文章编号: 1000-5862(2012)03-0280-04

一种基于 PCA 的时间序列异常检测方法

郭小芳[1], 李　锋[2], 宋晓宁[1]

(1. 江苏科技大学计算机科学与工程学院, 江苏 镇江 212003; 2. 江苏科技大学电子信息学院, 江苏 镇江 212003)

摘要: 在 *k*-近邻局部异常检测算法的基础上, 采用基于主成分分析的多元时间序列的降维方法, 依据累积贡献率选择主成分序列, 给出了一种效率较高的多元时间序列异常检测算法. 实验结果表明: 该算法可以较好地提高多元时间序列异常检测的效率.

关键词: 多元时间序列; 主成分分析; *k*-近邻; 异常检测
中图分类号: TP 391　　文献标志码: A

0　引言

异常检测(outlier detection)也称为异常挖掘、孤立点分析, 其目标是在数据集中发现不正常的数据点[1]. 目前异常检测方法有基于距离的异常点检测方法、基于密度的异常检测方法、基于模型的异常检测方法[2]. 在现有的检测方法中, 基于距离的异常点检测方法效率较高, 但对数据分布不同的数据集效果较差; 基于异常密度的方法虽然检测精度好, 但复杂度较高, 响应速度较慢; 基于模型的方法具有理论上的严密性, 但对于数据分布的识别和模型参数的估计存在一定的困难[3].

多元时间序列同时具有数据量大、维度高、变量相关性高、大量噪声干扰等特点, 使异常检测更加困难[4]. 本文在*k*-近邻局部异常检测算法的基础上, 结合基于主成分分析的多元时间序列的降维方法, 按照累积贡献率选择主成分序列, 利用局部异常检测方法对多元时间序列进行异常检测. 最后以股票数据异常检测实验验证了算法的有效性和合理性.

1　相关概念

在统计学中, 异常是指那些不服从序列分布、与其他数据点距离较远的数据点; 在回归模型中, 异常是指与给定模型偏离很大的数据点. 时间序列中的异常点通常是指在偏离数据集中大部分数据的数据, 这些偏离数据可能是由完全不同的机制产生的, 而非随机偏差[5].

按照异常的表现形式不同, 时间序列的异常可以分为序列异常、点异常及模式异常3种情况: (1)序列异常, 在时间序列数据集中与其它时间序列显著不同的、来源于不同产生机制的时间序列; (2)点异常, 在一条时间序列上与其它序列点存在显著差异的、具有异常特征的序列点; (3)模式异常, 在一条时间序列上与其它模式存在显著差异的、具有异常行为的模式.

时间序列 X 的模式可以表示为

$$X =< (l_1,k_1),(l_2,k_2),\cdots,(l_c,k_c) > . \tag{1}$$

模式 $p_1=(l_1,k_1)$ 和 $p_2=(l_2,k_2)$ 之间的距离[6]

$$d(p_1,p_2)=\frac{|l_1-l_2|}{\min\{l_1,l_2\}}+\frac{|k_1-k_2|}{\min\{k_1,k_2\}}, \tag{2}$$

其中, 二元组(l_i, k_i)中的 $l_i,k_i(i=1,2)$ 分别表示模式的长度和斜率, 即模式变化的长短和变化趋势.

2　多元时间序列异常检测算法

多元时间序列异常检测流程图如图1所示, 即首先利用主成分分析对多元时间序列进行降维处理, 得到多元时间序列的主成分序列, 在此基础上找出每个 MTS 的 *k*-近邻序列, 计算各 MTS 序列的异常因子 $LOF(q)$, 并对异常因子进行排序, 输出 λ 最异常的 MTS 序列.

收稿日期: 2011-11-28
基金项目: 国家自然科学基金(51008143)资助项目.
作者简介: 郭小芳(1974-), 女, 陕西商洛人, 讲师, 硕士, 主要从事数据挖掘方面的研究.

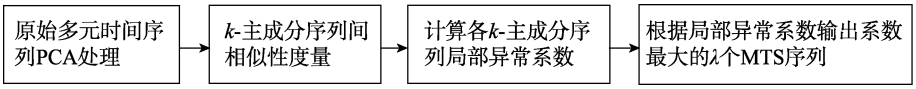

图 1　多元时间序列异常检测流程

2.1　PCA 主成分分析

主成分分析技术PCA可以有效地找出数据中最“主要”的元素和结构，对原有数据进行简化，并揭示隐藏在复杂数据背后的简单关系[6]，其基本思想是对原始变量的适当线性组合，形成少数几个原始变量主要信息的新变量，并采用新变量来分析问题和解决问题，其原理如图 2 所示. 原始变量 X_1 和 X_2 相关性很强(点分布在倾斜的椭圆内)，在适当的坐标变换下(如逆时针旋转一个角度 θ)，则新旧坐标之间关系

$$\begin{cases} Z_1 = \cos\theta X_1 + \sin\theta X_2, \\ Z_2 = \cos\theta X_1 + \sin\theta X_2. \end{cases} \tag{3}$$

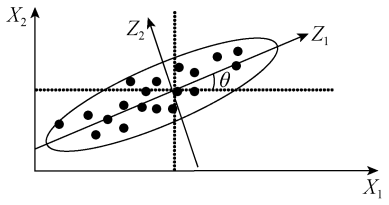

图 2　主元分析的几何意义

从图2可以看出 n 个点的波动主要在 Z_1 方向，Z_2 方向上的波动可以忽略，这样可以将2维问题降为1维处理，达到降维的目的[7-8].

对于代表MTS项的2个矩阵***A***和***B***(要求列数相同)，首先先通过奇异值分解获得每个矩阵的主成分，然后试探性选择最初的 z 个主成分(如选取代表变化95%的前 z 个主成分)，其相似性矩阵为[12]

$$S_{\mathrm{PCA}}(\boldsymbol{A},\boldsymbol{B}) = \mathrm{trace}(\boldsymbol{L}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{L}^{\mathrm{T}}) = \sum_{i=1}^{z}\sum_{j=1}^{z}\cos^2\theta_{ij}, \tag{4}$$

其中 ***L*** 和 ***M*** 包含矩阵 ***A*** 和 ***B*** 的前 z 个主成分，θ_{ij} 为 ***A*** 的第 i 个主成分与 ***B*** 的第 j 个主成分之间的夹角. S_{PCA} 从0到 z 变化，通过计算2个矩阵的前 z 个主成分的所有组合的余弦平方值来测量其相似性[9-10].

2.2　*k*-近邻局部异常检测

(1)对于给定一个正整数 k 和一个数据点集合 D，在 D 中 p 点的 k 近邻距离 $k-dist(p)$ 满足：(i) 至少有 k 个点 $o\in D\setminus\{p\}, d(p,o)\leqslant k-dist(p)$, (ii) 最多有 $k-1$ 个点 $o\in D\setminus\{p\}, d(p,o)<k-dist(p)$，那么称 $dist(p,o)$ 是 p 的 k^{th} 距离. 在图3中当 $k=3$ 时，$k-dist(p)=d(p,o)$，其中 $d(p,o)$ 表示 p 点到 o 点的距离.

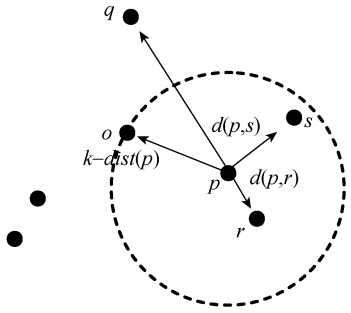

图 3　***k*=3** 时的 ***k–dist*(*p*)=*d*(*p*,*o*)**

(2) 若q点到 p 点的k-近邻距离满足

$$r-dist_k(q,p) = \max(d(q,p), k-dist(p)), \tag{5}$$

则称其为 p 点的k-近邻可达距离 $r-dist_k(q,p)$.

在图2中，因为 $d(q,p)>k-dist(p)$，所以 q 点到 p 点的 $r-dist_k(q,p)=d(q,p)$；而 r 点到 p 点的 $d(r,p)<k-dist(p)$，因此，$r-dist_k(r,p)=k-dist(p)$.

(3) 点 q 的 k 局部可达密度为

$$lrd(q) = \frac{k}{\sum_{p\in K(q)} r-dist_k(q,p)}, \tag{6}$$

其中 $K(q)$ 表示在数据集 D 中与对象 q 的距离不超过 $k-dist(p)$ 的所有点的集合. $lrd(q)$ 反映了 q 点周围点分布密度. 如果 $lrd(q)$ 较小，说明 q 点成为局部异常点的可能性比较大.

(4) q 点的局部异常系数为

$$LOF(q) = \frac{1}{k}\sum_{p\in K(q)} lrd(q) \Big/ lrd(q), \tag{7}$$

$LOF(q)$ 的大小值反映 q 点在其k领域内所含点稀疏程度，$LOF(q)$ 值越大，该点的所在的局部范围点越稀疏，则该点异常的可能性高.

值得注意的是，这里的对象间距离不是计算MTS对象间距离，而是MTS对象经过主成分分析后的主成分序列间距离. 各主成分所包含的信息占原来变量所包含信息的比重可以通过计算贡献率获得

贡献率大的权值大，权值一般由特征值获得.

2.3 异常检测算法

多元时间序列异常检测算法：输入多元时间序列集MTS，近邻数 k，异常点个数 λ，输出多元时间序列集MTS的 λ 个异常模式，具体过程如下：

(1)[*a*. *zcf*, *w*]←PCA(MTS); //主成分分析给出主成分
//和权向量
(2)*sn*←*length*(MTS); //计算序列长度
(3)for *i*=1: *sn*
(4) *data*(*i*). *point*←*compute*(*a*(*i*). *zcf*, *k*);
(5) *data*(*i*). $dist_k$←$dist_k$(*i*); //计算每个模式子序列
//的*k*近距离
(6) *s*←0, *t*←0;
(7) for *j*=1: *k*
(8) *v*←*data*(*i*). *point*(*j*, 1);
(9) *s*←*s*+max(*data*(*i*). *point*(*j*, 2), *data*(*v*). $dist_k$);
(10) *data*(*i*). *lrd*←*k*/*s*; //计算局部可达密度
(11) *t*←*t*+*data*(*v*). *lrd*;
(12) end;
(13) *data*(*i*). *lof*←*t*/*k*/*data*(*i*). *lrd*; //局部异常系数
(14)end.

算法的复杂度分析：对于由*n*个 $p\times p$ 矩阵构成的MTS序列(一般情况下 $p\ll n$)，奇异值分解的时间复杂度为 $O(n\times p^3)$，查找*k*-近邻序列采用分层顺序扫方法，时间复杂度为 $O(n\times p^2)$，采用(6)、(7)式计算各MTS序列的异常因子并对其进行排序，时间复杂度为 $O(k\times n)$. 考虑 $p\ll n$ 的情况，该算法的总体复杂度为 $O(n\times p^3)$.

3 实验结果与分析

3.1 实验数据集

在某股市的股票数据集中选取了300家上市公司的交易情况组成一个MTS数据集，MTS的长度为300个交易日的观测值，对各股票的起始收盘价、结束收盘价、股票波动的最高价、最低价进行长时间跨度趋势分析. 选定参照股并对其他股票数据进行标准化，以标准化后的最高与最低指数的差值为基准，定义一个差值上下波动区间，如果一段时间某只股票最高和最低价出现的先后顺序与参照股的相异或两者的差值超出上下波动区间，则可认为这只股票属于异常股票. MTS经过主成分分析后，得到特征值矩阵，按照累计方差贡献率需达到95%以上的要求，选定主成分个数*k*=2，然后对每个维度主成分序列分别进行BU分段[11]，然后根据多元时间序列相似性度量公式计算任意2只股票的相似度(相似距离)，根据相似度通过异常检测算法计算每只股票的局部异常系数. 所有实验过程在Matlab平台上完成.

3.2 实验结果

取 $k=8$，主成分取 $z=2$，前 2 个主成分的累积贡献率达 99%以上[12]，通过该算法计算每个 MTS 序列的异常因子见图 4. 根据本实验的异常定义，得到的预期异常股票序号分别是 24、62、123、174、212、279.异常检测算法的计算结果是每只股票的局部异常系数，局部异常系数越大，则该股票是异常的可能性也就越大.

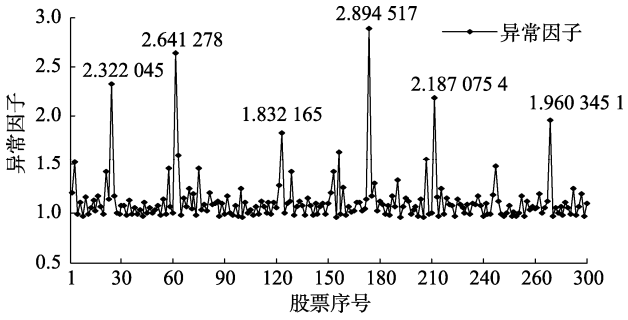

图 4 MTS 异常检测结果

MTS 的长度和取值都有可能影响到算法效率，从图 5 可以看出，随着 MTS 数据序列个数的增加，算法消耗时间是逐步增加的；图6给出了近邻值 k 对算法执行速度的影响，随着 k 的增加，算法消耗时间也是逐步增加的.

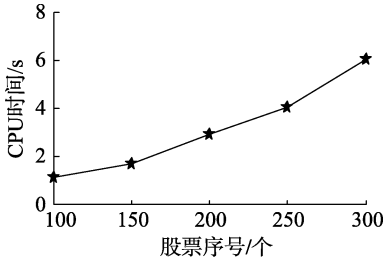

图 5 序列个数对算法的影响

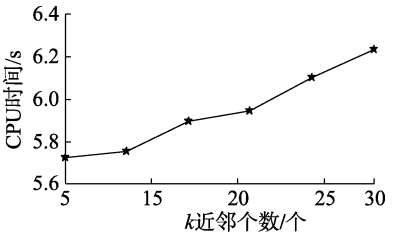

图 6 *k* 近邻个数对算法的影响

基于k-近邻的多元时间序列的局部异常检测算法不是对原始多元时间序列的直接处理，而是在k等于主成分序列基础上进行的异常检测．这种方法去除了多变量相关性对异常检测的影响，减少了参与异常检测的变量数，提高了异常检测精度．

4　结论

时间序列由于数据量大、维数高，如何检测时间序列中的异常数据是当前时间序列挖掘中一项具有重要意义的研究课题，本文在 k-近邻局部异常检测算法的基础上，结合基于主成分分析的多元时间序列的降维方法，给出了一种高效率的多元时间序列异常检测算法．实验通过股票数据验证了算法的有效性和合理性．但在线实时地进行多元时间序列异常检测，还是是今后进一步研究的内容．

5　参考文献

[1] Rahmani B, Markazi A H D, Mozayani N. Real time prediction of time delays in a networked control system [EB/OL]. [2011-11-15]. http: //ieeexplore. ieee. org/xpl/login. jsp?tp=&arnumber=4537416&url=http%3A%2F%2Fieeexplore. ieee. org%2Fiel5%2F4531369%2F4537177%2F04537416.pdf%3Farnumber%3D4537416.

[2] Sadeghzadeh N, Afshar A, Menhaj M B. An MLP neural network for time delay prediction in networked control systems [EB/OL]. [2011-11-15]. http: //ieeexplore. ieee. org/xpl/login. jsp?tp=&arnumber=4598345&url=http%3A%2F%2Fieeexplore. ieee. org%2Fxpls%2Fabs_all. jsp%3Farnumber%3D4598345.

[3] Liu Jianggang, Liu Biyu, Zhang Ruifang, et al. The new variable-period sampling scheme for networked control systems with random time delay based on BP neural network prediction [EB/OL]. [2011-11-15] http: //ieeexplore. ieee. org/xpl/login. jsp?tp=&arnumber=4347048&url=http%3A%2F%2Fieeexplore. ieee. org%2Fxpls%2Fabs_all. jsp%3Farnumber%3D4347048.

[4] Keogh E, Lin J. Finding unusual medical time-series subsequences: algorithms and applications [EB/OL]. [2011-11-15]. http: //www. cse. cuhk. edu. hk/~adafu/Pub/titb05.pdf.

[5] 林果园，郭山清．基于动态行为和特征模式的异常检测模型 [J]．计算机学报, 2006, 29(9): 1553-1559.

[6] 翁小清，沈钧毅．基于滑动窗口的多变量时间序列异常数据的挖掘 [J]．计算机工程, 2007, 33(12): 102-104.

[7] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reductionfor fast similarity search in large time series databases [J]. Journal of Knowledge and Information Systems, 2001, 3(3): 263-286.

[8] 肖辉．时间序列的相似性查询与异常检测 [D]．上海：复旦大学, 2005.

[9] 曲吉林，时间序列挖掘中索引与查询技术的研究 [D]．天津：天津大学, 2006.

[10] 陆声链，孤立点挖掘及其内涵知识发现的研究与应用 [D]．南宁：广西师范大学, 2005.

[11] 罗超，郭晨，梁家荣．确定性树突状细胞算法的异常检测系统 [J]．江西师范大学学报：自然科学版, 2011, 35(2): 170-172.

[12] 邱舟强，滕少华，李振坤，等．数据挖掘技术在网络入侵检测中的应用 [J]．江西师范大学学报：自然科学版, 2006, 30(1): 157-159.

The Outlier Detection Approach for Multivariate Time Series Based on PCA Analysis

GUO Xiao-fang[1], LI Feng[2], SONG Xiao-ning[1]

(1. School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang Jiangsu 212003, China;
2. School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang Jiangsu 212003, China)

Abstract: By means of cumulative contribution rate, dimension reduction method based on principal component analysis and principal components of multivariate time series was selected, an efficient multivariate time series outlier detection algorithm was provided based on the k-nearest neighbor local outlier detection algorithm was provide here,. the experimental results show that the algorithm can morely improve the efficiency of multivariate time series outlier detection.

Key words: multivariate time series; principal component analysis; k-nearest neighbor; outlier detection

(责任编辑：冉小晓)