计算机研究与发展
Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP
44(Suppl.): 111~116, 2007

基于 ACF 非线性趋势特征的时间序列聚类

管河山[1] 姜青山[2] 王声瑞[2,3] 郑宇泉[4]
[1](厦门大学计算机科学系 厦门 361005)
[2](厦门大学软件学院 厦门 361005)
[3](加拿大舍尔鲁克大学计算机系 加拿大魁北克 J1k2R1)
[4](厦门大学数学科学学院 厦门 361005)
(guanheshan@yahoo.com.cn)

Cluster Time-Series Based on the Non-Linear Trend Character of ACF

Guan Heshan[1], Jiang Qingshan[2], Wang Shengrui[2,3], and Zheng Yuquan[4]
[1](*Department of Computer Sciences, Xiamen University, Xiamen* 361005)
[2](*School of Software, Xiamen University, Xiamen* 361005)
[3](*Department of Computer Science, University of Sherbrooke, Quebec, Canada* J1K2R1)
[4](*School of Mathematical Science, Xiamen University, Xiamen* 361005)

Abstract There are lots of study focused on providing a new similarity measure in the mining of time series, including clustering and classification. The non-linear trend character is fully taken into the mining of time series. ACF of time series is first calculated, and then the non-linear trend character of ACF is constructed. The character is taken as the similarity measure, which provides a new way to classify the time series as stationary or as non-stationary. In the experiments two datasets, one natural and one synthetic are utilized. Using the non-linear trend character of ACF, less character values are taken to depict the stationarity of time series than that of some common similarity measures. Experiments are conducted with K-means in the paper.

Key words stationary time series; non-stationary time series; non-linear trend character of ACF; K-means

摘 要 在时间序列挖掘工作中,比如聚类和分类,需要计算距离来衡量时间序列样本之间的相似性,有许多研究都致力于时间序列相似性度量的研究.充分利用非线性趋势特征来进行时间序列挖掘.首先计算时间序列的 ACF,进而构造 ACF 的非线性趋势特征,利用该特征作为时间序列相似性度量来进行聚类,它给时间序列平稳性的判定提供了一种新的途径.列举了一个模拟数据和一个实际数据来进行实例验证,实验结果表明,ACF 非线性趋势特征作为一种新的相似性度量,相对已有的一些相似性度量而言,ACF 非线性趋势特征通常只需计算少量的若干特征值就能更合理地刻画时间序列的平稳性特征.借助 K-means 进行聚类实验.

关键词 平稳时间序列;非平稳时间序列;ACF 非线性趋势特征;K-means

中图法分类号 TP311

时间序列挖掘(特别是时间序列的分类和聚类等)得到了广泛的研究[1-4].在时间序列的分类和聚类的研究中,相似性度量(similarity measure)得到了广泛的关注[2,4].Euclidean 距离常被用来进行时间

收稿日期:2007-03-05
基金项目:国家"九八五"工程二期基金项目(0000-X07204)

序列的聚类研究,但是其鲁棒性不好,而且没有考虑到时间序列自身的关联特征. Piccolo 利用时间序列 ARIMA 模型的系数作为时间序列的相似性度量[5]. 另外还有其他关于时间序列的相似性度量,比如自相关系数(autocorrelation function, ACF)、逆相关系数(inverse autocorrelation function, IACF)和偏相关系数(partial autocorrelation function, PACF)等(详见本文第 1 节).

本文旨在判定时间序列的平稳性. 考虑到时间序列自身的关联特征,本文采用 ACF 非线性趋势特征(记为 ACF2)作为判定时间序列平稳性的标准,以此量化刻画时间序列的平稳性程度,并以此建立一个时间序列相似性度量. 本文列举了一个模拟数据和一个实际数据来进行实验,实验结果表明,采用 ACF 趋势特征只需要取少数几个特征值就可以得到更为合理的结果,这一特点与其他的一些相似性度量如 ACF, PACF 和 IACF 等相比,在准确率和简化计算方面都具有相当的优势.

1 时间序列相似性度量

目前有许多关于时间序列相似性度量的研究, Jorge 等人对一些常用的相似性衡量尺度进行了分析和总结[6]. 本文旨在探求平稳时间序列和非平稳时间序列之间的划分,不仅限于划分结果的准确性,而且力求采用少量特征值来简化计算(采用特征值个数少,有利于建立时间序列的索引). 本文在进行实验的过程中采用了常用的 3 个相似性度量(ACF, PACF 和 IACF)来进行实验结果的对比,下面先简单介绍这 3 种相似性衡量尺度.

1.1 ACF, PACF 和 IACF

Galeano 等人采用了时间序列的 ACF 来衡量时间序列样本之间的相似度[7]. Cleveland 采用时间序列的 PACF 或 IACF 作为其相似性衡量尺度[8]; Chatfield 进一步完善了这方面的工作[9]. 假定一个时间序列 $X=\{X_1, X_2, \cdots, X_n\}$,计算其 ACF, PACF 和 IACF,

并取前若干个特征值,得 $\hat{\rho}_x=(\hat{\rho}_{x,1}, \hat{\rho}_{x,2}, \cdots, \hat{\rho}_{x,m})$, $\hat{\rho}_x^{(p)}=(\hat{\rho}_{x,1}^{(p)}, \hat{\rho}_{x,2}^{(p)}, \cdots, \hat{\rho}_{x,m}^{(p)})$ 和 $\hat{\rho}_x^{(I)}=(\hat{\rho}_{x,1}^{(I)}, \hat{\rho}_{x,2}^{(I)}, \cdots, \hat{\rho}_{x,m}^{(I)})$,通常 $m<n$. 此时两个时间序列 x 和 y 之间的相似程度可以用式(1)~(3)来衡量. 其中 $\boldsymbol{\Omega}$ 为一个权重矩阵,当取 $\boldsymbol{\Omega}=\boldsymbol{I}$ 时为 Euclidean 距离.

$$d_{\mathrm{ACF}}(X,Y)=\sqrt{(\hat{\rho}_X-\hat{\rho}_Y)'\boldsymbol{\Omega}(\hat{\rho}_X-\hat{\rho}_Y)}, \tag{1}$$

$$d_{\mathrm{IACF}}(X,Y)=\sqrt{(\hat{\rho}_X^{(I)}-\hat{\rho}_Y^{(I)})'\boldsymbol{\Omega}(\hat{\rho}_X^{(I)}-\hat{\rho}_Y^{(I)})}, \tag{2}$$

$$d_{\mathrm{PACF}}(X,Y)=\sqrt{(\hat{\rho}_X^{(p)}-\hat{\rho}_Y^{(p)})'\boldsymbol{\Omega}(\hat{\rho}_X^{(p)}-\hat{\rho}_Y^{(p)})}. \tag{3}$$

2 ACF 非线性趋势特征

采用 ACF 作为时间序列的相似性度量,通常有两个问题至今尚未得到完美的解决,其一, ACF 特征值的取值个数,即 m 的取值个数难以决定,不同 m 的取值将得到不同的结果;其二,按照时间序列的平稳性定义,根据 ACF 收敛状况来判定时间序列的平稳性,只能得到近似结果,因而判定准确率仍然是个关键问题. 本文通过 ACF 非线性趋势特征分析给出这两个问题的一个解答.

趋势分析的应用极为广泛[3],本文的研究充分考虑到了非线性趋势因素. 在时间序列的相似性度量中,通常可以构造出不同的度量标准,而且大多研究都是直接采用度量的真实值进行分析[7-9],比如采用 ACF 的若干个特征值来度量时间序列样本间的相似程度. 本文进一步完善 ACF 方法,从 ACF 本身的非线性趋势特征出发来考虑时间序列相似性度量,并采用 ACF 非线性趋势特征来刻画一个时间序列的平稳性,避免直接采用 ACF 的真实值,通常能得到与 ACF 一样甚至更好的结果,而且在特征值选取个数方面具有更大的优势,通常只需要选择少数几个特征值就能够得到好的结果,这是 ACF 所不能达到的. 下面先具体介绍采用 ACF 非线性趋势特征度量时间序列样本之间相似性的做法.

任意选取时间序列库中的一个样本,假定为 $X=\{X_1, X_2, \cdots, X_n\}$,首先计算其 ACF,具体的计算如式(4),得 ACF 的真实值 $\hat{\rho}_X=(\hat{\rho}_{X,0}, \hat{\rho}_{X,1}, \cdots, \hat{\rho}_{X,n-1})$,显然 $\hat{\rho}_{X,0}=1$.

$$\hat{\rho}_{X,k}=\frac{\sum_{t=1}^{n-k}(x_t-\bar{x}_t)(x_{t+k}-\bar{x}_{t+k})}{\sqrt{\sum_{t=1}^{n-k}(x_t-\bar{x}_t)^2\sum_{t=1}^{n-k}(x_{t+k}-\bar{x}_{t+k})^2}}, \tag{4}$$

其中 $(k=0,1,\cdots,n-1)$, $\bar{x}_t=\frac{1}{n-k}\sum_{t=1}^{n-k}x_t$, $\bar{x}_{t+k}=\frac{1}{n-k}\sum_{t=1}^{n-k}x_{t+k}$. 得到 ACF 的真实值后再进行非线性变换,进而构造 ACF 非线性趋势特征,具体的非

线性变换可以分为以下两步:

第 1 步,进行 ACF 真实值的绝对值累加,得到 $\bar{\boldsymbol{\rho}}_X=(\bar{\rho}_{X,0},\bar{\rho}_{X,1},\cdots,\bar{\rho}_{X,n-1})$,其中 $\bar{\rho}_{X,k}=\sum_{i=0}^{k}|\hat{\rho}_{X,i}|(k=0,1,\cdots,n-1)$;

第 2 步,对累加 ACF 序列进行非线性变换,得 $\hat{\boldsymbol{\rho}}'_X=(\hat{\rho}'_{X,0},\hat{\rho}'_{X,1},\cdots,\hat{\rho}'_{X,n-1})$,其中,$\hat{\rho}'_{X,k}=\bar{\rho}'_{X,k}-\ln(\bar{\rho}'_{X,k})(k=0,1,\cdots,n-1)$.

对变换后的序列 $\hat{\rho}'_X$,提取前若干个特征值构建一个特征向量(可不提取第 1 个特征值,因 $\hat{\rho}'_{X,0}=1$),得特征向量 $\hat{\boldsymbol{\rho}}''_X=(\hat{\rho}'_{X,0},\hat{\rho}'_{X,1},\hat{\rho}'_{X,2},\cdots,\hat{\rho}'_{X,m})$,其中 m 为特征值的个数,且 $m<n$,以此特征向量来衡量时间序列之间的相似度,该特征向量中的元素可以反映出 ACF 非线性变换后的趋势特征. 通过非线性变换变化,可以将时间序列的 ACF 连线图(见图 1)变换成过点(0,1)的一条单调递增曲线(见图 2).

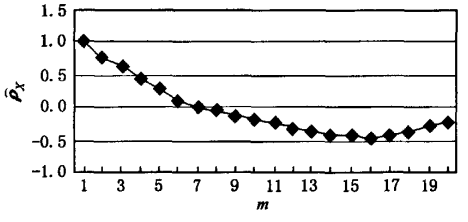

图 1 ACF 真实值分布(m 取值为 0~19)

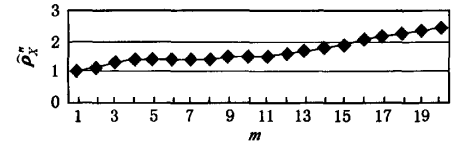

图 2 ACF 非线性趋势特征分布图(m 取值为 0~19)

2.1 ACF 非线性趋势特征向量的特征

本文采用特征向量 $\hat{\boldsymbol{\rho}}''_X$ 来刻画时间序列的平稳性,该特征向量具有以下几个特征:

1) 非负性,即 $\hat{\rho}'_{X,k}>0(k=0,1,2,\cdots,m)$;

2) 特征向量 $\hat{\boldsymbol{\rho}}''_X$ 中的元素具有单调递增性;

3) $\hat{\boldsymbol{\rho}}''_X$ 中的元素具有上确界,其最大值不超过 m,即 $\hat{\rho}'_{X,k}<m(k=1,2,\cdots,m)$恒成立.

通过 ACF 的非线性变换,将原来位于[-1,1]区间的 ACF 值映射到(0,m)区间,进而得到 ACF 非线性趋势特征向量,并用做时间序列平稳性判定的度量标准,此时两个时间序列之间的相似度转化为相应的两个 ACF 非线性趋势特征向量之间的 Euclidean 距离.

2.2 ACF 非线性趋势特征的收敛性

通常可以根据 ACF 的收敛状况来判定时间序列的平稳性特征,某个时间序列 X 的 ACF 序列 $\{\hat{\rho}_{X,0},\hat{\rho}_{X,1},\cdots,\hat{\rho}_{X,n-1}\}$快速收敛于 0,则原时间序列 X 是平稳的;此时对于 ACF 非线性趋势特征序列$\{\hat{\rho}'_{X,0},\hat{\rho}'_{X,1},\cdots,\hat{\rho}'_{X,n-1}\}$而言,如果其收敛,则可以推导出 ACF 收敛于 0.

定理 1. 对于时间序列 X 而言,其 ACF 非线性趋势特征序列$\{\hat{\rho}'_{X,0},\hat{\rho}'_{X,1},\cdots,\hat{\rho}'_{X,n-1}\}$收敛,则可以推导出 ACF 序列$\{\hat{\rho}_{X,0},\hat{\rho}_{X,1},\cdots,\hat{\rho}_{X,n-1}\}$收敛于 0.

证明. 如果 ACF 非线性趋势特征序列收敛,即易得出非线性趋势特征序列是有界的,而 $\hat{\rho}'_{X,k}=\bar{\rho}_{X,k}-\ln(\bar{\rho}_{X,k})$单调递增,故而 $\bar{\rho}_{X,k}$必定是有界的;而序列$\{\bar{\rho}_{X,0},\bar{\rho}_{X,1},\cdots,\bar{\rho}_{X,n-1}\}$中的所有元素满足 $\bar{\rho}_{X,k}=\sum_{i=0}^{k}|\hat{\rho}_{X,i}|$,也是单调递增的,所以序列$\{\hat{\rho}_{X,0},\hat{\rho}_{X,1},\cdots,\hat{\rho}_{X,n-1}\}$必定是收敛于 0 的. 即 ACF 序列收敛于 0. 证毕.

由以上的分析容易看出,ACF 非线性趋势特征序列收敛,则可以推导出 ACF 序列收敛于 0,它是一个更强的判定条件. 大量实验验证了该方法具有相当的优越性,本文列举两个数据来进行分析.

3 实 验

本文列举一个模拟数据和一个实际数据来进行实验,并选择常用的 ACF,PACF 和 IACF 3 种方法来进行实验结果的对比. 实验结果表明,采用 ACF 趋势特征通常只需要提取少数几个特征值就可以得到更为合理的结果. 而 ACF,PACF,IACF 等需要更多的特征值才能得到合理的结果. 本文是围绕平稳时间序列和非平稳时间序列之间的分类展开,所有数据都分成 2 类,在此采用了 K-means 方法.

3.1 结果的评价

对于事先知道分类结果的时间序列数据,可以采用客观的判定方法. 以指标 $Sim(G,A)$来度量结果[10]. 而事先不知道分类结果的数据可采用主观的方法进行判定,将分类后的两组时间序列的折线图分别绘出,并进行直观分析[4].

3.2 模拟数据

在此参考了文献[6]中的实验做法. 首先生成 6 组平稳时间序列[(a)~(f)]和 6 组非平稳时间序列

[(g)~(l)],每组时间序列各有 1000 个时间序列样本. 本文共进行了 6 次模拟数据,每次模拟的时间序列的维度大小分别为 50,100,200,500,1000,5000 等.

(a) $AR(1)$,其中 $\varphi_1=0.9$;

(b) $AR(2)$,其中 $\varphi_1=0.95,\varphi_2=-0.1$;

(c) $ARMA(1,1)$,其中 $\varphi_1=0.95,\varphi_2=0.1$;

(d) $ARMA(1,1)$, 其中 $\varphi_1=-0.1,\varphi_2=-0.95$;

(e) $MA(1)$, 其中 $\theta_1=-0.9$;

(f) $MA(2)$, 其中 $\theta_1=-0.95,\theta_2=-0.1$;

(g) $ARIMA(1,1,0)$, 其中 $\varphi_1=-0.1$;

(h) $ARIMA(0,1,0)$;

(i) $ARIMA(0,1,1)$, 其中 $\theta_1=0.1$;

(j) $ARIMA(0,1,1)$, 其中 $\theta_1=-0.1$;

(k) $ARIMA(1,1,1)$, 其中 $\varphi_1=0.1,\theta_1=-0.1$;

(l) $ARIMA(1,1,1)$, 其中 $\varphi_1=0.05,\theta_1=-0.05$.

在模型的参数选择上,尽量使得生成的时间序列不容易区分,即使平稳序列和非平稳序列之间难以判定.

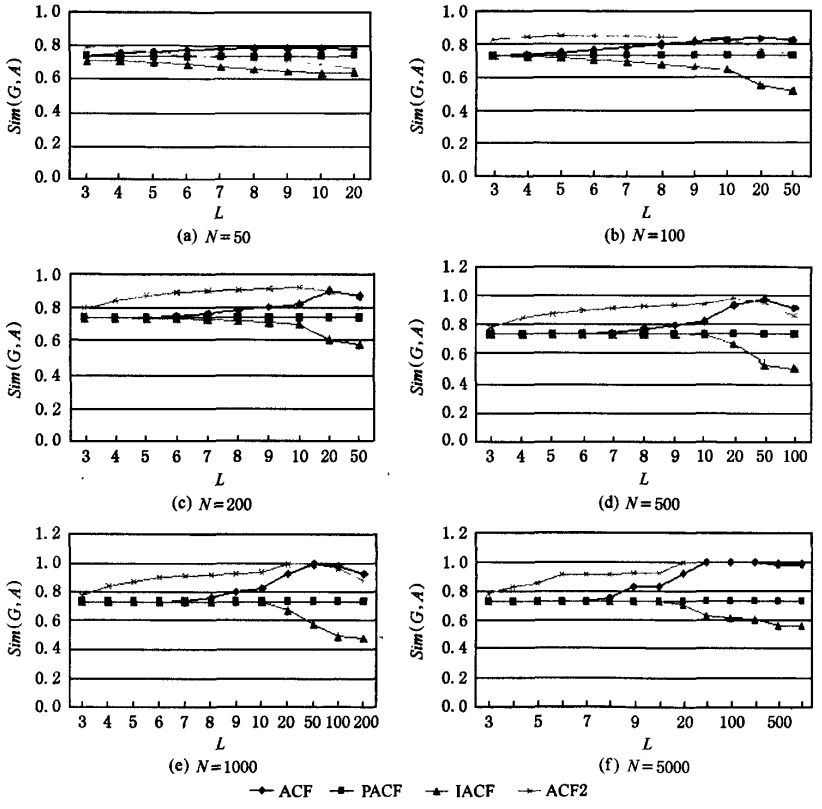

图 3 不同维度时 4 种相似性度量的 $Sim(G,A)$分布(L 表示特征值的个数)

实验时将时间序列分成 2 类,然后用 $Sim(G,A)$指标来判定结果. 在 6 种不同维度的情况下,绘出分类准确率指标 $Sim(G,A)$的分布图,如图 3 所示. 从图 3 来看,L(即 m,称之步长)取值为(3~10),时间序列维度为 100 以上,采用 ACF 非线性趋势特征作为衡量尺度来进行分类,其效果比其他 3 种尺度都要好;这表明如果采取少量特征值来进行分析时,ACF 非线性趋势特征更能准确刻画时间序列的平稳性特点. 当时间序列维度为 50 时时间序列样本本身的信息量过少,采用不同的方法进行实验,得到的结果都不太理想($Sim(G,A)<0.8$).

进一步分析,当维度为 50 和 100 的时间序列时,采用不同的相似性度量来进行计算,通常得到的准确率结果不太高(小于 90%),此时若采用 ACF 非线性趋势特征,只需要采用 5 个特征值可以得到较为合理的结果(见表 1);当维度为 200,500,1000 和 5000 的时间序列,此时采用 ACF 非线性趋势特征,通常只需要采用 10 个特征值就可以得到较为合理的结果,此时聚类的准确率都可以超过 90%(见表 1);特别的,对于维度为 500,1000 和 5000 的等

高维度的时间序列,采用 ACF 非线性趋势特征的 20 个特征值进行计算,其准确率高达 0.97 以上. 再次考虑选取的步长 L,当步长为 5 和 10 时,可以发现,对于不同维度的时间序列,维度越高分析的结果越准确(图 4),这可能是时间序列本身能提供足够的信息量,有利于进行平稳性分析. 此时若提取的特征值个数过多,准确率会有稍微下降的趋势.

表 1 ACF2 的 $Sim(G,A)$值分布

L	T50	T100	T200	T500	T1000	T5000
3	0.78496	0.7739	0.74761	0.73514	0.73343	0.73333
4	0.79151	0.80846	0.78498	0.7662	0.76086	0.73635
5	0.78422	0.83643	0.83288	0.82203	0.82521	0.82857
6	0.77372	0.84874	0.86173	0.85658	0.85029	0.83147
7	0.76164	0.85427	0.87983	0.88086	0.88059	0.87894
8	0.75032	0.8568	0.89347	0.89908	0.90234	0.91583
9	0.73918	0.85292	0.9034	0.91498	0.91472	0.91608
10	0.72771	0.84943	0.91121	0.92559	0.92101	0.91625
20	0.69678	0.79091	0.91367	0.9734	0.98741	0.99992
50		0.76457	0.8153	0.95752	0.99658	1
100				0.87363	0.96763	1
200					0.88077	0.99967
500						0.97882
1000						0.97882

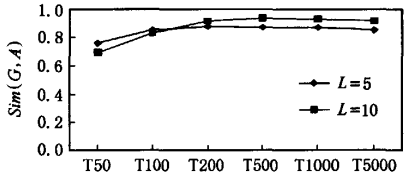

图 4 不同维度的时间序列,*ACF2* 在 L 为 5 和 10 时的 Sim(G,A)分布的折线图

综上所述,采用 ACF 非线性趋势特征能很好地描述出时间序列的平稳性特征,对于不同维度的时间序列库,通常只需要选取少数几个特征值(5 个或 10 个)就可以合理地描述出原时间序列的平稳性特征,这一特征有利于建立时间序列的索引(关于平稳性判定). 进一步,本文采用了一个真实数据来进行分析.

3.3 股票数据

该数据收集了中国股市[11]中上证指数的 9 个股票的历史数据(股票代号为 600739.ss,600031.ss,600550.ss,600262.ss,600030.ss,600151.ss,600155.ss,600152.ss,600193.ss),为了确保数据的时效性,以 2006 年 9 月 8 日为中止日,向前取股票每日的历史数据,每个股票取 200 个数据,构建一个时间序列数据集. 然后进行实验,并描绘出分类后的股票数据折线图.

股票数据的维度为 200,取步长 $L=10$,分成 2 类,得到的结果如图 5 所示. 可见上部分的 3 个时间序列形状总体上没有明显的单调性,局部波动性较大,表现出一定的平稳性;而下部分的 6 个时间序列都有显著的单调趋势(先增后减),表现出明显的非平稳性. 可见,ACF 非线性趋势特征能很好地刻画出时间序列的平稳性特点.

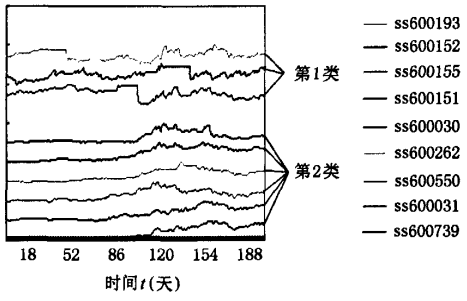

图 5 股票数据的聚类结果示意图

4 总 结

时间序列之间的相似性度量是研究时间序列挖掘的基础,好的度量对完成挖掘任务的质量和效率起着重要的作用. 本文利用不同的相似性度量来进行时间序列平稳性的判定,从客观和主观两个方面来进行实验结果分析. 实验表明,通过将 ACF 的值从区间$[-1,1]$映射到区间$[0,m]$,进而构造 ACF 非线性趋势特征,此时只需要少数几个特征值(步长取值较小)就可以得到好的结果,这给时间序列的平稳性判定提供了一条新的途径;特别是在特征值个数 m 尽量小的前提下,相对其他方法而言,可以得到更为准确的判定结果,这也给建立时间序列索引提供了一条新的途径.

目前的实验只是局限于时间序列的平稳性判定方面,下一步可就时间序列的索引和波动性分析展开研究.

参 考 文 献

[1] 陈佐, 谢赤, 陈晖. 基于小波聚类方法的股票收益率序列时间模式挖掘. 系统工程, 2005, 23(11): 102-107

[2] 汤胤. 时间序列相似性分析方法研究. 计算机工程与应用, 2006, 42(1): 68-71

[3] 周明磊. 非参数估计与小波分析在股市趋势线中的应用. 数理统计与管理, 2005, 25(4): 70-75

[4] E Keogh, S Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. The 8th ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002

[5] D Piccolo. A distance measure for classifying ARIMA models. Journal of Time Series Analysis, 1999, 11: 152-164

[6] C Jorge, C Nuno, P Daniel. A periodogram-based metric for time series classification. Computational Statistics & Data Analysis, 2006, 50(10): 2668-2684

[7] P Galeano, D Pe. Multivariate analysis in vector time series. Journal of the Institute of Mathematics and Statistics of the University of Sao Paolo, Resenhas, 2000, 4: 383-404

[8] W S Cleveland. The inverse autocorrelations of a time series and their applications. Technometrics, 1972, 14(2): 277-293

[9] C Chatfield. Inverse autocorrelations. Journal of Royal Statistical Society, A, 1979, 142: 363-377

[10] M Gavrilov, D Anguelov, P Indyk, *et al*. Mining the stock market: Which measure is best? The 6th ACM Int'l Conf on KDD, Boston, 2000

[11] http://cn.finance.yahoo.com/

管河山 男,1981 年生,博士研究生,主要研究方向为数据挖掘、金融数据处理、统计学、数学建模.

姜青山 男,1962 年生,教授,博士生导师,主要研究方向为数据挖掘、数据库系统、聚类分析、模糊集理论与应用.

王声瑞 男,1963 年生,教授,博士生导师,主要研究方向为模式识别、人工智能、数据挖掘、图像处理和理解等.

郑宇泉 男,1981 年生,硕士研究生,主要研究方向为数据挖掘、金融数据处理.