

基于 K-Means 聚类的瓦斯浓度预测

穆文瑜¹, 李茹^{1,2}

(1. 山西大学 计算机与信息技术学院, 太原 030006; 2. 山西大学 计算智能与中文信息处理教育部重点实验室, 太原 030006)

(wenyu1069@126.com)

摘要:提出一种基于 K-Means 聚类的非线性时间序列预测模型。利用混沌时间序列短期可以预测的特点,对选取的某两处煤矿构建了瓦斯浓度预测模型。采用关联积分方法确定相空间时间延迟 τ 和相空间嵌入维数 m 。然后在重构相空间中,运用基于 K-Means 聚类的加权一阶局域法构建煤矿瓦斯浓度的预测模型。结果表明:在预测间隔 1 min 的数据时,使用 200 个连续的数据进行训练,预测效果较好,误差达到最小值 0.034 1;在预测间隔多分钟的数据时,使用 200 个 15 min 间隔的数据进行训练,预测效果较好,误差达到最小值 0.043 7,可见该瓦斯浓度时序在间隔 15 min 后又恢复了初始的混沌性。

关键词:瓦斯浓度;相空间;时间延迟;嵌入维;加权一阶局域法

中图分类号: TP391.8; TP311.13 **文献标志码:** A

Prediction of gas concentration based on K-Means clustering

MU Wen-yu¹, LI Ru^{1,2}

(1. School of Computer and Information Technology, Shanxi University, Taiyuan Shanxi 030006, China;

2. Computer Intelligent and Chinese Information Processing of the Ministry Education Key Laboratory Built Together by Province and Department, Shanxi University, Taiyuan Shanxi 030006, China)

Abstract: A prediction model of non-linear time series based on K-Means clustering was proposed. Using the ability of short-term predicting for chaotic time series, the paper constructed a gas concentration prediction model for certain coal mines. Correlation integral method was used to determine the time delay τ and dimension m . After the phase space was reconstructed, the weighted one-rank local-region method based on K-Means clustering was used to construct prediction model. The experimental results show that, if next one minute data will be forecasted, it is more appropriate to use a continuous 200 training data to determine parameters τ and m for predicting better results, the error reaches 0.034 1; if next few minutes data will be forecasted, it is more appropriate to use a 200 training data with 15 minutes intervals for predicting better results, the error is 0.043 7. It shows that the timing of the gas concentration restores the initial chaos after 15 minutes.

Key words: gas concentration; phase space; time delay; embedded dimension; weighted one-rank local-region method

0 引言

我国煤炭生产主要为地下作业,由于煤炭赋存的地质条件复杂多变,经常受到瓦斯、水、火、粉尘和顶板等自然灾害的威胁,加之抗灾能力较弱,煤矿事故时有发生,其中瓦斯事故尤为严重。瓦斯事故的发生,不仅使国家的生命财产遭受重大损失,而且影响煤炭生产正常进行。瓦斯浓度随时间动态发展是一个复杂的非线性过程,即瓦斯浓度时序是非线性序列,采用传统方法分析其影响因素来预测是不易实现的。研究采用一些结构比较简单,但却能反映非线性本质特性的模型来表示复杂的非线性系统,是很有必要的。

在非线性时序建模及预测技术的研究过程中,国内外学者做了大量研究工作,提出了很多有价值的可应用的模型和建模方法。1987 年 Lapedes 等人^[1]首先应用神经网络对 Mackey-Glass 时间序列进行预测。小波网络是结合小波变换的时频局域化性质与传统人工神经网络的自学习功能而形成的,用其所建预测模型可以取得更好的预测效果^[2]。支持向量机(Support Vector Machine, SVM)是 20 世纪 90 年代中期

提出的一种机器学习算法^[3],该方法具有自学习自调整模型的特点,能对各种非线性系统产生较好的预测效果。

自美国气象学家 Lorenz^[4]在 1963 年首次发现混沌以来,混沌现象一直是国内外研究的热点。混沌是非线性耗散系统中产生的一种貌似随机的不规则现象,它具有对初值极端敏感性和系统长期行为不可预测性^[5]。在非线性的领域,从单一的时间序列获取系统的动态信息,分析混沌时间序列的前提是相空间重构,即由低维时间序列重构出一个多维的确定性相空间。相空间重构技术又称嵌入技术,它是由 Packard 等人^[6]首次提出并经 Takens^[7]数学完善的。该技术用系统的一个或几个元素的时间序列来重构相空间,所得的新系统在拓扑结构和概率特性等本质特征上与原系统保持一致,可通过分析研究新系统的特征来确定原系统特征。在混沌时间序列预测方法中,局域线性化方法是最通用的有效方法。

本文提出了一种基于 K-Means 聚类的加权一阶局域预测模型, K-Means 算法具有简单、快速并且能够有效地处理大量数据库的优点。从混沌系统的本质特性出发考查煤矿瓦斯浓度随时间的走向对初值的极端敏感性等特征,从而证明瓦斯浓

收稿日期:2010-08-30;修回日期:2010-10-19。

基金项目:太原市科技局专项(08121005);山西省高等学校中青年拔尖人才基金资助项目(2007)。

作者简介:穆文瑜(1987-),女,山西吕梁人,硕士研究生,主要研究方向:数据挖掘;李茹(1963-),女,山西太原人,教授,主要研究方向:智能信息处理。

度时间序列是一种混沌序列,进一步采用基于 K-Means 聚类的加权一阶局域预测模型进行预测。实验结果表明,预测效果较好。该方法克服了传统的加权一阶局域法中确定离中心点最近的点集的标准选取的困难。利用 K-Means 聚类法,可以比较精确地找到离中心点最近的点集,即属于中心点所在的类的点集。而传统的加权一阶局域法没有明确给出在什么条件下就满足离中心点最近,没有最合适的衡量标准。

1 瓦斯浓度预测模型

1.1 相空间重构参数(时延与嵌入维数)的确定

先利用 Wolf 方法^[8]计算出 Lyapunov 指数^[9]来判别瓦斯浓度时间序列是否具有混沌性,然后确定参数。

利用非线性系统输出的部分混沌时间序列考查系统中奇异吸引子的方法是分析混沌时间序列的常用方法,目前广泛采用的是 Packard 等人^[6]提出的延迟坐标状态空间重构法。由 Takens^[7]定理证明,只要找到一个合适的嵌入维,即如果延迟坐标的维数 $m \geq 2d+1$ (d 为原系统的阶数),在这个嵌入维空间里可以把有规律的轨线(吸引子或奇异吸引子)恢复出来。在重构相空间中,时间延迟 τ 和嵌入维 m 的选取十分重要,其精度直接关系到相空间重构后描述奇异吸引子特征的不变量的准确度。

在选择时延参数 τ 时,对于无限长、没有噪声的数据序列,原则上对其没有限制,但实际上由于噪声和计算误差问题,相空间的特征量依赖于 τ 的选择:若 τ 太小,坐标相关性太强,而且相空间轨迹沿同一方向挤压,信息不容易泄露;若 τ 太大,在混沌和噪声情况下,导致某一时刻的动力学形态与后一时刻动力学形态变化剧烈,导致信息丢失。对于嵌入维 m 的选择,理论上只有下限要求。但实际上,如果 m 太大,会引入噪声,并增加额外的计算量。

1996 年, Kugiumtzis 提出了相空间重构的嵌入窗法^[10],指出时延 τ 的选取不应独立于嵌入维 m , 而应依赖于嵌入窗 $\tau_w = (m-1)\tau$, 并且要求 $\tau_w \geq \tau_p$, 这里 τ_p 为混沌系统的平均轨道周期。1999 年, Kim 等人提出了 C-C 方法^[11], 在使用关联积分的同时估计出时延与嵌入窗。该方法描述如下。

对瓦斯浓度时间序列 $X = \{x_i | i = 1, 2, \dots, N\}$, 以时延 τ ($\tau = t$), 嵌入维 m , 重构相空间 $X = \{X_i\} = \{X_1, X_2, \dots, X_M\}$, X_i 为相空间中的点, 其中 M 为相点个数, $M = N - (m-1)\tau$ 。

$$\begin{cases} X_1: (x_1, x_{1+\tau}, x_{1+2\tau}, \dots, x_{1+(m-1)\tau}) \\ X_2: (x_2, x_{2+\tau}, x_{2+2\tau}, \dots, x_{2+(m-1)\tau}) \\ \vdots \\ X_M: (x_M, x_{M+\tau}, x_{M+2\tau}, \dots, x_{M+(m-1)\tau}) \end{cases}$$

其中: m 称为嵌入相空间的维数, τ 称为时间延迟。

定义瓦斯浓度序列的关联积分:

$$C(m, N, r, t) = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} \theta(r - d_{ij}); r > 0 \quad (1)$$

其中: $d_{ij} = \|X_i - X_j\|_{(\infty)}$; 若 $x < 0$, $\theta(x) = 0$; 若 $x \geq 0$, $\theta(x) = 1$ 。

关联积分是个累积分布函数, 表示相空间中任意两点之间距离小于 r 的概率。这里点与点之间的距离用矢量之差的无穷范数表示。

定义检验统计量为:

$$S(m, N, r, t) = C(m, N, r, t) - C^m(1, N, r, t) \quad (2)$$

选择最大和最小的两个半径 r , 定义差量:

$$\Delta S(m, t) = \max\{S(m, N, r_j, t)\} - \min\{S(m, N, r_j, t)\} \quad (3)$$

$\Delta S(m, t)$ 度量了 $S(m, N, r, t) \sim t$ 对所有半径 r 的最大偏差。

根据 BDS (Brock-Dechert-Scheinkman) 统计结论, 并结合实验可以得到 N , m 和 r 的合理估计, 这里取 $m = 2, 3, 4, 5$, $r_i = 0.5 \times i\sigma$, $\sigma = \text{std}(x)$ (σ 为时间序列的标准差), $i = 1, 2, 3, 4$, $N = 100$ 。计算:

$$\bar{S}(t) = \frac{1}{16} \sum_{m=2}^5 \sum_{i=1}^4 S(m, r_i, t) \quad (4)$$

$$\Delta \bar{S}(t) = \frac{1}{4} \sum_{m=2}^5 \Delta S(m, t) \quad (5)$$

$\Delta \bar{S}(t)$ 的第一个局部极小点作为最佳时延 τ 。此外, 定义指标:

$$S_{\text{cur}}(t) = \Delta \bar{S}(t) + |\bar{S}(t)| \quad (6)$$

$S_{\text{cur}}(t)$ 的全局最小点即可获得嵌入窗 τ_w 。由最大时间窗口 $\tau_w = (m-1)\tau$, 得出 $m = \tau_w/\tau + 1$ 。

1.2 基于 K-Means 聚类的加权一阶局域预测模型

嵌入维数 m 和时间延迟 τ 选好后, 就可以对一个混沌时间序列进行预测, 目前混沌时间序列的预测方法主要包括全局预测法、局域预测法、自适应预测法和局域自适应预测法。数值实验结果表明预测效果从低到高为: 全域法、局域法、加权局域法, 其中一阶局域法效果好于零阶局域法, 因此选用基于 K-Means 聚类^[12]的加权一阶局域法^[13]来对瓦斯浓度进行预测。

利用已经确定好的嵌入维数 m 和时间延迟 τ 进行相空间重构, 对于瓦斯浓度时间序列 (x_1, x_2, \dots, x_N) 来说, 构造一个 m 维相空间 $\{X_1, X_2, \dots, X_M\}$, 其中 M 为相点个数, $M = N - (m-1)\tau$, 这里 m 称为嵌入相空间的维数, τ 称为时间延迟。用嵌入相空间就代替了状态空间, 将这 M 个点依次连接, 就形成了该 m 维嵌入相空间的轨道。

把相空间轨迹的最后一相点 X_k 作为中心点, 把离中心点最近的若干轨迹点作为相关点, 要选离中心点最近的轨迹点, 传统的方法是确定一个标准, 即半径 r , 满足距中心点小于 r 的相点均作为研究对象, 但是这里的 r 值如何选取, 只是一个经验值, 存在一定的偏差性, 受主观因素的影响较大。采用 K-Means 聚类算法^[12]不需要确定这一标准。K-Means 是一种聚簇算法, 也是一种最简单的无监督学习算法之一, 也称为动态聚类或逐步聚类方法, 基本思想是开始先粗略地分类, 然后按照某种最优的原则修改不合理的分类, 直至类分得比较合理为止, 形成最终分类结果。基本步骤如下:

步骤 1 确定 k 值以及初始化聚类中心, 选择 k 个初始凝聚点, 作为欲形成类的中心点;

步骤 2 计算每一个观测到 k 个初始凝聚点的距离, 将每个观测和最近的凝聚点分到一组, 形成 k 个初始分类;

步骤 3 计算初始分类的重心(或均值), 作为新的凝聚点, 重新计算每个观测到初始分类重心的距离, 将每个观测和最近的凝聚点分为一组;

步骤 4 重复步骤 2 和步骤 3 直至初始分类的重心或均值没有明显变化为止。

对于以上找离中心点最近的轨迹点的情况来说,按逐个修改凝聚点的方法聚类较适合,因为初始中心点是确定的。

计算出各点到中心相点 X_k 之间的欧氏距离,找出 X_k 的局域参考向量集 $X_{k_i} (i = 1, 2, \dots, q)$ 及点 X_{k_i} 到 X_k 的距离为 d_i , 设 d_m 是 d_i 中的最小值, 定义点 X_{k_i} 的权值为:

$$P_i = \frac{\exp(-c(d_i - d_m))}{\sum_{i=1}^q \exp(-c(d_i - d_m))}; i = 1, 2, \dots, q \quad (7)$$

其中 c 为系数, 一般取 $c = 1$, 则一阶局域线性拟合为:

$$X_{k_{i+1}} = ae + bX_{k_i}; i = 1, 2, \dots, q, e = (1, 1, \dots, 1)^T \quad (8)$$

其中 a, b 为待定系数。当嵌入维数 $m = 1$ 时 ($m > 1$ 的情况类似), 为了使平方误差达到最小, 使预测模型与实验数据达到最佳的拟合, 应用加权最小二乘法有:

$$J = \sum_{i=1}^q P_i (X_{k_{i+1}} - ae - bX_{k_i})^2 \quad (9)$$

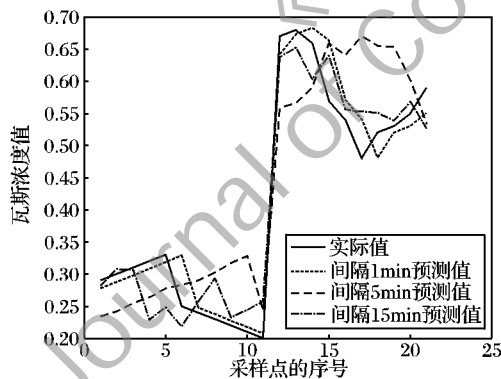
则 J 为待定系数 a, b 的函数, 两边分别对 a, b 求偏导并令其为零, 整理得:

$$\begin{cases} \sum_{i=1}^q P_i (X_{k_{i+1}} - a - bX_{k_i}) = 0 \\ \sum_{i=1}^q P_i (X_{k_{i+1}} - a - bX_{k_i}) X_{k_i} = 0 \end{cases} \quad (10)$$

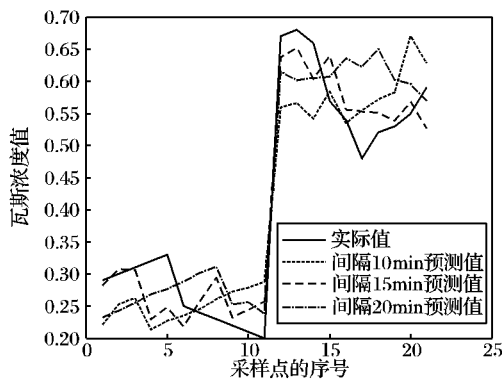
化简式(10)得到关于系数 a, b 的方程组为:

$$\begin{cases} a + b \sum_{i=1}^q P_i X_{k_i} = \sum_{i=1}^q P_i X_{k_{i+1}} \\ a \sum_{i=1}^q P_i X_{k_i} + b \sum_{i=1}^q P_i X_{k_i}^2 = \sum_{i=1}^q P_i X_{k_i} X_{k_{i+1}} \end{cases} \quad (11)$$

利用线性回归算法解方程组(11)则可得到 a, b , 然后代入式(8), 即得预测公式。

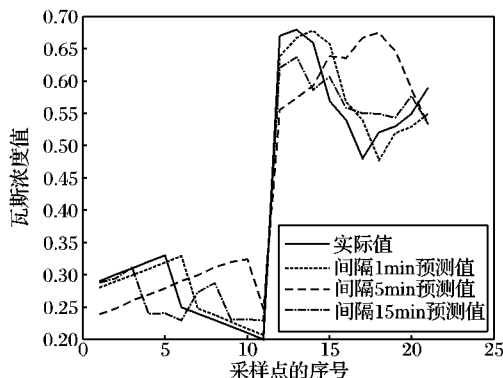


(a) 间隔为1、5、15min预测值与实际值的对比

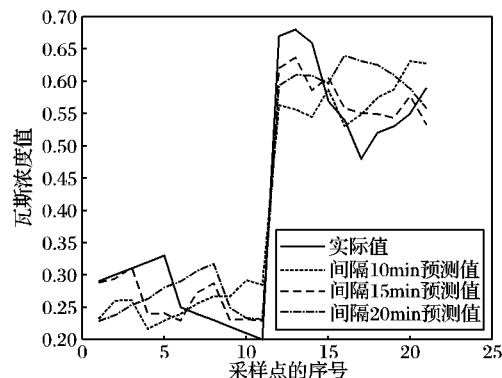


(b) 间隔为10、15、20min预测值与实际值的对比

图1 $N = 100$ 时不同时间间隔预测值对比



(a) 间隔为1、5、15min预测值与实际值的对比



(b) 间隔为10、15、20min预测值与实际值的对比

图2 $N = 200$ 时不同时间间隔预测值对比

根据预测公式进行预测, 显然参数向量集为 $X_{k_i} (i = 1, 2, \dots, q)$ 的一步预测为 $X_{k_{i+1}} (i = 1, 2, \dots, q)$ 。

因为使用的关系式(8), 阶数为1 所以称为一阶近似预测。加权一阶局域法能够保证每一个相点的实际值和计算值的误差达到最小, 同时, 轨迹点的权值越大说明距离中心点越近, 即中心点的下一步的轨迹越接近于该点的轨迹, 于是加入权值, 体现出轨迹点对中心点下一步走向的支持程度。这样能够更准确地预测出中心点的下一步。K-Means 聚类算法又较精确地找到了离中心点最近的点集, 所以 K-Means 聚类算法在加权一阶局域法的实现中起到了关键的作用。

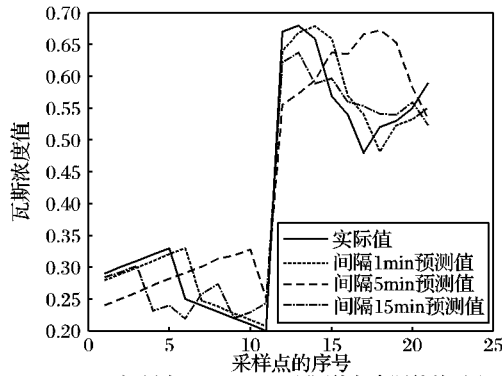
2 仿真实验及结果分析

本文实验选取的煤矿位于某城寨村境内, 资源储量丰富, 煤炭发热量高, 煤层气含量少, 属低瓦斯安全矿井, 采集了容量约9 GB 的数据进行了大量的实验, 并随机地选取了若干瓦斯浓度监测数据做了实验分析; 另外, 又选取了另一个煤矿, 亦属于高瓦斯煤矿, 同样, 采集了容量约9.8 GB 的数据进行实验, 并随机地选取了若干瓦斯浓度监测数据做了实验分析。这样, 代表性地选取了一个低瓦斯和一个高瓦斯煤矿的数据进行实验, 以证明预测模型的通用性, 仿真实验主要以各矿点的传感器类型为1 (即瓦斯浓度传感器) 的数据进行实验。分别以100, 200, 300 个连续数据为一组进行训练预测, 并且每隔1 min 采一次数据。训练以时间间隔1 min、5 min、10 min、15 min、20 min 进行实验, 用 Wolf 方法^[8] 计算其最大 Lyapunov 指数, 全部大于0, 可知其全为混沌时间序列。然后采用关联积分方法确定最佳时延与嵌入维数。在时间间隔不同的情况下, 瓦斯浓度值的预测结果与其实际监测值的部分对比图, 如图1~3 所示。

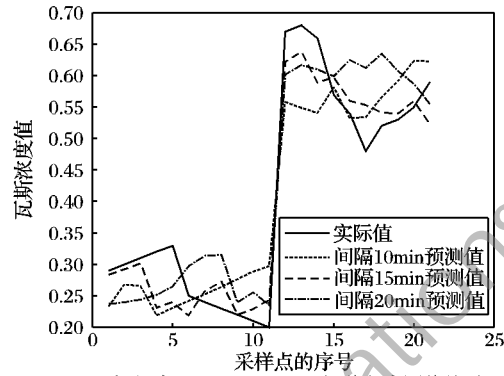
3 误差分析

为衡量不同情况对预测结果的影响,用预测值与实际值的均方差作为评判预测效果的一个指标:

$$ESS = \left[\frac{1}{L} \sum_{i=1}^L (x(i) - \hat{x}(i))^2 \right]^{\frac{1}{2}} \quad (13)$$



(a) 间隔为1、5、15min预测值与实际值的对比



(b) 间隔为10、15、20min预测值与实际值的对比

图3 $N=300$ 时不同时间间隔预测值对比

表1 仿真结果的误差分析 ($N=100$)

训练数据 长度 N	时间间隔/min				
	1	5	10	15	20
100	0.0349	0.0908	0.0737	0.0469	0.0687
200	0.0341	0.0901	0.0708	0.0447	0.0688
300	0.0342	0.0902	0.0711	0.0449	0.0680

从表1中可以看出,对于所测瓦斯浓度时间序列来说,时间间隔为1min时,预测效果最佳,并且当训练数据长度从100增大到200时,ESS减小;当 N 从200增大到300时,ESS又增大;因此,训练数据为200个时的ESS最小。另外,当增大时间间隔时噪声可能占主导地位,其混沌性发生变化,使得均方ESS变大;当时间间隔增大到15min时,ESS明显比5、10、20min的小,此时吸引子又能够很好地展开,再次恢复原系统的动力学特性,其中训练数据为200个时的ESS最小。因此,在预测间隔1min时,最佳选择是训练数据 $N=200$;在预测更远数据时,最佳选择仍是训练数据 $N=200$,而时间间隔应为15min,即为了让矿工在瓦斯灾难发生前有充足时间做逃离准备,需要预测更远时刻的瓦斯浓度值的时候,较宜采用间隔15min的时序数据进行预测估计。

4 结语

瓦斯浓度是影响煤矿安全生产的一个重要指标,其及时准确的预测能为煤矿的安全生产提供重要保障。本文在前期成果的基础上又提出基于K-Means聚类的加权一阶局域预测模型,并做了进一步的实验。运用相空间重构理论对未来某一时刻的瓦斯浓度进行预测,对选取的某两处煤矿监测到的瓦斯浓度时间序列进行了相空间重构,通过关联积分方法确定最优时延和嵌入维,然后用基于K-Means聚类的加权一阶局域法对下一时刻的瓦斯浓度进行预测。实验分别以100、200、300个训练数据进行预测,不仅预测下一分钟的数据,还预测接下来5、10、15、20min的数据。实验表明,对于该煤矿,当时间序列的长度 N 为200时,可以比较准确地预测未来

其中: $x(i)$ 表示实际值, $\hat{x}(i)$ 表示预测值。ESS小,说明预测值偏离实际值的程度小,预测效果较好;ESS大,说明预测值偏离实际值的程度大,预测效果较差。实验结果表明,影响预测效果的主要因素有被预测序列的长度、序列选取的时间间隔等。改变时间间隔和训练数据长度其均方差ESS值如表1所示。

1min的瓦斯浓度数据,误差可达0.0341。当希望预测未来几分钟的瓦斯浓度值时,可以选择以200个数据为训练对象,间隔为15min,误差可达0.0435,这样工作人员可以在发生瓦斯爆炸前迅速撤离,为保障其生命安全提供充足的时间。

参考文献:

- [1] LAPADES A, FARBER R. Nonlinear signal processing using neural networks: Prediction and system modeling[R]. Los Alamos, NM: Los Alamos National Laboratory, 1987.
- [2] 谷松, 张振文, 李国军. 矿井瓦斯涌出量的灰色小波神经网络预测模型[J]. 煤炭技术, 2009, 28(10): 123-125.
- [3] 董辉, 傅鹤林, 冷伍明. 支持向量机的时间序列回归与预测[J]. 系统仿真学报, 2006, 18(7): 1785-1788.
- [4] LORENZ E N. Deterministic nonperiodic flow[J]. Journal of Atmospheric Science, 1963, 20(2): 130-141.
- [5] 黄润生. 混沌及其应用[M]. 武汉: 武汉大学出版社, 2001.
- [6] PACKARD N H, GRUTCHFIELD J P, FARMER J D, et al. Geometry from a time series[J]. Physical Review Letters, 1980, 45(9): 712-715.
- [7] TAKENS F. Detecting strange attractors in turbulence[C]// Dynamical Systems and Turbulence, LNM 898. New York: Springer, 1981: 366-381.
- [8] 王妍, 徐伟. 基于时间序列的相空间重构算法及验证(二)[J]. 山东大学学报, 2005, 35(6): 91-92.
- [9] 马军海, 陈予恕, 季进臣. 三种动力系统Lyapunov指数的比较[J]. 天津大学学报, 1999, 32(2): 190-193.
- [10] 吕金虎, 陆君安, 陈士华. 混沌时间序列分析及其应用[M]. 武汉: 武汉大学出版社, 2002.
- [11] KIM H S, EYKHOLT R, SALAS J D. Nonlinear dynamics, delay times and embedding windows[J]. Physica D: Nonlinear Phenomena, 1999, 127(1/2): 48-60.
- [12] 赵国富, 周雪芹. 基于聚类的空间数据挖掘系统设计与实现[J]. 山东理工大学学报, 2005, 19(6): 41-44.
- [13] 程健, 白静宜, 钱建生, 等. 基于混沌时间序列的煤矿瓦斯浓度短期预测[J]. 中国矿业大学学报, 2008, 37(2): 232-234.