文章编号: 1001-0920 (2013) 11-1718-05

时间序列数据降维和特征表示方法

李海林[1,2], 杨丽彬[1]

(1. 华侨大学 工商管理学院，福建 泉州 362021；2. 大连理工大学 系统工程研究所，辽宁 大连 116024)

摘 要: 数据降维和特征表示是解决时间序列维灾问题的关键技术和重要方法, 它们在时间序列数据挖掘中起基础性作用. 鉴于此, 提出一种新的时间序列数据降维和特征表示方法, 利用正交多项式回归模型对时间序列实现特征提取, 结合特征序列长度对时间序列的拟合分析结果, 运用奇异值分解方法对特征序列进一步降维处理, 进而得到保存大部分信息且维数更低的特征序列. 数值实验结果表明, 新方法可以在维度较低的特征空间下取得较好的数据挖掘聚类和分类效果.

关键词: 时间序列；数据降维；特征表示；数据挖掘

中图分类号: TP273 **文献标志码:** A

Method of dimensionality reduction and feature representation for time series

LI Hai-lin[1,2], *YANG Li-bin*[1]

(1. College of Business Administration，Huaqiao University，Quanzhou 362021，China；2. Institute of Systems Engineering，Dalian University of Technology，Dalian 116024，China. Correspondent：LI Hai-lin，E-mail：hailin@mail.dlut.edu.cn)

Abstract: Dimensionality reduction and feature representation are the key technique and important methods to address the issue of dimensionality curse for time series. Meanwhile, they are a basis task in the field of time series data mining. Therefore, a novel method of dimensionality reduction and feature representation is proposed. An orthogonal polynomial regression model is used to obtain a feature sequence from an original time series. Furthermore, singular value decomposition combining with the fitting results of the feature sequence to time series is used to reduce the dimensionality of feature sequence and obtain another feature sequence with lower dimension to retain most of the information. The results of numerical experiments demonstrate that the novel method can obtain a good effect of clustering and classification in time series data mining under the space with lower dimensionality.

Key words: time series；dimensionality reduction；feature representation；data mining

0 引 言

时间序列是一种与时间直接或间接有关的高维数据类型, 它广泛存在于金融、经济、气象等科学工程领域. 时间序列数据挖掘能够利用相关技术从时间序列中发现潜在有价值的知识和信息, 是目前数据挖掘领域最具有挑战性的十大问题之一[1]. 时间序列数据的高维特性给数据挖掘工作带来了极大的困难, 若直接对时间序列进行挖掘, 则不但需要消耗较大时间和空间, 而且最终也很难得到较为科学合理的结果, 因此, 对时间序列进行数据降维和特征表示有重要的实际意义. 目前存在一些较为成熟的数据降维和特征表示方法, 如分段线性近似[2-3]、符号化表示方法[4-5]、分段聚合近似[4,6]、离散傅里叶变换[7]、离散小波变换[8]和奇异值分解[9]等.

近年来, 基于时间序列的多项式回归分析模型能够直观有效地对时间序列进行特征表示. 例如, 李爱国等[10]提出了一种分段的多项式回归分析模型, 不但实现了对时间序列的分段表示, 而且提出的距离度量满足下界要求, 避免了在相似性检索中发生漏报现象. Fuchs 等[11]提出了基于正交多项式的时间序列表示方法, 利用最小二乘法结合正交多项式拟合时间序列, 并利用正交多项式基向量形成特征空间, 选取数

收稿日期: 2012-07-13；**修回日期:** 2012-11-28.

基金项目: 中央高校基本科研业务费项目(12SKGC-QG03)；福建省社会科学规划项目(2013C018).

作者简介: 李海林(1982−), 男, 讲师, 博士, 从事数据挖掘与人工智能等研究；杨丽彬(1982−), 女, 讲师, 硕士, 从事数据挖掘的研究.

值较大的坐标系数作为特征序列, 已成功应用于时间序列在线分割[12]和主题发现[13]领域. 然而, 这些成果缺少确定多项式最高项次数的研究, 虽然 Fuchs 等建议事先设定一个较大的最高项次数, 保留较大的系数作为特征序列, 但因为较小系数通常出现在较高次数的正交多项式基向量中, 它们也包含了原始时间序列的重要信息. 所以, 对较小系数的分析是本文的重要工作. 由于对原时间序列具有最佳拟合效果的系数具有较大的长度, 利用奇异值分解方法可以对这些系数作投影变换, 选取包含绝大部信息的变换特征作为新的特征序列, 进而实现了时间序列的数据降维和特征表明. 数值实验结果表明, 该方法仅用较小的特征序列即可充分反映原始表示时间序列的主要信息, 能够提高时间序列数据挖掘算法的性能.

1　正交多项式回归分析模型

利用正交多项式回归分析模型可以对时间序列进行数据降维和特征表示. 假设长度为 M 的时间序列为 $Q=\{q_0,q_1,\cdots,c_M\}$, 通过正交多项式回归分析模型, 可以用一个最高项次数为 K 的多项式 $F_K(t)$ 近似表示. 同时, 该多项式 $F_K(t)$ 可以由 $K+1$ 个正交多项式基向量 $f_k(t)$ 线性表示, 即

$$F_K(t)=\sum_{k=0}^{K}a_k f_k(t). \tag{1}$$

其中: $t=0,1,\cdots,M$; $f_k(t)=t^k+r_{k,k-1}t^{k-1}+\cdots+r_{k,1}t+r_{k,0}$, 且任意两个正交多项式基向量的内积为 0, 即对于任意 $i\neq j$, 有

$$\langle f_i(t)|f_j(t)\rangle=\sum_{t=0}^{M}f_i(t)f_j(t)=0. \tag{2}$$

这样, 时间序列 Q 可以用特征序列 $A=a_0,a_1,\cdots,a_K$ 表示. 该特征序列可以视作时间序列在特征空间 $F=\{f_0(t),f_1(t),\cdots,f_K(t)\}$ 下的坐标, 同时有

$$a_k=\frac{1}{\|f_k\|^2}\sum_{t=0}^{M}q_t f_k(t), \tag{3}$$

$$f_{k+1}(t)=\alpha f_k(t)+\beta f_{k-1}(t). \tag{4}$$

其中

$$\alpha=t-\frac{M}{2},\ \beta=\frac{k^2(M+1)^2-k^4}{4-16k^2},$$

$$0\leqslant k\leqslant K,\ f_{-1}(t)=0,$$

$$f_1(t)=1,\ \|f_k\|^2=\frac{(k!)^4}{(2k)!(2k+1)!}\Pi(M+1+i).$$

由上述分析可知, 特征空间中的任一坐标 $f_k(t)$ 只与时间序列长度 $M+1$ 和当前坐标序列次序 k 有关, 与时间序列 Q 无关. 这表明具有相同长度的时间序列 Q 和 C 可以使用同一个坐标空间来度量, 能够得到与时间序列具体值相关的特征序列 $Q_A=\{q_{a_0},q_{a_1},\cdots,q_{a_K}\}$ 和 $C_A=\{c_{a_0},c_{a_1},\cdots,c_{a_K}\}$. 最终利用特征序列度量原时间序列的距离, 并称为形态空间距离度量 (SSD), 即

$$\mathrm{SSD}(Q_A,Q_C)=\sqrt{\sum_{k=0}^{K}(q_{a_k}-c_{a_k})^2}. \tag{5}$$

这是一个未证明满足下界要求的距离度量函数, 在相似性检索中可能会产生漏报现象[11].

2　数据降维和特征表示新方法

时间序列数据降维和特征表示的主要目的是提高数据挖掘的效率, 使得利用尽可能少的特征充分反应原时间序列的主要信息[4-6,14]. 通常情况下, 特征越多, 包含的信息量越大. 但在某些情况下, 由于这些特征之间可能存在潜在的冗余, 特征越多也不能体现所包含的信息越大, 需要对这些特征实现进一步降维和新的特征表示.

2.1　数据降维和特征分析

利用正交多项式回归分析模型可以实现时间序列数据降维和特征表示. 理论上, 这些特征拟合原时间序列的质量取决于特征序列的长度, 即特征序列的元素越多, 拟合时间序列的性能越好. 然而, 过高次数的多项式拟合时间序列会产生过拟合现象 (如图 1 所示, 实线为时间序列, 虚线为拟合曲线), 使得最高项次数 K 越大, 其拟合质量越差. 由图 1 可见, 随着 K 的增大, $K+1$ 个特征拟合原时间序列的性能先越来越好, 到达一定程度后 (如 $K=23$), 拟合时间序列的性能变差. 因此, 在对时间序列进行特征表示之前, 先要确定拟合性最好的最高项次数 K 值.

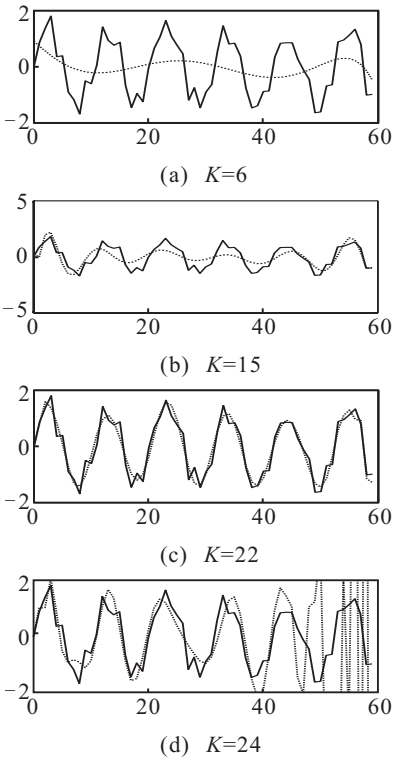


图 1　特征序列根据不同 K 值的拟合现象

由于正交多项式回归分析模型能快速计算出最高项次数为 K 所对应 $K+1$ 维特征空间下的特征系数 A, 可以通过分析 $K+1$ 个特征系数拟合原时间序列的误差来实现对最佳 K 值的选定. 通过对图 1 中时间序列数据进行拟合误差分析得知: 当 $K<23$ 时,相应的特征序列拟合时间序列越来越好; 当 $K\geqslant 23$ 时, 出现过拟合现象, 拟合性能越来越差. 从拟合误差曲线的单调性容易得到最佳的最高次项 $K=22$.

通过上述分析, 可以将最佳拟合原时间序列的 $K+1$ 个特征作为原时间序列的特征序列, 进而实现时间序列的有效降维和特征表示.

2.2 特征表示新方法

虽然通过正交多项式回归分析模型, 选取了较为合适的 $K+1$ 特征序列 $A=\{a_0,a_1,\cdots,a_K\}$, 但由图 1 可知, 对于长度为 60 的时间序列, 最终用 23 个特征来表示原时间序列, 降维幅度不大. 同时, 在 $K+1$ 个特征中, 随着 $k(0\leqslant k\leqslant K)$ 值的增大, a_k 值会急剧减小 (如表 1 所示), 所以 SSD 方法中的特征序列中元素值会很小. 文献 [11] 建议丢弃数值较小的特征元素 a_k, 但通过拟合误差分析可知, 增加较小的特征 a_k 也能使拟合性能越来越好, 保留这些数据较小的特征元素与较大元素对时间序列拟合效果具有同样重要的贡献, 因此应保留数值较小的特征元素.

表 1 SSD 与 LBD 对应特征序列中任选的 5 个特征数值

	a_1	a_3	a_8	a_{12}	a_{15}
SSD	5×10^{-3}	-8.5×10^{-6}	-1.0×10^{-10}	1.0×10^{-15}	-5.9×10^{-19}
LBD	-0.666	-0.269	-2.458	1.171	-2.084

由于 SSD 所对应的特征数值之间的差异性较大, 导致难以判别哪些特征可以丢弃. 即使这些特征全部被保留, 不同时间序列在同一坐标上的特征数值也存在很大的差异, 不利于时间序列特征之间的相似性比较. 因此, 提出另一种时间序列特征数值表示方法, 用于构造下界性距离 (LBD). 该特征表示方法使得不同特征的具体数据不会存在太大的差异性, 如表 1 中 LBD 所在行显示, 在 SSD 中特征数值存在很大差异, 但在 LBD 中差异性较小. LBD 中的特征序列通过对 SSD 中相应特征序列实现进一步转化得到, 即

$$a'_k=\sqrt{\|f_k\|^2}\times a_k. \tag{6}$$

因此两个时间序列 Q 和 C 所对应的特征序列距离度量函数为

$$\mathrm{LBD}(Q_{A'},C_{A'})=\sqrt{\sum_{k=0}^{K}(Q_{a'_k}-C_{a'_k})^2}. \tag{7}$$

该距离度量函数满足下界要求, 在相似性检测中不会产生漏报现象.

对于大部分时间序列而言, 利用正交多项式回归分析模量对时间序列进行特征表示时, 当最高项次数超过约 20 时, 利用相应的特征序列进行原时间序列拟合时会发生过拟合现象. 这样, 被保留的特征序列的维度即在 20 维左右, 其维度仍然过高, 不利于时间序列相似性比较和后期的数据挖掘. 因此, 提出利用奇异值分解方法 (SVD), 对 LBD 中得到的特征序列数值实现进一步降维处理和特征抽取, 使得这些特征的信息集中在几个基于 SVD 的新特征中.

若一组时间序列 $S=\{Q_1,Q_2,\cdots,Q_N\}$, 经过正交多项式回归分析模型, 则选取适合于所有时间序列拟合性能的 K 值, $K=\min_k\{K_1,K_2,\cdots,K_N\}$, K_n 为最佳拟合第 n 个时间序列对应多项式的最高次数. 对每个时间序列保留 $K+1$ 个特征, 形成基于 SSD 特征的矩阵 $A_{N\times(K+1)}$, 通过 LBD 特征数值处理, 得到一组各特征数据之间差异性较小的矩阵 $A'_{N\times(K+1)}$. 最终利用奇异值分解, 对矩阵 $A'_{N\times(K+1)}$ 进行分解, 即

$$A'_{N\times(K+1)}=U\times\Lambda\times V^{\mathrm{T}}. \tag{8}$$

其中: T 为矩阵转置操作, U 为列正交的 $N\times L$ 的矩阵, Λ 为 $A'_{N\times(K+1)}$ 对应特征值按从大到小排列所形成的 $L\times L$ 对角矩阵, V 为 $A'_{N\times(K+1)}$ 的列正交 $(K+1)\times L$ 特征向量矩阵.

根据主成分分析思想, $A'_{N\times(K+1)}$ 的大部分信息集中在前 l 个特征值所对应的主成分中, 因此, 可以得到由 l 个主成分所张成的特征空间对应的新特征矩阵 $A''_{N\times l}$, 即

$$A''_{N\times l}=U(:,1:l)\times\Lambda_{l\times l}. \tag{9}$$

其中: $U(:,1:l)$ 为 U 中前 l 列向量所形成的矩阵, $\Lambda_{l\times l}$ 为前 l 个较大特征值所形成的对角矩阵.

根据矩阵 $A''_{N\times l}$ 可以实现任意一对时间序列 Q_i 和 Q_j 的相似性比较, 即

$$\mathrm{SVD\_LBD}=\sqrt{\sum_{k=1}^{l}(A''_{N\times l}(i,k)-A''_{N\times l}(j,k))^2}. \tag{10}$$

3 数值实验

为了进一步表明新方法的可行性和优越性, 通过时间序列聚类和分类实验检验 3 种方法 (SSD, LBD 和 SVD_LBD) 的性能.

3.1 聚类实验

从 UCR 时间序列数据集 Synthetic_Control[15]中任意选择 15 个时间序列, 长度为 60, 15 条时间序列的分类情况是 {1,2,3}, {4,5}, {6,7,8}, {9,10,11}, {12,13,14} 和 {15} 各为一类. 由于层次聚类能直观有效地对算法的性能作出评价, 利用层次聚类结合 3 种时间序列特征表示方法和相应的距离度量函数 (SSD, LBD 和 SVD_LBD) 对 15 个时间序列进行聚类分析. 同时,

根据被选取特征值与所有特征值比值的不同进行层次聚类分析, 聚类结果如图 2 和图 3 所示.

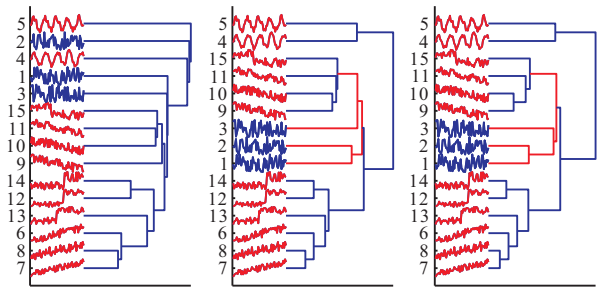


图 2　层次聚类结果比较 (信息量占 0.925 6)

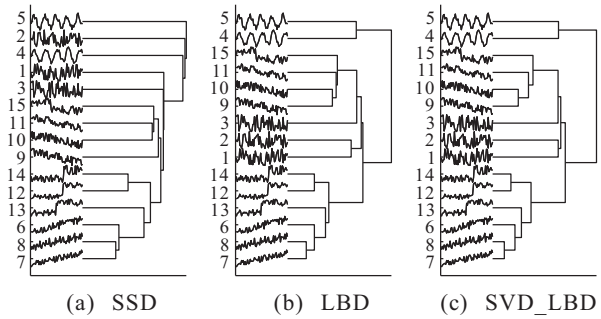


图 3　层次聚类结果比较 (信息量占 0.946 1)

图 2 中, SVD_LBD 被选用的特征值在所有特征值中所占的信息量比为 0.925 6 (即特征序列长度 l 为 5), 图 3 对应的信息量比为 0.946 1.

从 {1,2,3} 和 {4,5} 的聚类结果看, LBD 和 SVD_LBD 的聚类性能明显优于 SSD. 对 LBD 和 SVD_LBD 进行比较, SVD_LBD 利用较少的特征 (信息量较少) 也能得到更好的聚类结果, 如图 2 中 {1,2,3} 的聚类情况. LBD 先将时间序列 3 与 {9,10,11,15} 聚类, 然后与 {1,2} 聚类, 而 SVD_LBD 能够将 {1,2,3} 先聚成一类. 随着特征序列 l 的增加, SVD_LBD 的聚类情况会衰退为 LBD 的情况. 如图 2 所示, 当 $l \leqslant 6$ 时, SVD_LDB 和 LBD 的聚类结果完全一样. 因此可以说, SVD_LBD 能够用最少的特征较好地反映原时间序列的信息, 是一种表现能力较强的时间序列特征表示方法.

3.2　分类实验

利用最近邻分类算法并结合 3 种方法的距离度量方法对时间序列数据集 Beef[15] 进行分类实验, 即利用最近邻分类算法从训练集中找出与测试集中每条时间序列的最相似对象. 若查询结果的类标与查询对象的类标一致, 则视为正确分类, 否则为错误分类. 最后取查询结果的平均值作为不同方法在该数据集下的分类结果.

从 2 个方面考查 3 种方法的分类性能, 分别为指定特征序列维度下不同主成分数目的分类性能和特征序列不同维度下的特定主成分数目的性能. 前者利用 SSD, LBD 和 SVD_LBD 三种方法先对时间序列实现基于正交多项式回归分析的特征表示, 并将其转化为长度为 20 的特征序列, 再考查 SVD_LBD 方法取不同数目主成分的分类性能, 实验结果如图 4 所示. 后者利用 3 种方法对时间序列转化为不同长度的特征序列, 并且进一步考查 SVD_LBD 在特定数目主成分下的分类性能, 实验结果如图 5 所示.

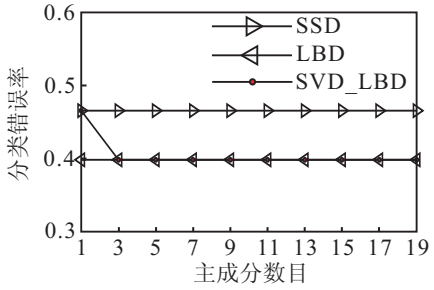
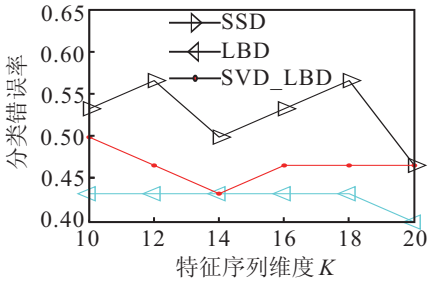
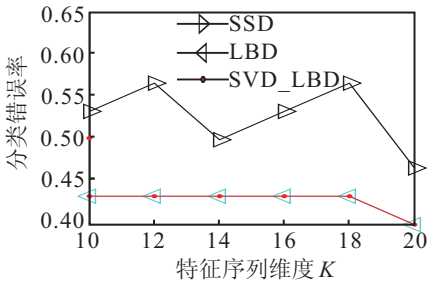


图 4　不同主成分下的分类结果



(a) 取第1个主成分的分类结果



(b) 取第2~9个主成分的分类结果

图 5　在特定数目主成分下的分类结果

由图 4 可见, 由于 SSD 和 LBD 与主成分数目无关, 对于特征序列长度为 20 的降维幅度, LBD 的分类结果要优于 SSD 的分类结果. 同时, 除了第 1 主成分外, SVD_LBD 和 LBD 具有相同的分类结果且优于 SSD. 图 5(a) 表明, 当取第 1 主成分时, SVD_LBD 的分类结果受特征序列维度 K 的影响, 其结果介于 SSD 和 LBD 之间. 图 5(b) 表明, 当取多于 1 个主成份时, SVD_LBD 的分类性能与 LBD 相同且优于 SSD; 同时也说明 SVD_LBD 仅需要 2 个主成分特征便可充分表示长度为 10～18 的特征序列. 因此, SVD_LBD 在不改变 LBD 分类效果的情况下, 可以使用较小数目的主成分来表示时间序列, 进而实现有效的数据降维和相似性度量.

4 结 论

本文基于正交多项式回归分析模型对时间序列进行特征表示, 提出了一种新的时间序列数据降维和特征表示方法. 通过正交多项式回归分析模型对时间序列进行数据降维, 对多项式不同最高项次数得到不同长度的特征序列拟合原时间序列情况进行分析, 最终确定最佳多项式最高项次数. 同时, 对特征序列的数值具有较大差异性实现了进一步处理, 减小了特征数值之间的差异, 并且利用奇异值分解方法进行更低维的数据降维和特征表示. 实验表明, 与传统方法相比, 新的数据降维和特征表示方法在时间序列层次聚类和分类数据挖掘任务中取得了较好的结果.

参考文献(References)

[1] Yang Q, Wu X D. 10 challenging problems in data mining research[J]. Int J of Information Technology & Decision Making, 2006, 5(4): 597-604.

[2] Keogh E, Chu S, Hart D, et al. An online algorithm for segmenting time series[C]. Proc of the 2001 IEEE Int Conf on Data Mining. NewYork: IEEE, 2001: 289-296.

[3] Li H L, Guo C H, Qiu W R. Similarity measure based on piecewise linear approximation and derivative dynamic time warping for time series mining[J]. Expert Systems with Applications, 2011, 38(12): 14732-14743.

[4] Lin J, Keogh E, Wei L, et al. Experiencing SAX: A novel symbolic representation of time series[J]. Data Mining and Knowledge Discovery, 2007, 15(2): 107-144.

[5] 钟清流, 蔡自兴. 基于统计特征的时序数据符号化算法[J]. 计算机学报, 2008, 31(10): 1857-1864.
(Zhong Q L, Cai Z X. The symbolic algorithm for time series data based on statistic feature[J]. Chinese J of Computers, 2008, 31(10): 1857-1864.)

[6] 李海林, 郭崇慧. 基于云模型的时间序列分段聚合近似方法[J]. 控制与决策, 2011, 26(10): 1525-1529.
(Li H L, Guo C H. Piecewise aggregate approximation method based on cloud model for time series[J]. Control and Decision, 2011, 26(10): 1525-1529.)

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

[8] Chan K P, Fu A C. Efficient time series matching by wavelets[C]. Proc of the 15th IEEE Int Conf on Data Engineering. NewYork: IEEE, 1999: 126-133.

[9] Korn F, Jagaciish H V, Faloutsos C. Efficiently supporting ad hoc queries in large datasets of times equences[C]. Proc of the 1997 ACM SIGMOD Int Conf on Management of Data. NewYork: ACM Press, 1997: 289-300.

[10] 李爱国, 覃征. 大规模时间序列数据库降维及相似搜索[J]. 计算机学报, 2005, 28(9): 1467-1475.
(Li A G, Qin Z. Dimensionality reduction and similarity search in large time series databases[J]. Chinese J of Computers, 2005, 28(9): 1467-1475.)

[11] Fuchs E, Gruber T, Nitschke J, et al. Temporal data mining using shape space representation of time series[J]. Neurocomputing, 2010, 74(1-3): 379-393.

[12] Fuchs E, Gruber T, Nitschke J, et al. On-line segmentation of time series based on polynomial least-squares approximation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(12): 2232-2245.

[13] Fuchs E, Gruber T, Nitschke J, et al. On-line motif detection in time series with SwiftMotif[J]. Patterns Recognition, 2009, 42(11): 3742-3750.

[14] Li H L, Guo C H. Piecewise cloud approximation for time series mining[J]. Knowledge-Based Systems, 2011, 24(4): 492-500.

[15] Keogh E, Xi X, Wei L, et al. The UCR time series classification & clustering hame page[EB/OL]. (2006-05-01)[2012-05-27]. http: //www.cs.ucr.edu/∼eamonn/time_series_data/.

(上接第1717页)

[11] Stenbakken G N, Souders T M, Stewart G M. Ambiguity groups and testability[J]. IEEE Trans on Instrumentation and Measurement, 1989, 38(5): 941-947.

[12] Janusz A Starzyk, Jing Pang, Stsfano Manetti, et al. Finding ambiguity groups in low testability analog circuits[J]. IEEE Trans on Circuits and Systems: Fundamental Theory and Applications, 2000, 47(8): 1125-1137.

[13] Stefano Manetti, Maria Cristina Piccirilli. A singular-value decomposition approach for ambiguity group determination in analog circuits[J]. IEEE Trans on Circuits and Systems: Fundmental Theory and Applications, 2003, 50(4): 477-487.

[14] Han Han, Wang Hou-jun. A new method for analog circuit fault diagnosis based on testability and chaos particle swarm optimization[J]. J of Convergence Information Technology, 2012, 7(14): 444-453.