

基于特征的时间序列聚类方法研究进展

宋 辞, 裴 韬

(中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101)

摘 要:时间序列聚类可以根据相似性将对象集分为不同的组,从而反映出同组对象的相似性特征和不同组对象之间的差异特征。当序列维度较高时,传统的时间序列聚类方法容易受噪声影响,难以定义合适的相似性度量,聚类结果往往意义不明确。当数据有缺失或不等长时,聚类方法也难以实施。基于上述问题,一些学者提出了基于特征的时间序列聚类方法,不仅可以解决上述问题,还可以发现序列本质特征的相似性。本文根据时间序列的不同特征,综述了基于特征的时间序列聚类方法的研究进展,并进行了分析和评述;最后对未来研究进行了展望。

关 键 词:时间序列;时间序列特征;数据挖掘;聚类

1 引言

随着传感器数量的不断增长以及遥感(RS)、地理信息系统(GIS)、全球定位系统(GPS)的广泛使用,地学研究邻域产生了大量的观测数据。这些数据不再局限于传统的静态空间中,而是逐渐向时间维扩展,形成了时间序列数据^[1]。时间序列中蕴藏着不同的模式,而不同的模式反映了不同的序列成因。因此,针对序列模式进行聚类,将其分为不同的类别成为我们认识序列数据,进而理解序列形成本质的重要手段。由此看来,针对时间序列数据的聚类方法研究具有非常重要的意义。

与传统的点数据聚类方法相比,针对时间序列的聚类具有一定复杂性。首先,时间序列数据具有高维性,在这巨大的维数中往往只有一小部分维度是与表现对象变化特征的簇结构密切相关的,而其他不相关或者相关性很小的维度会产生大量的噪声,从而掩盖了真实的簇结构^[2]。其次,由于维度较高,数据稀疏,维度之间也很可能会有相关性^[3],传统的相似性度量方法难以发现真实的结果^[4]。第三,时间序列相似性的定义多种多样,基于观测值的相似性度量只能发现表面的变化,没有体现事物的内在机制。两条序列即使观测值相差很小,也不

代表序列就很相似(图1a);同样,观测值完全不同,两条序列也有可能在某方面具有相似之处(图1b)。

目前,一些学者提出了许多方法来解决不同类型的时间序列聚类问题。这些方法大致可分为两种:①对现有的静态数据聚类方法进行改进使其能处理时间序列数据;②将时间序列数据转换为静态数据的形式,然后直接用静态数据聚类方法来进行聚类^[5]。按照这个思路,时间序列聚类方法可分为基于原始测度数据的时间序列聚类和基于特征的时间序列聚类。基于原始测度数据的时间序列聚类,直接根据原始数据定义相似度,如欧氏距离,相关系数,DTW距离等,然后进行聚类。Liao总结了用于时间序列聚类的各种相似性度量方法^[5];Díaz根据相似性度量的定义中是否需要估计模型参数,将时间序列聚类方法分为有参数的聚类方法和无参数的聚类方法^[6]。这些方法在现实生活中都有广泛的应用。然而,采用基于原始测度数据的时间序列聚类方法,不可避免地要面对高维数据的问题;此外,基于原始数据仅能发现序列表面的相似性,没有触及序列本身的内在机制,聚类结果有很大的局限性。基于特征的时间序列聚类方法,先对原始数据进行降维,抽取表征其内在变化机制的特征作为相似性度量的基础,然后运用各种聚类方法对这

收稿日期:2011-10; 修订日期:2012-03.

基金项目:中国科学院知识创新工程重要方向项目(KZCX2-YW-QN303);中国科学院地理资源所自主部署创新项目(200905004);863项目(2009AA12Z227)。

作者简介:宋辞(1986-),男,博士研究生,主要研究方向为空间数据挖掘。E-mail: songc@lreis.ac.cn

通讯作者:裴韬(1972-),男,副研究员,主要从事空间数据挖掘和空间信息统计等方面的研究。E-mail: peit@lreis.ac.cn

些特征进行聚类,不仅减少了计算量,解决了时间序列高维数据问题,而且还可以处理有数据缺失、不等长或采样不均匀的时间序列;最重要的是,基于特征的时间序列可以根据不同的应用问题选取合适的特征,从而发现时间序列内在机制中不同方面的相似性。

本文根据时间序列的不同特征,系统综述了基于特征的时间序列聚类方法的研究进展。首先介绍了时间序列的定义,概念以及各类特征;然后对基于特征的时间序列聚类方法进行了分析和评述;最后讨论了现有方法的问题和挑战,并对未来时间序列聚类方法研究进行了展望。

2 时间序列数据及特征

时间序列也称为动态序列,由一组随时间变化的观测值组成。与传统静态数据不同,时间序列是一类复杂的数据对象,描述了事物变化过程。

2.1 时间序列类型

时间序列有很多种。根据数据类型不同,可以分为数值型时间序列和类别型时间序列;根据采样时间不同可以分为均匀采样时间序列和非均匀采样时间序列;根据观测值维度不同可以分为单维时间序列和多维时间序列;根据统计特征不同可以分为平稳型时间序列和非平稳型时间序列。不同的时间序列具有的特征也不同,本文主要针对数值型时间序列,如果没有特殊说明,下文中出现的“时间

序列”均指数值型时间序列。

2.2 时间序列特征

通常时间序列具有多个特征,每个特征刻画了时间序列的一个方面。从对时间序列不同层次上的认知可将时间序列特征分为3种:形态特征、结构特征以及模型特征。这种分类体现了人们对时间序列认识逐步深化的过程。

2.2.1 形态特征

时间序列的形态特征主要指时间序列的形状变化特征,包括全局特征和局部特征。全局特征描述了时间序列的起伏变化,如上升、下降、头肩模式(图2)等;局部特征则表现为时间序列局部时间点上的异常观测值,如不连续点、极值点、突变点、转折点等。在时间序列最开始的研究中,人们通常是先将时间序列画出来,然后直观地通过观察来研究时间序列的起伏变化或异常点。这类反映时间序列整体变化或局部异常,可以直观看出特征,称为时间序列的形态特征。基于形态特征的时间序列聚类,可以发现具有相同形状的时间序列簇,寻求时间序列的起伏变化规律。

时间序列形态特征可以在一定程度上表现时间序列的特性,通常适用于描述短时间序列^[4]。当序列较长时,起伏变化往往比较复杂,难以用简单的“上升,下降”描述。虽然可以采用分段描述的方法^[7-8],但这割裂了时间序列的整体性,不能很好地反映时间序列的全局特征;异常点特征主要描述时间序列上的某些特殊点的特征,同样难以反映其全

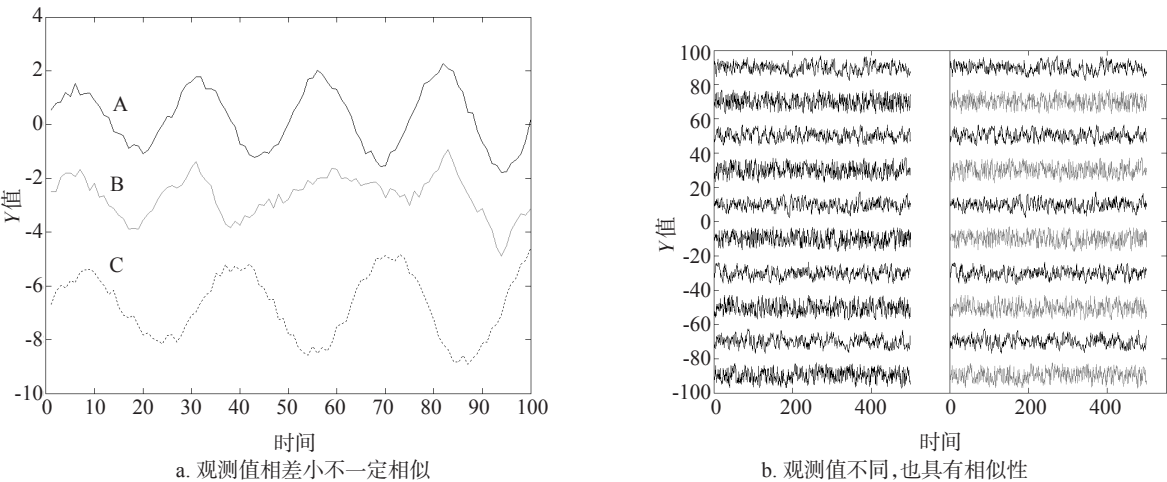


图1 观测值与相似性的关系

Fig.1 Relationship between observations and similarity of time series

注:a. A, B, C 三条序列,计算基于标准化后观测值的欧氏距离, $d(A,B) < d(A,C)$;但是直观上看, A 和 C 两条序列显然更相似;b. 序列的观测值之间难以看出关系,两两之间距离大体相等,但这些序列来自相同的创建机制:黑色序列来自系数为 0.55,噪声方差为 4 的 AR(1)模型,灰色序列来自系数为 0.35,噪声方差为 6 的 AR(1)模型。

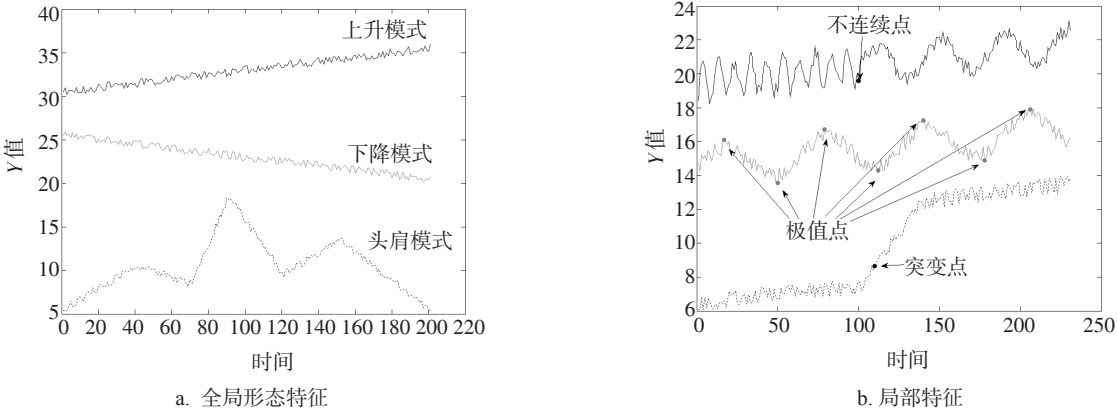


图2 时间序列形态特征
Fig.2 Shaped characteristics of time series

局特征。

2.2.2 结构特征

时间序列的结构特征是对时间序列全局构造或内在变化机制的描述,它可以很好的表现时间序列全局特点。时间序列的结构特征一般难以直观的看出,需要对原始数据进行统计或者转换得出。

时间序列结构特征通常包括以下3类:基本统计特征、时域特征和频域特征。

(1) 基本统计特征

基本统计特征是描述时间序列全局结构的一些统计量,它不是时间序列特有的特征,而是可用于描述任何一组数据的特征,包括均值、方差、偏度、峰度等(表1)。

均值和方差是用来描述数据的中心及其偏差的。偏度是表征概率分布密度曲线相对于均值不对称程度的特征指数,直观看就是函数曲线尾部的相对长度;峰度,则是表征概率密度分布曲线在平均值处峰值高低的特征指数,直观上反映了函数曲线尾部的厚度。

(2) 时间序列时域特征

时间序列时域特征是时间序列在时间域上表现出的全局结构特征,它反映了时间序列随时间变化的规律。时间序列时域特征包括:趋势、季节波动、时间序列的自相关、混沌等(表2)。

趋势是描述时间序列长期变化情况;季节性反映了时间序列周期内的波动情况;自相关性是时间序列特有的性质,表现为时间序列的观测值依赖于之前观测值的情况;混沌则表示时间序列受其初值影响的敏感程度。

(3) 时间序列频域特征

时间序列频域特征是时间序列在频率域上表现出的结构特征,它描述了时间序列的组成成分。

表1 时间序列基本统计量
Tab.1 Statistic indices of time series

统计量	计算方法
均值	$\mu = \sum_{i=1}^n y_i / n$
方差	$D = \sum_{i=1}^n (y_i - \bar{y})^2$
偏度	$S = (n\sigma^3)^{-1} \sum_{i=1}^n (y_i - \bar{y})^3$
峰度	$K = (n\sigma^4)^{-1} \sum_{i=1}^n (y_i - \bar{y})^4$

注: y_t 表示 t 时刻的观测值; $t = 1, 2, \dots, n$; \bar{y} 表示 y_t 的平均值; σ 表示 y_t 的标准差。

表2 时间序列时域特征^[4,9]
Tab.2 Characteristics of time series in time domain

时域特征	表现指标	计算方法
趋势	T 趋势项	对滑动平均值进行回归
季节波动	S 季节项	不同周期内同相位的观测值取平均数
	k 滞后自协方差函数	$r_k = \sum_{i=1}^{n-k+1} (y_i - \bar{y})(y_{i+k} - \bar{y})$
自相关	Box-Pierce 指数	$Q_k = n \sum_{i=1}^k \hat{\rho}_i^2$
	Hurst 指数	重极差(R/S)方法, 回归求 $H^{[10]}$ $\log((R/S)_n) = \log(K) + H \log(n)$
混沌	李雅普诺夫指数	$LE = e^{-\sum_{i=1}^N \lambda_i} / (1 + e^{-\sum_{i=1}^N \lambda_i})^{[11]}$

注: y_t 表示 t 时刻的观测值, $t = 1, 2, \dots, n$; k 表示滞后系数; \bar{y} 表示 y_t 的平均值; $\hat{\rho}_k$ 表示序列 y_t 的 k 滞后自相关系数; R 表示极差; S 表示标准差; K 为常数; H 表示 Hurst 指数; n 表示时间窗长度; Y_t 表示 t 时刻的观测值; Y_i^* 则是与 Y_t 最接近的点; N 表示观测点总数; λ 为常数。

一条时间序列可以看成由多个不同频率的振荡序列叠加而成^[1]。时间序列频域特征主要包括周期解析强度和谱密度。

周期强度是在频率 $\omega_j = j/n$ 下, 正弦振荡数据相关的平方度量, 为时间序列离散傅里叶转换的模的平方, 具体如下式:

$$P(j/n) = (2n^{-1} \sum_{i=1}^n x_i \cos(2\pi j i/n))^2 + (2n^{-1} \sum_{i=1}^n x_i \sin(2\pi j i/n))^2 \quad (1)$$

式中: j 表示 n 个数据点的 j 次循环; $P(j/n)$ 表示频率为 j/n 下的周期强度。

谱密度, 或者称(功率谱密度)是平稳随机过程中频率的一个正值函数, 可以看作是时间序列自相关函数的傅里叶变换。当且仅当时间序列是宽平稳的时候, 才存在功率谱密度。谱密度通常采用傅里叶变换技术来计算。

$$f(\omega) = \sum_{h=-\infty}^{+\infty} \gamma(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2 \quad (2)$$

式中: ω 表示频率; h 表示滞后系数; $\gamma(h)$ 表示时间序列的自协方差函数, 要求满足 $\sum_{h=-\infty}^{+\infty} |\gamma(h)| < \infty$ 。

2.2.3 模型特征

时间序列模型特征描述了事物变化潜在的运动规律。人们通过对大量时间序列的研究, 基于某种假设推理, 总结出的表达事物变化规律的抽象数学公式就是时间序列模型。模型特征一般表现为不同的参数特征, 不同的时间序列是具有不同参数的模型表达。描述时间序列的模型多种多样, 通常是将时间序列看成是一个随机过程, 用不同的随机过程去模拟时间序列。这些模型包括: 高斯过程模型、ARMA(自回归滑动平均模型)以及 ARIMA 模型(差分自回归移动平均模型)、马尔科夫链模型、隐马尔科夫模型等。

(1) 高斯过程模型: 假设各个时间点上的观测值相互独立, 且都服从高斯分布, 其模型表达为: $X_t \sim N_{iid}(\mu, \sigma)$, 主要参数特征包括均值 μ 和方差 σ 。

(2) ARMA 模型: 假设序列的当前观测值 x_t 与之前的 p 个值有线性关系, 因此只要知道原始序列的观测值, 就可对未来进行预测, 其模型表达式为:

$$x_t = a + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \quad (3)$$

该模型特征表现为上式中参数 a, ϕ, θ 。 x_t 表示 t 时刻观测值, w_t 表示 t 时刻的高斯噪声。

(3) 马尔科夫链模型: 常用来描述类别型时间序列, 也可以通过离散化处理数值型时间序列。它

将时间序列看成某系统“状态”的演变过程, $X_t = x$ 表示系统在 t 时刻处于状态 x 。假设时间序列的当前状态只与前一个状态有关, 记为: $P\{X_t = x_t | X_1 = x_1 \cdots X_{t-1} = x_{t-1}\} = P\{X_t = x_t | X_{t-1} = x_{t-1}\}$, 则时间序列 x_t 是一个马尔科夫链。马尔科夫链模型特征表现为概率转移矩阵。

(4) 隐马尔科夫模型: 由初始状态概率向量 π , 状态转移概率矩阵 A 和观察值概率矩阵 B 组成。表示无法直接观察到马尔科夫链的状态序列, 但是可以观察到其输出序列, 是一个双重随机过程, 其模型特征表现为 $O(\pi, A, B)$ 。

上述这些模型都体现了不同的时间序列特征, 在时间序列聚类方法中广泛使用。

3 基于特征的时间序列聚类方法

聚类分析根据对象之间的相似性, 将其分成不同的组, 其中组内对象之间距离最小, 而组间对象之间距离最大。传统的静态数据聚类方法分为 5 类: 基于划分的聚类、基于层次的聚类^[12-13]、基于密度的聚类^[14]、基于格网的聚类^[15]以及基于模型的聚类^[2, 16]。

基于特征的时间序列聚类, 在传统静态聚类方法的基础上引入了时间序列特征的相似性。通过不同的特征来研究时间序列的内在变化机制, 从而发现其相似规律。依据聚类问题所针对的不同特征, 可以将时间序列聚类分为 3 类: 基于形态特征的时间序列聚类、基于结构特征的时间序列聚类、基于模型特征的时间序列聚类。

3.1 基于形态特征的时间序列聚类

基于形态特征的时间序列聚类可以揭示时间序列中相似的起伏变化或其异常点。前者表明序列整体趋势变化相似, 后者则是序列局部相似的表现。基于这点考虑, 可将基于形态特征的时间序列聚类分为全局形态特征聚类和局部形态特征聚类。

3.1.1 基于全局形态特征的时间序列聚类

基于全局形态特征的时间序列聚类方法适用于处理短时间序列, 如基因谱聚类^[4], 发现序列的整体相似性。采用原始时间序列的欧氏距离或 Pearson 相关系数距离可以从一定程度上反映全局形态特征^[3], 但无法发现具有尺度拉伸、位移, 强度拉伸、位移的相似形态特征(图 3a)。此外欧氏距离和 Pearson 相关系数对噪声相当敏感(图 3b), 难以处理

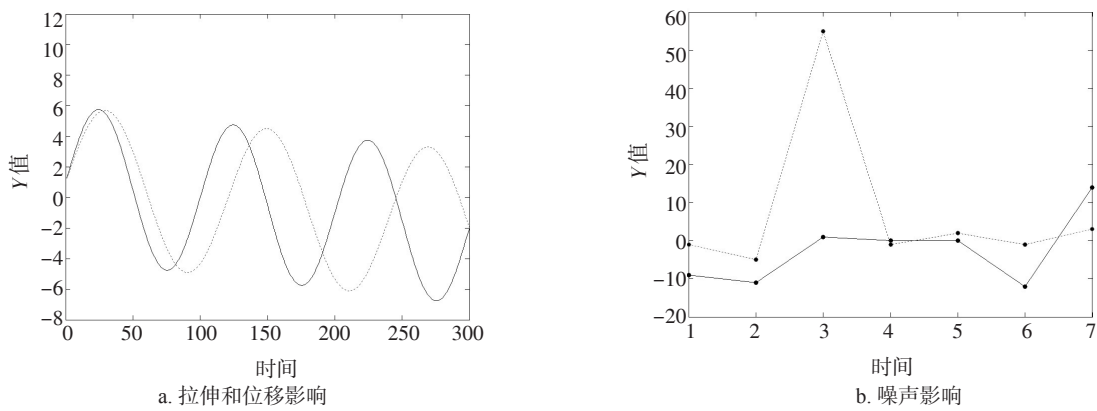


图3 受拉伸、位移或噪声影响的相似形态特征

Fig.3 Similar shaped characteristics affected by noise, shifts and scales

注：a. 具有尺度拉伸、位移的相似形态特征；b. 相似序列受噪声影响，Pearson 相关系数仅 0.3，Spearman 相关系数为 0.93。

不等长、非均匀采样或有数据缺失的时间序列。

DTW 距离^[17]放宽了全局形态特征相似性在尺度上的限制，可以处理不等长的时间序列。它在一定程度上克服了尺度位移的问题，但依然无法发现具有强度拉伸或位移的相似形态特征，此外该方法的计算量往往比较大，不适合长时间序列聚类问题。针对噪声问题，Balasubramaniyan 提出 Spearman 相关系数作为基因谱序列的相似性度量^[4]，采用观测值大小的排名来描述时间序列全局形态特征，而忽略序列观测值取值本身。Spearman 相关系数的计算如下式：

$$SRC(X,Y)=1-\frac{6}{n(n^2-1)}\sum_{i=1}^n(r_x(x_i)-r_y(y_i))^2 \quad (4)$$

式中： $r_x(x_i)$ 表示 i 时刻的观测值 x_i 在时间序列 (x_1,x_2,\dots,x_n) 中的排名； $r_y(y_i)$ 类似； n 表示观测总数。Spearman 相关系数的显著度通过经验分布计算。实验证明该方法在对具有相似全局形态特征的短时间序列聚类上可以在一定程度克服噪声和形状位移等问题，优于传统的采用欧氏距离和 Pearson 相关系数函数作为相似性度量的方法(图 3b)。但由于这种方法忽略观测值本身，聚类结果往往比较粗糙。

Möller-Levet 等定义了短时间序列距离来描述短时间序列全局形态特征的相似性^[18-19]。每条时间序列的形态特征用一组分段斜率代替，这样可以减弱拉伸或位移所带来的影响。该方法也可以处理非均匀采样的时间序列数据，但要求数据是等长的。下式是该方法的距离度量：

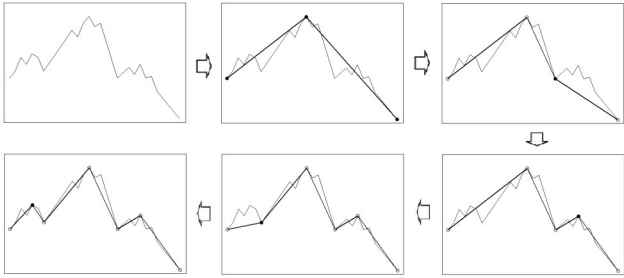


图4 重要点的提取过程^[20]

Fig.4 Process of extract PIPs

$$d_{STS}^2(x,v)=\sum_{k=0}^{n_t-1}\left(\frac{v_{(k+1)}-v_k}{t_{(k+1)}-t_k}-\frac{x_{(k+1)}-x_k}{t_{(k+1)}-t_k}\right)^2 \quad (5)$$

式中： x_k 和 v_k 表示不同序列第 k 次观测时刻的观测值； t_k 表示第 k 次观测的时间； n_t 表示观测时间点总数。

长时间序列由于维数很高，其全局形态特征的描述容易受维度之间的相关性及噪声的影响。对此，Fu 等对长时间序列进行了简化，采用序列的 PIP 点(Perceptual Important Point)来表征其全局形态特征，并进行聚类^[20]。这种方法很好的克服了噪声问题，可以发现表征大尺度变化的相似形态特征。序列的简化过程采用道格拉斯压缩算法，大大提高了聚类算法的效率。

3.1.2 基于局部形态特征的时间序列聚类

局部形态特征可以体现时间序列局部的异常值。针对序列的局部形态特征，Keogh 等提出了分段线性分割的方法，将原始序列分为多个子序列，通过各个子序列的相似性来度量时间序列的整体相似性^[7]。每段子序列采用 5 个参数来表示：

$A \equiv \{AXL, AYL, AXR, AYR, AW\}$, 分别表示线段的左点 x 坐标, 左点 y 坐标, 右点 x 坐标, 右点 y 坐标以及该段线段权重, 采用分段加权距离计算序列之间的相似度。Chen 等采用与 Keogh 等类似的方法, 也对时间序列进行了分段处理。它认为一条时间序列由一组局部模式组成^[8], 每个模式可以用 5 个参数表示: $lp = (\theta_{pos}, \theta_{amp}, \theta_{shp}, \theta_{sc1}, \theta_{sc2})$, 分别表示局部模式在原始时间序列中的起始位置, 平均振幅, 形状参数, 时间尺度和振幅尺度。随后他定义了局部模式的综合相似度——SpADe 距离。实验证明采用 SpADe 距离聚类可以很好的解决拉伸和位移问题, 其结果精度比欧氏距离, DTW 距离以及 EDR 距离都要高。

小波变换具有多尺度效应, 基于这点考虑, Hsu 对原始序列进行小波处理, 采用多尺度的小波系数表征原始序列的特征, 既要突出全局整体特征, 又表现局部序列特征。聚类结果表明采用小波系数聚类可以很好的发现降水时间序列局部奇异值和**锐转变点**的相似特征以及整体周期变化的特征^[21]。

表 3 给出了基于形态特征时间序列聚类方法中各种相似性度量的特点。该方法适用于短时间序列聚类问题, 多用于基因序列聚类问题以及一些轨迹聚类问题^[22]。当处理长时间序列聚类问题时, 往往需要进行特殊处理, 对序列本身形式有要求, 有一定的局限性。

3.2 基于结构特征的时间序列聚类

基于形态特征的时间序列聚类停留在序列表面形状的相似上, 没有考虑其内部结构的相似性。这类方法适用于短时间序列聚类, 对于长时间序列往往有一定的局限性。基于结构特征的时间序列聚类能够揭示时间序列潜在的相似变化机制和结

构, 从而发现更有意义的聚类结果。根据聚类结构特征的不同可以分为基于统计特征的时间序列聚类、基于时域特征的时间序列聚类、基于频域特征的时间序列聚类(表 4)。

3.2.1 基于统计特征的时间序列聚类

基于统计结构特征的时间序列聚类采用描述一般序列的基本统计量作为时间序列特征来进行聚类。

Nanopoulos 等最早提出了一种基于统计结构特征的时间序列聚类方法^[23], 它选取了时间序列的均值、标准差、偏度、峰度 4 个基本统计量表征时间序列的结构特征, 偏度和峰度包含观测值分布的形状信息。分别计算了原始序列及其一阶差分序列的均值、标准差、偏度和峰度值, 采用神经网络的方法对这些特征进行了聚类。实验表明, 基于这些统计特征的时间序列聚类在一定程度上克服了噪声问题, 并大大提高了计算效率。Ouyang 选取了时间序列的最大值、最小值、均值以及标准差作为时间序列的结构特征, 对塔里木流域的单一水文站点不同月份的流量序列进行了聚类, 从而发现了该地区的不同水文流量时期^[24]。

3.2.2 基于时域特征的时间序列聚类

基于时域特征的时间序列聚类采用时间序列在时域上特有的一些全局结构特征, 如: 趋势、周期、自相关等, 进行聚类。

Kontaki 等^[25]和 Kumar 等^[26]考虑用时间序列的趋势结构特征和季节性结构特征进行了聚类。前者采用分段线性概化的方法, 定义了 DPLA 距离表示为分段趋势距离之和, 作为相似性度量; 后者在考虑季节性相似度量时, 不仅计算了季节特征波动部分, 而且还考虑其误差, 采用两个季节模式具有

表 3 基于形态特征时间序列聚类的相似性度量的特点

Tab.3 Properties of similarity in shaped characteristics based time series clustering							
克服问题	全局	局部	噪声	尺度位移	强度位移	不等长	非均匀采样
距离度量	特征	特征	影响	或拉伸	或拉伸	时间序列	时间序列
欧氏距离	O	×	×	×	×	×	×
Pearson 相关系数	O	×	×	O	×	×	×
DTW	O	×	×	O	×	O	×
Spearman 相关系数	O	×	O	O	O	×	×
PIP 距离	O	×	O	×	×	O	O
短时间序列距离(STS)	O	×	O	×	×	×	O
分段加权距离	×	O	O	O	O	O	×
小波变换	O	O	O	O	O	×	×

注: O 表示可以克服该种问题; × 表示难以克服该种问题。

相同均值的零假设的显著性作为季节性相似的程度。该方法用来对零售商品数据的季节性模式进行聚类,发现了零售商品中具有相似均值分布的季节模式。Wang等在上述两人的基础上又加入了一部分时间序列特有的特征,包括周期、自相关性、非线性以及混沌性等共9个特征,采用层次聚类方法和SOM方法,对其进行时间序列聚类^[9]。实验结果表明,用9个特征代表原始数据进行时间序列聚类,不仅可以提高计算效率,而且可以得到更高精度的聚类结果。此外通过特征选取步骤,可以发现不同意义的聚类结果,Wang等将此方法用在对人类行为的聚类研究上^[27]。

3.2.3 基于频域特征的时间序列聚类

随后,更多时间序列特有的结构特征被引入时间序列聚类,以发现其不同方面的内在变化机制。基于频域特征的时间序列聚类可以发现具有相似周期或谱密度等频域特征的时间序列。

Caiado等提出用周期解析强度作为时间序列的结构特征^[28],定义了标准化周期解析强度的对数距离作为时间序列的相似性度量。实验表明基于该特征聚类可以区分具有不同ARMA或ARIMA模型的时间序列。Shumway等则对多维时间序列的谱密度特征进行聚类^[29-30],其相似性度量采用了谱矩阵的两种拟距离:Kullback-Liebler信息散度与Chernoff对称信息散度。基于该相似性度量,文中

采用层次聚类的方法将地震时间序列和爆炸的时间序列数据进行了分组。

3.2.4 基于其他结构特征的时间序列聚类

时间序列的结构特征多种多样,基于不同的特征可以发现不同方面的序列机制。Alonso等先对时间序列进行预测,采用时间序列在未来时段预测值的概率密度分布作为时间序列的特征,然后对其进行聚类^[31]。两条序列的距离度量采用了各自概率密度函数差的积分。Singhal等对多维时间序列聚类,采用多维时间序列的主成分以及其各维数据的质量精度来进行聚类^[32]。文中定义3个基础距离度量,分别表示为主成分的夹角、多维数据集的马氏距离以及数据质量精度差异,最终多维时间序列的距离采用3个基础距离的加权和。

Díaz等考虑对时间序列的多种特征聚类^[6],这些特征包括时间序列的自相关函数,部分相关函数,周期解析强度,谱密度等。文中对比了基于不同特征的相似性度量,将其分为了有参数和非参数的聚类方法。有参数的方法先对时间序列的模型参数进行估计,然后基于这些参数计算时间序列的相似度;非参数的方法则采用统计检验,将两条序列来自同一参数模型作为零假设,检验其显著性,作为时间序列之间的相似性度量。实验证明,选择这些时间序列结构特征聚类,可以解决3种时间序列的聚类问题,包括平稳与非平稳时间序列区分,

表4 基于结构特征的时间序列聚类

Tab.4 Structural characteristics based time series clustering				
文献	采用的结构特征	相似性度量	聚类算法	应用数据
Nanopoulos, 等 ^[23]	均值、标准差、偏度、峰度	×	神经网络	×
Ouyang, 等 ^[24]	最大值,最小值,均值,标准差	欧氏距离、DTW距离	K-means	水文数据
Kontaki, 等 ^[25]	趋势特征	DPLA距离	层次聚类	股票数据
Kumar, 等 ^[26]	季节特征	基于数据误差符合高斯模型的假设检验p值	层次聚类	零售商品数据
Wang, 等 ^[19, 27]	均值、标准差、偏度、峰度、趋势、周期、自相关、混沌性	标准化后欧氏距离	层次聚类 神经网络	基准数据 ^[33] 人类行为数据
Caiado, 等 ^[28]	周期解析强度	标准化周期解析强度的对数数据距离	层次聚类 K均值聚类	经济数据
Shumway, 等 ^[29-30]	谱密度	Kullback-Liebler散度 Chernoff对称信息散度	层次聚类	地震和煤矿爆炸数据
Alonso, 等 ^[31]	预测值的概率密度分布	概率密度差的积分	层次聚类	各国CO ₂ 排放量
Singhal, 等 ^[32]	主成分及其数据质量	主成分夹角距离, 原始数据距离, 数据质量距离	K均值聚类	×
Díaz, 等 ^[6]	ACF, PACF, 周期解析强度, 谱密度	欧氏距离, 马氏距离, Piccolo距离 ^[34] , Maharaj距离 ^[35] , Caiado距离等 ^[28]	层次聚类	×

不同 ARMA 过程的时间序列区分以及一些非平稳时间序列之间的区分。

基于结构特征的时间序列聚类可对原始时间序列降维,找出具有相同结构特征的时间序列,从而发现其潜在机制的相似性。同时它很好的解决了噪声问题,并可以处理不等长以及非均匀采样的时间序列数据。但由于结构特征种类繁多,具体选择哪种特征聚类往往与实际问题的密切相关,因此还需对如何选取合适的结构特征作进一步的研究^[9]。

3.3 基于模型特征的时间序列聚类

基于模型特征的时间序列聚类,假设不同簇的时间序列是由具有不同参数的模型创建而来的,而具有相同模型特征的时间序列就认为是相似的。给定一组时间序列,聚类问题就是找出具有代表性的参数模型,根据该模型特征将时间序列分配到相应的组中。这种聚类方法往往更能反映时间序列的自然特性,产生有意义的结果。

基于模型特征的时间序列聚类方法可以分为两种:基于模型参数特征的时间序列聚类和基于混合模型的时间序列聚类(表 5)。基于模型参数特征的聚类对时间序列建立模型,然后将该模型的参数或者拟合的残差作为时间序列的模型特征,以此定义合适的相似性度量进行聚类;基于混合模型的时间序列聚类将时间序列看成由多个模型组件组成的混合模型,计算模型各组件的后验概率或对数似然,根据最大后验概率或最大似然的原则对混合模型各组件中的模型参数进行估计,从而确定时间序列各簇对应的模型。

3.3.1 基于参数特征的时间序列聚类

基于参数特征的聚类方法,与之前基于形态特征和基于结构特征的聚类方法思路大体相同,主要还是建立模型,用模型参数定义序列之间的相似性

度量。Maharaj^[35]针对平稳型时间序列建立了自回归模型(AR),对自回归系数 π 进行估计。采用零假设: $\pi_x = \pi_y$ 的显著性作为两个时间序列的相似性度量,聚类结果可以发现具有相同自回归模型的时间序列。随后 Maharaj 将该方法扩展到多维时间序列聚类上,建立了向量自回归滑动平均模型 VAR-MA^[36],同样采用零假设: $\pi_x = \pi_y$ 的显著性作为两条序列的相似性度量。Ramoni 则对时间序列建立马尔科夫链模型^[37],将每条时间序列看成是一个马尔科夫链,估计其概率转移矩阵,然后定义了转移矩阵的 Kullback-Liebler 距离,作为序列之间相似性度量。通过层次聚类法,结合最大后验概率的原则对时间序列进行聚类。Ramoni 等也将该方法扩展到了多维时间序列聚类上^[38]。

3.3.2 基于混合模型的时间序列聚类

基于混合模型的时间序列聚类核心问题在于对模型参数的估计,参数估计过程中,初始值的选取也往往对聚类结果有一定的影响。目前有很多种参数估计的方法:Xiong 等随机选择初始值,采用 EM 算法对 ARMA 模型的混合模型参数进行了估计^[39],应用于人口数据,气温数据的聚类等。Bicego 等则建立隐马尔科夫模型,先选择 R 条时间序列作为“参考”时间序列^[40],然后通过 Baum-Welch 算法以及前向后项算法^[41]对参数进行估计,其方法优于标准的隐马尔科夫链聚类方法,但还是存在隐马尔科夫链隐状态数未知的问题。Oates 等则针对此问题,采用 DTW 距离先对原时间序列进行聚类找出初始划分,从而推断出隐状态数的初始值,然后通过迭代计算找出最优的隐马尔科夫模型^[42],但是他并没有对聚类簇数的选择问题进行探讨。Li 等则依据最大后验概率的原则,对隐马尔科夫混合模型 4 个层次的参数特征进行估计^[43]——包括聚类簇

表 5 基于模型特征的时间序列聚类

Tab.5 Model characteristics based time series clustering					
文献	采用的模型	模型特征	相似性度量	聚类算法	应用数据
Maharaj ^[35-36]	AR 模型, VAR 模型	自回归系数	模型相同的假设检验 p 值	层次聚类	×
Ramoni, 等 ^[37-38]	马尔科夫链模型	概率转移矩阵	Kullback-Liebler 散度	层次聚类	机器人传感器数据
Xiong, 等 ^[39]	ARMA 混合模型	ARMA 模型参数	后验概率	EM 算法	公共数据
Bicego, 等 ^[40]	离散隐马尔科夫模型	隐马尔科夫模型的参数	对数似然	层次聚类	2 维形状数据
Oates, 等 ^[42]	离散隐马尔科夫模型	隐马尔科夫模型的参数	对数似然	初始聚类 DTW 点操作	机器人传感器数据
Li, 等 ^[43-44]	连续隐马尔科夫模型	聚类簇数,划分的结构, 隐马尔科夫模型的结构 隐马尔科夫模型的参数	对数似然	四层搜索 BIC 准则	生态数据

数,划分的结构,隐马尔科夫模型的结构和隐马尔科夫模型的参数,从而对时间序列进行聚类。随后,Li等在该方法基础上加入了BIC准则,用来更准确的选择聚类簇数和隐马尔科夫模型结构^[44]。

与基于结构特征的时间序列聚类类似,基于模型特征的聚类对噪声不敏感,可以处理不等长或非均匀采样的数据。更重要的是它可以发现序列中具有相同潜在运动规律的过程,更接近时间序列的自然本质,发现事物的动态变化规律。

4 结语与展望

本文系统综述了基于特征的时间序列聚类方法,将其分为基于形态特征的聚类、基于结构特征的聚类和基于模型特征的聚类。根据形态特征的全局性和局部性,可以分为基于全局形态特征的聚类和基于局部形态特征的聚类;根据结构特征的不同来源,可以分为基于统计特征的聚类、基于时域特征的聚类和基于频域特征的聚类;根据模型特征的不同原理,可以分为基于参数特征的聚类和基于混合模型特征的聚类。这样的分类可以让我们们在研究中清楚的认识时间序列不同层次上蕴藏着的模式特征,从不同的角度反映时间序列的成因,加深对时间序列本质的理解。

目前,国际上关于时间序列聚类的方法已经成为热点研究领域,取得一定的研究进展。然而,时间序列聚类方法仍面临许多困难与挑战,其热点和难点问题主要包括以下4个:

(1) 如何结合实际应用问题,充分利用先验知识,选择适当的时间序列聚类方法,使聚类算法更容易产生好的聚类结果^[45]。目前一些学者提出了许多时间序列聚类方法,然而没有哪一种可以说是最好的,使用不同方法产生结果的好坏往往取决于实际应用问题,而少量的先验知识往往比复杂的聚类算法更容易产生有意义的聚类结果。因此如何根据实际问题,结合先验知识,选择最优时间序列聚类方法是有待于进一步研究的问题^[45-46]。

(2) 建立时间序列聚类的基准测试数据^[33]。许多时间序列聚类方法有所缺陷,但由于实验数据本身有一定的特殊性,导致该时间序列聚类算法在实验中效果很好,而在解决实际问题时却得出很糟糕的结果。因此需要建立基准测试数据集,保证时间序列聚类方法的有效性。

(3) 对算法中参数的选择或估计问题^[37, 47]。目前许多聚类算法都需要预先设定参数,如聚类簇数,最小距离阈值等,如何根据实际问题合理的设定或估计这些参数,使聚类结果更好,也是热点研究问题之一^[48]。

(4) 提高算法的效率。海量时间序列数据不断的产生,随着处理的数据量越来越大,时间序列聚类方法也要提高计算效率。

参考文献

- [1] Shumway R H, Stoffer D S. Time Series Analysis and Its Applications with R Examples. New York: Springer, 2009.
- [2] Han J W, Kamber M. Data Mining: Concepts and techniques. Singapore: Elsevier, 2006.
- [3] Košmelj K, Batagelj V. Cross-sectional approach for clustering time varying data. *Journal of Classification* 1990, 7: 99-109.
- [4] Balasubramaniyan R, Hüllermeier E, Weskamp N, et al. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* 2005, 21 (7): 1069-1077.
- [5] Liao T W. Clustering of time series data: A survey. *Pattern Recognition* 2005, 38(11): 1857-1874.
- [6] Diaz S P, Vilar J A. Comparing several parametric and nonparametric approaches to time series clustering: A simulation study. *Journal of Classification*, 2010, 27(3): 333-362.
- [7] Keogh E J, Pazzani M J. An enhanced representation of time series which allows fast and accurate classification, Clustering and Relevance Feedback//*Procs. of the 4th Conference on Knowledge Discovery in Databases*, 1998: 239-241.
- [8] Chen Y G, Nascimento M A, Ooi B C, et al. SpADe: On shape-based pattern detection in streaming time series// *Proceedings of the 23rd International Conference on Data Engineering*, IEEE, 2007: 786-795.
- [9] Wang X Z, Smith K, Hyndman R. Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 2006, 13(3): 335-364.
- [10] Rose O. Estimation of the Hurst Parameter of Long-Range Dependent Time Series. Research Report, 1996.
- [11] Hilborn R C, Ottino J M, Shinbrot T. Chaos and nonlinear dynamics: An introduction for scientists and engineers. *AIChE Journal* 1995, 41(7): 1831-1832.

- [12] Tian Z, Raghu R, Miron L. BIRCH: An efficient data clustering method for very large databases. *SIGMOD Rec*, 1996, 25(2): 103-114.
- [13] Karypis G, Han S, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 1999, 32(8): 68-75.
- [14] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure. *SIGMOD Rec*, 1999, 28(2): 49-60.
- [15] Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining//*Proceedings of the 23rd Conference on VLDB*, 1997: 186-195.
- [16] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans*, 2000, 22(7): 719-725.
- [17] Keogh E, Ratanamahatana C A. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 2005, 7(3): 358-386.
- [18] Möller-Levet C S, Klawonn F, Cho K H, et al. Clustering of unevenly sampled gene expression time-series data. *Fuzzy Sets and Systems*, 2005, 152(1): 49-66.
- [19] Möller-Levet C S, Klawonn F, Cho K H, et al. Fuzzy clustering of short time-series and unevenly distributed sampling points//*Proceedings of the 5th International Symposium on Intelligent Data Analysis*, Berlin, Germany, August 28-30, 2003.
- [20] Fu T C, Chung F L, Vincent N, et al. Pattern discovery from stock time series using self-organizing maps//*KDD 2001 Workshop on Temporal Data Mining*. San Francisco, 2001: 27-37.
- [21] Hsu K C, Li S T. Clustering spatial-temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources* 2010, 33(2): 190-200.
- [22] Lee J G, Han J W, Whang K Y. Trajectory clustering: a partition-and-group framework. *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2007: 593-604.
- [23] Nanopoulos A, Alcock R, Manolopoulos Y. Feature-based classification of time-series data. *International Journal of Computer Research*, 2001: 49-61.
- [24] Ouyang R, Ren L, Cheng W, et al. Similarity search and pattern discovery in hydrological time series data mining. *Hydrological Processes*, 2010, 24(9): 1198-1210.
- [25] Kontaki M, Papadopoulos A N, Manolopoulos Y, et al. Continuous trend-based clustering in data streams. *Data Warehousing and Knowledge Discovery*, 2008, 5182: 251-262.
- [26] Kumar M, Patel N R, Woo J. Clustering seasonality patterns in the presence of errors. in *ACM KDD Conference Proceedings*, 2002: 557-563.
- [27] Wang X, Wirth A, Wang L. Structure-based statistical features and multivariate time series clustering//*Proceedings of the Seventh IEEE International Conference on Data Mining*, 2007: 351-360.
- [28] Caiado J, Crato N, Peña D. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis* 2006, 50(10): 2668-2684.
- [29] Kakizawa Y, Shumway R H, Taniguchi M. Discrimination and Clustering for Multivariate Time Series. *J. Amer. Stat. Assoc*, 1998, 93(441): 328-340.
- [30] Shumway R H. Time-frequency clustering and discriminant analysis. *Statistics & Probability Letters*, 2003, 63(3): 307-314.
- [31] Alonso A M, Berrendero J R, Hernández A, et al. Time series clustering based on forecast densities. *Computational Statistics & Data Analysis*, 2006, 51(2): 762-776.
- [32] Singhal A, Seborg D E. Clustering multivariate time-series data. *Journal of Chemometrics*, 2005, 19(8): 427-438.
- [33] Keogh E, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery* 2003, 7(4): 349-371.
- [34] Piccolo D. A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 1990, 11(2): 153-164.
- [35] Maharaj E A. Cluster of time series. *Journal of Classification*, 2000, 17(2): 297-314.
- [36] Maharaj E A. Comparison and classification of stationary multivariate time series. *Pattern Recognition*, 1999, 32(7): 1129-1138.
- [37] Ramoni M, Sebastiani P, Cohen P. Bayesian Clustering by Dynamics. *Machine Learning*, 2002, 47(1): 91-121.
- [38] Ramoni M, Sebastiani P, Cohen P. Multivariate clustering by dynamics//*Proceedings of the Seventeenth National Conference on Artificial Intelligence*, 2000: 633-638.
- [39] Xiong Y, Yeung D Y. Mixtures of ARMA Models for Model-Based Time Series Clustering. *Proceedings of IEEE International Conference on Data Mining*, 2002: 717-720.
- [40] Bicego M, Murino V, Figueiredo M A T. Similarity-based clustering of sequences using hidden Markov models. *Machine Learning and Data Mining in Pattern Recognition*, 2003, 2734: 86-95.
- [41] Rabiner L R. A tutorial on hidden Markov models and

- selected applications in speech recognition. Proceedings of the IEEE 1989, 77(2): 257-286
- [42] Oates T, Firoiu L, Cohen P R. Clustering time series with hidden markov models and dynamic time warping. Proceedings of the IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Learning Methods for Sequence Learning, 1999.
- [43] Li C, Biswas G. Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification. Advances in Intelligent Data Analysis., 1999: 245-256.
- [44] Li C, Biswas G, Dale M, et al. Building models of ecological dynamics using HMM based temporal data clustering: A preliminary study. Advances in Intelligent Data Analysis, 2001: 53-62, doi: 10.1007/3-540-44816-0_6.
- [45] Jain A K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2009, 31(8): 651-666.
- [46] Wang N Y, Chen S M. Temperature prediction and TAI-FEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series. Expert Systems with Applications, 2009, 36(2): 2143-2154.
- [47] Fr uhwirth-Schnatter S. Model-based clustering of time series: A review from a Bayesian perspective. Manuscript, 2011.
- [48] Pakhira M K, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition 2004, 37(3): 487-501.

Research Progress in Time Series Clustering Methods Based on Characteristics

SONG Ci, PEI Tao

(State Key Lab of Resources and Environmental Information System,

Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China)

Abstract: As terabyte time series data pour into the world, more and more attentions have been paid to the technique of analyzing this data. To understand discrepancy between these data, time series clustering methods have been used to divide them into different groups by similarities. Due to high dimension of time series, the traditional clustering methods for static data is not valid for time series clustering problem when they are susceptible to noise, and can hardly define suitable similarity which are prone to a meaningless result. It is also vexatious for many other methods to solve the clustering problem with missing or unequal data. Time series clustering methods based on characteristics could deal with these problems and discover the essential similarities of time series in all directions. According to characteristics of time series, this paper aimed to review the research progress of characteristics-based clustering methods for time series. Firstly, we introduced the definition and classified the different characteristics of time series. Then we reviewed different time series clustering methods based on characteristics and summarized the generality of each method. Finally we discussed some deficiencies of existing methods, and predicted the future of the relative research.

Key words: time series; characteristics of time series; data mining; clustering

本文引用格式:

宋辞, 裴韬. 基于特征的时间序列聚类方法研究进展. 地理科学进展, 2012, 31(10): 1307-1317.