小波变换的离群时序数据挖掘分析

文 琪[1]，彭 宏[2]

(1. 西华大学科技处 成都 610039; 2. 西华大学计算机与数理学院 成都 610039)

【摘要】针对时序数据进行离群数据挖掘方法的研究。通过对时序数据进行离散小波变换，将其从时域空间变换到频域空间，使时序数据映射为多维空间的点。该方法具有多尺度、时移不变性等特点，经离群时间序列进行离散小波变换后，不仅具有良好的保距性又达到降低维数目的。然后提出一种基于距离的离群时序数据挖掘算法。仿真试验表明了该方法的有效性。

关 键 词 小波变换; 时序数据; 离群数据; 数据挖掘

中图分类号 TP311 文献标识码 A

Analysis of Time Series Outlier Mining Based on Wavelet Transform

WEN Qi[1]，PENG Hong[2]

(1. Dept. of Science and Technology, Xihua University；2. School of Computer & Mathematical-Physical Science, Xihua University, Chengdu 610039)

Abstract In this paper, the outlier mining method for time series data is investigated. DWT is used to transform the time series data from time domain to frequency domain. The time series data can be mapped into the multidimensional points in multidimensional space. We proposed a distanced-based algorithm to mine the outliers. The simulation results show the effectiveness of the method.

Key words wavelet transform; time series; outlier data; data mining

离群数据挖掘是指从大量数据中挖掘明显偏离、不满足一般行为模式的数据。对离群数据挖掘的研究往往可以发现一些潜在的有用信息，如股票交易数据异常、银行信用卡欺诈行为的检测等。离群数据已在统计学领域得到广泛研究[1]，但使用统计方法建立概率分布模型时需要事先知道数据集的分布和分布参数等信息。Knorr和Ng提出基于距离的离群数据挖掘方法[2]，但这种方法中的距离难以确定，而且没有离群数据的离群衡量测度。Arning等提出一种基于偏离的离群数据发现方法[3]，通过确定相异函数来进行离群数据挖掘，但当相异函数选取不合适，则会得不到满意的结果。在对时序数据进行离群数据挖掘时，由于其长度较长，直接处理整个时序数据会带来较大的计算复杂性，因此往往需要对数据进行降维，即在进行模式匹配之前，需要对数据维度进行约简。因此本文提出了使用小波变换来进行离群时序数据挖掘的方法。

1 时序数据的离散小波变换

假设小波变换满足条件 $\int_R \phi(x)\mathrm{d}x = 0$，且基本小波函数 $\phi(x)$ 的平移和伸缩构成的一簇小波函数系表示(逼近)一个函数。二进制小波是由伸缩因子和平移因子满足一定条件的一组函数：

$$\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k),\ j,k \in Z \tag{1}$$

对任意平方可积函数 $f(x)$ 来说，其离散小波变换(DWT)为：

收稿日期：2005－05－10

作者简介：文 琪(1967－)，女，硕士，工程师，主要从事数据挖掘等方面的研究.

$$W_f(j,k)=<f,\phi_{j,k}>=\int_{-\infty}^{\infty}f(x)\phi_{j,k}(x)\mathrm{d}x,\ f(x)\in L^2(R) \tag{2}$$

假设$\{f(x)\}_{x\in Z}$有n个非零样本值，Mallat提出的离散二进制小波变换的计算程序为：

$$c_{j,n}=\frac{1}{\sqrt{2}}\sum_k c_{j-1,k}\overline{h_{k-2n}}\quad ,\quad d_{j,n}=\frac{1}{\sqrt{2}}\sum_k c_{j-1,k}\overline{g_{k-2n}} \tag{4}$$

式中　h_k，g_k是离散滤波器。

由于小波变换将时间序列分为尺度部分C与细节部分D，尺度部分通过待分析序列卷积低通滤波器得到，反映了原序列的大致趋势和走向；而细节部分通过待分析序列卷积高通滤波器得到，表示信号在细节上的差异。由于小波的多尺度分解性质，对尺度部分进一步分解得到更详细的尺度部分和细节部分，这个过程可以一直继续下去。对长度为n的序列实施DWT，经过下取样之后，得到2个长度为$n/2$的尺度序列C'与D'，再进行一次DWT的话，得到的尺度序列的长度为原序列的长度的1/2。可以看出，随着尺度数越高，得到的尺度序列的长度越短，信号越模糊。因此，当进行时间序列相似性匹配时，可以只考虑尺度序列，可达到大幅度约简数据量，而信号量相对丢失较少。本文只考虑取前k个尺度序列c_j作为从时间序列上提取的特征，将时序数据的子序列映射为k维空间上的点。

2　基于距离的离群时序数据挖掘

时序数据的离群数据挖掘就是在数据库中发现一部分与其余数据有明显不同的例外数据。为了评价这部分离群数据与其余数据有明显不同，需要采用距离函数作为判别函数，序列间的距离函数常用欧氏距离：

$$dist(X,Y)=(\sum_{i=0}^{n-1}(X_i-Y_i)^2)^{\frac{1}{2}} \tag{5}$$

由Parseval定理[4]，如果采用正交小波，那么小波变换后得到的尺度序列之间的欧氏距离将不超过原始序列之间的欧氏距离，即：

$$dist(x,y)>dist(X,Y) \tag{6}$$

因此，若$dist(X,Y)>M$，则$dist(x,y)>M$。这一保距性质保证了尺度序列的判别的离群数据包含了所有的正确结果。

在进行时序数据的离群挖掘时，由于时序数据的长度一般较长，直接对整个时序数据进行离群数据挖掘，其计算复杂度增大。因此，把时序序列划分为一系列子序列，用DWT变换将子序列时序数据从时域变为频率空间，由于小波变换的多尺度分解性质，可以用尺度序列代替原序列，并且取前k个尺度作为序列的特征，这样从每个序列获得k个特征，进一步把它们作为到k维空间上的一个映射，即将时序数据子序列映射为k维空间的点。这样保留了时序数据的主要特征，但降低了时序数据的维数，降低了计算的复杂度。

对于映射到k维空间的时序点，采用一种$D^k(p)$作为衡量离群数据度量的离群数据定义。

定义 1　假设k维空间Ω有m个点，任取$p\in\Omega$，且记$d(p,x)$为p与$x\in\Omega$的点的距离。把Ω中不包含p数据点集合$d(p,x)$依从小到大顺序排列，得到序列

$$d(p,x_1),\ d(p,x_2),\ldots,\ d(p,x_{m-1}) \tag{7}$$

记序列中第k个值$d(p,x_k)$为点p的第k个最近邻点与点p的距离，记为$D^k(p)$。

定义 2　对于集合Ω中的所有点，给定整数n与k，将点$p\in\Omega$的$D^k(p)$按从达到小顺序排列，其中前n个点为离群数据点。

为了对k维空间上的时序数据点进行离群数据挖掘，采用点$p\in\Omega$的第k个最近邻点与点p的距离$D^k(p)$作为衡量离群数据的度量。基于$D^k(p)$的离群数据距离定义对时序数据点进行k近邻查询，再根据离群数据的定义，搜索到的输入数据集合中前n个$D^k(p)$最大的数据点即为离群数据点。

对于小波变换后的m个点，采用最小边界矩形MBR进行相似性数据挖掘。采用R+树来存储这些MBR，用R+树进行检索。其算法分为：(1) 求取数据集合的每个点p的第k个最近邻与该点的距离$D^k(p)$；(2) 根据$D^k(p)$挖掘离群数据点。计算点p的$D^k(p)$与参考点r的$D^k(r)$的差，将其依从小到大的顺序排列，选取前n个差对应的点为相似数据点。

算法：离群数据查询。

输入：经DWT变换后的k维空间的时序点集P。

输出：离群数据点集C。

(1) 设置离群数据点集C初始为空。

(2) 将P中的每个点 p 和MBR插入到R+树中。

(3) 在k树中调用快速最近邻算法[5]求得P中每个点 p 的第k个最近邻距离 $D^k(p)$ 。

(4) 将点 p 的 $D^k(p)$ 按从大到小顺序排列，将前n个 $D^k(p)$ 对应的点插入到离群数据点集C中。

(5) 返回C。

3 实验结果

为了说明本文算法的有效性，针对某冶金企业的电力负荷时序数据库来进行分析。以电力负荷的功率因子为指标，来进行离群数据挖掘。图1是某月的时序曲线。

由仿真得到的电力负荷离群数据曲线中含有零值或负值功率的离群数据点(图2中最强的两个时间点即第14 d与第19 d的功率点)，这明显偏离正常的功率值。通过查询该企业的电力能量数据库与设备检修数据库，发现出现电力负荷离群数据点时，均有设备故障等情况记录，表明本文的时序数据离群挖掘算法的有效性。

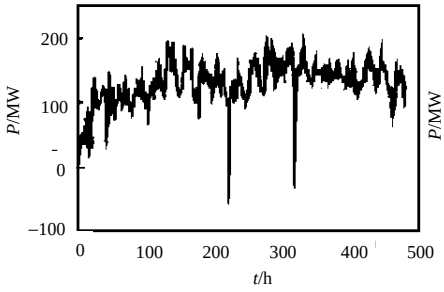

图1 电力负荷功率曲线

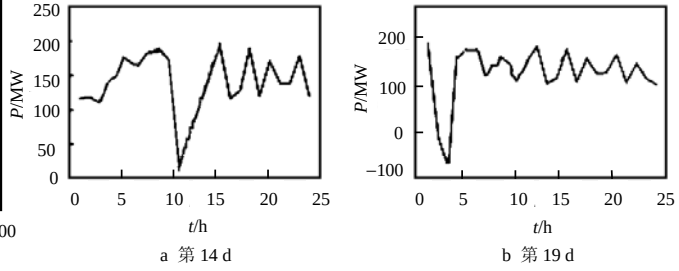

图2 电力负荷离群数据曲线

4 结 论

本文针对时序数据进行离群数据挖掘方法的研究，提出一种基于距离的新的离群时序数据挖掘算法，这种算法对时序序列进行小波变换，并对维数进行约简。采用第k个最近邻的距离 $D^k(p)$ 作为衡量离群数据的度量，本文的方法无须确定数据点的概率分布模型或相异函数，从而克服了以往数据挖掘方法的局限性。

参 考 文 献

[1] Fayyad U, Piatetsky S G, Smyth P. From data mining to knowledge discovery: An overview[A]. Advances in Knowledge Discovery and Data Mining [C]. USA:AAAI/MIT Press, 1996

[2] Edwin K, Roymond N. Algorithms for mining distance-based outliers in large database[A]. Proc of the VLDB Conf[C]. New York: 1998. 392-403

[3] Arning A, Agraal A, Raghavan, P. A linear method for deviation detection in large database[A]. Int Conf on Knowledge Discovery in Databases and Data Mining[C]. Portland, 1996. 169-184

[4] Oppenheim A V, Schafer R W. Digital Signal Processing[M]. New Jersey, USA: Prentice Hall, 1975

[5] Roussopoulos N, Kelley S, Vincent F. Nearest neighbor queries[A]. In: Proceedings of ACM SIGMOD International Conference on Management of Data[C]. San Jose, CA, 1995, 71-79

[6] 郑斌祥, 杜秀华, 席裕庚. 时序数据相似性挖掘算法研究[J]. 信息与控制, 2002, 31(3): 264-271

编 辑 徐安玉