# 机械监测大数据清洗出现错误数据的方法

Xuefang Xu , Yaguo Lei , Member, IEEE, and Zeda Li

## 概要：

错误数据的存在导致状态监测的大数据质量下降。于是，通过分析这些质量较差的数据可能会获得不可靠或误导性的结果。在本文中，为了提高数据质量，提出了一种基于改进的局部离群因子（LOF）的数据错误检测方法。首先，使用滑动窗口技术将数据划分为不同的段。这些段被认为是不同的对象，它们的属性由从每个段提取的时域统计特征组成，例如平均值，最大值和峰间值。其次，使用这些属性计算基于内核的LOF（KLOF），以评估每个段作为不正确数据的程度。第三，根据这些KLOF值和阈值，就能够检测到错误的数据。

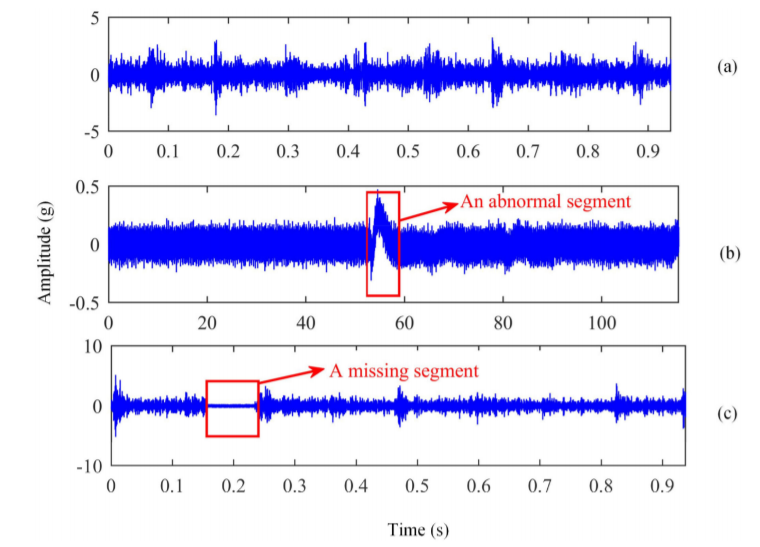
## 索引词：

状态检测大数据，数据清洗，数据质量，错误数据，局部离群因子（LOF）。

## 介绍

随着从监控系统中收集海量数据，机械状态监控已进入大数据时代[1]。作为状态监测的重要扩展，可以通过使用获得的状态监测大数据来检测，定位和识别机械中的缺陷来进行故障诊断[2]。监视大数据具有五个属性，包括高容量，高多样性，高速度，低准确性和低值密度[3]，[4]。结果，基于信号处理技术的传统故障诊断方法显示出它们在处理状态监测大数据方面的弱点。为了解决这个问题，许多研究人员非常重视智能故障诊断，它能够更快，更有效地处理海量数据。特别是，广泛研究了一些基于深度学习的高级智能诊断方法。例如，基于深度信念网络[5]，递归神经网络[6]和卷积神经网络[7]提出了不同的智能诊断方法。从状态监视数据中更有效地提取故障特征。尽管这些智能诊断方法在提取故障特征方面具有许多优势，但是这些方法在处理大数据方面仍然存在许多关键和开放的问题，包括处理大规模数据，处理异构类型的数据，高速处理流数据，以及处理劣质数据[4]。对于大规模数据的处理问题，在分布式并行框架中执行智能故障诊断可能是解决该问题的一种有效的方法。例如，在Hadoop平台下，可以将海量数据分成小块，然后可以使用MapReduce以分布式和并行方式处理这些块。此外，在创建智能故障诊断方法之前，可以将某些数据集成方法应用于处理异构类型的数据。为了满足高速传输数据的需求，可以使用一些在线智能故障诊断方法。无疑，与这三个问题相比，数据质量对智能故障诊断的性能影响最大。因此，本文重点讨论了处理低质量数据的问题，而其他三个问题没有考虑在内。

数据质量反映了数据与机器健康状况之间的相关程度，通常用四个维度来描述，包括准确性，完整性，一致性和及时性[8]。其中，准确性衡量收集的数据与健康状况之间的一致性程度。完整性衡量收集的数据是否包含整个健康状况信息。一致性表示数据在格式和结构方面的一致性程度。及时性衡量收集的数据是否最新[9]。低时效性数据通常会导致结果过时，无法代表当前的健康状况。自然，质量差的数据表明这些数据是不准确，不确定，不完整，



**图一. (a)从有内圈故障的轴承获得的正确数据，**

**(b)有内圈故障的低质量数据，(c)有缺失段的低质量数据。**

时间短或多格式的。结果，智能诊断方法很难从这些质量差的数据中提取有用的信息。此外，由于“输入的是垃圾，输出的也是垃圾” [10]，使用劣质数据创建的智能诊断模型可能会导致诊断结果不可靠或产生误导。因此，有必要在将数据用作创建智能诊断模型的训练数据之前提高数据质量。机械状态监测数据质量的下降归因于以下四个原因：恶劣环境中的干扰[11]，数据获取设备故障[12]，数据传输故障以及由于各种格式而导致的数据多样性[13]。特别是，由于前三个原因，会出现许多不正确的数据，从而导致数据质量的严重下降。因此，应该检测这些不正确的数据，然后将其清除以提高数据质量。但是，关于机械状态监测的数据清理或不正确数据检测的文献很少。此外，其他领域的数据清理方法可能不适合状态监视。原因之一是不同字段之间错误数据的类型不同。例如，重复值被视为错误数据或脏数据，应在许多字段中清除[14]。众所周知，当旋转机械发生故障时，重复的瞬变是常见的组成部分，因此重复值不能视为状态监测的不正确数据。各个字段的数据都有自己的区别特征。难以产生一个通用的可行的过程来检测错误的数据。

本文的目的是基于错误数据的特征，通过将它们视为异常值来检测错误数据，以解决大数据的质量问题。具体来说，高质量数据的特征被视为上下文属性[15]。与这些高质量数据相比，不正确的数据不符合高质量数据的模式。例如，它们的属性不再是确定性的，而是服从某些随机或可能性分布[4]，与上下文属性有很大不同。因此，不正确的数据可以视为上下文异常值。为了进一步解释这一点，从轴承和齿轮箱中收集了振动数据，这是机械的关键组成部分[16]。图1（a）显示了从具有内圈缺陷的轴承收集的高质量数据。这些数据表现出一些共同的特征，例如，长时间内在看起来都是连续的，并且它们的幅度在合理范围内[11]。此外，可以观察到由缺陷引起的周期性脉冲。这些特征表明，振动数据与被监测轴承的健康状况密切相关。如图1（b）所示，由不稳定电流的干扰而从变速箱获取的数据，作为一种典型的不正确数据，数据中出现了异常段。异常段的预期平均值不为零，并且此异常段似乎与其他段明显不同。产生这种差异的原因是异常段的生成机制与其他段的生成机制不同。通常，此异常段是由振动传感器的饱和或干扰引起的。图1（c）显示了另一个错误数据，在从轴承收集的数据中模拟了一个明显的缺失部分，这破坏了数据的连续性。实际上，由于网络数据传输丢失或电缆故障，通常通过将破坏数据完整性的实际数据值替换为0，NaN，NULL值或环境噪声[17]。可以看出，这些错误数据的某些特征与高质量数据的特征有很大的不同，并且与大量高质量数据相比，这些错误数据可以被认为是离群值。因此，要检测不正确的数据，一个简单的想法就是可以使用异常值检测方法。离群检测[19]，被大量使用，也被称为异常检测，在许多领域得到了广泛的研究，例如信用卡欺诈检测，网络安全，流量和医疗领域[20]。上述领域中的多种离群检测方法以及有人提出了。例如，基于统计的方法假设数据的分布或概率模型，然后将低概率区域中的点确定为异常值[21]。该方法需要有关数据集某些参数的知识[22]，不适用于多变量数据的异常检测。作为一种无监督的方法，可以使用基于聚类的方法，例如支持向量域描述[23]，具有噪声的基于密度的应用程序空间聚类[24]和*k-*均值[25]。检测与聚类相对的离群值。由于聚类方法的主要目标是寻找聚类[26]，因此基于聚类的方法可能无法检测到异常值。基于距离的方法计算每个对象与远离大多数对象的对象之间的距离，这些值被视为离群值[27]，[28]。此方法可能无法检测到来自不同密度区域的异常值。基于密度的方法能够通过将数据点周围的密度与其本地邻居的密度进行比较来有效地检测异常值。作为基于经典密度的方法，在[29]中首次提出了局部离群因子（LOF）方法。通过给出不受监督的异常值的程度，LOF方法能够识别其他方法可能无法做到的异常值。因此，对LOF方法的研究引起了很多关注，并且开发了LOF算法的许多变体，例如GridLOF [30]和基于局部距离的离群因子[31]。

从上面的陈述中，应该使用LOF方法通过将错误数据视为异常值来检测错误数据。但是，不正确的数据包含许多异常值，而不是一个或几个异常值，这些异常值可能属于一类，因此无法直接使用LOF方法检测到这些异常值。因此，本文提出了一种基于改进LOF的错误数据检测方法。在所提出的方法中，基于从每个分段提取的时域统计特征，使用改进的LOF定性地评估了由滑动窗口划分的分段的不正确程度。使用所提出的方法可以有效地从收集的数据中检测出不正确的数据，并分别通过仿真和实验进行了验证。

基于以上陈述，本文的贡献总结如下。

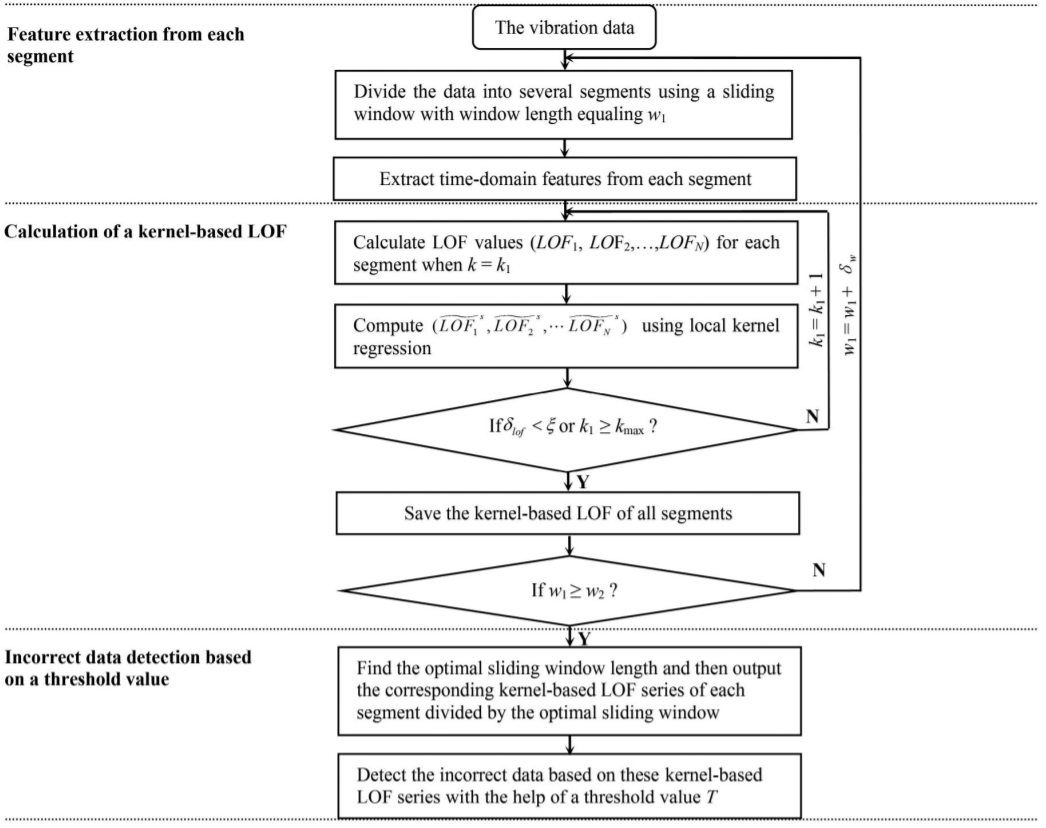
1. 最初说明了机械状态监控领域的数据质量概念。此外，讨论了数据质量下降的原因和错误数据的特征。另外，基于这些讨论，提出了一种不正确的数据检测方法，用于对机械状态监测的大数据清洗。
2. 该提议的方法扩展了LOF的应用，因为它很少应用于状态监视数据或异常段检测。此外，基于内核的LOF（KLOF）通过克服LOF的两个缺点，可以更准确地检测错误数据。通过考虑参数的不同值构造KLOFķ，还可以提高错误数据检测的准确性。
3. 可以使用KLOF值来评估不正确的数据程度，该值可以对不正确的数据进行定量描述。构造阈值以选择不正确数据的程度。

本文的其余部分安排如下。第二节详细介绍了提出的错误数据检测方法。第三节专门通过从故障轴承的模拟数据中检测缺失的段来验证所提出方法的有效性。在第四节中，用固定轴变速箱，风力涡轮机和行星齿轮箱收集的真实数据来证明所提出方法的性能。最后，第五节总结了本文。

## 方法

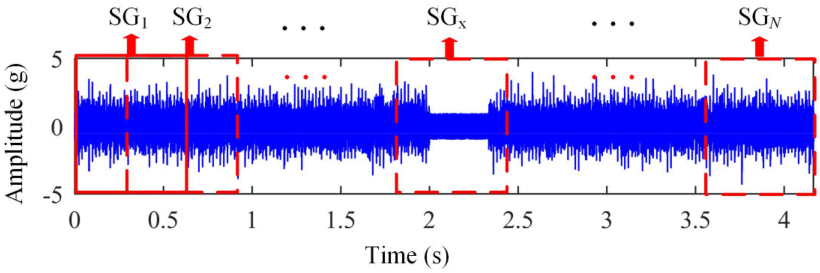
在本节中，将详细说明所提出的方法。首先，通过滑动窗口将数据划分为几个片段，然后从每个片段中提取时域统计特征。其次，基于这些特征计算KLOF。第三，借助于大于阈值的KLOF值检测到不正确的数据。以下三个部分详细介绍了每个步骤的实现和说明。

基于以上陈述，所提出的方法的流程图如图2所示。



**图二.方法的流程图**

### 从每个片段中提取特征



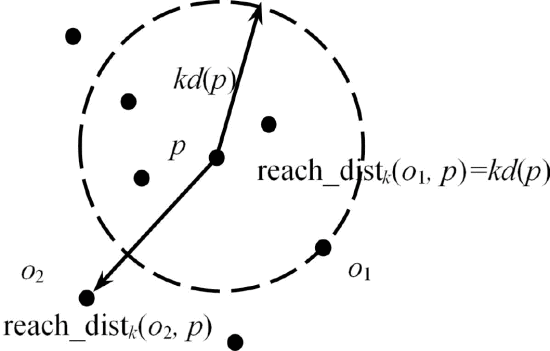
**图3.使用滑动窗口的数据片段**

如图3所示，首先，滑动矩形窗口的长度为ws，用于将整个数据划分为几个段，从而允许相邻窗口之间有90％的重叠。使用滑动窗口将数据转换成几个段。窗长不同w1个 至 w2 步长 δw 考虑选择最佳窗口长度 wØ。从w1到w2以δw为步长的窗口长度看作是最优的选择的窗口长度wo . 初始窗口长度w1和δw都取500，在所有数据长度范围内，取决于2s到2.333s的w2为

w2 = l/kmin (1)

其中l代表全部数据的长度，kmin代表KLOF中参数k的最小迭代值。

然后，时域统计特征包括均值，最大值，最小值，峰-峰值，波峰因数，方差，均方根，平均振幅，根振幅，形状因数，冲动因数，清除因数，峰度，峰和偏度[32]可以轻松地从每个细分中提取出来以计算以下KLOF。这些时域统计特征已广泛用于机械状态监测，并且对于描述不同方面的振动数据的特征都很重要。例如，可以使用间隙因子，形状因子，脉冲因子等来评估时域数据的时间序列分布。其他一些特征（例如，根振幅，均方根和峰值）可以用来反映振动振幅和数据的能量。由于振动数据总是很复杂，因此很难在错误和正确数据之间选择明显不同的特征。此外，某些数据段的每个特征中可能存在很小的差异，但是考虑所有特征后，数据段的差异可能非常大。因此，错误选择功能或使用较少功能可能会导致性能不佳。总之，在计算KLOF时应考虑所有特征，以实现良好的检测性能。



**图4。当*k* = 4 时，*o*1和*o*2到对象*p的*可达距离。**

### KLOF的计算

所收集的数据由许多部分组成，这些部分由一个滑动窗口划分，可以视为一个数据集，每个部分都被视为一个对象。从每个线段提取的各种特征均视为其对象。基于这些属性，可以为每个段计算LOF值。传统的LOF如下所示。

第一步：首先，对于每个o ∈ D计算对象p的距离k，其中D表示数据集，kd(p)表示p到o的距离k。*kd*（*p*）满足以下两个条件。首先，至少有*k个*对象o‘满足d(p, o‘) ≤ d(p, o) and o‘ ∈ D\{p}。其次，最多有k-1个对象o‘满足g d(p, o‘) ≤ d(p, o) and o‘ ∈ D\{p}。d(p, o)是对象和对象o之间的距离，可由下列公式计算得出：

 （2）

其中fi(p) 和 fi(o)是p和o的不同的特征值，特别的，dim代表的是特征的维度。

第二步：k的最相邻的对象p可以有该公式得到：

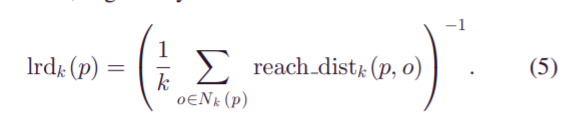
Nk (p) = {q |d(p, q) ≤ kd(p)}. （3）

第三步：对象p到对象o的可达距离用下列公式得到：

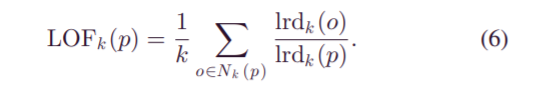
reach distk (p, o) = max{kd(o), d(p, o)} （4）

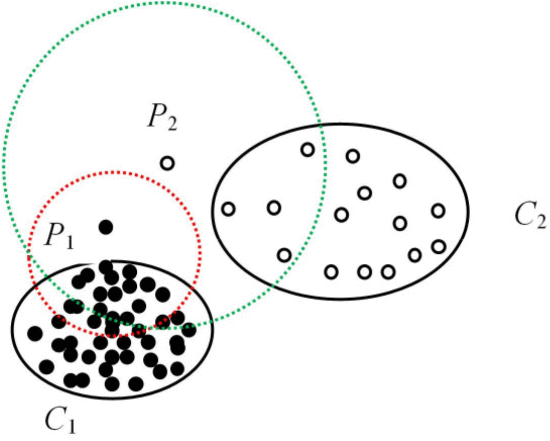
图4显示了当k=4时对象o1和o2与对象p之间的可达距离。可以看出当对象靠近p时，从一个对象到p之间的可达距离p等于它们的实际距离，而当这个对象靠近p时，可达距离变成了第k个距离p。利用这一可达距离，接近到o的距离的对象的统计学波动将会被减小，平滑效果的强度在很大程度上取决于参数k。具体来说，k越大，同一邻域内的对象的可达距离越相似[29]。

第四步：根据上述定义，代表k的平均可达距离的局部可达密度p(lrd)可以由以下公式得到：



第五步：最后，计算出对象p的LOF来评估p的离群值程度。LOF是p的lrd与p的k个近邻的lrd之比的平均值，公式如下：



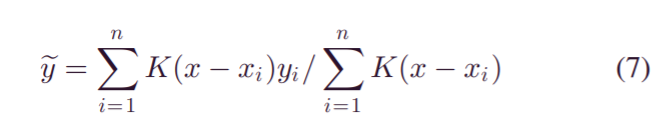


**图五.数据集分布：∙：簇*C*1中的点，○：簇*C*2中的点。**

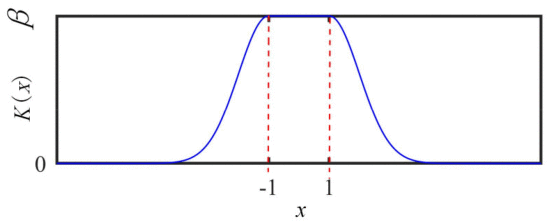
尽管可以使用LOF评估离群值，但是LOF还有两个缺点，这可能导致对离群值的错误评估。首先，LOF可能无法从复杂的大型数据集中检测异常值。LOF基于局部密度估计来检测离群值，这表明准确的密度估计可导致良好的检测性能。通过计算给定对象与其邻居之间的平均到达距离的倒数，可以粗略地计算出局部密度。因此，将无法准确估计局部密度，然后LOF无法找到真实的异常值。例如，图5显示了包括聚类1（*C*1）和聚类2（*C*2）的数据集的分布），它们具有不同的密度。显然，数据点*P*1接近密度较大的*C*1，而数据点*P*2接近*C*2。虽然*P*2个从偏离更大的距离*c ^*2比的偏离距离*P*1从*Ç*1，*P*1更可能比异常值*P*2。这是因为*C*1的密度远高于*C*2的密度。但是，*P*使用LOF 比使用*P*1更容易将2视为异常值，因为计算LOF时无需考虑数据分布。具体地，除了*C*2的点之外，在*P*2附近还存在许多*C*1的点，这增加了*P*2的离群因子。

其次，LOF是在单个*k*近邻或*k*距离邻域中计算的，因此它对参数*k*非常敏感。特别是，如果离群值以不同的密度随机分布在几个微簇中，则LOF可能会实现较差的检测性能。原因是参数*k*不能满足所有离群值的检测要求。

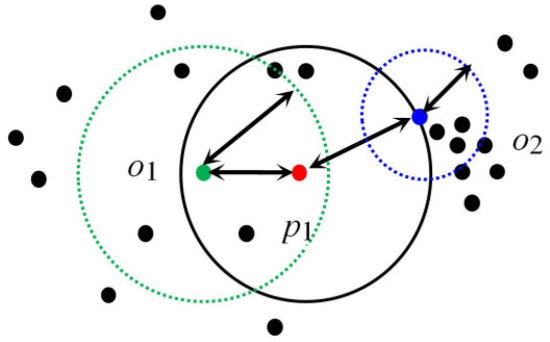
为了克服这两个缺点，基于上述传统的LOF，可以通过基于Nadaraya-Watson卷积核的回归来计算k个相邻距离中的i回归估计量来构造更健壮的KLOF [33]。设yi(i = 1,2，…n)为参数xi的回归，对于x的Nadara - Watson核回归y如下:



其中K(.)代表卷积核功能，n式yi的总数。

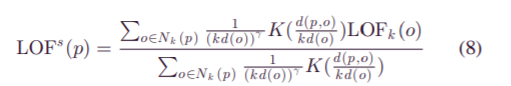


**图六.提出的方法中使用的卷积核曲线**



**图七.当*k*=5时，*p*1（红点），*o*1（绿点）和*o*2（蓝点）的*邻域*。**

根据(7)，我们可以获得一个对象的KLOFp在迭代*s处*，相应的方程式定义如下：

****

其中γ是一个敏感参数，