时间序列的模式距离

王 达,荣 冈
(浙江大学 工业控制技术国家重点实验室,浙江 杭州 310027)

摘 要: 为了有效度量时间序列变化趋势的相似性,基于时间序列的分段线性表示,针对变化趋势,提出时间序列的模式模型表示.该模式模型表示不对测量尺度进行标准化处理,实现了模式距离的快速计算.序列模式距离克服了以点距离为基础的时间序列误匹配以及物理概念不明确等缺陷.对应于时间序列线性分段数目的不同,模式距离体现了多分辨特性,可以有效反应不同分析频率下时间序列的相似程度.

关键词: 时间序列;序列模式;序列距离
中图分类号: TP183 文献标识码: A 文章编号: 1008-973X(2004)07-0795-04

Pattern distance of time series

WANG Da, RONG Gang
(*National Key Laboratory of Industry Control Technology, Zhejiang University, Hangzhou* 310027, *China*)

Abstract The pattern model representation (PMR) of time series was proposed for measuring the trend similarity of the time series. PMR is based on piecewise linear representation and is effective at describing the tendency of time series. Because normalization was unnecessary, series pattern distance (SPD) was calculated rapidly. SPD overcomes the problem of time series mismatch based on point distance. According to the numbers of time series′ segmentations, pattern distance has multi scale feature and can reflect different similarity of time series under various analyzing frequency.

Key words time series; sequence pattern; sequence distance

时间序列的知识发现是数据挖掘的一个重要部分.特别是时间序列的变化趋势,反应了序列的动态特性,具有更高的使用价值.但目前的序列匹配,一般是对序列采用基于点距离的方法或改进方法进行匹配计算[1,2].为了防止测量中度量不同造成的误差,必须先对比较序列进行标准化处理,这大大增加了处理的计算量.而且采用不同的标准化方法会得到不同的距离,使得比较结果的物理概念不明确.以点距离为基础的方法对以“变化趋势”为兴趣的时间序列匹配存在本质的缺陷,因为点距离是一种静态的度量,无法有效体现时间序列的动态特性.如图 1所示,实际序列 1.3有着相近的变化趋势,而 1.2的变化趋势的相似性相对较小.但是基于距离的判别,则会把 1.2分在一组.基于点距离的方法不具备多分辨率特性,不能有效反应时间序列在不同分析频率下的相似性.以股票指数为例,基于点距离的分析无法对不同频率,比如以一周或一个月为周期的数据进行各种比较.

本文在时间序列的分段线性表示的基础上,进一步提出时间序列“模式”概念.它根据时间序列的变化趋势,将序列分成若干个子集,每个子集代表一种模式.在此基础上定义时间序列模式距离.

收稿日期: 2003-07-18. 浙江大学学报(工学版)网址: www.journals.zju.edu.cn/eng
基金项目: 国家“863”高技术研究发展计划资助项目(2001AA411210, 2001AA413220).
作者简介: 王达(1976—),男,浙江湖州人,博士生,从事数据挖掘,数据校正等方面研究. E-mail: dwang@iipc.zju.edu.cn
通讯联系人:荣冈,男,教授,博导. E-mail: grongapc@hzcnc.com

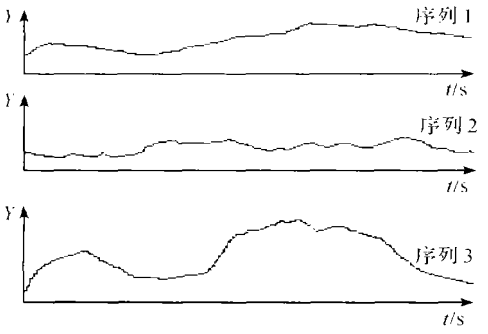

图 1 点距离分类缺陷示例

Fig. 1 Example of fault classification based on point distance

1 时间序列的模式表示

1.1 分段线性表示和模式表示

时间序列的分段线性表示(piecewise linear representation, PLR)是将时间序列表示成一段段相邻的直线,如图 2所示.时间序列 S的分段线性表示为

$$S=\{(y_{1L},y_{1R},t_1),\cdots,(y_{iL},y_{iR},t_i),\cdots,(y_{KL},y_{KR},t_K)\}. \quad (1)$$

式中: y_{iL}、$y_{iR}(i=1,2,\cdots,K)$分别表示第 i段直线起始值(左端)和终值(右端), t_i表示第 i段结束的时刻, K表示整个时间序列划分的直线段数目. Keogh[3]提出自底向上算法,很好地解决了时间线性表示的开放性问题,即如何选择合适的直线段数 K.

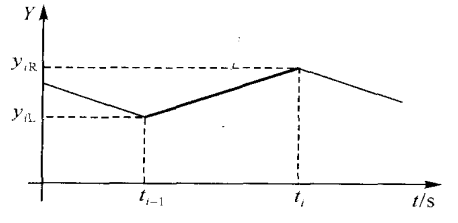

图 2 时间序列的分段线性表示

Fig. 2 Piecewise linear representation of time series

时间序列模式表示时间序列中某个子集单一的变化趋势.模式是三元集合{上升,保持,下降}.为便于计算,表示为 $M=\{1,0,-1\}$. $m\in M$表示模式集合中的一个元素.一个时间序列 S的模式表示为(模式,时刻)对的形式:

$$S=\{(m_1,t_1),\cdots,(m_i,t_i),\cdots,(m_N,t_N)\}. \quad (2)$$

式中: $m_i\in M, i=1,2,\cdots,N$; $t_1,t_2,\cdots,t_N$为模式结束时间; N为时间序列 S模式的划分数量.式(2)表示时间序列 S在 0~t_1时刻是模式 m_1, t_1~t_2时刻是模式 m_2,…, t_{N-1}~t_N是模式 m_N.

1.2 PLR2PMR算法

本算法将分段线性模型转化为时间序列模式.在计算 PLR每段模式并合并相同模式后,得到具有 N个模式表示的时间序列.其中 $N\leqslant K$,这使得针对模式匹配算法的搜索空间比 PLR更小.算法如下:

输入 1) PLR表示时间序列 S,如式(1)所示;
2)模式区分阈值 Th ($Th>0$).

输出 PMR表示时间序列 S',如式(2)所示.

```
S' = O
for i = 1: K
{
m = (y_iL − y_iR) / (t_i − t_{i−1});
if m ≤ − Th
    m = − 1;
else if m ≥ Th
    m = 0;
if LastMod(S') ≠ m
S' = [S', (m, t_i)];
else LastTime(S') = t_i
}
```

其中, LastMod(S')表示 S'的最后一个模式, LastTime(S')表示 S'的最后一个时间点.该算法仅对时间序列 PLR模型扫描一次,具有很高的计算效率.

2 时间序列的模式匹配计算

2.1 模式距离

用 s_i表示时间序列 $S=\{(m_1,t_1),\cdots,(m_N,t_N)\}$中第 i个模式,即 $s_i=(m_i,t_i)$, $i=1,2,\cdots,N$,则第 k个和第 p个模式差别可以用"模式距离"来度量.模式距离是表示具有相同保持时间长度的两个模式的距离.

$$D_M(s_k,s_p)=|m_k-m_p|. \quad (3)$$

式中: $s_k=(m_k,t_k)$, $s_p=(m_p,t_p)$.显然, $D_M(s_k,s_p)$是离散的,有

$$D_M(s_k,s_p)\in\{0,1,2\}. \quad (4)$$

当且仅当 $m_k=m_p$时,有 $D_M(s_k,s_p)=0$.

2.2 时间序列的模式距离

时间序列的模式距离是表示具有相同长度两个序列趋势的差异程度,是模式距离在时间序列上的应用.

S_1、S_2表示两个等长的待匹配时间序列模式模型, s_{1i}、s_{2j}分别表示序列 S_1和 S_2中的第 i、j个模式.

$$S_1 = \{(m_{11}, t_{11}), \cdots, (m_{1N}, t_{1N})\}, \quad (5)$$

$$S_2 = \{(m_{21}, t_{21}), \cdots, (m_{2M}, t_{2M})\}. \quad (6)$$

式中: $(m_{1i}, t_{1i}) = s_{1i}$, $i = 1, 2, \cdots, N$; $(m_{2j}, t_{2j}) = s_{2j}$, $j = 1, 2, \cdots, M$; $t_{1N} = t_{2M}$.

定义待匹配序列 S_1 和 S_2 的任意模式的保持时间:

s_{1i}的保持时间　$t_{1ih} = t_{1i} - t_{1(i-1)}$,

s_{2j}的保持时间　$t_{2jh} = t_{2j} - t_{2(j-1)}$.

当 $i = j$ 时,两者的保持时间相等,即

$$t_{1ih} = t_{2jh}. \quad (7)$$

在满足式(7)时,有 $M = N$.此时,定义 S_1 和 S_2 的序列模式距离为

$$D_M(S_1, S_2) = \frac{1}{t_N}\sum_{i=1}^{N} t_{ih} D_M(s_{1i}, s_{2i}). \quad (8)$$

且有

$$\frac{1}{t_N}\sum_{i=1}^{N} t_{ih} = 1. \quad (9)$$

式中: t_N 表示序列长度,t_{ih}表示第 i 个模式的保持时间,N表示序列模式的划分数量.

式(8)中的保持时间 t_{ih}的作用是对不同保持时间进行加权.保持时间越长,该模式在序列模式距离中所占的比重就越大.

由式(8)可知,S_1、S_2 的序列模式距离 $D_M(S_1, S_2) \in [0, 2]$.这个值可以反应如下状况:

① 序列模式距离越接近 0,表示待匹配序列的模式越接近;

② 在实际的序列中,模式 0很少(参考第 3节的计算用实例数据),$D_M(S_1, S_2)/2$表示了序列 S_1 和 S_2 趋势的差异程度,即当 S_1 在某种模式下时,S_2 有 $1 - D_M(S_1, S_2)/2$的可能性处于同样的模式下.$D_M(S_1, S_2)$越接近 0或 2时,程度值就越可靠;

③ 距离模式距离越接近 2,表示待匹配序列的模式越接近相反.

然而在实际的匹配序列中,满足式(7)的概率很小.大多数序列的模式表示模型中,模式的结束时间不等.此时无法直接用式(8)来计算时间序列间的模式距离.需要作进一步的处理,定义为等模式数(EPN)过程.

经过 EPN后,待匹配时间序列模式表示具有相同的模式结束时间,也就保持了对应的模式保持时间相等.以图 3中的两个时间序列为例:

$$S_1 = \{(1, t_{11}), (-1, t_{12}), (0, t_{13})\}, \quad (10)$$

$$S_2 = \{(1, t_{21}), (-1, t_{22})\}. \quad (11)$$

经过等模式数处理后得到

$$S_1' = \{(1, t_1), (-1, t_2), (-1, t_3), (0, t_4)\}, \quad (12)$$

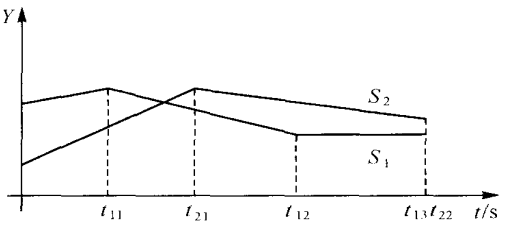

图 3　时间序列的模式模型

Fig. 3　Pattern model of time series

$$S_2' = \{(1, t_1), (1, t_2), (-1, t_3), (-1, t_4)\}. \quad (13)$$

式中: $t_1 = t_{11}$, $t_2 = t_{21}$, $t_3 = t_{12}$, $t_4 = t_{13} = t_{22}$.则式(10)和(11)的模型所表示的时间序列可以用式(8)计算它们之间的模式距离.

显然,待匹配序列间的模式距离计算可以与 EPN过程同时进行,减少数据库扫描次数,提高计算效率.

2.3　模式距离分析

时间容噪性: 由于采样网络时延等问题,实际的采样时间和记录时间会有一定的差距.可以把采样获得的(Value, Time)对表示为(Value, Time+ X),X表示采样时间误差.基于点距离的序列匹配方法要求计算待配序列所有点之间的距离,因此这些方法对时间噪声的要求十分苛刻.而时间序列的模式距离(8),由于具有模式保持时间加权因子 t_{ih},并且只有在模式切换点的时间噪声才可能改变时间序列的模式距离,同时,时间噪声引起的模式保持时间改变量 X远小于整个序列的时间长度,即 $X \ll t_N$,根据式(9),它对整个距离的影响很小,使得它对时间噪声具有天然的免疫力.

多分辨率特性: Pavilidis等人[4]指出 PLR表示有数据压缩和过滤作用.可以认为 PLR模型的不同段数具有不一样的滤波作用,称之为多分辨率特性.分段数目大,对应 PLR刻画信号细节部分;分段数目小,对应 PLR刻画信号长期表现,有低通滤波器作用.因此基于 PLR的模式距离可以用于度量时间序列在不同分辨率下的相似程度.

3　计算实例

采用图 4中从 1986年 6月 6日开始的 2 700个工作日的三种股票指数作为试验数据,分别为: S_1, Hong Kong (Hang Seng); S_2, New York (S& P 500); S_3, Paris (CAC40)①.由于要进行点距离比

① http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/data/FVDI.dat. 2002-12.

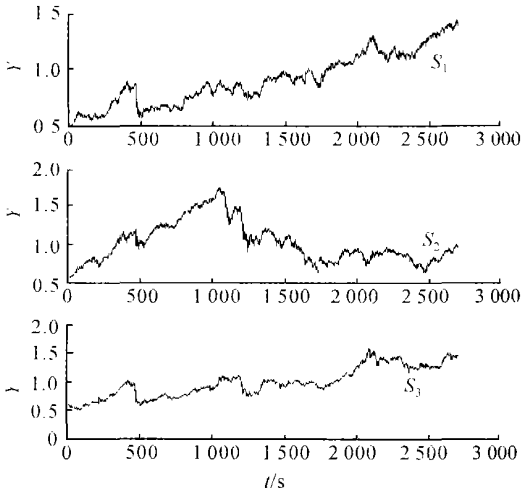

图 4 三种股票指数的标准化

Fig. 4 Normalization of three stock-indexes

较,对数据进行了相对均值的标准化处理.点距离采用被广泛使用的欧几里德距离,即

$$D(s_1, s_2) = \sqrt{\sum_i (s_{1i} - s_{2i})^2}.$$

3.1 点距离和模式距离比较

序列的 PLR模型的分段数目为 50.实验数据相互间的模式距离和点距离比较结果如表 1所示.

表 1 模式距离和点距离比较

Tab. 1 Comparison of pattern distance and point distance

	S_1与 S_2	S_1与 S_3	S_2与 S_3
模式距离	0.427	0.413	0.465
点距离	21.197	5.443	21.076

根据点距离划分,S_1和 S_3区别于 S_2,属于一类.这个划分无法说明类的具体物理属性,对序列趋势的相似程度也无法正确判断.但根据序列模式距离,可以得到如下结果:三种指数模式相互间趋势(上升、保持或下降)的相似程度十分接近,相互间有 80%的相似性,可以划分为一类.

3.2 多分辨率下模式距离的比较

在不同的分段数目下,比较实验数据间的模式距离,结果如表 2所示.

表 2 多分辨模式距离比较

Tab. 2 Comparison of multi scale pattern distance

段数	S_1与 S_2	S_1与 S_3	S_2与 S_3
10	0.358	0.724	0.628
20	0.823	0.375	0.942
50	0.427	0.413	0.465

从表 2中可以发现:当实验数据被分为 10段时,S_1与 S_2的模式非常接近;当分为 20段时,S_1与 S_3的模式距离最小,此时它们的模式很接近.对应不同的"分辨率",时间序列模式的相似程度会不一样.这个特性为数据的挖掘应用提供了新的属性,更有利于潜在信息的知识发现.

4 结 语

时间序列模式距离能够有效地度量序列变化趋势的相似程度,同时还可以表示序列趋势的相反程度,这是基于点距离的分析方法所不具备的.基于 PLR的滤波作用,模式距离还具有多分辨率特性,可以在不同的分辨率下有效地度量时间序列的模式距离,为数据挖掘提供隐藏信息.基于时间序列的模式模型,还可以在关联规则挖掘、序列的周期发现等方面作进一步的研究.

参考文献(References):

[1] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence database [A]. **Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithm** [C]. New York: Springer, 1993.

[2] CHAN K, FU W. Efficient time series matching by wavelets [A]. **Proceedings of the 15th IEEE International Conference on Data Engineering** [C]. Sydney: IEEE, 1999.

[3] KEOGH E. Fast similarity search in the presence of longitudinal scaling in time series databases [A]. **Proceedings of the 9th International Conference on Tools with Artificial Intelligence** [C]. Newport Beach: IEEE, 1997.

[4] PAVLIDIS T, HOROWITZ S. Segmentation of plane curves [J]. **IEEE Trans on Computation**, 1974, C23(8): 859-870.