时间序列挖掘中一种新的相似性度量

管河山[1],姜青山[2],Wang Shengrui[2,3]
GUAN He-shan[1],JIANG Qing-shan[2],WANG Shengrui[2,3]

1.厦门大学 计算机系,福建 厦门 361005
2.厦门大学 软件学院,福建 厦门 361005
1.Department of Computer Science,Xiamen University,Xiamen,Fujian 361005,China
2.School of Software,Xiamen University,Xiamen,Fujian 361005,China
3.Department of Computer Science,University of Sherbook,Quebec,Canada
E-mail:guanheshan@yahoo.com.cn

GUAN He-shan,JIANG Qing-shan,WANG Shengrui.New similarity measure for mining time series.Computer Engineering and Applications,2007,43(26):152-155.

Abstract: Proposes a new similarity measure-global characters for whole clustering of time series,that replaces the raw data with 11 global characteristics,from the aspects of statistical distribution,non-linear and Fourier transformation,thus can get a characteristic vector,which can hold most information of the original time seiries and reduce the calculating complexity.Experimentally compares the four similarity measures on three database under group-ward hierarchical clustering,evaluates the results objectively and subjecttively respectively,and is shown to yield useful and reasonable clustering,especially for economic time series.
Key words: time series;cluster;Euclidean distance;autocorrelation function;cepstrum;global characteristics;group-ward hierarchical clustering

摘 要:针对时间序列的全序列聚类展开,提出一种新的相似性度量——全局特征,即从时间序列的统计分布特征、非线性和Fourier频谱转换等3个方面提取11个全局特征构建特征向量。利用特征向量来描述原时间序列,不仅保留了大部分原有的信息,还能加快聚类计算的速度。经过大量的实验验证表明,基于全局特征提取的相似性度量能得到合理的聚类结果,特别是对经济领域的时间序列效果更为明显。例举了2个数据进行实验,并从主观和客观两个角度对聚类结果进行评估。
关键词:时间序列;聚类;Euclidean距离;自相关系数;谱系数;全局特征;层次聚类
文章编号:1002-8331(2007)26-0152-04 **文献标识码**:A **中图分类号**:TP311

1 序言

时间序列是一种在科学研究、商业应用中普遍存在的数据形式。比如经济、生物、医学、物理、商业等多种领域;经济学中通常可以根据一个国家的某些经济指标的时间序列,比如国民收入、就业率、通货膨胀程度等,来判定一个国家的经济状况;医学中,通常可以根据病人的脑电图、心电图等来判定病人的病况。目前时间序列挖掘的工作主要集中在以下4个方面:索引(Indexing)、聚类(Clustering)、分类(Classification)和分割(Segmentation)。其中时间序列聚类在数据挖掘领域中得到了广泛地应用[1-4]。本文主要针对时间序列的全序列聚类展开探讨(Whole Clustering),且主要针对经济领域的时间序列聚类进行分析,给出适合经济领域的时间序列挖掘的一种新方法,并通过经济数据分析给出一些合理的信息和决策,从实践中体现出时间序列挖掘的重要作用。所以本文在实验时只是采用了一些在经济领域应用较多的方法进行对比。

时间序列由于其特定的形状特征,使得目前常用的一些相似性度量和聚类方法失去了原有的优越性,而几乎所有的时间序列聚类算法都涉及到计算序列之间的相似性问题。例如,Euclidean距离仍然是常用的相似性度量[5,6],但是Euclidean距离的鲁棒性较差,而且采用Euclidean距离进行聚类,其计算的复杂度很高;Wang,C[7]提取时间序列的自相关函数作为时间序列的相似性度量,该方法的处理效果很大程度上受到ACF系数收敛性的影响;Kalpakis,K[8]提取时间序列的谱系数(Cepstrum)来衡量时间序列之间的相似度,该方法很大程度上也受到模型系数的影响。此外,一些传统的经典方法,例如PCA和小波变化(wavlet)等方法在众多文章中都有采用[8],本文中就不作详细论述。总之,时间序列的相似性度量直接影响到聚类的结果,这使得寻找一个合理的相似性度量成为了时间序列聚类(甚至是分类、索引)问题中一个重要的步骤。

目前,提取时间序列的特征来描述原时间序序列成为了一个重要的途径[4],这样不仅可以得到更为合理的结果,而且通过用少量特征来描述大型时间序列,可以大大降低聚类计算过程中的复杂度。在目前对时间序列的研究中,统计分布特征、非线性分析和Fourier频谱转换得到了广泛关注,本文正是从这三

基金项目:厦门大学985二期信息创新平台项目(No.0000-X07204)。

个方面提取时间序列的全局特征,例如趋势特征、周期特征、自相关系数(ACF)、逆自相关系数(IACF)、偏自相关系数(PACF)、偏度特征、峰度特征、LE 指数和时间序列从时间域向频谱域转换(Fourier 变换)后的截尾系数(前两阶)作为时间序列的特征,共计 11 个全局特征,进而构建时间序列的全局特征向量,用全局特征向量来描述原时间序列,不仅可以保留原始时间序列的信息,而且可以大大降低聚类计算过程的复杂度。利用全局特征向量来进行时间序列的聚类,在很大程度上避免了直接利用 Euclidean 距离鲁棒性差的弊端。同时,全局特征的提取不需要对数据作出过多的假设,这使得该方法在很大程度上不受到领域知识的限制,适合在多种领域内运用。实验证明,该方法进行时间序列聚类可以得到合理的结果,特别是对经济领域的时间序列的处理效果更佳。

本文的结构如下,第 2 章具体讨论了全局特征的提取和建立相似性度量模型;第 3 章采用了一个经济领域的数据和一个股票数据来进行实验,从主观和客观两个方面进行实验结果的评估[6],并采用了常用的 Euclidean 距离、ACF 系数法和谱系数(Cepstrum)来进行实验结果对比。第 4 章主要对研究的工作做一些总结,并指出今后一些可行的研究方向。

2 时间序列全局特征提取

从不同的角度出发来考虑时间序列本身特征,通常可以得到一些不同的特征。本文力求从多个角度出发提取时间序列的全局特征,并确保这些特征不重复描述原时间序列的信息,以少量的特征来准确地描述出原时间序列的信息。

统计特征在许多时间序列的分析过程中都必须考虑。统计模型在时间序列的研究中得到普遍应用,特别是加法模型和 ARIMA 模型,本文从加法模型角度出发,提取趋势特征和周期特征;从时间序列的 ARIMA 模型角度出发,提取自相关函数、逆自相关系数和偏自相关系数等特征,从时间序列本身数据分布特征角度出发,提取偏度特征和峰度特征;并且考虑到时间序列的非线性特征,提取 LE 指数作为一个全局特征;此外进行时间序列的转换,即采用傅立叶变换(Fourier)将时间域转换到频谱域,并提取变换后的截尾系数(前两阶,共 3 个特征)作为时间序列的特征。这样可以得到时间序列的 11 个全局特征并构成特征向量,每个时间序列对应一个特定的特征向量,用特征向量来衡量时间序列之间的相似度,进而进行聚类分析。下面就各个全局特征的提取展开论述。

2.1 趋势特征

时间序列的一个直观特征就是趋势特征(单调特征),利用趋势特征来刻画时间序列是一种常用的方法,例如时间序列的加法模型和乘法模型都利用了趋势项特征。本文提取趋势特征作为时间序列的一个全局特征。即用线性函数拟合时间序列,其参数的估计采用最小二乘法,得到一个斜率(系数 β),用斜率来衡量时间序列的趋势特征。对于时间序列 $Y=\{Y_1,Y_2,\cdots,Y_n\}$,其关于时间的一元回归模型如式(1):

$$Y=\alpha+\beta*T \tag{1}$$

根据最小二乘法容易得到自变量 t 的系数 β 的估计值:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(T_i-\bar{T})}{\sum_{i=1}^{n}(T_i-\bar{T})^2} \tag{2}$$

其中 $\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$,$\bar{T}=\frac{1}{n}\sum_{i=1}^{n}T_i$。

2.2 周期性特征

时间序列另一个直观特征就是表现出一定的周期性(或者季节性特征),特别是经济时间序列的一些研究中对周期性特别重视,所以本文也采用了周期性作为时间序列的一个全局特征。在此对于没有表现出明显周期性的时间序列,其周期性定为 1,其他情况下,时间序列的周期可以通过一个固定的算法得到,算法如下:

(1)计算时间序列的自相关函数,为了保证能准确地找到周期性,计算自相关系数时,取滞后步长为(1–1/2 个时间序列长度的自相关系数)并找出所有的波峰和波谷;

(2)求出波峰和最近邻的波谷之间的差,要求波峰为正,前面有一个波谷,而且差值大于某个阀值(该阀值需根据数据的特性来决定,通常取 0.1[4]),取出满足条件的第一个波峰,此时波峰所对应的自相关系数定义为周期的值。

其中自相关函数的计算公式如下

$$r_k=\frac{\sum_{t=1}^{n-k}(y_t-\bar{y}_t)(y_{t+k}-\bar{y}_{t+k})}{\sqrt{\sum_{t=1}^{n-k}(y_t-\bar{y}_t)^2\cdot\sum_{t=1}^{n-k}(y_{t+k}-\bar{y}_{t+k})^2}} \tag{3}$$

其中 $\bar{y}_t=\frac{1}{n-k}\sum_{t=1}^{n-k}y_t$,$\bar{y}_{t+k}=\frac{1}{n-k}\sum_{t=1}^{n-k}y_{t+k}$。

2.3 峰度特征和峰度特征

时间序列的数据分布状况是很多研究中需考虑的因素之一。可以采用偏度和峰度来刻画时间序列的数据分布特征。峰度用于度量总体分布尾部"粗细"状况的数字特征,这里的"粗细"是与正态分布的尾部而言的;偏度用于度量总体分布状况偏斜程度的数字特征。本文提取了这两方面的特征值作为时间序列的全局特征。其中偏度的计算公式如式(4),峰度计算公式如式(5)。

$$G_1=\frac{E(x-\mu)^3}{\sigma^3} \tag{4}$$

$$G_2=\frac{E(x-\mu)^4}{\sigma^4}-3 \tag{5}$$

其中 n 为样本个数,μ 表示均值,σ 为标准差。

2.4 自相关函数

ARIMA 模型是时间序列常用的模型之一,其中 MA 阶数通常可以根据自相关函数的数字特征进行判定,特别的,可以根据其收敛速度来判定。本文采用了 $Q_h=\sum_{k=1}^{h}r_k^2$ 统计量[4]作为时间序列的一个全局特征。

2.5 逆相关系数和偏自相关系数

ARMA 模型中的 AR 阶数通常可以根据逆相关系数和偏相关系数来进行判定,所以本文构造了两个统计量,即 $Q_h=\sum_{k=1}^{h}I_k$ 统计量(I_k 表示逆相关系数)和 $Q_h=\sum_{k=1}^{h}P_k$ 统计量(偏相关系数)作为时间序列的全局特征,其中 k 的取值跟 2.4 节中的 Q_k 量一样。

2.6 李雅普诺夫指数(LE 指数)

通常时间序列是一个复杂的序列,单纯的线性模型难以描

述。时间序列的非线性模型得到了广泛的研究和应用。李雅普诺夫指数(Lyapunov指数)可以确定产生时间序列的过程是否是混沌的，本文采用LE指数来刻画时间序列的非线性状况，其计算公式采用Hilborn[9]的算法，算法如下：

(1)假定时间序列 $Y=\{Y_1, Y_2, \cdots, Y_n\}$；

(2)假定为h个滞后阶，找出点 y_i 最近邻点 y_j，其中 $i\neq j$ 且 $\| y_i - y_j \|$ 很小，定义 $\lambda(y_i, y_j)=\frac{1}{h}\ln\frac{\| y_i - y_j \|}{\| y_{i+h} - y_{j+h} \|}$；

(3)计算出所有点 y_i 的上述 $\lambda(y_i, y_j)$ 值，对i取平均值，估计LE指数取值 $\lambda=\frac{1}{n}\sum_{i=1}^{n}\lambda(y_i, y_j)$

2.7 Fourier变换系数

Fourier变换的基本思想是将信号分解成一系列不同频率的连续正弦波的叠加，对信号 $f(x)$，可以按照式(6)进行Fourier变换。将信号从时域转化到频谱域，可以说它是时域信号的一个总体统计。时间序列可以看成时间域的某个信号，因此本文采用Fourier变换后的系数(前两阶共3个系数)作为时间序列的三个全局特征。

$$F(\omega)=\int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}dx \qquad (6)$$

2.8 建立模型

提取时间序列的全局特征构建时间序列的特征向量，此时每个时间序列将对应一个特征向量，事先对特征向量集进行标准化，然后再进行建模。

正如上文所述，相似性度量在时间序列挖掘中发挥着重要的作用。进行时间序列聚类，首先要确定其聚类的距离函数，或者说一个合理的样本之间相似性度量函数。本文利用时间序列的特征向量进行计算，采用Euclidean距离来计算两个时间序列之间的相似度，即对任意两个时间序列特征向量 $X=\{X_1, X_2, \cdots, X_m\}$ 和 $Y=\{Y_1, Y_2, \cdots, Y_m\}$，可以按照式(7)计算之间的相似度

$$d_{X,Y}=\sum_{i=1}^{m}(X_i - Y_i)^2 \qquad (7)$$

其中m为特征向量的维度(本文m取值为11)。然而根据研究者不同的目的，可能会各有所侧重，比如经济时间序列，更侧重其周期性；股票价格的时间序列，更侧重其趋势性；所以本文针对具体的问题处理时，通过设计一个加权函数来构建时间序列相似度的衡量尺度模型，如式(8)

$$d_{X,Y}=\sum_{i=1}^{m} u_i(X_i - Y_i)^2 \qquad (8)$$

其中 $\sum_{p=1}^{m} u_p=1$。本文的计算是均等考虑各个特征在聚类中所发挥的作用，为了进行比较采取了相同的权值来进行聚类计算。

3 时间序列聚类实验

大量的实验表明，本方法给时间序列挖掘提供了一个合理的途径。本文例举了一个真实数据(美国个人收入增长数据[8])和中国股市9个股票的历史价格数据来进行测试，其中前一个数据是已经分类的数据，可以从客观角度进行评估聚类结果，如3.1节所述。而第二个数据是采用股票某历史时间段内的数据，事先不知道分类结果，但是可以根据聚类的图像来主观判定聚类效果，继而判定相似度衡量尺度的性能[6]。本文采用了经济领域应用较多的方法，如Euclidean距离、ACF系数法和谱系数(Cepstrum)来进行实验结果的比较分析，采用Euclidean距离计算时，事先将数据进行了标准化。本文采用了ward的层次聚类。

3.1 聚类结果的评价

假定两个聚类的结果 $G=G_1, G_2, \cdots, G_k$(真实结果)和 $A=A_1, A_2, \cdots, A_k$(某种方法得到的聚类结果)，采用下面的指标(式(9))来评估聚类结果[8]。当 $Sim(G, A)$ 越大时，说明采用某种方法得到的聚类结果越合理，反之，聚类方法的结果越不合理

$$Sim(G, A)=\frac{\sum_{i=1}^{k}\max_j Sim(G_i, A_j)}{k} \qquad (9)$$

其中 $Sim(G_i, A_j)=\frac{2|G_i\cap A_j|}{|G_i|+|A_j|}$。

3.2 美国个人收入增长数据

该数据收集了美国各洲自1929年至1999年逐年的收入情况，该数据可以较好的用ARIMA模型来拟合，该数据也被Cepstrum系数的作者用来实验[8]，他采用了25个州的数据进行实验，进而评价聚类方法的实际效果，本文也采用了该数据，以便更好的进行实验结果的比较。

采用上述4种不同的相似性度量进行聚类，得到的结果如表1。可见此时基于全局特征提取的相似性度量的 $Sim(G, A)$ 指标的取值接近1，实际处理效果比Cepstrum系数更为合理。而ACF系数在刻画时间序列的特征时，$Sim(G, A)$ 指标的取值仅有0.627，效果不是很明显，尽管ACF系数在区分时间序列的平稳和非平稳问题中[10]得到了广泛地应用。Euclidean距离的方式也能得到较好的结果，但是比Cepstrum系数和全局特征度量的结果要稍差。从实际的聚类结果中可以发现，本文结果跟实际结果非常近似，能很好地将经济发达地带和经济滞后地带区分开来，这给经济决策提供了强大的理论支持。

表1 4种不同相似性度量的聚类结果

相似性度量方式	$Sim(G, A)$
Euclidean距离	0.851
全局特征	0.851
ACF	0.627
Cepstrum系数	0.844

3.3 股票数据

该数据收集了中国股市中上证指数的9个股票的历史数据(股票代号为600739.ss、600031.ss、600550.ss、600262.ss、600030.ss、600151.ss、600155.ss、600152.ss、600193.ss)，本文为了确保数据的时效性，以2006年9月8日为中止日，向前取股票每日的历史数据，共计200个数据，构建一个数据集。然后进行实验，由于所采取的股票事先并没有一个准确的分类(或者说一个标准的分类结果)，因而本文将聚类的结果描述成图像，进而主观分析聚类的结果。

采用上述4种不同的相似性度量进行聚类，分别得到的聚类系统树形图，如图1。实验数据表现出明显的两种模式，分2类时，ACF和全局特征度量都可以准确的将样本归类，图中上面6个股票的数据表现出明显的趋势特征，即先递增后递减；而此时Euclidean和Cepstrum的结果就不如前两种度量方式的好，在它们的结果中，样本表现出混杂不分的情形，没有很好地区别两类样本。本文方法给股票分类提供了一种新的途径，同时利用全局特征来建立一个股票数据的索引机制，将有利于

股票的查询、检索等工作。

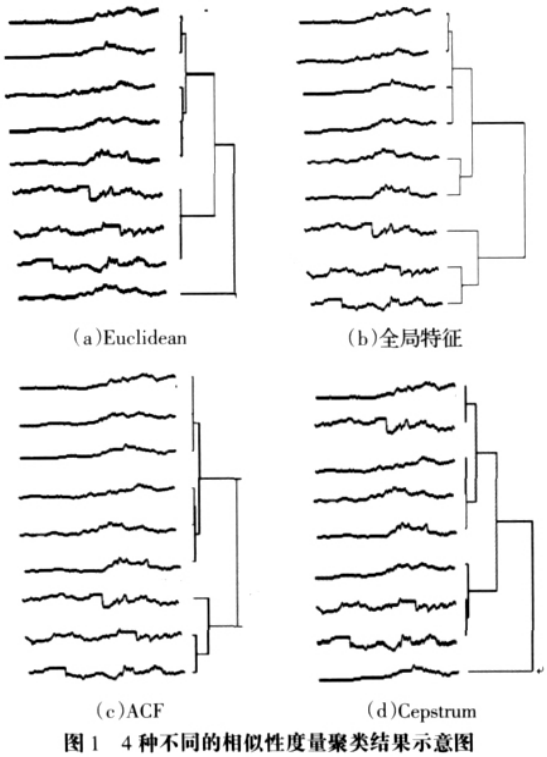

图1 4种不同的相似性度量聚类结果示意图

由上述分析可见,基于时间序列全局特征提取的相似性度量能很好地保留原时间序列的信息,适合时间序列的聚类,特别是对经济领域的时间序列,更能体现其聚类的优势。这给经济领域的时间序列提供了一种新的途径,这也给经济决策提供一个合理的理论支持。

4 结论

时间序列挖掘的研究已经得到广泛重视,目前关于时间序列相似性度量的研究在时间序列挖掘的工作中占据了主要地位。本文提出的基于时间序列全局特征提取的相似性度量能处理多领域的时间序列聚类,特别对经济领域的时间序列,更能体现起优势。跟传统的降维方法相比,该方法通常不需要太多的假设和领域知识,通用性很好。同时该方法也可以处理不同长度的时间序列的特征提取,适合不同纬度的时间序列的聚类,这是直接采用 Eclidean 距离所不能的;根据特征向量来处理时间序列的聚类可以大大降低聚类计算的复杂度。实验结果表明,该方法可以得到较好的聚类结果。此外该方法还可以在一定程度上避免缺失数据对时间序列聚类的影响。下一步工作就是进一步探讨含缺失数据的时间序列聚类问题。

(收稿日期:2007 年 1 月)

参考文献:

[1] 翁颖钧,朱仲英.基于动态时间弯曲的时序数据聚类算法的研究[J].计算机仿真,2004.

[2] 段江娇,薛永生,林子雨,等.一种新的基于隐 Markov 模型的分层时间序列聚类算法[J].计算机研究与发展,2006.

[3] Keogh E, Lin J, Truppel W.Clustering of time series subsequences is meaningless: implications for past and future research[C]//The 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA, 2003.

[4] Wang Xiao-zhe, Smith K A, Hyndman R J.Dimension reduction for clustering time series using global characteristics[C]//LNCS 3516: ICCS 2005, 2005: 792-795.

[5] Popivanov I, Miller R J.Similarity search over time series data using wavelets[C]//The 18th International Conference on Data Engineering, San Jose, CA, USA, 2002.

[6] Keogh E, Kasetty S.On the need for time series data mining benchmarks: a survey and empirical demonstration[C]//The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002.

[7] Wang C, Wang X S.Supporting content-based searches on time series via approximation[C]//Proceedings of the 12th Int'l Conference on Scientific and Statistical Database Management, Berlin, Germany, Jul 26-28, 2000: 69-81.

[8] Kalpakis K, Gada D, Puttagunta V.Distance measures for effective clustering of ARIMA time-series[C]//IEEE International Conference on Data Mining, San Jose, CA, USA, 2001.

[9] Hilborn R C.Chaos and nonlinear dynamics: an introduction for scientists and engineers[M].New York: Oxford University Press, 1994.

[10] Caiado J, Crato N, Pe D.A periodogram-based metric for time series classification[J].Computational Statistics & Data Analysis, 2006, 50: 2668-2684.

(上接113 页)

因素,本文在原有的分配及调度算法的基础上引入了节点生命周期的概念,使最终选择的节点主机具有更好的稳定性,并结合原有机制中的基于文件分段的存贮策略,取得了令人满意的结果。实验证明,在维持原有的初始播放等待时延不变的前提下,系统的服务容量及服务完成率都有较大的改进,分别提高了 10%~20%之多。(收稿日期:2007 年 5 月)

参考文献:

[1] Fei Z, Yang M.A segmentation-based fine-grained peer sharing technique for delivering large media files in content distribution networks[J].IEEE Transactions on Multimedia, 2006, 8(4): 821-829.

[2] Kim T, Ammar M.A comparison of layering and stream replication video multicast schemes[C]//Proc NOSSDAV'01, 2001.

[3] Zhang Z L, Wang Y, Du D H, et al.Video staging: a proxy-server-based approach to end-to-end video delivery over wide-area networks[J].IEEE/ACM Trans Networking, 2000, 4(8): 429-442.

[4] Ramesh S, Rhee I, Guo K.Multicast with cache(mcache): an adaptive zero-delay video-on-demand service[C]//Proc IEEE Infocom'01, 2001.

[5] Santos J R, Muntz R.Comparing random data allocation and data striping in multimedia servers[C]//Proc ACM Sigmetrics, Santa Clara, CA, 2000.

[6] Drapeau A L, Chen P M, Hartman J H, et al.RAID-II: a high-bandwidth network file server[C]//Proc 21st Int Symp Computer Architecture, Chicago, IL, Apr 1994.

[7] Gao L, Kurose J, Towsley D.Efficient schemes for broadcasting popular videos[C]//Proc NOSSDAV'98, 1998.

[8] Banerjee S, Bhattacharjee B, Kommareddy C.Scalable application layer multicast[C]//Proc of ACM SIGCOMM, Aug 2002.

[9] Bustamante F E, Qiao Y.Friendships that last: peer lifespan and its role in P2P protocols[C]//Proc of 8th WCW Workshop, 2003.