

基于异常检测的时间序列研究

陈运文,吴 飞,吴庐山,刘 博

(上海工程技术大学 电子电气工程学院,上海 201620)

摘 要:时间序列是一种重要类型的时态数据,广泛应用于科研、经济和军事等各个领域,而针对时间序列的异常检测研究是近年来大数据挖掘的热点与难点。文中以国内外最近的研究成果和时间序列检测的研究价值为基础,探讨了时间序列异常检测的定义并对相关异常检测方法进行归类研究与总结,同时指出每种异常检测方法的优缺点,并进一步分析部分具有代表性的时间序列异常检测的相关研究成果,尤其是讨论了多元时间序列异常检测研究所面临的难题,并给出解决此难题的思路和方法。最后总结归纳时间序列异常检测的几点建议与未来研究方向,以期对相关研究提供有益的参考。

关键词:时间序列;异常检测;数据挖掘;多元时间序列

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2015)04-0166-05

doi:10.3969/j.issn.1673-629X.2015.04.038

Research on Time Series Based on Anomaly Detection

CHEN Yun-wen, WU Fei, WU Lu-shan, LIU Bo

(College of Electronic and Electrical Engineering, Shanghai University of Engineering Science,
Shanghai 201620, China)

Abstract: Time series is a temporal data of important class which are widely used in scientific research, economics, and military and other fields, but the study aiming at anomaly detection of time series in recent years is the hot and difficult of data mining. In this paper, based on the recent research achievements at home and abroad and the research value of time sequential detection, discuss the definition of anomaly detection of time series and the research on related anomaly detection methods are classified and summarized, at the same time point out the advantages and disadvantages of each type of anomaly detection methods, and further analyze relevant research achievements of some typical time series of the anomaly detection, especially discussing the multivariate time series, the challenges faced by the institute of anomaly detection, and giving ideas and methods to solve this problem. Finally, some suggestions about anomaly detection and future research trends are also summarized, which is hopefully beneficial to the researchers of time series and other relative domains.

Key words: time series; anomaly detection; data mining; multivariate time series

0 引言

随着信息化的推进,大量数据不断涌现。据美国南加州大学 2011 年最新研究报告显示:自 20 世纪 80 年代起,全球信息量每隔十几个月甚至几个月就要增加一倍,呈爆炸式增长,至 2007 年,全球数据信息总量已达到 295 EB。因此,如何从海量的数据中挖掘感兴趣的知识是亟待解决的问题。而在这些数据中有非常大一部分是时间序列。所谓的时间序列就是一系列按照时间先后顺序记录的各个观测值。时间序列在军事、经济以及科学观测等各个社会领域中都广泛存在,

引起了相关研究学者的极大关注。

相对正常时间序列数据而言,尽管异常数据数量很少,然而这并不代表异常数据不重要,相反,这些少数的异常数据却可能隐藏着重要的信息^[1]。例如在医学领域,对病人进行心电图观测时,要求能检测出心脏的异常跳动,以便医生及时发现病情。此外,时间序列的异常检测研究还可以广泛应用于发动机状态监测、网络入侵检测、反洗钱侦测、网络舆情监控、信用卡反欺诈、股市分析、不正当税务行为监测、重大建设项目稽查以及自然灾害分析等领域。可见,进行时间序列

收稿日期:2014-04-02

修回日期:2014-07-06

网络出版时间:2015-02-23

基金项目:国家自然科学基金资助项目(61272097)

作者简介:陈运文(1987-),男,硕士研究生,研究方向为嵌入式智能系统;吴 飞,教授,CCF 高级会员,研究方向为计算机并行处理与节能控制。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150223.1233.006.html>

异常检测研究具有重要的理论价值与实际意义。

1 时间序列异常检测的定义

时间序列广泛存在于各种大型的金融、医学、工程和社会科学数据库中,与其他数据类型相比具有两个特点:一是时间属性,其每个变量的记录都必须有时间维并按时间先后顺序进行排列,但是像市场货篮数据就没有这种属性;二是序列属性,即其记录值是变量在某一时间段内连续的记录值并具有一定的规律。按其变量数可分为一元时间序列和多元时间序列,具体定义如下:

定义1 时间序列:一系列按照时间先后顺序记录的值为 $S = \{v_i(1), v_i(2), \dots, v_i(t), \dots, v_i(n)\}$ 称为时间序列。其中, t ($t = 1, 2, \dots, n$) 表示时刻, i ($i = 1, 2, \dots, m$) 表示变量, $v_i(t)$ 表示第 i 个变量在 t 时刻上的记录值。当 $m = 1$ 时, S 为一元时间序列 (Univariate Time Series, UTS); 当 $m > 1$ 时, S 为多元时间序列 (Multivariate Time Series, MTS)。

同时,为使得对 MTS 的异常检测有意义,MTS 还需要满足以下三个要求:

(1) 对于 MTS 数据集,各序列的变量维数相同,变量之间一一对应且表示相同的含义;

(2) 对某一 MTS 样本,各变量数值的记录时刻对应,且具有相同的时间粒度;

(3) MTS 需经正则化处理。

研究时间序列异常检测,首先还要明确的就是异常的定义,即什么才是异常?对于异常,学术界并未给出统一的定义。统计学中,将那些不服从序列分布,距离其他对象较远的数据点视为异常^[2];回归模型中,将与指定模型偏离较大的数据点视为异常;众多研究者也依据自己理解提出了一些相似的术语:异常 (Anomaly)^[3]、奇异 (Surprise)^[4]、例外/离群 (Outlier)^[5-6]、新颖性 (Novelty)^[7]、偏离差 (Deviant)^[8] 等。一般比较认可 Hawkins^[9] 给出的定义:异常是在数据集中与众不同的数据,使人怀疑其并非为随机误差所致而是由完全不同的机制而产生。因此,多数学者都是基于此再结合实际的应用领域给出时间序列的异常定义。

2 时间序列异常检测的主要方法

自 20 世纪 80 年代, Barnett 发表第一本关于异常检测的著作《Outliers in statistical data》以来,许多国外的学者纷纷加入,例如: M Breunig, E M Knorr, E Keogh, Portnoy, J Takeuchi, M Agyemang, M Markou, V Chandola 等。国内的相关研究起步较晚但发展迅速,进行相关研究的科研院校主要有清华大学、西安交通大学、天津大学、复旦大学、香港科技大学等等。由于

时间序列异常检测的研究价值和应用前景,大量学者投身其中,近十年来著名的国际学术会议如 PAKDD、PKDD、SIGKDD、VLDB 以及期刊如 IEEE TKDE、Neural Computation、Computational Statistics and Data Analysis 等也都呈现了一些高水平的研究成果,但主要还是针对一元时间序列的异常检测。

异常检测作为数据挖掘领域的一个重要分支,正受到越来越多的关注和研究。国内外学者也提出了许多异常检测的方法,可以分为以下五类:基于统计的异常检测方法^[10]、基于聚类的异常检测方法^[11]、基于偏差的异常检测方法^[12-13]、基于距离的异常检测方法^[14]和基于密度的异常检测方法^[15]。

2.1 基于统计的异常检测方法

基于统计的异常检测方法是研究最早、最多的方法,认为如数据与给定统计分布或模型的显著性差异超过某一特定数值或范围即为异常^[16]。该方法分为两类:基于分布的方法和基于深度的方法。前者为数据集指定一个分布(如正态分布、泊松分布等),然后用一致性检验的方法发现异常。由于现实中大量的异常数据挖掘都是在多维空间中进行的,而基于分布的异常检测方法的绝大多数检验是针对于单个属性;另外,实际数据集的分布常常是未知的,并且也很难估计高维空间下的数据分布;后者将每个对象视为 n 维空间中的一个点,每个点对应有一个点集深度,认为异常很可能就存在于那些具有较低深度的数据对象中。该方法避免了基于分布方法中的数据分布拟合问题,对 UTS 的检测效果较好,但该方法检测时需要计算 n 维空间的凸闭包,计算复杂度较高,只适合于二、三维的低维数据,对于四维以上的大型数据集效率较低^[17]。

2.2 基于聚类的异常检测方法

基于聚类的异常检测方法就是直接或间接地利用已有的诸如 DBSCAN、 K -means、ROCK 等聚类算法对数据进行聚类^[18],将那些数据点少的类或不能被聚类的数据视为异常。该方法简单直观且可以利用大量已有的研究成果,但聚类分析与异常检测是有较大区别的,前者的目的在于寻找聚类的类别;后者在于发现异常的数据。异常检测只是聚类的“附属产品”,一般算法中也未对异常检测作特殊的优化,导致算法效率不高。而且多数情况下,异常的定义和检测标准是隐含的,无法在聚类过程中明确体现。

2.3 基于偏差的异常检测方法

基于偏差的异常检测方法主要分为三类:序列异常检测方法、OLAP 数据立方体方法和预测模型方法。

(1) 序列异常检测方法。Agrawal 等人于 1996 年提出 Sequential exception,即序列异常的概念。采用一种机制扫描数据集,将与相邻序列有明显偏差的数据

点视为异常,此算法计算复杂度与数据集大小呈线性关系,计算效率较高,但 Sequential exception 对异常存在的假设太过理想化,在概念上有缺陷,容易遗漏很多真正有实际意义的异常数据,对现实的复杂数据效果不佳。

(2) OLAP 数据立方体方法。Sarawagi 等在大规模多维数据中用数据立方体技术来确定异常区域^[19]。如一数据立方体的单元值显著不同于统计模型的期望值则该单元数值被视为异常。据此,分析人员可按数据的层次结构逐层进行钻取以找出发生异常的原因。此方法考虑了涉及一个单元所属的所有维的度量值中的变化以及隐藏在数据立方体集合分组操作后面的异常情况。但由于搜索空间很大,尤其当存在许多涉及多层概念层次的维的时候,人工探测非常困难^[20]。

(3) 预测模型方法。国内外许多学者采用 Bayesian 网络、ARMA、神经网络、支持向量机等模型对时间序列数据中未知的内质关系进行学习,而后建立预测模型,通过各时间点的预测值与实际值的偏离来判断异常^[21-22]。此方法对于变量维数较低的数据集效果较好,但当变量维数较高时,效果较差以至于训练过程不收敛,不适宜处理多元时间序列的异常检测问题。

2.4 基于距离的异常检测方法

基于距离的异常检测方法的基本思想是通过设定某种距离函数,计算数据空间中数据点之间的距离,当一个数据对象与其他对象存在较大距离时,将其视为异常。Knorr 等首先提出基于距离的异常检测方法^[23]。他们认为如果一数据集中,至少有 p 个对象与对象 o 的距离大于 d 则将 o 定义为 $DB(p, d)$ 异常。随后,此距离的概念被推广为 k 近邻距离, Ramaswamy 等计算所有对象的 k 近邻距离并从小到大排序,将距离最大的前 n 个对象视为异常^[24]。目前基于距离的异常检测方法主要有基于索引的算法、循环嵌套的算法和基于单元的算法^[25]。基于距离的异常检测方法综合了基于分布的思想,克服了基于分布的异常检测方法的主要缺陷,而且其易于实现与理解,被广泛研究和应用。但其本身也存在一些不足:首先,算法复杂性较高,不能兼顾适用数据集的数据规模和维数的可扩展性。基于索引的算法和循环嵌套算法的时间复杂度可达到 $O(mn^2)$, 基于单元的算法的时间复杂度为 $O(cm + n)$ 。其中, m 表示维数, n 表示数据集中的数据对象, c 表示单元数; 其次, Breunig 等也指出基于距离的方法在处理内部密度差异明显的数据集时存在缺陷,要么将密度稀疏的区域中的数据都判断为异常,要么无法发现某些异常。这主要是因为基于距离的方法是从全局角度考虑的,当数据集含有多种分布或数据集是由不同密度子集混合而成时,异常检测的效果就不好;另

外,基于距离的异常检测方法对参数 p 和 d 非常敏感,这就要求用户具有一定的专业领域知识以设置合理的参数,实际应用受到一定的限制。

2.5 基于密度的异常检测方法

上述的异常检测方法都有一个共同的不足:采用了一个全局的距离标准作为异常检测的依据。而实际上,异常往往是从局部的视角出发的,即某点异常是指该点与其相邻的聚类相对较远,因此完全采用全局距离是不适宜的。针对此问题, Breunig 等提出了基于密度的异常检测算法,其基本思想是通过比较对象与其领域对象的密度关系来检测异常,即引入局部异常因子 (Local Outlier Factor, LOF) 的概念,并认为异常并非二值属性而是一种度量,数据对象的 LOF 值越高则越有可能是异常。此后, Agyemang 等对该方法进行了改进,将对象最近的 k 个对象的最大距离作为 k -距离,提出用局部稀疏距离 (Local Sparsity Coefficient, LSC)^[26] 来检测异常,减少了计算复杂度。进一步地, Papadimitriou 等通过比较数据对象的 r -领域中所含的数据对象个数与其领域中的所有对象的 r -领域中所含的对象个数的平均值得到多粒度的偏差因子 (Multi-granularity DEviation Factor, MDEF)^[27], 并以其作为数据对象异常程度的度量,该方法无需直接计算数据点的密度,计算效率比 LOF 高。基于密度的异常观点比基于距离的异常观点更贴近 Hawkins 的异常定义,因此能够检测出基于距离异常算法所不能识别的一类异常数据—局部异常,即减少了当数据集含有多种分布或数据集是由不同密度子集混合而造成的检测错误,检测精度比较高^[28], 这点较之基于距离的方法有较大优势。但同时基于密度方法也存在一些不足:首先是其时间复杂度依然较高;其次是检测结果对异常因子阈值等参数的选择较为敏感且没有一种统一、简单、有效的参数确定办法。

3 典型方法分析

对于时间序列,它的一个重要特点是具有时间属性,即序列值之间具有严格的顺序,属有序数据。针对时间序列的特点,中外学者也进行了大量的异常检测研究,现列举一些有代表性的研究成果。

E. Keogh 等将一元时间序列符号化,通过符号检索出时间序列中差异最为显著的子时间序列^[29],算法简单且易于实现,但如何更好地描述原始数据也是值得思考的问题。

S. Sadik 等在文献[30]中将所有接收到的数据点视为全局环境 (global context), 将暂时封闭的数据点视为局部环境 (local context), 通过数据点相对全局和局部环境的偏离检测异常,提出一种数据流的自动异

常检测方法 A-ODDS。

C. Shahabi 等在文献[4]中提出了改进的 TSA-Tree 算法并通过小波系数的局部极大值实现了序列的异常模式查找,但其异常模式是建立于小波分解的基础上,因此会遗漏掉一些异常模式^[31]。

V. Chandola 在文献[32]中针对 MTS 的异常检测,利用子空间跟踪(Subspace monitoring)将 MTS 转换为一个 UTS,再利用一种 WINCsvm 方法实现异常检测,能同时兼顾 MTS 的多元和时序的特点,但其滑动窗口特征向量的计算是其计算效率进一步提高的瓶颈,当 MTS 规模较大时尤其如此。

利用 Voronoi 图的基本原理,提出了一种基于密度的异常检测方法并应用到一元时间序列的点异常和线性模式异常检测,算法复杂度降至 $O(n \log n)$,但对于多元时间序列的异常检测却并未涉及。

文献[33]基于 UTS 的 GMBR(Grid Minimum Bounding Rectangle)表示,用“异常特征值”来衡量时间序列模式的异常程度,提出了一种基于距离和密度的异常检测方法,但正如文中所述该方法的理论基础还有待进一步研究,并可向 MTS 异常检测的方向发展。

文献[34]采用扩展的 Frobenius 范数计算 2 个 MTS 子序列之间的相似性,通过 k -均值聚类得到模式的集合,然后计算每个模式的例外支持度和频率,从而检测出异常,取得了较好的应用效果。但如何进一步提高算法效率以适应 MTS 异常的在线识别还需进一步研究。

文献[35]提出一种两阶段的 MTS 异常检测算法。通过有界坐标系统技术计算 MTS 样本之间的相似性,后又采用基于距离的方法实现了异常检测,算法分两个阶段:第一阶段采用 K -means 算法聚类,并估计每个簇包含异常点的可能性;第二阶段在循环嵌套算法的基础上增加了剪枝规则以提高算法效率。但该方法仍是一种基于距离的异常检测方法,对由不同密度子集混合而成的时间序列数据集的检测效果不好。

文献[36]提出一种基于 KPCA 的 MTS 异常检测方法,通过核函数隐性地地将 MTS 数据映射到高维特征空间中,并采用 KPCA 方法获取数据的主成分方向矢量作为数据的特征表达,可用于不同的数据的异常检测。但该方法中子序列长度和核函数选择对检测结果影响较大,需进一步研究。

可见,时间序列的异常检测研究并不成熟,对于 MTS 的异常检测尤其如此,这主要是因为其面临如下难题:

(1)异常的度量问题。由于 MTS 数据的稀疏性,在传统的距离意义下,所有的对象都可能是异常。因

此,如何对多元时间序列的异常进行定义和度量就成为首要难题。

(2)数据类型复杂。MTS 类型的数据同时具有时序数据、变量维数高、变量相关性复杂、噪声干扰等特点,使得对其的异常检测更加困难。

(3)维灾的影响。维灾主要表现在两个方面:维数增大时,索引结构的修剪效率迅速下降,以至于到最后还不如顺序扫描;在高维空间中,最近邻的概念可能会失去意义,因为很多情况下,某点与其最近邻和最远邻几乎是等距的^[37]。这就给时间序列异常的快速检测带来了不小的难题。

(4)计算复杂度高。快速的异常检测算法大都依赖于索引结构和网格的划分,但由于在高维空间中索引结构有可能失效,网格划分的数目也随维数呈指数式的增长,使得算法时间复杂度增加,效率下降。

笔者认为解决多元时间序列异常检测的一种重要思路就是将 MTS 数据从多维空间映射到低维子空间,此种映射变换要保持空间聚集特性不变,在低维空间中就可再采用传统的异常检测方法检测异常,并可通过剪枝、过滤等方法以减少重复计算,进一步提高效率;另一种重要思路是可认为与绝大多数 MTS 序列最不相似的 MTS 序列为异常序列,因此可以结合多元时间序列相似性度量和索引查询的研究成果和具体的 MTS 异常的定义提出 MTS 的异常检测方法。

4 结束语

综上所述,在此提出时间序列异常检测的几点建议:

(1)异常是个相对概念。将异常看作一种二元特性,即要么正常,要么异常的认识是不准确的。

(2)不同领域,异常的概念和意义不同。例如一个人的身高为 568 cm 显然属异常,而一公司 CEO 月工资几十万,远远高于绝大多数员工的工资几千元,但不能说 CEO 的工资数据属异常。因此,异常产生的原因各异,异常检测算法所检测的是否为具有实际意义的真正异常只能由该领域专家来解释,算法仅是为用户提供可疑的数据以引起其注意。因此,从此角度上讲,异常检测也可认为是一种特殊条件和意义下的查询。

(3)MTS 的异常检测与传统的多变量数据和 UTS 的异常检测是有较大区别的。因为 MTS 具有多变量和有序数据的双重特点。通常多元时间序列的异常检测要通过各变量序列的综合分析才能检测出来。

总之,时间序列的异常检测研究并不成熟,尤其在多元时间序列的异常检测方面。此外,现行的异常检测算法的效率还不尽人意,因此如何进一步降低算法

复杂度以适应动态时间序列异常挖掘的需要有待进一步深入研究。

参考文献:

- [1] 周大镛,刘月芬,马文秀.时间序列异常检测[J].计算机工程与应用,2008,44(35):145-147.
- [2] 杨越,胡汉平,熊伟,等.一种基于超统计理论的非平稳时间序列异常点检测方法研究[J].计算机科学,2011,38(6):93-95.
- [3] Whitehead B, Hoyt W A. Function approximation approach to anomaly detection in propulsion system test data[J]. Journal of Propulsion and Power, 1995, 11(5):1074-1076.
- [4] Shahabi C, Tian X, Zhao W. TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries[C]//Proc of the 12th international conference on scientific and statistical database management. Washington: [s. n.], 2000:55-68.
- [5] Takeuchi J, Yamanishi K. A unifying framework for detecting outliers and change points from time series[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(4):482-492.
- [6] 于浩,王斌,肖刚,等.基于距离的不确定离群点检测[J].计算机研究与发展,2010,47(3):474-484.
- [7] Ma J, Perkins S. Online novelty detection on temporal sequences[C]//Proc of the international conference on knowledge discovery and data mining. Washington, DC: ACM Press, 2003.
- [8] Jagadish H V, Koudas N, Muthukrishnan S. Mining deviants in a time series database [C]//Proc of the 25th international conferences on very large data bases. Edinburgh: Morgan Kaufmann Publishers, 1999:102-113.
- [9] Hawkins D. Identification of outliers [M]. London: Chapman and Hall, 1980:20-35.
- [10] Johnson T, Kwok I, Ng R. Fast computation of 2-dimensional depth contours[C]//Proc of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM Press, 1998:224-228.
- [11] He Zengyou, Xu Xiaofei, Deng Shengchun. Discovering cluster based local outliers[J]. Pattern Recognition Letters, 2003, 24(9/10):1641-1650.
- [12] Arning A, Agrawal R, Ragaran P. A linear method for deviation detection in large database [C]//Proc of the 2nd international conference on knowledge discovery in databases and data mining. Portland, Oregon: Morgan Kaufmann Publishers, 1996:164-169.
- [13] Jagadish H V, Nick K, Muthukrishnan S. Mining deviants in a time series database [C]//Proc of 25th international conference on very large data bases. Edinburgh: Morgan Kaufmann Publishers, 1999:7-10.
- [14] Knorr E M, Ng R T, Tucakov V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal, 2000, 8(3):237-253.
- [15] Breunig M, Kriegel H P, Ng R, et al. LOF: identifying density-based local outliers [C]//Proc of the 2000 ACM SIGMOD international conference on management of data. Dallas, Texas: ACM Press, 2000:93-104.
- [16] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3):1-58.
- [17] Ruts I, Rousseeuw P. Computing depth contours of bivariate point clouds[J]. Computational Statistics and Data Analysis, 1996, 23(1):153-168.
- [18] Hardin J S, Rocke D M. Outlier detection in the multiple cluster Setting using the minimum covariance determinant estimator[J]. Computational Statistics and Data Analysis, 2004, 44(4):625-638.
- [19] Sarawagi S, Agrawal R, Megiddo N. Discovery driven exploration of OLAP data cubes [C]//Proc of the 6th international conference on extending database technology. Valencia: Springer Verlag, 1998:168-182.
- [20] 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.
- [21] Markou M, Singh S. Novelty detection: a review part2: neural network based approaches [J]. Signal Processing, 2003, 83(12):2499-2521.
- [22] 杨虎, 王会琦, 程代杰. 基于预测的序列异常数据挖掘[J]. 计算机科学, 2004, 31(4):117-119.
- [23] Knorr E M, Ng R T. A unified notion of outliers: properties and computation [C]//Proc of the 3rd international conference on knowledge discovery and data mining. Newport Beach, CA: AAAI Press, 1997:219-222.
- [24] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets [C]//Proc of the 2000 ACM SIGMOD international conference on management of data. New York: ACM Press, 2000:427-438.
- [25] 周大镛. 多变量时间序列的聚类、相似查询与异常检测 [D]. 天津: 天津大学, 2008.
- [26] Agyemang M, Ezeife C I. Lsc-mine: algorithm for mining local outliers [C]//Proc of the 15th international conference on information resource management association. New Orleans: [s. n.], 2004:5-8.
- [27] Papadimitriou S, Kitagawa H, Gibbons P B, et al. LOCI: fast outlier detection using the local correlation integral [C]//Proc of ICDE. [s. l.]: [s. n.], 2002.
- [28] 蒋盛益, 李庆华, 王卉, 等. 一种增强的局部异常挖掘方法[J]. 计算机研究与发展, 2005, 42(2):210-216.
- [29] Keogh E, Lin J, Fu A W, et al. Finding unusual medical time-series subsequences: algorithms and applications [J]. IEEE Transactions on Information Technology in Biomedicine, 2006, 10(3):429-439.
- [30] Sadik S, Gruenwald L. An adaptive outlier detection technique for data streams [C]//Proc of the SSDBM. Portland: [s. n.],

大数据环境下减少存储空间消耗和提高匹配速度。另外,基于留空 q -gram 的过滤算法具有错误容忍度高、匹配速度快等优势,是近似串匹配当前研究中的热点,也是未来的研究趋势。

参考文献:

- [1] Navarro G. A guided tour to approximate string matching[J]. ACM Computing Surveys, 2001, 33(1): 31–88.
- [2] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals[J]. Soviet Physics Doklady, 1966, 10(8): 707–710.
- [3] Burkhardt S. Filter algorithms for approximate string matching [D]. Saarland: Saarland University, 2002.
- [4] Navarro G, Baeza-Yates R, Sutinen E, et al. Indexing methods for approximate string matching[J]. IEEE Data Engineering Bulletin, 2001, 24(4): 19–27.
- [5] Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins [J]. Journal of Molecular Biology, 1970, 48(3): 443–453.
- [6] Smith T F, Waterman M S. Identification of common molecular subsequences [J]. Journal of Molecular Biology, 1981, 147(1): 195–197.
- [7] Altschul S F, Madden T L, Alejandro A S, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389–3402.
- [8] Pearson W R, Lipman D J. Improved tools for biological sequence comparison[J]. Proceedings of the National Academy of Sciences of the United States of America, 1988, 85(8): 2444–2448.
- [9] Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search[J]. Bioinformatics, 2002, 18(3): 440–445.
- [10] Giladi E, Walker M G, Wang J Z, et al. SST: an algorithm for finding near-exact sequence matches in time proportional to

the logarithm of the database size[J]. Bioinformatics, 2002, 18(6): 873–879.

- [11] Wu S, Manber U. Fast text searching allowing errors[J]. Communications of the ACM, 1992, 35(10): 83–91.
- [12] Chang Y I, Chen J R, Hsu M T. A hash trie filter method for approximate string matching in genomic databases [J]. Applied Intelligence, 2010, 33(1): 21–38.
- [13] Jokinen P, Ukkonen E. Two algorithms for approximate string matching in static texts[C]//Proceedings of the 16th international symposium on mathematical foundations of computer science. Berlin, Germany: Springer-Verlag, 1991: 240–248.
- [14] Burkhardt S, Crauser A, Ferragina P, et al. Q-gram based database searching using a suffix array[C]//Proceedings of the annual international conference on computational molecular biology. New York, USA: ACM, 1999: 77–83.
- [15] Rasmussen K R, Stoye J, Myers E W. Efficient q-gram filters for finding all epsilon-matches over a given length [J]. Journal of Computational Biology, 2006, 13(2): 296–308.
- [16] Sutinen E, Tarhio J. Filtration with q-samples in approximate string matching[C]//Proceedings of 7th annual symposium on combinatorial pattern matching. Berlin, Germany: Springer-Verlag, 1996: 50–63.
- [17] Navarro G, Baeza-Yates R. A hybrid indexing method for approximate string matching[J]. Journal of Discrete Algorithms, 2000, 1(1): 205–239.
- [18] Navarro G, Sutinen E, Tarhio J. Indexing text with approximate q-grams [C]//Proceedings of CPM'2000. Berlin, Germany: Springer, 2000: 350–363.
- [19] Sutinen E, Szpankowski W. On the collapse of the q-gram filtration[C]//Proceedings of international conferences on FUN with algorithms 1998. Waterloo, Canada: Carleton Scientific, 1998: 178–193.
- [20] Egidi L, Manzini G. Better spaced seeds using quadratic residues[J]. Journal of Computer and System Sciences, 2013, 79(7): 1144–1155.

(上接第 170 页)

- 2011: 596–597.
- [31] Keogh E, Lonardi S, Chiu W. Finding surprising patterns in a time series database in linear time and space[C]//Proc of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM Press, 2002: 550–556.
- [32] Chandola V. Anomaly detection for symbolic sequences and time series data [D]. Minnesota: The University of Minnesota, 2009.
- [33] 孙梅玉. 基于分形的非平稳时间序列挖掘关键技术研究

[D]. 上海: 东华大学, 2009.

- [34] 翁小清, 沈钧毅. 多变量时间序列例外模式的识别[J]. 模式识别与人工智能, 2007, 20(3): 336–342.
- [35] 王 欣. 两阶段的多元时间序列异常检测算法[J]. 计算机应用研究, 2011, 28(7): 2466–2469.
- [36] 李 权, 周兴社. 基于 KPCA 的多变量时间序列数据异常检测方法研究[J]. 计算机测量与控制, 2011, 19(4): 822–825.
- [37] Baragona R, Battaglia F. Outlier detection in multivariate time series by independent analysis[J]. Neural Computation, 2007, 19(7): 1962–1984.

基于异常检测的时间序列研究

作者：[陈运文](#)，[吴飞](#)，[吴庐山](#)，[刘博](#)，[CHEN Yun-wen](#)，[WU Fei](#)，[WU Lu-shan](#)，[LIU Bo](#)
作者单位：[上海工程技术大学 电子电气工程学院, 上海, 201620](#)
刊名：[计算机技术与发展](#)
英文刊名：[Computer Technology and Development](#)
年，卷(期)：2015(4)

引用本文格式：[陈运文](#). [吴飞](#). [吴庐山](#). [刘博](#). [CHEN Yun-wen](#). [WU Fei](#). [WU Lu-shan](#). [LIU Bo](#) [基于异常检测的时间序列研究](#)[期刊论文]-[计算机技术与发展](#) 2015(4)