

基于应急报警事件的先重构后建模 预测法的探索研究

薛京生^{1,2}, 李瑞卿²

(1. 天津大学 电子信息工程学院, 天津 300072; 2. 天津市公安局信息通信处, 天津 300020)

摘要: 通过大量实验证明了公安110报警事件具备可预测性。该预测模型具备适用性特点,可应用于公安机关实战指挥及政府应急管理相关部门的预测应用,并辅助决策支持。该创新建模预测过程称为先重构后建模预测法。

关键词: 小波; 自回归; 数据挖掘; 时间序列; 应急预测

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)07-2488-03

doi:10.3969/j.issn.1001-3695.2012.07.023

Exploration and research for forecasting method of reconstructed first before modeling based on emergency-alarm events

XUE Jing-sheng^{1,2}, LI Rui-qing²

(1. School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China; 2. Information & Communication Department of Tianjin Public Security Bureau, Tianjin 300020, China)

Abstract: This research proved that 110-alarm events have predictability; the applicability of forecast models that could be applied to government emergency management departments and applied for forecasting of public security organ in actual command and supports assistant-decision. This innovative method used high-order AR to synthesize time series of 110-alarm forecasting result to the original signal forecasting result and to be proved through experiments. This innovative modeling and forecasting process is called forecasting method of reconstructed first before modeling.

Key words: wavelet; auto-regressive; data-mining; time-series; emergencies-forecasting

时间序列模型包括AR、MA和ARMA等模型。从理论上讲,任何ARMA模型或AR模型都可以用无穷阶的AR模型来逼近,因而有的学术观点认为应当尽可能用AR模型来描述时间序列,即使真实模型是ARMA或MA模型,也可以用高阶AR模型来近似^[1,2]。公安机关“110”数据采集过程是真实可靠的,因此在“110”报警事件预测中可以有大量“廉价”的数据样本提供建模使用,便于用高阶AR模型来逼近任何一类时间序列模型。为此,本研究确定全部使用高阶AR模型进行建模和实验比较,从而得到了一些创新性的结论。

1 研究过程

本实验建模数据取自某大城市公安局2008年全年110报警数据,取样间隔1h。截取其中连续8784项(366d×24h)数据,预测数据使用截取数据后连续3d的110报警数据。每个数据点代表每1h接警数之和。

1.1 实验数据与图形

该110报警数据时间序列为8784个样本,利用Daubechies小波系 $N=3$,即db3作为尺度函数进行多尺度分析,其结果如图1所示。为便于比较,将分解得到的各个尺度上近似信号、误差信号分别重构回到原尺度,并给出各信号样本自相

关系系数。图2是分解后各尺度上近似信号的重构结果。

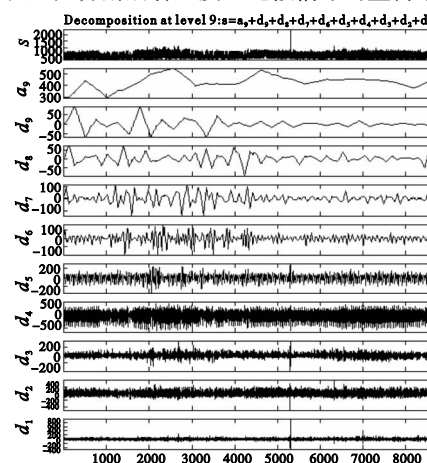


图1 Daubechies小波系($N=9$)共8784项数据多尺度分析

1.2 实验特征

a)这是由于多个低通滤波器的作用,将较高频信号不断消去,形成平滑稳定信号,且在不同尺度上的相似性很强。以物理上看,误差信号是被低通滤波器过滤掉的高频信号,每一尺度上的误差信号是频率相近的一组信号叠加;而不同尺度上的误差信号,频率相差较大,它们之间不存在相似性,也互不相

收稿日期: 2011-12-31; 修回日期: 2012-02-09

作者简介: 薛京生(1963-), 副研究员, 客座教授, 博士(后), 主要研究方向为数据挖掘、决策分析(xuejingsheng@163.com); 李瑞卿, 科长, 工程师, 硕士。

关。于是,一组信号(110报警)经过多次滤波,随后得到平稳的基本趋势信号(近似信号)和一系列噪声干扰信号(误差信号)。

b)通过观察各层细节分量可以看到,干扰信号与基本趋势信号的区分是相对的。这与实验选取的预测模型的区分精度有关,只有区分精度高,才能将干扰信号从基本趋势信号中分离出来。

c)这种自相关性随着尺度的减小而减弱,在某个尺度上自相关接近0,说明接近成为白噪声信号。

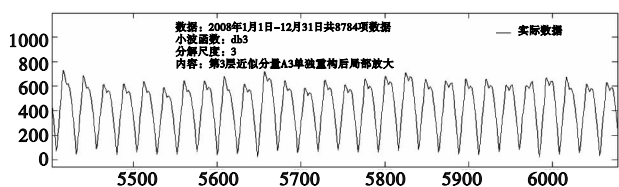


图2 近似分量单独重构局部放大

1.3 运用多尺度分析方法开展时间序列预测

以上实验为利用多尺度方法进行时间序列预测提供了一种创新性思想。即通过多尺度分析方法将表达110报警事件流的本质变化趋势的基本趋势信号(近似信号)与高频干扰信号(误差信号)相分离,这种干扰信号可能有多组。对基本趋势信号与干扰信号利用现有的时间序列方法进行分析,得到外推预测结果;然后将基本趋势信号和不同尺度干扰信号及其各自预测结果再利用重构算法合成到原始尺度上的信号序列中,这一序列中就包含了对原始信号的预测结果。

假设110报警事件数量以 T 为周期变化。在每个周期 T 内的相同时刻服从相同的分布,可构造如下随机序列^[3,4]:

$$v_k(t) = \{v(t + kT) | k = 1, 2, \dots, t \in [0, T)\} \quad (1)$$

其中: T 是 $v(t)$ 的周期,对110报警事件时间序列一般有 $T = 24$ h。

鉴于原始信号具备强烈的周期特征,为严谨起见,把每一个周期内同一时刻的样本点提取出来,作为一个独立的子时间序列单独加以分析预测。

2 子时间序列特征分析

如果有周期特征,那么周期 T 是多少,以此来确定每个细节分量需要分为几个子时间序列。计算过程如下:

a)第三层细节分量 $D3$

$D3$ 是所有细节分量中频率最低,各细节分量中噪声成分最少的,它最有可能与近似分量具有同样的时序特征。为此,采用小波函数, $N=3$,即db3作为尺度函数,分解深度取3进行多尺度分析,得到细节分量 $D3$ 的自相关系数,如表1所示。

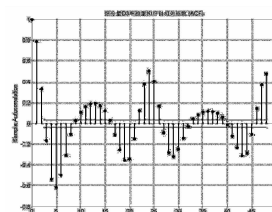
图3是细节分量 $D3$ 自相关系数统计图。可以看出,未考虑周期性而使用原分量重构序列计算的自相关系数,如图3(a)所示,它呈现明显的周期性,周期 T_{D3} 与 T_{A3} 相等,都是24,证明了前面的推测。然后把该序列按照时段划分为24个子时间序列,考察 $t=0$ 时刻的子时间序列的自相关系数,如图3(b)所示,自相关系数呈现缓慢下降趋势。这说明该子序列属于不稳定序列,且属高阶AR,阶数 $p \geq 2$ 模型,可以使用高阶AR模型进行预测。

b)第二层细节分量 $D2$

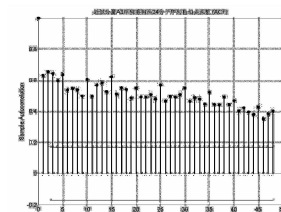
对 $D2$ 的特征判断则稍微复杂。表2是细节分量 $D2$ 的自相关系数(小波函数db3,分解深度3)。

表1 细节分量 $D3$ 的自相关系数

K	ACF	K	ACF	K	ACF	K	ACF
1	0.81683	13	0.45576	25	0.52765	37	0.37275
2	0.41173	14	0.22299	26	0.27996	38	0.16271
3	-0.04919	15	-0.048121	27	-0.0066398	39	-0.077212
4	-0.43553	16	-0.28669	28	-0.25775	40	-0.28442
5	-0.61668	17	-0.41954	29	-0.39721	41	-0.39448
6	-0.63425	18	-0.45497	30	-0.43875	42	-0.4161
7	-0.55466	19	-0.41771	31	-0.41203	43	-0.37256
8	-0.37347	20	-0.27857	32	-0.29606	44	-0.23858
9	-0.10516	21	-0.025228	33	-0.084965	45	-0.004012
10	0.18495	22	0.26585	34	0.1576	46	0.26158
11	0.43711	23	0.52011	35	0.37033	47	0.49083
12	0.55442	24	0.63628	36	0.46662	48	0.59117



(a)未考虑24 h周期自相关系数

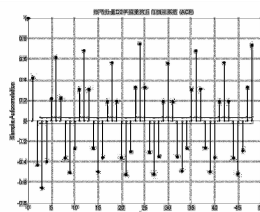


(b) $t=0$ 时考虑24 h周期自相关系数

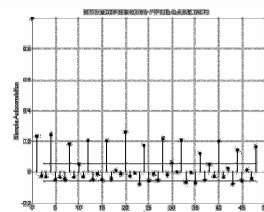
图3 细节分量 $D3$ 自相关系数

表2 细节分量 $D2$ 的自相关系数

K	ACF	K	ACF	K	ACF	K	ACF
1	0.44537	13	0.23578	25	0.36923	37	0.22647
2	-0.38438	14	-0.24677	26	-0.29143	38	-0.24345
3	-0.66203	15	-0.40877	27	-0.55335	39	-0.3992
4	-0.46952	16	-0.2547	28	-0.42062	40	-0.24573
5	0.13534	17	0.23425	29	0.12231	41	0.23496
6	0.58048	18	0.53751	30	0.52892	42	0.53035
7	0.26852	19	0.12688	31	0.23447	43	0.1206
8	-0.24715	20	-0.42295	32	-0.2446	44	-0.42382
9	-0.41636	21	-0.55542	33	-0.39886	45	-0.55042
10	-0.25947	22	-0.29101	34	-0.24426	46	-0.28446
11	0.22675	23	0.36988	35	0.22582	47	0.3684
12	0.55767	24	0.80087	36	0.54265	48	0.79249



(a)未考虑周期性原序列自相关系数

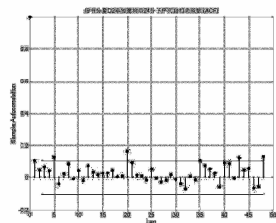


(b)考虑6 h周期自相关系数 $t=0$ 子序列

图4 细节分量 $D2$ 自相关系数

同理,图4(a)是未考虑周期性的细节分量 $D2$ 自相关系数统计图。可以看出未考虑周期性而使用原分量重构序列计算的自相关系数呈现明显的周期性,自相关系数分布呈现除衰减幅正弦下降趋势,周期 T_{D2} 似表现为6。然后把该序列按照时段划分为6个子时间序列,考察 $t=0$ 时刻的子时间序列的自相关系数,如图4(b)所示,自相关系数呈现比较特殊的状态:子序列仍旧表现出周期性显著值,周期 $T_{D2} T=6 | t=0 = 4$ 。这表明 $T=6$ 不能正确描述该子序列的周期特征,于是取 $T_{D2} = 24$ (即 4×6),将原序列划分为24个子序列;然后计算 $t=0$ 时的序列自相关系数,如图5所示。这说明相关系数呈现随机状,且满

足95%的值均处在上下控制限度内;也表明使用周期24描述该序列这种特征是适当的,且该序列属高阶AR,可以使用高阶AR模型进行预测。



(考虑周期为24 $t=0$ 子序列自相关系数)

图5 细节分量D2自相关系数

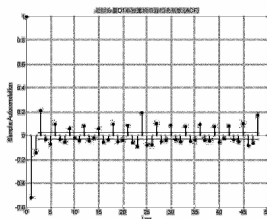
c) 第一层细节分量D1

D1,理论上特征判断属于高频噪声,域D3包含的噪声成分更多。表3是细节分量D1的自相关系数(小波函数db3,分解深度3)。

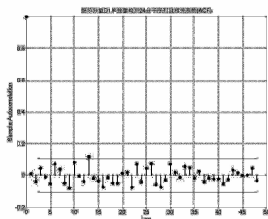
表3 细节分量D1的自相关系数

K	ACF	K	ACF	K	ACF	K	ACF
1	-0.45302	13	-0.067728	25	-0.10706	37	-0.089274
2	-0.27807	14	-0.059359	26	-0.13642	38	-0.055815
3	0.28782	15	0.11071	27	0.14915	39	0.12263
4	-0.018282	16	-0.081451	28	-0.046001	40	-0.096246
5	-0.11081	17	-0.035042	29	-0.067656	41	-0.043008
6	0.12589	18	0.12334	30	0.12032	42	0.14658
7	-0.021586	19	-0.070261	31	-0.031951	43	-0.065046
8	-0.097496	20	-0.044488	32	-0.086662	44	-0.073564
9	0.11261	21	0.15012	33	0.12369	45	0.1483
10	-0.046425	22	-0.13914	34	-0.072209	46	-0.10846
11	-0.078439	23	-0.10508	35	-0.076923	47	-0.1056
12	0.15309	24	0.28684	36	0.179	48	0.25426

同理,图6(a)是未考虑周期性的细节分量D1自相关系数统计图。可以看出未考虑周期性而使用原分量重构序列计算的自相关系数除了 $r1$ 、 $r2d$,均表现出显著性,应该属于白噪声序列。但是仔细观察后发现,多数自相关系数均超出上下控制限,所以该序列还有未被反映出来的周期性,单纯只观察自相关系数不能判断出周期。联想到D2、D3的周期性,遂大胆假设周期 $T_{D1}=24$,然后把该序列按照时段划分为24个子时间序列,考察 $t=0$ 时刻的子时间序列的自相关系数,如图6(b)所示。说明相关系数呈现随机状,且满足95%的值均处在上下控制限内,这表明使用周期24描述该序列特征是适当的;且该序列属高阶AR,完全可以使用高阶AR模型进行预测^[5,6]。



(a)未考虑周期性原序列自相关系数



(b)假设24 h周期特征时自相关系数 $t=0$ 子序列

图6 细节分量D1自相关系数

d) 序列重构

本研究使用了常用的AIC(Akaike information criterion)方法确定AR模型阶数。AIC准则的一般形式^[7,8]为

$$AIC = (-2) \log \text{极大似然函数} + 2(\text{独立参数个数}) \quad (2)$$

使用AIC准则确定各子时间序列的最佳回归阶数,可以利用已有的ARMA方法或AR(MA)方法进行建模和预测。本研究采用高阶AR方法,再将各序列的预测结果合成为对原始信号的预测结果,称这一创新的建模预测过程为先重构后建模预测法。

与确定分解尺度相似,可以将 $v_k^N, w_k^j (k=1, 2, \dots, N)$ 的 $N+1$ 个序列单独重构到原尺度上,得到 $N+1$ 在原始尺度上的时间序列 $\tilde{v}_k^N, \tilde{w}_k^j (k=1, 2, \dots, N)$,这 $N+1$ 个时间序列具有相同的样本数,它们的代数和等于原信号^[9,10],即

$$v_k^0 = \tilde{v}_k^N + \sum_{j=1}^N \tilde{w}_k^j \quad (3)$$

可以在原尺度上对 $N+1$ 个重构后的时间序列分别建模,得到以下模型:

$$\varphi_v^-(B) \tilde{v}_i^N = \theta_v^-(B) \tilde{a}_i^v \quad (4)$$

$$\varphi_w^-(B) \tilde{w}_i^N = \theta_w^-(B) \tilde{a}_i^w \quad j=1, 2, \dots, N \quad (5)$$

利用上述模型进行预测,并将预测结果累加,得到原始信号的预测结果。可见,各个尺度上分解信号的自相关性一般低于该信号在原始尺度上重构信号的自相关性,若采取先建模预测后重构的做法,对分解信号建模效果将劣于对其重构信号建模的效果。

3 结束语

实验证明:

a) 对同一个时间序列使用不同的小波函数进行预测会有不同的精度。最适合为110报警电话时间序列建模的小波函数是Bior1.3(Bior族小波),其各项误差相对最小。由此计算出来的样本各分量所获得的各高阶AR模型系数能够较准确地表述110报警事件时间序列的时序特征,因而能够更准确地预测该报警事件时间序列的发展趋势。

b) 110报警事件具备可预测性,其预测模型具备适用性特点,可以用于政府应急管理部和公安机关指挥实战的预测应用,辅助决策支持。

参考文献:

- [1] 杨增武. 统计预测原理[M]. 北京:中国财政经济出版社,1990.
- [2] 于锦华,石北源,杨维权. 概率论与数理统计[M]. 广州:中山大学出版社,2000.
- [3] 汪荣鑫. 数理统计[M]. 西安:西安交通大学出版社,2004.
- [4] 布洛克威尔·戴维斯,田铮. 时间序列的理论与方法[M]. 北京:高等教育出版社,2001.
- [5] PERCIVAL D B, WALDEN D B, WALDEN A T. Wavelet methods for time series analysis[M]. 北京:机械工业出版社,2004.
- [6] KANTZ H, SCHREIBER T. Nonlinear time series analysis[M]. 北京:清华大学出版社,2000.
- [7] 杨虎,王会琦,程代杰. 基于预测的序列异常数据挖掘[J]. 计算机科学,2004,31(4):117-120.
- [8] 贺国光,马寿峰,李宇. 基于小波分解与重构的时间序列预测法[J]. 自动化学报,2002,28(6):1012-1014.
- [9] 虞湘滨,徐伟业. 小波快速分解和重构的一般公式[J]. 河海大学常州分院学报,2001,15(3):6-10.
- [10] 鲁万波,余竞. 用AIC准则识别最佳因子分析模型[J]. 成都大学学报:自然科学版,2003,22(2):28-32.