基于时间序列演变分析的有效相似性定义和聚类

周原冰,左新强,顾 杰,赵春晖
ZHOU Yuan-bing,ZUO Xin-qiang,GU Jie,ZHAO Chun-hui

国网北京经济技术研究院,北京 100761
State Power Economic Research Institute,Beijing 100761,China
E-mail:zhouyuanbing@chinasperi.com.cn

ZHOU Yuan-bing,ZUO Xin-qiang,GU Jie,et al.Effective similarity definition and clustering through time series evolution analysis.Computer Engineering and Applications,2008,44(10):138-141.

Abstract: Time series is one of the most widely-used data in business applications,e.g.power load sequence,web log etc.It is very important to mine time series for supporting decision-making.Especially,determining the similarity of time series plays a key part in various problems,e.g.analyzing the features of eletricity demand for each district.The previous methods,in the content of managing and mining data,hardly or do not enough use the evolution specialty of time series to measure similarity.This paper proposes an unexplored and effective approach based on evolution analysis of time series,and this approach quantifies the evolution trend to construct effective similarity definition,termed Similarity with Evolution Analysis (SEA).The clustering strategy based on SEA is also provided.The superior experimental results of compared methods on real data sets demonstrate the effectiveness of the method proposed,and thus imply the important significance of evolution analysis for similarity measure of time series.

Key words: time series;similarity definition;evolution analysis;clustering

摘 要:时间序列广泛存在于商业应用中,比如电力负荷序列、网络日志等。挖掘时间序列数据对决策分析非常重要,特别地,决定时间序列的相似性在各种实际问题中起关键的作用,比如分析各个区域的电力需求特征。以前的相似性度量方法从未使用过演变这种特性去度量时间序列的相似性,基于演变分析提出了有效的时间序列相似性度量方法(SEA),该方法通过量化演变趋势构建了有效的相似性定义,并且提出了基于该方法的聚类策略。通过在实际数据集上和其它方法的实验比较,证明了提出方法的有效性,因此也证明了时间序列演变分析对相似性度量的重要意义。

关键词:时间序列;相似性定义;演变分析;聚类
文章编号:1002-8331(2008)10-0138-04 **文献标识码**:A **中图分类号**:TP391.41

1 前言

时间序列作为一种有时间标志的数据类型,在各种领域中都有应用,如电力负荷序列、股票数据、传感器数据、室内温度数据、人口数据等。对于时间序列数据管理和挖掘系统来说,相似性定义是一个基本的需求,它决定了范围查询和 KNN(K-Nearest Neighbor)查询的返回结果,同时严重地影响分类和聚类等应用的精度,因此吸引了大量的研究工作者设计有效的相似性定义(或者是距离公式)。在相关的搜索和数据挖掘领域,以前的方法大多数将时间序列直接作为高维数据来度量,但是都没有涉及到时间序列的本质的演变特性,这种特性描述了一个变量随时间变化的趋势,比如每天的股票价格。

演变分析已经研究了很多年[1],然而没有任何工作通过时间序列的本质演变去提高相似度量的效果,在本文中,提出了一个新的基于演变分析的时间序列相似性定义, 主要的思想如下:

(1)时间序列 X 和 Y 的相似性描述了 X 演变为 Y 的概率(或者 Y 演变为 X);

(2)X 的演变具有方向和大小,也就是说 X 的每一维具有不同的变化权重,在有些方向上,X 不能到达;

(3)在某一时刻,数据的演变趋势可以通过下一时刻的数值得到。

由于演变趋势是时间序列所固有的本质,因此如果一个时间序列演变为另一个的概率很大, 那么它们就具有相似的本质,这是使用演变分析来度量时间序列相似性的核心思想。明显地,在某个确定的潜在规则下,将一个给定的时间序列变为一个随机的序列(潜在随机变化)几乎是不可能的,除非这个规则是完全随机的,一般来说,任何时间序列的变化都具有一定的限制,比如每天的温度几乎不可能高于 60℃,所以时间序列

作者简介:周原冰(1971-),男,高级工程师,主要研究领域为电力信息资源管理与决策支持、电力经济技术分析等;左新强(1982-),男,硕士,主要研究领域为数据挖掘、数据库;顾杰(1983-),硕士,主要研究领域为数据挖掘;赵春晖(1974-),女,高级工程师,主要研究领域为软件工程。

收稿日期:2007-07-23 **修回日期**:2007-10-24

的演变是具有一定规则或者限制的。提出的方法使用方向和大小量化地描述演变趋势,它们由时间序列在所有维度上的变化权重以及分布决定,时间序列中随时间连续变化的值满足分析趋势的需求,因此使用连续的值来判断前一时刻的演变趋势。

在提出的方法中,首先用一个映射的方法将时间序列转换为一个向量,该向量在高维空间中有一个起始点代表数据本身,它的方向和模代表其演变的方向和大小,则在其它方向上演变的大小可以通过投影来计算得到。时间序列 X 和 Y 的相似性距离是由 X 的演变向量在 $\boldsymbol{Y}_s-\boldsymbol{X}_s$ 方向上的投影和它们之间空间距离决定的,其中 $\boldsymbol{X}_s$ 和 $\boldsymbol{Y}_s$ 分别是 X 和 Y 映射向量。基于相似性定义,又提出了一个有效的聚类方法,最后,给出了实验,结果证明了提出方法的有效性。本文的主要贡献主要包括:

(1)在演变分析和时间序列相似性定义之间构建了一个有意义的连接;

(2)通过量化时序数据的演变趋势,设计了一个基于演变分析的有效相似性度量;

(3)提出了一个有效的聚类方法。

2　相关工作

近些年,在时间序列研究方面,有很多的工作和成果,如时间序列的分类[13]、聚类[3,8]、变化监测[10]、降维等[2,12]。在这些应用中,相似性度量是一个基本的过程,而且对系统的性能有很大影响。欧式距离(Euclidean Distance)是最流行的度量方法,因为它简单而且快速,并且能够支持各种索引技术,然而它不能处理局部的时间弯曲,这对时间序列度量方法的可用性是非常重要。Berndt[4]等提出用 DTW(Dynamic Time Warping)去度量时间序列的形似性,它允许时间轴上的弯曲,但是它对噪音很敏感,主要是由于它的连续弯曲路径造成的;LCSS(Longest Common Subsequence)具有抗噪音能力,但是由于不同的间隔存在于相似的形状中,造成了度量时的不准确[7];近来,Lei 等分别提出了 ERP(Edit distance with Real Penalty)[5]和 EDR(Edit Distance on Real sequence)[6]来度量多维时间序列的相似性,ERP 对噪音也很敏感,EDR 虽然不满足三角不等式,但是给出了一个近似的三角不等式。近期,Megalooikonomou 等提出了一种时间序列的多分解表示方法,并使用多层直方图模型来度量新的表示之间的距离[14],Vlachos 等通过萃取的周期特征来度量相似性,而不是时间点上的数值。在相似性度量之前,有一个简单但是非常重要的预处理,就是归一化[15],它是为了去除时间序列上值的绝对大小对相似性度量的影响[8,9]。

虽然已经提出了多种相似性度量方法来提高效果,但是度量的结果还是要依赖于数据和任务本身,造成很多方法在实际中不可用,这主要是因为以前的方法从未考虑过数据的本质特征,也就是说它们的度量过程和数据本身是无关的。演变分析是一种发现事物发展规律或者趋势的技术,可以被用来分析时间序列的本质趋势[11]。本文通过将的演变分析引入到度量过程中来提高相似性度量的效果。

3　基于演变分析的相似性:SEA

通过估计时间序列之间的演变概率,提出的方法 SEA 设计了全新的度量机制,其中演变概率是通过时间序列的演变趋势和普通的距离公式计算得到的。方法的主要步骤包括:

步骤 1　映射:一个长度为 n 的时间序列被映射到 $n-1$ 维的空间中;

步骤 2　演变趋势分析:用一个向量去表示时间序列的演变趋势;

步骤 3　相似性距离:设计一个有效的公式去计算两个时间序列的距离。

3.1　演变趋势映射

$X=x_1,x_2,\cdots,x_n$ 表示一个时间序列,其中每个值 x_i 对应一个时间点,$X[i]=x_i$ 表示在第 i 个取样时间点上的值,$|X|=n$ 表示时间序列 X 的长度为 n,映射过程通过以下步骤完成:

(1)X 被映射为 $n-1$ 维空间中的点,定义为 $\boldsymbol{X}_s=(x_1,x_2,\cdots,x_{n-1})$(or $\boldsymbol{X}_s$)。这个简单的过程主要是为了保留原始序列的静态信息,它定义了演变的基本点,也就是开始点。

(2)演变向量被形式化地定义为 $\boldsymbol{X}_v=(x_2-x_1,x_3-x_2,\cdots,x_n-x_{n-1})$。这一步使用一阶差分来描述 $\boldsymbol{X}_s$ 的趋势,考虑时间序列的意义,x_{i+1} 正是 x_i 的下一个抽样值,因此使用 $x_{i+1}-x_i$ 来描述 x_i 的动态趋势是有道理的。在 SEA 中,使用 $\boldsymbol{X}_v$ 来表示 $\boldsymbol{X}_s$ 拥有最大演变概率的方向,也就是说 $\boldsymbol{X}_s$ 最容易向这个方向演变,然后,通过在某个方向映射来得到 $\boldsymbol{X}_s$ 在下一个时间点向该方向演变的概率。

(3)X 被表示为一个特殊的向量,定义为 $\boldsymbol{X}_r=(\boldsymbol{X}_s,\boldsymbol{X}_v)$,它拥有一个确定的起始点 $\boldsymbol{X}_s$,而不允许平移,$\boldsymbol{X}_r$ 的方向和模定义为 $\boldsymbol{X}_v$ 的方向和模。事实上,$\boldsymbol{X}_r$ 的结束点就是$(x_2,x_3,\cdots,x_n)$,也就是 $\boldsymbol{X}_s$ 在下一个时刻的值。

以上的定义可以通过以下的例子说明:$X=(4,5,6,7)$的长度等于 4,然后 X 将被映射到三维空间中,其中 $\boldsymbol{X}_s=(4,5,6)$,$\boldsymbol{X}_v=(1,1,1)$。

3.2　演变相似性距离

在给定时间序列的岸边趋势映射后,本节将定义两个序列间的演变相似性距离(Evolution Similarity Distance,ESD)。给定两个时间序列 X 和 Y,长度均为 n,通过以下三步来计算它们之间的相似性(或者说不相似性),如图 1 所示。

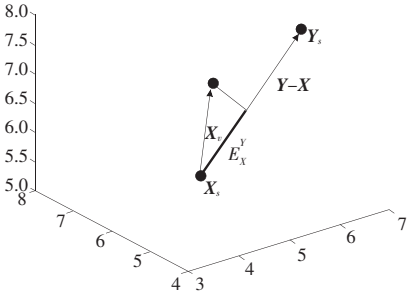


图 1　展示 SEA 方法的例子

(1)将 X 映射到多维空间,表示为 $\boldsymbol{X}_r=(\boldsymbol{X}_s,\boldsymbol{X}_v)$。

(2)将 $\boldsymbol{X}_r$ 投影到 $\boldsymbol{Y}_s-\boldsymbol{X}_s$ 方向上,投影的长度定义为 E_X^Y,它用来表示 X 在 $\boldsymbol{Y}_s-\boldsymbol{X}_s$ 方向上的演变概率,所以 E_X^Y 可以通过下式计算得到:

$$E_X^Y=\frac{\boldsymbol{X}_s\cdot(\boldsymbol{Y}_s-\boldsymbol{X}_s)}{|\boldsymbol{Y}_s-\boldsymbol{X}_s|} \tag{1}$$

其中 $\boldsymbol{X}_s\cdot(\boldsymbol{Y}_s-\boldsymbol{X}_s)$表示 $\boldsymbol{X}_s$ 和$(\boldsymbol{Y}_s-\boldsymbol{X}_s)$的叉积,$|\boldsymbol{Y}_s-\boldsymbol{X}_s|$是$(\boldsymbol{Y}_s-\boldsymbol{X}_s)$的模。

(3)计算 $\boldsymbol{X}_s$ 和 $\boldsymbol{Y}_s$ 的空间距离,记为 $\mathrm{d}(\boldsymbol{X}_s,\boldsymbol{Y}_s)$,其中 d 是欧

式距离公式。

定义 1 演变相似性距离(ESD)。时间序列 X 和 Y 的距离定义如下:

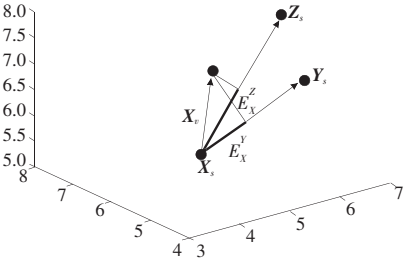
$$ESD(X,Y)=\Theta(E_X^Y,d(\boldsymbol{X}_s,\boldsymbol{Y}_s)) \quad (2)$$

函数 $\Theta(\alpha,\beta)$:$R^+\times R^+\to R^+$满足以下属性,其中 R^+是非负实数集合。

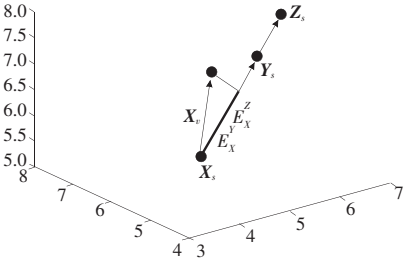
属性 1 $\Theta(\alpha,\beta)$是一个关于 α 的递减函数,也就是说 $\forall\alpha1,\alpha2,\beta\in R^+$,当 $\alpha1<\alpha2$ 时,$\Theta(\alpha1,\beta)>\Theta(\alpha2,\beta)$。

属性 2 $\Theta(\alpha,\beta)$是一个关于 β 的递增函数,也就是说 $\forall\alpha,\beta1,\beta2\in R^+$,当 $\beta1<\beta2$ 时,$\Theta(\alpha,\beta1)<\Theta(\alpha,\beta2)$。

假如 Θ 是可导的,则以上属性可以被描述为$\frac{d\Theta}{d\alpha}<0$和$\frac{d\Theta}{d\beta}>0$。给定三个时间序列 X、Y 和 Z,通过在以下特殊情景(如图 2 所示)中比较 $ESD(X,Y)$和 $ESD(X,Z)$,可以解释以上属性的合理性。



(a)情景 1



(b)情景 2

图 2　展示 ESD 方法的例子

情景 1 $E_X^Y<E_X^Z\wedge d(\boldsymbol{X}_s,\boldsymbol{Y}_s)=d(\boldsymbol{X}_s,\boldsymbol{Z}_s)$。在这种情况下,$\boldsymbol{X}_s$ 到 $\boldsymbol{Y}_s$ 和 $\boldsymbol{Z}_s$ 的距离相同,但是 $\boldsymbol{X}_s$ 更倾向于变为 $\boldsymbol{Z}_s$,因此 $ESD(X,Y)>ESD(X,Z)$。

情景 2 $E_X^Y<E_X^Z\wedge d(\boldsymbol{X}_s,\boldsymbol{Y}_s)<d(\boldsymbol{X}_s,\boldsymbol{Z}_s)$。在这种情况意味着 $\boldsymbol{X}_s$ 变为 $\boldsymbol{Y}_s$ 和 $\boldsymbol{Z}_s$ 的倾向度是相同的,但是 $\boldsymbol{Y}_s$ 更接近于 $\boldsymbol{X}_s$,因此 $ESD(X,Y)<ESD(X,Z)$。

从 ESD 的定义过程中可以看出,该度量不是对称的,也就是说 $ESD(X,Y)\neq ESD(Y,X)$,$ESD(X,Y)$和 $ESD(Y,X)$的最小正值可以用来作为最后的结果,函数 Θ 可以简单的设置为 $\Theta(\alpha,\beta)=c\cdot\frac{\beta}{\alpha}$,在实验中采用该式来计算 ESD,其中 c 是一个正的常数;其它的常用函数也可以被用来实现 Θ 以满足实际的需要,比如指数函数 $\Theta e(\alpha,\beta)=e^{\beta-\alpha}$,其中 $e>1$。

4　基于 SEA 的层次聚类方法

本章介绍了一种新的基于 SEA 的层次聚类策略,它是基于传统层次凝聚聚类(Hierarchical Agglomerative Clustering, HAC)[8]的,首先,两种时间序列关系定义如下:

定义 2 可达性。假如 $E_X^Y>0$,则定义为 X 是可达 Y 的,否则,X 是不可达 Y 的。

定义 3 连通性。假如 X 是可达 Y 的或者 Y 是可达 X 的($E_X^Y>0$ 或者 $E_Y^X>0$),则定义为 X 和 Y 是连通的,否则它们是不连通的。这个关系满足对称性,但是不满足传递性,可以通过以下的例子证明:

$X=(0,0,1)$,则 $\boldsymbol{X}_s=(0,0)$,$\boldsymbol{X}_v=(0,1)$;$Y=(2,1,1)$,则 $\boldsymbol{Y}_s=(2,1)$,$\boldsymbol{Y}_v=(-1,0)$;$Z=(3,4,4)$,则 $\boldsymbol{Z}_s=(3,4)$,$\boldsymbol{Z}_v=(1,0)$那么可以得到,$E_X^Y>0$,$E_X^Z>0$,但是 $E_Y^Z<0$ 且 $E_Z^Y<0$,根据连通性的定义,虽然 X 和 Y 连通,X 和 Z 连通,但是 Y 和 Z 是不连通的。

不同于普通的层次聚类,提出的聚类策略中增加了以下约束:

聚类约束:在聚类凝聚过程中,X 可以被分到类 C 中,仅当 C 中至少有一个时间序列和 X 连通。

那么可以通过设置不连通的序列间的距离为无穷大(∞)来的得到新的层次聚类算法,因为无限的距离存在于聚类过程中,则需要重新设计方法去计算类之间的距离以及聚类结束条件。首先定义两个类之间的连通度(Connective-Degree, ConnD),给定两个类 $A=A_1,A_2,\cdots,A_n$ 和 $B=B_1,B_2,\cdots,B_m$,其中 $A_i(1\leqslant i\leqslant n)$和 $B_j(1\leqslant j\leqslant m)$是类中的时间序列,$n$ 和 m 分别是类 A 和类 B 的大小,则 A 和 B 的连通度定义如下:

$$ConnD(A,B)=\frac{|\{(A_i,B_j)|A_i \text{ and } B_j \text{ are connective}\}|}{n\times m} \quad (3)$$

要计算两个类之间的距离,需要度量两个因素,主要包括连通度(Connective-Degree, ConnD)和一般化距离(General-Distance, GenD),其中 GenD 的计算和普通的 HAC 中的类间距基本相似,比如最小距离、最大距离、完全距离和平均距离。给定类 A、B、C 和 D,通过以下方法,比较 A、B 和 C、D 的类间距(CDistance),找到距离较小的两个类,在聚类过程中将它们合并:

(1)$CDisttance(A,B)<CDistance(C,D)$,假如 $ConnD(A,B)>ConnD(C,D)$或者 $ConnD(A,B)=ConnD(C,D)$而且 $GenD(A,B)<GenD(C,D)$;

(2)$CDisttance(A,B)=CDistance(C,D)$,假如 $ConnD(A,B)=ConnD(C,D)\wedge GenD(A,B)=GenD(C,D)$。

也就是说,计算类间距的时候,ConnD 比 GenD 的优先级高;此外,计算 GenD 时,取消遇到的无限距离(两个时间序列是不连通的),以便得到有限的结果。在聚类中,为了满足提出的聚类约束,除了传统 HAC 的类数和最大距离外,增加了新的结束条件如下:

新的聚类连通性结束条件:对于当前任意的两个类 A 和 B,$Conn(A,B)=0$。

完整聚类算法(使用类数作为结束条件之一)如下:

```
Function:基于 SEA 的层次距离算法
输入:T,n,k;// T 是包含所有时间序列的集合;n 是 T 的大小;k 是预设的类数
输出:{C1,C2,…,Cm} //Ci 是第 i 个类,m≥k
1.for i=1:n
2.    初始化 Ci={Ti} //Ti 是 T 中第 i 个序列
3.end
4.while n>k
```

```
5.      计算两两类之间的 GonnD 和 GenD
6.      if(连通性结束条件满足)
7.              return 当前的类集合
8.      else
9.              合并 CDistance 最小的两个类
10.             n=n-1
11.     end
12.end
13.return 当前的类集合
```

5　实验评价

在本章中,通过聚类实验(clustering)来测试提出的方法和其他方法的性能,公用的实际数据集被用在测试中。

5.1　实验框架

比较的方法主要有:欧式距离(ED)、DTW、LCSS、EDR 以及本文提出的方法 SEA。

实验数据:TEMP 是一个温度数据集[1],包括美国 30 个地方 2000 年的温度,每一个时间序列代表一个地方 366 天的温度变化,根据地理位置 30 个时间序列被分为 4 类,作为度量标准结果。

评价方法:凝聚的层次聚类用来实现其它方法的聚类实验,其中簇间距使用完全距离(complete distance)计算,使用一种常见方法来度量聚类的精度,给定每个方法的聚类结果 $C'=C_1',C_2',\cdots,C_k'$和从先验的分类信息中得到的标准结果 $C=C_1,C_2,\cdots,C_k$,则聚类精度可由以下公式计算[8]:

$$\mathrm{Sim}(C_i,C_j')=\frac{2|C_i\cap C_j'|}{|C_i|+|C_j'|}$$

$$\mathrm{Sim}(C,C')=\frac{\sum_i \max_j \mathrm{Sim}(C_i,C_j')}{k}$$

$\mathrm{Sim}(C',C)$的计算和 $\mathrm{Sim}(C,C')$类似,因为 $\mathrm{Sim}(C,C')$和 $\mathrm{Sim}(C',C)$不对称,故都被作为最终结果。

5.2　实验结果

图 3 给出了在 TEMP 数据集上的聚类结果,从图中可以看出,SEA 的精度比其它方法都要高,这说明了提出的方法能够更准确地度量时间序列的相似性,进一步表明了演变分析对时间序列度量和聚类的重要提高。

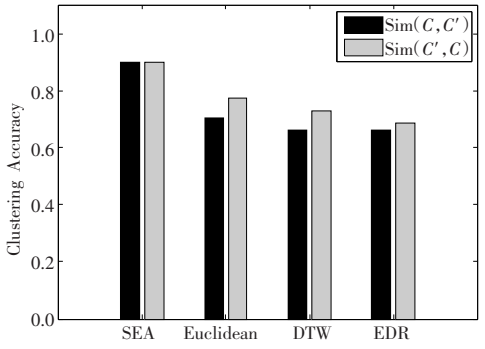


图 3　聚类精度比较

6　结论和总结

本文通过研究演变分析,提出了有效的时间序列相似性定义,量化了演变趋势,将每一个时间序列转换为一个向量,该向量表示了演变的方向和大小。通过将时间序列的演变向量映射到指向其它序列的方向,给出了新的相似性度量公式,并且提出了基于该方法的新的聚类策略。通过在真实数据集上的实验,结果证明了提出方法的有效性,因此也证明了演变分析对构建有力的相似性度量方法的重要性。

[1] http://www.csee.umbc.edu/~kalpakis.

参考文献:

[1] Gustavo D,Bernd S.Learning time series evolution by unsupervised extraction of correlations[J].Physical Review E,1995,51(3):1780-1790.

[2] Agrawal R,Faloutsos C,Swami A N.Effcient similarity search in sequence databases[C]//International Conference of Foundations of Data Organization and Algorithms,Chicago,Illinois,1993:69-84.

[3] Bagnall A J,Janacek G J.Clustering time series from arma models with clipped data [C]//ACM International Conference on Knowledge Discovery and Data Mining,Seattle,2004:49-58.

[4] Berndt D J,Clifford J.Using dynamic time warping to find patterns in time series[C]//AAAI Workshop on Knowledge Discovery in Database,Washington,1994:359-370.

[5] Chen Lei,Ng R.On the marriage of l_p norms and edit distance[C]//International Conference on Very Large Data Bases,Toronto,2004:792-803.

[6] Chen Lei,Tamer Ozsu M,Oria V.Robust and fast similarity search for moving object trajectories[C]//ACM SIGMOD International Conference on Management of Data,Baltimore,Maryland,2005:491-502.

[7] Gautam Das,Dimitrios Gunopulos,Heikki Mannila.Finding similar time series[C]//European Symposium on Principles of Data Mining and Knowledge Discovery,London,1997:88-100.

[8] Gavrilov M,Anguelov D,Indyk P,et al.Mining the stock market:which measure is best? [C]//ACM International Conference on Knowledge Discovery and Data Mining,Boston,2000:487-496.

[9] Goldin D Q,Kanellakis P C.On similarity queries for time-series data:constraint specification and implementation [C]//International Conference on Principles and Practice of Constraint Programming,Cassis,France,1995:137-153.

[10] Valery Guralnik,Jaideep Srivastava.Event detection from time series data[C]//ACM International Conference on Knowledge Discovery and Data Mining,New York,1999:33-42.

[11] Han Jia-wei,Kamber M.Data mining concepts and techniques[M].CA:Morgan Kaufmann Publisher,2000.

[12] Chakrabarti K,Keogh E,Mehrotra S,et al.Locally adaptive dimensionality reduction for indexing large time series databases [J].ACM Trans Database Syst,2002,27(2):188-228.

[13] Lee S L,Chun S J,Kim D H,et al.Similarity search for multidimensional data sequences[C]//IEEE International Conference on Data Engineering,2000:599-608.

[14] Megalooikonomou V,Wang Q,Li G,et al.A multiresolution symbolic representation of time series[C]//Proceedings of the 21st International Conference on Data Engineering.Washington,DC:IEEE Computer Society,2005:668-679.

[15] Vlachos M,Yu P S,Castelli V.On periodicity detection and structural periodic similarity[C]//SIAM International Conference on Data Mining,Newport Beach,CA,2005.