ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.1, January 2009, pp.67−79
doi: 10.3724/SP.J.1001.2009.03450

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于点分布特征的多元时间序列模式匹配方法*

管河山[1], 姜青山[2+], 王声瑞[3]

[1](厦门大学 计算机科学系,福建 厦门 361005)

[2](厦门大学 软件学院,福建 厦门 361005)

[3](Department of Computer Science, University of Sherbook, Quebec, Canada)

Pattern Matching Method Based on Point Distribution for Multivariate Time Series

GUAN He-Shan[1], JIANG Qing-Shan[2+], WANG Sheng-Rui[3]

[1](Department of Computer Sciences, Xiamen University, Xiamen 361005, China)

[2](School of Software, Xiamen University, Xiamen 361005, China)

[3](Department of Computer Science, University of Sherbook, Quebec, Canada)

+ Corresponding author: E-mail: qjiang@xmu.edu.cn

Guan HS, Jiang QS, Wang SR. Pattern matching method based on point distribution for multivariate time series. *Journal of Software*, 2009,20(1):67−79. http://www.jos.org.cn/1000-9825/3450.htm

Abstract: Common methods for matching multivariate time series such as the Euclid method and PCA method have difficulties in taking advantage of the global shape of time series. The Euclid method is not robust, while the PCA method is not suitable to deal with the small-scale multivariate time series. This paper proposes a pattern matching method based on point distribution for multivariate time series, which is able to characterize the shape of series. Local important points of a multivariate time series and their distribution are used to construct the pattern vector. To match pattern of multivariate time series, the Euclid norm is used to measure the similarity between the pattern vectors. The global shape characteristic is used in the method to match patterns of series. The results of experiments show that it is easy to characterize the shape of multivariate time series with this method, with which various scales can be dealt with in series data.

Key words: multivariate time series; local important point; point distribution; shape characteristic; similarity measure; pattern matching

摘 要: 多元时间序列模式匹配的常用方法难以刻画序列的全局形状特征,比如,Euclid 方法的鲁棒性不够强;而 PCA 方法不适合处理小规模多元时间序列.基于点的统计分布提出了一种能够有效刻画多元时间序列形状特征的模式匹配方法.首先,提取多元时间序列样本的局部重要点,作为模式描述的方式;然后,根据重要点的统计分布特点构建特征模式向量,并借助 Euclid 范数来度量两个特征模式向量之间的相似程度,进而进行多元时间序列模式匹配.采用该方法进行模式匹配,充分利用了序列的全局形状特征.实验结果表明,基于点分布特征的多元时间序列模式匹配能够有效地刻画序列的形状特征,且能处理多种规模的序列数据.

* Supported by the National Natural Science Foundation of China under Grant No.10771176 (国家自然科学基金); the National 985 Project of China under Grant No.0000-X07204 (国家"九八五"工程二期基金)

Received 2007-11-21; Accepted 2008-08-07

关键词: 多元时间序列;局部重要点;点分布;形状特征;相似性度量;模式匹配
中图法分类号: TP311 文献标识码: A

多元时间序列包括了医学、音频、视频和过程监控等方面的数据.随着相关领域发展的需求和计算机技术的进步,多元时间序列数据的收集变得越来越庞大,而多元时间序列挖掘的研究工作也得到了极大的挑战和发展,比如多元时间序列的分类、预测和模式挖掘等都得到广泛的研究[1−5].从其应用研究角度来看,多元时间序列挖掘的相关技术也得到广泛的应用.比如,医生可以根据心电图来判定病人的状况[1−2];生产过程控制中,监控人员可以根据历史数据来形成经验,并在监控过程中及时发现错误并予以纠正[3−5];此外,音频检索和视频检索也可借助多元时间序列模挖掘的相关技术.目前,时间序列挖掘的相关研究中,大多数研究都利用了相似性度量的技术,其中,有很多一元时间序列的相似性度量研究,也产生了一套比较成熟的理论[6−13],而多元时间序列的相似性度量研究相对较少[14−21].可以说相似性度量的研究是时间序列挖掘的核心技术之一,也是时间序列挖掘的重大挑战之一,时间序列的模式匹配与相似性度量也是紧密关联的.

时间序列模式匹配主要是指从时间序列中寻找变化规律并预测未来的发展趋势,从而有效地对客观事物规律进行预报和控制.我们主要针对小规模多元时间序列来展开模式匹配的研究,比如Robot Execution Failures数据[22](每个序列样本为 15×6 阶的矩阵).Robot的监控数据可以分为正常状态和非正常状态两种类型,通过对Robot收集相应的监控数据,并借助已有的决策系统可以对其实现实时监控、错误诊断和修复等相关工作.本文提出基于点分布特征的模式匹配方法(point distribution,简称PD方法),该方法可以对Robot监控这样的小规模多元时间序列数据进行有效的模式匹配,也可以处理其他领域的小规模多元时间序列数据,而且对大规模的多元时间序列进行模式匹配也有较好的效果.

给定一个多元时间序列 $X_t=(X_{1t},X_{2t},...,X_{lt})'$,其中l为一个正整数,时间t=1,2,…,n,该序列为一个l元的时间序列,其中序列规模可以根据$l\times n$的取值来判定,若取值较小,则定义小规模的多元时间序列.多元时间序列模式匹配的流程如图 1 所示.通常,模式匹配需要解决两大关键问题:模式的定义方式(模式表示)和相似性度量的方式.多元时间序列由于其不同维度之间的关系复杂,使得模式匹配的研究工作开展得相对缓慢,特别是在相似性度量的研究方面,多元时间序列远远落伍于一元时间序列.比如,在多元时间序列模式匹配的研究当中,PCA(principal component analysis)是一种常用的方法[14,15],然而主成分的求解通常需要大量的数据才能得到合适的解,在处理小规模的多元时间序列数据时,PCA方法难以得到合理的结果;Euclid模式匹配方法是指采用Euclid函数[21]直接计算模式之间的距离,然后进行模式匹配,这也是一种可行的方法,然而该方法的稳健性不够,特别是它只能处理规模相同的多元时间序列样本.PCA方法是采用主成分作为模式表示的准则,而Euclid方法则直接采用原数据作为模式表示的准则.

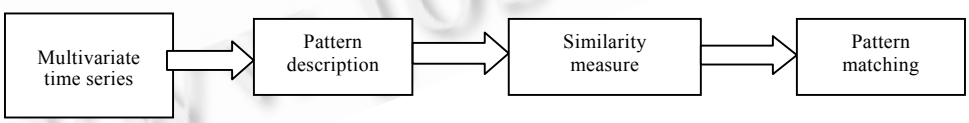


Fig.1 Process of pattern matching for multivariate time series

图 1 多元时间序列相似模式匹配流程

针对多元时间序列模式匹配的研究,我们提出一种新的多元时间序列模式表示方法,该方法能够刻画不同规模的多元时间序列的形状特征.具体做法通过两步来完成,首先抽取多元时间序列的局部重要点作为多元时间序列的模式表示;然后,根据局部重要点的统计分布特征构建一个特征模式向量,并借用 Euclid 距离函数来刻画特征模式向量之间的差异,建立新的多元时间序列相似性度量方法.基于点分布特征的模式表示可以很好地刻画多元时间序列的形状特征,并且能够处理多种规模的序列数据,即 $l\times n$ 可以等于任意的正整数,特别是对小规模的多元时间序列数据处理更能发挥其独特的优势.我们将在实验部分详细地分析该方法对不同规模的多

元时间序列进行模式匹配时的性能.

本文第1节介绍时间序列模式匹配的相关研究,特别是多元时间序列的模式表示和相似性度量的研究.第2节详细介绍提取多元时间序列局部重要点的方法,并构建统计特征向量作为多元时间序列的相似性度量.第3节列举5个数据集进行实验,并与Euclid方法和PCA方法加以对比.第4节进行总结,并提出展望.

1 相关研究

时间序列挖掘究其具体的研究工作而言,包括了时间序列聚类、分类、检索、分割、预测、可视化和模式匹配等多方面的内容;究其研究对象而言,可以分为一元时间序列挖掘和多元时间序列挖掘.多元时间序列的相关研究(包括多元时间序列模式匹配)在很大程度上受制于多元时间序列相似性度量,由于多元时间序列样本不同维度之间的相关性,使其相似性度量的研究成果远不及一元时间序列.多元时间序列的相关研究中常有的方法包括参数法和非参数法.参数化方法主要是指对多元时间序列进行建模,然后利用模型参数来衡量不同序列之间的相似程度,比如,Raquel[20]通过建立多元时间序列的VAR模型(vector autoregressive),提取模型的系数作为相似性度量的依据.常用的非参数化方法有Euclid方法、PCA方法[14−16]、修正的PCA方法(modified PCA[17])、基于概率分布的距离方法[15]、多重自相关函数的距离方法[18]和形状特征向量方法[19]等,后几种方法有其独特的应用背景,其中,Euclid方法和PCA方法是本文的分析重点.

多元时间序列的参数化分析方法主要是指针对多元时间序列样本建立相应的模型,并提取模型的系数作为多元时间序列的模式表示方式,该方法的性能在很大程度上取决于所建立模型的合理性.我们主要是针对多元时间序列的非参数化分析方法展开研究和对比分析,多元时间序列非参数化分析方法研究颇多,比如,Wang[21]对一个模拟的流动接触反应数据(fluid catalytic cracking)进行多元时间序列聚类研究,他将数据张开成一个长的行向量(a long row vector),使用张开后的向量作为特征,然后采用Euclid距函数计算特征之间的相似性程度[21].该方法要求被分析的数据具有相同的观察个数,且难以有效地刻画多元时间序列的整体形状特征.Krzanowski [16]提出一种基于PCA的多元时间序列相似性度量,如式(1):

$$S_{PCA}=\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\cos^2\theta_{ij} \tag{1}$$

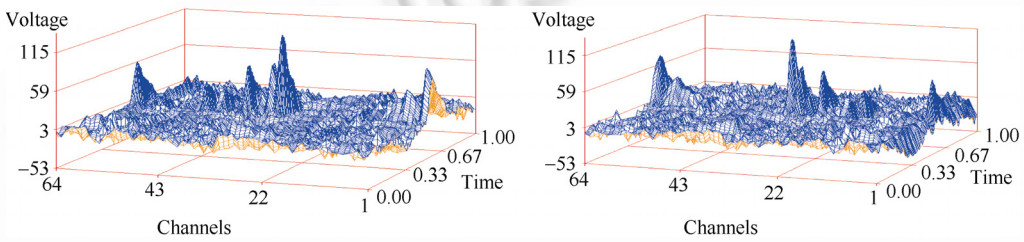
其中,θ_{ij}表示第i个主成分和第j个主成分之间的夹角,m表示主成分的个数,m的取值是根据方差(或者特征根)来决定的,通常要求所选取的前m个主成分所对应的特征根的和占所有特征根总和的95%以上[15].在PCA的相关研究中,张军等人[14]在研究多变量时间序列的模式挖掘时将多变量的数据集分段平均为连续矩阵,并采用基于主成分分析和奇异值分解的方法来对矩阵进行相似性比较,最后通过相邻片断的合并来组成更高层次的时序片断,以提高模式的匹配范围.

众多的研究表明,Euclid方法的鲁棒性不好,它对时间序列在垂直方向波动和水平方向波动的鲁棒性都不好[10,11,19],因此对时间序列的形状描述能力很有限.目前,时间序列挖掘的相关研究中通常避免直接采取Euclid方法进行分析.然而,PCA方法通常要求足够的样本点才能有效地求解得到其主成分向量,而且PCA方法在计算夹角余弦$\cos^2\theta_{ij}$时,并不考虑主成分向量的正负方向(比如θ_{ij}=30°和θ_{ij}=50°两种情况,S_{PCA}的结果是相同的),而且PCA方法计算特征根、主成分、夹角余弦和除以m时,通常需要精确到较高的小数位(本文精确到10位小数).本文提出的PD方法,从“样本点构建多元时间序列的形状特征”的角度出发,通过提取多元时间序列的局部重要点集来初步描述多元时间序列的样本点特征,然后通过“9维向量”来进一步提取样本点的分布特征,并建立相应的相似性度量.该方法通过捕捉多元时间序列的样本点主要分布特征来描述多元时间序列的形状特征,这符合人们对时间序列形状特征的直观认识.该方法对小规模的多元时间序列处理具有良好的性能,优于PCA方法和Euclid方法,为小规模多元时间序列模式匹配提供了一条新的途径,同时,该方法对大规模的多元时间序列处理也能得到较好的结果,但稍逊于PCA方法的性能.

2 基于点分布的多元时间序列模式匹配(PD)

多元时间序列可以通过三维空间来描述其图形,其形状特征是由所有的样本点来形成的,因此,利用样本点的分布特点来刻画多元时间序列的形状特征是一条可行的途径,我们正是从多元时间序列的样本点的统计特征来展开模式表示和相似性度量的相关研究的,提出了一种新的多元时间序列模式表示(基于局部重要点的模式表示),并提出新的相似性度量方式(基于局部重要点分布的特征模式向量).

任意给定的一个多元时间序列 $X_t=(X_{1t},X_{2t},...,X_{lt})'$,其中,$l$>1,$t$=1,2,…,$n$.对其不同维度的 l 个一元时间序列,可以按照 l 取值的某种方式排序,本文称其为 l 排序,之所以要考虑多元时间序列不同维度之间的排序关系,是因为不同的排序会造成不同的序列图像,造就不同的形状特征.对所有给定的多元时间序列,应该按照同一种固定的 l 排序方式来处理所有的序列样本,此时,所有序列样本的形状特征之间才具有可比性.大致来说,某种 l 排序形式下的两个相似多元时间序列,在另一种 l 排序形式下也呈现出相当的相似程度,并在三维空间中描绘出一个多元时间序列图像,以此来探讨多元时间序列的形状特征.比如给定医学中一个 EEG 数据(electroencephalogram,即一个多元时间序列样本,该数据来自文献[23],编号为 co2c0000337 的第 5 个样本),可以按照 l 递增顺序和 l 递减顺序分别绘制其 3D 图像,如图 2 所示.



(a) Sort l increasingly
(a) l 递增的顺序

(b) Sort l decreasingly
(b) l 递减的顺序

Fig.2 Illustration of multivariate time series

图 2 多元时间序列图像

2.1 基于局部重要点的多元时间序列模式表示

多元时间序列模式表示的方法有很多,最直接的办法就是用原数据来表示多元时间序列,该方法能够精确地刻画多元时间序列所有的特征,保留完整的信息.然而,有时候更需要关注于多元时间序列的形状特征概貌,而不过多地注重其细节的形状特点,为此,采用某种合适的方式来进行多元时间序列模式描述,也显得格外重要.我们根据多元时间序列样本点的分布特征来进行模式抽取,即提取多元时间序列的局部极大值点和极小值点(称为局部重要点),利用点集作为多元时间序列模式表示的方法.

对多元时间序列局部重要点的定义,需借助函数局部极小值点和极大值点的概念.给定一个多元时间序列 $X_t=(X_{1t},X_{2t},...,X_{lt})'$,$t$=1,2,…,$n$,则样本点的取值是时间$t$和维度数$l$的一个函数,记为式(2);根据函数极值点的定义,函数在点$F(t,l)$的邻域G(对应于数据区间$X[i_1{:}i_2,j_1{:}j_2]$)内有定义,且对该领域内任意点$(t+h,l+h)$,满足式(3):

$$X=F(t,l) \tag{2}$$

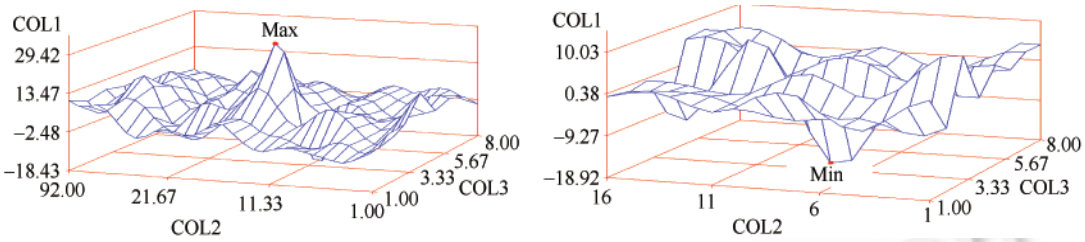
$$F(t+h,l+h)\geq F(t,l) \text{ 或者 } F(t+h,l+h)\leq F(t,l) \tag{3}$$

称为极小值点(或极大值点),其中 h 为任意小的数.给定一个多元时间序列,下面具体地介绍一下局部重要点的提取方法.首先给出多元时间序列局部重要点的定义如下:

定义 1. 给定一种分割方式,点$x[i,j]$的邻域G(即$X[i_1{:}i_2,j_1{:}j_2]$),i为i_1和i_2的均值,表示矩阵的行数;j为j_1和i_2的均值,表示矩阵的列数.如果点$x[i,j]$为邻域G中的最大值点,则称为局部极大值重要点,如图 3(a)所示;如果点$x[i,j]$为邻域G的最小值点,则称为局部极小值的重要点,如图 3(b)所示.

要提取多元时间序列样本中所有的局部重要点,必须提取所有的局部极大值点和极小值点,此时,必须首先

定义时间序列样本的分割方式.假定已有的一个分割方式为$X[i_1:i_2,j_1:j_2]$,可采用一种算法来求得所有的局部重要点,具体描述如下:



(a) Maximum point
(a) 极大值点

(b) Minimum point
(b) 极小值点

Fig.3 Local important point

图 3 局部重要点示意图

算法 1. 局部重要点提取算法 Loc_Imp_Point.

输入:多元时间序列 X.

输出:多元时间序列 X 的局部重要点.

Step 1. 给定一个多元时间序列X,为$l\times n$阶;并给定一种分割方式$X[i_1:i_2,j_1:j_2]$,且其中心点为$x[i,j]$;

Step 2. 如果点$x[i,j]$为小块$X[i_1:i_2,j_1:j_2]$中的最大值,则记为X的局部极大值点;如果点$x[i,j]$为小块$X[i_1:i_2,j_1:j_2]$中的最小值,则记为X的局部极小值点;

Step 3. 遍历 X 中所有点,找出所有的局部极值点 $x[i,j]$,作为 X 的局部重要点.

提取所有的局部重要点后,构成局部重要点的点集,以该点集作为多元时间序列模式表示(如图 4 所示).此外,提取多元时间序列的局部重要点需要事先对样本进行分割.分割块的大小由局部重要点的保留率λ来决定,保留率是指多元时间序列的局部重要点个数与该序列的总样本点个数的比率.对多元时间序列的纵横分割比例λ_1和λ_2,分别采用式(4)和式(5)来定义和计算:

$$\lambda_1=(i_2-i_1)/l \tag{4}$$

$$\lambda_2=(j_2-j_1)/k \tag{5}$$

纵横分割比例越小,表示每个分割块越细,观察的粒度越细,此时,保留的重要点个数越多,反之保留的重要点个数越少.比如,采用EEG数据中的一个多元时间序列样本[23](编号为co2c0000337 的第 5 个样本),该序列样本大小为 256×64,在不同分割方式下,计算纵横分割比例、保留率λ 和重要点个数的值,见表 1.为了描述方便,对分割方式采用了另一种描述方式,见表 1 中的第 1 列.显然,分割块越细,保留的重要点个数就越多.

Table 1 Segmentation approach of EEG data

表 1 EEG 数据分割方式

Segmentation	λ_1 and λ_2	Ratio λ	Number of important points
$X[i-32:i+32,j-8:j+8]$	$\lambda_1=1/4$, $\lambda_2=1/4$	0.002 197 3	36
$X[i-16:i+16,j-8:j+8]$	$\lambda_1=1/8$, $\lambda_2=1/4$	0.003 295 9	54
$X[i-16:i+16,j-4:j+4]$	$\lambda_1=1/8$, $\lambda_2=1/8$	0.006 713 9	110
$X[i-8:i+8,j-4:j+4]$	$\lambda_1=1/16$, $\lambda_2=1/8$	0.011 352 5	186
$X[i-8:i+8,j-2:j+2]$	$\lambda_1=1/16$, $\lambda_2=1/16$	0.021 911 6	359
$X[i-4:i+4,j-2:j+2]$	$\lambda_1=1/32$, $\lambda_2=1/16$	0.040 954 6	671
$X[i-4:i+4,j-1:j+1]$	$\lambda_1=1/32$, $\lambda_2=1/32$	0.068 542 5	1 123

实验时,对所有多元时间序列样本应采用同一种分割方式,才能确保局部重要点模式之间的可比性.本文研究的重点是针对小规模的多元时间序列数据,此时,多元时间序列本身的数据点个数较少,即$l\times n$为一个较小的

数,为了保留足够多的局部重要点以进行分析,可直接设定为最细分割形式$X[i_1{:}i_2,j_1{:}j_2]$,即$i_2-i_1=2$, $j_2-j_1=2$.此外,对X的行边缘点,比如行的某个边缘点$x[1,j]$,小块$X[i_1{:}i_2,j_1{:}j_2]$中的i_1可能出现负值,此时分割块定义为$X[1{:}i_2,j_1{:}j_2]$;同理,对列边缘点也采用同样的处理方式.

如果处理的是大规模的多元时间序列数据集,设共有q个序列样本构成的数据集,则此时按照纵横等分割比的原则,可以采取多种分割方式对所有序列样本进行分割,假设得到的重要点个数分别为$I=\{I_1,I_2,\ldots,I_q\}$,我们定义重要点个数分布的极值如式(6)所示:

$$r=\max_{i=1,2,\ldots,q} I_i-\min_{i=1,2,\ldots,q} I_i \tag{6}$$

本文优先选取重要点个数为二位数以上,集合中的最大值和最小值相差一个数量级的情况,然后取 r 值最小的分割方式作为该数据集中样本的分割方式.

2.2 基于点分布特征的相似性度量

抽取局部重要点作为多元时间序列模式表示,还需要提供一个合适的相似性度量(或者说提供一个合适的距离函数)来刻画模式之间的相似程度.常用的相似性度量有 Euclid 距离、夹角余弦和相关系数等方法,但这些方法都不适合局部重要点模式之间的相似性刻画,因为不同的多元时间序列样本在同一种分割方式下所提取的重要点的个数可能不同.为了能够发挥局部重要点的多元时间序列模式表示的优势,我们提出了一种新的相似性度量,该度量结合了点分布特征和 Euclid 距离函数的优势.

通常提取的局部重要点只是占原数据点数的小百分比,即保留率λ取值往往较小,设该点集为$P=\{P_1,P_2,\ldots,P_h\}$,h为所提取的局部重要点的个数.图 4 是一个EEG数据重要点提取的示意图(该数据来自文献[23],编号为co2c0000337 的第 5 个样本).所提取的重要点只是col1=0(根据上文,将多元时间序列看成是时间(t)和维度数(l)的一个函数F,则为$F(t,l)=0$ 平面)平面上方、下方和平面上的一些点集.此时,需提取这些局部重要点的统计分布特点,以构建特征模式向量作为相似性度量.

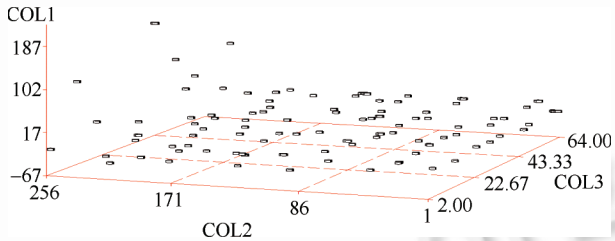


Fig.4 Distribution of local important points

图 4 局部重要点分布示意图

采用分位点来描述点集的分布特征是统计学中一种常用的思路[24,25],本文采用 9 个常用的分位点来描述多元时间序列局部重要点集P的分布特征.考虑到多元时间序列的局部重要点集P的样本点规模和实际分析的需求,具体采用的分位点包括盒子图中的 5 个分位点,即极大值(max)、75%分位点、50%分位点、25%分位点和极小值(min)5 个分位点,并结合常用的 95%分位点和 5%分位点、90%分位点和 10%分位点等 4 个对称的百分位点[24,25],共采用 9 个特征来分析点集P.理论上来说,采取的分位点的个数越多,则对点集P的特征描述得越精确,然而考虑到小规模多元时间序列的局部重要点个数通常较少(比如第 3 节中的robot实验数据,其局部重要点个数通常是十几个、或者几十个的数量级),因此,提取的分位点个数不宜过多,提取这 9 个特征也是与小规模多元时间序列的实际应用相符合的.本文采用这 9 个特征来构建特征模式向量,并建立相似性度量.

$$d(X,Y)=\sum_{i=1}^{9}(F_i^{(X)}-F_i^{(Y)})^2 \tag{7}$$

3 实验与结果分析

基于点分布特征的模式匹配能够有效地刻画出多元时间序列的整体形状特征,且对各种规模的序列数据都能得到较好的结果.这里,我们列举 4 个小规模数据和 1 个大规模数据来进行实验,并对实验结果进行详细的分析.所列举的 5 个数据都已知分类结果,采用 k-近邻的方法进行实验,具体描述如下:

假定待分析的实验数据集中含有n_2个多元时间序列样本,任意从该数据集中抽取一个样本,记为输入样本X.提取该数据集中所有样本的局部重要点,并建立相应的特征模式向量,然后从该数据集中找出与输入样本X最相似的"k个样本",比如k取 10 个、5 个或 1 个最相似性的样本.统计这"k个样本"与输入样本X类别相同的样本个数n_1,按照式(8)计算准确率:

$$e = n_1 / k \tag{8}$$

对其他任意一个样本,都一一作为输入样本,然后重复以上实验,并计算相应的准确率,这样就可以得到n_2个模式匹配的准确率.对PCA方法[15,16]和Euclid[21]方法重复以上实验,分别得到准确率以进行比较分析.

采用 k-近邻的方法进行实验,根据公式(8)计算模式匹配的准确率,然后根据公式(9)进一步计算这些准确率的期望值.按照本文的实验方式,所有的准确率取值 e 可能为 $\{0,0.1,0.2,...,1.0\}$,共 11 个可能值(见实验部分),记为 $\{e_1,e_2,e_3,...,e_{11}\}$.将准确率作为一个离散随机变量$\varepsilon$,此时,该随机变量的期望值可按照式(9)来确定.

$$P = \sum_{i=1}^{11} p(\varepsilon = e_i)e_i \tag{9}$$

3.1 Robot Execution Failure(REF)数据

数据Robot Execution Failure共有 5 个子数据集[22],采用其中的 3 个子数据集分别进行实验.我们借助多元时间序列模式匹配的方法来对Robot进行监控,此时,匹配的准确率越高,说明该方法对过状态的识别能力越高(判别过程是否正常).我们首先采用第 1 个子数据集LP1 进行实验,该数据已知分为 4 类,即normal类、collision类、fr_collision类和obstruction类,共 88 个样本.每个序列样本为 15×6=90 阶的矩阵,为小规模的多元时间序列.分割形式为 $X[i-1:i+1, j-1:j+1]$.

分别采用 PD,PCA 和 Euclid 这 3 种方法进行模式匹配,并计算相应的准确率,见表 2,所有的准确率保留 2 位有效数字.总体看来,在成功率为小概率事件的情况下(比如取值为 0 和 0.1 等),PD 方法和 Euclid 方法所对应的次数都比 PCA 方法要少,而在成功率为大概率事件的情况下(比如取值为 0.8 和 0.9 等),PD 方法和 Euclid 方法所对应的次数都比 PCA 方法要多.特别地,当准确率为 1(即 100%)时,在 3 种模式匹配的情况下(即取 1 个相似样本、5 个相似样本或 10 个相似样本),PD 方法和 Euclid 方法所对应的次数都远多于 PCA 方法.从准确率分布情况来看,PCA 方法在处理该数据集时,得不到理想的结果,这表明,在多元时间序列规模较小的情况下,PCA 方法不再是一种合适的模式匹配方法.

Table 2 Experimental results of LP1 dataset (N represents number, R represents ratio)

表 2 LP1 数据集的实验结果(N 表示个数,R 表示比率)

Parameter k	PD						PCA						Euclid					
	1		5		10		1		5		10		1		5		10	
e	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R
0	12	0.14	4	0.05	3	0.03	29	0.33	9	0.10	5	0.06	12	0.14	4	0.05	3	0.03
0.1	0	0	0	0	2	0.02	0	0	0	0	6	0.07	0	0	0	0	4	0.05
0.2	0	0	5	0.06	6	0.07	0	0	13	0.15	3	0.03	0	0	5	0.06	8	0.09
0.3	0	0	0	0	3	0.03	0	0	0	0	8	0.09	0	0	0	0	5	0.06
0.4	0	0	5	0.06	4	0.05	0	0	10	0.11	17	0.19	0	0	8	0.09	7	0.08
0.5	0	0	0	0	2	0.02	0	0	0	0	22	0.25	0	0	0	0	8	0.09
0.6	0	0	6	0.07	6	0.07	0	0	24	0.27	9	0.10	0	0	8	0.09	4	0.05
0.7	0	0	0	0	6	0.07	0	0	0	0	6	0.07	0	0	0	0	7	0.08
0.8	0	0	12	0.14	9	0.10	0	0	20	0.23	7	0.08	0	0	14	0.16	6	0.07
0.9	0	0	0	0	2	0.02	0	0	0	0	4	0.05	0	0	0	0	12	0.14
1	76	0.86	56	0.64	45	0.51	59	0.67	12	0.14	1	0.01	76	0.86	49	0.56	24	0.27

进一步地,分别计算在 3 种相似性模式匹配的情况下,3 种方法的准确率期望值,见表 3.显然,PD 方法和

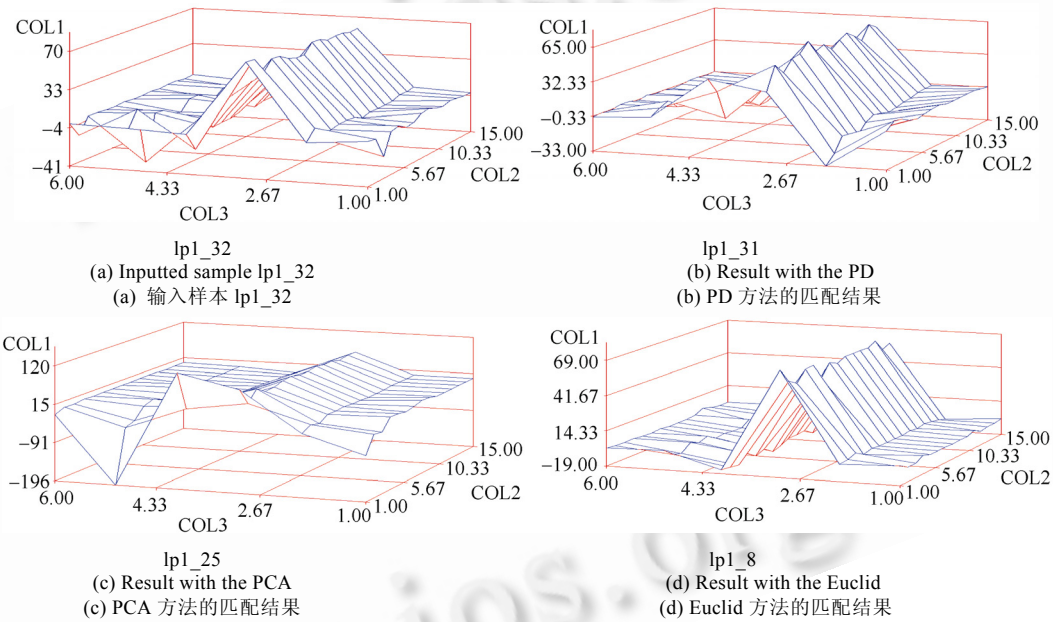
Euclid 方法的准确率期望值高于 PCA 方法.在 3 种模式匹配的情况下,PD 方法的准确率期望值约为 80%左右,而 PCA 方法的准确率期望值约为 50%左右,如此低的准确率通常不能满足实际应用的需求,这也进一步表明,PCA 方法在处理小规模多元时间序列时,其性能远不及 PD 方法.在匹配出的相似样本个数较多(比如取 10 个)时,Euclid 方法的准确率期望值约为 65%左右,这远远不如 PD 方法.

Table 3 Accuracy expectations of LP1 datatset

表 3 LP1 数据集的准确率期望值

Parameter k	PD	PCA	Euclid
10	0.76	0.47	0.65
5	0.82	0.56	0.79
1	0.86	0.67	0.86

为了形象地说明不同方法的实验效果,我们采用第 32 个样本,记为 lp1_32(其他多元时间序列样本也采用同种方式命名),称为输入样本,采用上述 3 种方法进行相似模式匹配.列举出最相似的 1 个模式所对应的样本,如图 5 所示.PD 方法所得到的样本,与输入样本 lp1_32 在形状上都具备很大的相似性,而 Euclid 和 PCA 两种方法所得到的结果在形状上与输入样本不具备良好的相似性,因此,其匹配效果不如 PD 方法.



lp1_32
(a) Inputted sample lp1_32
(a) 输入样本 lp1_32

lp1_31
(b) Result with the PD
(b) PD 方法的匹配结果

lp1_25
(c) Result with the PCA
(c) PCA 方法的匹配结果

lp1_8
(d) Result with the Euclid
(d) Euclid 方法的匹配结果

Fig.5 Results of matching similar pattern for the LP1 dataset

图 5 LP1 数据的相似模式匹配结果

此外,再考虑采用其他 4 个子数据集进行实验,其中第 2 个子数据集 lp2 有 5 种类别,而其样本总个数仅 47 个,此时,某些类(比如,collision to the right 类)的样本个数还不到 10 个,按照我们设计的实验方式,该数据不适合进行实验,所以在此我们不采取第 2 个子数据集 lp2;同理,也不采取第 3 个子数据集 lp3.我们采用第 4 个子数据集 lp4 和第 5 个子数据集 lp5 进行实验,它们的样本总数分别为 117 个和 164 个,类别数分别为 3 类和 5 类.简单起见,我们只是给出了 3 种方法实验结果的准确率期望值,见表 4 和表 5,而不详细列出具体的准确率分布表.从表 4 和表 5 可以看出,在 3 种模式匹配的情况下,PD 方法和 Euclid 方法的准确率期望值比 PCA 方法要高,这进一步表明 PCA 方法在处理小规模多元时间序列时,其性能远不及 PD 方法.此外,Euclid 方法的准确率期望值也较高,这也表明,在处理小规模多元时间序列时,如果所有样本的规模都一样,则该方法能够得到较好的匹配结果,然而该方法不能处理多种规模的样本,见第 3.2 节.

Table 4 Accuracy expectations of LP4 dataset

表 4 LP4 数据集的准确率期望值

Parameter k	PD	PCA	Euclid
10	0.85	0.63	0.74
5	0.88	0.68	0.79
1	0.90	0.73	0.88

Table 5 Accuracy expectations of LP5 dataset

表 5 LP5 数据集的准确率期望值

Parameter k	PD	PCA	Euclid
10	0.54	0.41	0.55
5	0.59	0.45	0.61
1	0.62	0.51	0.68

3.2 apanese Vowel(JV)数据

PD方法也可以处理样本规模不同的小规模多元时间序列,我们列举数据Japanese Vowel[26]来进行实验,该数据常用于多元时间序列的分类研究.采用其中的训练子数据进行实验,该数据已知分为 9 种类别,共 270 个样本.每个样本序列含有 12 个一元序列,时间长度位于 7~29 的区间内,为小规模的多元时间序列.该数据采用的分割形式为 $X[i-1:i+1, j-1:j+1]$.

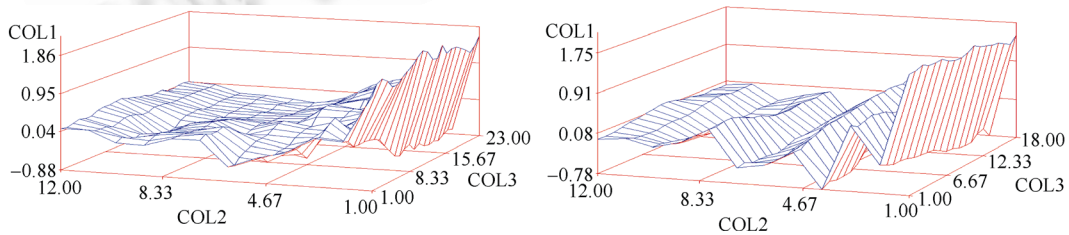
该时间序列数据的时间长度不一致,因此 Euclid 和 PCA 方法不能处理该数据.我们在此直接采用 PD 进行模式匹配,并计算相应的准确率,见表 6 和表 7,所有的准确率保留 2 位有效数字.该数据集共计 9 个类别,因此,对每个样本而言,随机寻找最相似的样本,其成功的概率为 1/9.而采用 PD 方法在处理该数据集时,从准确率期望值来看,远远大于 1/9.这进一步表明 PD 方法在处理小规模多元时间序列时所具有的优势.该实验也表明,PD 方法能够同时处理规模不同的多元时间序列数据集.

为了形象地说明不同方法的实验效果,我们采用了第 6 个样本,记为 jv_6(其他多元时间序列样本也采用同种方式命名),称为输入样本,采用上述 3 种方法进行相似模式匹配.列举出最相似的 1 个模式所对应的样本,如图 6 所示.PD 方法所得到的样本(jv_8)与输入样本 jv_6 在形状上都具备较好的相似性.

Table 6 Experimental results of JV dataset (N represents number, R represents ratio)

表 6 JV 数据集的实验结果(N 表示个数,R 表示比率)

Parameter k	PD						PCA	Euclid
	1		5		10			
e	N	R	N	R	N	R		
0	125	0.46	28	0.10	9	0.03		
0.1	0	0	0	0	23	0.09		
0.2	0	0	53	0.20	34	0.13		
0.3	0	0	0	0	39	0.14		
0.4	0	0	57	0.21	41	0.15		
0.5	0	0	0	0	30	0.11	–	–
0.6	0	0	66	0.24	27	0.10		
0.7	0	0	0	0	25	0.08		
0.8	0	0	38	0.14	25	0.08		
0.9	0	0	0	0	11	0.04		
1	145	0.54	28	0.10	6	0.02		



(a) Inputted sample jv_6
(a) 输入样本 jv_6

(b) Result with the PD, jv_8
(b) PD 方法的匹配结果,jv_8

Fig.6 Results of matching similar pattern for JV dataset

图 6 JV 数据的相似模式匹配结果

Table 7 Accuracy expectations of JV dataset

表 7 JV 数据集的准确率期望值

Parameter k	PD	PCA	Euclid
10	0.45	–	–
5	0.49		
1	0.54		

3.3 EEG数据

PD方法在对大规模的多元时间序列进行处理时,较Euclid方法而言也能达到较好的效果,这里,我们列举了EEG数据来进行实验分析[23].该脑电图用 256Hz的电极同时在 64 个部位测量得到的一组数据,数据收集来源于两种人群:alcoholic subjects和control subjects,一共有 122 个测试者的数据,每个测试者共 120 次测试.本文只是采用了前 2 位测试者的 2 个数据,编号为:co2a0000364 和co2c0000337,2 种类别,共 166 个样本.每个序列样本为256×64 阶的矩阵,是大规模的多元时间序列.该数据采用的分割形式为 $X[i-16:i+16, j-4:j+4]$.

分别采用 PD,PCA 和 Euclid 这 3 种方法进行模式匹配,并计算相应的准确率,见表 8,所有的准确率保留 2 位有效数字.总体看来,在成功率为小概率事件的情况下(比如:取值为 0 和 0.1 等),PD 方法和 PCA 方法所对应的次数都比 Euclid 方法要少,而当成功率为大概率事件的情况下(比如取值为 0.8 和 0.9 等),PD 方法和 PCA 方法所对应的次数都比 Euclid 方法要多.特别地,当准确率为 1(即 100%)时,在 3 种模式匹配的情况下(即取 1 个相似样本、5 个相似样本或 10 个相似样本),PD 方法和 PCA 方法所对应的次数都多于 Euclid 方法.从准确率分布情况来看,PCA 方法在处理该数据集时能够得到理想的结果,这表明,在多元时间序列规模较大的情况下(256×64 阶矩阵),PCA 方法是一种合适的模式匹配方法.另外,PD 方法在处理该数据集时优于 Euclid 方法.

Table 8 Experimental results of EEG dataset (N represents number, R represents ratio)

表 8 EEG 数据集的实验结果(N 表示个数,R 表示比率)

Parameter k	PD						PCA						Euclid					
	1		5		10		1		5		10		1		5		10	
e	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R	N	R
0	29	0.17	1	0.01	1	0.01	25	0.15	1	0.01	0	0	31	0.19	5	0.03	1	0.01
0.1	0	0	0	0	0	0	0	0	0	0	1	0.01	0	0	0	0	5	0.03
0.2	0	0	9	0.05	8	0.05	0	0	2	0.01	1	0.01	0	0	10	0.06	8	0.05
0.3	0	0	0	0	2	0.01	0	0	0	0	4	0.02	0	0	0	0	7	0.04
0.4	0	0	18	0.11	8	0.05	0	0	15	0.09	6	0.04	0	0	18	0.11	7	0.04
0.5	0	0	0	0	8	0.05	0	0	0	0	10	0.06	0	0	0	0	15	0.09
0.6	0	0	21	0.13	17	0.10	0	0	15	0.09	8	0.05	0	0	26	0.16	16	0.10
0.7	0	0	0	0	8	0.05	0	0	0	0	20	0.12	0	0	0	0	12	0.07
0.8	0	0	24	0.14	15	0.09	0	0	33	0.20	17	0.10	0	0	30	0.18	18	0.11
0.9	0	0	0	0	15	0.09	0	0	0	0	27	0.16	0	0	0	0	17	0.10
1	137	0.83	93	0.56	84	0.51	141	0.85	100	0.60	72	0.43	135	0.81	77	0.46	60	0.36

采用 k-近邻方法进行多元时间序列的模式匹配,根据公式(8)分别计算 3 种方法的准确率分布,然后根据公式(9)分别计算 3 种方法的准确率期望值,得到的结果见表 9.其中 PCA 方法的准确率期望值最高,PD 方法的准确率期望值稍低于 PCA 方法,而 Euclid 方法的准确率期望值都低于前两种方法.PD 方法的准确率期望值都在80%以上,这表明 PD 方法在处理大规模的多元时间序列数据时有较好的效果,但却不及 PCA 方法.为了形象地说明不同方法的实验效果,我们采用 co2c0000337 中的第 47 个样本,记为 Co2c0000337_47(其他多元时间序列样本也采用同种方式命名),称为输入样本,采用上述 3 种方法进行相似模式匹配.列举出最相似的 1 个模式所对应的

Table 9 Accuracy expectations of EEG dataset

表 9 EEG 数据集的准确率期望值

Parameter k	PD	PCA	Euclid
10	0.82	0.84	0.73
5	0.82	0.86	0.76
1	0.84	0.86	0.81

样本,如图 7 所示.PD 和 PCA 两种方法得到的样本,与输入样本 Co2c0000337_47 在形状上都具有很大的相似性,而 Euclid 方法所得到的结果在形状上与输入样本不具备良好的相似性,因此,其匹配效果不如前两种方法.从多元时间序列的形状特点来看,基于点分布的模式表示能够很好地刻画多元时间序列的形状特征.

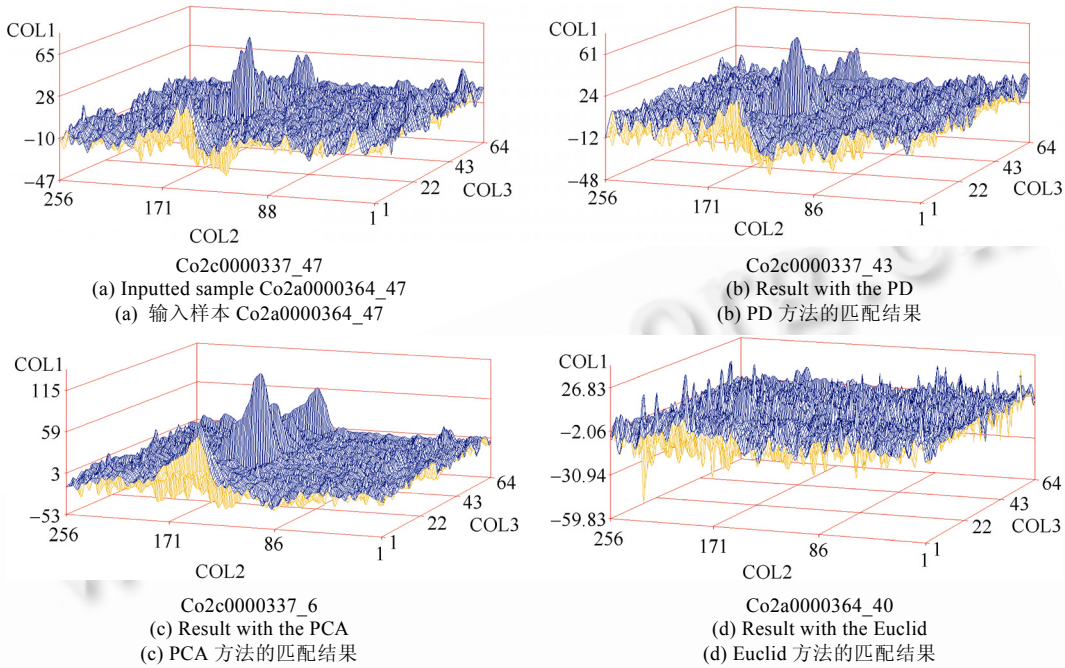


Fig.7 Results of matching similar pattern for EEG dataset

图 7 EEG 数据的相似模式匹配结果

总之,在进行小规模多元时间序列模式匹配时,PD 方法能够有效地刻画出多元序列的形状特征,且能处理多种规模的序列数据.这是 PCA 方法和 Euclid 方法所不及的.从 PD 方法的计算过程来看,该方法充分利用了多元时间序列的样本点的统计分布特征来进行形状刻画,这样更能从全局来刻画其形状特征.这 3 种模式匹配方法的详细对比见表 10,我们从模式表示、相似性度量及所处理的数据规模这 3 个角度对上述 3 种方法进行了详细的对比.

Table 10 Comparison of pattern matching methods

表 10 模式匹配方法的对比

Methods	PD	PCA	Euclid
Patter description	Important points	Principal component	Original data
Similarity measure	Fractile	Cosine	Euclid
Data scale	Be able to deal with multi-scale data, especially small-scale data	Be able to deal with large-scale data	Data scale must be same

4 结论与展望

模式匹配是时间序列挖掘研究中的重要一支.多元时间序列的模式匹配与模式定义是直接关联的.本文从形状特征来定义多元时间序列的模式,并提出基于点分布特征的模式表示方法和相似性度量方式来进行模式匹配.该方法能够有效地刻画出多元时间序列的形状式特征,在一定程度上不受多元时间序列规模大小的制约.我们所提出的 PD 方法对小规模的多元时间序列进行模式匹配具有较好的性能,对某些大规模的多元时间序列的处理性能稍逊于 PCA 方法,因此,借鉴 PD 方法和 PCA 方法的各自优势,建立一种新的方法以便处理各种规模的多元时间序列数据,这将是我们下一步研究工作的重点.

致谢 感谢给予本文有价值建议的匿名审稿人,感谢厦门大学国际数据挖掘中心的同学们对本文提供的帮助.

References:

[1] Zhou Y, Zhao Y, Xie LL, Zhou LM, Chen ZY. Computation and analysis of parameters in phase space reconstruction of epileptic EEG signal. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2007,46(3):5−9 (in Chinese with English abstract).

[2] Wang XY, Luo C, Qiu TS. Nonlinear dynamic research on EEG in HAI experiment. Chinese Journal of Biomedical Engineering, 2007,24(4):408−415 (in Chinese with English abstract).

[3] Singhal A, Seborg DE. Matching patterns from historical data using PCA and distance similarity factors. In: Krogh BH, ed. Proc. of the 2001 American Control Conf. Arlington, 2001,2:1759−1764.

[4] Liu B, Liu J. Multivariate time series prediction via temporal classification. In: Rakesh A, ed. Proc. of the 18th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society, 2002. 268.

[5] Camarinha-Matos LM, Seabra Lopes L, Barata J. Integration and learning in supervision of flexible assembly systems. IEEE Trans. on Robotics and Automation, 1996,12(2):202−219.

[6] Liu HT, Ni ZW, Li JY. An effective algorithm to match similar time series pattern. Journal of Computer-Aided Design & Computer Graphics, 2007,19(16):725−729 (in Chinese with English abstract).

[7] Huang H, Huang K, Hang XS, Xiong FL. Algorithm for fast time-series pattern recovery in a long sequence. Computer Engineering and Applications, 2003,39(21):192−194 (in Chinese with English abstract).

[8] Ge XP, Padhraic S. Deformable Markov model templates for time-series pattern matching. In: Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2000. 81−90.

[9] Pratt KB, Fink E. Search for patterns in compressed time series. Int'l Journal of Image and Graphics, 2002,2(1):89−106.

[10] Wang XH. Study on time series similarity and trend prediction [Ph.D. Thesis]. Tianjin: Tianjin University, 2003 (in Chinese with English abstract).

[11] Dong XL, Gu CK, Wang ZG. Research on shape-based time series similarity measure. Journal of Electronics & Information Technology, 2007,29(5):1228−1231 (in Chinese with English abstract).

[12] Wu SC, Wu GF, Wang W, Yu ZC. A time-sequence similarity matching algorithm for seismological relevant zones. Journal of Software, 2006,17(2):185−192 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/185.htm

[13] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series. In: Nick C, ed. Proc. of the IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society Press, 2001. 273−280.

[14] Zhang J, Wu SC, Wang W. Research of data mining method on multivariate time series. Computer Engineering and Design, 2006, 27(18):3364−2266 (in Chinese with English abstract).

[15] Singhal A, Seborg DE. Pattern matching in multivariate time series databases using a moving window approach. Ind. Eng. Chem. Res., 2002,41(16):3822−3838.

[16] Krzanowski WJ. Between-Groups comparison of principal components. Journal of the American Statistical Association, 1979, 74(367):703−707.

[17] Singhal A, Seborg DE. Pattern matching in historical batch data using PCA. IEEE Control Systems Magazine, 2002,22(5):53−63.

[18] Guan HS, Jiang QS, Wang SJ. A new similarity measure for clustering multivariate time series. Journal of Computational Information Systems, 2007,3(5):2031−2036.

[19] Huang H, Shi ZZ, Zheng Z. Similarity search based on shape *k-d* tree for multidimensional time sequences. 2006, 17(10):2048−2056 (in Chinese with English abstract).

[20] Raquel P, Francisco M, Gabriel H. Multivariate time series modeling and classification via hierarchical VAR mixtures. Computational Statistics & Data Analysis, 2006,51(3):1445−1462.

[21] Wang XZ, McGreavy C. Automatic classification for mining process operational data. Industrial & Engineering Chemistry Research, 1998,37(6):2215−2222.

[22] http://kdd.ics.uci.edu/databases/robotfailure/robotfailure.html. 1999.

[23] Archive P. 1999. http://kdd.ics.uci.edu/databases/eeg/eeg.html

[24] Ruan GH, *et al*. SAS Statistic Analysis and it Application. Beijing: Tsinghua University Press, 2003. 48−57 (in Chinese).

[25] Bernstein S, Wrote; Shi DJ, Trans. Theory of Statistic. Beijing: Science Press, 2002. 139−139 (in Chinese).

[26] http://kdd.ics.uci.edu/databases/JapaneseVowels/JapaneseVowels.html. 2000.

附中文参考文献:

[1] 周毅,赵怡,解玲丽,周列民,陈子怡.癫痫 EEG 信号相空间重构参数的计算和分析.中山大学学报(自然科学版),2007,46(3):5−9.

[2] 王兴元,骆超,邱天爽.HAI 实验中 EEG 信号的非线性动力学研究.中国生物医学工程学报,2005,23(4):408−415.

[6] 刘慧婷,倪志伟,建洋.时间序列相似模式的有效匹配.计算机辅助设计与图形学学报,2007,19(16):725−729.

[7] 黄河,黄轲,杭小树,熊范纶.时间序列中快速模式发现算法的研究.计算机工程与应用,2003,39(21):192−194.

[10] 王晓华.时间序列数据挖掘中相似性和趋势预测的研究[博士学位论文].天津:天津大学,2003.

[11] 董晓莉,顾成奎,王正欧.基于形态的时间序列相似性度量研究.电子与信息学报,2007,29(5):1228−1231.

[12] 吴绍春,吴耿锋,王炜,蔚赵春.寻找地震相关地区的时间序列相似性匹配算法.软件学报,2006,17(2):185−192. http://www.jos.org.cn/1000-9825/17/185.htm

[14] 张军,吴绍春,王炜.多变量时间序列模式挖掘的研究.计算机工程与设计,2006,27(18):3364−2266.

[19] 黄河,史忠植,郑征.基于形状特征 *k-d* 树的多维时间序列相似搜索.软件学报,2006,17(10):2048−2056. http://www.jos.org.cn/1000-9825/17/2048.htm

[24] 阮桂海,等.SAS 统计分析实用大全.北京:清华大学出版社,2003.48−57.

[25] Bernstein S,著;史道济,译.统计学原理.北京:科学出版社,2002.139−139.

管河山(1981－),男,湖南衡阳人,博士生,主要研究领域为数据挖掘,时间序列挖掘,统计学,数据分析,数学建模.

王声瑞(1963－),男,博士,教授,博士生导师,主要研究领域为模式识别,人工智能,数据挖掘,图像处理和理解.

姜青山(1962－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,数据库系统,聚类分析,模糊集理论与应用.