时间序列异常检测

周大镯[1,2],刘月芬[2],马文秀[2]
ZHOU Da-zhuo[1,2],LIU Yue-fen[2],MA Wen-xiu[2]

1.天津大学 管理学院,天津 300072
2.河北经贸大学 计算机中心,石家庄 050061
1.School of Management,Tianjin University,Tianjin 300072,China
2.Computer Center,Hebei University of Economics and Trade,Shijiazhuang 050061,China
E-mail:zhou_zhuo@163.com

ZHOU Da-zhuo,LIU Yue-fen,MA Wen-xiu.Effective time series outlier detection algorithm based on segmentation. Computer Engineering and Applications,2008,44(35):145-147.

Abstract: A new time series outlier detection algorithm of high-efficiency is proposed for the foundation of k-nearest local outlier detection algorithm based on segmentation.Firstly,series important point as segmentation point can compress high-proportionally time series data in this algorithm;Secondly,the outlier pattern of time series can be detected by local outlier detection technique.Experimental results on electrocardiogram(ECG) data show that the algorithm is effective and reasonable.
Key words: time series;outlier pattern;local outlier factor;series important point

摘 要:在k-近邻局部异常检测算法的基础上,结合时间序列的分割方法,提出了一种高效的时间序列异常检测算法。该算法首先把序列重要点作为数据的分割点,对时间序列数据进行高比例压缩;其次利用局部异常检测方法检测出时间序列中的异常模式。通过心电图(ECG)数据实验验证了算法的有效性和合理性。
关键词:时间序列;异常模式;局部异常因子;序列重要点
DOI:10.3778/j.issn.1002-8331.2008.35.044 **文章编号**:1002-8331(2008)35-0145-03 **文献标识码**:A **中图分类号**:TP311.13

时间序列是一类重要的数据对象,在经济、气象、医疗等领域都普遍存在,它们具有数据量大、维数高、更新速度快等特点。近年来许多学者在时间序列的挖掘方面做了很多工作,相关的研究主要集中在时间序列分割、序列聚类和分类、相似查询、模式发现等研究方向。在时间序列挖掘中,大部分挖掘任务的目的是为了发现那些频繁出现的模式,期望发现某种规律,异常数据通常被作为噪声而忽略,而在另外一些领域,尽管异常数据与正常数据相比是不经常发生的事件,但信息背后可能隐藏着一些重要信息,异常数据的发现往往能带给人们更有价值的知识。例如在金融领域,跟踪信用卡顾客的使用情况,当顾客在某段时期内的信用卡使用情况异常时,能够及时报告,预防信用欺诈。首先提出序列分段点的概念,描述了局部异常检测方法,其次利用异常检测算法计算出最异常时间序列模式,最后分析了算法的性能和有效性。

1 相关工作

目前,时间序列的异常还没有一个公认的定义,研究也比较少,人们普遍采用的是 Hawkin 给出的定义[1]:异常是在数据集中偏离大部分数据的数据,使人怀疑这些数据是由不同的机制产生的,而非随机偏差。从 20 世纪 80 年代起,异常检测问题在统计学领域里得到了广泛研究,随着其应用领域的不断扩展,以及其它领域方法和技术的融合,研究人员提出了许多不同的检测方法。假设检验是最早用来发现异常样本的基于统计学原理的方法[2],它基于对小概率事件的判别来实现对数据样本异常性的鉴别,主要缺陷是事先要假定数据集符合特定的分布模型,针对大量分布特征未知数据时,这种先验假设存在很大的局限性。近年来,基于数据挖掘的异常检测研究取得了一定的进展,Knorr 等[3]首先提出了基于距离的异常检测方法,从全局角度考虑通过计算数据点或对象之间的距离来检测孤立点,当数据集含有多种分布或数据集由不同密度子集混合而成时效果不好。Breunig 等[4]提出了基于密度的异常检测算法 LOF(Local Outlier Factor),这种算法克服了不同密度子集混合而造成的检测错误,检测精度比较高,但在处理大数据集时复杂度过高,无法获得令人满意的响应速度。

以上研究大多是检测时间序列中存在的异常点,近年来针对时间序列异常模式的研究也有了一些成果。Eamonn K 等[5]研究如何检测时间序列中最异常的时间子序列,算法首先对时间序列符号化,通过符号检索出时间序列中的最不寻常的子序

基金项目:河北省科技攻关计划项目(No.062135140)。
作者简介:周大镯(1971-),女,副教授,博士研究生,主要研究方向:数据挖掘;刘月芬(1970-),女,讲师,硕士研究生,主要研究方向:概率统计。
收稿日期:2007-12-20 **修回日期**:2008-03-20

列。林果园等[6]研究了在改进的隐马尔科夫方法的基础上,将动态行为和全局特征结合起来进行异常检测的方法。翁小清等[7]提出了基于滑动窗口的异常检测方法,该方法使用滑动窗口对原始时间序列进行分割,利用扩展的 Frobenius 范数来计算两个子序列之间相似性,利用局部异常检测方法实现了时间序列异常子序列(含异常数据)挖掘,该方法能有效发现异常子序列,但面对大数据量的时间序列,算法的时间复杂度过高。

2 问题描述及相关定义

时间序列包含数据量大、维数高、数据更新快,若直接在原始时间序列上进行异常模式检测非常困难。解决这个问题的方法是将一个长时间序列分割为若干个相对短但不重叠的子序列,并将各个子序列转换为某种高级数据表示形式。分段线性表示 PLR[8]比较符合人们直观经验,通常索引结构维数低、计算速度较快,所以被许多人采用。以往的 PLR 方法过分注重拟和原始数据,在拟和的过程中,平滑掉了原始数据的一些重要特征,为了保证异常模式的有效检测,采用序列重要点作为时间序列的分割点。

定义 1 序列重要点。给定时间序列,误差最大的区域中,距离区域端点最远的点,这样的点称为序列重要点(Series Important Point,SIP)。序列重要点作为分割点,其所在区域误差最大,同时由于距离端点最远,对原始数据的形状影响大,依次选定这样的点作为序列的分割点。

定义 2 时间序列 X 的模式表示[9]。每个直线段采用如下二元组表示,其中 l_i 为 X 第 i 段的长度,代表了趋势变化的长短,m_i 为每个直线段的斜率,表示变化趋势:

$$X=<(l_1,m_1),(l_2,m_2),\cdots,(l_c,m_c)> \tag{1}$$

定义 3 模式距离[9]。经分段表示的每个直线段长度和斜率可能不等,为了有效检测时间序列中的异常模式,定义模式 $p(l_1,m_1)$ 和模式 $q(l_2,m_2)$ 之间的距离为:

$$d(p,q)=\frac{|l_1-l_2|}{\min\{l_1,l_2\}}+\frac{|m_1-m_2|}{\min\{m_1,m_2\}} \tag{2}$$

定义 4 p 点的 k 近邻距离 $k\text{-}dist(p)$[4]。给定一个正整数 k 和一个数据点集合 D,在 D 中 p 点的 $k\text{-}dist(p)$ 满足下面两点:(1)至少有 k 个点 $o\in D\backslash\{p\}$,$d(p,o)\leqslant k\text{-}dist(p)$;(2)最多有 $k-1$ 个点 $o\in D\backslash\{p\}$,$d(p,o)<k\text{-}dist(p)$。

在图 1 中当 k=3 时,$k\text{-}dist(p)=d(p,o)$,其中 $d(p,o)$ 表示 p 点到 o 点的距离。

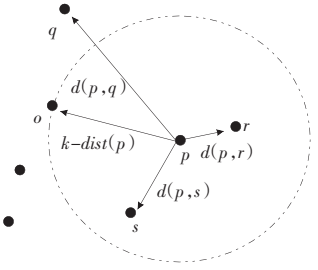


图 1 k=3 **时的** $k\text{-}dist(p)=d(p,o)$

定义 5 q 点到 p 点的 k 近邻可达距离 $r\text{-}dist_k(q,p)$:

$$r\text{-}dist_k(q,p)=\max(d(q,p),k\text{-}dist(p)) \tag{3}$$

图 1 中因为 $d(q,p)>k\text{-}dist(p)$,所以 q 点到 p 点的 $r\text{-}dist_k(q,p)=d(q,p)$;而 r 点到 p 点的 $d(r,p)<k\text{-}dist(p)$,因此,$r\text{-}dist_k(q,p)=k\text{-}dist(p)$。

定义 6 点 q 的 k 局部可达密度 $lrd(q)$[4]:

$$lrd(q)=\frac{k}{\sum_{p\in k(q)} r\text{-}dist_k(q,p)} \tag{4}$$

其中 $k(q)$ 表示 q 的 k 近邻范围,局部密度反映了该点的周围点的分布密度,具有较小局部密度的点成为局部异常点的可能性比较大,反之亦然。

定义 7 q 点的局部异常系数 $LOF(q)$[4]:

$$LOF(q)=\frac{\frac{1}{k}\sum_{p\in k(q)} lrd(p)}{lrd(p)} \tag{5}$$

如果点 q 的局部异常系数较大,就意味着该点的局部范围所含点比较稀疏,说明该点是异常的可能性比较大。

定义 8 异常模式。异常模式是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。通过求各个模式的局部异常系数,局部异常系数的值较大的是异常模式。

3 基于序列重要点分割的异常检测算法

3.1 异常检测算法的设计模型

本算法主要包括以下几个子模块,如图 2 所示。

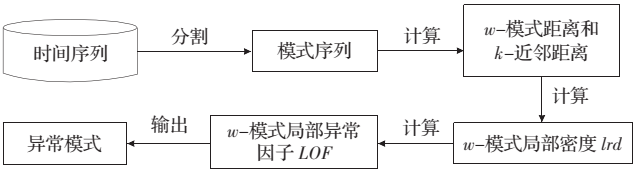


图 2 基于分割的时间序列异常检测模型

3.1.1 时间序列进数据分割

序列重要点在选择过程中首先把开始点和结束点作为初始序列重要点,在区域内按垂直距离计算每个点与近邻端点的距离和区域误差,误差最大的区域内距离端点最远的点作为序列重要点加入到分割点中,直到所有区域误差小于输入误差为止。

3.1.2 计算模式的 k 近邻距离

分割后的时间序列被描述为模式序列,为了发现最不寻常的异常子序列模式,由公式(3)计算模式之间的 k 近邻距离,为输出 k 个异常系数最大的模式做准备。

3.1.3 局部异常检测

通过第二步计算出的每个模式子序列的 k 近邻距离,由公式(4)可以计算出每个模式子序列的局部稀疏密度 lrd,再根据公式(5)计算出它们的局部异常因子 LOF,异常因子大的模式子序列成为异常模式子序列的可能性最大,反之亦然。

3.2 算法描述

输入:时间序列 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,局部误差 e,k 近邻,w 模式子序列长度。

输出:异常模式子序列。

(1)$[T,L,M]\leftarrow Segmentation(X,e)$;//对时间序列进行分割,直到每个时间段的误差都不超过输入的 e 为止

(2)For each point $p \in [T,L,M]$

(3){$m \leftarrow <p_1, p_2, \cdots, p_{i+w-1}>$;//由输入的 w 形成模式子序列

(4)$A \leftarrow k\text{-}dist(m_i)$;}//计算每个模式子序列的 k 近距离,同时存储该模式子序列的 k 近邻距离所包含的模式子序列位置

(5)For each point $m \in [T,L,M]$

(6){$B \leftarrow lrd(m)$;}//计算局部可达密度

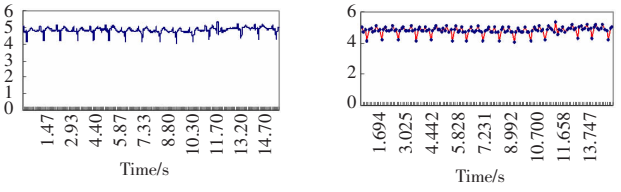
(7)For each point $m \in [T,L,M]$

(8){$C \leftarrow LOF(m)$;}//计算局部异常因子

(9)$q \leftarrow \max(C)$;//局部异常因子最大者作为异常子序列模式输出

4 实验与结果分析

为了验证算法,使用工具软件 Matlab 7.1 在 CPU 1.4 GHz、内存 256 MB、硬盘 40 GB、Windows XP 操作系统的计算机上验证该算法,实验采用了一个包含 5 400 个数据的心电图 ECG数据集。将数据分割的局部区域误差设置为 e=3,时间序列被 138 个直线段描述,压缩比例为 138/5 400,尽管压缩比例很高,但仍保持了原始数据的整体特征,时间序列分割前后比较见图 3。



(a)原始时间序列 (b)误差 e=3,分段数为 138

图 3 时间序列分割前后比较

4.1 实验结果

固定分割误差 e=3、滑动窗口 w=1 的情况下,不同的 k-近邻(k 分别取值为 7、9、11)取值,局部异常因子最大的前 3 个模式序列见表 1,模式下标表示分割点位置。实验结果表明,在对时间序列进行分割的基础上,利用局部异常检测方法得到的异常数据与实际是一致的。

表 1 心电图 ECG 数据集上的实验结果

k-近邻	模式序列	模式局部异常因子	时间范围/s
7-近邻	P_{4118}	10.324 7	11.436~11.447
	P_{4104}	6.025 3	11.397~11.436
	P_{1752}	4.432 4	4.864~4.928
9-近邻	P_{4118}	8.816 9	11.436~11.447
	P_{4104}	5.193 9	11.397~11.436
	P_{1752}	4.246 0	4.864~4.928
11-近邻	P_{4118}	5.772 6	11.436~11.447
	P_{4104}	4.303 2	11.397~11.436
	P_{1752}	3.742 1	4.864~4.928

4.2 实验分析

算法的时间复杂度主要由 4 部分组成:(1)时间序列分割,由于采用的是序列重要点作为的分段点,时间复杂度为 O($c\log n$)其中 c 为分段数、n 为时间序列长度,分段数与分割误差有关,e 越小,得到的分段数 c 越大。当固定 k=9 时不同的输入误差对算法执行效率的影响见图 4 所示;(2)计算模式子序列之间的距离,时间复杂度与时间序列分割的分段数 c 有关,时间复杂度为 O(c^2);(3)计算模式的局部可达密度 lrd,该部分要用到模式子序列之间的距离、r 可达距离和 k 近邻内的模式子序列的位置,此数据在上面已存储,时间复杂度为 O(ck);(4)计算模式的局部异常因子 LOF, 时间复杂度和计算 lrd 是一样的 O(ck)。c 作为数据分割的分段数满足 $c<<n$,总体上算法的时间复杂度不会超过 O(n^2),从以上分析该算法在保证执行效率高的前提下,可以得到有效的异常检测结果。

5 结语

时间序列由于数据量大、维数高,直接检测其中的异常模式非常困难。提出了基于序列重要点分割的时间序列异常检测算法,可以高效地检测出数据中最异常的模式。利用序列重要点对数据进行分割,能很好地描述时间序列的整体形态,分割后的数据使用斜率和分段长度表示,反映了数据在不同时间段内的变化趋势,利用局部异常检测方法可以有效地检测到时间序列中的异常数据。

参考文献:

[1] Hawkins D.Identification of outliers[M].London:Chapman and Hall, 1980.

[2] Billor N,Hadi A,Velleman P.BACON:blocked adaptive computationally-efficient outlier nominators[J].Computational Statistics & Data Analysis,2000:279-298.

[3] Knorr E M,Ng R T.A Unified notion of outliers:properties and computation[C]//ICDM'97.[S.l.]:AAAI Press,1997:219-222.

[4] Breunig M M,Kriegel H P,Ng R,et al.LOF:identifying density-based local outliers[C]//ACM SIGMOD,2000:93-104.

[5] Keogh E,Lin J.Finding unusual medical time-series subsequences: algorithms and applications[C]//IEEE Transactions on Information Technology in Biomedicine,2006:429-439.

[6] 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.

[7] 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.

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

[9] 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.