ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 99－105
DOI:10.16337/j.1004-9037.2015.01.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 一种基于 DTW 的新型股市时间序列相似性度量方法

冯　钧　陈焕霖　唐志贤　吴　德

(河海大学计算机与信息学院,南京,211100)

**摘　要:** 现有时间序列相似性度量方法在进行股市序列相似性分析时,通常忽略成交量等其他重要因素对股价的影响,从而导致序列聚类、分类不精确。针对这一问题,本文提出了新的股市时间序列相似性度量方法。该方法在动态时间弯曲算法的基础上,通过引进时间衰竭因子,并结合成交量因素,给出了股市序列的最终度量公式。为了证明提出方法的可行性和有效性,本文实验部分通过选取家电等三个行业中的股票数据进行测试。实验结果表明,基于动态时间弯曲(Dynamic time warping,DTW)的新型股市时间序列相似性度量方法能够在保持股票序列形态特征的基础上,较好地解决股市技术分析中量价关系问题,从而更有效地应用于股市技术分析里关于模式发现等领域。

**关键词:** 时间序列;动态时间弯曲;分类

**中图分类号:** TP391.4　　**文献标志码:** A

## Similarity Measurement Method Based on DTW for Stock Time Series

Feng Jun, Chen Huanlin, Tang Zhixian, Wu De

(College of Computer and Information, Hohai University, Nanjing, 211100, China)

**Abstract:** The existing similarity measurement methods for stock time series always ignore the trading volume and other important factors influencing the stock price. This phenomenon results in inaccuracy when clustering and classifying the series. To solve the problem, a new similarity measurement method for stock time series is proposed. The method which is based on dynamic time warping(DTW) introduces time-exhaustion factor and trading volume factor, and puts forward the ultimate similarity measurement formula for stock time series. To prove the feasibility and validity of the method, the stock time series in the household appliances and two others in the experiment of this paper are tested. The test result indicates that the new similarity measurement method based on DTW can maintain the shape features of stock series. On this basis, the method can solves the problem of price-volume relationship in the stock technical analysis well. Thus the method can be applied to pattern discovery and other fields in the stock technical analysis for more effective results.

**Key words:** time series; dynamic time warping; classification

## 引　　言

随着海量数据日益增加,人们面临的主要问题是如何有效地利用这些数据。面对这一挑战,数据挖

**基金项目:** 国家自然科学基金(61370091)资助项目;江苏省科技支撑计划(工业)(BE2012179)资助项目;江苏省普通高校研究生科研创新计划(CXZZ12－0229)资助项目。
**收稿日期:** 2014-06-05;**修订日期:** 2014-06-30

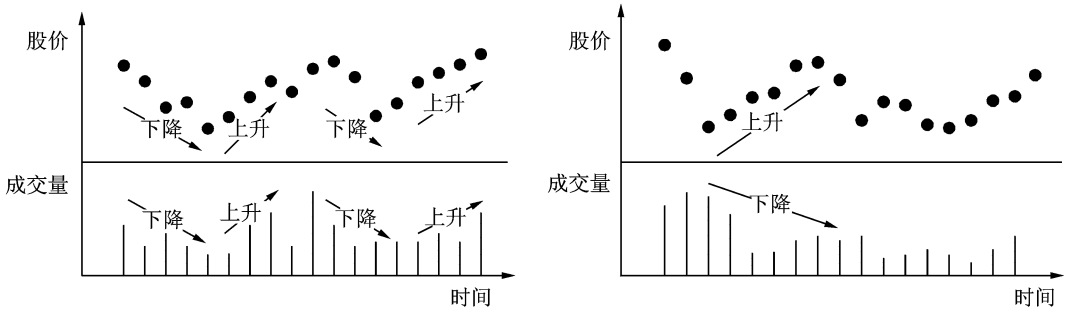
掘技术应运而生。面向股市数据的研究与分析一直是数据挖掘领域的研究热点。股市技术分析是基于“历史可以重演”这一原则才有效，而在现实的行情走势中历史不会简单地重复，却往往是惊人地相似。因此，研究其相似性度量是解决对其分类、聚类等诸多数据挖掘问题的基础，是进行技术分析的关键所在。

长期以来，相关领域学者对时间序列相似性度量方法进行大量研究，针对不同应用背景提出许多度量方法。然而，趋势是股票序列最重要的特征之一，而“价格”“成交量”“时间”“空间”是体现技术分析全局思想的四大要素，是反映趋势特性的基本要素，是技术分析的支柱[1]，因此必须予以综合考察。“价”“量”“时”“空”四大要素相互依存、相互影响，而现有的很多成熟技术如基于欧氏距离[2]、基于夹角余弦距离[3]等虽然简单易行，但由于其没有综合考虑“价”“量”“时”“空”四大要素，并未能保持其技术形态特征。针对以上问题，本文提出了适用于股市技术分析的序列相似性度量方法。

## 1 相关研究

国内外研究工作者对时间序列进行深入挖掘，针对不同研究重点提出各种行之有效的相似性度量方法。文献[4]提出模式距离，物理概念明确，划分合理，但表示方法粗糙，结论不够精确；基于欧氏距离[2]、基于夹角余弦距离[3]等虽然计算简单，结论明确，但前提是匹配序列长度必须相同；基于斜率距离[5]反映股价变化趋势，但在保持股价整体形态特性上有所欠缺，对整个股价序列信息的表达具有局限性，会出现匹配不精确等现象；而语音识别中的动态时间弯曲(Dynamic time warping, DTW)距离[6]有效地解决非等长序列的匹配问题，且能较好地保持股价趋势特征。而 Nanopoulos 等通过采集序列均值、标准差、偏度以及峰度，对序列模式进行分类[7]；文献[8]则提出先根据原始序列提取大量特征属性，然后过滤这些特征属性，最终根据过滤后的特征属性对原始序列进行分类。

然而，由于股市技术分析中“价”“量”“时”“空”四大要素相互依存、相互影响，但传统相似性度量方法通常只用于股价序列之间的匹配，往往容易忽略其他重要因素(如成交量)而导致最终的模式匹配不准确，如图 1，两个股虽然都录得“W”底的 k 线技术形态，若采用传统动态时间弯曲算法通常会聚为同一类，但从量价关系角度上看，图 1(a)量价配合良好，而图 1(b)则出现量价背离现象[9]，因此从股市技术分析上看这两个序列并非相似。




(a) 合肥三洋(600983)2009年8月13日至2009年9月7日收盘价及成交量
(a) Closing price and volume of stock 600983 from August 13th 2009 to September 7th 2009

(b) 阳光照明(600261)2009年2月24日至2009年3月19日收盘价及成交量
(b) Closing price and volume of stock 600261 from February 24th 2009 to March 19th 2009

图 1 两个个股的量价关系走势对比图

Fig. 1 Comparison of volume-price relation between two stock

针对以上问题，本文在传统 DTW 基础上，引进时间衰竭因子以反映 DTW 迭代过程中不同时间戳

上序列值的影响程度,同时采用“跳跃步长”为 2 与限制弯曲路径相结合的方法以防止“病态匹配”现象,并根据原始序列分布情况构造新序列,以反映量价配合情况。

## 2　基于 DTW 的新型股市时间序列相似性度量方法

### 2.1　动态时间弯曲距离

由于本文的核心算法是动态时间弯曲算法,因此,在详细介绍本文提出方法之前,首先介绍动态时间弯曲距离的定义。

**定义 1**　给定两个时间序列 $X=\{x_1,x_2,\cdots,x_m\}$ 和 $Y=\{y_1,y_2,\cdots,y_n\}$,构造累积距离

$$D(i,j)=d(x_i,y_j)+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\} \tag{1}$$

式中:$d(x_i,y_j)=\sqrt{(x_i-y_i)^2}$,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant n$,初始条件为 $D(1,1)=d(x_1,y_1)$。求其最终累积距离的过程实际上是在距离矩阵 $\boldsymbol{B}$ 中找到一条最佳的弯曲路径 $W$,使得累积距离最小,其中距离矩阵 $\boldsymbol{B}$ 为

$$\boldsymbol{B}=\begin{bmatrix} d(x_1,y_n) & d(x_2,y_n) & \cdots & d(x_m,y_n) \\ \vdots & \vdots & \cdots & \vdots \\ d(x_1,y_2) & d(x_2,y_2) & \cdots & d(x_m,y_2) \\ d(x_1,y_1) & d(x_2,y_1) & \cdots & d(x_m,y_1) \end{bmatrix} \tag{2}$$

弯曲路径 $W=(w_1,w_2,\cdots,w_k)$ 带约束条件,$\max(m,n)\leqslant k\leqslant m+n-1$,元素 $w_r=(i,j)$ 表示序列 $X$ 中的第 $i$ 个点和序列 $Y$ 中的第 $j$ 个点匹配,并且满足:(1)$W$ 起于矩阵 $\boldsymbol{B}$ 的左下角,止于矩阵 $\boldsymbol{B}$ 的右上角,即 $w_1=(1,1)$,$w_k=(m,n)$;(2)$W$ 上任意两个相邻的元素在距离矩阵 $\boldsymbol{B}$ 中也相邻,并向前发展,即若有 $w_k=(a_k,b_k)$,$w_{k+1}=(a_{k+1},b_{k+1})$,则满足 $0\leqslant a_{k+1}-a_k\leqslant 1$ 以及 $0\leqslant b_{k+1}-b_k\leqslant 1$。

距离矩阵中弯曲路径 $W$ 可到达的范围称为弯曲窗口。最终的度量距离定义为

$$D(X,Y)=\min\left(\sqrt{\sum_{(i,j)\in W}d(x_i,y_j)^2}/K\right) \tag{3}$$

其中 $K$ 为弯曲路径 $W$ 的长度。

### 2.2　基于 DTW 的新型股市时间序列相似性度量方法

#### 2.2.1　波动空间统一化

为了消除高价股与低价股在度量相似性时所产生的差异,在进行序列相似性计算之前,需要对序列进行预处理。

(1)关于“股价”序列的预处理:对两个股价序列 $P_1$ 和 $P_2$,分别求其序列均值 $E_1$ 和 $E_2$,将其比值 $\rho=\dfrac{E_2}{E_1}$ 作为调整因子,并进行如下调整

$$P_1'=\rho * P_1 \tag{4}$$

用新生成的序列 $P_1'$ 取代原始序列 $P_1$。

(2)关于“成交量”序列的预处理:对两个成交量序列 $V_1=\{v_{11},v_{12},\cdots,v_{1m}\}$ 和 $V_2=\{v_{21},v_{22},\cdots,v_{2n}\}$,从第二个时间点开始,分别求其相对于前一时间点的变化幅度,即

$$av_{1i}=(v_{1(i+1)}-v_{1i})/v_{1i} \tag{5}$$

式中 $i=1,2,\cdots,m-1$。

$$av_{2i}=(v_{2(i+1)}-v_{2i})/v_{2i} \tag{6}$$

式中:$i=1,2,\cdots,n-1$,分别形成新的序列 $AV_1=\{av_{11},av_{12},\cdots,av_{1(m-1)}\}$ 和 $AV_2=\{av_{21},av_{22},\cdots,av_{2(n-1)}\}$,为了解决新序列与原始序列的“时间点错位问题”(新序列的第 $i$ 个点表

示的是原始序列的第 $i+1$ 个点相对前一时间点的变化幅度),对每个新序列进行“首位添 0”操作(表示首个交易日成交量不涨不跌),即

$$AV_1=\{0,av_{11},av_{12},\cdots,av_{1(m-1)}\} \tag{7}$$

$$AV_2=\{0,av_{21},av_{22},\cdots,av_{2(n-1)}\} \tag{8}$$

同样地,对股价序列 $P_1$ 和 $P_2$ 进行以上操作,得到

$$AP_1=\{0,ap_{11},ap_{12},\cdots,ap_{1(m-1)}\} \tag{9}$$

$$AP_2=\{0,ap_{21},ap_{22},\cdots,ap_{2(n-1)}\} \tag{10}$$

2.2.2 改进的动态时间弯曲算法

DTW 算法最早应用于语音识别领域,文献[6]提出采用 DTW 解决时间序列模式发现问题。但是,DTW 进行序列匹配时,若弯曲路径 $W$ 过于偏离距离矩阵 $\boldsymbol{B}$ 对角线,则易出现“病态匹配”[10]现象。为避免这种情况,Keogh 等人[11]采用带约束弯曲窗口以限制弯曲路径 $W$ 偏离范围的解决方案。此外,当时间序列长度较短时,DTW 易受异常点影响而导致分类精确度不高。文献[12]采用“跳跃步长”为 2 的 DTW,以避免中间异常点对最终累积距离的影响。本文在前人研究基础之上,提出引进时间衰竭因子和成交量序列的 DTW 算法。具体步骤如下:

(1)对弯曲路径范围的限制

对于序列 $P'_1=\{p'_{11},p'_{12},\cdots,p'_{1M}\}$ 和 $P'_2=\{p'_{21},p'_{22},\cdots,p'_{2N}\}$,构造矩阵 $\boldsymbol{D}$,令

$$D(i,j)=+\infty \tag{11}$$

式中:$i=1,2,\cdots,\min\{\lfloor M/2\rfloor,\lfloor N/2\rfloor\}$;$j=\lfloor N/2\rfloor+i,\lfloor N/2\rfloor+i+1,\cdots,N$ 或 $i=\lfloor M/2\rfloor+1,\lfloor M/2\rfloor+2,\cdots,M$;$j=1,2,\cdots,i-\lfloor M/2\rfloor$。通过对弯曲路径的改进,避免弯曲路径过于偏离距离矩阵 $\boldsymbol{D}$ 的对角线而导致病态匹配的问题。

(2)根据距离矩阵 $\boldsymbol{D}$ 的不同区域,设置动态时间弯曲的“跳跃步伐 $r$”为 1 或 2

$$D(i,j)=d(p'_{1i},p'_{2j})+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\} \tag{12}$$

式中:定义 $d(x,y)=\sqrt{(x-y)^2}$,则 $d(p'_{1i},p'_{2j})=\sqrt{(p'_{1i}-p'_{2j})^2}$,下文同。$i=1,2;j=1,2,\cdots,\lfloor N/2\rfloor+i-1$ 或 $i=1,2,\cdots,\lfloor M/2\rfloor+j-1;j=1,2$。

$$D(i,j)=d(p'_{1i},p'_{2j})+\min\begin{Bmatrix}D(i-1,j),D(i-2,j),D(i,j-1),D(i-1,j-1)\\D(i-2,j-1),D(i,j-2),D(i-1,j-2),D(i-2,j-2)\end{Bmatrix} \tag{13}$$

式中:$i=3,4,\cdots,\lfloor M/2\rfloor$;$j=\lfloor N/2\rfloor,\lfloor N/2\rfloor+1,\cdots,\min\{(\lfloor N/2\rfloor+i-3),N\}$ 或 $i=\lfloor M/2\rfloor,\lfloor M/2\rfloor+1,\cdots,M$;$j=\lfloor N/2\rfloor,\lfloor N/2\rfloor-1,\cdots,\min\{(i-\lfloor M/2\rfloor+1),\lfloor N/2\rfloor\}$ 或 $i=\lfloor M/2\rfloor,\lfloor M/2\rfloor+1,\cdots,M$;$j=\lfloor N/2\rfloor,\lfloor N/2\rfloor+1,\cdots,N$。

(3)引进时间衰竭因子

由于时间因素的远近对股价影响程度不同(离当天越近影响越大,反之则相反),因此引进时间衰竭因子 $a_k=\dfrac{(M+N)/2-k}{(M+N)/2}$,$k$ 为距离当前时间点的时间长度,$k$ 越大,对当前时间点的影响越小,$k$ 越小,则对当前时间点的影响越大,基于此规则,改进式(12,13),即

$$D(i,j)=d(p'_{1i},p'_{2j})+\min\{a_1*D(i-1,j),a_1*D(i,j-1),a_1*D(i-1,j-1)\} \tag{14}$$

式中:$i=1,2;j=1,2,\cdots,\lfloor N/2\rfloor+i-1$ 或 $i=1,2,\cdots,\lfloor M/2\rfloor+j-1;j=1,2$。

$$D(i,j)=d(p'_{1i},p'_{2j})+\min\begin{Bmatrix}a_1*D(i-1,j),a_2*D(i-2,j),a_1*D(i,j-1),\\a_1*D(i-1,j-1),a_2*D(i-2,j-1),a_2*D(i,j-2),\\a_2*D(i-1,j-2),a_2*D(i-2,j-2)\end{Bmatrix} \tag{15}$$

其中,$i=3,4,\cdots,\lfloor M/2\rfloor$;$j=\lfloor N/2\rfloor,\lfloor N/2\rfloor+1,\cdots,\min\{(\lfloor N/2\rfloor+i-3),N\}$或 $i=\lfloor M/2\rfloor,\lfloor M/2\rfloor+1,\cdots,M;j=\lfloor N/2\rfloor,\lfloor N/2\rfloor-1,\cdots,\min\{(i-\lfloor M/2\rfloor+1),\lfloor N/2\rfloor\}$或 $i=\lfloor M/2\rfloor,\lfloor M/2\rfloor+1,\cdots,M;j=\lfloor N/2\rfloor,\lfloor N/2\rfloor+1,\cdots,N$。

(4)采用式(14,15),求其给定两序列 $P'_1$ 和 $P'_2$ 之间的最小累积距离 $D(m,n)$,同时记录最佳匹配路径 $W_o$ 和累积次数 $K$,并计算最小平均累积距离

$$D_P=(D(m,n)/E_2)/K \tag{16}$$

(5)沿最佳弯曲路径 $W_o$,结合序列 $AP_1$ 和 $AV_1$ 与 $AP_2$ 和 $AV_2$,按照以下规则统计各自量价关系分布情况:若某个时间点上成交量相对于前一个时间点的变化方向与股价变化方向一致且变化幅度大于或等于股价变化幅度,则视为量价配合程度良好;若某个时间点上成交量相对于前一个时间点的变化方向与股价变化方向一致且变化幅度小于股价变化幅度,则视为量价配合一般;若某个时间点上成交量相对于前一个时间点的变化方向与股价变化方向相反,则视为量价背离。

基于以上规则,分别构造两个记录量价关系分布情况的序列 $C_1$ 和 $C_2$,即

$$C_1=[c_{11},c_{12},\cdots,c_{1K}]$$
$$C_2=[c_{21},c_{22},\cdots,c_{2K}]$$

其中,若 $ap_{1i}\neq 0$,则

$$c_{1i}=\begin{cases}1.0 & av_{1i}/ap_{1i}\geqslant 10\\ 0.5 & 0\leqslant av_{1i}/ap_{1i}<10\\ 0.0 & av_{1i}/ap_{1i}<0\end{cases} \tag{17}$$

若 $ap_{2j}\neq 0$,则

$$c_{2j}=\begin{cases}1.0 & av_{2j}/ap_{2j}\geqslant 10\\ 0.5 & 0\leqslant av_{2j}/ap_{2j}<10\\ 0.0 & av_{2j}/ap_{2j}<0\end{cases} \tag{18}$$

若 $ap_{1i}=0$,则

$$c_{1i}=\begin{cases}1 & |av_{1i}|<0.01\\ 0 & \text{其他}\end{cases} \tag{19}$$

若 $ap_{2j}=0$,则

$$c_{2j}=\begin{cases}1 & |av_{2j}|<0.01\\ 0 & \text{其他}\end{cases} \tag{20}$$

其中 $(i,j)\in W_o$,计算序列 $C_1$ 与 $C_2$ 的"差值"$C'$

$$C'=\sum_{i=1}^{K}|c_{1i}-c_{2i}| \tag{21}$$

基于以上定义,有 $C'\in[0,K]$。

(6)根据 $C'$ 对最小平均累积距离 $D_P$ 进行修正,得到最终的相似性度量距离

$$D_M=D_P*(K-C')/K \tag{22}$$

## 3　实验方案

选取"W"底与"圆弧顶"两大类 k 线技术形态,行业包括沪深 A 股市场中房地产、家电、文化传媒三个行业,以当日收盘价为准,分别采集满足两类技术形态之一的个股,并记录股价及相应的成

交量,存储到相应的数据集中。

本实验方案采用 KNN 分类算法,数据集样本空间分为 4 类,分别为量价配合"W"底、量价背离"W"底、量价配合"圆弧顶"和量价背离"圆弧顶",每一类中,按照序列长度范围(分别为 1～20 组、21～40 组、41～60 组、61～80 组)抽取相同数量数据,以保证样本分布均匀,最后分别在每一组中随机抽取数据进行试验,最终分类准确度如图 2 所示。

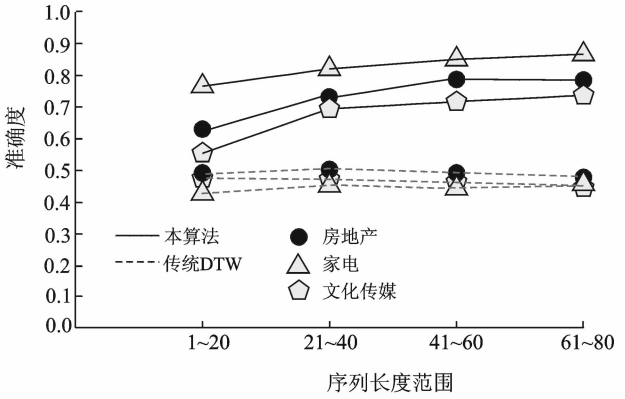



图 2 不同序列长度范围的分类准确度对比

Fig. 2 Comparison of classification accuracy among different sequence length

实验结果表明,相比于传统 DTW 算法,本文提出的算法分类准确度有较大提高。对于本算法而言,分类准确度与序列长度范围基本上呈正相关性,随着序列长度增加,对各数据集中各类数据进行分类的准确度会有所提高,但准确度增加幅度也会随之减缓。当序列长度比较大时,对其分类的准确度较高,从而有效解决了上文提到的股市序列匹配过程中通常遇到的量价关系问题。此外,不同行业的分类准确度对序列长度的敏感性也会有所不同,家电行业受其影响程度较低,而文化传媒行业则比较敏感,其分类准确度波动较大。

## 4 结束语

传统的时间序列相似性度量方法在计算两个序列的相似性时,通常忽略了除股价之外的其他重要因素,如成交量,这就导致了量价关系问题无法得到有效解决。因此,传统的相似性度量方法并不适用于股市技术分析。为了解决以上问题,本文提出了一种基于 DTW 算法的新型股市度量方法,并通过实验验证了这一度量方法的有效性和可行性。实验发现,当股市序列长度较长时,采用本文提出的新型度量方法进行相似性分析,能够有效解决不同量价关系的匹配问题,从而提高股市技术形态匹配的准确度和精确度。

**参考文献:**

[1] Michael N K. 技术分析入门[M]. 3 版. 北京:机械工业出版社,2011.
Michael N K. Technical analysis plain and simple[M]. The third edition. Beijing: China Machine Press, 2011.

[2] 刘懿,鲍德沛,杨泽红,等. 新型时间序列相似性度量方法研究[J]. 计算机应用研究,2007,24(5):112-114.
Liu Yi, Bao Depei, Yang Zehong, et al. Research of new similarity measure method on time series data[J]. Application Research of Computers, 2007, 24(5): 112-114.

[3] Tan P N, Steinbach M, Kumar V, et al. Wikipedia[EB/OL]. http://zh.wikipedia.org/wiki, 2013.

[4] 王达,荣冈. 时间序列的模式距离[J]. 浙江大学学报:工学版,2004,38(7):795-798.

Wang Da, Rong Gang. Pattern distance of time series[J]. Journal of Zhejiang University:Engineering Science, 2004,38(7):795-798.

[5] 张建业,潘泉,张鹏,等.基于斜率表示的时间序列相似性度量方法[J].模式识别与人工智能, 2007,20(2):271-274.
Zhang Jianye, Pan Quan, Zhang Peng, et al. Similarity measuring method in time series based on slope[J]. Pattern Recognition and Artificial Intelligence, 2007,20(2):271-274.

[6] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]//KDD workshop. [S. l.]: AAAI Press, 1994,10(16):359-370.

[7] Nanopoulos A, Alcock R, Manolopoulos Y. Information processing technology feature based classification of time-series data [C]//Commack. NY, USA: Nova Science Publishers Inc,2001:49-61.

[8] Fulcher B D, Jones N S. Highly comparative, feature-based time-series classification[J]. arXiv Preprint arXiv,2014,1401:3531.

[9] Cabbage, Dan Yixi. Wikipedia[EB/OL]. http://wiki.mbalib.com/wiki/, 2013.

[10] Chen Q, Hu G, Gu F, et al. Learning optimal warping window size of DTW for time series classification[C]//Information Science, Signal Processing and Their Applications (ISSPA), 2012 11th International Conference on. [S. l.]: IEEE, 2012:1272-1277.

[11] Ratanamahatana C A, Keogh E. Making time-series classification more accurate using learned constraints[C]// Proceedings of SIAM International Conference on Data Mining. Lake Buen Vista,Florida: [s. n.],2004:11-12.

[12] Feng L, Zhao X, Liu Y, et al. A similarity measure of jumping dynamic time warping[C]//Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on. [S. l.]: IEEE, 2010,4:1677-1681.

作者简介:**冯钧**(1969-),女,教授,研究方向:数据管理、时空索引和搜索方法、智能交通系统和领域数据挖掘,E-mail:fengjun@hhu.edu.cn;**陈焕霖**(1989-),男,硕士,研究方向:时间序列相似性度量方法;**唐志贤**(1983-),男,博士,研究方向:时空索引与搜索方法;**吴德**(1978-),男,博士,研究方向:时间序列相似性度量方法。