第33卷 第12期 计 算 机 工 程 2007年6月
Vol.33 No.12 Computer Engineering June 2007
·软件技术与数据库· 文章编号：1000—3428(2007)12—0102—03 文献标识码：A 中图分类号：TP391

基于滑动窗口的多变量时间序列异常数据的挖掘

翁小清[1,2]，沈钧毅[1]
(1. 西安交通大学计算机软件与理论研究所，西安 710049；2. 河北经贸大学计算机中心，石家庄 050061)

摘 要：与其它多变量时间序列(MTS)子序列显著不同的子序列，称为异常子序列(含异常数据)。该文提出了一种基于滑动窗口的MTS异常子序列的挖掘算法，使用扩展的Frobenius范数来计算两个MTS子序列之间相似性，使用两阶段顺序查询来进行K-近邻查找，将不可能成为候选异常子序列的MTS子序列剪去，对上海证券交易所股票交易情况MTS数据集进行了异常子序列(含异常数据)挖掘，结果表明了算法的有效性。

关键词：多变量时间序列；滑动窗口；局部稀疏系数；扩展的Frobenius范数；异常数据挖掘

Outlier Mining for Multivariate Time Series Based on Sliding Window

WENG Xiaoqing[1,2], SHEN Junyi[1]
(1. Institute of Computer Software, Xi'an Jiaotong University, Xi'an 710049;
2. Computer Center, Hebei University of Economics and Trade, Shijiazhuang 050061)

【Abstract】 Multivariate time series (MTS) subsequences, which differ significantly from the remaining MTS subsequences, are referred to as outlier subsequences. Tthe mining method for MTS outlier subsequences based on sliding window is proposed. An extended Frobenius norm is used to compare the similarity between MTS subsequences, K-NN searches are performed by using two-phase sequential scan, and MTS subsequences which are not possible outlier candidates are pruned which reduce the number of computations and comparisons. The MTS datasets of stock market is used for outlier mining, the results show the effectiveness of the algorithm.

【Key words】 Multivariate time series; Sliding window; Local sparsity coefficient; Extended frobenius norm; Outlier mining

多变量时间序列(MTS)在各个领域中是非常普遍的，如在金融领域，上市公司的股票交易情况可以用6个变量的MTS描述；在多媒体领域，CyberGloves作为人与计算机的接口[1]，有22个传感器，可以用22个变量(传感器)的MTS来描述。MTS通常用$m\times n$矩阵来表示，其中，m是观测值的个数，n是变量的个数。

如果一个观测值偏离其它的观测值太远，以至于怀疑该观测值是由不同的机制产生的，这样的观测值称为异常数据(或异常点)[2]。

异常点(outlier)挖掘(或检测)已经越来越受到人们的重视，异常点挖掘方法大体上可以分为以下几类：基于分布的异常点检测方法，基于密度的异常点检测方法[4]，基于距离的异常点检测方法以及基于偏离的异常点检测方法[5]等。

针对单变量时间序列，文献[6]提出了一种基于离群指数的离群数据(异常数据)的挖掘方法；文献[7]提出了一种基于小波变换的异常数据挖掘方法；然而针对多变量时间序列，如何从中挖掘出异常数据(或异常点)，目前这方面的研究成果很少。

1 问题的提出及相关定义

1.1 问题的提出

本文要解决的问题是在一个给定的多变量时间序列(MTS)中，找出含有异常数据的MTS子序列，具体描述如下：

给定一个长度为m的多变量时间序列$x(t)=(x_1(t),x_2(t),...,x_n(t))$，$1\le t\le m$，以及长度为$l$的滑动窗口，$l<<m$，将滑动窗口放在MTS的起始位置，此时滑动窗口对应MTS上长度为l的一段子序列，然后时间窗口向后移动，再以序列的第2个点为起始点，形成另一个长度为l的子序列；依此类推，总共可以得到$m-l+1$个长度为l的MTS子序列，这些MTS子序列用

$$s=(s_1,s_2,...,s_{m-l+1})$$

来表示。其中，$s_i=(x_i,x_{i+1},...,x_{i+l-1})$。要解决的问题是找出含有异常数据的MTS子序列。如果将每一个MTS子序列看成是一个样本，需要解决的问题就转化为在$m-l+1$个样本中找出异常样本。

1.2 相关定义

定义1 扩展的Frobenius范数(extended frobenius norm, eros)距离[3]

设A和B是2个MTS子序列，分别为$m_A\times n$和$m_B\times n$，对它们的协方差矩阵进行奇异值分解(SVD)得到的右特征向量矩阵，分别记为$\boldsymbol{V_A}=[a_1,a_2,\ldots,a_n]$，$\boldsymbol{V_B}=[b_1,b_2,\ldots,b_n]$，其中$a_i$和$b_i$为长度为$n$的正交列向量，$A$与$B$之间的Eros距离定义为

$$D_{Eros}(A,B,w)=\sqrt{2-2\sum_{i=1}^{n}w_i\,|<a_i,b_i>|}$$
$$=\sqrt{2-2\sum_{i=1}^{n}w_i\,|\sum_{i=1}^{n}a_{ij}\times b_{ij}|} \quad (1)$$

基金项目：国家自然科学基金资助项目(60173058)
作者简介：翁小清(1965－)，男，博士生、副教授，主研方向：数据挖掘；沈钧毅，教授、博导
收稿日期：2006-09-05 **E-mail：**xqweng@mail.xjtu.edu.cn

其中，$<\boldsymbol{a}_i,\boldsymbol{b}_i>$是列向量$\boldsymbol{a}_i$与$\boldsymbol{b}_i$的内积，$\boldsymbol{w}$是权重向量，$\sum_{i=1}^{n} w_i = 1$。

定义 2 局部稀疏比率(local sparsity ratio)[4, 6]

样本p的局部稀疏比率$\mathrm{lsr_k(p)}$定义为p与其K-近邻内的样本的平均距离的倒数。

$$\mathrm{lsr_k(p)} = \frac{|N_k(p)|}{\sum_{o\in N_k(p)} distofN_k(p)} \tag{2}$$

其中，$|N_k(p)|$为样本p的K-近邻内样本的个数；$\mathrm{distofN_k(p)}$是样本p的K-近邻内的样本与p的实际距离。

局部稀疏比率反映了样本 p 的周围样本的分布密度，具有较小局部稀疏比率的样本成为异常样本的可能性比较大，即异常样本的局部稀疏比率的比较小，反之亦然。

定义 3 剪切因子Pf (pruning factor)定义为所有样本与其K-近邻内样本的平均距离的倒数[4]。

$$Pf = \frac{\sum |N_k(p)|}{\sum\sum_{o\in N_k(p)} distofN_k(p)} \tag{3}$$

如果一个样本是异常样本，则它的稀疏比率应该小于剪切因子 Pf，将所有稀疏比率小于剪切因子 Pf 的样本作为侯选的异常样本。

定义 4 样本p的局部稀疏系数$LSC_k(p)$ (local sparsity coefficient)定义为[4]

$$LSC_k(p) = \frac{\sum_{o\in N_k(p)} \frac{lsr_k(o)}{lsr_k(p)}}{|N_k(p)|} \tag{4}$$

如果样本p的局部稀疏系数$LSC_k(p)$较大，就意味着样本p的邻域中所含样本的个数比较少，不拥挤，说明样本p是异常样本的可能性比较大。

2 算法

MTS 异常子序列挖掘算法主要包括 3 个方面：(1)对每个 MTS 子序列的协方差矩阵进行奇异值分解，求出其特征值、右特征向量矩阵，并计算权重 w；(2)对每个 MTS 子序列找出其 K-近邻样本；(3)计算侯选异常子序列的局部稀疏系数，然后进行排序，具有最大局部稀疏系数的 MTS 子序列就是异常子序列，它含有异常数据。

算法 MTS 异常子序列挖掘算法

输入 长度为 m 的 MTS， 滑动窗口的长度为 l, 以及需要输出的 MTS 异常子序列的个数 Num。

输出 含有异常数据的 MTS 子序列。

(1)将 MTS 划分成 $m-l+1$个长度为 l 的 MTS 子序列。

(2)对每个 MTS 子序列 p，计算其协方差矩阵。

(3)对每个 MTS 子序列 p 的协方差矩阵进行奇异值分解，求其特征值、右特征向量矩阵。

(4)计算权重 w。

(5)对每个 MTS 子序列 p，采用两阶段顺序查找方法，获得其 K-近邻子序列，计算并保存 K-近邻内的各个子序列与 p 之间的 D_{Eros} 距离。

(6)计算每个MTS子序列p的局部稀疏比率$\mathrm{lsr_k(p)}$

(7)计算剪切因子 Pf。

(8)获得侯选 MTS 异常子序列。

(9)计算侯选 MTS 异常子序列的局部稀疏系数 LSC。

(10)对局部稀疏系数 LSC 排序，局部稀疏系数最大的前 Num 个 MTS 子序列进入异常子序列集合，返回异常子序列集合。

在算法的第(4)步中，所有MTS子序列的协方差矩阵的特征值向量的平均值向量(或最大值或最小值)，经过单位化以后，都可以作为权重向量w[3]。

在算法的第(5)步中，对每个 MTS 子序列 p，采用文献[3]提出的两阶段顺序查找方法，获得其 K-近邻子序列。

算法分析：花费时间最多的是第(3)步和第(5)步，对每个MTS子序列p的协方差矩阵进行奇异值分解，由于对 $n\times n$ 矩阵进行奇异值分解的时间复杂度为 $O(n^3)$[3]，第(3)步的时间复杂度是$\mathrm{O}(m\times n^3)$，其中，n是多变量时间序列中变量的个数，m是MTS的长度，一般情况下，$n<<m$；在算法的第(5)步中，对每个MTS子序列p，采用两阶段顺序查找方法，获得其K-近邻子序列的时间复杂度是$\mathrm{O}(m^2)$[3]，所以整个算法的时间复杂度为$\mathrm{O}(m\times n^3+m^2)$。

3 实验

3.1 数据集

数据集为上海证券交易所股票交易情况 MTS 数据集，该数据集是由 6 个变量组成的多变量时间序列组成，这 6 个变量分别为开盘指数、最高指数、收盘指数、最低指数、成交量、成交金额；时间范围是 2003 年 1 月 2 日~2004 年 12 月 31 日，共 484 个交易日；用 $x(t)=(x_1(t),x_2(t),...,x_6(t))$，$1\le t\le 484$，表示该 MTS，其中，$x_1(t)$表示开盘指数；$x_2(t)$表示最高指数；$x_3(t)$表示收盘指数；$x_4(t)$表示最低指数；$x_5(t)$表示成交量(单位为手)；$x_6(t)$表示成交金额(单位为万元)。表 1 给出了该数据集合的部分数据。图 1 给出了变量 $x_3(t)$收盘指数的示意图。

表 1 上海证券交易所一周的交易情况

日 期	开盘	最高	收盘	最低	交易量/手	交易金额/万元
2003-1-6	1 319	1 334	1 334	1 311	58 505	397 456
2003-1-7	1 335	1 346	1 332	1 326	65 470	441 225
2003-1-8	1 331	1 373	1 373	1 330	84 773	558 174
2003-1-9	1 374	1 402	1 397	1 365	161 857	1 121 300
2003-1-10	1 398	1 411	1 384	1 384	146 391	1 055 797

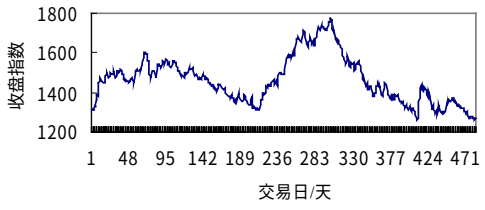

图 1 2003 年 1 月 2 日~2004 年 12 月 31 日收盘指数

3.2 实验结果

用Matlab 6.1 编写了所有的程序，并在清华同方笔记本(CPU 1.5GHz, 内存 112MB，硬盘 40GB，Windows XP操作系统)上实现。将本文算法应用于股票交易情况MTS数据集，将滑动窗口的长度 l 固定为 20，可以得到 $m-l+1$=484-20+1=465 个MTS子序列，即$s=(s_1,s_2,...,s_{465})$；实验结果见表 2，在表 2 中列出了不同的K-近邻(K的取值分别为 11、13、15)时，局部稀疏系数最大的前 4 个MTS时间子序列，当K的取值发生变化时，子序列s_{286}(时间范围是 2003 年 3 月 16 日~2004 年 4 月 12 日)，子序列s_{287}(时间范围是 2004

年 3 月 17 日~2004 年 4 月 13 日)，子序列s_9(时间范围是 2003 年 1 月 14 日~2003 年 2 月 19 日)，这 3 个子序列的局部稀疏系数，都排在前 3 位。在这个数据集中，可以认为这 3 个子序列含有异常数据，是MTS异常子序列。上证收盘指数如图 1 所示，在 2004 年 4 月 8 日，收盘指数达到了最大值 1 771.22，子序列s_{286}包含这个最大值(异常数据)。子序列s_{286}收盘指数 $x_3(t)$ 如图 2 所示。

表 2 在股票交易情况数据集上的实验结果

K-近邻	MTS 子序列编号	局部稀疏系数
11-近邻	s_{286}	2.534 1
	s_{287}	2.407 9
	s_9	2.142 1
	s_{447}	1.939 7
13-近邻	s_{286}	2.539 1
	s_{287}	2.417 9
	s_9	2.064 1
	s_{447}	1.841 2
15-近邻	s_{286}	2.379 3
	s_{287}	2.268 1
	s_9	1.955 0
	s_{69}	1.820 8

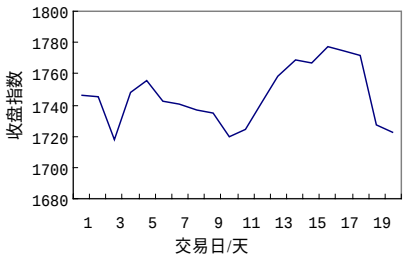

图 2 2004 年 3 月 16 日~2004 年 4 月 12 日收盘指数

当 K 取值为 11 时，在图 3 中按时间顺序给出了所有 465 个 MTS 子序列的局部稀疏系数；从图 3 可以看出绝大多数 MTS 时间子序列的局部稀疏系数在 1.5 以下，只有 3 个 MTS 时间子序列的局部稀疏系数超过了 2，可以认为这 3 个 MTS 子序列是异常子序列，它们含有异常数据。

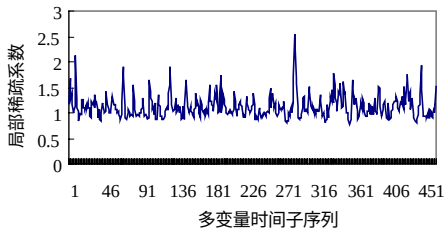

图 3 多变量时间子序列的局部稀疏系数

3.3 算法的性能

将 *K*-近邻中 *K* 值固定为 10，让滑动窗口长度 *l* 从 10 增加到 20，执行算法所花费的时间见图 4，从图 4 可以看出，当 *l* 的取值增加时，执行算法所花费的时间在一定的区域内波动，说明 *l* 的取值对算法执行的时间影响不大。

将滑动窗口长度 *l* 固定为 20，让 *K*-近邻中 *K* 值从 5 增加到 15，执行算法所花费的时间见图 5，从图 5 可以看出，*K* 的取值对执行算法所花费的时间影响不大。

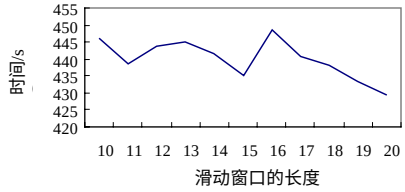

图 4 滑动窗口的长度与执行算法所花费的时间

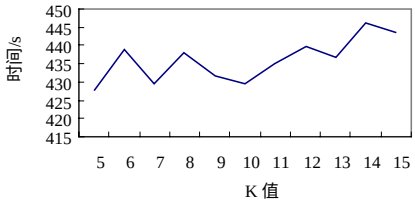

图 5 K 值与执行算法所花费的时间

将滑动窗口长度 *l* 固定为 20，K-近邻中的 K 值固定为 10，让 MTS 的长度 *m* 从 50 时增加到 450，执行算法所花费的时间见图 6，从图 6 可以看出，当 MTS 的长度 *m* 增加时，算法执行时间与 MTS 的长度 *m* 呈近似二次方关系 O(m^2)。

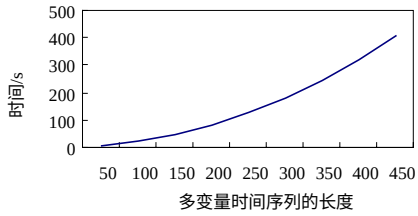

图 6 多变量时间序列的长度与执行算法所花费的时间

4 结语

针对 MTS，本文提出了一种基于滑动窗口的 MTS 异常子序列的挖掘算法，实验表明，该算法可以挖掘出含有异常数据的 MTS 子序列。

研究更有效的 MTS 子序列的索引查找的方法，是提高挖掘算法性能的关键；另外，在线实时地挖掘含有异常数据的 MTS 子序列，应用前景较好，也值得今后进一步的研究。

参考文献

1 Li C, Pradhan G, Zheng S Q, et al. Indexing of Variable Length Multi-attribute Motion Data[C]//Proc. of MMDB'04, 2004: 75-84.

2 Hawkins D. Identification of Outliers[M]. London: Chapman and Hall, 1980.

3 Yang K, Shahabi C. A PCA-based Similarity Measure for Multivariate Time Series[C]//Proc. of MMDB'04. 2004: 65-74.

4 Agyemang M, Ezeife C I. LSC-Mine: Algorithm for Mining Local Outliers[C]//Proc. of the 15th Information Resources Management Association International Conference, New Orleans, Louisiana. 2004.

5 Han J W, Kamber M. Data Mining Concepts and Technique[M]. Beijing: China Machine Press, 2001.

6 郑斌祥, 席裕庚, 杜秀华. 基于离群指数的时序数据离群挖掘[J]. 自动化学报, 2004, 30(1): 70-77.

7 文 琪, 彭 宏. 小波变换的离群时序数据挖掘分析[J]. 电子科技大学学报, 2005, 34(4): 556-558.