非线性时间序列的符号化分析方法研究*

金宁德 李伟波
(天津大学电气与自动化工程学院,天津 300072)

摘要 符号时间序列分析方法是近年来新兴的一种数据处理方法,已经被广泛的应用于各个领域.采用符号化分析方法能够从动力系统中快速有效地提取有用定量信息,计算简单快捷,而且能够有效的抑制噪声.本文采用三种统计量用于表征符号化时间序列的特性,用Henon方程作为算例验证了该方法的可靠性,并且将此方法应用于垂直上升管中油水两相流流型分析,结果表明从符号时间序列计算的统计量对油水两相流过渡流型变化敏感.

关键词 符号时间序列,时间不可逆转性,统计量

前言

人类对自然界研究的根本目的在于揭示客观事物基本不变的性质,但是描述事物本质的特征量为数不多,所以对原始数据进行约化,可以取得更好的效果.符号化时间序列分析方法是一种数据分析的新工具,而且在实际中得到了验证,现在已经应用于象天文学、地理学、医药生物、化工、机械、人工智能、控制通讯、数据挖掘等各种领域,研究证明利用这种方法可以大大缩短计算时间,能够很好的抑制噪声.

符号时间序列分析起源于上世纪90年代中期,它是由符号动力学理论、混沌时间序列分析和信息理论发展起来的一种新的信息分析方法,它为强噪声工程对象提供了一种简单、快速且有效的处理方式[1].Tang等[2]运用符号时间序列分析方法从噪声信号中重构混沌系统,他们研究了在时空系统中此方法的参数选择问题并且证明可以将此方法运用于不规则时间序列处理[3].Daw等[4]将此方法应用于复杂的多相流测量信号分析.Lehrman等[5]对湍流波动的混沌信号采用了符号分析方法.Godelle等[6]在分析水和甘油混和物喷射状态时运用了符号化方法.最近,Daw等[7,8]提出了如何针对时间不可逆转性指标进行符号划分方法,并且系统地综述了实验数据符号化分析方法.

在石油工业中对油井生产及动态监测技术有重要影响的油水两相流流型辨识一直是没有很好解决的问题.在早期研究中,Govier等[9]研究了垂直上升管中油水两相流流型,在内径为26.4 mm管中观察到有四种流型(泡状、段塞、泡沫、雾状)并建立了相应的流型图.Vigneaux等[10]在内径为20 cm垂直上升管中观察到了油水两相流过渡流型其持水率变化范围为0.20~0.30,并且没有出现段塞流及泡状流,Zavarch等[11]也作了此类研究.近年来Flores等[12]在内径为50.8 mm的垂直管中观察油水两相流有六种流型,同时提出了流型转变机理模型.尽管在油水两相流流型辨识领域已经取得了很大的进展,但是很多问题仍然没有解决.首先,不同研究者对油水两相流流型辨识的结果差异很大,而且观点也不尽一致;其次,不同研究者采用的流动工况及试验条件不同,所给出的流型转换准则与研究者对流型主观认识及仪器测量精度有很大关系,因此对流型进行客观的辨识仍有待于进一步探讨.

本文采用非线性时间序列的符号化分析方法,利用时间不可逆转性 T_{fb},χ^2 统计量及修正的香农熵 H_s 对符号化的序列进行特征量提取.重点讨论了符号时间序列分析方法中的关键参数选择问题,并将此方法应用于油水两相流流型表征.

1 符号时间序列分析

对时间序列进行符号化分析分为两步:先将时间序列转化为符号序列,再对符号序列进行统计学分析.

1.1 符号时间序列定义

将原始时间序列转化为符号序列的形式,首先需要将测得数据分割为离散的数字.通常采用两种方法划分原始数据[13]:分割区间法及差值法.如图

2004-07-03收到第1稿,2004-09-02收到修改稿.
* 国家自然科学基金(60374041)和教育部留学回国人员科研启动基金资助项目.

1 所示为最简单的分割区间法,测量值在分割线之上为 1,之下为 0.图 2 中差值法规定相邻的两点差值为正,则为符号 1,如果为负,则为符号 0,这种差值的符号划分方法对于突发的大噪声不敏感,因此通常采用分割区间法来划分符号序列.

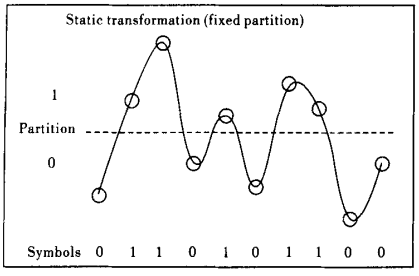

图 1　连续模拟序列分割区间的离散化方法示意图[13]
Fig. 1　Fixed partition transformation of a data series into a symbol series [13]

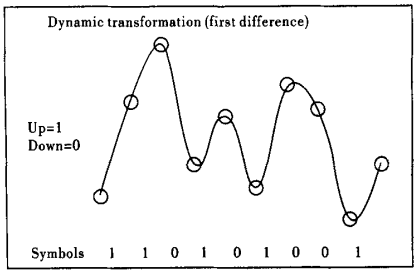

图 2　连续模拟序列差值式的离散化方法示意图[13]
Fig2　First difference transformations of a data series into a symbol series[13]

对于分割区间法,如果分割区间越多,原始数据转化为符号时间序列就越细化,但是也会带来一些负面的影响,例如对噪声的抑制作用减弱,因此需要选择合适的分割区间.

Daw 等人[7] 基于符号区间法给出了一种全面的生成符号时间序列的方法,数据符号化的基本思想就是在几个可能值上对时间序列进行离散化,把有许多可能值的数据序列变换为只有几个互不相同值的符号序列.这是一个粗粒化过程,将原始时间数据状态空间划分为一系列区间,每一个区间分配不同的符号,根据原始数据落在不同的区间,将它们转化为不同的符号,从而将一个连续模拟的时间序列转换为一个符号序列.图 3 解释了如何得到划分区间及原始时间序列如何转化为符号序列.

原始数据转化为符号形式后,就要提取其特征量.一种特别有用的办法就是选择一个标准长度 L,L 个连续的符号组成一个字,每一个字被编码为十进制数,形成了新的序列.上述过程与时间延迟嵌入类似,即利用离散的符号替代连续的原始测量值.然后以出现的字的频率作为时间序列分析的一个指标,如图 4 所示就是符号序列柱状图,符号序列柱状图直观地描述了每一个字出现的频率.

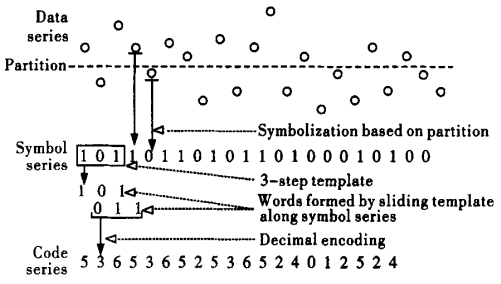

图 3　符号化时间序列生成示意图[7]
Fig. 3　Process of symbolizing a time series[7]

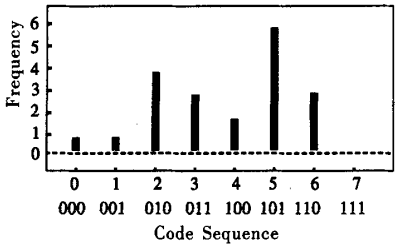

图 4　符号序列柱状图[7]
Fig. 4　Symbol sequence histogram[7]

1.2　符号序列统计量

符号时间序列分析的重点在于分析每一个字的相对频率,通过对出现的字进行分析,达到揭示系统动力学特性目的.一般来说,表征系统主要特性的字出现的频率往往较高,而那些出现频率很低的字不能反映系统的动力学特性.已经有多种统计方法用于频率统计分析.迄今为止,三种主要的统计方法被证明是非常有用的,它们分别是:修正的香农熵 H_s,时间不可逆转性指标 T_{fb} 及 χ^2 统计量.

修正的香农熵是基于信息理论的统计量分析,其定义如下[4]

$$H_s = -\frac{1}{\log N_{obs}} \sum_i p_i \log p_i \qquad (1)$$

其中 p_i 是第 i 个字出现的概率,N_{obs} 是在符号序列中出现的不同字的数量.对于完全随机的数据,修正的香农熵等于 1.对于不完全随机的数据,修正香农熵的范围 0 ~ 1.它与香农熵是一致的,熵值越大,说明系统的规律性越差,熵值接近于 0 时,系统的规律性很好.

对同一个符号序列,正向划分原始数据与反向划分就得到两个不同的符号序列,分别求这两个序

列中字出现的概率,然后计算两种情况下字的概率的欧氏范数,就可计算出时间不可逆转性.时间序列不可逆转性定义为[4]

$$T_{fb}=\sqrt{\sum_i (P_{f,i}-P_{b,i})^2} \tag{2}$$

其中 $P_{f,i}$ 和 $P_{b,i}$ 分别为前向序列中符号串与后向序列符号串的概率.时间序列不可逆转性就是对字进行计算,尤其适用于多维空间.

我们也可以利用 χ^2 统计量来计算前向符号序列与后向符号序列的差别,其效果与时间序列不可逆转性指标相同,其定义如下[4]

$$\chi^2=\sum_i \frac{(P_{f,i}-P_{b,i})^2}{P_{f,i}+P_{b,i}} \tag{3}$$

时间不可逆转性 T_{fb} 及 χ^2 统计量是基于传统的统计分析,若系统呈规律性变化,则 T_{fb} 与 χ^2 值趋近于0,而 T_{fb} 与 χ^2 值越大,则表明系统的动力学特性越复杂.

1.3 参数选择

在有的实验中,采集信号最多可达每秒几千个采样点,若采样点过大,则会导致连续多个点的符号相同.因此需要控制采样点不能太多,通常来说把连续的实验数据转化为符号序列后,将包含大量连续的相同符号.符号重复的频率过高将会导致原始数据的过采样,另外,采样时间如果大于采样定律规定的时间,将会发生混叠现象并且丢失信息量.它与延迟时间的选取类似,因此,混沌相空间重建中的延迟时间选取方法可以用于符号时间序列分析,在构造符号序列时减少符号冗余的方法就是增加符号间的时间间隔.通常的方法就是利用互信息极小值求取合理的时间延迟,互信息方程定义如下

$$I(\tau)=\sum p_{i,j}(\tau)\log_2\frac{p_{i,j}}{p_i p_j} \tag{4}$$

其中 τ 是应用于原始连续测量数据特定的时间延迟.

分割区间的大小 n 及标准长度 L 决定了出现的符号序列长度,序列长度由下式计算

$$N_{seq}=n^L \tag{5}$$

上述三个统计量值随着分割区间 n 与字标准长度 L 的变化而变化.对于特定的时间序列,应当选择适当的 n 与 L 才能更好地揭示系统的动力学特性.迄今为止,研究人员仍然没有从理论上找到一种有效的方法,但是 Finney 等[14] 根据实际经验发现,当修正的香农熵为最小值时,对应的 n 与 L 就是最佳的参数.Daw 等[7] 用时间不可逆转性指标 T_{fb} 来选择最佳参数,当此指标值为最大时,对应的 n,L 值就是最佳的参数值.这两种方法都可以用来作为 n 与 L 参数选择的标准.

1.4 算例验证

我们利用混沌研究中常用的 Henon 方程对以上介绍的方法进行了初步验证,所取的数据为方程中 x 变量的前3000个点,Henon 方程定义为

$$\begin{cases} x_{i+1}=y_i+1-ax_i^2 \\ y_{i+1}=bx_i \end{cases} \tag{6}$$

取方程中 $a=1.4, b=0.3$,这样就可以产生混沌序列,图5所示为Henon方程中 x 变量得到的3000个数据点中的前200点,并且比较了Henon方程中 x 变量的前向符号序列及后向符号序列,其中分割区间 $n=4$ 及标准长度 $L=3$.图6揭示了前向序列与后向序列的差异.图7所示为此数据的时间不可逆转性指标计算结果,对于试验数据可以通过计算其指标值的大小,来揭示不同数据背后的动力学特性.

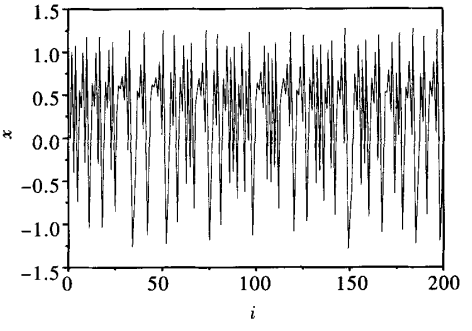

图5 Henon映射数据
Fig.5 Data of Henon map

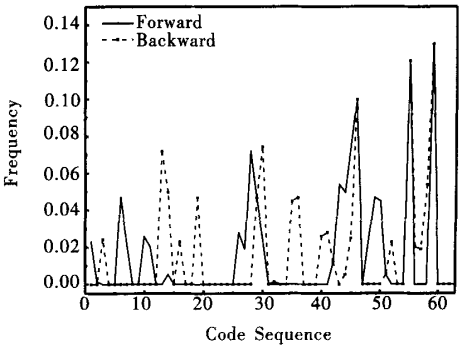

图6 前向序列与后向序列的比较
Fig.6 Comparison of forward and backward sequence

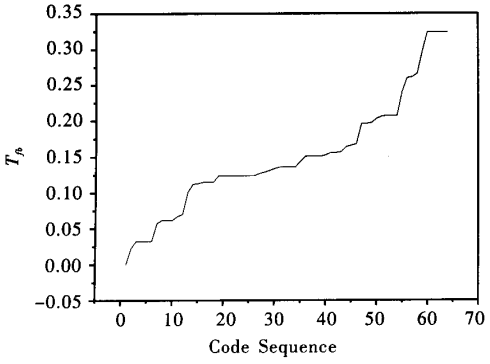

图 7　由 Henon 映射数据计算的时间不可逆转性结果
Fig. 7　Temporal irreversibility of Henon data

2　油水两相流流型表征

我们利用符号化时间序列分析方法分析了垂直上升管中油水两相流的 16 组不同流动工况下的实验数据,实验在大庆油田生产测井研究所垂直上升管中油水两相流流动环中进行,数据采集系统为电导式相关流量计[15],其含水率 K_w 实验范围为 51 % ～ 91 %,油水两相流总流量 Q_t 实验范围为 10 ～ 60(m^3/d),油相密度 ρ_o 为 0.82(g · cm^{-3}),水相密度 ρ_w 为 1.0(g · cm^{-3}),油水粘度比为 3.26. 所考察的油水两相流流型是内径为 18 mm 的垂直上升管中油水两相流流动特性.

首先我们利用修正的香农熵选取最佳参数,如图 8 所示,在流动工况条件 Q_t = 20(m^3/d),K_w = 91 % 情况下,确定的参数最佳值为 n = 3 及 L = 5. 表 1 给出了对 16 种流动工况油水两相流电导波动信号实验数据处理的全部结果.

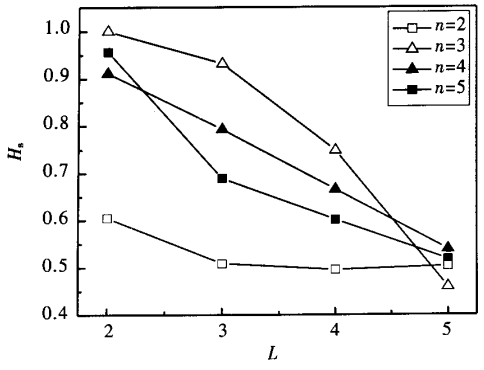

图 8　修正香农熵与符号化参数间关系
Fig. 8　Modified Shannon entropy *versus* symbolization parameters

表 1　实验数据最佳参数选择结果
Table 1　Optimization of symbolic parametersand n and L

Q_t/(m^3 · d^{-1})	K_w/%	n	L
10	91	5	5
20	51.5	5	5
20	81	3	5
20	91	3	5
30	61	5	5
30	71	2	5
30	81	5	5
40	51	3	5
40	51.5	5	4
40	71	5	5
40	81	5	5
50	51	3	5
50	61	3	4
50	71	2	4
60	51	3	5
60	61	3	5

图 9,10 为对全部 16 种流动工况符号时间序列分析结果,从图中我们可以看出,当含水率 K_w 在 61 % ～ 91 % 之间变化时,时间不可逆转性指标 T_{fb} 及 χ^2 统计量随总流量 Q_t 变化很小,当含水率 K_w 为 51 % 时,两个指标都随着总流量的变化而突

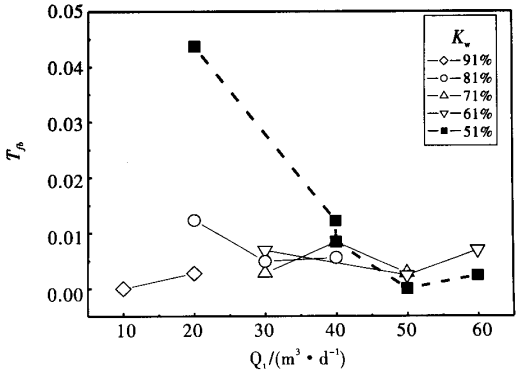

图 9　时间不可逆转性随总流量及含水率的变化
Fig. 9　Temporal irreversibility versus total flowrate Q_t and water cut K_w

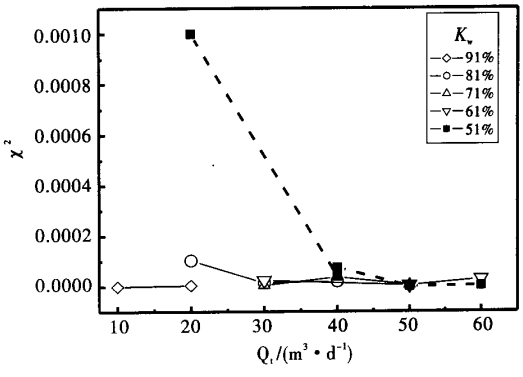

图 10　统计量随总流量及含水率的变化
Fig. 10　Chi − square statistics versus total flowrate Q_t and water cut K_w

变,这是因为当含水率为51%时,这种流动状态属于油包水及水包油流型共存的不稳定过渡流型[15].结果表明符号时间序列计算的统计量对油水两相流流型变化敏感.

3 结论

符号化时间序列分析方法是近年来新出现的一种数据分析方法,它在许多方面存在很大的优势:计算简单快捷、有效抑制噪声干扰,降低仪器抗干扰的要求等.随着数据采集速度的加快及动力学系统的日趋复杂,符号化方法将会是一个重要的信息分析工具.随着这种方法的日趋完善,此方法将会被广泛地应用于各个工程应用领域.

参 考 文 献

1 Crutchfield JP, Packard NH. Symbolic dynamics of noisy chaos. *Physica D*, 1983, 7: 201~223

2 Tang XZ, Tracy ER, Boozer AD, Brown R. Symbol sequence statistics in noisy chaotic signal reconstruction. *Physical Review E*, 1995, 51(5): 3871~3889

3 Tang XZ, Tracy ER, Brown R. Symbol statistics and spatio - temporal systems. *Physica D*, 1997, 102: 253~261

4 Daw CS, Finney CEA, Tracy ER. Symbolic statistics: A new tool for understanding multiphase flow phenomena. ASME International Congress & Exposition. Anaheim, California, USA, November, 1998: 15~20

5 Lehrman M, Rechester AB, White RB. Symbolic analysis of chaotic signals and turbulent fluctuations. *Physical Review Letters*, 1997, 78 (1): 54~57

6 Godelle J, Letellier C. Symbolic sequence statistical analysis for free liquid jets. *Physical Review E*, 2000, 62 (6): 7973~7981

7 Daw CS, Finney CEA, Kennel MB. Symbolic approach for measuring temporal 'irreversibility'. *Physical Review E*, 2000, 62 (2): 1912~1921

8 Daw CS, Finney CEA, Tracy ER. A review of symbolic analysis of experimental data. *Review of Scientific Instruments*, 2003, 74(2): 915~930

9 Govier GW, Sullivan GA, Wood RK. The Upward Vertical Flow of Oil - Water Mixtures. *The Canadian Journal of Chemical Engineering*, 1961: 67~75

10 Vigneaux PG, Chenais P, Hulin JP. Liquid - Liquid Flows in an Inclined Pipes. *AICHEJ*, 1988, 34(5): 781~789

11 Zavareh F, Hill AD, Podoa AL. Flow Regimes in Vertical and Oil/Water Flown Pipes. 63rd Annual Technical Conference and Exhibition of the Society of Petroleum Engineers. *Houston TX Society of Petroleum Engineers*, 1988, (2-5): 361~318

12 Flores JP, Chen XT, Sarica Cem, Brill JP. Characterization of oil - water flow patterns in vertical and deviated wells. SPE Annual Technical Conference and Exhibition in San Antonio. *Texas Society of Petroleum Engineers*, 1997, (5-8): 601~610

13 Finney CEA, Nguyen K, Daw CS, Halow JS. Symbolic sequence statistics for monitoring fluidization. International Mechanical Engineering Congress & Exposition. Anaheim, California, USA, November, 1998: 15~20

14 Finney CEA, Green JB, Daw CS. Symbolic time series analysis of engine combustion measurement. *SAE Paper*, 1998: 980624

15 Jin ND, Nie XB, Ren YY, Liu XB. Characterization of oil/water two - phase flow patterns based on nonlinear time series analysis. *Flow Measurement and Instrumentation*, 2003, 14(4,5): 169~175

STUDY ON THE ANALYSIS METHOD OF NONLINEAR SYMBOLIC TIME SERIES*

Jin Ningde　Li Weibo

(*School of Electrical Engineering and Automation, Tianjin University, Tianjin* 300072, *China*)

Abstract Symbolic time series analysis is a new tool for analysis of experimental data, which has been successfully applied in many fields, bacause it can increase the efficiency of finding and quantifying information from dynamic systems, and reduce sensitivity to measurement noise. In this paper, first we proposed a method, which used three statistical quantities to characterize the symbolic time series, and has been validated by Henon equation. Then we applied the method to analyze the experimental data of oil/water two-phase flow in vertical upwards pipes. The analysis showed that the calculated statistical quantities from the symbolic time series were sensitive to the transitional flow pattern variations of oil/water two phase flow.

Key words symbolic time series, temporal irreversibility, chi-square statistics

Received 03 July 2004, revised 02 September 2004

* The project supported by the National Natural Science Foundation of China(60374041)and the Scientific Research Foundation for the Returned Oversear Chinese Scholars, State Education Ministry of China.