山东农业大学学报(自然科学版),2017,48(2):287-292 VOL.48 NO.2 2017
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2017.02.026

基于相似性分析的时间序列异常检测方法

孙 焱,林 意
江南大学 数字媒体学院, 江苏 无锡 214000

摘 要: 时间序列数据是按照时间顺序在不同的时间点采集的数据，反映了某一对象随时间的变化状态和程度。由于时间序列的海量性及复杂性，我们采用频域表示时间序列，并以此为基础提出了基于相似性分析的时间序列异常检测方法。将动态模式匹配距离作为衡量相似性的指标，计算每一个模式同其余各模式之间的相似性，据此确定异常状态。该方法大大降低了数据搜索复杂度，提高了系统效率与准确度。

关键词: 时间序列; 相似性分析; 动态模式匹配; 异常检测
中图法分类号: TP39 **文献标识码:** A **文章编号:** 1000-2324(2017)02-0287-06

The Detection Method on Abnormal Time Series on account of Similarity Analysis

SUN Yan, LIN Yi
School of Digital Media/Jiangnan University, Wuxi 214000, *China*

Abstract: Time series data are collected at different time points according to the time order to reflect an objective variation states and contents with time. This paper took frequency domain to express the time series in consideration of the magnanimity and complexity of time series and propose the detection method on abnormal time series on account of similarity analysis to take a dynamic pattern matching distance as a index to calculate the similarity between every model and others hereby to ensure the abnormal state. This method could greatly reduce the complexity in data search and improve the efficiency and accuracy of the system.

Keywords: Time series; similarity analysis; dynamic pattern matching; abnormal detection

近年来伴随着社会经济的高速发展以及科学技术的巨大进步，计算机技术也获得了巨大的进步。随着信息时代的到来，人类社会对于外界信息的依赖变得越来越重要同时也产生了海量的信息数据。另一方面这些数据的产生速度可以用惊人来形容，只需要短短一年甚至几个月整个人类社会的信息量就可以增长一倍。而采用哪种有效手段管理并挖掘出这些海量数据中所隐藏的规律与知识则成为了当代研究者们所关注的热点话题，因此数据挖掘技术在当代社得越来越重要。

时间序列作为一种广泛存在于医学、工程、商业以及自然科学等领域数据库中的常见数据形式，近年来得到了研究人员越来越多的关注[1]。对其进行相关性分析并在此基础上进一步进行数据搜索匹配成为了数据挖掘的重要步骤。时间序列异常是指在数据集中某一数据偏离大部分数据，其数值特性已经超出了随机偏差的范围，而更有可能是由不同机制产生的[2-4]。为了能够有效检测出时间序列数据中的异常数据，本文提出了基于相似性分析的时间序列异常检测方法。首先通过建立合适的时间序列模型来抽象化数据，降低搜索复杂度，便于检索；随后通过一个滑动的窗口平滑处理时间序列，计算其和其他模式的相似度以确定其是否异常。

本文的结构安排如下：首先给出了时间序列模型对数据进行抽象化处理，随后介绍对时间序列进行相似性度量的各类方案，紧接着提出了时间序列异常检测的方案，最后通过仿真实验来验证分析该方案的可行性。

1 时间序列的模型

因为时间序列一般都是高维数的，假如直接要在原始的数据中进行数据挖掘需要付出高昂的代价，其复杂性高效率低下，也会降低算法的准确性和可靠性。针对这个问题，必须采用合适的抽象方法对时间序列进行抽象建模，以达到简化数据模型，去除冗余，搭建出符合要求的数据库索引的目的[5-7]。虽然直接采用传统时间序列分析方法理论上同样可以解决相似性分析的问题，但是实际运

收稿日期: 2016-08-23 **修回日期:** 2016-10-02
作者简介: 孙 焱(1992-),男,研究生.研究方向:时间序列数据挖掘. E-mail:thierry14henry12@163.com

数字优先出版:2017-04-11 http://www.cnki.net

用时间复杂度。因此，必须能够建立一个合适的数据模型，以同时具有 高鲁棒性和低复杂度的优中效果差，因为相似性度量的计算依赖于时间序列的表示方式，这会大大影响计算过程的点，这会大大提高数据索引的效率。

在这里我们采用频域表示法，由于离散傅里叶变换 DFT 开头的几个系数表现突出，能够保留信号的绝大部分能量，因此可以只留存 DFT 头几个系数而直接删除其余系数同时保留下了数据的大部分特征来达到数据压缩的目的，这样做也保留了原始时间序列的基本特性[8]。

离散傅里叶变化 DFT 的具体使用方法如下：给定一个信号 $\vec{x}=[x_t]$， $t=0,1,2,...,n-1$, 则其 n 个点的 DFT 变换可以定义为是 n 个复数 X_f表示，f=0,1,2,…,n-1，组成序列 $\vec{X}$ ：

$$X_f = 1\Big/\sqrt{n}\sum_{t=0}^{n-1} x_t \exp(-j2\pi ft/n), \text{其中 } f=0,1,2,...,n-1$$

其中 j 是虚数单位， $j=\sqrt{-1}$ 。同时信号 $\vec{x}$ 通过傅里叶逆变换后可以恢复出来：

$$X_f = 1\Big/\sqrt{n}\sum_{t=0}^{n-1} x_f \exp(j2\pi ft/n), \text{其中 } t=0,1,2,...,n-1$$

DFT 变换同时对原始时间序列中的局部极大值与局部极小值都进行了数据平滑，这样做使得数据的部分重要信息丢失。除此之外 DFT 还对序列的平稳性有着较高的要求，其对于非平稳的序列并不适用。而且 DFT 变换以相同长度的系数来度量所有长度的时间序列，降低了方法的合理性。因此，在此基础上我们提出使用滑动窗口的简单平滑方法，不但可以去除噪声，也较为真实地反映出数据的实际特性。具体操作如下：

给定长度为 n 的时间序列 $T(t_1,t_2,...,t_n), t_i$ 的键值为 $kp_i(i=1,2,...,n)$; 设定滑动窗口的大小为 w，则根据平滑计算公式计算得出平滑后的键值 k。 $k_i=(kp_{i-w+1}+kp_{i-w+2}+...+kp_i)/w$

2　时间序列的相似性度量

2.1　Minkowski 距离

欧几里得距离是平时最常用的一种距离计算方式。假定长度为 n 的时间序列看作是一个 n 维欧式空间中的点，它的坐标点则对应着时间序列在各个时间点的取值，则两条长度为 n 的时间序列之间的欧氏距离就是这个 n 维空间中两点的距离[9-11]。其可以作如下描述：

假定两条时间序列为 $X=\langle x_1,x_2,...,x_n\rangle, Y=\langle y_1,y_2,...,y_n\rangle$, 则这两个时间序列间的欧式距离为：

$$d(X,Y)=(\sum_{i=1}^{n}|x_i-y_i|^2)^{\frac{1}{2}} \tag{1}$$

Minkowski 距离作为相似性度量距离，其是欧氏距离的推广，如下定义：

$$L_p(X,Y)=(\sum_{i=1}^{n}|x_i-y_i|^p)^{\frac{1}{p}} \tag{2}$$

当 p=2 时，Minkowski 距离就成为了欧式距离，而在 p=1 时则变成了曼哈顿距离；当 p 趋于无穷大时则称为最大距离。由于 Minkowski 距离同样满足非负性即所有值不小于零、对称性以及距离三角不等式，所以该距离也能够作为一种度量距离。

正是由于 Minkowski 距离具有满足三角不等式的特性，所以在基于索引进行数据查询时，能够根据这一特性将其作为索引距离，快速过滤某些不符合索引条件的节点，从而提高索引速度。

Minkowski 距离应用于数据索引的相似性度量时具有诸多优点，其简单直观，计算简便，具有非常高的可扩展性，同样可以应用于数据地查询以及聚类等方面。然而 Minkowski 距离应用在时间序列数据挖掘时却不具备很好的可靠性，其对于时间序列自身的噪声以及波动不具备很好的鲁棒性，相似的时间序列也会存在着多种变形，例如振幅平移与伸缩、线性飘逸、不连续、时间轴伸缩等等。

2.2　动态模式匹配距离

虽然 Minkowski 距离计算简便，在索引查询以及聚类领域有很优秀大表现，但是其对于时间序列的时间轴弯曲以及伸缩并不友好。所以为了能够更好的进行时间序列的相似性分析，这一节提出使用动态模式匹配距离，同传统距离所不同的是，动态模式匹配距离并不是根据两个目标点之间的距离进行计算，而是通过模式匹配进行。这样做一方面是因为模式的定义较为灵活，同时因为时间序列的模式一般远远小于序列的长度，这样可以降低计算的数据量，提高算法效率。模式之间的距离使用加权欧氏距离进行定义：

假定给定两个模式 $p_1=(l_1,k_1)$和 $p_2=(l_2,k_2)$，其中 l 和 k 分别表示模式的长度与斜率，则两个模式之间的距离可以如下定义：

$$d(p_1,p_2)=\frac{|l_1-l_2|}{\min\{l_1,l_2\}}+\frac{|k_1-k_2|}{\min\{k_1,k_2\}} \tag{3}$$

在以上定义中，分母的作用是将长度和斜率这两个不同的量纲进行统一，而取最小值则是为了能够突出短模式的重要性。

假定时间序列 $X=\langle x_1,x_2,...,x_m\rangle$，其模式可以表示为 $P(X)=\langle px_1,px_2,...,px_u\rangle$，另一个时间序列 $Y=\langle y_1,y_2,...,y_n\rangle$，同样的其模式可以表示为 $P(Y)=\langle py_1,py_2,...,py_v\rangle$，则其两个模式之间的动态模式匹配距离为：

$$D(X,Y)=d(px_1,py_1)+\min\begin{cases}D(P(X)-px_1,P(Y))\\D(P(X),P(Y)-py_1)\\D(P(X)-px_1,P(Y)-py_1)\end{cases} \tag{4}$$

公式中 $d(px_1,py_1)$表示的是 px_1 与 py_1 之间的模式距离，而 $P(X)$-px_1 和 $P(Y)$-py_1 分别表示 $P(X)$和 $P(Y)$去除了第一个元素后的序列。

从上述公式可以看出，模式是由他的长度和斜率这两个特征表示。由于模式的长度与时间序列的振幅大小无关，而其斜率则体现了时间序列振幅的相对大小，所以所提的动态模式匹配距离可以克服时间序列的振幅平移与伸缩变换。除此以外，因为采用了模式的动态匹配方法，可以实现时间序列在时间轴上的伸缩和弯曲。

动态模式匹配距离可以使用累积距离矩阵的方法进行计算，这样的话其时间复杂度就为 $O(mn/uv)$，这其中的 m 和 n 分别表示两个时间序列的长度，而 u 和 v 则表示模式的平均长度。由此可见，如果模式的平均长度越长的话，动态模式匹配的时间复杂度就越低。进一步可以看出，采用动态模式匹配距离的计算方法要远远优于 Minkowski 距离计算。

3　时间序列的异常检测

3.1　时间序列的异常模式

异常可以简单理解为在一个时间序列数据集中，其某一个数据点的值与其他数据点值存在非常明显的差别，超出了随机产生的可能，有可能是因为不同的机制而产生的，这一类数据就称为异常。如图中就是一种直观的异常模式。其中的点 3,4,5,6,7 单独来看时，其值与整体数据而言并没有什么差异，然而当这些数据在时间上连续出现时就形成了整个时间序列中的异常数据。

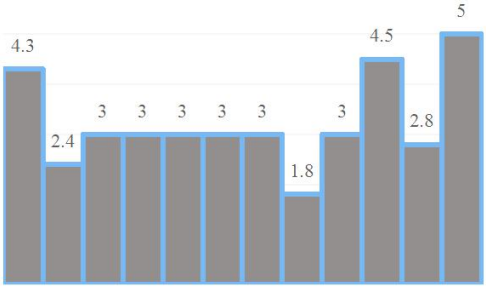

图 1 时间序列异常

Fig.1 The abnormal time series

3.2　*K*-近邻原理

某一点 *p* 的 *k*-近邻距离（*k*-dist（*p*））可以如下定义[12-14]：假定 *k* 是一个正整数，*D* 则是一个数据点的集合，而 *p* 为改数据集中的一个点，*p* 点的 *k*-近邻距离应当满足以下两个条件：

（1）数据集 *D* 中至少有 *k* 个数据点（*p* 点除外），这些数据点到 *p* 点距离不大于 *k*-dist（*p*）。

（2）数据集 *D* 中至多有 *k*-1 个数据点（*p* 点除外），这些点到 *p* 点的距离小于 *k*-dist（*p*）。

如图所示，当 *k*=4 时，点 *p* 的 *k*-近邻距离 *k*-dist(*p*)=*d*(*p*,*u*),*d*(*p*,*u*)即表示点 *p* 到 *u* 的距离。

在数据集 *D* 中，点 *p* 到点 *t* 的 *k*-近邻可达距离 *r*-dist(*p*,*t*)可以定义为：

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

点 *q* 的 *k*-局部可达密度 lrd(*q*)可以定义为：

$$lrd(q)=\frac{k}{\sum_{x\in k(q)} r-dist(q,x)} \tag{6}$$

以上公式中，*k*(*q*)表示 *q* 点的近邻范围，由局部密度的定义方式，可以看出该密度反应的是点 *q* 周围的数据点分布情况，如果密度越高表示在数据集中类似于点 *q* 的点越多，同时也表明点 *q* 是异常数据的概率也越小。

数据集 *D* 中，点 *q* 的局部异常系数 LOF(*q*)可以如下定义：

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

如上公式所示，如果局部异常系数越大则表明 *q* 点的局部范围内数据点较为稀疏，则其为异常点的可能性也就越大[15,16]。

3.3　基于相似性分析的异常检测算法

基于相似性分析的异常检测算法不是直接对比目标两个点，而是采用 2.2 节中提出的动态模式匹配距离，将两个模式进行比较。由于模式的数据量远远小于原始数据量，这样就极大地降低了需要检测的数据量，降低了算法的复杂度。同时也对噪声进行了过滤。并且使用这种方式计算出的异常是一个目标范围而不再是单单的某一个数据点，这极大地提高了算法的鲁棒性与合理性而且也更加符合实际。

该方法的流程图如图所示，第一步将目标时间序列进行频域抽象化表示，形成模式化后的序列数据；第二步计算每一个模式同其余模式之间的模式距离分析其相似性，计算 *k*-近邻距离，紧接着根据公式 6 和公式 7 计算出每一个模式的局部密度以及局部异常因子；最后选取具有较大局部异常因子的模式判定为异常模式。

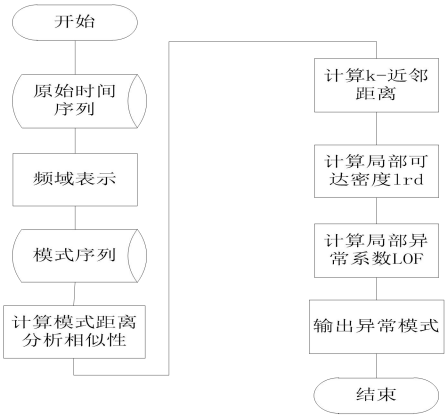

图 2 时间序列异常检测流程

Fig.2 The detection process of abnormal time series

4　实验结果与分析

4.1　方法验证

这部分我们对所设计的方法进行仿真验证，采用 dell 便携机作为仿真主机，频率 2.27 GHz，内存 4 G，基于 MATLAB 进行仿真分析。验证分为两个部分，分别是对该方法的可行性以及可靠性进行测试。可行性测试的方法是通过 MATLAB 仿真对一系列随机产生的数据进行模式距离计算后计算出每一个点的局部异常系数，并输出异常数据。

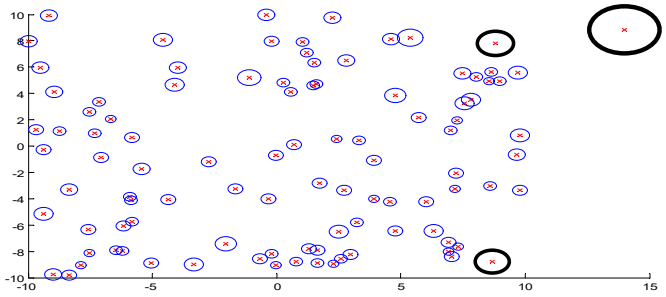

图 3　可行性验证结果图

Fig.3 The verification results for feasibility

如图所示，对产生的一系列时间序列模式化后计算出了每一个数据点的拒不异常系数，并将其直观地用圆的半径表示，半径越大则该点的异常系数越大，其为异常模式的可能性也就越大。设置合适的阈值之后就可以通过比较判别出异常模式，如图中黑色圆所示。

随后我们对该方法的可靠性进行验证，探讨其在不同数据量下的检测准确性与时间消耗。

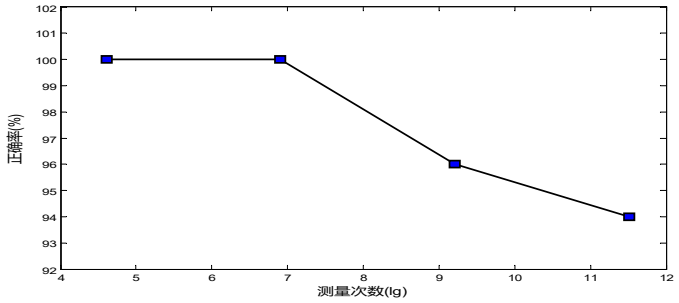

图 4　不同数据量下准确度（%）

Fig.4 The detection accuracy on different data

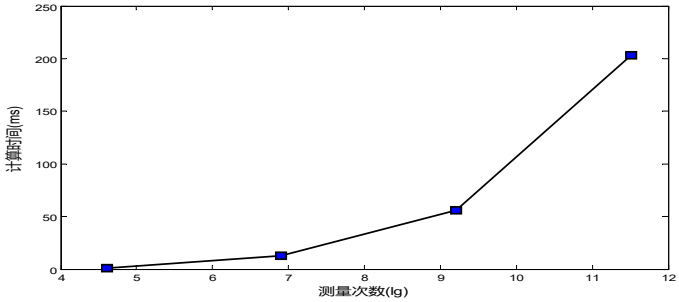

图 5　不同数据量下检测时间

Fig.5 The detection time on different data

如图 4 所示为改方法在不同数据量下的准确度，可以看出随着检测数据量的不断增加，该方法虽然准确度有所下降但是依然保持着非常高的准确性。从图 5 中可以看出，随着数据量的增多，方案的检测时间迅速增加，表明该方法对于算法的复杂度优化方面还存在着缺陷。综上所述，该方法可以有效对于时间序列的异常进行检测，并且在大数据量下依然具有非常高的准确性，但是其时间消耗方面需要进一步优化。

4.2　方案不足

该方法基于相似性的分析设计了时间序列异常检测的方法，虽然可以有效完成异常检测的目标，但是也存在着一些不足之处：

（1）对于时间序列的模式化还有着其他许多更优秀的方法，而不单单是频域表示，如分段线性表示法等，可以采用其他抽象方法进行对比仿真。

（2）虽然动态模式匹配距离获得了很高的准确性，但是其算法的速度依然有待提高，可以对其进行进一步的优化。

5　结 论

本文基于相似性分析，结合局部密度计算数据点的数据密度设计了基于相似性分析的时间序列异常检测方法。提出了动态模式匹配距离，使用模式而不是空间中的两个点进行距离计算，模式的距离就直观反映了每一个模式与其余模式之间的相似性，该距离很好的克服了时间序列振幅平移、时间伸缩等困难，采用频域表示法模式化时间序列极大地降低了数据量，提高了算法效率。同时结合局部密度计算算法有效检测了每一个目标数据的异常情况。总结而言，该方法不但可以有效检测出异常数据，而且极大地降低了数据量提高了算法效率，同时可以克服时间序列的伸缩等缺陷，具有很好的扩展性。

参考文献

[1] 陈海燕,刘晨晖,孙 博.时间序列数据挖掘的相似性度量综述[J].控制与决策,2017,2(1):1-11
[2] 陈 然.基于相似性分析的时间序列异常检测研究[D].重庆:西南交通大学,2011:30-33
[3] 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
[4] 曹文平,熊启军,罗 颖,等.基于相关性分析的时间序列异常检测方法[J].信息系统工程,2012,22(10):131-132
[5] 闫明月.时间序列相似性与预测算法研究及其应用[D].北京:北京交通大学,2014:18-30
[6] 李 权,周兴社.一种新的多变量时间序列数据异常检测方法[J].时间频率学报,2011,34(2):154-158
[7] 唐 亮.时间序列挖掘和相似性查找技术的研究[D].上海:上海师范大学,2004:18-20
[8] 方加果.基于相似性分析的时间序列数据挖掘算法研究[D].杭州:浙江大学,2011:33-35
[9] 杨永强.基于相似性分析的时间序列数据挖掘研究[D].重庆:西南交通大学,2007:22-25
[10] 曾海泉.时间序列挖掘与相似性查找技术研究[D].上海:复旦大学,2003:20-30
[11] 冯 钧,陈焕霖,唐志贤,等.一种基于 DTW 的新型股市时间序列相似性度量方法[J].数据采集与处理,2015,30(1):99-105
[12] 张 军.基于时间序列相似性的数据挖掘方法研究[D].南京:东南大学,2006:33-39
[13] 邱均平,王菲菲.时间序列相似性查询与索引方法研究[C].中国索引学会年会暨学术研讨会,2009:4-8
[14] 门连生,卫婧菲,李 中.基于形态相似距离的时间序列相似性度量[J].计算机工程与应用,2015,51(4):120-122
[15] 刘 杰.时间序列相似性查询的研究与应用[D].北京:北方工业大学,2016
[16] 刘永志.基于两点的时间序列相似性研究[J].盐城工学院学报:自然科学版,2014,27(4):1-4

（上接第 286 页）

参考文献

[1] 纪荣芳.主成分分析法中数据处理方法的改进[J].山东科技大学学报:自然科学版,2007,26(5):95-98
[2] 刘洪琼,刘知贵,张活力.视频稳定系统的角点跟踪算法[J].计算技术与自动化,2012,31(2):78-81
[3] 穆欣侃,徐德江,罗海波,等.基于新模板匹配的运动目标跟踪算法[J].火力与指挥控制,2013,38(3):22-26
[4] 徐 珩,贺飞越.模板匹配跟踪的哈希增强算法[J].计算机应用与软件,2016,33(7):167-171
[5] 赵径通,吴钟建,李 黎.一种抗遮挡的相关滤波器跟踪算法[J].科技经济导刊,2015(18):15-16
[6] 田 立,周付根,孟 偲,等.互相关跟踪算法的多核 DSP 快速实现[J].高技术通讯,2013,23(12):1248-1253
[7] 叱干鹏飞.基于分布场模型的目标跟踪方法研究[D].西北农林科技大学,2014:52-56
[8] 徐德江,史泽林,罗海波.零均值归一化互相关跟踪算法特性研究[J].红外与激光工程,2010,39(S):485-489
[9] 钟 韬.Spearman 秩相关系数计算方法的一点注记[J].读写算:教师版,2016(35):22-23