

基于嵌入式索引的水文时间序列预测模型

周金玉, 万定生, 肖 艳

(河海大学 计算机与信息学院, 江苏 南京 211100)

摘 要: 提出了基于嵌入式索引的相似性搜索和 LM 算法改进的双隐含层 BP 神经网络的水文时间序列预测模型. 利用相似性搜索从大量的历史数据中挖掘出相似的信息, 去除掉历史数据中的冗余和错误信息, 从而减少训练集的数量, 提高预测准确性.

关键词: 时间序列; 相似性分析; 水文预测; BP 神经网络

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2018)01-0118-06

Hydrological Time Series Forecasting Model Based on Embedded Index

ZHOU Jin-yu, WAN Ding-sheng, XIAO Yan

(College of Computer and Information, HoHai University, Nanjing 211100, China)

Abstract: In this paper, we propose a hydrological time series forecasting model based on embedded index similarity search and LM algorithm improved BP neural network. The similarity search is used to mine the similar information from a large number of historical data, which can eliminate the redundancy and error information in the historical data, so as to reduce the number of training sets and improve the prediction accuracy.

Key words: time series; similarity analysis; hydrological forecasting; Bp neural network

1 引言

水文现象随着时间而变化, 水文数据则是对这一变化过程的离散记录^[1]. 水文数据是典型的时间序列数据, 其中包含了大量信息, 并且能够反映很多有用的规律. 作为数据挖掘研究领域的重要分支, 时间序列数据挖掘^[2] (Time Series Data Mining, TS-DM) 是从时序数据中进行模式发现和知识挖掘的过程, 主要通过不同的数据挖掘技术^[3] 分析时间序列的局部特征从而挖掘出时间序列内在的规律. 从庞杂的水文数据中提取事先未知的但含有重要信息的内容或规律, 这将有助于分析水文数据以及研究水文预报. 文献^[4] 提出了基于 LM 算法^[5] 改进的双隐含层 BP 神经网络^[6] (LMDBP) 进行水文预测, 本文在此基础上引入嵌入式索引搜索^[7] 方法思想, 根

据水文数据的特点, 改进得到快速水文时间序列相似性搜索, 并将其与 LMDBP 模型结合得到一种新的预测模型.

2 模型框架

该模型主要分为三个模块: 快速相似性搜索、相关性分析和水位预测. 具体流程如图 1 所示.

(1) 快速相似性搜索模块: 为了在不降低相似性搜索精确度的前提下提高相似性搜索的速度, 基于 EBSM 算法提出基于嵌入式索引的快速水文时间序列相似性搜索算法 (HSS_EBSM). 该算法主要包括离线和在线过程, 在离线部分确定参考序列并且建立嵌入式索引, 在线过程主要对查询序列进行相似性搜索.

(2) 相关性分析: 将上一步得到的序列作为初始

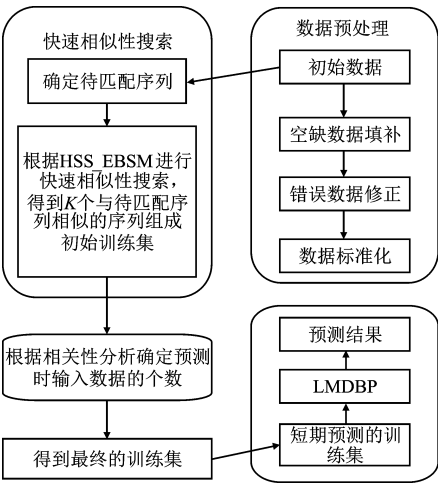


图 1 模型流程图

训练集,根据这些初始训练集利用相关性分析确定在预测时所使用的输入(特征)的个数,从而确定最终的训练集。

(3)水位预测:根据相关性分析所确定的最终的训练集,利用改进的双隐含层 BP 神经网络进行水位预测. 水位预测一般分为短期预测和实时预测,两者对预测的精度和预测所消耗的时间有不同的要求,该模型针对短期预测进行实验。

2.1 基于嵌入式索引的快速水文时间序列相似性搜索

嵌入式索引搜索方法通过把整个搜索过程的任务分解成两个过程,通过离线过程的嵌入索引和索引向量优化,提高了在线相似性查询的搜索速度. 但EBSM 算法存在一些缺陷,并且它不能很好地利用水文数据的周期性的特征. 因此,本文提出了符合水文时间序列特征的改进嵌入式索引搜索方法 HSS_EBSM. 其离线过程和在线过程的具体步骤如下所示。

- (1)离线过程:水位时间序列分割、聚类得到初始参考集,对初始参考集进行训练得到最终的参考集,根据最终的参考集对原时间序列建立嵌入索引.
- (2)在线过程:计算待匹配序列的索引,选出与待匹配序列相似的序列,得到预测将要使用的初始训练集.

首先需要对这些数据进行预处理:填补空缺数据、利用 3σ 准则(拉依达准则)^[8]找出错误数据、进行 Min-Max 标准化.

Min-Max 标准化转换函数如式(1)所示.

$$x' = \frac{x - \min}{\max - \min}$$

(1)

式中,max 是这些数据中的最大值 min 是这些数据中的最小值.

2.1.1 参考集生成

离线过程的第一步就是获取初始参考序列集,在 EBSM 中,初始参考序列集是随机选取的,这忽略了数据集本身的特点,为了能够获得更高质量的初始参考序列集,HSS_EBSM 使用先找到极小值点^[9]并且存储,每个极小值点作为分割的断点来分割原始时间序列,并用串行聚类来减少相似度高的重复序列的方法来得到初始的粗糙参考序列. 但在基于嵌入式索引的相似性方法中,参考集的选择是非常重要的,仅利用分割和分类算法选出的初始参考集是远远达不到要求的。

Venkateswaran 提出了参考序列选择的启发式规则^[10],第一个是最大方差规则,它能很好地指示在数据库中如何选择参考集,第二个是最大剪枝规则,它的意思是尽可能地减少参考集的维度。

HSS_EBSM 在遵从该启发式规则的基础上针对水文数据的特点提出了适合水文时间序列的参考集选择法。

首先将水文时间序列集中每个序列的方差定义为两部分:

- (1)集合中的某个序列与其他序列之间的距离数组的方差作为其量度的一部分;
- (2)该序列与其他序列之间波峰或者波谷的数目之差所组成的数组的方差作为其量度的一部分.

具体过程见算法 1,其中 SPs 是分割后得到的时间序列集, k 为将要选出的参考序列的个数.

虽然算法 1 获得的 k 个参考序列 refR 符合原始时间序列的分布,且它们相互之间具有显著差异,但是,算法 1 中得到的 k 个序列并没有通过训练集进行过训练,还不能完全符合查询序列的要求,因此还需要对参考序列集进行训练学习. 因而,本文还采用了另一种学习方法(算法 2)来获得最终的 d 个参考序列. 算法 2 定义了一个学习误差函数 $LE(Q)$,该函数表示当使用该参考集对查询序列 Q 进行查询时产生的误差. 其定义为欧氏向量空间搜索后得到的候选集与正确匹配位置不同的个数.

算法 1 HSS_MaxVariance (k , SPs)

输入: k : 需要选出的参考序列的个数

SPs:初始参考序列集

输出:refR:含有 k 个参考序列的集合

1. for ($i=1$ to k) do //按最大方差规则获取最后 k 个序列

```

2. index = 1, MaxVar = 0; //初始化下标和方差
3. for (each in SPs) do //选取当前 SPs 中具有最大方差的子序列,  $j$  代表当前的下标
4.     newMaxVar = calculateMaxVar (SPs ( $j$ )); //计算 SPs 中第  $j$  个序列的方差, SPs ( $j$ ) 表示第  $j$  个序列
5. if (newMaxVar > MaxVar) then
6.     MaxVar=newMaxVar, index= $j$ ;
7. end if //在 MaxVar 和 index 中存放最大方差的方差值和具有最大方差的序列标号
8.     end for
9. refR  $\leftarrow$  SPs (index); //在输出集合中加入当前具有最大方差的序列
10. 从 SPs 中删除 SPs (index);
11. end for //结束基于最大方差原则的筛选

```

2.1.2 欧氏向量空间的优化

在水文领域的相似性搜索过程中,重点关注的是相似度比较高的洪峰或者波谷. 查询序列大部分都是按周、旬、月、季、年这样的时间段^[11]. 所以在选择初始参考序列集时可以遵循了这样的经验,从波峰和波谷处选择,并且尽量包含一个洪峰或者一个干旱的过程.

对于索引后的向量空间,向量的个数与原始序列的长度是一致的,实验发现,对于相邻点的索引,它们的向量是比较相近的,随着原始序列长度的增加,索引空间的冗余度就会增加. 这将影响空间内向量的搜索效率. 本文提出了水文时间序列分段近似算法 (Hydrological Time Series Piecewise Aggregate Approximation, HTS_PAA) 来提高搜索的速度,该算法包括降维方法和抽样过程两个部分. HTS_PAA 方法借用了 PAA^[12] 方法的优点,其过程如下:不同于 PAA 的等长段划分和均值作平滑处理,在 HTS_PAA 方法中设置了一个阈值 φ , 对于一个向量 $\mathbf{V}$, 后续的向量只要与 $\mathbf{V}$ 的距离量度小于 φ 则加入当前段. 整个向量序列都按照这个原则划分成了 P 段. 假设之前长度为 N , 则压缩率为 P/N . 对于每一段的最终表示,首先选择与其他向量间距离方差较小的个体,将交叉运算作用到这些个体上,并且再次经过选择,交叉得到最终的向量表示. 经过 HTS_PAA 降维后,长度减少的向量序列成为了最后待搜索空间.

算法 2 HSS_ReferenceTraining(d , refR, X , Q)

输入: d : 最终参考集中序列的个数

refR: 算法 1 得到的参考集

X : 原始时间序列

Q : 训练用到的查询序列

输出: R : 最终的参考集

```

1.  $R = \{R_1, R_2, \dots, R_d\}$ ; // 将从 refR 中随机选择的  $d$  个序列加入  $R$  中
2. EmbeddingVector = Embedding( $R, X$ ); // 基于  $R$  计算序列  $X$  的嵌入索引从而生成索引向量空间
3. oldLE = ErrorCalculate ( $Q$ , EmbeddingVector); // 计算使用当前的参考序列集对查询序列  $Q$  训练时具有的总体误差
4. index = 1;
5. while(true) do
6. remainR = refR -  $R$ , exchange=0; // remainR 中存放 refR 中除了  $R$  中的序列集的其他序列集
7. for ( $i=1$  to |remainR|) do // 将 remainR 中的每个序列替换  $R$  中的序列
8. EmbeddingVector=Embedding( $R - R(\text{index}) + \text{remain}(i), X$ ); //  $R$  中的第 index 个序列被 remainR 中的第  $i$  个序列替换, 并计算时间序列  $X$  的嵌入索引
9. newLE=ErrorCalculate( $Q$ , EmbeddingVector); // 计算新的误差
10. if (newLE < oldLE) then
11.  $R(\text{index}) = \text{remain}(i)$ , oldLE=newLE;
12.     exchange=1; // 当有交换发生时改变 exchange 的值
13. end if
14. end for
15. index= (index mod  $d$ ) + 1;
16. if (exchange > 0) then
17. break; // 循环到每次替换  $R$  中的序列都不能减少误差 LE
18.     end if
19. end while

```

通过 φ 可以控制压缩程度和变换后序列的局部特征, φ 越大表示对数据的压缩程度越大, 失真会变大, 查询的结果准确度就会降低. 因此可以通过设置 φ 的值来控制搜索的准确度. 本文在实验的基础上设置 $\varphi=1.5$ (在向量空间中距离量度小于 1.5 的向量作为 HTS_PAA 中的一段进行降维).

2.2 相关性分析

预测时比较重视的是下一时刻的预测值以及之后几个时刻的趋势, 因此可以利用相关性分析^[13] 从训练集中找出与下一时刻的预测值比较相关的前几个时刻.

首先, 计算初始训练集中每条序列的最后一个时刻以及该序列对应的下一个时刻 (即待预测时刻) 的值之间的相关系数, 计算公式如式 (2) 所示.

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

式中, n 为初始参考集中序列的数目,代表这 n 条序列的最后一个对应的值; x_i 代表这 n 条序列的最后一个时刻对应的值; $\bar{x}$ 是这些最后时刻值的平均值; y_i 代表这 n 条序列对应的下一个时刻的值; $\bar{y}$ 是这些值的平均值; r 就是 x 与 y 之间的相关系数的值.

计算完最后一个时刻与待预测时刻对应的值之间的相关系数后,继续向前计算前一个时刻与待预测时刻对应的数据之间的相关性.当 $r \geq 0.9$ 时,将该时刻视为与待预测时刻相关,从最后一个时刻开始往前面若干个时刻进行计算,当遇到第一个时刻的 $r < 0.9$ 时停止寻找,得到的就是最终训练集的输入的数目.

2.3 短期预测模型

在水文领域,所谓短期预测是指利用大量的历史日平均水位值来预测后面三到七天的日平均水

位,主要的目的是观察这个测站的未来水位的趋势,但由于日平均水位本身就是对测站一天的水位的平均,本身就是一个不太精确的值,这会影响短期预测的效果,因此用于短期预测的模型必须有很高的预测精度,即使增加预测的时间消耗.

针对上述短期预测的特点以及需求,本文基于相似性搜索和 LMDBP 提出一个动态建模的预测模型.所谓的动态建模就是每次预测都会重新建一次预测模型然后预测,并不是单单基于历史数据所建的不变的模型,每次预测都会加入待预测日一日的新得到的日平均水位.这样每次预测所消耗的时间主要包括更新新加入的前一日的水位的索引所需的时间、建模所需的时间以及利用建好的模型预测所需的时间.短期预测的训练集构成如图 2 所示.

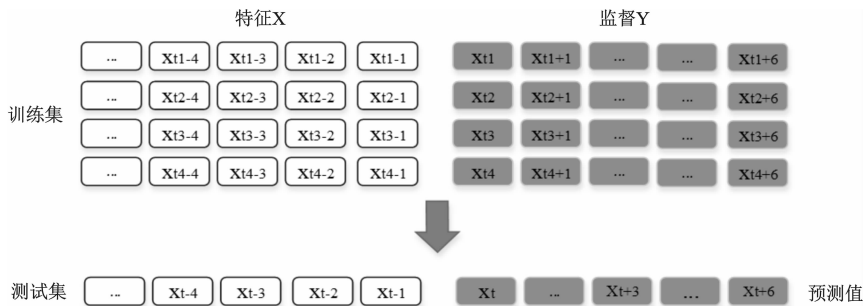


图 2 短期预测训练集结构

表 1 两个算法的短期预测结果及对应的预测精度的三个指标

日期	实际值	LMDBP		FSS_LMDBP	
		预测值	绝对误差	预测值	绝对误差
2010-9-20	17.44	17.795 69	0.355 69	17.449 31	0.009 31
2010-9-21	17.4	17.783 2	0.383 2	17.497 75	0.097 75
2010-9-22	17.46	17.764 37	0.304 37	17.546 77	0.086 77
2010-9-23	17.46	17.743 15	0.283 15	17.587 62	0.127 62
2010-9-24	17.22	17.727 31	0.507 31	17.579 91	0.359 91
2010-9-25	17.08	17.717 37	0.637 37	17.565 36	0.485 36
2010-9-26	17.03	17.708 93	0.678 93	17.545 56	0.515 56
MRE		0.026 102		0.014 011	
NMAE		0.026 014		0.013 875	
Q		100%		100%	

3 实验分析

本文使用数据是江西省外洲站从 1950 到 2010 年这 61 年间的日平均水位值,下面展示利用本文提出的 FSS_LMDBP 预测 2010 年 9 月 20 日至 26 日

的日平均水位的值,并与直接采用 LMDBP 算法的预测效果进行比较,比较过程主要从两个方面:预测精确度和消耗的时间.

由图 3 可以看出,两个算法都能将前 4 天的趋势很好的预测出来,尤其 FSS_LMDBP 算法预测的

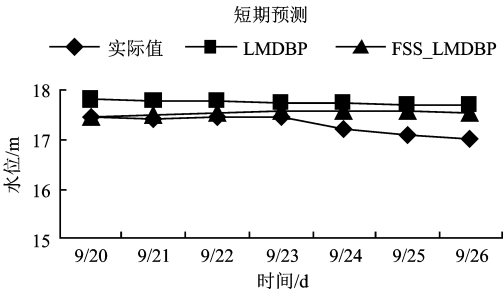


图 3 2010 年 9 月 20 日至 26 日短期预测结果

趋势更加贴合,后 3 天实际值有一个下降的趋势, FSS_LMDBP 预测出了下降的趋势,但幅度不明显, LMDBP 则没有预测出下降的趋势. 由表 1 可以看出 FSS_LMDBP 的 MRE 和 NMAE 都小于 LMDBP,意味着 FSS_LMDBP 的预测精度高于 LMDBP,同时两者的预测合格率都为 100%,这也证明了基于神经网络的水文时间预测的有效性.

由于再进行下一天的预测时,需要将前一天的真实值加入到历史数据中,从而达到动态建模的目的,但更换少量数据不会对历史参考集造成过大改变,因此不需要更新参考集,只需要计算新加入的数的索引,这样预测时间虽然有所增加,但不会有很大的影响,预测时间的对比如表 2 所示.

表 2 短期预测时间 s

模型	LMDBP	FSS_LMDBP
时间 t_1	22.11	7.02
时间 t_1	21.26	8.69

9 月 21 日到 9 月 27 日的水位的预测结果如图 4 所示. 从预测的趋势来看, FSS_LMDBP 优于 LMDBP,并且即使增加了计算前一天索引的时间,总的预测时间依旧有很大的优势.

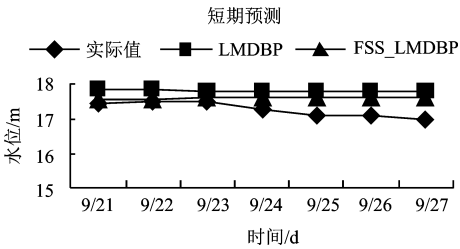


图 4 2010 年 9 月 21 日至 27 日短期预测结果

洪峰或干旱的水文趋势需要重点关注,本文使用循环连续的加入实际值预测出之后的若干天的水位值,并逐渐替换除了第一日以外的其他水位值,得到三个月的预测值,以此来分析该模型是否能够很

好地预测出洪峰或干旱的水文趋势. 图 5 所示为利用本文所提出的 FSS_LMDBP 与 LMDBP 模型预测的外洲站 2010 年 6 月 1 日至 8 月 31 日之间的日平均水位值.

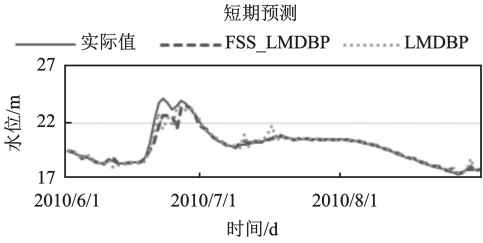


图 5 2010—06—01 至 08—31 日的短期预测结果

如图 5 所示, FSS_LMDBP 在水位平坦时,曲线的拟合程度较好, LMDBP 则出现了一小部分的尖峰,分析原因可能是受历史数据中的异常数据的影响,甚至有可能是出现了过拟合. 由于 FSS_LMDBP 先进行了相似性搜索,去除了大量的冗余信息以及异常信息,虽然在 8 月 25 日也出现了很明显的尖峰,但是其他曲线拟合程度还是很高的. 其在汛期的预测能力明显优于 LMDBP,尤其趋势预测得很好,只是波峰的值始终达不到实际水位值那样的高度. 综上所述,在趋势预测、预测精度和时间上 FSS_LMDBP 均有优势.

4 结束语

本文根据水文数据在短期预测中的特点提出了基于嵌入式索引的相似性搜索和 LM 算法改进的双隐含层 BP 神经网络的水文时间序列预测模型,从而实现了从大量的历史水文数据中挖掘出相似的有用的信息用于训练改进后的神经网络模型,以达到提高预测精度的目的. 并将其应用到水文时间序列预测中,证明了模型的可行性、高效性、准确性和稳定性.

参考文献:

[1] 艾萍,倪伟新. 我国水文数据挖掘技术研究的回顾与展望[J]. 计算机工程与应用, 2003, 40(28): 13-17.

[2] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surveys (CSUR), 2012, 45(1): 12.

[3] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2004, 5(4): 1-18.

[4] Wan D, Xiao Y, Zhang P, et al. Hydrological big data prediction based on similarity search and improved BP neural network[C]// IEEE International Congress on

Big Data. IEEE, 2015;343-350.

[5] Moré J J. The levenberg-marquardt algorithm: Implementation and theory[J]. Lecture Notes in Mathematics, 1977, 630:105-116.

[6] Lacher R C, Hruska S I, Kuncicky D C. Back-propagation learning in expert networks[J]. IEEE Transactions on Neural Networks, 1992, 3(1):62-72.

[7] Athitsos V, Papapetrou P, Potamias M, et al. Approximate embedding-based subsequence matching of time series [C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM,2008;365-378.

[8] 何世彪,敖仙丹,吴乐华. 3σ 准则在基于谱相关信号检测中的应用[J]. 通信技术, 2008, 41(6): 52-54.

[9] 刘永志,皮德常,贾学萍. 基于三点的时间序列关键点研究[J]. 微电子学与计算机, 2015,32(1):45-47.

[10] Xu ZS. On similarity measures of interval-valued intuitionistic fuzzy sets and their application to pattern recognitions[J]. Journal of Southeast University (English Edition) 2007; 23: 139-143.

[11] 李婕好,陈志军,闫学勤. LM-BP 算法在径流预测中的应用[J]. 人民黄河, 2015, 37(6):29-31.

[12] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge and information Systems, 2001, 3(3): 263-286.

[13] 刘峰,瞿俊. 基于聚类分析和神经网络的时间序列预测方法[J]. 微电子学与计算机, 2006, 23(9):85-87.

作者简介:

周金玉 女,(1995-),硕士研究生. 研究方向为数据挖掘. E-mail: pattyzhou@hhu. edu. cn.

万定生 男,(1963-),教授. 研究方向为信息处理与信息系统.

肖 艳 女,(1991-),硕士研究生. 研究方向为数据挖掘.

(上接第 117 页)

参考文献:

[1] Mentor Graphics Corporation. Mentor graphics verification academy cookbook[M/OL]. [2017-06-01]. <https://www.mentor.com/company/news/mentor-verification-academy-launches-coverage-cookbook>.

[2] Accellera. Universal Verification Methodology (UVM) 1.2 User's Guide[M/OL]. [2017-06-01]. <http://www.accellera.org/downloads/standards/uvm/>.

[3] 彭宏涛,朱德智. 基于 Farrow 滤波器的宽带数字波束形成技术研究及实现[J]. 舰船电子对抗, 2015, 38(2):37-41.

[4] 宋何娟,马强. 应用于数字阵列雷达的数字波束形成(DBF)的设计与实现[J]. 中国集成电路, 2010, 205(4):42-45.

[5] 张强, UVM 实战卷 1. 2014[M]. 北京:机械工业出版社, 2014.

[6] 成超,李会勇,何子述. 基于子阵时延的数字阵列宽带波束形成[J]. 雷达科学与技术, 2008, 6(6), 459-462.

作者简介:

邓庆勇 男,(1987-),博士,工程师. 研究方向为雷达系统、雷达信号处理、FPGA 测试. E-mail:dqy19870509@sina. com.

朱 鹏 男,(1987-),硕士,工程师. 研究方向为雷达系统、雷达信号处理、FPGA 测试.

习建博 男,(1986-),博士,工程师. 研究方向为雷达系统、雷达信号处理、FPGA 测试.