

文章编号:1672-3031(2020)01-0031-09

基于孤立森林算法的取用水量异常数据检测方法

赵臣啸¹, 薛惠锋¹, 王磊¹, 万毅²

(1. 中国航天系统科学与工程研究院, 北京 100048; 2. 水利部水资源管理中心, 北京 100053)

摘要: 水资源管理系统中储存着海量的取用水量数据, 通过筛选数据中的异常值定位异常取水行为, 是水资源监管的重要手段。对取用水量数据中的异常值普遍缺乏明确定义, 传统的异常值检测算法在实时性和稳定性方面存在不足。在总结归纳现阶段取用水量异常数据种类、特点的基础上, 首先运用平均插值法对可直观识别异常值进行预处理, 在预处理后的数据中随机取样训练, 建立多个孤立二叉树形成孤立森林, 以此为工具对数据样本进行异常值检测。对某供水公司连续两年日取水量监测数据的实证分析结果表明, 基于孤立森林算法的异常值检测方法将数据样本的特征通过非监督学习方式存储在森林中, 具有更高的稳定性; 能够准确检测出数据样本中的异常值, 相比于传统最小二乘拟合方法具有更高的检出率。

关键词: 水资源监测; 异常数据; 平均插值; 孤立森林; 最小二乘拟合

中图分类号: N37

文献标识码: A

doi: 10.13244/j.cnki.jiwhr.2020.01.004

1 研究背景

水是基础性的自然资源和战略性的经济资源, 是生态环境的控制性要素, 是经济社会发展的重要支撑和保障, 水资源供需矛盾突出是制约我国可持续发展的主要瓶颈之一。在当前全国范围内进行水资源税费改革的大背景下, 水资源监控能力越来越受到社会各界的关注, 保障取用水数据的准确性, 对用水总量控制和水资源税征收具有重要意义。取用水量异常值的检测是保障取用水数据准确的重要手段之一。

异常值检测是数据挖掘中十分重要的部分, 国内外学者在该领域提出了一系列的思路和方法, 形成了较为完整的体系。目前, 主要的异常值检测方法按照检测原理可分为基于偏差、基于统计、基于密度、基于聚类以及基于距离等方法^[1]。传统的异常值检测方法以基于密度和基于偏差的方法为主。侍建国等^[2]采用拉依达准则处理区域水文数据异常值, 肖树臣等^[3]运用格拉布斯法对实验数据进行分析 and 筛选, 但此类异常值检测方法, 都默认样本数据符合某种概率分布模型如正态分布、高斯分布等, 而水资源监测数据随机性强、易受外界影响, 仅简单将其归纳为某种分布缺乏科学性、严谨性。基于小波变换、基于最小二乘拟合法的异常值检测方式^[4-6]本质上都是基于偏差的异常值检测方法, 即首先利用小波变换或最小二乘拟合法对已有数据进行处理, 再对处理后的数据与原始样本数据进行残差分析, 彭小奇等^[7]、方海泉等^[8]在此基础上进行了研究。这类算法的主要问题在于, 数据拟合本身以已有数据作为样本, 拟合结果受已有数据影响较大, 数据拟合中存在很多参数, 不同的参数选择会对拟合结果产生较大影响。近年来, 随着数据挖掘理论与实践水平的不断提升, 越来越多的仿生算法被引入到异常值检测中。王琰等^[9]将 Bayes 方法引入时间序列异常值检测, 韩旻等^[10]采用仿生学中的阴性选择算法对飞行数据异常值进行检测。文献[11]于2007年提出 Isolation Forest 算法并将其应用到异常值检测, 张荣昌^[12]、张为金^[13]等分别将其应用到用电数据的分析方面, 朱佳俊等^[14]将孤立森林算法应用到用户画像的异常行为检测, 均取得了很好的检测效果。

作为数据处理分析的第一步, 数据异常值的检测一直以来都是数据挖掘领域研究者所关注的重

收稿日期: 2018-10-18

基金项目: 国家自然科学基金重点项目(U1501253)

作者简介: 赵臣啸(1996-), 硕士生, 主要从事水资源系统工程研究。E-mail: m18600493854@163.com

点。水资源的开放性特点给水资源的监测和管理带来很多困难，对取用水量数据异常值的检测不能简单沿用传统的异常值检测方法和模型，大数据时代的到来推动了数据挖掘领域向纵深发展，这为取用水量数据的分析和利用提供了新的思路和方法。本文以某取水户日取水量数据为研究对象，以基于Ensemble的孤立森林算法为主要方法进行异常值的检测。

2 取用水量数据特征分析

本文以实际取水户的实际监测数据为试验样本，在实际取水过程中，数据存在很大的随机性，且易受外界不确定因素影响。无论基于哪种异常值检测方法，单纯依靠数据特征筛选异常值往往都是不全面的，且都存在一定程度的误报。对于取用水量数据而言，基于数据特征只能找出“疑似异常值”，准确判定是否为异常值，还需要结合取水点其他信息以及专家知识，为方便理解，本文仍然使用“异常值”一词，但应该明确文中异常值与实际异常值存在的差异。

2.1 取用水量数据特征 取用水量数据属于典型的时间序列数据，数据整体具有明显的趋势性、周期性、随机性、综合性等特点。数据维度方面，以水利部组织建设的水资源管理系统为例，数据维度包含小时取水量(m^3/h)、日取水量(m^3/d)、年取水量(m^3/a)等。不同维度下呈现的数据虽然紧密关联，但数据特征具有很大区别，在进行分析时，只能对同一维度的数据进行分析比较，不同维度间的数据不存在可比性。

不同于电力、石油等相对封闭、来源相对单一的资源，水资源具有开放、分散和不确定的特点，易受环境及人为因素的影响。这些特点给水资源的监控和管理带来挑战，行政部门如何在海量监测数据中甄别有效、真实数据，并通过对数据的分析支撑决策，是当前水资源监控能力建设需要解决的重要问题。

2.2 异常数据定义及分类 在数据挖掘领域，通常将数据中的异常点定义为离群点(outlier)，将异常检测定义为偏差检测(deviation detection)或例外挖掘(exception mining)。异常数据具有以下基本特点：①在数据样本中占比很少；②相比于样本中的正常数据，异常数据具有明显不同的属性。作为时间序列数据，取用水量数据异常值还具有复杂性、多样性、滞后性和被动性的特点。

根据异常的成因进行分类，取用水量异常数据可分为两大类：主体异常和客体异常。主体异常是指取水行为本身存在异常，在数据特征上表现为数据突然上升或下降、与相邻时间数据规律不符，通常不会连续出现；客体异常是指数据采集、传输、交换和存储的过程存在异常，在数据特征上表现为连续出现极大数据或极小数据，甚至出现负值。

根据异常数据特点分类，取用水量异常数据可分为异常大值、异常小值、零值、负值、缺报值等类型。零值、负值成因复杂，需要筛选出来进行人工鉴别，考查数据中的零值是否为异常时，需结合取水户类型，季节性取水的灌区、企业等连续出现零值不应判定为异常；异常大值、异常小值是指有违于样本数据正常取水规律的值，不能简单理解为某一阈值之外的数据，取水量处于正常范围但与临近时间点取水规律不一致的数据应判断为异常数据；缺报值一般是由客体异常造成的，若对缺报值进行简单的删除或置零处理，将对缺报值附近数据的准确性造成影响，因此在数据处理时，应采用统计方法处理缺报值。

根据异常数据识别难易度分类，取用水量异常数据还可分为可直观识别的数据异常值和不能直观识别的数据异常值，具体类型见表1。

表1 取用水量数据异常值分类

取用水监测数据异常值分类			
按异常成因分类	主体异常	按数据特点分类	异常大值
			异常小值
	客体异常		零值、负值
			缺报值

3 数据预处理与数据异常值筛选方法

异常值筛选是数据进行分析处理的前提，数据预处理则是数据异常值筛选的重要基础。由上文可知，取用水监测数据异常值可分为可直观识别的数据异常值和难以直观识别的数据异常值，数据预处理的目即区分这两类异常值，并首先对可直观识别的数据异常值进行处理，以减小甚至消除此类数据异常值对周围数据的影响，从而提高难以直观识别的数据异常值的检出率，降低检错率。

3.1 数据预处理方法 通常，可直观识别的数据异常值是指数据中的负值、缺报值等，这些异常值往往难以通过某种特定方式修正，需要结合专家知识进行人工判断。

在已有研究中，通常将可直观识别的异常值直接剔除或置零，这种处理方法简单易实现，但忽略了被剔除、置零数据点对其他数据点产生的影响。若使用基于偏差的异常值检测算法进行异常值检测，可直观识别的异常值处理不当将大大提高算法的误检率。在取用水监测数据分析中，经常会对数据未来趋势进行预测，基于拟合的预测方法十分依赖已有数据的数据特征，若只对可直观识别的异常值进行简单的剔除或置零，将严重影响拟合精度。

本文采用均值法对可直观识别异常值进行处理，这种方法虽然会影响数据的方差，损失数据信息，但保证了数据的连续性、平稳性和合理性，极大地方便了后续分析。

可直观识别异常值的处理一般可分为两种情况：

(1)异常值单独出现。若 x_i 为可直观识别异常值点， x_{i-1} 、 x_{i+1} 所对应的数据可做进一步处理，则可将 x_i 点对应的数据量 y_i 定义为 $y_i = \frac{y_{i-1} + y_{i+1}}{2}$ ，对 (x_i, y_i) 进行标注，若进一步分析后判定 y_i 为异常值则不再计入。

(2)异常值连续出现。若 $x_i, x_{i+1}, \dots, x_{i+n}$ 这 $n+1$ 个点为可直观识别异常值， x_{i-1} 、 x_{i+n+1} 所对应的数据可做进一步处理，则可将 $x_i, x_{i+1}, \dots, x_{i+n}$ 对应的数据点分别定义为 $y_i = y_{i-1} + \frac{y_{i-1} + y_{i+n+1}}{n+2}$ ， $\dots$ ， $y_{i+n} = y_{i+n-1} + \frac{y_{i-1} + y_{i+n+1}}{n+2}$ ，分别对以上各点进行标注，若进一步分析后判定以上某点为异常值则不再计入。

3.2 基于孤立森林的数据异常值检测算法 孤立森林(Isolation Forest)是一种由周志华等人提出的基于 Ensemble 的快速异常值检测算法，具有线性时间复杂度和高精度度，是符合大数据处理要求的神经网络算法(图 1)。与本文对异常值的定义一致，孤立森林算法将异常值定义为“容易被孤立的离群点”，即分布稀疏且离密度高的群体较远的点。孤立森林算法的基本思想是，对描述同一对象的不同维度的数据构建一系列的随机二叉树。这些随机二叉树每个节点或有两个子节点，或为叶子节点。



图1 孤立森林算法原理

通过在取值范围内随机取值, 将该范围内的数据划分为两个分支, 再在两个分支中继续随机取值进行划分, 不断重复, 直到不可分割或者树的高度达到上限。相对于数据样本中的正常点, 异常点通常表现出稀少的特性, 因此在随机树中异常数据会很快被划分到叶子节点中, 即异常数据在随机树中的深度较浅; 相反, 正常数据由于集中为簇且密度较大, 往往通过多次分割才能划分为叶子节点。因此, 该算法通过叶子节点到根节点之间的路径长度, 可以快速判断一条数据是否为异常数据, 将多维数据的分割结果相综合, 则可以得知某一对象是否为异常对象。例如, 图 1 中, x_i 为正常对象, 需经过多次切割, 才能将其从所有数据中孤立出来; x_o 为异常对象, 只需经过较少次数的切割即可将其孤立。

孤立森林算法的实现过程可以分为两个阶段。

(1) 构建 t 个孤立二叉树(Isolation Tree)组成的孤立森林。孤立二叉树是构成孤立森林的基本元素, 由于孤立森林算法的学习过程属于无监督学习, 即不需要专门的训练集对其进行训练, 因此构造孤立二叉树的过程大大简化: ①从待检测数据中随机选择 φ 个样本点作为子样本集, 放入树的根节点; ②随机选取一个数据维度, 在当前节点数据中随机产生一个切割点 p ——切割点产生于当前节点数据中指定维度的最大值和最小值之间; ③以此切割点为基础形成一个超平面, 将当前节点的数据空间划分成 2 个子空间, 把指定维度中小于 p 的数据放在当前节点的左边, 把大于等于 p 的数据放在当前节点的右边; ④在子节点中递归步骤②、③, 不断构造新的子节点, 当数据本身不可再分或已经达到树的最大深度 $\log_2 \varphi$ 时, 递归过程结束。

生成一个孤立二叉树的伪代码如表 1 所示。

表 1 孤立二叉树伪代码

Algorithm 1 孤立二叉树(X,e,h)
Input: X - input data; e - current height; h - height limit.
Output: an 孤立二叉树.
1: if $e \geq h$ OR $ X \leq 1$ then
2: return $\text{exNode}\{\text{Size} \leftarrow X \}$;
3: else
4: Randomly select an attribute q;
5: Randomly select a split point p between min and max value of attribute q in X;
6: $X_l \leftarrow \text{filter}(X, q < p)$, $X_r \leftarrow \text{filter}(X, q \geq p)$;
7: return $\text{inNode}\{\text{left} \leftarrow \text{iTree}(X_l, e+1, h)$,
$\text{right} \leftarrow \text{iTree}(X_r, e+1, h)$,
$\text{SplitAttr} \leftarrow q, \text{SplitValue} \leftarrow p\}$;
8: end if

注: X 为独立抽取的训练样本; 参数 e 的初始值为 0; h 是树可以生成的最大高度。

(2) 对被检测样本计算异常分值。获得 t 棵孤立二叉树后, 孤立森林形成, 训练过程结束。由于孤立二叉树的形成具有随机性, 单独一棵树的结果并不可靠, 因此对于待测数据样本, 令其遍历孤立森林中的每一棵树, 计算数据样本中的每一个样本值落在每棵孤立二叉树的第几层, 最后得出样本 x 在每棵树的平均深度 $h(x)$ 。异常分值与样本在孤立二叉树的深度有关, 当样本在孤立二叉树中的深度越小, 则异常分值越高, 即该样本为异常样本的概率越大。

对 n 个数据样本, 将其路径长度记为 $h(n)$, 则其平均路径长度 $c(n)$ 为:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (1)$$

其中 $H(i)$ 为谐波数, 等于 $\ln(i) + \text{欧拉常数}$ 。

通过对孤立二叉树的长度进行归一化处理, 可以得到介于 0 ~ 1 之间的数即为被检测样本的异常

分值。记 $s(x, n)$ 为异常指数，有：

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2)$$

式中： $E(h(x))$ 为对某一个给定值的路径长度的期望； $s(x, n)$ 为对该值所对应的路径的归一化。

$$\begin{cases} s(x, n) \rightarrow 1, & \text{异常} \\ s(x, n) \rightarrow 0, & \text{正常} \\ s(x, n) \rightarrow 0.5, & \text{无明显异常} \end{cases} \quad (3)$$

孤立森林具有以下特点：①孤立森林具有线性时间复杂度，不需要计算距离或者密度来寻找异常数据；②抗噪能力强；③模型稳定性好；④可用于分布式系统，运算效率高；⑤不善于处理特别高维的数据，且仅对全局稀疏点敏感。

4 实例分析

为充分验证孤立森林算法的有效性并比较该算法与传统异常值检测算法的性能差异，本文以传统的基于偏差的最小二乘拟合算法作为对比项。在数据预处理完成后，首先运用孤立森林算法对数据样本进行异常值检测，分析检测结果的合理性和准确性；再对两种异常值检测方法的检测结果进行对比，验证孤立森林算法在处理此类问题方面的优越性。

4.1 数据说明 试验采用水利部水资源管理中心提供的广东省辖区内国家重点监控用水户某城市供水企业 2016 年日取水量数据 366 条、2017 年日取水量数据 365 条，主要研究供水企业取水量变化情况。供水企业担负着保障本地区生活、生产用水的直接责任，在水资源监测系统中处于主体地位。

4.2 可直观识别异常值的处理 以该城市供水企业 2017 年日取水量监测数据为例。在全部 365 条数据中，共有有效数据 350 条，缺失数据 15 条。数据中不存在负值、零值、连续不变值等情况，补充缺失数据后可直接进行分析检测异常值，数据预处理结果如表 2 所示。在此基础上，对预处理后的样本数据进行特征分析，结果如表 3 所示。

表 2 2017 年数据预处理

日期	补充缺失数据/m ³	日期	补充缺失数据/m ³	日期	补充缺失数据/m ³
2017/2/23	5158530	2017/10/3	5541460	2017/10/10	6182085
2017/2/27	3920057	2017/10/6	5649977	2017/11/6	6831077
2017/2/28	4246283	2017/10/7	5783004	2017/11/7	6509623
2017/6/27	4707455	2017/10/8	5916031	2017/12/11	112980
2017/9/19	5249280	2017/10/9	6049058	2017/12/31	5003278

表 3 样本数据特征分析

数据样本	平均值/m ³	5004569
	中位数/m ³	5280000
	标准差/m ³	1567249

图 2 为 2017 年数据预处理前后对比。从图 2 中可以看出，经过数据预处理，数据波形中的断点消失，而数据的总体特征得以完整保留，为后续的数据分析奠定了基础。应该注意，进行补充的数据点本身已经属于可直观识别的异常值，因此在进一步分析时，即使异常点中存在修补数据，也不应纳入统计，以避免重复统计，从而降低误检率。

同上，对该城市供水企业 2016 年数据进行可直观识别异常值的处理。在全部 366 条数据中，共有有效数据 364 条，缺失数据 2 条。数据中不存在负值、零值、固定值等情况，补充缺失数据后可直接进行分析检测异常值，数据预处理结果如表 4 所示，2016 年数据预处理前后效果对比见图 3。与对

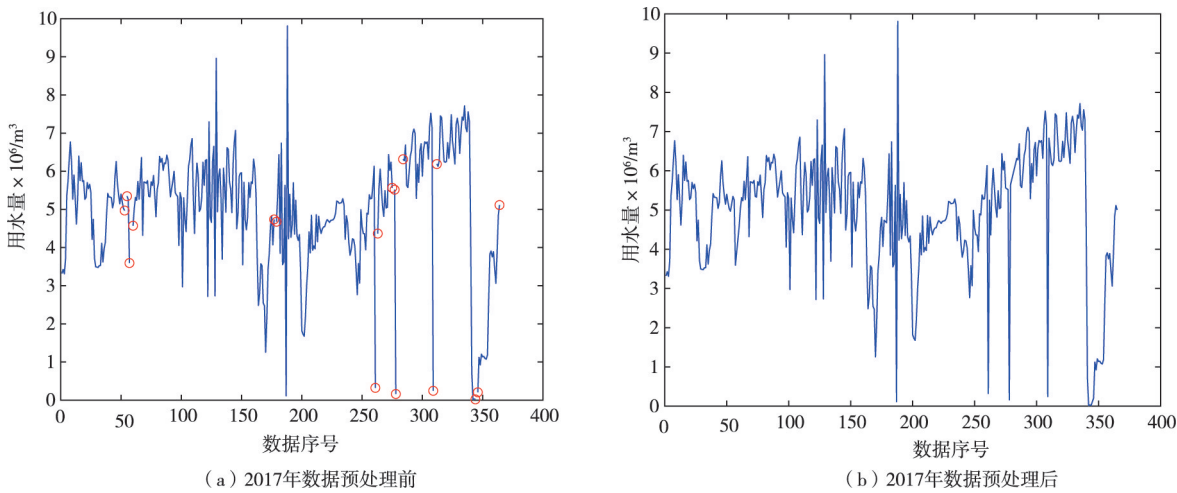


图2 2017年数据预处理前后效果对比

表4 2016年数据预处理

日期	补充缺失数据/m ³
2016/10/9	6391590
2016/12/31	3503680

表5 样本数据特征分析

数据样本	平均值/m ³	4528316
	中位数/m ³	4456410
	标准差/m ³	1377378

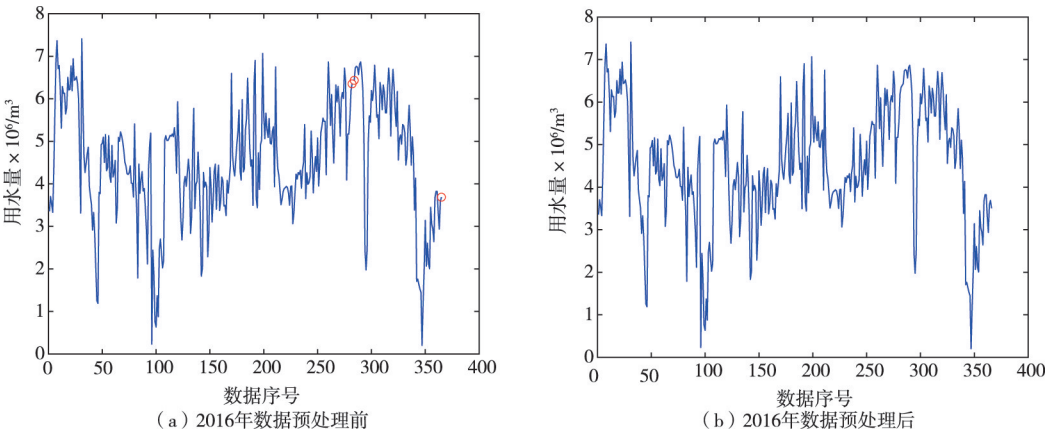


图3 2016年数据预处理前后效果对比

2017年数据的处理方式相同，对预处理后的2016年样本数据进行特征分析，结果如表5所示。

4.3 非可直观识别异常值的处理 孤立森林算法运用了集成学习的思想，在基于孤立森林的取用水量数据异常值算法中有两个重要参数：孤立二叉树采样数 φ ，称为采样规模；孤立二叉树数量 t ，称为集成规模。

采样规模和集成规模的确定遵循以下规则：①孤立森林算法的计算时间随着采样规模和集成规模的增加呈现线性增长趋势；②当孤立二叉树数量 t 达到一定值以后，模型精度的提升十分有限；③当采样规模过大时，模型的性能会明显下降，查准率、查全率均会受到影响。

以此为基础，结合已有研究的经验^[15]，本文将采样规模设定为256；集成规模设定为100。将预处理后的2016年、2017年数据分别输入试验程序，将得到全部数据点在100棵孤立二叉树上的平均

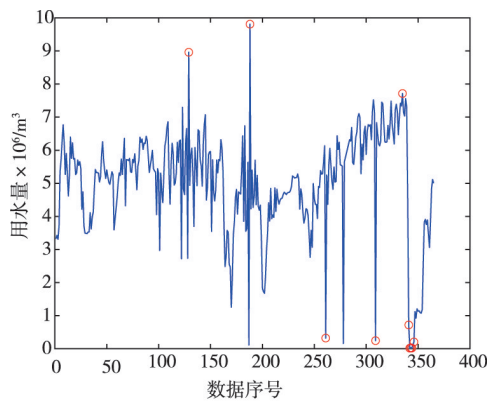
路径长度。

对各数据点的平均路径长度进行归一化处理，得归一化处理后所有数据的平均路径长度 $\text{ave}_{2017}=0.82$ $\text{ave}_{2016}=0.84$ 。定义任一数据点的异常率得分为： $s(x)=2^{-\frac{h(x)}{\text{ave}}}$ ，其中 $h(x)$ 为样本 x 在全部 100 棵树的平均路径长度。 $h(x)$ 越短，则异常率得分越高，即该数据点越有可能为异常值点。

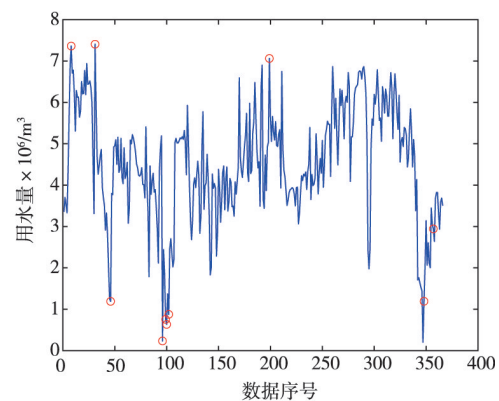
表 6 给出了数据样本中异常概率最高的 10 个数据。如表 6 所示，这些数据在数据样本中的分布情况见图 4。

表 6 数据异常值检测结果(孤立森林算法)

日期	监测水量/ m^3	异常率得分	异常类型	日期	监测水量/ m^3	异常率得分	异常类型
2017/7/7	9808720	0.817	异常大	2016/12/12	203330	0.781	异常小
2017/5/9	8960430	0.791	异常大	2016/4/5	232893	0.778	异常小
2017/9/18	322130	0.714	异常小	2016/1/31	7407526	0.773	异常大
2017/12/8	16950	0.705	异常小	2016/1/8	7363471	0.759	异常大
2017/12/1	7713440	0.702	异常大	2016/4/9	633521	0.738	异常小
2017/12/10	20300	0.702	异常小	2016/4/8	753944	0.729	异常小
2017/12/7	718960	0.701	异常小	2016/4/11	875975	0.720	异常小
2017/12/9	30810	0.700	异常小	2016/7/17	7063570	0.685	异常大
2017/11/5	244990	0.696	异常小	2016/2/15	1190052	0.669	异常小
2017/12/12	205660	0.680	异常小	2016/12/13	1191590	0.669	异常小



(a) 2017年数据异常值检测结果



(b) 2016年数据异常值检测结果

图 4 数据异常值检测结果(孤立森林算法)

孤立森林算法本身并不会对样本数据中存在的异常值数量和规模进行限定，仅按照异常概率进行排序。在本文中，异常值的确定结合了专家经验。首先，对数据异常值进行检测的首要目的是监控取用水量，因此数据异常值中最需要关注的是极大值和极小值。当出现极大值时，应考虑取水户是否存在超量取水，或计量统计过程是否产生偏差；当出现极小值时，应考虑取水户是否存在偷采行为。

4.4 算法效果评估 为验证孤立森林算法的检测效果和性能，引入传统的最小二乘拟合异常值检测算法，这一算法的原理是基于已有数据进行曲线拟合，计算得到的拟合曲线与原有数据之间的偏差，通过残差分析得出残差较大的数据点，将其认定为异常值。采用最小二乘法拟合法获得的异常值检验结果如表 7 所示，异常值在数据样本中的分布情况见图 5。

将孤立森林算法的检测结果与最小二乘拟合法的检测结果进行对比。以孤立森林算法的检测结果为基准，两种算法的检测重合度为 50%。与最小二乘拟合法相比，孤立森林算法对连续出现的异常值具有较高的检出率，最小二乘拟合法虽然能检出部分异常数据，但对于连续出现的异常值缺乏有效检测，且对某些接近均值或中位数的正常波动数据存在误检。

表 7 异常值检测结果(最小二乘拟合法)

日期	监测水量/m ³	日期	监测水量/m ³	日期	监测水量/m ³	日期	监测水量/m ³
2017/7/7	9808720	2017/10/5	160840	2016/4/4	5190000	2016/4/5	232893
2017/5/9	8960430	2017/9/18	322130	2016/7/10	6900070	2016/4/1	2117755
2017/12/4	7556100	2017/7/6	114340	2016/5/14	5773914	2016/3/23	1787580
2017/12/5	7306510	2017/12/8	16950	2016/10/19	5111440	2016/5/22	2010544
2017/11/5	244990	2017/12/9	30810	2016/4/17	5030000	2016/5/27	2286174

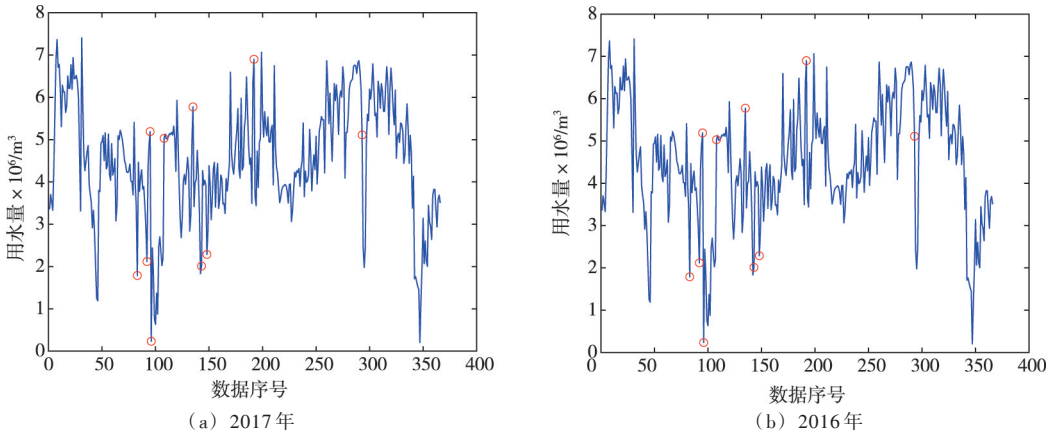


图 5 数据异常值检测结果(最小二乘拟合法)

此外，传统的异常值检测算法通常具有滞后性，无法做到对异常值的实时监测，孤立森林算法在这一方面进行了优化，已有数据的特征已经存储在各个孤立二叉树中，即孤立森林已经涵盖了已有数据的基本特征，以此为基础可为按时间顺序出现的新数据进行判断。

4.5 讨论 取用水量监测数据存在多个维度，而不同维度之间的数据不具备可比性，这一特点限制了本文对于数据的应用范围。如本文所使用的样本数据，当数据维度缩减到一维时，孤立森林算法实际上将异常值的筛选问题抽象为数据出现的频次问题，异常值出现频次较低，分布稀疏，因此更容易被区分出来，孤立森林算法在本研究中的应用正是基于这一原理。但这种单一维度的数据处理并没有发挥出孤立森林算法的最佳性能。

若能将数据维度进行拓展，运用孤立森林算法就能实现对数据更立体、更全面的分析。假设将取水点的日取水数据细化为小时取水数据，则 1 维数据被拓展到 24 个维度，这 24 个值共同描绘了某一取水点某日的取水行为。将该取水点一年的小时取水量数据进行综合，可以得到一个 365×24 的矩阵，在此基础上运用孤立森林算法，算出平均路径，则可以对 365 d 的取水行为进行排序，得到最可能属于异常取水行为的数据。这种方式基于对小时取水量的全面分析，数据量大，孤立森林的优势得以完全发挥。

5 结论

对取用水量数据进行分析，得出取用水量数据具有趋势性、周期性、随机性、综合性的特点，且数据维度多元，不同维度的数据变化趋势不同，缺乏可比性；按照数据特点，将取用水量数据异常值分为异常大值、异常小值、零负值、缺报值等 4 类；孤立森林算法基于数据本身的位置特征筛选异常值，具有抗噪能力强、模型稳定性好、运算效率高的特点，该方法可用于处理水量数据异常值检测。通过实证分析，以最小二乘法为代表的传统的基于偏差的数据异常值检测算法易受异常数据干扰，检测结果不稳定，无法发现连续出现的异常值，孤立森林算法基于数据整体特征，不易受数据中异常值的影响，对各类异常值有较高的检出率。

水资源管理系统中的数据以小时为单位进行上报，将日水量数据分解为小时水量数据，可以在

更高维度描述一天内的取水行为；在处理高维数据异常值问题时，孤立森林算法使用超平面对高维数据进行分割，相比传统方法效率更高，筛选结果更加准确。

参 考 文 献：

- [1] PANG-NING TAN . 数据挖掘导论：完整版[M] . 北京：人民邮电出版社，2011 .
- [2] 侍建国，张亦飞 . 拉依达准则在处理区域水文数据异常值中的应用[J] . 海河水利，2016(5)：49-51 .
- [3] 肖树臣，秦玉勋，韩吉庆 . 基于格拉布斯法的试验数据分析方法[J] . 弹箭与制导学报，2007(1)：275-277 .
- [4] 张利红 . 基于小波分析的时间序列异常值检验[D] . 湘潭：湘潭大学，2017 .
- [5] 尚华，张贝贝，纪宏 . 一种新的基于回归分析的异常值检测[J] . 河南大学学报(自然科学版)，2015，45(6)：635-639 .
- [6] 王爱军，李宏，张小桃 . 一种基于小波变换的超高压线路故障选相方法[J] . 电力系统保护与控制，2013，41(12)：92-97 .
- [7] 彭小奇，宋彦坡，唐英 . 基于小波分析的异常样本处理[J] . 信息与控制，2005(6)：676-679 .
- [8] 方海泉，薛惠锋，蒋云钟，等 . 基于 EEMD 的水资源监测数据异常值检测与校正[J] . 农业机械学报，2017，48(9)：257-263 .
- [9] 王琰，张倩倩，车通宇，等 . 时间序列异常值探测的 Bayes 方法及其 GNSS 数据质量控制中的应用[J] . 中国惯性技术学报，2016，24(1)：45-50 .
- [10] 韩旻，赵清洲 . 一种基于阴性选择算法的飞行数据异常值检测方法[J] . 航空计算技术，2010，40(4)：53-55，58 .
- [11] LIU F T, KAI M T, ZHOU Z H . Isolation Forest[C]//Eighth IEEE International Conference on Data Mining . IEEE, 2009 .
- [12] 张荣昌 . 基于数据挖掘的用电数据异常的分析与研究[D] . 北京：北京交通大学，2017 .
- [13] 张为金 . 基于机器学习的电力异常数据检测[D] . 成都：电子科技大学，2018 .
- [14] 朱佳俊，陈功，施勇，等 . 基于用户画像的异常行为检测[J] . 通信技术，2017，50(10)：2310-2315 .
- [15] AGGARWAL C C . Outlier Analysis[M] . Springer Publishing Company, Incorporated, 2015 .

Water Consumption Abnormal Data Detection Method based on Isolation Forest

ZHAO Chenxiao¹, XUE Huifeng¹, WANG Lei¹, WAN Yi²

(1.China Aerospace Academy of Systems Science and Engineering, Beijing 100048, China;

2.Water Resources Management Center, The Ministry of Water Resources of the People's Republic of China, Beijing 100053, China)

Abstract: Water resource management system store hugs amounts of data on water consumption, and it is an important means of water resource regulation to locate abnormal water intake behavior by screening the abnormal values in the data. These outliers lack effective classification. The traditional outlier detection algorithm has shortcomings in real-time and stability. On the basis of summarizing the types and characteristics of abnormal data of water consumption at the present stage, firstly, the average interpolation method is used to pre-process the outliers, and random sampling training is performed in the pre-processed data to establish multiple isolated binary trees to form isolation forest. The forest is used to perform outlier detection on data samples. The empirical analysis of the daily water intake monitoring data of a water supply company shows that the outlier detection method based on the isolation forest algorithm stores the characteristics of the data samples in the forest through unsupervised learning, which has higher stability and can accurately detect. The outliers in the data samples have a higher detection rate than the traditional least squares fitting method; they are suitable for real-time monitoring of water resources data.

Keywords: water resources monitoring; abnormal data; average interpolation; isolation forest; least squares

(责任编辑：王学风)