时间序列序列模式的相似性研究

林 珣[1,2] 李志蜀[2] 周 勇[3]
(西南财经大学经济信息工程学院 成都 610074)[1] (四川大学计算机学院 成都 610064)[2]
(华兴职业技术学院 成都 610000)[3]

摘 要 时间序列序列模式相似性的度量是从时间序列中获取时序关联规则的重要环节。一般情况下,距离度量法只能度量相同长度序列模式的相似性。借用动态时间弯曲距离的思想,这种基于非线性弯曲技术的算法可以获得很高的识别、匹配精度。在定义元模式相似性的基础上,定义了序列模式的动态时间弯曲距离,最后用两个不同时间序列进行仿真实验,可以得到不同长度的序列的相似度。

关键词 时间序列,序列模式,相似性
中图法分类号 TP391 **文献标识码** A

Study on Pattern Similarity of Time Series

LIN Xun[1,2] LI Zhi-shu[2] ZHOU Yong[3]
(School of Economic Information Engineering, Southwestern University of Finance and Economics(SWUFE), Chengdu 610074, China)[1]
(School of Computer, Sichuan University(SCU), Chengdu 610064, China)[2]
(Huaxing Vocational and Technical College, Chengdu 610000, China)[3]

Abstract Under normal circumstances, the same distance measurement method can only measure the similarity of the same length sequence mode. Based on the similarity with definition meta model, this article drawed the idea of dynamic time warping distance and gave the definition of a sequential pattern of dynamic time warping distance, and finally conducted simulation experiments with two different time series.

Keywords Time series, Pattern, Similarity

时间序列时序关联规则是隐含在时间序列中的重要而有用的信息,这些信息能够很好地为时间序列的使用者提供决策者服务。时间序列序列模式相似性的度量是从时间序列中获取时序关联规则的重要环节。

1 时间序列元模式的概述

时间序列 $X=(x_1,x_2,\cdots,x_n)$,用所有局部极值点 $S=\{s_1,s_2,\cdots,s_m\}$ 把 X 分成 $m+1$ 个子序列(假设分割点中不含两个端点),简记为:

$$X=\begin{cases} l(x_1,s_1)[t_{x_1},t_{s_1}] \\ l(s_1,s_2)[t_{s_1},t_{s_2}] \\ \cdots \\ l(s_m,x_n)[t_{s_m},t_{x_n}] \end{cases} \quad (1)$$

用最小二乘法把每个子序列拟合成直线,其结果为:

$L_i(t)=a_it+b_i(i=1,2,\cdots,m+1)$

解出相邻两条直线的交点,可以得到 $m+1$ 条线段,取出每条线段的斜率 a_i 和交点的时间值 t_i,用斜率 a_i 和线段左右端点的时间值 $\Delta t_i=t_{ir}-t_{il}$ 把这条线段表示成模式:

$M_i=(a_i,\Delta t_i)(i=1,2,\cdots,m+1)$ (2)

并得到时间序列 $X=(x_1,x_2,\cdots,x_n)$ 模式序列为:

$S_X=\{(a_1,\Delta t_1),(a_2,\Delta t_2),\cdots,(a_{m+1},\Delta t_{m+1})\}$ (3)

定义1 称式(2)表示的模式为时间序列 X 的元模式,称式(3)表示的模式为时间序列 X 的模式序列。

定义2 时间序列 X 的模式序列为 $S_X=\{M_1,M_2,\cdots,M_m\}$,如果 S 是 S_X 的相邻元模式组成的子序列,即有 $S=\{M_i,M_{i+1},\cdots,M_{i+l}\}$,$(1\leqslant i,i+l\leqslant m)$,称 S 为 S_X 的序列模式。

定义3 时间序列 X 的模式序列为 $S_X=\{M_1,M_2,\cdots,M_m\}$,任意两个序列模式 $S_1,S_2\in S_X$ 的相似性函数为 $\mathrm{Sim}(S_1,S_2)$,对于给定正常数 ε,如果有:

$\mathrm{Sim}(S_1,S_2)\leqslant\varepsilon$

成立,那么称元模式 S_1 和 S_2 在以 ε 为界的情况下相似。

相似性函数 $\mathrm{Sim}(S_1,S_2)$ 对模式序列 S_X 中任意元模式 S_1、S_2 和 S_3 必须满足:

(1)正定性 $\mathrm{Sim}(S_1,S_2)\geqslant 0$;

(2)对称性 $\mathrm{Sim}(S_1,S_2)=\mathrm{Sim}(S_2,S_1)$;

(3)三角不等式 $\mathrm{Sim}(S_1,S_2)\leqslant\mathrm{Sim}(S_1,S_3)+\mathrm{Sim}(S_2,S_3)$。

2 时间序列的动态时间弯曲距离

动态时间弯曲(Dynamic Time Warping, DWT) 起初被应

到稿日期:2010-11-12 返修日期:2011-02-22 本文受国家自然科学基金(60803106)资助。
林 珣(1973-),女,博士生,讲师,主要研究方向为数据挖掘和商务智能,E-mail:linx_t@swufe.edu.cn;李志蜀(1947-),男,教授,博士生导师,主要研究方向为计算机网络、智能控制等;周 勇(1970-),男,副教授,主要研究方向为数据挖掘、粗糙集等。

用于文本数据匹配和视觉模式识别的研究领域。研究表明这种基于非线性弯曲技术的算法可以获得很高的识别、匹配精度。Berndt 和 Clifford 提出把 DWT 引入时间序列分析领域，在初步实验中取得了较好的结果[1-3]。

时间序列 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_m)$，定义距离矩阵：

$DM=(a_{ij})_{n\times m}$

式中，$a_{ij}=(x_i-y_j)^2$，a_{ij} 体现数据 x_i、y_j 之间的相异程度。

在距离矩阵 $DM=(a_{ij})_{n\times m}$ 中，把一组相邻矩阵元素的集合称为弯曲路径，设为 $W=w_1,w_2,\cdots,w_K$，W 的第 k 个元素为 $w_k=(a_{ij})_k$，这条路径满足下列条件：

(1) 有界性：$\max\{m,n\}\leqslant K\leqslant m+n-1$；

(2) 边界性：$w_1=a_{11}$，$w_K=a_{mn}$；

(3) 连续性：在弯曲路径中，相邻两个元素 $w_k=a_{ij}$，$w_{k-1}=a_{i'j'}$ 必须满足 $i-i'\leqslant 1$，$j-j'\leqslant 1$，即弯曲路径中的元素相互连续；

(4) 单调性：在弯曲路径中，相邻两个元素 $w_k=a_{ij}$，$w_{k-1}=a_{i'j'}$ 必须满足 $0\leqslant i-i'$，$0\leqslant j-j'$，即路径 W 通过点的同时必须至少通过点 $(i-1,j)$，$(i-1,j-1)$ 或 $(i,j-1)$ 中的一个，以保证弯曲路径是单调的。

整个矩阵表示两个时间序列的距离矩阵，一个方格就表示距离矩阵中的一个元素。如图 1 所示，从 a_{11} 到 a_{mn} 的路径就是一条弯曲路径。在两个时间序列的距离矩阵中，从 a_{11} 到 a_{mn} 弯曲路径并不唯一。不过，总可以找到一条弯曲路径 $W=w_1,w_2,\cdots,w_K$，使 $\sum_{i=1}^{K}w_i$ 达到最小，称这条路径 $W=w_1,w_2,\cdots,w_K$ 为距离矩阵的最佳弯曲路径[4-6]；把这个最小值称为时间序列 X 和 Y 动态时间弯曲距离，简记为 $D_{dtw}(X,Y)$，即有：

$$D_{dtw}(X,Y)=\min\{\sqrt{\sum_{i=1}^{K}w_i},W=w_1,w_2,\cdots,w_K\}$$

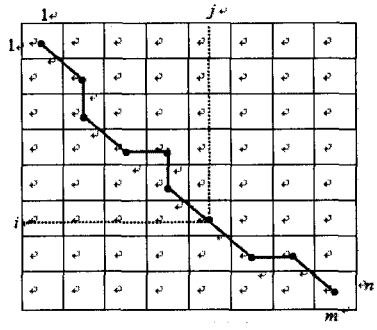

图 1　动态弯曲路径

两个时间序列的动态时间弯曲距离 $D_{dtw}(X,Y)$ 具有以下性质：

(1) 当 $m=n=0$ 时，$D_{dtw}(X,Y)=0$；

(2) 当 $m=n$ 时，$D_{dtw}(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$。

上面仅给出两个时间序列的动态时间弯曲距离，但并没有给出怎样求这个距离。理论上，可以利用穷举搜索法寻找最佳弯曲路径，但是完全穷举模式序列很长的情况下往往不切实际。弯曲路径有很多条，并且与距离矩阵中的元素数量成指数关系，这样时间的开销也非常大。如果点 a_{ij} 在最佳时间弯曲路径上，那么从点 a_{11} 到点 a_{ij} 的子路径也是局部最优解，即从点 a_{11} 到点 a_{mn} 的最佳路径可以由起始点 a_{11} 到终点 a_{NM} 之间的局部最优解通过递归搜索获得，即：

$$\begin{cases}s(1,1)=a_{11}\\ s(i,j)=a_{ij}+\min\{s(i-1,j-1),s(i,j-1),s(i-1,j)\}\end{cases}\tag{3}$$

式中，$i=2,3,\cdots,n$，$j=2,3,\cdots,m$。在距离矩阵中，弯曲路径的最小累加值 $s(m,n)$，可以从 a_{ij} 起沿弯曲路径按最小累加值倒退直到起始点 $s(1,1)$ 就可以找到最佳弯曲路径[7,8]，并求出这两个模式序列的动态时间弯曲距离 $D_{dtw}(X,Y)$。

从时间序列的动态时间弯曲距离 $D_{dtw}(X,Y)$ 中可以得到：(1) 动态时间序列弯曲距离可以求任意两个长度时间序列间的距离；(2) 在时间轴上的扭曲具有不敏敏感性[9,10]。

3　序列模式相似性的动态时间弯曲距离法

不论是序列模式的元模式单调距离和元模式向量距离，还是加权模式距离都只能求相同长度的两个序列模式间的距离[11]；动态时间弯曲距离法能够求两个不同维数的点之间的距离，把这种思想用于求不同长度序列模式间的距离。

时间序列 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_m)$，通过得到的模式序列分别为：

$S_X=\{M_1,M_2,\cdots,M_t\}$

$S_Y=\{N_1,N_2,\cdots,N_p\}$

式中，$M_i=(a_i,\Delta t_i)$ 是 S_X 的元模式，$N_j=(a_j,\Delta t_j)$ 是 S_Y 的元模式 $(i=1,2,\cdots,t,j=1,2,\cdots,p)$。建立这两个模式序列间的模式距离矩阵：

$DM=(a_{ij})_{t\times p}$

式中，$a_{ij}=D_{wei}(M_i,N_j)$ 表示模式序列 S_X 的第 i 个元模式 M_i 和 S_Y 的第 j 个元模式 N_j 的加权距离。a_{ij} 体现元模式 M_i、N_j 之间的相异程度：当元模式 M_i、N_j 越相似，a_{ij} 值越接近 0；当元模式 M_i、N_j 越不相似，a_{ij} 越大[11]。

与时间序列间的动态时间弯曲距离相似，在距离矩阵 $DM=(a_{ij})_{t\times p}$ 中，搜索[1] 最佳弯曲路径 $W=w_1,w_2,\cdots,w_K$，并把计算最佳弯曲路径的这个最小值称为模式序列 S_X 和 S_Y 的动态时间弯曲距离，简记为 $D_{dtw}(S_X,S_Y)$，即有：

$$D_{dtw}(S_X,S_Y)=\min\{\sqrt{\sum_{i=1}^{K}w_i},W=w_1,w_2,\cdots,w_K\}$$

两个模式序列[2] 的动态时间弯曲距离 $D_{dtw}(S_X,S_Y)$ 具有以下性质：

(1) 当 $M=N=0$ 时，$D_{dtw}(S_X,S_Y)=0$；

(2) 当 $M=N$ 时，$D_{dtw}(S_X,S_Y)=D_{wei}(S_X,S_Y)$。

当然 $D_{dtw}(S_X,S_Y)$ 也是用递归的方式进行搜索，即有：

$$D_{dtw}(S_X,S_Y)=D_{wei}(M_1,N_1)+\min\begin{cases}D_{dtw}(S_X,S_Y[2,|S_Y|])\\ D_{dtw}(S_X[2,|S_X|],S_Y)\\ D_{dtw}(S_X[2,|S_X|],S_Y[2,|S_Y|])\end{cases}$$

式中，$D_{dtw}(S_X,S_Y[2,|S_Y|])$ 表示模式序列 S_X 与 S_Y 中从第 2 个到最后一个元模式组成的序列模式之间的动态时间弯曲距

[1] 用递归的方法进行搜索。

[2] 如果取成模式序列的子序列，那么就成为序列模式。

离。

根据模式序列的动态时间弯曲距离可以求任意两个长度的模式序列间的距离。模式序列间的动态时间弯曲距离满足正定性、对称性和三角不等式性，因而，可用它定义两个模式序列的相似性。

定义 4 压缩时间序列 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_m)$ 之后得到的模式序列分别为：

$S_X=\{M_1,M_2,\cdots,M_t\}$

$S_Y=\{N_1,N_2,\cdots,N_p\}$

称 $D_{drw}(S_X,S_Y)$ 为模式序列 S_X 和 S_Y 的相似性函数，即有：

$\mathrm{Sim}(S_X,S_Y)=D_{drw}(S_X,S_Y)$

在判断两个模式序列是否相似时，只需要计算出它们之间的动态时间弯曲距离，再给定正常数 ε，如果相似性函数值小于 ε 时，那么它们就相似，否则它们不相似。

定义 5 把时间序列 $X=(x_1,x_2,\cdots,x_n)$ 转换成模式序列，结果分别为：

$S_X=\{M_1,M_2,\cdots,M_t\}$

任意序列模式 S_{X1}，$S_{X2}\in S_X$，$S_{X1}=\{M_j,M_{j+1},\cdots,M_{j+l_1}\}$，$S_{X2}=\{M_i,M_{i+1},\cdots,M_{i+l_2}\}$，称 $D_{drw}(S_{X1},S_{X2})$ 为序列模式 S_{X1} 和 S_{X2} 的相似性函数，即有：

$\mathrm{Sim}(S_X,S_Y)=D_{drw}(S_X,S_Y)$

式中，$1\leqslant i,i+l_1\leqslant t,1\leqslant j,j+l_1\leqslant t$。

4 仿真实验

用 2006 年沪深 300 指数的全部数据，以及 2007 年沪深 300 指数的全部数据作为实验对象，如图 2、图 3 所示。分别把它们转换成模式序列，分别含有 80 和 88 个元模式，应用模式序列动态时间弯曲距离方法计算它们的距离为 28.4。因而这两个时间序列的模式序列并不具有较好的相似性，这是因为 2006 年沪深 300 指数温和增长，而到了 2007 年几乎全民进入股市，促使 2007 年沪深 300 指数急剧增长，使它们具有很大的差异性。

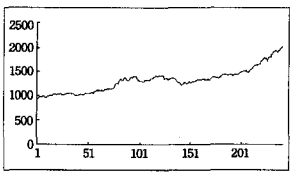

图 2 2006 年沪深 300 指数

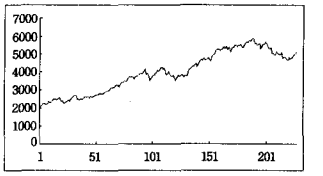

图 3 2007 年沪深 300 指数

参 考 文 献

[1] 翁颖钧，朱仲英. 基于分段线性动态时间弯曲的时间序列聚类算法研究. 研究与设计[J]. 微型电脑应用，2003，19(9)

[2] Berndt D J，Clifford J. Finding patterns in time series：DB，2000. A dynamic programming approach[Z]. Advances in Knowledge Discovery and Data Mining，1996

[3] Berndt J，Clifford D. Using dynamic time warping to find patterns in time series[C]// AAAI-94 Workshop on Knowledge Discovery in Database. 1994：229-248

[4] 曲文龙，张德政，杨炳儒. 基于小波和动态时间弯曲的时间序列相似匹配[J]. 北京科技大学学报，2006，28(4)

[5] Berndt D，Clifford J. Using dynamic time warping tp find patterns in time series[C]// AAAI Workshop on Knowledge Discovery in Databases. 1994：229-248

[6] Aach J，Church G. Alidning gene expression time series with time warping algorithms[J]. Bioinformatics，2001，17

[7] Kim S W，Park S，Chu W W. Efficient processing of similarity search under time warping in sequence databases：an index-based approach[J]. Inf. Syst.，2004，29(5)：405-420

[8] Das G，Lin K I，Marmila H，et al. Rule Discovery from Time Series [A]// Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining[C]. [S. l.]：AAA I Press，1998：16-22

[9] 张保稳. 时间序列数据挖掘研究[D]. 西安：西北工业大学，2002

[10] 王亮，姜丽红. 快速挖掘最大频繁模式算法[J]. 计算机工程与应用，2006(17)

[11] 周勇. 时间序列时序关联规则挖掘研究[D]. 成都：西南财经大学，2008(6)：40-60

(上接第 229 页)

[10] Sim K M，Guo Y Y，Shi B Y. BLGAN：Bayesian learning and genetic algorithm for supporting negotiation with incomplete information [J]. IEEE Transactions on Systems，Man，and Cybernetics，Part B：Cybernetics，2009，39(1)：198-211

[11] 李剑，牛少彰. 一种基于混合遗传算法的双边多议题协商[J]. 北京邮电大学学报，2009，32(2)：1-4

[12] Ng S C，Sulaiman M N，Selamat M H. Machine learning approach in optimizing negotiation agents for E-Commerce [J]. Information Technology Journal，2009，8(6)：801-810

[13] 程昱，高济，古华茂，等. 基于对手态度学习的协商决策模型[J]. 浙江大学学报：工学版，2008，42(10)：1676-1680

[14] 程昱，高济，古华茂，等. 基于机器学习的自动协商决策模型[J]. 软件学报，2009，20(8)：2160-2169

[15] Real C，Kersten G E，Vahidov R. Predicting opponent's moves in electronic negotiations using neural networks[J]. Expert Systems with Applications，2008，34(2)：1266-1273

(上接第 244 页)

[2] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州：苏州大学，2010

[3] Zhou Li-li，LI Fan-zhang. Research on Mapping Mechanism of Learning Expression[C]// Jian Yu，et al. Proceeding of Rough Set and Knowledge Technology. Beijing，China，October 2010：298-303

[4] 贺伟. 范畴论[M]. 北京：科学出版社，2006

[5] 李凡长，何书萍，钱旭培. 李群机器学习研究综述[J]. 计算机学报，2010，33(7)：115-1126

[6] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州：苏州大学，2010

[7] Jolliffe I T. Principal Component Analysis [M]. New York：Springer，1989

[8] Tenenbaum J B，de Silva V，et al. A global geometric framework for nonlinear dimensionaility reduction [J]. Science，2000，290(5500)：2319-2323

[9] 赵连伟，罗四维，等. 高位数据流形的低维嵌入及嵌入维数研究[J]. 软件学报，2005，16(8)：1423-1430

[10] Roweis ST，Saul LK. Nonlinear dimensionality analysis by locally linear embedding[J]. Science，2000，290(12)：2323-2326

[11] Tenenbaum J B，de Silva V，et al. A global geometric framework for nonlinear dimensionaility reduction [J]. Science，2000，290(5500)：2319-2323