第 33 卷 第 23 期 计 算 机 工 程 2007 年 12 月
Vol.33 No.23 Computer Engineering December 2007
·博士论文· 文章编号：1000—3428(2007)23—0019—03 文献标识码：A 中图分类号：TP311

# 基于奇异事件特征的时间序列相似模式匹配

曲文龙[1,2]，杨炳儒[2]，贺毅朝[1]
（1. 石家庄经济学院计算机系，石家庄 050031；2. 北京科技大学信息工程学院，北京 100083）

**摘 要：**现有的时间序列特征提取方法多为单尺度方法，导致特征点的时间定位不准确，从而影响模式发现的质量。该文基于小波奇异检测理论，提出了一种多尺度时间序列特征提取方法，利用奇异特征将时间序列压缩为事件序列表示，定义了事件序列动态时间弯曲相似度量，给出了基于事件序列相似模式匹配算法。实验表明，该方法具有较高的匹配精度和较低的计算代价。

**关键词：**时间序列；相似匹配；奇异事件；知识发现

# Time Series Similar Pattern Matching Based on Singularity Event Features

QU Wen-long[1,2], YANG Bing-ru[2], HE Yi-chao[1]
(1. Department of Computer Science, Shijiazhuang University of Economics, Shijiazhuang 050031;
2. Information Engineering College, University of Science and Technology Beijing, Beijing 100083)

**【Abstract】** The state-of-art features extraction methods from time series are single-scale methods that result in the location of features imprecision and suppress the quality of discovered pattern. A novelty multi-scale features extraction methods from time series is proposed based on the principle of wavelet singularity detection. The time series are compressed into event sequence using singularity features and a dynamic time warping similarity measure of event sequenced is defined. The proposed algorithm is used to similarity pattern matching for event sequence. The experimental result shows that it has higher matching precision and lower computing cost.

**【Key words】** time series; similarity matching; singularity event; knowledge discovery

时间序列分析与挖掘广泛应用于众多领域，相似模式匹配是时间序列分类、聚类、规则获取和预测的基础。典型的相似性测度多采用欧氏距离，但欧氏距离对数据在时间轴上的形变缺乏辨识能力。动态时间弯曲(dynamic time warping, DTW)[1]测度可以处理时间轴的变形，但在原始空间中计算代价很高。研究表明，人更善于识别时间序列的结构和定性相似，利用线性分段可以处理时间轴的变形、消噪和降维，减小计算复杂度，得到广泛应用。Guralnik提出了基于统计方法进行时间序列的变化点检测方法[2]，Keogh采用分段拟合方法进行线性分段[3]，两方法均以线性拟合的均方误差$\varepsilon$值作为时间序列的分割标准。Perng采用界标点作为特征进行相似匹配[4]，Pratt提出重要点相似匹配方法[5]。以上方法均为单尺度特征提取方法，而复杂时间序列受多种因素影响，难以在单一尺度获取其重要特征及其准确的发生时间。

本文依据信号的小波奇异检测理论[6]，提出了一种时间序列多尺度事件特征提取方法，准确地提取时间序列的形态特征和发生时间，将时间序列抽象为事件序列。定义了事件弯曲形态相似度量，给出了基于事件的相似匹配算法。

## 1 奇异事件特征的提取

奇异性是指信号本身或它的某阶导数在某一时刻存在突变，奇异点携带有比较重要的信息。本文利用从时间序列中抽取奇异特征实现对时间序列的事件分割和压缩表示，利用事件序列进行相似模式匹配。

### 1.1 基本概念及理论

数学上用李氏指数(Lipschitz)来描述函数或信号的奇异性[6]，定义如下：

**定义 1** 设 $n$ 是一非负整数，$n<a\le n+1$，如果存在两个常数 A 和 $h_0>0$，及 $n$ 次多项式 $p_n(h)$，使得对任意的 $h\le h_0$，均有

$$|f(x_0+h)-P_n(h)|\le A|h|^a \tag{1}$$

则称 $f(x)$ 在点 $x_0$ 为李氏指数 $a$。李氏指数越大函数越光滑。函数在一点连续、可微，则在该点的李氏指数为 1；在一点可导而导数有界但不连续时，李氏指数仍为 1。如果 $f(x)$ 在 $x_0$ 李氏指数不为 1，则称函数在 $x_0$ 点是奇异的。

**定义 2** 如果对任意 $x\in(a,b)$，$x^*\in(a,b)$，且 $x-x^*\in(a,b)$ 有式(1)成立，则称 $f(x)$ 在 $(a,b)$ 上是一致 *Lipschitz* $\alpha$ 的。

分析函数 $f(x)$ 的奇异性的传统方法是考察其傅立叶变换 $f(\omega)$ 的渐近衰减性。但傅立叶变换只给出 $f(x)$ 在全域的奇异性度量。利用小波可分析这种局部奇异性，小波系数的值取决于 $f(x)$ 在 $x_0$ 的邻域内的特性及小波变换所选取的尺度，在比较小的尺度上它提供了 $f(x)$ 的局部化性质。

**定义 3** 在尺度 $s_0$ 下，称点 $(s_0,x_0)$ 是局部极大值点，若

$$\frac{\partial Wf(s_0,x_0)}{\partial x}$$

**基金项目：**北京市自然科学基金资助项目(4022008)；河北省教育厅科研基金资助项目(Z2006313)
**作者简介：**曲文龙(1970－)，男，博士，主研方向：复杂类型数据挖掘；杨炳儒，教授、博士生导师；贺毅朝，硕士、副教授
**收稿日期：**2007-01-15 **E-mail：**quwenl@sohu.com

在 $x=x_0$ 有一过零点，则称 $(s_0,x_0)$ 为小波变换的模极大值点。若对属于 $x_0$ 的某一邻域内的任意点 $x$ ，有

$$|Wf(s_0,x)|\le|Wf(s_0,x_0)|$$

则尺度空间 $(s,x)$ 中所有模极大值点连线称为模极大值线。

**定义 4** 称小波 $\varphi(x)$ 具有 $n$ 阶消失矩，如果对于一切正整数 $k<n$ ，有

$$\int_{-\infty}^{\infty}x^k\varphi(x)\mathrm{d}x=0 \tag{2}$$

下面的定理给出了小波变换的模极大值与函数的奇异性 *Lipschitz* 指数的关系。

**定理** 设 $n$ 为一严格正整数，$\varphi$ 为一有 $n$ 阶消失矩、$n$ 次连续可微具有紧支集的小波 $f(x)\in L^1([a,b])$ ，则式(1)若存在尺度 $s_0>0$ ，使得 $\forall s<s_0$ ， $x\in(a,b)$ ， $|Wf(s,x)|$ 没有局部极大值点，则 $\forall\varepsilon>0$ 和 $a<n$ ， $f(x)$ 在区间 $(a+\varepsilon,b-\varepsilon)$ 是一致 *Lipschitz* $\alpha$ 。如果 $\varphi(x)$ 是某个平滑函数的 $n$ 阶导数，则 $f(x)$ 在该区间 $(a+\varepsilon,b-\varepsilon)$ 上是一致李氏指数 $n$。

定理证明了如果小波变换在更精细的尺度上没有模极大值，那么函数在该处任何邻域中无奇异性[6]。尺度从大到小变化，其模极大值点会聚为奇异点，构成一条模极大值线。不同的尺度上小波变换系数的模极大值对应于原始信号相应频率范围内的局部奇异点。尺度越大，小波系数模极大值越反映信号低频分量的局部奇异性。由于小波基函数本身具有传输延迟，因此在大尺度上实现奇异点定位后，应在精细尺度进行时移补偿以得到原始信号的准确定位。

### 1.2 奇异事件特征提取算法

为从时间序列中抽取特征，实现对时间序列的事件压缩表示，首先确定一个分析尺度 $s_A$ ，将该尺度 $s_A$ 下的模极值点数目作为事件特征数。对 $s_A$ 尺度的奇异特征点，从大尺度沿模极大值线向小尺度追踪，在时域对奇异事件特征点进行精确定位，从而将时间序列转化为事件序列。以下给出奇异事件特征提取算法：

**step1** 对待处理序列 $X=\{x_t\}$ $(t=1,2,\cdots,n)$ 。根据的分析尺度 $s_A$ ，计算最大二进分解层数L=$\lfloor\ln s_A/\ln 2\rfloor$。用选取的小波对序列实施 $2^0,2^1,\cdots,2^L$ , $s_A$ 尺度小波变换，得到小波变换系数矩阵

$$Wcoefs=\{w_{i,t}\}\ i=1,2,\cdots,L,L+1\ t=1,2,\cdots,n$$

$\{w_{L+1,t}\}$ 为 $s_A$ 尺度小波分解系数。

**Step2** 对 1～$L$+1 每一层小波分解系数序列 *Wcoefs* ，按定义 3 计算每层的模极大值点，得到模极大值标记矩阵

$$\boldsymbol{LocalMax}=\{LM_{i,t}\}$$

$$LM_{i,t}=\begin{cases}1, & if\ \ w_{i,t}\ \text{是模极值}\\ 0, & if\ \ w_{i,t}\ \text{非模极值}\end{cases}\quad i=1,2,\cdots,L,L+1\quad t=1,2,\cdots,n \tag{3}$$

$\sum_{t=1}^{n}LM_{i,t}$

为 $2^i(i=1,2,\cdots,L)$ 尺度和 $s_A(i=L+1)$ 的模极大值点数目。

**Step3** 对 $s_A$ 尺度的每一模极大值点 $(LM_{i,t}=1)$ ，利用 ***LocalMax***标记矩阵依次寻找其在 $2^L$, $2^{L-1},\cdots,1$ 尺度和原始序列 $X=\{x_t\}$ 的繁殖点，得到事件序列

$$EventSeq=\{(t_i,x_{t_i},w_{L+1,t_i})\}\quad i=1,2,\cdots,k$$

其中， $t_i$ 为第 $i$ 个特征点的水平位置； $x_{t_i}$ 为第 $i$ 个特征点的值； $w_{L+1,t_i}$ 为第 $i$ 个特征点在特征尺度 $s_A$ 的模极大值，可作为事件特征点的重要性度量；$k$ 为由 $s_A$ 尺度确定的事件特征点数目

$$k=\sum_{t=1}^{n}LM_{L+1,t}$$

为提高运算效率，算法采用了二进小波变换，对于尺度 $2^j$ 上一个模极大值 $x_1$，若它与尺度 $2^{j-1}$ 上的一个模极大值 $x_2$ 有相同的符号，位置也比较靠近且有较大的幅值，则认为 $x_2$ 为 $x_1$ 的繁殖点，利用这种规律可估计信号的模极大值线。

### 1.3 奇异事件检测小波的选取

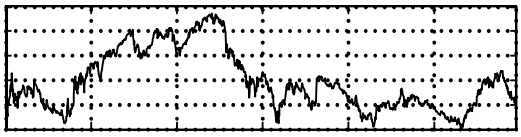
为检测时间序列的事件特征点(趋势拐点)，即一阶导数的间断点(其李氏指数 $a<2$ )，用于奇异事件检测的小波应具有对称性且要有二阶消失矩。选择具有低通特性的平滑函数 $\theta(x)$ 的二次导数做小波，对 $f(x)$ 实施卷积形式小波变换：

$$\begin{aligned}W_f^2(s,x)&=f(x)*\varphi_s^2(s)\\&=s\frac{\mathrm{d}}{\mathrm{d}x}[f(x)*\varphi_s^1(s)]=s^2\frac{\mathrm{d}}{\mathrm{d}x}[f(x)*\theta_s(x)]\end{aligned} \tag{4}$$

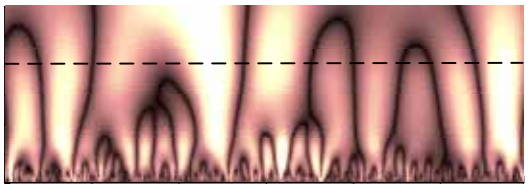
信号 $f(x)$ 的小波变换可表示为 $f(x)$ 在尺度 $s$ 被 $\theta_s(x)$ 平滑后信号的二阶导数，其模极值点即 $f(x)$ 在尺度 $s$ 被 $\theta_s(x)$ 平滑后信号的一阶导数的突变点，也就是时间序列的趋势拐点。

本文选择 Mexican-hat 作为奇异事件检测小波，它是由高斯平滑函数的二次导数导出，具有二阶消失矩。当尺度从小到大变化时，采用高斯平滑函数不会引入新的奇异点。试验发现 Mexican-hat 小波能有效检测和准确定位峰值奇异点，其时域解析式如下：

$$\psi(x)=\left(\frac{2}{\sqrt{3}}\pi\right)^{-\frac{1}{4}}(1-x^2)\exp(-\frac{1}{2}x^2),\ t\in\mathrm{R} \tag{5}$$



(a)原始时间序列



(b)连续小波变换表示

(c) 64 尺度奇异事件的模极大值线

**图 1 上证指数时间序列的奇异事件检测**

理论上利用模极大值线对奇异特征点进行精确定位，需对信号实施连续小波变换，为提高处理速度采用二进离散小波变换并采用启发式搜索模极大值线。图 1 为对 1999 年 6 月至 2003 年 5 月上证指数序列进行奇异特征提取，图 1(a)为原始时间序列，图 1(b)为采用 Mexican-hat 小波进行连续小

波变换的结果，图 1(c)在 64 尺度识别的奇异事件。

图 2 为对上证指数时间序列进行奇异事件特征提取(事件用内插直线段表示)，得到原始序列在 16 和 32 尺度事件序列近似表示。可见随着尺度的减小奇异事件数目逐步增加，近似精度随之逐步提高。

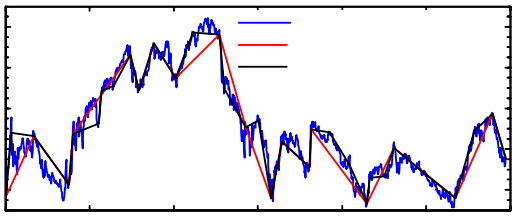


**图 2　利用 16 和 32 尺度的事件序列近似上证指数时间序列**

## 2　事件序列相似匹配算法

### 2.1　事件弯曲相似度量

利用提取的事件特征点进行事件序列的相似匹配，原始时间序列表示为如下事件(特征点)序列 $EventSeq=\{(t_i,x_{t_i},w_{L+1,t_i})\}(i=1,2,\cdots,k)$。相邻两点间的子序列表示具有相同趋势的事件，事件 $i$ 有一定持续时间：$t_{i+1}-t_i$，$(i=1,2,\cdots,k-1)$，每个事件可利用多项式回归模型表示，这里利用相邻点的内插直线简单表示。为进行事件序列相似匹配需定义两事件序列的相似度量，这里采用事件动态时间弯曲(event dynamic time warping, EDTW)测度来定义两事件序列的相似性。两个特征点之间的距离采用加权欧氏距离计算，权值 $\gamma$ 表示对特征点时间偏移的惩罚因子。

给定两事件序列 $Q=\{(qt_i,qx_{t_i})\}$ $(i=1,2,\cdots,k)$ 和 $C=\{(ct_j,cx_{t_j})\}$ $(j=1,2,\cdots,k^*)$，两个事件特征点距离为

$$d(q_i,c_j)=(qx_{t_i}-cx_{t_j})^2+\gamma(qt_i-ct_j)^2 \tag{6}$$

事件序列 $Q$ 和 $C$ 的最佳对准路径可以由时间起始点(1,1)到终点$(m, n)$之间的局部最优解递归获得：

$$S(1,1)=d(q_1,c_1)$$
$$S(i,j)=d(q_i,c_j)+\min\big(S(i-1,j),S(i,j-1),S(i-1,j-1)\big) \tag{7}$$

$S(i, j)$ 为累积距离，由当前对准点的距离和相邻点的累积 DTW 距离计算得到，则事件序列 $Q$ 和 $C$ 的事件动态时间弯曲距离定义为

$$EDTW(Q,C)=\min\left(\sqrt{\sum_{k=1}^{K}w_k}\right)\Big/K=\sqrt{S(n,m)}/K \tag{8}$$

其中，$K$为弯曲路径总数，分母中$K$是为消除不同的弯曲路径长度引起的偏差所进行的调整。采用事件相似度量，相对于原始时间序列的DTW距离具有较小的计算复杂度，设序列的平均长度为$n$，事件的平均长度为$d$，则DTW距离的运算时间为$O(n^2)$，事件DTW距离的运算时间为$O((n/d)^2)$，其运算速度提高$d^2$倍。定义EDTW是DTW距离的近似，采用这种事件表示还可以在一定程度上消除噪声的影响。

### 2.2　事件序列 k-最近邻相似匹配算法

**输入**　时间序列集合 *Set*，待匹配序列 $Q$，事件抽取尺度 $S$，最近邻数 $K$。

**输出**　$Q$ 的 $K$ 最近邻集 *Result Set*

(1) for each　$R \in Set$ ;
(2)对 $R$ 实施二进小波分解直到 $S$ 尺度；
(3)按式(4)提取事件特征，得到事件序列
　$ER=\{(rt_j,rx_{t_j})\}$ $(j=1,2,\cdots,k^*)$
(4)ADD *ER* to *EventSeqSet*;
(5)对　$Q$ 实施二进小波分解直到 $S$ 尺度；
(6)按式(4)提取事件特征，得到事件序列
　$EQ=\{(qt_i,qx_{t_i})\}$ $(i=1,2,\cdots,k)$
(7) for each $ER \in EventSeqSet$
(8)　　$DistR$＝EDTW($EQ$, $ER$)
(9)　　ADD $DistR$ to $DistSet$
(10) for $i$= 1 to $K$
(11)　　FIND *k-th* minimum　$DistR$　in $DistSet$
(12)　　ADD $R$ to *Resul Set*
(13) return *ResultSet*

## 3　试验结果

试验目的主要考察利用 EDTW 距离进行相似匹配的效率和质量。试验数据选择 UCI KDD 数据库中的 Synthetic ControlChart (Robert1999) 时间序列数据集(http://www.cs.ucr.edu/~eamonn/TSDMA /index.html)，包含 600 个长度为 60 的样本。数据分为：正常，周期，上升，下降，向上跳变，向下跳变 6 类，每类含 100 个样本。

设计 $K$ 最近邻分类试验，从每类中随机抽取 20 个序列共 120 个序列作为分类样本，其余 480 个序列共 60 个序列作为测试样本。分别采用 Euclidian,DTW 和 EDTW 距离作为相似度量，进行分类实验，$K$ 值取为 13。对于 EDTW 距离，选择 Mexican-hat 小波，信号延拓方式为 0 阶平滑；分析尺度为 2。EDTW 的时间偏差惩罚因子设为 $\gamma$ =0.2，实际计算中 DTW 限制最大弯曲度点数为 20，EDTW 限制最大弯曲事件数为 5。

表 1 是分别采用 3 种距离作为相似度量进行 K 最近邻分类的运行时间和准确率，结果表明： Euclid 计算速度快，但分类结果较差，DTW 和 EDTW 都具有很高的分类准确率，但 EDTW 的的计算时间比 DTW 快约 11 倍。因此采用事件序列的抽象表示可以大大降低 DTW 的计算量，同时保持很高的近似精度。

**表 1　3 种距离聚类结果的比较**

| 距离度量 | 计算时间/s | 正确数 | 准确率/(%) |
|---|---|---|---|
| Euclidean | 1.48 | 50 | 83.3 |
| DTW | 25.34 | 59 | 98.3 |
| EDTW | 2.29 | 58 | 96.7 |

表 2 结果表明：采用 EDTW 距离作为相似度量时，惩罚因子 $\gamma$ 值对分类准确率有一定影响。$\gamma$ 过小会导致时间轴过度弯曲，$\gamma$ 过大则使时间轴弯曲受到抑制，实验中 $\gamma$ =0.2 时得到了较高的分类正确率。

**表 2　不同的 $\gamma$ 值对 EDTW 分类正确率的影响**

| 惩罚因子 | 正确分类数 | 准确率/(%) |
|---|---|---|
| 0 | 57 | 95 |
| 0.2 | 58 | 96.7 |
| 0.4 | 56 | 93.3 |
| 0.6 | 55 | 91.7 |
| 1 | 54 | 90 |

## 4　结束语

针对复杂时间序列相似匹配问题，本文提出了一种多尺度事件特征提取方法。利用小波奇异检测方法提取时间序列的重要奇异特征点，作为事件划分依据，利用线性模型表示事件。方法利用了小波的多尺度特性，识别时间序列中奇异事件的准确定位。采用 $K$ 最近邻分类实验检验 EDTW 相似匹配的效率和准确性，结果表明该算法具有较小计算代价和很

(下转第 24 页)