DOI: 10.19797/j.cnki.1000-0852.20200124

基于优化 DTW 算法的水文要素时间序列数据相似性分析

陈春华[1], 李 薇[2], 陈雅莉[1]

(1. 长江水利委员会水文局,湖北 武汉 430010;
2. 水利部信息中心(水利部水文水资源监测预报中心),北京 100053)

摘 要:本文在分析基于日平均水文要素时间序列数据相似性分析的基础上,提出基于水文要素摘录数据的相似性分析优点和存在问题,引入动态时间弯曲距离算法(DTW)解决水文要素摘录数据时序间距不一致,时常不一致的问题,并根据水文时间序列数据周期性起伏变化频繁的特征,基于最长公共子串,对 DTW 算法进行优化改进,从而提高相似性匹配精度。以汉口(武汉关)水文站洪水水文要素摘录水位过程数据为例,基于优化后的算法,进行相似性应用分析。应用分析表明,优化后的 DTW 算法能够较好的解决水文时间序列数据的相似性匹配问题,且提高了精度。

关键词:DTW;水文统计;时间序列;相似性

中图分类号:TV11;P333 **文献标识码**:A **文章编号**:1000-0852(2021)03-0044-05

0 引言

时间序列数据是由时间和对应数值元组组成的、按观测时间排列的有序集合,从统计意义上讲,时间序列数据是指某个指标在不同时间点上的数值根据时间先后顺序排列而成的数列, 由时间序列数据绘制的曲线称为过程曲线。水文现象具有时变现象,水文时间序列数据是在不同时间点上的水文要素监测值根据时间的先后顺序排列而成的数列, 其表征的水文知识包括气候及下垫面的演变过程及趋势。在防汛抗旱指挥和预报分析中经常要知道相似的水文过程,以便做出相应的防汛抗旱决策,就需要对水文降雨、水位、流量等过程线进行相似性分析,查找相似的水文过程。

1 水文时间序列数据相似性度量的相关研究

水文时间序列数据是由时刻和相应的水文要素监测值组成的、按观测时刻排列的水文数据有序集合。水文时间序列数据相似性就是比较基于水文监测数据绘成水文过程曲线的相似程度。

判断基于水文时间序列数据过程线相似度有许多较为直接有效的方法, 最常见的是通过比较两个时间序列的距离来衡量,距离越小,两个时间序列越相似,常用的时间序列相似性衡量指标是欧式距离。长期以来,相关领域研究人员对水文时间序列数据相似性度量方法进行大量研究,针对不同应用背景和应用场景提出许多分析方法。李薇等[1]通过引入数据仓库和数据挖掘的理论与技术,以漯河站、何口站汛期日平均降雨量相似性查询为例,为解决水文时间序列相似性研究提供了新的思路;朱跃龙等[2]以太湖流域大浦口站日平均水位数据为例,提出基于语义相似的水文时间序列相似性挖掘;杨艳林等[3]提出一种基于 DTW 聚类的水文时间序列相似性挖掘方法,通过凝聚层次聚类法分别对单一语义的符号集进行聚类,根据聚类结果实现时间子序列的符号化,从而提高水文时间序列的语义相似性查找效率。

一般研究水文过程相似性都是基于整编的日平均水文要素数据,如日平均水位、流量、降雨量等,由于日平均水文数据时间间距一致,时长相等,一般为365d,采用欧式距离指标分析过程线相似性,具有很好的时刻对应性。可是一般水文要素的日平均水文数据不能代表该水文要素的实际变化过程时,而水文要素摘录数据可以完整的反映水文要素的实际变化过

收稿日期:2020-05-06
作者简介:陈春华(1981—),男,江苏海安人,硕士,高级工程师,主要从事水文信息方面的工作。E-mail:chench@cjh.com.cn

程。但水文要素摘录数据的时间间距不一致,序列数据时间点也不是一一对应的，目前主要依靠专家经验查找对比,由于数据量大、时间跨度大等原因,造成查找效率低且准确性不高。

下面以长江干流汉口(武汉关)水文站 1997 年和 2018 年的水位监测数据为例,说明在进行水文过程相似性分析过程中时间间距不一致，序列数据时间点不是一一对应的情况。

图 1 为长江干流汉口(武汉关)水文站 1997 年和 2018 年日平均水位过程线。两条过程线数据时间间距为 1d,两个时间序列数据的时序数都为 365,且时间点一一对应,直接计算两条曲线 365 个时序点距离,就可以直接基于各类距离算法计算两条曲线的相似度。

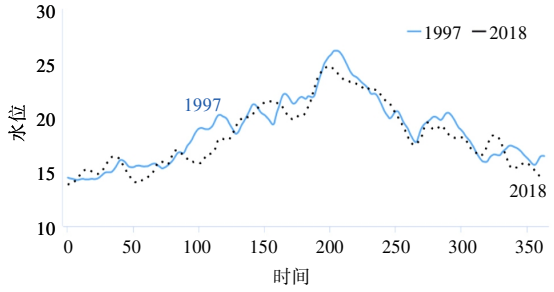

图 1　日平均水位过程线

Fig.1 The daily average water level process line

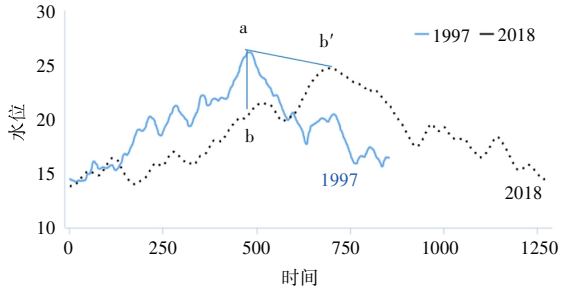

图 2　洪水水文要素摘录水位过程线

Fig.2 The process line of water level extracted from flood hydrological elements

图 2 为汉口(武汉关)水文站 1997 年与 2018 年洪水水文要素摘录水位过程线，两条过程曲线具有一定的相似性,从语义上分析,在曲线峰谷数一致,区域单调性相似,但数据长度也不一样,序列数据时间点不是一一对应的,2018 年 1278 个数据,1997 年 891 个数据，由于摘录数据的特征，其时间间距也不一样,有 1h,2h,4h 等等,但在时间轴上不是对齐的。在衡量曲线距离时，欧氏距离中上面曲线的 a 点会对应于下面曲线的 b 点,而实际应该对应到 b'点才是正确的。显然,这样传统的通过比较距离来计算两条曲线的相似性是不合理的,就需要将其中一个(或者两个)序列在时间轴下适当的“扭曲”[4],以达到更好的对齐,从而适合欧氏距离指标计算。

本文引入动态时间弯曲距离算法(Dynamic Time Warping,DTW),对时间序列数据进行规整(延伸、缩短)使得两个序列的时刻对应在逻辑上尽可能一致。动态时间弯曲距离算法(DTW)本质上是通过动态规划来计算这两个序列的相似距离,是实现序列在时间轴下“扭曲”的一种有效方法[4],最早应用于语音识别领域[5],可以弥补欧氏距离的只能计算时间间距一致、数据点数一致的缺陷。DTW 通过把时间序列进行延伸和缩短,来解决两个时间序列中各个点对应时间的迟滞问题,它不考虑时间序列的长度,通过时间规整来计算两个一一对应的时间序列性之间的相似性。

2　基于 DTW 的水文要素时间序列数据相似性度量方法

动态时间弯曲距离算法 DTW 是一个典型的优化问题,它用满足一定条件的时间规整函数描述测试模板和参考模板的时间对应关系,求解两模板匹配时累计距离最小所对应的规整函数。

2.1　DTW 算法的定义

假设给定连续时间序列 $X=(x_1,x_2,\cdots,x_n)$ 和连续时间序列 $y=(y_1,y_2,\cdots,y_m)$,函数 $d(i,j)=f(x_i,y_i)\geqslant 0$ 为序列中点到点的距离函数，一般使用欧式距离公式,构建序列 X 和序列 Y 的距离矩阵 D。

基于构建的距离矩阵 D，找到一条从左上角到右下角的路径,使得路径经过的元素值之和最小,即求解扭曲曲线 $\varphi(t)=(\varphi_x(t),\varphi_y(t)),\varphi_x(t)\in[1{:}n],\varphi_y(t)\in[1{:}m],t\in[1{:}k]$。也就是说,求出 k 个从 X 序列中点到 Y 序列中点的对应关系,若 $\varphi_x(t)=(1,1)$,就称 X 序列的第一个点与 Y 序列的第一个点是一个对应。

给定了 $\varphi(t)$,我们可以求解两个时间序列的累积距离 D_φ,如公式(1),最终找到一个最合适的扭曲曲线 $\varphi(t)$,使得累积距离最小,如式(2)所示。

$$D_\varphi(X,Y)=\sum_{t=1}^{k} d(\varphi_x(t),\varphi_y(t)) \tag{1}$$

$$DTW(X,Y)=\min D_\varphi(X,Y) \tag{2}$$

2.2　算法设计

DTW 距离允许序列点自我复制后再进行对齐匹配,能够很好地支持时间轴弯曲,并且它可以对非等长时间序列进行度量,也支持时间轴伸缩[6]。实际计算过

程中,采用“准对称步模式”基于距离矩阵 D,生成损失矩阵 M,步模式一定程度上涵盖了不同的约束。M[i][j]=min(M[i−1][j],M[i][j−1],M[i−1][j−1])+D[i][j],则损失矩阵的最后一行最后一列的值即为最小累积距离。

根据定义,结合约束条件,DTW 的算法设计如式(3)所示。

$$M[i][j]=\begin{cases}(x[0]-y[0])^2 & i=0,j=0\\ (x[0]-y[0])^2+M[0][j-1] & i=0\\ (x[0]-y[0])^2+M[i-1][0] & j=0\\ (x[0]-y[0])^2+\min(M[i-1][j],M[i][j-1], & \\ \quad M[i-1][j-1]) & i,j>0\end{cases} \tag{3}$$

所以 M[len(X)−1][len(Y)−1]就是 X 序列和 Y 序列相似距离 dist(X,Y),再把相似距离通过 $\omega(X,Y)=\frac{1}{1+\mathrm{dist}(X,Y)}$转化为相似度,这也就是 DTW 的算法核心。

2.3 算法应用

选取长江干流汉口(武汉关)水文站 2018 年洪水水文要素摘录水位摘录过程数据作为匹配对象,采用 DTW 算法,在历年(1865~2017 年)数据中,查找与 2018 年最相似的汛期水位摘录变化过程。

针对匹配的两个序列数据,建立距离矩阵 D,从左上角到右下角找一条路径,使得路径经过的元素值之和最小,采取步模式算法,计算过程中各参数及结果数据如表 1(选取相似度最高的 3 个数据成果)。

表1 计算过程各参数及结果

Table1 The parameters and results of the calculation process

年份	时间序列数	峰数	谷数	*dist*	ω
2018	1278	27	27	–	–
2014	1333	25	26	14.24	0.066
1978	415	19	20	16.95	0.056
1965	978	16	17	17.59	0.054

3 基于最长公共子串的 DTW 算法优化

在基于 DTW 算法的水文时间序列数据相似性匹配计算中,同步分析了曲线峰谷的语义特征,发现水文时间序列数据具有周期性起伏变化频繁特征,曲线单调性推进速率不一致,在步模式计算推进过程中,路径方向只能维持短暂的单调性,调整频繁,使得在计算曲线距离时,容易出现波峰波谷走势不一致的“病态匹配”[7]现象。

以图 3 中画圈位置为例,由于在此处单点距离过大,拉高了区域距离平均距离,在过程线匹配过程中出现了“病态匹配”,如果曲线通过适当迁移或后移,即可以满足区域走势相似,从而消除“病态匹配”对相似度的影响。基于水文时间序列数据具有周期性起伏变化频繁特征,两个数值序列的最长公共子串的长度对相似度的影响很大,即两个数值序列的最长公共子串越长,偏差越小,需要调整的幅度越小。所以,我们定义一个系数,称之为惩罚系数,利用惩罚系数对两个水文时间序列的距离进行调整,用惩罚系数乘以距离得到新的距离。具体算法如下:

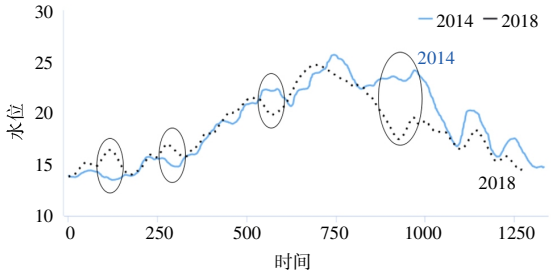

图 3 洪水水文要素摘录水位摘录过程线

Fig.3 Process line of water level extracted from flood hydrological elements

(1)计算最大标准差 $sd_{\max}$。设 x 表示数值序列 X 的平均数,n 表示数值序列 X 的数量,X 标准差 sd 的计算,如式(4)~(5)所示。

$$sd_X=\sqrt{\frac{\sum_{i=1}^{n}(x_i-x)^2}{n}} \tag{4}$$

同理:

$$sd_Y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y)^2}{n}} \tag{5}$$

所以,最大标准差取两个时间序列数据标准差较大的一个,如式(6)所示。

$$sd_{\max}=\max(sd_X,sd_Y) \tag{6}$$

(2)求解最长公共子串及其长度 l。因为 X 和 Y 是数值序列,在求最长公共子串时,将最大标准差设置为偏移容忍。也就是说,两个数值在这个标准差内,认为也是公共子串中的一部分。

设序列 X 的长度为 a,设序列 Y 的长度为 b。根据公式(7)定义矩阵 dp[i][j](0≤i<a,0≤j<b),其中:

$$dp[i][j]=\begin{cases}0 & X[i]-Y[j]\geqslant sd_{\max}\\ 1 & X[i]-Y[j]<sd_{\max}\ i\|j=0\\ dp[i-1][j-1]+1 & X[i]-Y[j]<sd_{\max}\ i,j>0\end{cases} \tag{7}$$

根据动态规划原理,可以求出 dp[i][j]从左上角,到右下角的最优路径,从而得到最长公共子串及其长度 l。

(3)计算惩罚系数 α,如式(8)所示。

$$\alpha=1-\frac{1}{\mathrm{len}(X)\times\mathrm{len}(Y)} \tag{8}$$

(4)按式(9)、(10)进行距离算法优化。

$$\mathrm{dist}_\alpha(X,Y)=\alpha\times\mathrm{dist}(X,Y) \tag{9}$$

$$\omega_\alpha(X,Y)=\frac{1}{1+\mathrm{dist}_\alpha(X,Y)} \tag{10}$$

4　应用与分析

应用基于最长公共子串的 DTW 改进算法，分析当前水位过程与历史上哪一时期的同类过程类似,对洪水水文要素摘录水位摘录数据水位过程线进行相似性分析,查找相似水位过程。选取长江干流汉口(武汉关)水文站 2018 年洪水水文要素摘录水位摘录过程数据作为匹配对象,在历年(1865~2017 年)数据中,查找与 2018 年最相似的水位摘录变化过程。

4.1　环境与流程

采用 Sqlserver2008 作为数据库,存储长江干流汉口(武汉关)水文站历年(1865~2017 年)洪水水文要素摘录水位摘录数据,采用 MyEclipse10 作为开发平台,利用 Java 程序设计语言编程。如图 4 所示,程序设计流程为:(1)获取 2018 年和历年(1865 年~2017 年)洪水水文要素摘录水位摘录数据;(2)以 2018 年为匹配对象计算每两年序列数据之间的最小距离和相似度;(3)计算两年序列数据的最大标准差;(4)计算两年序列数据最长公共子串及其长度;(5)计算两年序列数据的优化惩罚系数;(6)优化计算两年序列数据之间的最小距离和相似度;(7)循环(2)~(6);(8)采用冒泡排序法对计算成果排序,输出结果。

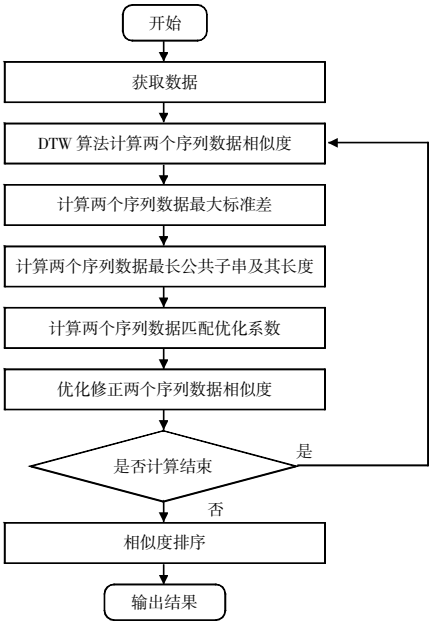

图 4　基于最长公共子串的 DTW 改进算法流程

Fig.4 Flow chart of DTW algorithm based on longest common substring

4.2　结果与分析

采用基于最长公共子串的 DTW 改进算法，长江干流汉口(武汉关)水文站 2018 年洪水水文要素摘录水位摘录过程数据作为匹配对象,在历年数据中进行曲线相似性匹配分析,计算过程中各参数及结果数据如表 2(选取相似度最高的 3 个数据成果)。

表2　基于优化DTW算法的计算过程各参数及结果

Table2 The parameters and results of the calculation process based on the optimized DTW algorithm

年份	时间序列数	峰数	谷数	sd_{max}	α	$dist_\alpha$	ω_α
2018	1278	27	27	–	–	–	–
2013	1434	27	26	3.27	0.30	7.14	0.123
2014	1333	25	26	3.68	0.54	7.63	0.116
2012	1281	23	24	3.76	0.50	9.72	0.093

通过计算结果,可以看出,采用基于最长公共子串的 DTW 改进算法，提高了过程线距离计算精度,2013 年与 2018 年洪水水文要素摘录水位摘录数据水位过程线距离最小,通过改进的 DTW 算法计算出来,$dist_\alpha$=7.14,相似度最高。

分别绘制 2012~2014 年洪水水文要素摘录水位过程线与 2018 年洪水水文要素摘录水位过程线,其对比见图 5。

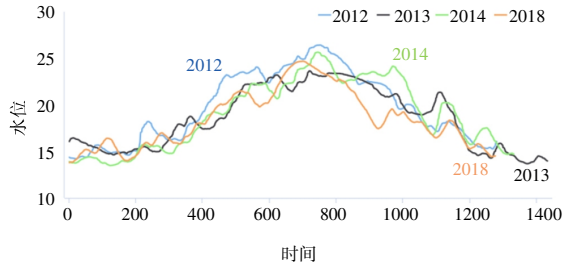

图 5　2018 年与 2012~2014 年洪水水文要素摘录水位过程线

Fig.5 The process line of water level extracted from flood hydrological elements in 2018 and 2012~2014

5　结论与展望

水文时间序列数据的相似性挖掘在水文研究领域具有重要意义,尤其是对洪水预报、防洪调度等方面的支撑作用。本文在分析了基于日平均水文要素时间序列数据和基于水文要素摘录时间序列数据进行相

似性分析的不同,提出基于水文要素摘录数据,进行水文过程相似性分析更能反映水文过程实际,并通过基于最长公共子串匹配,对 DTW 优化进行算法,旨在解决周期性频繁起伏变化水文时间序列数据距离计算中的"病态匹配"问题,并通过实验验证了这一度量方法的有效性和可行性,提高了水文时间序列数据相似性挖掘的精度。但优化同时也增加算法的复杂性,下一步将重点研究如何通过大数据、云计算等方式提高海量历史水文数据中的相似性匹配计算效率和并行计算机能力。

参考资料:

[1] 李薇,孙洪林.水文时间序列相似性查询的分析与研究[J]. 水文,2009,29(6):76-80. (LI Wei, SUN Honglin. Analysis and study on hydrological time series similarity search [J]. Journal of China Hydrology, 2009,29(6):76-80. (in Chinese))

[2] 朱跃龙,王咏梅,万定生,等.基于语义相似的水文时间序列相似性挖掘[J]. 水文, 2011,31(1):35-40. (ZHU Yuelong, WANG Yongmei, WAN Dingsheng, et al. Similartymining of hydrological time series base on semantic similarity measures [J]. Journal of China Hydrology, 2011,31(1):35-40. (in Chinese))

[3] 杨艳林,叶枫,吕鑫,等. 一种基于 DTW 聚类的水文时间序列相似性挖掘方法[J]. 计算机科学, 2016,43(2):245-249. (YANG Yanling, YE Feng, LV xing, et al. A method for similarity mining of hydrological time series based on DTW clustering is presented [J]. Computer Science, 2016,43(2):245-249. (in Chinese))

[4] GerHard Z. 动态时间规整 (DTW)[EB/OL]. 2017-12-21. https://blog.csdn.net/guohao_zhang/article/details/78303879. (GerHard Z. Dynamic Time Warping [EB/OL]. 2017-12-21. https://blog.csdn.net/guohao_zhang/article/details/78303879. (in Chinese))

[5] 吴晓平,崔光照,路康. 基于 DTW 算法的语音识别系统实现[J]. 电子工程师,2004,30(7):17-19. (WU Xiaoping, CUI Guangzhao, Lu kang. Realization of speech recognition system based on DTW algorithm [J]. Electronics Engineer, 2004,30(7):17-19. (in Chinese))

[6] 李正欣,张凤鸣,李克武. 一种支持 DTW 距离的多元时间序列索引结构[J]. 软件学报,2014(3):560-575. (LI Zhengxin, ZHANG Fengming, LI Kewu. A multivariate time series index structure that supports DTW distance [J]. Journal of Software, 2014(3):560-575. (in Chinese))

[7] Chen Q, Hu G, Gu F, et al. Learning optional warping window size of DTW for time series classification [A]. International Conference on Information Science, 2012.

Similarity Analysis of Hydrological Time Series Based on Optimized DTW Algorithm

CHEN Chunhua[1], LI Wei[2], CHEN Yali[1]

(1. *Bureau of Hydrology, Changjiang Water Resources Commission, Wuhan* 430010, *China*;
2. *Information Center* (*Hydrological Monitor and Forecast Center*), *Ministry of Water Resources, Beijing* 100053, *China*)

Abstract: Based on the similarity analysis of time series data of daily mean hydrological elements, this paper puts forward the advantages and problems of similarity analysis based on extracted data of hydrological elements, and introduces the Dynamic Time Warping (DTW) distance algorithm so as to solve the problems of inconsistent time series interval and duration extracted data of hydrological elements. According to the characteristics of frequent periodic fluctuation of hydrological time series data, and based on the longest common substring, this paper reports a definite improvement of DTW algorithm, so as to promote the similarity matching precision. Taking the water level process data extracted from hydrological elements of Hankou (Wuhan Customs) hydrological station as an example, this paper analyzes similarity application based on the optimized algorithm, the result of which indicated that the optimized DTW algorithm was able to solve the similarity matching problem of hydrologic time series data to promote precision.

Key words: DTW (dynamic time warping); hydrologic statistics; time series; similarity

(上接第 62 页)

Abstract: This paper introduces the latest distributed hydrological model (HYPE) developed by the Swedish Meteorological and Hydrological Institute and applied it to Heihe River basin, Shuiyangjiang River basin and Da River basin in different climate regions of China. The Nash - Sutcliffe Eficiency (*NSE*) is selected as statistical index for precision evaluation. The HYPE model shows good simulation results in different climatic regions, where the monthly runoff simulation results are significantly better than the daily time scale simulation results. Consequently, it can promote the use of this model in other similar regions, afford relevant simulation data information for areas where the analysis date is insufficient and provide services for water resource management.

Key words: HYPE model; Heihe River basin; Shuiyangjiang River basin; Da River basin; runoff simulation