文章编号: 1001-0920 (2011) 10-1525-05

基于云模型的时间序列分段聚合近似方法

李海林, 郭崇慧

(大连理工大学 系统工程研究所，辽宁 大连 116024)

摘 要: 针对时间序列数据的高维特性, 提出一种基于云模型的时间序列分段聚合近似方法. 利用云模型的熵评判分段聚合后各子序列的数据稳定性, 选取稳定性最弱的子序列再分段聚合, 最终得到云模型序列, 同时给出了云模型序列的相似性度量. 该方法对时间序列能够有效降维, 并能够自适应地识别和描述其基本特征. 实验结果表明, 数据压缩较大时, 所提出方法能够较好地保证近似的准确性, 并提高时间序列数据挖掘的效率.

关键词: 时间序列；云模型；相似性；分段聚合近似

中图分类号: TP18 **文献标识码:** A

Piecewise aggregate approximation method based on cloud model for time series

LI Hai-lin, GUO Chong-hui

(Institute of Systems Engineering，Dalian University of Technology，Dalian 116024，China．Correspondent：LI Hai-lin，E-mail：hailin@mail.dlut.edu.cn)

Abstract: This paper proposes a technique of piecewise aggregate approximation based on cloud model to resolve the high dimensionality of time series. The entropy of cloud model is used to evaluate the stability of data points in a subsequence and choose the subsequence with lower stability to further divide so that a series of cloud models can be obtained to approximate time series. The similarity between two cloud model series is calculated. The proposed method can reduce the dimensionality, and also can adaptively recognize and represent the essential features of time series. The results of experiments indicate that the proposed method can guarantee the accuracy of similarity and improve the efficiency of time series data mining under larger compress ratio.

Key words: time series；cloud model；similarity；piecewise aggregate approximation

1 引 言

时间序列是一类与时间相关且具有高维特性的数据对象, 在金融、气象、经济等领域普遍存在. 数据挖掘技术可以从时间序列数据中发现有用的潜在的信息和知识, 为生产实践提供有利的决策支持, 即为时间序列数据挖掘 (TSDM)[1-2]. 国外许多学者开展了该领域的相关研究, 国内起步则相对较晚, 一些高校虽然已经开展了一些相关研究, 但总体上缺乏一定的组织性, 而且研究成果相对于国外也略显不足. 直接对原始数据进行挖掘, 其效率较低, 时间序列的高维特性始终是时间序列数据挖掘的瓶颈, 因此数据约减和降维便成为当前研究时间序列数据挖掘的首要任务和重要课题. 目前已有不少关于的时间序列数据约简和降维技术, 如离散傅里叶变换 (DFT)[3], 离散小波变换 (DWT)[4], 奇异值分解 (SVD)[5], 分段聚合近似 (PAA)[6]和分段线性近似 (PLR)[7]等. 不同方式的降维技术需要有各自的相似性度量方法, 如 DFT 和 DWT 各自系数之间的度量以及基于 PAA 的符号化表示方法 (SAX)[8]的相似性度量等.

根据时间序列基本特征的重要性, 本文提出一种基于云模型的时间序列分段聚合近似方法. 利用云模型的熵来评价分段序列数据的稳定性, 自适应地对时间序列的基本特征进行识别和表示, 同时给出了云模型相似性计算方法来度量时间序列降维后的关系. 实验结果表明, 数据压缩较大时, 该方法能够较好地保证近似的准确性, 并能够提高数据挖掘的效率.

收稿日期: 2010-06-02；**修回日期:** 2010-10-29.

基金项目: 国家自然科学基金项目(10571018, 70871015)；国家 863 计划项目(2008AA04Z107).

作者简介: 李海林(1982−), 男, 博士生, 从事人工智能、时间序列数据挖掘等研究；郭崇慧(1973−), 男, 教授, 博士生导师, 从事数据挖掘、决策分析等研究.

2 云模型和经典分段聚合近似表示

2.1 云 模 型

云模型[9]是定性概念与定量表示之间相互转化的一种模型, 主要反映了人类知识中概念的模糊性和随机性. 其中, 以正态云为代表的云模型已经在知识发现[10]、智能控制[11]以及决策分析[12]等领域得到了广泛的应用并取得了良好的效果. 下面简要介绍云模型的相关理论知识[9].

定义 1 设 U 为定量数据论域, C 为 U 上的定性概念, 若定量值 $x \in U$, 且 x 是概念 C 的一次随机实现, x 对于 C 的确定度为 $y = \mu_C(x)$ 是具有稳定倾向的随机数, 则云滴 (x, y) 在论域 U 上的分布称为云.

定义 2 由 3 个参数 $(\mathrm{E}x, \mathrm{E}n, \mathrm{H}e)$ 表示云的数字特征的模型称为云模型 $C(\mathrm{E}x, \mathrm{E}n, \mathrm{H}e)$. $\mathrm{E}x$ 表示云滴在论域空间分布的期望值; 熵 $\mathrm{E}n$ 表示定性概念的不确定性度量, 反映了云滴的离散程度; 超熵 $\mathrm{H}e$ 是熵的不确定性度量, 可以用来描述云的厚度.

定义 3 若 x 满足 $x \sim N(\mathrm{E}x, \mathrm{E}n'^2)$, 其中 $\mathrm{E}n' \sim N(\mathrm{E}n, \mathrm{H}e^2)$, 对于定性概念 C 的确定度满足

$$\mu_C(x) = \mathrm{e}^{\frac{-(x-\mathrm{E}x)^2}{2\mathrm{E}n'^2}}, \tag{1}$$

则 x 在论域 U 上的分布称为正态云.

正向正态云发生器 $\mathrm{cloud}(\mathrm{E}x, \mathrm{E}n, \mathrm{H}e, n)$ 和逆向正态云发生器 $\mathrm{back\_cloud}(X)$ 分别可以实现定性概念与定量数据之间的映射过程[9]. 特别地, 正向正态云发生器产生的云滴对概念的贡献不同, 对定性概念有重要贡献的云滴主要落在区间 $[\mathrm{E}x - 3\mathrm{E}n, \mathrm{E}x + 3\mathrm{E}n]$, 区间外的云滴贡献可以忽略, 这就是正态云的“$3\mathrm{E}n$”规则.

2.2 经典分段聚合近似表示

定义 4 (经典 PAA) 将时间序列 $Q = \{q_1, q_2, \cdots, q_m\}$ 用另外一个序列 $\bar{Q}$ 表示, 有 $\bar{Q} = \{\bar{q_1}, \bar{q_2}, \cdots, \bar{q_w}\}$, 其中 $w < m$, 且 m 被 w 整除. 令 $k = m/w$, 则 $\bar{Q}$ 中元素 $\bar{q_i}$ 满足

$$\bar{q_i} = \frac{1}{k} \sum_{j=k*(i-1)+1}^{k*i} q_j,\ 1 \leqslant i \leqslant w. \tag{2}$$

PAA 将长度为 m 的时间序列 Q 转变成长度 w 的序列 $\bar{Q}$, 进而实现了从高维到低维的降维过程. 降维后的序列可以粗糙地表示原时间序列的基本形态, 反映了时间序列的整体变化趋势, 但容易忽视局部数据的分布形态.

定义 5 时间序列数据点 $q_1, q_2, \cdots, q_m$ 服从正态分布, 若将数据分布空间按等概率划分 alphabet_size 个区域, 每个区域用 1 个字符来表示, 则把序列 $\bar{Q}$ 中各个数据点用所在区域的字符来表示的过程称为 SAX.

基于 PAA 方法的 SAX 在时间序列数据挖掘中发挥了重要的作用, 具有较大的影响力. Lin 等人[8]对 PAA 方法得到的序列 $\bar{Q}$ 按等概率分布实现了符号化表示, 并提供了基于 SAX 的时间序列相似度计算方法.

利用 SAX 方法, 由 PAA 得到的序列分别被 alphabet_size (如 $\mathrm{alphabet} = 4$) 个符号 (如 A,B,C,D) 表示成字符串, 其具体相似度计算方法可参阅文献[8]. 从转换后的数据分布来看, PAA 和 SAX 仅反映了原时间序列的整体形态变化趋势, 却忽视了各个子序列的局部变化趋势.

3 云模型分段聚合近似及相似性度量

为了更好地进行分段聚合近似和相似度计算, 提出一种基于云模型的时间序列分段聚合近似方法. 首先利用云模型中的熵来评价分段聚合近似的稳定性, 得到稳定性较好的云模型序列. 同时, 给出云模型的相似度计算方法, 以衡量云模型序列之间的相似性.

3.1 云分段聚合近似

时间序列的云模型表示过程实际上是一种从高维到低维的转化过程, 其主要步骤如下:

Step 1: 对当前的各个分段序列数据进行云模型表示.

Step 2: 利用各个云模型的熵来评价所在子序列的数据稳定性, 选取稳定性最差 (熵最小) 的子序列 (记为 $Q(i_0 : j_0)$) 再进一步分段聚合.

Step 3: 在子序列 $Q(i_0 : j_0)$ 中找到一个数据点作为关键点 $q_k, i_0 < k < j_0$, 该关键点 q_k 能使被它分开的两个子序列 ($Q(i_0 : k)$ 和 $Q(k : j_0)$) 的云模型熵之和与子序列 $Q(i_0 : j_0)$ 的云模型熵之间的差值最大, 同时删除子序列 $Q(i_0 : j_0)$, 记录子序列 $Q(i_0 : k)$ 和 $Q(k : j_0)$.

Step 4: 重复 Step 1∼Step 3, 到满足停止条件为止.

云分段聚合近似算法 (CPAA) 如下:

输入: 时间序列 $Q = \{q_1, q_2, \cdots, q_m\}$, 降维后的维度 w;

输出: 云模型序列 $C_q = \{c_{q_1}, c_{q_2}, \cdots, c_{q_w}\}$, 其中 $c_{q_i} = [\mathrm{E}x_i, \mathrm{E}n_i, \mathrm{H}e_i]$.

Step 1: 初始化数据. 用矩阵 $V_{w\times 5}$ 记录聚合分段信息, $V(k, 1:3)$ 记录第 k 个子序列云模型的 3 个数字特征, $V(k, 4:5)$ 记录第 k 个子序列云模型在原时间序列的始末位置. k 记录当前子序列个数, 初始设置 k =1. $V(k, 1:3) = \mathrm{back\_cloud}(Q)$, $V(k, 4:5) = [1, m]$.

Step 2: 若 $k > w$, 则停止程序, $C_q = V(:, 1:3)$, 输出 C_q; 否则, 执行 Step 3.

Step 3: 对 V 的第 2 列 $V(:, 2)$ 查找最大熵值, 记

$k_0 = \arg\max_i V(i,2)$.

Step 4: 令 $t_1 = V(k_0,4)$, $t_2 = V(k_0,5)$. 从子序列 $Q(t_1:t_2)$ 中搜索 $t_0, t_1 < t_0 < t_2$, 计算

$$L(1:3) = \text{back\_cloud}(Q(t_1:t_0)),\ L(4:5) = [t_1, t_0];$$

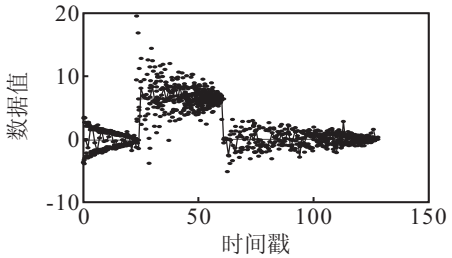
$$R(1:3) = \text{back\_cloud}(Q(t_0:t_2)),\ R(4:5) = [t_0, t_2]$$

使得下式最大:

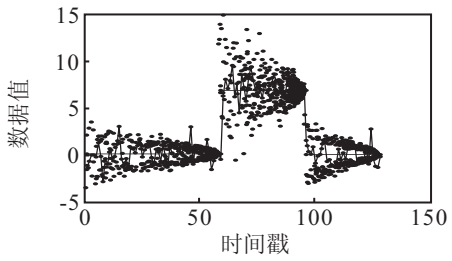
$$\begin{aligned}\Delta En =& V(k_0,2)(V(k_0,5) - V(k_0,4)) - \\ & (L(2)(L(5) - L(4)) + R(2)(R(5) - R(4))).\end{aligned}$$

Step 5: 令 $V(k_0,:) = L$, $V(k+1,:) = R$, $k = k+1$, 返回 Step 2.

通过上述算法, 可以自适应地对时间序列进行基本特征识别和表示. 如图 1 所示, 具有相同基本特征的时间序列, 算法能够自适应地识别飘移的基本特征, 并对其进行云模型表示.



(a) 情况1



(b) 情况2

图 1　CPAA 识别时间序列 Cylinder 的飘移基本特征

3.2　相似性度量

降维后的时间序列转变成云模型序列后, 需要利用相应的相似度度量方法来衡量云模型之间的关系. 由于正态云模型水平方向形成的期望曲线能够较好地反映云模型的几何特征[9], 通过期望曲线相交形成的面积 S 来衡量两个云模型的相似度不失为一个好的方法, 如图 2 所示.

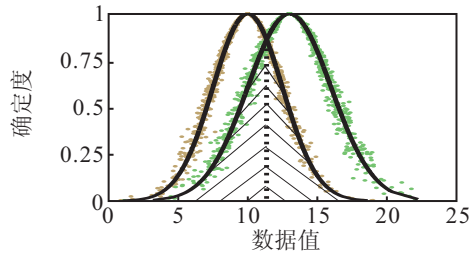
由定义 3 可以推出云模型 $C(\text{E}x, \text{E}n, \text{H}e)$ 的期望曲线方程为

$$y = \text{e}^{-\frac{(x-\text{E}x)^2}{2\text{E}n^2}}. \tag{3}$$

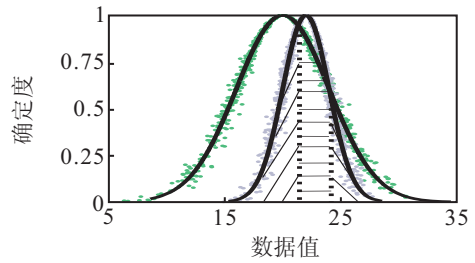
进而得到两个云模型 C_1 和 C_2 相交的面积为

$$S = \int_{-\infty}^{x_0} y_2(x)\text{d}x + \int_{x_0}^{\infty} y_1(x)\text{d}x, \tag{4}$$

其中 x_0 为两条期望曲线 y_1 和 y_2 的交点. 结合式 (3), 可知式 (4) 是不可积的. 变换式 (3), 有



(a) 情况1



(b) 情况2

图 2　两个云模型的相似度面积 S 的构成情况

$$\begin{aligned}y =& \sqrt{2\pi}\text{E}n\frac{1}{\sqrt{2\pi}\text{E}n}\text{e}^{-\frac{(x-\text{E}x)^2}{2\text{E}n^2}} = \\ & \sqrt{2\pi}\text{E}n\frac{1}{\sqrt{2\pi}\text{E}n}f(z),\end{aligned} \tag{5}$$

其中 $f(z)$ 是标准正态分布的概率密度函数. 将式 (4) 转换成标准正态分布下的积分形式为

$$S = \sqrt{2\pi}\text{E}n_2\int_{-\infty}^{z_2} f_2(z)\text{d}z + \sqrt{2\pi}\text{E}n_1\int_{z_1}^{\infty} f_1(z)\text{d}z. \tag{6}$$

其中: $z_1 = (x_0 - \text{E}x_1)/\text{E}n_1$, $z_2 = (x_0 - \text{E}x_2)/\text{E}n_2$. 若知道 x_0, 则可得 z_1 和 z_2, 再通过查询标准正态分布表, 便可解得相交面积 S. 两条期望曲线的相交点为

$$\begin{aligned}x_0^{(1)} &= \frac{\text{E}x_2\text{E}n_1 - \text{E}x_1\text{E}n_2}{\text{E}n_1 - \text{E}n_2}, \\ x_0^{(2)} &= \frac{\text{E}x_1\text{E}n_2 + \text{E}x_2\text{E}n_1}{\text{E}n_1 + \text{E}n_2}.\end{aligned} \tag{7}$$

在式 (7) 中, 选取满足 3 En 规则的 x_0 值来求解相交面积 $S = S_1 + S_2$, 如图 2(a) 所示. 若两点均满足 "3 En" 规则 (如图 2(b) 所示), 则利用上述同样方法来分段求得相交面积 S, 即 $S = S_1 + S_2 + S_3$. 两个云模型的相似度归一化可表示为

$$\text{ECM}(C_1, C_2) = \frac{2S}{\sqrt{2\pi}\text{E}n_1 + \sqrt{2\pi}\text{E}n_2} \in [0,1], \tag{8}$$

其中 $\sqrt{2\pi}\text{E}n_1$ 和 $\sqrt{2\pi}\text{E}n_2$ 分别表示两个云模型与横坐标所形成的面积.

若给定长度为 m 的两个时间序列 Q 和 C, 通过 CPAA 可以得到长度为 w 的云模型序列 C_q 和 C_c, 则降维后的时间序列相似度为

$$D(C_q, C_c) = \sqrt{\frac{1}{w}\sum_{i=1}^{w}\text{ECM}(c_{qi}, c_{ci})}. \tag{9}$$

4　实验分析

为了更好地理解和分析 CPAA 算法的性能, 分别从 3 个方面与经典的分段聚合方法 PAA 以及 SAX 方

法作比较, 即数据拟合误差、时间序列层次聚类和时间序列分类结果比较.

4.1 拟合误差比较

从某一股票时间数据[13]中任选取长度为2 000的时间序列作为研究对象, 分别利用PAA和CPAA方法对其进行数据降维处理, 比较各种处理维度下的拟合误差情况. 处理后维度w越小, 则数据压缩比$k = m/w$越大, 其中$m(m = 2\,000)$表示时间序列的长度. 最终的实验结果如图3所示.

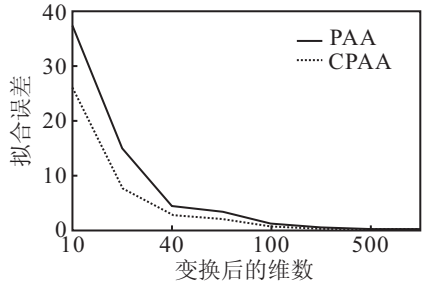
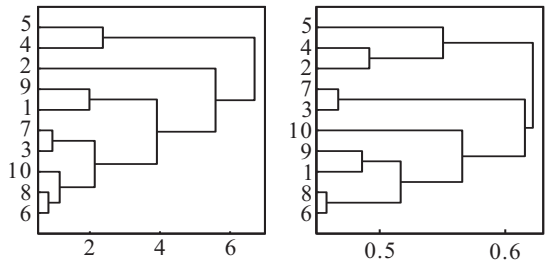


图 3 CPAA与PAA对股票数据的拟合误差比较

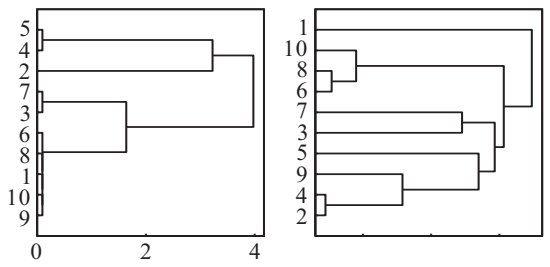
由图3可见, CPAA的数据拟合误差要比PAA小, 且压缩比越大, CPAA相对于PAA的拟合性能越显著.

4.2 时间序列层次聚类

层次聚类方法可以较为客观详尽地描述对象之间的相似性距离, 能够全面地对时间序列相似度的准确性作出评价. 由于基于PAA的符号化表示方法SAX已在时间序列挖掘中表现出良好的性能, 利用CPAA和SAX同时对时间序列长度为128的UCI数据集CBF (Cylinder-Bell-Funnel)[14]在不同降维强度下进行聚类, 分别比较它们的聚类结果. 在数据集CBF中, 随机选取10个时间序列作为聚类对象, 其中$C\{3, 7\}, B\{1, 2, 4, 5, 9\}, F\{6, 8, 10\}$各为一类, 实验结果如图4所示.



(a) w=16



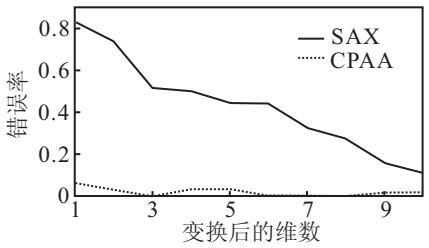
(b) w=4

图 4 两种方法降至不同维度的层次聚类结果

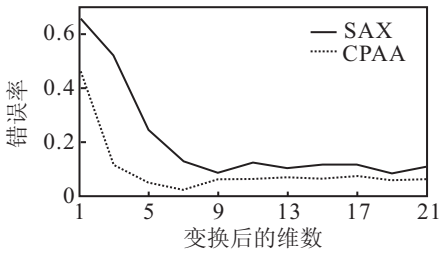
当维度较大时, SAX的聚类结果要略优于CPAA的聚类结果(如图4(a)). SAX首先将时间序列6,8,10聚合成一类; 而CPAA却先将序列6,8,1,9聚合成一类后才与序列10聚合. 但在较低维空间里, CPAA的聚类结果明显优于SAX的聚类结果(如图4(b)). CPAA能够在降至4维的情况下正确归类数据对象, 而SAX出现错误聚类. 因此, CPAA方法更适合时间序列数据降维后的聚类分析.

4.3 时间序列分类

为了进一步验证CPAA方法的数据挖掘性能, 令CPAA和SAX方法采用最近邻分类算法对两个UCI数据集CC(control chart)[15]和数据集CBF (Cylinder-Bell-Funnel)[14]实现分类. 数据集CC有600个长度为60的时间序列, 包含6类数据, 每类有100个时间序列. 在本次实验中, 分别取每类的最后11个时间序列组合成测试集(共66个), 其余时间序列作为训练集(共534个). 同样, 选取长度为128的数据集CBF中的900个时间序列作为测试集, 30个时间序列作为训练集. 对这两个数据集在不同的维度下进行分类实验, 分类结果如图5所示.



(a) 数据集CC



(b) 数据集CBF

图 5 CPAA与SAX在不同维度下对CC和CBF的分类错误率

分类实验结果表明, CPAA方法的分类性能要优于SAX方法. 特别地, 当维度较低时, CPAA的分类性能更为显著. 同时也说明, 数据压缩比较大时, CPAA方法能表现出更好的分类性能.

5 结 论

本文提出的CPAA算法利用云模型的熵来评价各个子序列的数据稳定性, 自适应地对时间序列实

现了分段聚合. 同时, 给出了云模型相似度求解方法, 进而客观有效地实现了降维后的云模型序列相似度计算. 实验结果表明, CPAA 方法不仅能够以较小的误差拟合原时间序列, 而且在时间序列挖掘技术中得到了良好的体现. 与基于经典分段聚合近似 PAA 的 SAX 相比, CPAA 具有较好的聚类性能和分类结果. 然而, 降维过程中 CPAA 需要利用云模型的熵进行评价, 并选取稳定性最差的子序列进行再分段聚合, 时间消耗要大于经典分段聚合近似方法, 因此, 如何提高 CPAA 方法的时间效率需要在今后的研究中进一步探讨.

参考文献(References)

[1] 冯玉才, 蒋涛, 李国徽, 等. 高效时序相似搜索技术[J]. 计算机学报, 2009, 32(11): 2108-2122.
(Feng Y C, Jiang T, Li G H, et al. Underlying techniques of efficient similarity search on time series[J]. Chinese J of Computers, 2009, 32(11): 2108-2122.)

[2] 贾澎涛, 何华灿, 刘丽, 等. 时间序列挖掘综述[J]. 计算机应用研究, 2007, 24(11): 15-29.
(Jia P T, He H C, Liu L, et al. Overview of time series data mining[J]. Application Research of Computers, 2007, 24(11): 15-29.)

[3] Agrawal R, Faloutsos F, Swami A. Efficient similarity search in sequence databases[C]. Proc of the 4th Int Conf on Foundations of Data Organization and Algorithms. Chicago: Springer-Verlag, 1993: 69-84.

[4] Chan K P, Fu A W. Efficient time series matching by wavelets[C]. Proc of the 15th IEEE Int Conf on Data Engineering. New York: IEEE Press, 1999: 117-126.

[5] Korn F, Jagadish H V, Faloutsos C. Efficently supporting ad hoc queries in large dataset of time sequences[C]. Special Interest Group on Management of Data. New York: ACM Press, 1997: 289-300.

[6] Hung N Q, Anh D T. An improvement of PAA for dimensionality reduction in large time series databases[C]. Proc of the 10th Pacific Rim Int Conf on Artificial Intelligence. New York: AAAI Press, 2008: 698-707.

[7] Keogh E, Chu S, Hart D, et al. An online algorithm segmenting time series[C]. IEEE Int Conf on Data Mining. New York: IEEE Press, 2001: 289-296.

[8] Lin J, Keogh E, Lonardi S, et al. A symbolic representation of time series with implications for streaming algorithms[C]. Proc of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. New York: ACM Press, 2003: 2-11.

[9] 李德毅, 杜鹢. 不确定性人工智能[M]. 北京: 国防工业出版社, 2005: 143-185.
(Li D Y, Du Y. Artificial intelligence with uncertainty[M]. Beijing: National Defense Industry Press, 2005: 143-185.)

[10] Li D Y, Di K C, Li D R, et al. Mining association rules with linguistic cloud models[J]. J of Software. 2000, 11(2): 143-148.

[11] Li D Y. Uncertainty reasoning based on cloud models in controllers[J]. Computers and Mathematics with Applications, 1998, 35(3): 99-123.

[12] 柳炳祥, 李海林, 杨丽彬. 云决策分析方法[J]. 控制与决策, 2009, 24(6): 957-960.
(Liu B X, Li H L, Yang L B. Cloud decision analysis method[J]. Control and Decision, 2009, 24(6): 957-960.)

[13] Stock data web page[EB/OL]. (2005-03-26)[2010-06-02]. http://www.cs. ucr.edu/ wli/FilteringData/stock.zip.

[14] Saito N. Local feature extraction and its application using a library of bases[z].

[15] Alcock R J, Manolopoulos Y. Time-series similarity queries employing a feature-based approach[C]. The 7th Hellenic Conf on Informatics. New York: ACM Press, 1999: 1-9.

(上接第1524页)

[9] Lien C H. Non-fragile guaranteed cost control for uncertain neutral dynamic systems with time-varying delays in state and control input[J]. Chaos, Solitons and Fractals, 2007, 31(4): 889-899.

[10] Li Y M, Xu S Y, Zhang B Y, et al. Delay-dependent guaranteed cost control for uncertain neutral systems with distributed delays[J]. Int J of Control, Automation and Systems, 2008, 6(1): 15-23.

[11] Gu K. An integral inequality in the stability problem of time-delay systems[C]. Proc of the 39th IEEE Conf on Decision Control. Sydney, 2000: 2805-2810.

[12] Petersen I R, Hollot C V. A riccati equation approach to the stabilization of uncertain linear systems[J]. Automatica, 1986, 22(4): 397-412.

[13] Crocco L. Aspects of combustion stability in liquid propellant rocket motors[J]. J of the American Rocket Society, 1951, 21(1): 163-178.

[14] Fiagbedzi Y, Pearson A. Feedback stabilization of linear autonomous time lag systems[J]. IEEE Trans on Automatic Control, 1986, 31(6): 847-855.