

基于加权 Euclid 范数的 MTS 异常检测

郭小芳¹ 李 锋² 宋晓宁¹ 刘庆华¹

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

摘 要 为了提高时间序列的异常检测算法的精度,根据主成分的累积贡献率选择序列及其主成分,在 k -近邻局部离群点检测算法中采用加权 Euclid 范数距离作为 k -近邻距离,从而实现对多变量时间序列的异常检测。为了验证算法的有效性,对测试数据进行了异常检测。实验结果表明,算法的精度和查全率比传统方法具有更大的优越性。

关键词 多变量时间序列,扩展 Frobenius 范数, k -近邻,异常检测

中图法分类号 TP391 文献标识码 A

Outlier Detection of Multivariate Time Series Based on Weighted Euclid Norm

GUO Xiao-fang¹ LI Feng² SONG Xiao-ning¹ LIU Qing-hua¹

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)¹

(School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China)²

Abstract In order to improve the precision of anomaly detection algorithm for multivariate time series, based on the cumulative contribution rate of principal components, sequence and its principal components are selected, and in the k nearest neighbor local outlier detection algorithm, the weighted Euclid norm distance is used as the k nearest neighbor distance, so as to realize the anomaly detection of multivariate time series. In order to verify the effectiveness of the algorithm, anomaly detection was carried out on test data. The experimental results show that the algorithm has more advantages than traditional methods in the precision and recall.

Keywords Multivariate time series, Extended Frobenius norm(Eros), K -nearest neighbor, Outlier detection

异常检测(Outlier Detection)也称为离群点检测、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]。与之相关的研究包括:事件检测(Event Detection)^[2]、变化点检测(Change Point Detection)^[3]、异常行为检测(Aberrant Behavior Detection)^[4]等。目前异常检测方法主要有:基于距离的异常点检测方法^[5]、基于密度的异常检测方法^[6]、基于模型的异常检测方法^[7]。基于距离离群点检测方法由于各种数据分布的影响,效果不是很好;基于密度的异常检测方法是高精度检测,但当大数据集时,计算量大,复杂度太高,响应速度慢;基于模型的异常检测方法将具有较低概率的数据点作为异常点检出,该方法具有坚实的数学理论基础,其难点在于数据分布的识别和模型参数的估计。由于多变量时间序列 MTS(Multivariate Time Series)具有数据量大、维度高、变量之间相互依赖、许多高噪声等特点,异常检测是比较困难的。林果园等^[8]将动态行为和全局特征的特点结合起来进行异常检测,算法的时间复杂度较高^[8];张力生^[9]提出的基于时间序列重要点分段检测异常序列,只适用于一维数据异常序列检测。本文根据主成分累积贡献率的大小,通过加权 Euclid 范数进行 k -近邻局部异常检测。实验结果表明,该方法在异常序列检测的有效性和准确性方面具有一定优势。

1 相关定义

1.1 主成分分析法

主成分分析法^[10]可以有效地对数据进行压缩,实现噪声的分离,去除变量之间的相关性。对于 MTS 项 A 和 B (矩阵列数相同),对 A 和 B 的协方差矩阵进行奇异值分解,选择代表主要变化(大于 95%)的前 z 个主成分,构造相似性矩阵:

$$S_{PCA}(A, B) = \text{trace}(LM^T ML^T) = \sum_{i=1}^z \sum_{j=1}^z \cos^2 \theta_{ij} \quad (1)$$

其中, L 为 A 的前 z 个主成分, M 为 B 的前 z 个主成分, L 的第 i 个分量与 M 的第 j 个分量之间的夹角为 θ_{ij} 。

1.2 加权 Euclid 范数距离

为了判断某个主成分模式是否是异常式,首先要解决的问题就是衡量该模式与周围模式的距离度量。 $m \times n$ 阶矩阵 A 的 Frobenius 范数定义为:

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n (a_{ij})^2 \right)^{1/2} \quad (2)$$

对于大小分别为 $m_A \times n$ 及 $m_B \times n$ 的 MTS 项 A 和 B ,对其协方差矩阵 M_A 和 M_B 进行奇异值分解(SVD),得到右特征向量矩阵 $V_A = [a_1, \dots, a_n]$ 和 $V_B = [b_1, \dots, b_n]$,则 A 和 B 的加权 F 范数(Eros) 定义为:

到稿日期:2013-05-16 返修日期:2013-08-30 本文受国家自然科学基金(51008143),江苏省高校自然科学基金项目(10JKB520006)资助。

郭小芳(1974—),女,硕士,讲师,主要研究方向为人工智能、数据挖掘, E-mail: gxfsl@126.com; 李 锋(1970—),男,副教授,主要研究方向为智能计算、可信计算。

$$Eros(A, B, w) = \sum_{i=1}^n w_i |\langle a_i, b_i \rangle| = \sum_{i=1}^n w_i |\cos \theta_i| \quad (3)$$

其中, 加权向量 w 满足 $\sum_{i=1}^n w_i = 1$, θ_i 代表 a_i 和 b_i 的夹角。加权 F 范数 $Eros$ 等于 1 代表最相似, 等于 0 表示完全不相似。由此可以构造 A 和 B 间的相似距离:

$$D_{Eros}(A, B, w) = \sqrt{2 - 2 \sum_{i=1}^n w_i |\langle a_i, b_i \rangle|} \\ = \sqrt{2 - 2 \sum_{i=1}^n w_i |\sum_{j=1}^n a_{ij} \times b_{ij}|} \quad (4)$$

$D_{Eros}(A, B, w)$ 的值越小, 则 A 和 B 越相似。然而多变量时间序列 A 和 B 两者之间 $Eros$ 距离 $D_{Eros}(A, B, w)$ 并不满足三角不等式, 为了提高相似查询的效率, 可通过 D_{Eros} 的上下界距离快速过滤部分不满足相似要求的时间序列^[10]:

$$\sum_{i=1}^n w_i \sum_{j=1}^n |a_{ij} b_{ij}| \geq \sum_{i=1}^n w_i |\sum_{j=1}^n a_{ij} b_{ij}| \geq \sum_{i=1}^n w_i \sum_{j=1}^n a_{ij} b_{ij} \quad (5)$$

由式(4)可知:

$$\sqrt{2 - 2 \sum_{i=1}^n w_i \sum_{j=1}^n a_{ij} b_{ij}} \geq D_{Eros}(A, B, w) \geq \\ \sqrt{2 - 2 \sum_{i=1}^n w_i |\sum_{j=1}^n a_{ij} b_{ij}|} \quad (6)$$

式(5)说明 $D_{Eros}(A, B, w)$ 存在上下界:

$$D_{\min}(A, B, w) \leq D_{Eros}(A, B, w) \leq D_{\max}(A, B, w) \quad (7)$$

其中

$$D_{\max}(A, B, w) = \sqrt{2 - 2 \sum_{i=1}^n w_i \sum_{j=1}^n a_{ij} b_{ij}} \quad (8) \\ D_{\min}(A, B, w) = \sqrt{2 - 2 \sum_{i=1}^n w_i |\sum_{j=1}^n a_{ij} b_{ij}|}$$

1.3 局部异常检测

(1) 设 $p \in D$, p 的 k -近邻距离 $k_dist(p)$ 满足: I) 至少有 k 个对象 $o \in D \setminus \{p\}$, $Dist(p, o) \leq k_dist(p)$; II) 至多有 $k-1$ 个对象, $o \in D \setminus \{p\}$, $Dist(p, o) < k_dist(p)$ 。

(2) 若 q 到 p 的 k -近邻距离^[11] 满足

$$r_dist_k(q, p) = \max\{d(q, p), k_dist(p)\} \quad (9)$$

则 $d(q, p)$ 为 q 到 p 的加权 Euclid 距离, $r_dist_k(q, p)$ 称为 q 的 k -近邻可达距离。

(3) q 的 k 局部可达密度 $lrd_k(q)$

$$lrd_k(q) = 1 / \left(\frac{\sum_{p \in N_k(q)} r_dist_k(q, p)}{|N_k(q)|} \right) \quad (10)$$

其中, $N_k(q)$ 为在数据集 D 中与对象 q 的距离不超过 $k_dist(p)$ 的所有对象的集合。 $lrd_k(q)$ 的值反映 q 点周围点密度分布。如果 $lrd_k(q)$ 较小, 说明 q 成为局部离群点的可能性比较大。

(4) 计算 q 的局部异常系数 $lof_k(q)$

$$lof_k(q) = \frac{\sum_{p \in N_k(q)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(q)|} \quad (11)$$

$lof_k(q)$ 的值接近 1, 数据 q 记录隶属于一个集群, $lof_k(q)$ 大于 1, 远离其数据的聚类, 并且值愈大则反映 q 的 k 邻域内的数据点愈稀疏(所包含的对象较少), 意味着 q 是异常的可能性较大。

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 对象的主成分间距离。

(5) 异常样本

MTS 的 PCA 模式映射到特征空间中的一个对象的集

合, 记录的数据集 DB , k 作为参数, 设 $p \in DB$, 则 DB 中的异常点 p 满足:

$$w_k(p, DB) = \sum_{i=1}^k D_{Eros}(p, nm_i(p), w) \quad (12)$$

式中, $nm_i(p)$ 为 DB 中 p 的第 i 个临近距离。

2 异常检测算法与分析

2.1 算法描述

算法主要包括 3 个方面:

(1) PCA 处理, 其基本步骤: MTS 序列 $\rightarrow$ 协方差矩阵 $\rightarrow$ 特征值 $\rightarrow$ 右特征向量矩阵, 权重 w ;

(2) 计算并保存 k -近邻内的各个子序列与 p 之间的 D_{Eros} 距离;

(3) 对各个子序列按照局部异常系数排序, 局部异常系数较大的 MTS 序列被认为是异常序列。

2.2 实验描述

主成分分析是把序列总方差分解成 m 个时间序列变量的方差之和, 用较少的 z 个主元代替原来的 m 个变量 ($z < m$)。算法参照如下:

算法 1 PCA(X)

输入: 主元数 z , MTS 数据集 X

输出: 主元, 累积贡献率 w

```
1. sn ← length(X);
2. cu ← 0;
3. for i = 1:sn
4.   za ← zsp(X(i)); // MTS 标准处理
5.   [a, zcf, score, ltent, tquare] ← promp(za); // 主元分析
6.   cu ← cu + ltent;
7.   A(i), zcf ← pc(i).data(1:z); // 取前 z 个主元
8. end
9. w ← cu/sn;
10. w ← w/sum(w); // 累计贡献率
11. retrn(a, w);
```

算法中根据给定累积贡献率确定 z 的取值, 如要求累积贡献率达到 98%^[12], 则取 $z=2$ 即可。从而输出相应的 MTS 主元序列。MTS 经过主成分分析后, 取各主成分的贡献率作为权重向量 w 。

算法 2 计算权重向量 w 。

要求: $m \times N$ 矩阵 S , m 是数据集的变量数, N 是数据库中 MTS 项目数。 s_{ij} 是 S 矩阵中第 i 列、第 j 行的矢量本征值, 列矢量 s_i 是第 i 个 MTS 项目的所有本征值。 s_{*i} 表示在 S 中的第 i 行, s_{i*} 表示在 S 中的第 i 列。

```
1. for i = 1:N
2.    $s_i \leftarrow s_i / \sum_{j=1}^m s_{ij}$ ; // 计算每个矢量本征值的比重
3. end
4. for j = 1:m
5.    $w_i \leftarrow f(s_{*i})$ ; // 第  $i$  行的权重向量
6. end
7. for i = 1:m
8.    $w_i \leftarrow w_i / \sum_{j=1}^m w_j$ ; // 计算权重向量的比重
9. end
```

算法 3 MTS 序列的异常检测

输入: MTS 数据集, 近邻数目 k , 异常点个数 λ

输出: λ 个异常模式

```

1. [a, zcf, w] ← PCA(MTS); //主成分分析给出主成分和权重向量
2. sn ← lth(MTS); //序列长度计算
3. For i = 1; sn
4.   data(i).point ← compute(a(i), PCA, k);
5.   data(i).DEros ← DEros(i); //计算每个模式子序列的 k_临近距离
6.   s ← 0, t ← 0;
7.   for j = 1; k
8.     v ← data(i).point(j, 1);
9.     data(i).lrdk ← k / (s + max(data(i).point(j, 2), data(v).DEros));
        //局部可达密度
10.    t ← t + data(v).lrdk;
11.  end
12.  data(i).lofk ← t / k / data(i).lrdk; //局部异常系数
13. end

```

通常 $p \ll n$, n 个 $p \times p$ 矩阵的奇异值分解的时间复杂度为 $O(n \times p^3)$, 分层顺序扫描查找 k _近邻序列的时间复杂度为 $O(n \times p^2)$, 计算各 MTS 的异常因子及因子排序的复杂度为 $O(k \times n)$ 。整体复杂度为 $O(n \times p^3)$ 。

3 实验结果与分析

3.1 实验过程

1. 主成分 z 的取值

实验在 Intel 双核, 1.6GHz, WindowsXP 操作系统主机上展开, 在 Matlab7.9a 平台上测试算法的实验过程。在实验中使用 UCI 数据集、数据集 a(乳腺癌数据包含 300 个检测实例, 每组记录 34 维数)和数据集 b(听力受损者的治疗方案数据, 包含 366 个检测实例, 每组记录 69 维)两组数据进行实验, 该数据集 MTS 经过主成分分析后, 得到特征值矩阵。为考察主成分数量 z 对结果的影响, 表 1 给出了主成分 $z=1 \sim 5$ 两组数据集的贡献率(样本序列数为 300)。

表 1 主成分算法中主元 z 的贡献率

主元 z	1	2	3	4	5
z 的贡献率 (数据集 a)	0.07332	0.01456	0.003326	0.00174	0.000532
z 的贡献率 (数据集 b)	0.06785	0.01396	0.003175	0.00152	0.000489

在实验中当主成分 $z=2$ 时, 两组实验的累积贡献率就分别高达 98.5% 和 98.6%, 这说明主成分取前 2 个就可以很好地代表原始数据。对于不同的数据集, 一般在验证算法时取 $z=2$ 即可。然后为每个主成分序列进行 BU 分段, 再根据加权 Euclid 范数距离公式计算任意两组数据的相似距离和局部离群系数。

2. 查全率和查准率

为了验证异常检测时的查全率和查准率, 进行了 Eros、ED、DTW 3 种相似距离方法的比较。针对数据集 a, 从图 1 可以看到, 当查全率接近于 1 时, 图 1 中 Eros 的查准率超过 ED 25%, 超过 DTW 60%; 对于数据集 b, 从图 2 看当查全率等于 0.2 时, Eros 查准率至少优于 ED 和 DTW 6.5%, 当查全率(recall)等于 1 时, Eros 超过 ED 33%, 超过 DTW 64%。

采用 Eros 相似距离的查全率、查准率高于其他两种方法, 这是因为 Eros 距离采用了协方差矩阵, 而其他相似度量未考虑到变量之间的相关性。可见 D_{Eros} 距离对不同属性的数据集都有较好的效果, 因此验证了所提出的相似度量测量方法的有效性。

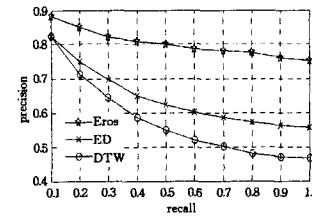


图 1 查全率查准率

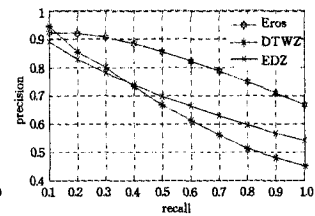


图 2 查全率查准率

3.2 实验结果

取 $z=2$ 时, 主成分累积贡献率达 98% 以上。取 $k=5$ 计算每个 MTS 序列异常因子, 如表 2、表 3 所列。计算结果数据集的局部离群系数, 局部离群系数越大, 则序列异常的可能性越大。从图 3、图 4 可以看出绝大多数 MTS 时间子序列的局部异常系数在 2.0 以下, 局部异常系数超过了 2 的有 4 个 MTS 子序列, 说明这 4 个 MTS 的子序列为异常子序列, 它们含有异常数据。

表 2 data(a) 的局部异常系数

异常序号	24	58	62	81	123
局部异常系数	2.32204	1.48963	2.64127	1.41562	1.86534
异常序号	174	203	212	247	279
局部异常系数	2.89451	1.53427	2.187075	1.46352	1.96034

表 3 data(b) 的局部异常系数

异常序号	10	35	45	83	100
局部异常系数	1.427	1.457	1.469	1.429	1.4294
异常序号	123	156	196	260	300
局部异常系数	1.3372	1.4821	1.4872	1.4826	1.484

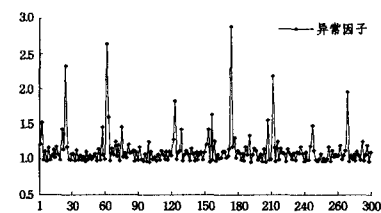


图 3 data(a) MTS 异常检测

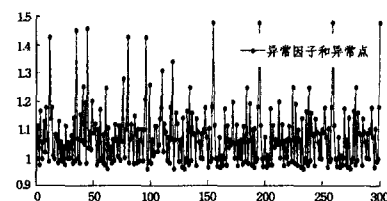


图 4 data(b) MTS 异常检测

本文对异常离群点检测算法的分析的主要不同就是两个多变量时间序列间距离的计算方法采用加权 Euclid 范数距离。本文根据主成分累计贡献率的大小选取 z 个特征向量, 运用加权 Euclid 距离进行离群点查找。将这 z 个特征向量与原始序列进行乘计算, 得到新的序列。新序列则是一个 $m \times z$ 的矩阵(m 为记录的时间点数, z 为新序列变量个数), 相较于于

(下转第 295 页)

的细胞图像。在这个样本提取的算法中,由于利用图像矩作为样本图像的特征,避免了对不同的细胞要采用不同的方法来测量图像特征的问题,因此,本文所介绍的算法可以适用于更多需要采样的木材细胞图像,可以推广到其他细胞图像的处理中。本算法由于只能对二值化后的细胞图像进行处理,因此一定程度上对细胞图像有错误处理的情况发生,故将本算法应用到彩色图像中将是下一步工作研究的重点。

参考文献

- [1] 关涛,周东翔,刘云辉,等.基于自适应阈值分割的宫颈细胞图像分类算法[J].信号处理,2012(09):1262-1270
- [2] 王晓华,王绍虎,余增亮.细胞图像的采集区域划分及细胞识别[J].计算机工程,2003(20):35-37
- [3] 张燕红,李瑛,张燕宁,等.基于数学形态学的细胞图像分割方法

(上接第 265 页)

右特征向量矩阵,新序列所含的信息更多,用于异常检测变量的参数个数减少,异常检测结果会更精确。证明了该方法的有效性和准确性。

多变量时间序列长度值和 k 的取值可能会影响算法的效率。当 $k=5$ 时,数据样本的数量从 100 增加到 300,执行算法消耗的 CPU 时间如图 5 所示。从图 5 可以看出,消耗的时间与样本个数 N 呈近似二次方关系;当固定 $N=100$ 时,让相邻值 k 从 3 增加到 13,花在算法的 CPU 时间如图 6 所示,可以看到当 k 增加时算法消耗时间逐渐增加。

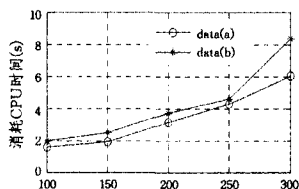


图 5 序列个数对算法的影响

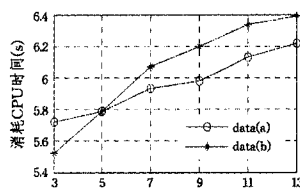


图 6 k 近邻个数对算法的影响

结束语 相似性度量方法是提高多变量时间序列异常检测算法性能的关键。主成分加权 Euclid 范数距离度量算法与 k 近邻局部异常检测算法相结合,能够有效去除异常检测中的多变量之间相关性的影响,减少参与异常检测变量的个数,提高异常检测的准确性,并且对于时间序列的长度没有特殊要求,不仅适用于多变量时间序列,如果将 $D_{Euc}(A, B, w)$ 距离替换为欧氏距离或动态时间弯曲距离,还可以计算单变量时间序列之间的相似性,用于单变量时间序列异常检测。针对数据的异常检测实验验证了异常检测的有效性和合理性。

参考文献

- [1] Tsai C Y, Chuang C-C. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm[J]. Com-

putational Statistics & Data Analysis, 2008, 52(10): 4658-4672

- [4] 刘艳丽,孟朝晖.层次聚类在细胞图像分析中的应用[J].计算机应用与软件,2013(05):287-290
- [5] 杨建菊,张贵英.基于自适应滤波的淋巴细胞图像分割算法研究[J].计算机仿真,2012(08):257-260
- [6] Colantonio S. Automatic fuzzy-neural based segmentation of microscopic cell images [J]. International Journal of signal and Imaging systems Engineering, 2008, 1(1): 18-24
- [7] 梁肖,胡贞,吕晓玲,等.基于自适应阈值的活体细胞图像分割改进方法[J].长春理工大学学报:自然科学版,2013(z2):138-140
- [8] 张鑫,陈伟斌.基于形态学重构的多结构元细胞图像边缘检测[J].计算机仿真,2009(08):216-219
- [9] 汪婧,曹益平,程旭升.一种基于 H 直方图变换的白细胞图像分割方法[J].光学与光电技术,2013(02):74-78
- [2] Kumpulainen P, Kimmo. Local anomaly detection for mobile network monitoring[J]. Information Sciences, 2008, 178(20): 3840-3859
- [3] Jiang F, Yuan J, Tsafaris S, et al. Anomalous video event detection using spatiotemporal context[J]. Computer Vision and Image Understanding, 2011, 115(3): 323-333
- [4] 肖辉.时间序列的相似性查询与异常检测[D].上海:复旦大学,2005:33-40
- [5] Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2011:3449-3456
- [6] Baragona R, Battaglia F. Outlier detection in multivariate time series by Independent Component Analysis [J]. Neural Computation, 2007, 19(1): 1962-1984
- [7] Yamanli K, Takeuchi J. A unifying framework to detecting outliers and change-points from nonstationary data [C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2002). New York: ACM Press, 2002: 676-681
- [8] 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
- [9] 张力生,杨美洁,雷大江.时间序列重要点分割的异常子序列检测[J].计算机科学,2012,39(5):183-186
- [10] 郭小芳,张绛丽.基于加权范数的多维时间序列相似性主成分分析[J].江苏科技大学学报:自然科学版,2011,25(5):466-469
- [11] Agyemang M, Ezeife C I. LSC-Mine: Algorithm for Mining Local Outliers[C]//Proceedings of the 15th Information Resource Management Association (IRMA) International Conference. New Orleans, 2004, 1: 5-8
- [12] Yang K, Shahabi C. An efficient k nearest neighbor search for multivariate time series [J]. Information Computation, 2007, 205(1): 65-98