

数据挖掘软件在水文预报中的应用实例

胡 艳¹ 孙洪林²

(1.太湖流域管理局水利发展研究中心,上海 200434;2.水利部水利信息中心,北京 100053)

摘要 水文统计方法因其简单、易于实现的优势在现代水文预报中有着广泛的应用,目前国内外很多数据挖掘软件集合了一些新的统计算法模型,使得统计建模计算过程更加便捷。把神经网络模型运用于流域站点水位预测,采用实测数据验证了 Clementine 数据挖掘软件在水文行业的运用,结果表明其预测结果能反映站点水位变化趋势,预测值接近实测值。

关键词 数据挖掘 水文预报 神经网络

中图分类号:P338 文献标识码:B 文章编号:1006-764X(2007)S2-0164-03

1 水文预报工作的分类

水文预报方法很多,可以粗略地分为过程驱动模型方法和数据驱动模型方法两大类。过程驱动模型指以水文学概念为基础,对径流的产流过程与河道演进过程进行模拟,从而进行流量过程预报的数学模型,比如新安江模型、HBV 模型、TANK 水箱模型等,这类模型将一些有物理依据的公式与经验性的公式结合起来,模拟流域水文过程,对于在不同流域的应用,需要对模型进行改进。例如在降雨产流模拟中,有些地区采用一水源模型,而有些地区可能采用三水源模型,区别就在于不同的下垫面类型。水文预报中还有一种常见的方法就是数据驱动模型方法,它基本不考虑水文过程的物理机制,而是以建立输入、输出数据之间的最优数学关系为目标的黑箱子模型,亦即常说的水文统计学方法,其中以回归模型方法最为常用^[1]。近年来新的预测方法,如神经网络模型、时间序列分析模型等在水文预报中的应用也得到很快发展^[2]。

2 统计分析在水文预报中的运用及存在问题

由于水文现象本身的复杂性,相对来说用过程驱动模型模拟比较难,而统计学模型因其简单、易于实现等特点在水文预报上得到广泛应用。例如早在 20 世纪 80 年代,为了进行径流还原计算,就有人建立了基于统计学原理的流域降雨径流模型,并且发现年降水径流关系十分复杂,可能与年降水深、上年降水

深、日降水超过某定数的累计降水深、基流变化指标、地下水位变化指标、年平均气温等有关,最终通过回归分析方法找出能基本反映降雨径流关系的主要参数,以此径流计算结果作为径流还原计算的依据。

统计学水文模型就是依据大量历史资料,应用数理统计方法分析水文要素自身的统计规律或要素与有关因子之间的统计关系,然后应用这些规律或关系制作预报的模型。一般而言,统计学水文模型是利用人为选定的统计公式和算法,对于已定的因变量和自变量进行分析预测工作,因此不能挖掘未知的、潜在的关系;而且由于水文数据的特殊性,其在时间序列上可能会出现数据的缺失和存在偏差,需要对水文数据进行查补展延,利用统计学模型进行预报工作之前必须进行人工数据处理和准备工作,数据前期处理工作量大;再次随着水文数据的不断更新和充实,不能实现由于数据库不断更新变化而对相关参数的实时修正功能。

3 数据挖掘软件的发展

近年来,统计软件在电信、金融系统得到广泛的运用,在目前数据量越来越多,潜在关系越来越复杂的情况下,数据挖掘软件也因此得到快速发展。比如美国的 SAS(Statistical Analysis System)即统计分析系统,软件是目前世界上应用最广泛的数据分析软件之一,不仅具有强大的统计功能及较好的人机交互性,而且提供了多种友好的用户界面,广泛适用于农业科学、生物科学、经济学、社会科学及医疗卫生等

基金项目:引进国际先进农业科学技术创新项目(948 项目)

作者简介:胡艳(1980—),女,安徽绩溪人,助理工程师,硕士,从事水资源与水环境研究。E-mail:jyhty@tba.gov.cn

领域的数据分析。又如 SPSS 公司的 Clementine 数据挖掘软件,它是世界上最早的统计分析软件,它们的应用分布于通讯、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业,是世界上应用最广泛的专业统计软件。

数据挖掘软件的兴起要归结于其中的三大支柱:数据库技术、人工智能技术和数理统计技术。它能进行数据的选择提取、净化与处理以及数据转化,在确定了挖掘目标(概念描述、识别、关联、分类、回归、聚类、异常分析等)后,根据功能类型可以应用决策树、神经网络、覆盖正例排斥反例、概念树、遗传算法、公式发现、统计分析、模糊论等理论与技术,并在可视化技术的支持下,构造满足不同目的的数据挖掘应用系统^[3]。

数据挖掘软件集合了近年来在水文学中用到的新预测手段,比如说人工神经网络(ANN),是近 10 多年来最广为关注的一种非线性预报方法,已被广泛应用于实时中长期水文预报中,实现了传统水文统计学方法与非线性科学方法的结合^[4],而且数据挖掘软件中对于数据的准备便捷,与所有主流数据库实现了无缝连接,可将预报中用到的多个数据源数据(比如实时报讯水雨情库和遥测库等)结合起来,能避免数据冗余和不一致;由于水文数据的特殊性,其在时间序列上可能会出现数据的缺失和存在偏差,数据挖掘软件提供了有效的差补和剔除方法,改进数据质量,提高预测预报的精度;数据挖掘软件还能定制任务,减少水文预测工作因年际变化造成的影响。

4 Clementine 数据挖掘软件在水利行业的应用实例

SPSS 公司的 Clementine 数据挖掘软件为用户提供功能强大、易用的数据挖掘工具平台。它包括 6 个节点区:源数据节点(Sources)、记录处理节点(Record ops)、字段(变量)处理节点(Field ops)、图形节点(Graphs)、建立模型节点(Modeling)和输出节点(Output),用户建立模型的过程就是把各个节点区的节点以连线的方式连在一起的过程。Clementine 数据挖掘软件因其功能强大的数据挖掘算法可以快速建立预测性模型,实现产品的风险预警、风险处理和监测,以及通过数据中隐含的模式来预测潜在的联系等。

近年来,数据挖掘在水文行业应用广泛,主要是预测式挖掘,通过模型对新数据集的行为做出预测。水文数据挖掘主要是对时序数据的挖掘,需要与水文领域数据分析处理技术紧密结合,同时也需要充分应用计算机软件技术和数据库、数据仓库及其他数据处理新技术,按其功能细分为相似搜索、序列模

式挖掘和周期分析等^[5]。
本文通过考虑流域控制性站点水文要素的相关性分析来说明 Clementine 数据挖掘软件的适用性。在流域防汛调度和水资源管理中,充分利用控制站与相关站的关系,来预测流域控制站的相应值,将比较简单易行。

琳桥站是太湖流域的防汛调度控制站,也是国家防汛抗旱的重要控制站点。通过多年的调度实践,可以大概判断在平原河网地区,测站水位的主要影响因素是地区降水和相邻测站的水位影响,但它们之间的影响到底有多大,或者说影响权重有多大还没有定论。考虑到相关性统计分析中参数会受到时段的影响,通常在一段时间后会进行参数的调整,因此本次采用最近的 2000 ~ 2004 年流域水文报讯资料,对琳桥站的水位进行相关性分析,并对琳桥站 2005 年的水位进行预测。

4.1 问题的描述

在利用数据挖掘软件进行站点水位预测之前,需要进行如下准备工作:预测站点的水位数据相关性分析(对相关站点的水位、降水因素的影响进行预测),相关站点的数据资料,包括观测站的水位数据,以及与观测站相关站点的水位和降雨数据。

通过表 1 可建立琳桥站水位函数关系如下:

$$Z_{\text{预测}} = F(Z_1, Z_2, P_1, P_2, \dots, P_n)$$

表 1 相关站点关系

相关站点	相关因素	记录数 (8 时)	相关站点	相关因素	记录数 (8 时)
陈 墅	水位	1575	江 阴	降雨	381
湘 城	水位	1841	青 阳	降雨	233
张家港闸	降雨	165	白茆山	降雨	231
十一圩港闸	降雨	527	无 锡	降雨	414
望虞闸	降雨	409	陈 墅	降雨	398

4.2 报讯站水位预测

根据 Clementine 数据挖掘软件的要求,其神经网络模型对于进行时间序列预测的应用占较大优势,故本次水位预测工作采用神经网络模型。其处理流程分以下几个步骤(图 1):

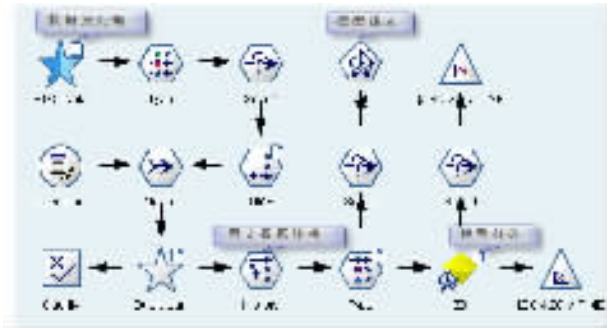


图 1 对观测站的水位预测数据处理流程

a. 数据预处理过程。从关系表中抽取与预测观测站相关的站点的站码,然后对于不同类型的数据分别抽取对应的历史数据(包括相关站点水位、相关站点降雨、相关站点流量),接着根据预测观测站的站码(为关键字段),将多个表中的数据进行融合汇总,形成统一的汇总数据,同时进行一些其他必要的数据库整理和转换工作,例如数字到字符型的转换、日期变换、数据筛选等,这部分工作需要根据具体的数据库情况进行设定。

b. 数据插补过程。由于实时报讯时经常出现数据不完整的情况,故在数据预处理时还要进行数据插补,为系统处理方便、简单,对水位、降雨数据采取以下补缺规则:降雨数据若没有则当作没有降雨处理,以0补缺;水位数据根据前后两天的平均值插补。

c. 模型建立过程。在数据预处理的基础上,利用 Clementine 中的历史节点,将数据构成历史数据序列,然后选择神经网络模型进行模型训练。

d. 模型评估过程。对建立好的数据模型,利用已有数据进行训练、评估打分,并和测试集中的已有数据比较,优化模型参数,可以得到分析图,对预测结果进行分析。

e. 预测过程。将相关站点的水位、降雨等实时数据应用到经评估训练得到优化参数的模型中,即可获得观测站的预测水位。

4.3 预测结果

a. 对模型的评估统计量如表 2 所示,从预测的结果可以看出,预测站点总体的水位涨落趋势以及水位的预测值准确度还是比较高的。从结果看来,预测模型能很好地逼近实际的观测量,说明预先根据经验假定的相关因素对于站点的水位影响还是比较显著的。

表 2 对所有时间点神经网络预测结果统计

统计量	统计值	统计量	统计值
最小误差	-0.396	标准差	0.059
最大误差	0.451	线性系数	0.935
平均误差	-0.002	记录数	2062
绝对误差	0.038		

b. 根据预测结果,峰值数据的预测值与实测值还是有一些差距,也是大部分预测的误差所在之处。所以,要提高峰值预测的准确度,需要能够找出产生峰值的真正原因所在,确定产生峰值的日期,根据记录日期找到当天对峰值产生影响的相关因素,整理这些可能因素的历史数据,尤其是在峰值发生当天的加报数据等。对于预测中出现的误差,可以根据数据挖掘软件的便捷性来进行实时校正,以减少误差。

5 结 语

a. 上面的预测仅做了 1 日的预测,在短期水文预报中时间步长可能会要求长一些,这就需要:①建立预测第 2 ~ N 天的模型;②对相关站的数据进行预测后,作为预测该数据的输入值。

b. 本次主要考虑相邻地区水位和地区降水对测站水位的影响,下阶段将细化流域站点影响因素,使得预测结果更加接近实测值。

c. 增加以下方法减少峰值误差:①利用加报数据;②将 1 天再分成 2 ~ 4 个时段,利用加报数据;③按线性补缺数据。

参考文献:

[1] 袁秀娟,夏军.径流中长期预报的灰色系统方法研究[J]. 武汉水利电力大学学报,1994,27(4):367-375.
[2] 王文,马骏.若干水文预报方法综述[J]. 水利水电科技进展,2005,25(1):56-60.
[3] 赵新生,赵杰吉,俊峰.浅谈数据挖掘与水文现代化[J]. 人民黄河,2005(27):28-29.
[4] 陈守煜.中长期水文预报综合分析理论模式与方法[J]. 水利学报,1997(8):15-21.
[5] 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(28):13-17.

(收稿日期:2007-01-23 编辑:熊水斌)

(上接第 99 页)

进行模拟。分析了大龙港泵站在完建期、正常设计水位正向和反向挡水工况时结构的受力特性,得出了泵房底板和运河侧挡水面板在各种工况下的应力分布规律。泵房底板在各工况下横河向应力基本以压应力为主,底板两侧靠近边墙中间则会出现一定程度的拉应力。完建期底板会出较小的顺河向拉应力,但在泵站挡水后拉应力区会逐渐减小。这些结论可以为泵站优化设计提供参考作用。

参考文献:

[1] 卜小明,龙驭球.一种薄板弯曲问题的四边形位移单元[J].力学学报,1991,23(1):53-60.
[2] 卜小明,龙驭球.一种高精度的矩形板弯曲单元[J].土木工程学报,1991,24(1):17-22.
[3] 朱伯芳.有限单元法原理与运用[M].2版.北京:中国水利水电出版社,1998.

(收稿日期:2007-03-15 编辑:熊水斌)