

DOI:10.3880/j.issn.1006-7647.2008.04.011

基于粗集和改进 BP 网络的城镇日用水量预测模型

张传建¹ 成 志²

(1. 雷山县水利局, 贵州 雷山 557100; 2. 凯里市水利局, 贵州 凯里 556000)

摘要 :针对城镇日用水量受某些影响因素冗余性、非定量性、非线性的影响以及这些影响在预测模型中很难体现等问题,分析了影响城镇日用水量的因素,利用粗集知识约简方法去除冗余,选择影响城镇日用水量的主要因素,结合改进的 BP 网络建立城镇日用水量预测模型,并将该模型的预测效果与未采用粗集方法去除因素冗余的模型预测效果进行比较,结果显示该模型的预测精度更高、所需时间更短、更加适用于影响因素较多的城镇年、月用水量的预测。

关键词 粗集;改进 BP 网络;日用水量;预测模型

中图分类号:TU991.31 文献标识码:A 文章编号:1006-7647(2008)04-0037-04

Forecasting model of urban daily water demand based on rough sets and improved back-propagation network//ZHANG Chuan-jian¹, CHEN Zhi²(1. Water Resources Department of Leishan County, Leishan 557100, China; 2. Water Resources Department of Kaili City, Kaili 556000, China)

Abstract :In view of the fact that urban daily water demand is subject to certain factors that are redundant, non-quantitative, nonlinear and not easily incorporated into a forecasting model, these factors are deleted with the rough sets theory and a forecasting model of urban daily water demand is established based on improved back-propagation (BP) network analysis. The results of an example simulation show that, after the redundant factors have been discarded with the rough sets theory, the prediction accuracy is higher and the forecasting time is shorter. Hence, the forecasting model based on rough sets and improved BP network analysis is valid and practical. It is especially appropriated for forecasting urban annual and monthly water demand, which is affected by many influencing factors.

Key words :rough sets; improved back-propagation network; daily water demand; forecasting model

在进行城镇水资源规划时,用水量预测是其重要内容之一。用水量预测结果直接影响到政府调度决策的正确性,也关系到城镇的可持续发展,有必要探求用水量的准确预测方法,揭示其变化规律和趋势,为水利部门制定供水规划提供科学的依据。

城镇用水量预测方法有回归分析法、趋势外推法、灰色预测法^[1-2]、用水量定额法和人工神经网络法^[3]。前 3 种方法把用水系统看成简单系统的延续,没有考虑社会因素,误差较大,因而失去了预测的价值。由于城镇日用水量的影响因素具有非线性、非定量性的特征,因此人工神经网络法得到了广泛的应用。当影响因素较多时,为避免网络规模过大影响计算速度,应选择主要因素,消除次要因素。在确定主要因素时,人们常带有一定的主观性,而粗集可以客观、有效地解决这一问题。

Matlab 是 MathWorks 公司于 1984 年推出的一种高性能计算机语言,作为一种优秀的数学软件,目前

被广泛应用于各个领域。Matlab 有 30 多个工具箱,如神经网络等。笔者在分析城镇日用水量影响因素的基础上,利用粗集方法去除冗余,选择主要因素,利用改进的 BP 网络建立预测模型,在 Matlab 中运行预测程序,得出预测结果。最后将预测效果与直接采用 BP 网络预测的效果进行比较,以显示该模型的有效性和实用性。

1 粗集理论

粗集^[4]的特点是利用数据本身所提供的信息,不需要任何附加信息或先验信息,在保持知识库分类能力不变的条件下,删除不相关或不重要的信息,降低信息冗余。

定义 1 区分矩阵:对于信息系统 $S = \langle U, A, V, f \rangle$ 其中 $A = C \cup D$ 是属性集合,子集 $C = \{a_i | i = 1, \dots, m\}$ 和 $D = \{d_j | j = 1, 2, \dots, k\}$ 分别称为条件属性集和决策属性集, $U = \{x_1, \dots, x_n\}$ 是论域, $|U| =$

n, S 的区分矩阵是一个 $n \times n$ 矩阵 其任一元素为

$$\alpha(x, y) = \{a \in A \mid f(x, a) \neq f(y, a)\} \quad (1)$$

因此 $\alpha(x, y)$ 是区别对象 x 和 y 的所有属性的集合。

显然, 区分矩阵是一个依主对角线对称的矩阵, 在考虑区分矩阵的时候, 只需要考虑其上三角(或下三角)部分就可以了。

定义 2 区分函数: 对于每个属性 $a \in A$, 指定一个布尔变量 a 。若 $a_1 \vee a_2 \vee \dots \vee a_k$, 则指定一个布尔函数 $\alpha(x, y) = \{a_1, a_2, \dots, a_k\} \neq \phi$, 用来表示 $\sum \alpha(x, y)$; 若 $\alpha(x, y) = \phi$, 则指定布尔常量 1。区分函数可定义如下:

$$\Delta = \prod_{(x, y) \in U \times U} \sum \alpha(x, y) \quad (2)$$

区分函数是一个范式, 由布尔变量和常量通过析取与合取运算来表达。

2 城镇日用水量的影响因素分析

我国城镇日用水包括居民生活用水、工矿企业生产用水、公共事业用水和农田灌溉用水^①等。工矿企业生产用水与生产规模和条件、季节和天气变化等有关; 居民生活用水直接受人口、天气、季节、生活习惯及社会活动等影响; 公共事业用水与节假日、社会活动等有关; 农田灌溉用水取决于灌溉面积、农忙季节、枯水期、天气等。城镇日用水量预测是根据过去几天或十几天内用水量的历史数据及影响因素的预报值对今后一天或几天的用水量做出预测。因此, 影响城镇日用水量的因素应该是近期对城镇日用水量影响较大的因素, 而企业规模、居民人口、灌溉面积等因素短期内不会有太大变化。

根据文献 [4] 日用水量影响因素分析, 天气变化对用水量的影响十分显著, 包括每日最高温度、最低温度及晴雨情况, 一般情况下, 温度越高, 用水量越大, 晴天比阴雨天用水量大; 一年四季因气温和日照等因素的影响, 工业生产条件和居民生活条件发生变化, 用水量也产生了差异, 一般冬季用水量比夏季少, 春、秋两季则基本接近, 随季节的变化不同月份日用水量变化明显; 由于人们工作以及日常生活以周为周期, 因此日用水量在每周呈现相同的变化趋势, 在周末由于居民用水量急剧增加, 总的用水量有所增加, 而在“国庆”、“春节”等长假, 由于部分生产部门停产, 用水量反而减少。同时, 在农忙季节和枯水期, 农田灌溉用水量也会有明显的变化。

综上所述, 影响城镇日用水量的因素有季节、气温、星期、法定节假日、农忙季节、枯水期、天气、高温

持续时间等。

3 基于粗集和改进 BP 网络的预测模型

去除冗余是粗集的长处, 对于 BP 网络来说, 数据越少, 神经网络就越容易学习; 神经元个数越多, 权值越多, 则训练时间就越长, 泛化能力就越差。用粗集约简后的数据作为改进 BP 网络变量数目的依据和训练样本, 可以减少网络神经元个数, 提高训练效率。因此, 可以将粗集和改进 BP 网络相结合, 用于城市日用水量的预测, 其基本步骤如下:

- 收集以往的日用水量实例, 构造初始样本数据表, 以影响因素作为条件属性, 以用水量作为决策属性。
- 将初始样本数据表中的条件属性进行人工分类, 并进行离散化处理。
- 构造一张与初始样本数据表大小相同的二维表格, 以离散化的属性值填充, 即可得到能采用粗集处理的决策表。
- 用粗集的基于区分矩阵方法对决策表进行属性约简, 以约简的属性集作为所提取的关键因素进行预测。
- 依据提取的关键因素, 以初始样本数据表中相应关键因素属性的值构造训练样本, 以预测日期关键因素属性值作为测试样本, 设置相关的改进 BP 网络参数, 编制 Matlab“*.m”程序, 并将程序复制到 Matlab 的工作目录“work”中。
- 在安装好 Matlab 软件系统的计算机中, 启动 Matlab, 打开并运行“*.m”程序文件, 即可得出预测结果。

4 城镇日用水量预测实例

4.1 决策表的建立

设季节为 S 、星期为 D 、法定节假日为 H 、农忙季节为 B 、枯水期为 G 、晴雨天为 W 、最高气温为 T 、高温持续时间为 h 、用水量为 Q , 以上所有因素均可由简单统计得到。对用水量进行特殊处理, 即通过单位的变化, 将用水量转化为 $[0, 1]$ 之间的数。整理的实例数据见表 1。

4.2 属性离散化

把表 1 中的数据看作一个信息系统, 影响城镇日用水量的因素组成条件属性集合 $R = \{S, D, H, B, G, W, T, h\}$, 10 组数据构成论域 $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, 对每个条件属性数字化, 作出以下数字化定义: S 中 1 表示春季, 2 表示夏季, 3 表示秋季, 4 表示冬季; D 中 0 表示星期一至星期四, 1 表示

①在我国南方中小城镇用水中农田灌溉用水所占比例较大, 因此, 笔者将农田灌溉用水纳入城镇用水进行分析。

表 1 某城镇 2006 年 10 月 1 ~ 10 日的历史统计数据									
日期	<i>S</i>	<i>D</i>	<i>H</i>	<i>B</i>	<i>G</i>	<i>W</i>	<i>T</i> / ℃	<i>h</i> / h	<i>Q</i> / 万 t
10-01	秋	星期日	国庆	非农忙	枯水期	阴	22	0	0.263
10-02	秋	星期一	国庆	非农忙	枯水期	晴	27	1.0	0.265
10-03	秋	星期二	国庆	非农忙	枯水期	晴	29	1.1	0.262
10-04	秋	星期三	国庆	非农忙	枯水期	晴转阴	29	0.9	0.275
10-05	秋	星期四	国庆	非农忙	枯水期	阴	20	0	0.277
10-06	秋	星期五	国庆	非农忙	枯水期	晴	26	1.2	0.262
10-07	秋	星期六	国庆	非农忙	枯水期	晴	25	2.5	0.279
10-08	秋	星期日	非假日	非农忙	枯水期	阴	22	0	0.296
10-09	秋	星期一	非假日	非农忙	枯水期	阴转晴	25	1.1	0.300
10-10	秋	星期二	非假日	非农忙	枯水期	阴	21	0	0.299

星期五至星期日 ;*H* 中 0 表示非节假日 ,1 表示节假日 ;*B* 中 0 表示非农忙季节 ,1 表示农忙季节 ;*G* 中 0 表示非枯水期 ,1 表示枯水期 ;*W* 中 0 表示阴天 ,1 表示晴天 ,2 表示晴转阴或阴转晴 ;*T* 中 0 表示低温天气(气温低于 25 ℃) ,1 表示高温天气(气温高于或等于 25 ℃) ;*h* 中 0 表示高温持续时间小于 1 h ,1 表示高温持续时间大于或等于 1 h。

根据以上数字化定义进行离散化处理。去掉属性值均相同的 *S* ,*B* ,*G* ,得到影响城镇日用水量的因素集{*D* ,*H* ,*W* ,*T* ,*h* }。去掉重复的纪录 3 和 6 ,并将 *D* ,*H* ,*W* ,*T* ,*h* 用 *a* ,*b* ,*c* ,*d* ,*e* 代替 ,得到最终决策表(表 2)。

表 2 最终决策结果					
<i>U</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
1	1	1	0	0	0
2	0	1	1	1	1
4	0	1	2	1	0
5	0	1	0	0	0
7	1	1	1	1	1
8	1	0	0	0	0
9	0	0	2	1	1
10	0	0	0	0	0

4.3 提取关键因素

采用基于粗集区分矩阵的约简方法计算的区分矩阵见表 3。基于粗集的区分矩阵约简方法简单地说 ,如表 3 中的单元(2 ,1) 的值“ *acde* ”,即是由表 2 中第一、第二条纪录中因素集{*a* ,*b* ,*c* ,*d* ,*e* }属性值不同的属性 *a* ,*c* ,*d* ,*e* 而得。为了使预测方法更加

表 3 区分矩阵								
<i>U</i>	1	2	4	5	7	8	9	10
1								
2	<i>acde</i>							
4	<i>acd</i>	<i>ce</i>						
5	<i>a</i>	<i>cde</i>	<i>cd</i>					
7	<i>cde</i>	<i>a</i>	<i>ace</i>	<i>acde</i>				
8	<i>b</i>	<i>abcde</i>	<i>abcde</i>	<i>ab</i>	<i>bcde</i>			
9	<i>abcd</i>	<i>bc</i>	<i>bc</i>	<i>bcde</i>	<i>abc</i>	<i>acde</i>		
10	<i>ab</i>	<i>bcde</i>	<i>bcde</i>	<i>b</i>	<i>abcde</i>	<i>a</i>	<i>cde</i>	

简便实用 ,根据粗集理论和数理逻辑中析取和合取运算 ,对区分矩阵进行简化。简化方法通俗地说就是先判断有没有单因素 ,如“ *a* ”或“ *b* ”等 ,如有 ,则将区分矩阵中包含“ *a* ”的“ *acde* ”等都去掉 ,保留“ *a* ”,然后再判断区分矩阵中的 2 个元素的包含关系 ,如“ *acde* ”包含“ *acd* ”,则在区分矩阵中去掉“ *acde* ”,保留“ *acd* ”。如此循环直到区分矩阵中的任意 2 个元素都不存在包含关系。最后 ,在区分矩阵中如有相同的单元值 ,则只保留 1 个。简化的区分矩阵见表 4。根据基于粗集的区分矩阵约简方法中区分函数的定义 ,以及数理逻辑中的析取、合取运算 ,则区分函数为

$$\Delta =(a)\wedge (b)\wedge (c\vee e)\wedge (c\vee d)=\\ (a\wedge b\wedge c\vee e)\vee (a\wedge b\wedge c\vee d) \quad (3)$$

因此 ,决策表有 2 个约简{*a* ,*b* ,*c* ,*e* }和{*a* ,*b* ,*c* ,*d* }核是{*a* ,*b* ,*c* }。对于该系统 ,约简后取其中一个约简 ,如表 5 所示的决策表。可见 ,影响因素由 8 个约简到 4 个。

表 4 简化的区分矩阵								
<i>U</i>	1	2	4	5	7	8	9	10
1								
2								
4			<i>ce</i>					
5	<i>a</i>		<i>cd</i>					
7								
8	<i>b</i>							
9								
10								

表 5 约简{ <i>a</i> , <i>b</i> , <i>c</i> , <i>d</i> }的决策结果					
<i>U</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>Q</i> /万 t
1	1	1	0	0	0.263
2	0	1	1	1	0.265
4	0	1	2	1	0.275
5	0	1	0	0	0.277
7	1	1	1	1	0.279
8	1	0	0	0	0.296
9	0	0	2	1	0.300
10	0	0	0	0	0.299

4.4 编制 Matlab 程序

采用 3 层改进 BP 网络模型 ,其中输入层有 4 个神经元 ,输出层为 1 个神经元。设置改进 BP 网络参数 :学习率为 0.1 ,动量因子为 0.9 ,步长为 5 000 ,目标误差为 0.000 01。在安装好的 Matlab 软件系统中 ,编制相关程序 ,运行即可得出结果。以论域 1 ,2 ,4 ,5 ,7 ,8 ,9 为训练样本 ,论域 3 ,6 ,10 为验证样本 ,由于 1 次训练存在偶然性 ,取 10 次预测的平均值为最终结果 ,见表 6。

4.5 预测结果及分析

约简前 ,样本 3 ,6 ,10 的预测值误差分别为

表 6 约简前后预测结果对比

次数	约简前				约简后			
	样本 3	样本 6	样本 10	步数	样本 3	样本 6	样本 10	步数
1	0.271	0.263	0.283	33	0.265	0.266	0.314	20
2	0.264	0.268	0.282	4	0.262	0.264	0.278	6
3	0.267	0.266	0.264	8	0.264	0.266	0.290	4
4	0.263	0.264	0.309	4	0.264	0.263	0.277	4
5	0.270	0.263	0.271	22	0.264	0.265	0.345	4
6	0.271	0.263	0.272	412	0.263	0.261	0.262	4
7	0.265	0.264	0.320	9	0.267	0.266	0.286	9
8	0.265	0.269	0.336	21	0.264	0.268	0.307	16
9	0.266	0.267	0.274	39	0.268	0.263	0.331	6
10	0.265	0.265	0.276	4	0.265	0.263	0.291	4
均值	0.267	0.265	0.289	55.6	0.265	0.265	0.298	7.7

1.78% ,1.12% , - 3.26% ,平均训练步数为 55.6 ,约简后 ,样本 3 ,6 ,10 的预测误差分别为 0.98% , 0.87% , - 0.13% ,平均训练步数为 7.7。即约简后经过 7.7 步训练后就收敛 ,约简前则需要经过 55.6 步才能收敛。结果表明 ,约简前和约简后的预测精度均在 90% 的精度范围内 ,属于高精度预测。但是约简后预测效果更加明显 ,预测精度比约简前的预测精度高得多 ,并且误差收敛的平均步数是约简前的 13.8% ,可以节省 86.2% 的时间。用水量的特殊处理在提高预测精度方面也起到了很大作用 ,以样本 10 为例 ,用水量未经过特殊处理时在同等条件下约简前后的预测误差分别为 4.28% 和 0.27% ;经过特殊处理后 ,约简前后的预测误差分别为 - 3.26% 和 - 0.13%。因此 ,笔者提出的基于粗集和改进 BP 网络的城镇日用水量预测模型具有快速、

高效的特点。粗集强大的约简能力和模型的强抗干扰性在去除干扰因素影响的同时 ,提高了预测精度 ,并节省了大量时间 ,这在影响因素较多的用水量预测中具有较大的优势。

5 结 语

本文利用粗集理论消除干扰项 ,确定影响城镇日用水量的主要因素 ,用改进的 BP 网络找出日用水量与影响因素间的非线性关系 ,通过与粗集相结合预测城镇日用水量 ,是城镇日用水量预测的新尝试 ,为城镇水资源调度提供了一种新的决策方法。该方法可以剔除次要因素 ,简化网络结构 ,提高预测效率和精度。当影响城镇日用水量的因素较多时 ,效率将会更高。该方法在影响因素较多的城镇年、月用水量预测方面有广泛的应用前景。

参考文献 :

[1] 王涛 ,张宏伟 ,牛志广 .GM 组合模型用于城市生活用水量预测[J].水资源保护 ,2007 ,23(5) :28-30.
[2] 楼玉 ,张清 ,刘国华 ,等 .城镇用水量预测中的多变量灰色预测模型[J].水资源保护 ,2005 ,21(1) :11-13.
[3] 路志强 ,李书全 ,赵良英 ,等 .改进 BP 网络模型在城镇用水量预测中的应用[J].水科学与工程学报 ,2005(1) :18-20.
[4] 张文修 .粗糙集理论与方法[M].北京 :科学出版社 ,2001 :22-25.

(收稿日期 2007-07-30 编辑 :方宇彤)

· 简讯 ·

河海大学英文期刊 *Water Science and Engineering*
(水科学与水工程) 创刊号出版发行

由国家新闻出版总署批准、河海大学主办的英文期刊《Water Science and Engineering》(水科学与水工程) 第一期日前出版发行。

《Water Science and Engineering》英文期刊为季刊 ,B5 开本 ,每期 120 页 ,以与水科学、水工程有关的研究人员、工程技术人员及高校师生为主要读者对象 ,面向国内外公开发行人。期刊主要收录国内外水科学研究、水工程建设、水资源保护、水生态修复等领域的研究成果及技术论文 ,报道国内外相关的学术会议及技术交流等消息动态 ,稿件来源主要是国内外高等院校、科研院所、规划设计和工程建设单位的专家、教授、工程技术人员和学生用英文撰写的高水平稿件。创刊号收录的稿件有河海大学水文水资源与水利工程科学国家重点实验室芮孝芳教授等合作的《基于网格水滴的流域汇流计算》、河海大学海洋学院郑金海教授和京都大学防灾研究所间濑肇教授等合作的《强沿岸流影响下的随机波浪传播变形模拟研究》、美国密歇根州环保署钱学德教授的《不同垃圾渗滤液条件下填埋场平移破坏的极限平衡分析》、河海大学岩土工程研究所刘汉龙教授等合作的《不同应力路径作用下筑坝粗粒料剪切特性试验研究》、美国乔治华盛顿大学 Michael M. Cernea 教授的《补偿和利益共享 :为什么必须改革移民安置的政策和实践》等。

(本刊编辑部供稿)