计算机科学与探索
Journal of Frontiers of Computer Science and Technology
ISSN 1673-9418,CN 11-5602/TP

《计算机科学与探索》网络首发论文

题目：基于相关性的多维时序数据异常溯源方法
作者：王沐贤，丁小欧，王宏志，李建中
网络首发日期：2020-11-06
引用格式：王沐贤，丁小欧，王宏志，李建中．基于相关性的多维时序数据异常溯源方法．计算机科学与探索．
https://kns.cnki.net/kcms/detail/11.5602.tp.20201105.1343.022.html

计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2008100

基于相关性的多维时序数据异常溯源方法

王沐贤，丁小欧，王宏志+，李建中
哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150000
+ 通信作者 E-mail:wangzh@hit.edu.cn

摘 要：本文提出一种基于统计学相关性分析的多维时序异常数据检测分析方法，以对检测中表现为异常的数据进行溯源：对反映系统故障的数据和传感器质量问题的数据进行分类，进而识别出真正的系统故障，避免误检。首先根据相关关系构建时序相关图，再进一步归纳为时序相关环模型，通过在时序相关图上搜索并确定时序相关环的过程提取时序相关环中的特征，得到时间序列相关性集合。进而利用时间序列相关性集合进行时序数据异常来源检测，根据检测结果评估时序传感器数据对应的系统故障的几率。本文在真实的工业设备传感器序列数据集上进行大量实验，实验结果验证了本文方法在高维时序数据的异常检测任务上的有效性。通过对比实验，验证了本文方法从稳定性和效率上优于基于统计和基于机器学习模型的基准算法，时间序列的维度越高，本文方法较基准算法的提升越明显。本文方法通过对多维时序数据相关性知识的挖掘，既节约了计算成本，又实现了对多维异常数据来源的精准识别。

关键词：多维时间序列；异常检测；相关性分析；图算法；工业大数据；溯源
文献标志码：A **中图分类号：**TP18

王沐贤，丁小欧，王宏志，等. 基于相关性的多维时序数据异常溯源方法.计算机科学与探索
WANG M X, DING X O, WANG H Z, et al. Correlation-based method for tracing multi-dimensional time series data anomalies. Journal of Frontiers of Computer Science and Technology

Correlation-based method for tracing multi-dimensional time series data anomalies

WANG Muxian, DING Xiaoou, WANG Hongzhi+, LI Jianzhong
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China

Abstract: A multi-dimensional time series anomaly data detection method based on correlation analysis is proposed to classify the cause of anomaly detection: system failure data and sensor quality problem data, and then identify real system failures to avoid false detections. Firstly, the time series correlation graph model is proposed, which is further summarized as the time series correlation loop model. The time series correlation set is obtained by extracting the features in the time series correlation cycle, and the cause of abnormality is detected, and the system failure

*The National Key R&D Program of China under Grant No.2018YFB1004700 (国家重点研发计划); NSFC Grant Nos. U1866602, 61602129, 61772157 (国家自然科学基金).

is judged according to the result. Through a large number of experiments on real industrial data sets, the effectiveness of the method in the detection of abnormal sources of high-dimensional time series data is verified.

Key words: multi-dimensional time series; abnormal detection, correlation analysis; graph algorithm; industrial big data; provenance

1 引言

近年来我国制造业持续快速发展。工业互联网的智能工厂已经积累并正在产生大量的工业时序数据,通过对基于采集时间点的多维时间序列数据的分析和挖掘,能够实现对系统的运行状态进行控制、分析、决策和规划[1]，形成了有效的工业知识产生、提取、应用的积极循环,进而实现了对工业大数据的智能分析[2]。

通过分析传感器组传回数据可以检测出工业系统及设备中存在的显性或隐性异常，例如产品质量缺陷、精度缺失、设备故障、加工失效、性能下降、环境突变等[1]。这些故障给企业带来了大量消耗损失。为了降低危害，工业系统和设备会加入一套故障处理方案，由系统运维人员或自动运维程序对数据中显示的异常状态进行判断并介入故障排查[3]。但通过调研发现，在很长一段时间中，工业生产的故障诊断存在着以下一些问题：

（1）工业时序数据具有数据量大、时效性强、模式多样的特点。传统的单维静态数据处理方法存在一定局限性。（2）利用工业时序数据判别异常类型为系统故障抑或是错误数据存在困难。（3）工业时序数据可能存在模式相关，这是工业系统的物理属性决定的。但数据可能仅是数学上表现出相关性，可能在整体上并不能表现出相关关系。

为了能够减少工业生产中将数据异常错误归类造成损失，本研究提出了一种基于时序相关环模型的异常来源检测算法。本文的创新点包括：

1 给出区别于传统单维数据的基于图论的多维时序相关性模型，将相关性分析放在图中进行解释，在不失去严谨性的同时更加直观；

2 设计了一种在时序相关图中提取最大时序相关环的方法对异常成因进行溯源。利用序列间的相关关系得到相关序列集合，通过在相关序列中定量分析序列相关性对异常来源进行分类；

3 在实际的多维工业数据中，通过与基准算法进行比较，本文方法在准确率和召回率以及稳定性上高于基准算法。

本文主要分为 5 个部分：第一部分介绍相关研究并对要解决的问题做一个全面的解释；第二部分介绍研究方法的预备知识和基本概念；第三部分对设计的算法进行介绍；第四部分分析实验数据；第五部分总结该研究。

2 相关工作

静态数据的异常检测（Anomaly Detection）研究相对起步较早。现在静态异常检测已经被应用到网络入侵检测、工业探伤、星云探测等多学科多领域,文献[4]识别网络中流量的特定异常分布模式,以确认监控的计算机是否将敏感数据发送给未经授权的其他计算机。文献[5]利用心电图中的正常模式进行匹配，若失配则认为对应患者的心脏存在病变。这是利用已知数据中的特征或固有统计模型挖掘数据中不符合该特征或模型的点或片段检测异常。

利用机器学习的模型进行异常检测已有很多研究，其中基于分类和基于聚类的方法得到了广泛应用。基于多类别分类的异常检测技术假定训练数据包含属于多个正常类别的标记实例[6]。这种异常检测技术可以学习得到一个多维分类器，以区分每个正常分类与其余分类。如果一个测试实例没有被任何分类器归类为正常，则该测试实例被认为是异常的。基于二分类的异常检测技术假定所有训练实

例都只有一个类标签。这类方法有使用二分类 SVM 判别一个正常模式的边界[7]，也有利用 Fisher 核做判别式指导分类器划分的方法[8]。基于聚类的异常检测技术有：正常数据实例属于数据中的一个聚类，而异常实例不属于任何聚类。基于上述假设的技术将已知的基于聚类的算法应用于数据集，并将不属于任何聚类的任何数据实例声明为异常[9]。正常数据实例位于其最接近的聚簇质心附近，而异常则离其最接近的聚簇质心很远。基于上述假设的技术包括两个步骤。在第一步中，使用聚类算法对数据进行聚类。在第二步中，对于每个数据实例，将其到其最接近的群集质心的距离计算为其异常分数[10]。注意，如果数据形式中的异常本身是聚簇的，则上述技术将无法检测到此类异常。

近些年数据的时序属性不断受到重视，基于时序数据的异常检测方法也得到了很大发展。在时序异常检测研究中，对于异常的检测对象而言，时序数据异常主要有毛刺异常（glitch），点异常（abnormality）和区间异常（interval abnormality）三种[11]；在时序异常检测方法上以机器学习方法为主，包括基于聚类和基于分类器的算法。基于聚类的方法将正常或异常数据点聚集并尽可能将二者的距离增加[12]。基于分类器的方法利用确定特征方程的系数得到正常和异常数据的分界。文献[13]利用 EM 算法做正常数据与异常预测数据的多分类器。文献[14]利用隐马尔科夫方法发现偏序序列中各个正常点或正常子序列所具有的特征，从而将异常点或异常子序列检测并标记处理。文献[15]利用速度约束的概念，配合最大似然估计得到正常情形下的数据值，以此检测异常数据并进行修复。在已有的异常检测方法中，基于机器学习的方法往往开销较大，基于统计模型的方法要求待测数据的分布模式已知，应用存在局限。而约束方法在挖掘较长区间的异常模式时受到计算方法的限制不能很好地使用。

3 研究内容介绍

在一个工业系统中，异常（anomaly）一般被定义为数据中不满足常态、约束、规则、给定模型的不寻常数据值或模式[16]。工业时序数据异常的溯源可能有两种，一种是指工业系统丧失其规定性能的状态，我们称之为故障；还有一种是指传感器失灵造成正确数据出现偏差，称之为错误数据。二者都可以引发数据异常，但造成的后果完全不同。

3.1 基本定义

本文的问题定义基于已有研究文献[17]。下面给出一些便于理解本文算法的关键基本定义：

定义 1(时间序列)：时间序列是由传感器采样的一系列连续的数据点。一条长度为N的时间序列表示为$S=\langle s_1,s_2,\cdots,s_N\rangle$，其中每个序列点表示为一个二元组$s_i=\langle x_i,t_i\rangle$，$x_i$是一个实数值，$t_i$是时间记录点。对于任意的整数$i$、$j$，若$ij$，则有$t_i<t_j$。$T=\{t_i\}$，$i=1\cdots N$记作时间序列$S$的时间点集合。

定义 2(多维时间序列)：$\boldsymbol{S}$是一个包含K条具有相同时间点集合T的时间序列集合，记为$\boldsymbol{S}=\{S_1,S_2,\cdots,S_K\}$。$\boldsymbol{S}$称为$K$维时间序列。

定义3(相关系数矩阵)：在K维时间序列$\boldsymbol{S}$的(默认长度为n,下同)时间序列组中,第k个时间序列表示为$\mathbf{S}_k=\{\mathbf{s}_k(1),\ldots,\mathbf{s}_k(n)\}$。在这个时间序列时间段中,我们在公式(1)中定义相关系数矩阵,用于测量传感器组$\boldsymbol{S}$上第l时间段内K条序列的相关性,表示为SCM。

$$\mathrm{SCM}=\begin{vmatrix} R_{11} & \cdots & R_{1K} \\ \vdots & \ddots & \vdots \\ R_{K1} & \cdots & R_{KK} \end{vmatrix} \qquad (1)$$

定义 4（时序相关图）：设有图 $G=(V,E)$，$V=\{v\mid v\in S_K\}$，$E=\{e\mid e\in(R_{ij}=1)\}$，则图 G 被称为时序相关图。

3.2 方法概述

由研究背景所述，现有时序异常检测算法仅能输出发生了异常和异常的表征，无法对异常的来源

是故障还是错误数据进行判断。由此，我们首先对待检测时间序列组的线性相关关系进行挖掘，进而在图论的思想下设计异常检测算法，达到对异常数据的来源进行识别的目的。

本文的时间序列异常来源检测的步骤如图 1 所示，主要包含异常相关模型训练阶段（简称为训练阶段）和异常来源判别检测阶段（简称为检测阶段）。

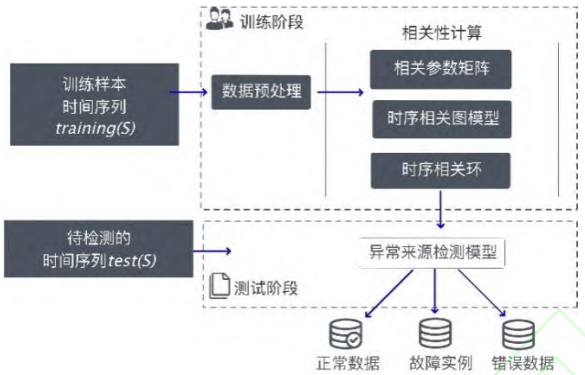


图 1　时间序列异常来源检测步骤示意图

Fig.1　The step from the cause of time series anomaly detection

训练阶段该阶段包含两部分，数据预处理和相关性计算。所以在数据预处理阶段需要对各序列数据的单位和量纲进行标准化。这里我们规定，训练阶段程序接收并分析的多维时序数据的标签标注都为正确，经调研表明工业生产的绝大多数时间产生的序列数据都是正常运转状态的，正确的工业多维时序数据获取难度不大。

在得到预处理的数据后，我们将在序列组内计算序列间的相关性关系，得到相关系数矩阵。之后将时序相关关系矩阵表达为一个时序相关图，在图中发掘出所有时序相关环并输出对应的相关序列集，完成对该多维时序数据的训练。

检测阶段该阶段将对输入的带有异常标签的相同类型序列组通过之前得到的相关序列集分析每个存在异常的序列的异常来源，输出异常的来源类型，以指导工厂对异常情况进行进一步处理。

4　多维时序相关性计算方法

在工业场景下的多维时间序列数据中，处于同一个系统或具有物理关系的序列间往往呈现出较强的相似性，这里我们定义序列相关性来定量计算序列之间的相似程度。基于多维序列相关性的异常来源检测方法的整个过程如图2所示：

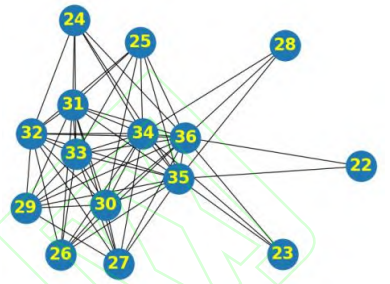


图 2　时序相关图某连通分量，点标号为对应序列号

Fig.2　A connected conponent of Time Series Correlation Graph

4.1　建立时序相关图

得到时序相关关系矩阵后，如果将传感器组的 K 维序列数据（也就是矩阵中的维度）作为节点，将序列间满足相关关系阈值的关系作为边，可以得到一个时序相关图 G。进而利用时序相关图做进一步的分析，把相关关系利用图模型联系起来。

4.2　构建时序相关环

在以每条序列为顶点，每两条相关序列间形成一条边的图 G 中，构成的连通图可能有一个或多个，这些连通图被称为图 G 中的连通分量。即在一个工业系统中，时间序列的相关关系可能存在着很多个。由时序相关图的概念，本文提出在时序相关图 G 的各连通分量中寻找最大时序相关环 C。某个时序相关图的其中一个连通分量的样例如图 2 所示。最大时序相关环的定义如下：

定义 5(最大时序相关环)：在时序相关图 G 的连通分量（Connected Component）$CC=(V,E)$，其中在 $|V|>2$ 中，若存在一条路径 $C=(V',E')$，使 $|V'|\le|V|(|V'|>2)$，$E'=\{e'|e'\in E$ 且 e' 首尾相接$\}$，且 $V-V'$ 的点中找不到首尾相接的路径或使已有路径的边增加，则路径 C 被称为**最大时序相关环**（Maximum Time Series Correlation Cycle）。

定理 1：每个时序相关图 *G* 的连通分量 *CC* 中至多存在一个最大时序相关环。

证明：假设 *CC* 中存在两个最大时序相关环 *C1*，*C2*。如果 C1、C2 存在重复路径，则 *C1*、*C2* 可以合并为一个更大的环 *C*，*C* 的顶点集合 $V = V1 \cup V2 - (V1 \cap V2)$，*C* 的边集合 $E = E1 \cup E2 - (E1 \cap E2)$；如果 *C1*，*C2* 没有相交路径，由题设，则 *C1*、*C2* 分属两个不同的连通分量，与前提在同一连通分量 *CC* 中不符合。即 *CC* 中不可能存在两个以上的最大时序相关环，得证。

可以看出在一个时序相关图 *G* 中可以找到若干个连通分量，在每个连通分量 *CC* 中至多存在一个最大时序相关环 *C*，它们包含的顶点总数的大小关系是 $|V(G)| \geq |V(\{CC\})| \geq |V(\{C\})|$。每个最大时序相关环 *C* 表示一组时间序列相关关系。

下面介绍最大时序相关环的搜索算法。由定理 1 可知，目标为找到一条在任意连通分量中经过每个点的路径[18]。据此本文提出在每个连通分量 *CC* 中搜索一条支撑树 T 的算法，将连通分量中的边集合 *CCEdge* 划分为：所有已知树边加入支撑树边集合 *TreeEdge* 以及所有已知非树边加入成环边集合 *FcEdge*，有 $CCEdge = TreeEdge \cup FcEdge$。在生成支撑树的过程中借鉴了最小生成树中的 prim 算法，最后得到的 T 表示成一个支撑树边的集合。如图 3 即为图 2 中时序相关图的连通分量生成的一棵支撑树，这棵支撑树只是其中的一种解，但同一连通分量生成的不同支撑树最后计算出的最大相关环是唯一的。这时引出定理 2，描述如何从一个支撑树找到一个环结构。

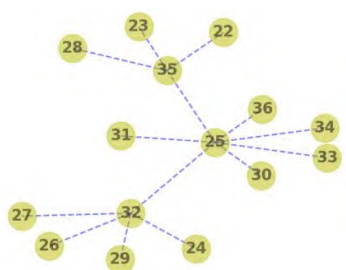


图 3　由图 2 连通分量生成的支撑树

Fig.3　Support Tree of the connected component in Fig.3

定理 2：一条属于 *FcEdge* 的边必与对应无向图连通分量的支撑树形成一个环

证明：设 $fcedge \in FcEdge$，则 *fcedge* 的两个顶点都在连通分量中，由连通图支撑树的定义可知，*fcedge* 的这两个顶点一定在支撑树的顶点集合中出现。又因为支撑树任意两点间必然存在一条路径，则一定存在一个环，以 *fcedge* 的一个顶点为起始点，沿支撑树中的一条路径到达 *fcedge* 的另一个顶点，再沿 *fcedge* 回到起始点形成一个环。得证。

下一步需要在支撑树中找到一条路径，该路径通过支撑树的任意两个叶节点和根节点。寻找该路径的步骤见算法 1。

算法 1：寻找支撑树中的最长路径

输入：一棵支撑树 T，根节点 root，叶节点集合 LeafNode，非树边集合 FcEdge

输出：该支撑树中包含的最长路径的顶点的集合 CycleNode

1 从 T 的 root 出发，搜索 root 的所有邻节点加入邻节点集合 NeighborNode;

2 对 NeighborNode 中的每一个节点 neighbornode，通过深度优先算法（DFS）找到 neighbornode 对应的深度搜索路径点集合 tmpLeafNode;

3 对每个路径点集合 tmpLeafNode，自 root 邻节点到叶节点由长到短依次对 tmpLeafNode 中的每个元素 tmpleafnode（路径点集合大小）排序;

4 按顺序每次取出最长的两条路径，判断两条路径各自的叶节点 leafnode1、leafnode2 是否在 FcEdge 中构成一条边，如果不构成则去除当前最长路径并从 tmpLeafNode 中加入一条在当前 tmpLeafNode 中最长的路径。重复该步骤直到满足上述判断即两个路径的叶节点构成边在非树边集合中存在;

5 在 CycleNode 中加入取出的两条路径中的所有节点和 root;

6 return CycleNode;

在找到支撑树中可以成环的最长路径后，对于还不属于这个环的其他顶点 *unfindnode*，需要对每个 *unfindnode* 进行遍历以确定是否可以将该点加入

环以形成一个更大的环，称为环的扩张。定理 3 在这里作为最大时序相关环的生成算法中的环扩张算法的一个理论依据。

定理 3：在支撑树中，若一个顶点不属于现有的环的顶点集合，且该顶点与环的顶点集合中至少两个顶点各自存在一条非树边，则该顶点加入环后可以将环的长度增加。

证明：设存在一棵支撑树 T 以及一个环顶点集合 *CycleNode*，一个不属于 *CycleNode* 的顶点 v。如果在 *CycleNode* 中存在着两个顶点 $c1$，$c2$，且 $(v, c1) \in FcEdge$，$(v, c2) \in FcEdge$。又因为存在 $(c1, c2) \in CycleNode$，则 v，$c1$，$c2$ 形成了一个三角形。由三角形两边之和大于第三边可知，v 加入环后环的长度增加，且环没有断裂，即环的结构是完整的。所以 v 的加入可以增加环的长度。

接下来用伪代码给出环扩张算法的实现步骤：

算法 2：环扩张算法

输入：已知环顶点集合 CycleNode，非环顶点集合 unFindNode，非树边集合 FcEdge

输出：加入了符合条件的新顶点后的环顶点集合 CycleNode

1 对任意非环顶点集合 unFindNode 中的顶点 unfindnode，重复步骤 2-3：

2 for c1, c2 in CycleNode：

3 if （unfindnode，c1），（unfindnode，c2） in FcEdge：then CycleNode.add(unfindnode):

4 return CycleNode;

示例说明：由上文中叙述的支撑树可以得到支撑树中由根节点的邻节点作为起始点，对应叶节点生成的路径，如例子可知其中的最长路径为 28-34-29-31-24，经过扩张算法后，*CycleNode*={23,24,25, 26,27,28,29,30,31,32,33,34,35,36}。为所求的时间序列相关集合 *CS*。

4.3 多维时间序列异常来源检测

在工业多维时间序列的数据分析中，由于单列异常检测无法很好区分故障和错误数据，本文提出了利用多维时间序列间相关性进行异常检测溯源，算法 3 即为该基于相关性的异常检测溯源算法。

算法 3：多维时序数据异常来源检测

输入：时序相关关系集合 CS，待检测 n 维时间序列数据 DTSD（n > 2）

输出：n 维时间序列数据异常来源标签数组 DTSDA

1 利用 Cycle 中包含的时序相关集合 CS 划分 DTSD 中序列为 m 个时序相关关系;

2 在每个时间序列相关关系 cs 包含的时间序列组 S：{S_1，S_2，…，S_K}中，重复步骤 3-4:

3if □Si，i∈K 存在异常，且相关关系超过阈值：then S 的所有序列 *S1-SK* 对应的异常来源标签数组 DTSTA 标为“故障”;

4 else if 存在 S_i 为异常序列且相关的其他序列无异常情形：then *Si* 对应的异常来源标签 DTSTA[i]标为“错误数据”;

5 return DTSDA;

4.4 算法效率分析

本段对第三章的上述算法进行效率分析，上述算法的主要的时间和空间开销产生于创建时序相关图和寻找最大时序相关环等两个步骤。

4.4.1 创建时序相关图

创建时序相关图时，设计算的时间序列长度为 n，序列数量为 K。则时序相关图的计算需要计算 K 维序列中每两列的相关性，在计算的过程中对序列中的每个点有一次遍历，总的时间复杂度为 $O(n*K^2)$

4.4.2 寻找最大时序相关环

设一个 K 个顶点的时序相关图 G 中可以划分出 N 个连通分量 CC，在每个连通分量中利用支撑树搜索算法后再使用最大时序相关环搜索算法。算法的快慢与连通分量划分相关，如果每个连通分量的顶点数都接近 K/N，则支撑树搜索算法的复杂度约为 $o(K^2/N)$；如果只有一个连通分量即 G 的 K 个顶点构成一个连通分量，则支撑树搜索算法的复杂度约为 $O(K^2)$。在每棵支撑树中搜索一个最大相关环的步骤中，在极限条件下即 G 的 K 个顶点构成一

个连通分量，时间复杂度约为 O（K^3）。实际工业系统中，一个系统往往 K 的值较小；而 K 值较大的系统又往往可以划分成多个内部相关性较强的子系统，所以时间复杂度的规模一般可以接受。

5 实验与结果

5.1 度量标准

本文的目的是追溯并区分异常序列的异常来源，即该异常属于系统故障还是错误数据。将系统故障作为正例，错误数据作为反例。设每个时间序列组为一个实例单位，实验结果可以分为 4 类，分别是：

1 预测正，实际正（True Positives，简称 TP）。

2 预测正，实际负（False Positives，简称 FP）。

3 预测负，实际正（False Negatives，简称 FN）。

4 预测负，实际负（True Negatives，简称 TN）。

利用以上 4 个分类给出计算公式：

准确率公式：$Precision = TP / (TP + FP)$

召回率公式：$Recall = TP / (TP + FN)$

实际上计算时还会出现算法判断不存在错误数据的情况，该情况是利用的单维时序异常检测算法造成，在计算准确率和召回率时不加入计算。

5.2 数据集

本文应用国内某大型火力发电厂的某发电机组数据进行实验。我们在研究了电厂发电机组的某型号引风机连续 6 个月的数据后，将其中连续某三个月的历史采样数据作为训练集。该设备共有 84 个传感器，分别检测引风机的轴应力、轴温、腔内气体温度、绕组线圈电流及温度等。传感器每 8 秒记录一次数据，一共记录了 84 列数据。每个列时间序列数据经预处理方法去除无效或低质量数据后，最终采用 37 列总计 74 万个时间点上数据进行实验。

5.3 对照算法

在实验中实现了本文提出的上述算法。为突出该算法与其他算法在异常来源检测上的上佳表现，本文采用两种算法与该算法做对照，进行算法性能对照：

1 基于约束的异常检测方法。通过检测异常值的波动检测异常并对故障或错误数据进行分类。

2 基于聚类的异常检测算法。通过提取时间序列中的特征区分故障或错误数据。即对潜在相关序列进行聚类，找到聚类中心和搜索半径并确定每条序列所在的分类。

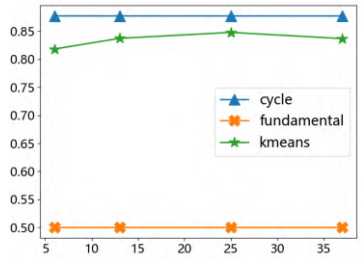
5.4 方法有效性分析

本文分别对测试序列组的序列维数总数、测试集规模和训练集规模对上述三种算法检测性能进行了测试，测试结果如下：

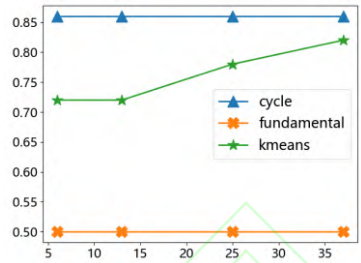
5.4.1 序列维数的影响

实验测试了序列维数从 6 列提高到 37 列的过程中，本文提出的最大相关环算法（cycle）和两种对比算法（单列检测算法 fundamental 和 k 聚类算法 kmeans）的准确率和召回率。由图 4(a)、4(b)（横轴为序列维数，纵轴 a 为准确率、b 为召回率）可以看出虽然在序列存在 kmeans 算法的准确率有较大上升，且伴随着召回率的上升。但本文提出算法的准确率、召回率以及不同维度下的稳定性都略高于 kmeans 算法，其中准确率保持在 87%左右，较 kmeans 算法提升 9%左右；召回率保持在 86%左右，平均较 kmeans 算法提升 11%左右。kmeans 算法的召回率能够随着维度上升有所提高是因为序列维度升高使数据量增加，减少了 kmeans 算法陷入局部最优解的可能。但是因为本文提出算法较 kmeans 算法多利用了序列间相关性这一特征，使该算法较 kmeans 的准确率和召回率都略高，也更加稳定。这里要说明一下，单列

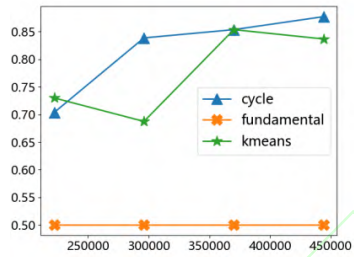
检测算法几乎无法判断出异常的来源，所以后续的测试将不再单独介绍该算法的效率。



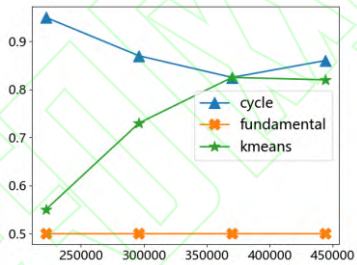
(a)Position rates comparison by dimension
(a)不同维度准确率比较图



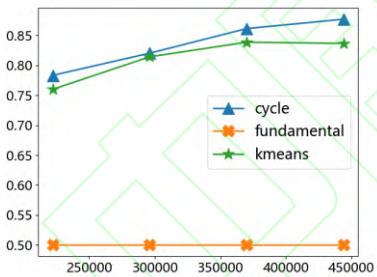
(b) Recall rates comparison by dimension
(b)不同维度召回率比较



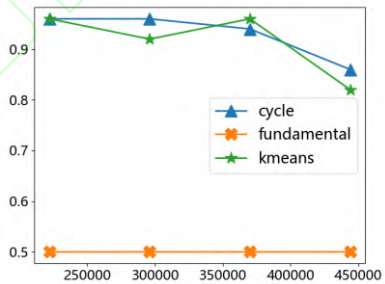
(c)Position rates comparison bysizeoftestset
(c)不同测试集规模准确率比较



(d) Recall rates comparison by size oftestset
(d)不同测试集规模召回率比较



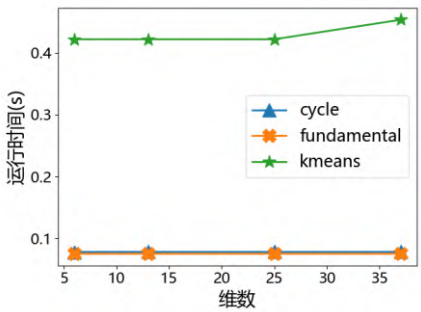
(e)Position rates comparison bysizeoftrainingset
(e)不同训练集规模准确率比较



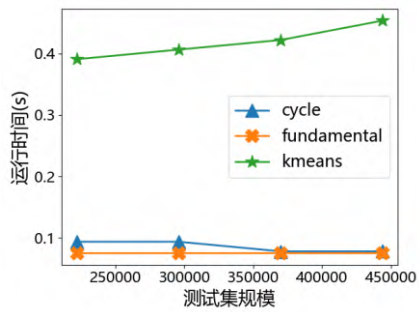
(f) Recall rates comparison by size oftrainingset
(f)不同训练集规模召回率比较

图 4　实验有效性分析图

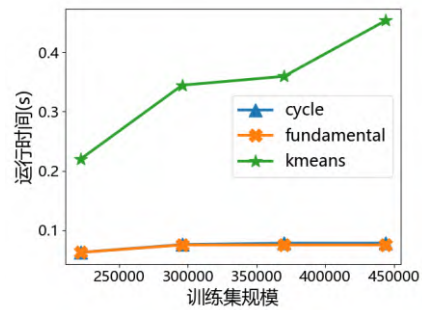
Fig.4　Effectivenessanalysisinexperiment



(a) Dimension comparisonbytime
(a)不同维数运行时间比较



(b) Testset comparisonbytime
(b)不同测试集规模运行时间比较



(c) Trainingset comparison bytime
(c)不同训练集规模运行时间比较

图 5　实验效率分析图

Fig.5　Efficiencyanalysisinexperiment

5.4.2　测试集规模的影响

在这里我们实验了不同测试集规模对上述算法

性能的影响。测试结果如图 4(c)、4(d)（横轴为测试集规模，纵轴 c 为准确率、d 为召回率）所示。从图中可以看出，随着测试集规模的增加，最大相关环算法和 kmeans 算法的准确率都有所提升，在某些测试集条件下 kmeans 的准确率可以达到甚至超过本文提出算法,但稳定性仍存在不足。而最大相关环算法的召回率则稳定优于 kmeans 算法。测试集规模增大后，本文提出算法的基于半监督的特性使该方法的准确率仍较优于 kmeans 算法。

5.4.3 训练集规模的影响

不同训练集规模对算法准确率和召回率的影响比较如图 4(e)、4(f)（横轴为训练集规模，纵轴 e 为准确率、f 为召回率）所示。可见随着训练集规模的提升，本文提出算法和 kmeans 算法的准确率都有所上升，且本文提出算法的准确率一直较 kmeans 算法高出 3%至 5%。而随着训练集规模的上升，二者的召回率都有所下降。但从两图仍可以看出，在训练集的规模增加或减少时，本文提出算法的稳定性都优于 kmeans 算法。

5.4.4 算法速率比较

如图 5(a)、5(b)、5(c)是上述变量的场景下运行时间的比较。如图可以看出本文提出算法的运行时间远小于 kmeans 算法。虽然本文算法在训练阶段需要一些时间，但利用训练步骤得到的环顶点集合可以使异常来源检测算法在 0.1s 内输出结果，该算法因为不需要在每次检测时对原数据进行额外计算而在流数据处理上较 kmeans 算法有更大优势。

6 总结

本文研究了基于统计相关性方法的异常检测问题，提出了解决异常来源检测问题的框架结构，利用实际遇到的问题进行示例分析，分别介绍了多维时间序列相关性计算算法和多维时序相关性的最大时序相关环算法以及基于前两种算法的多维时序异常来源检测算法。本文在实验部分说明了该方法的稳定性、运行速度和性能都较传统的朴素基于机器学习的异常检测算法有所提高。

References:

[1] Hao S., Li G., Feng J. et.al., Survey of structured data cleaning Methods[J]. J Tsinghua Univ Sci & Technol. Vol.58, No.12, 2018:1037-1050

[2] M.Gupta, J.Gao, C.Aggarwal, J.Han. Outlier Detection for Temporal Data[J]. IEEE Transactions on Knowledge & Data Engineering. VOL.25. 2014:1

[3] A.Chalamalla, I.Ilyas, M.Ouzzani. Descriptive and prescriptive data cleaning[J]. ACM SIGMOD International Conference on Management of Data. 2014: 445-456

[4] Kumar, V. Parallel and distributed computing for cybersecurity[J]. Distributed Systems Online, IEEE, 2005; 6-10

[5] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, et.al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals[J]. Circulation, 2000, 101(23): 215-220.

[6] Stefano, C., Sansone, C., and Vento, M.. To reject or not to reject: that is the question-an answer in case of neural classifiers[J]. IEEE Transactions on Systems, Management and Cybernetics, Vol.30, No.1. 2000. 84-94

[7] SchÄolkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution[J]. Neural Comput. Vol.13, No.7. 2001. 1443-1471

[8] Roth, V. Kernel Fisher discriminants for outlier detection[J]. Neural Computation, Vol.18, No.4. 2006. 942-960

[9] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise[J]. Proceedings of Second International Conference on Knowledge Discovery and Data Mining. 1996. 226-231

[10] Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski, B. Clustering approaches for anomaly based intrusion detection[J]. Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. 2002. 579-584

[11] Wang.X., Wang.C. Time Series Data Cleaning: A Survey[J]. IEEE Access, Vol8. 2019. 1866-1881

[12] Hartigan, J. A. and Wong, M. A. A k-means clustering algorithm[J]. Applied Statistics 28. 1999. 100-108

[13] R.Fujimaki，T.Nakata，H.Tsukahara，A.Sato，K.Yamanishi. Mining Abnormal Patterns from Heterogeneous Time-Series with Irrelevant Features for Fault Event Detection[J]. Statistical Analysis and Data Mining. 2009. 1-17

[14] Y.Qiao，X.Xin，Y.Bin，S.Ge. Anomaly Intrusion Detection Method based on HMM[J]. Electronics Letters. 2002, VOL.38. 663-664

[15] S.Song, A.Zhang, J.Wang. SCREEN: Stream Data Cleaning under Speed Constraints[C]. ACM SIGMOD International Conference on Management of Data. 2015. 827-841

[16] V.Chandola，A.Banerjee，V.Kumar. Anomaly Detection:A Survey[J]. ACM Computing Surveys. 2009 : 41

[17] X.Ding, S.Yu, M.Wang. Anomaly detection on industrial time series based on correlation analysis[J]. Journal of Software, 2019 (in Chinese). http://www.jos.org.cn/1000-9825/0000.htm

[18] P. Chaudhuri. A Self-Stabilizing Algorithm for Detecting Fundiamental Cycles in a Graph[J]. Journal of Computer and System Sciences, Vol.59. 1999. 84-93.

WANG Muxian is currently working toward the master degree in the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. As a CCF student member, his research interests include data cleaning, anomaly detection and IoT database system.
王沐贤(1997-),男,硕士生,CCF 学生会员,主要研究领域为数据清洗,异常检测,时序数据库系统。

DING Xiaoou is currently working toward the PhD degree in the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. She is a student member of CCF. Her research interests include data cleaning, temporal data quality management, and temporal data mining and analysis. She is also interested in IoT data cleaning and anomaly detection.
丁小欧(1993-),女,博士生,CCF 学生会员,主要研究领域为数据质量,数据清洗,时序数据挖掘,异常检测。

WANG Hongzhi received the PhD degree in computer science from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2008. Since 2015, he has been a professor and doctoral supervisor of the Department of Computer Science and Technology. His research interests include big data management, data quality, graph data management, and Web data management. He is an advanced member and a recipient of the Outstanding Dissertation Award of CCF.
王宏志(1978–),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,大数据,数据质量。

LI Jianzhong received the BS degree from Heilongjiang University, Harbin, China in 1975. Since 1998, he has been a professor and doctoral supervisor with the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His current research interests include database management systems, data warehousing and data mining, wireless sensor network, and data intensive super computing.
李建中(1950-),男,博士,教授,博士生导师,主要研究领域为海量数据管理与计算,无线传感器网络,数据质量