost)等。本文使用 XGBoost 作为机器算法对水文趋势进行数据挖掘，XGBoost 是由 CHEN 等[12]提出的一个开源机器学习项目。XGBoost 的目标损失函数是由模型的训练误差和模型空间复杂度组成的，即 XGBoost 能够同时兼顾模型的泛化能力以及运行速度。同时，它是一种集成学习算法，即通过构建多个 CART 树对数据集进行预测，然后将多个树模型预测的结果集成起来，作为最终的预测结果。与 GBDT 算法一样，XGBoost 作为集成学习中的 Boosting 流派，其每一次的计算都是为了减少上一次的残差，进而在残差减少(负梯度)的方向上建立一个新的树模型，也就是说，前面决策树的训练和预测效果会影响建立下一棵树模型时的样本输入。但是，XGBoost 在高效实现 GBDT 算法的同时对其进行了算法和工程上的许多改进，与 GDBT 算法相比，XGBoost 使用了二阶的泰勒展开式逼近目标函数的泛化误差部分，有效简化了目标函数的计算；XGBoost 还可通过在目标函数中加入正则项来降低模型预测的波动性以及改善模型过拟合现象；XGBoost 还支持 GPU 并行运算，可节省大量计算成本。

2.2.3　未来水文趋势预测

XGBoost 回归模型训练好后，选取一段时间序列长度与参照期相同的当前时刻水文时间序列，按照与训练样本中单列时间序列相似性度量算法相同的计算方法，计算其与每个参照样本参照期各特征的相似度，并将基作为模型的预测输入，模型输出即为水文趋势预见期各水文特征与各历史时间序列数据各个水文特征的相似度，再利用 BORDA 计数法综合各特征的相似度，BORDA 分最高的历史样本序列即是与未来水文时间序列最相似的序列，我们即可认为历史样本序列所对应的水文趋势与未来水文趋势具有相似性，从而为调度决策提供了一定的参考。

3　实验验证与分析

本文以太湖水文数据为研究对象，验证所提出的基于机器学习的多元水文时间序列数据挖掘的有效性和可行性。

太湖水文趋势预报对太湖流域的防洪调度有着至关重要的作用，因此对太湖水文时间序列的相似性挖掘的研究非常重要。本文数据采用太湖日平均水位、太湖流域日平均降雨、常熟枢纽日平均流量、望亭立交日平均流量、太浦闸日平均流量这 5 个水文特征的逐日记录值，时间跨度从 2006 年 8 月至 2018 年 11 月，共 4458 条记录，选取 2006 年 8 月至 2018 年 9 月数据作为模型的训练数据，2018 年 10 月至 2018 年 11 月数据作为袋外测试数据。

考虑到降雨对洪峰的形成具有一定的时间积累以及太湖流域的产汇流特性，本文以时长为 15 天为参照期，以时长为 3 天为预见期，并将预见期的日平均降雨以降雨预报的形式考虑到参照期内，以及考虑到水位、流量等水文特征的即时性，取每个样本参照期内水文特征为预见期前一个时间段的太湖日平均水位、常熟枢纽日平均流量、望亭立交日平均流量、太浦闸日平均流量以及整个样本时间序列长度内的太湖流域日平均降雨，预见期内的水文特征取逐日太湖日平均水位以及预见期内常熟枢纽、望亭立交、太浦闸的平均流量。首先以欧氏距离算法为单列时间序列的相似性度量，通过训练数据计算出总训练数据集。在对模型进行训练时，学习曲线[13]是一种能够将模型的训练效果可视化的方法，通过绘制学习曲线，我们能够比较直观地了解模型过拟合或欠拟合的状态。学习曲线主要刻画的是模型的偏差和方差，模型的偏差即模型的预测值与真实值之间的差异度量，通常用来衡量模型的预测精度；模型的方差被定义为多次用大小相同、样本不同的训练集训练模型时，其每一次的预测结果与平均预测结果之间的误差，其衡量了模型受训练数据扰动的影响。常用的衡量模型偏差的指标有均方误差(Mean Square Errror)、平均绝对误差(Mean Absolute Error)、平均绝对百分比误差(Mean Absolute Percentage Error)和 R^2，本文定义 Score=R^2 衡量模型的偏差，其越接近于 1 代表模型的预测越精准，其计算方法为：

$$R^2=1-\frac{u}{v} \tag{1}$$

$$u=\sum_{i=1}^{N}(f_i-y_i)^2 \tag{2}$$

$$v=\sum_{i=1}^{N}(y_i-\bar{y})^2 \tag{3}$$

其中，N 是样本数量；i 是每个数据样本；f_i 是模型回归出的数值；y_i 是样本点 i 实际的数值标签；$\bar{y}$ 是真实数值标签的平均数。

模型的学习曲线计算步骤为：

1)从总训数据集中按照从 10%到 100%的等差数列比例取得大小不同的 5 份数据，将其作为每次的训练数据集。

2)对于大小不同的训练数据集，将训练集再次按照 20%的测试集、80%的训练集随机划分，用训练集训练好模型后，将训练集与测试集输入模型，分别得出训练集 Score 与测试集 Score，重复执行 5 次。计算出 5 组训练集、测试集的 Score 后，分别对其求平均值与方差得到模型的偏差与方差。

表 1　方差与偏差对模型的影响

Table 1　Influence of variance and deviation on model

	Score 大	Score 小
方差大	模型过拟合，对某些数据集预测准确，对某些数据集预测糟糕	模型不适合此数据，更换模型
方差小	模型泛化误差小，模型效果好	模型欠拟合，预测稳定，但对所有数据预测都不太准确

图 2 展示的模型学习曲线中，红色线条表示训练数据集对于模型的 Score，红色阴影部分表示训练数据集对应的方差，绿色线条表示测试数据集对于模型的 Score，绿色阴影部分表示测试数据集对应的方差。可以看出，随着样本数量的增加，模型对于训练数据的方差降低，预测精度有所下降，说明模型对于训练数据集的过拟合程度在逐渐降低；对于预测

数据集，模型的预测精度不断提升，方差不断减小，说明模型的泛化误差在不断降低。最终两条线趋于平行达到平衡，由模型在预测数据集上的表现可看出，最终模型的预测精度趋于85%。

模型训练好后，为了直观地展示模型的真实效果，从未接触过模型的袋外测试数据中随机选取目标样本，将其输入回归模型中计算出各历史序列样本与其对应的预见期各水文特征的相似度；最后利用 BORDA 计数法综合特征相似度选出最相似序列，实现多元水文时间序列相似性数据挖掘。

在测试数据中随机选取 3 个目标样本，通过基于机器学习的多元水文时间序列相似性数据挖掘，在历史数据中找出与其预见期水文趋势最相似的相似序列。图 3 给出了 3 例预见期水文趋势各特征与相应的相似序列特征的对比可视化，其中图 3(a-1)、图 3(b-1)、图 3(c-1)表示太湖水位的相似程度，图 3(a-2)、图 3(b-2)、图 3(c-2)表示常熟枢纽、望亭立交、太浦闸的流量相似程度，即右图中各线条越水平，模型所找出来的历史序列与输入状态越接近，通过观察可以发现真实水文趋势与所查找出的相似序列水文趋势较为接近，因此利用 XGBOOST 机器学习算法对多元时间水文序列进行相似性数据挖掘的方法是可行的。

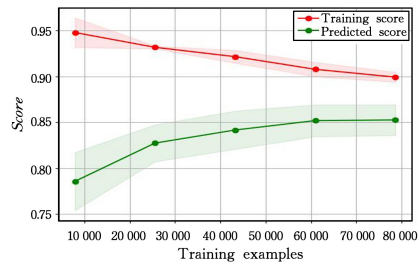
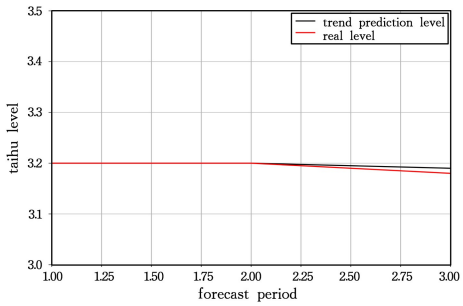
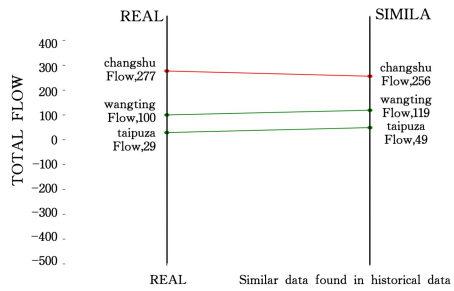


图 2 XGBoost 模型的学习曲线(电子版为彩色)

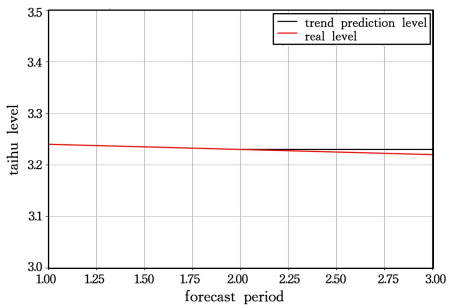
Fig. 2 Learning curve of XGBoost model



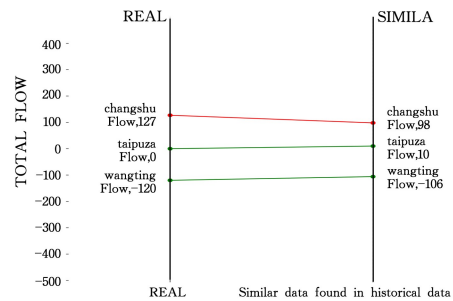
(a-1)



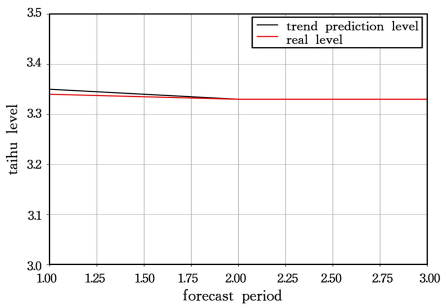
(a-2)



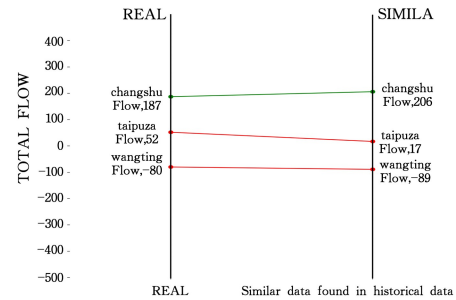
(b-1)



(b-2)



(c-1)



(c-2)

图 3 水文趋势特征相似性对比

Fig. 3 Similarity comparison of hydrological trend characteristics

结束语 针对多元时间序列的相似性数据挖掘，本文提出了一种基于 XGBoost 机器学习算法的多元时间序列相似性度量，通过建立参照期与预见期的水文趋势相似度之间的映射关系，来实现利用现有的水文时间序列数据找出历史上与之最相似的水文场景，达到预测预见期水文趋势的目的，并以太湖流域日平均降雨、太湖日平均水位、常熟枢纽日平均流量、望亭立交日平均流量、太浦闸日平均流量为数据源训练模型，通过交叉验证得出模型在测试数据集上的预测精度为85%左右，最后通过可视化水文特征输出验证了该方法的可行性。该方法实现了机器学习与多元水文相似度量的有机结合，通过匹配未来水文趋势与历史最相似序列，来作为未来水文趋势的强有力参照，更据说服力。

目前，利用人工智能技术在多元水文时间序列上的数据挖掘研究还不多，如何提高算法的准确率和效率以及制定更

加符合水文领域的相似性度量将是未来的研究重点。

参 考 文 献

[1] ZHANG J Y，PAN Q，ZHANG P，et al. Time series similarity measurement method based on slope representation [J]. Pattern Recognition and Artificial Intelligence，2007，20(2)：271-274.

[2] DONG X L，GU C K，WANG Z G. Research on morphology-based time series similarity measurement [J]. Journal of Electronics and Information Technology，2007，29 (5)：1228-1231.

[3] LI H L，YANG L B. Time series similarity measurement method based on incremental dynamic time warping [J]. Computer Science，2013，40(4)：227-230.

[4] BAGNALL A，HILLS J，LINES J. Finding Motif Sets in Time Series[J]. Bmc Public Health，2014，12(1)：1-11.

[5] LI Z X，LI K W，WU H S. Similarity measure for multivariate time series based on dynamic time warping[C]// The 2016 International Conference. 2016.

[6] DUCHNE F，GARBAY C，RIALLE V. Similarity measure for heterogeneous multivariate time-series[C]// Proceeding of the 12th European Signal Processing Conference. 2004：7-1.

[7] SHEN J Y，HUANG W P，ZHU D Y，et al. A Novel Similarity Measure Model for Multivariate Time Series Based on LMNN and DTW[J]. Neural Processing Letters，2017，45(3)：925-937.

[8] LI S J，ZHU Y L，ZHANG X H，et al. Similarity analysis of multiple hydrological time series based on BORDA counting method [J]. Journal of Hydraulic Engineering，2009，40(3)：378-384.

[9] KARAMITOPOULOS L，EVANGELIDIS G，DERVOS D. PCA-based Time Series Similarity Search[C]// Data Mining. 2010：255-276.

[10] VAN DER MAATEN L. Accelerating t-SNE using tree-based algorithms[J]. The Journal of Machine Learning Research，2014，15(1)：3221-3245.

[11] TANG J，LIU J Z，ZHANG M，et al. Visualizing Large-scale and High-dimensional Data[C]// International Conference on World Wide Web. 2016：287-297.

[12] CHEN T，GUESTRIN C. XGBoost：A Scalable Tree Boosting System[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[13] MURATA N，YOSHIZAWA S，AMARI S. Learning curves，model selection and complexity of neural networks[C]// Proceedings of the 1992 Conference. San Mateo. CA. Morgan Kaufmann，1993：607-614.

DING Wu，born in 1996，postgraduate. His main research interests include hydrological big data analysis and optimized operation of hydropower.

WANG Chao，born in 1989，Ph.D，senior engineer. His main interests research include basin water resources scheduling and intelligent water conservancy.

（上接第 458 页）

[30] BAI B G，ZHU H L，FAN Q X. Research on Early Warning of Dairy Product Quality and Safety Risk Based on GeneticOptimization BP Neural Network[J/OL]. Food Science. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=SPKX2020032000O&v=17WwU59A5kA%25mmd2FsWQldVPlWn%25mmd2FoewnrOzprziVfNRH9%25mmd2FVKtFqM2kjlkDOesG4Rrkydj.

[31] LI Y F，LI K W，PAN Y T，et al. A Dynamic Fusion Algorithm of Path Planning Based on Genetic andAnt Colony for Ground Autonomous Combat Robot[J]. Journal of Gun Launch & Control，2019(4)：42-46，50.

[32] LIU J W，CHANG Z G，DENG H B，et al. Energy-saving operation model for urban rail train based onimproved genetic algorithm[J]. Journal of Railway Science and Engineering，2019 (11)：2881-2888.

[33] CHEN Z X，DONG R X，HAO Y N. Modeling and Optimization of Picking Location Allocation in AutomaticPicking System Based on Improved Genetic Algorithm[J]. Industrial Engineering Journal，2019(6)：40-44，56.

[34] SHEN W S，ZHAO H C，SUI Y W. Sales Forecasting Model Based on BP Neural Network Optimized by Improved Genetic Algorithms[J]. Computer Systems & Applications，2019，(12)：200-204.

[35] MO T P，JIN H，SHI K，et al. The Fault Diagnosis of Analog Circuit Based on Wavelet Packet and SGD-XGBoost [J]. Microelectronics & Computer，2019(4)：38-42.

WANG Xiao-hui，born in 1998，undergraduate. His main research interests include machine learning and so on.

LI Jun-qing，born in 1984，postgraduate，associate professor. His main research interests include artificial intelligence and bigdata.