查看全部 39 个回答

阿里云云栖号
已认证账号

+ 关注

282 人赞同了该回答 〉

最近很多小伙伴都比较关注异常值检测的方法，接下来小编就为大家介绍几种，希望能帮到大家！！

摘要： 本文介绍了异常值检测的常见四种方法，分别为Numeric Outlier⁺、Z-Score、DBSCA以及Isolation Forest⁺

在训练机器学习算法或应用统计技术时，错误值或异常值可能是一个严重的问题，它们通常会造成测量误差或异常系统条件的结果，因此不具有描述底层系统的特征。实际上，最佳做法是在进行下一步分析之前，就应该进行异常值去除处理。

在某些情况下，异常值可以提供有关整个系统中局部异常的信息；因此，检测异常值是一个有价值的过程，因为在这个工程中，可以提供有关数据集的附加信息。

目前有许多技术可以检测异常值，并且可以自主选择是否从数据集中删除。在这篇博文中，将展示KNIME分析平台中四种最常用的异常值检测的技术。

数据集和异常值检测问题

本文用于测试和比较建议的离群值检测技术的数据集来源于航空公司数据集，该数据集包括2007年至2012年间美国国内航班的信息，例如出发时间、到达时间、起飞机场、目的地机场、播出时间、出发延误、航班延误、航班号等。其中一些列可能包含异常值。

从原始数据集中，随机提取了2007年和2008年从芝加哥奥黑尔机场（ORD）出发的1500次航班样本。

为了展示所选择的离群值检测技术是如何工作的，将专注于找出机场平均到达延误的异常值，这些异常值是在给定机场降落的所有航班上计算的。我们正在寻找那些显示不寻常的平均到达延迟时间的机场。

四种异常值检测技术

数字异常值|Numeric Outlier

数字异常值方法是一维特征空间中最简单的非参数异常值检测方法，异常值是通过IQR⁺（InterQuartile Range）计算得的。

计算第一和第三四分位数（Q1、Q3），异常值是位于四分位数范围之外的数据点x i：

$$x_i > Q3 + k(IQR) \bigvee x_i < Q1 - k(IQR),$$
$$\text{where } IQR = Q3 - Q1 \text{ and } k \geq 0.$$

使用四分位数乘数值k=1.5，范围限制是典型的上下晶须的盒子图。这种技术是使用KNIME Analytics Platform内置的工作流程中的Numeric Outliers节点实现的（见图1）。

Z-score

Z-score是一维或低维特征空间中的参数异常检测方法。该技术假定数据是高斯分布，异常值是分布尾部的数据点，因此远离数据的平均值。距离的远近取决于使用公式计算的归一化数据点z i的设定阈值Zthr：

关于作者

阿里云云栖号
阿里云官网内容平台

已认证账号

回答	文章	关注者
499	7,052	334,645

关注　发私信

被收藏 479 次

机器学习 tom pareto 创建	2409 人关注
数据分析 Linglai Li 创建	1779 人关注
机器学习 李开国 创建	298 人关注
数据分析师 求知鸟 创建	16 人关注
大数据 劳风雷 创建	12 人关注

相关问题

异常检测（anomaly/ outlier detection）领域还有那些值得研究的问题？ 52 个回答

$$z_i = \frac{x_i - \mu}{\sigma},$$

其中xi是一个数据点，μ是所有点xi的平均值，δ是所有点xi的标准偏差。

然后经过标准化处理后，异常值也进行标准化处理，其绝对值大于Zthr：

$$|z_i| > z_{thr}$$

Zthr值一般设置为2.5、3.0和3.5。该技术是使用KNIME工作流中的行过滤器节点实现的（见图1）。

DBSCAN

该技术基于DBSCAN聚类方法，DBSCAN是一维或多维特征空间中的非参数，基于密度的离群值检测方法。

在DBSCAN聚类技术中，所有数据点都被定义为核心点（Core Points）、边界点（Border Points）或噪声点（Noise Points）。

- 核心点是在距离ε内至少具有最小包含点数（minPTs）的数据点；
- 边界点是核心点的距离ε内邻近点，但包含的点数小于最小包含点数（minPTs）；
- 所有的其他数据点都是噪声点，也被标识为异常值；

从而，异常检测取决于所要求的最小包含点数、距离ε和所选择的距离度量，比如欧几里得或曼哈顿距离。该技术是使用图1中KNIME工作流中的DBSCAN节点实现的。

孤立森林|Isolation Forest

该方法是一维或多维特征空间中大数据集的非参数方法，其中的一个重要概念是孤立数。

孤立数是孤立数据点所需的拆分数。通过以下步骤确定此分割数：

- 随机选择要分离的点“a”；
- 选择在最小值和最大值之间的随机数据点“b”，并且与“a”不同；
- 如果“b”的值低于“a”的值，则“b”的值变为新的下限；
- 如果“b”的值大于“a”的值，则“b”的值变为新的上限；
- 只要在上限和下限之间存在除“a”之外的数据点，就重复该过程；

与孤立非异常值相比，它需要更少的分裂来孤立异常值，即异常值与非异常点相比具有更低的孤立数。因此，如果数据点的孤立数低于阈值，则将数据点定义为异常值。

阈值是基于数据中异常值的估计百分比来定义的，这是异常值检测算法的起点。有关孤立森林技术图像的解释，可以在此找到详细资料。

通过在Python Script中使用几行Python代码就可以实现该技术。

```
from sklearn.ensemble import IsolationForest
import pandas as pd

clf = IsolationForest(max_samples=100, random_state=42)
table = pd.concat([input_table['Mean(ArrDelay)']], axis=1)
clf.fit(table)
```

```
output_table = pd.DataFrame(clf.predict(table))```python
```

Python Script节点是KNIME Python Integration的一部分，它允许我们将Python代码编写/导入到KNIME工作流程。

在KNIME工作流程中实施

KNIME Analytics Platform是一个用于数据科学的开源软件，涵盖从数据摄取和数据混合、数据可视化的所有数据需求，从机器学习算法到数据应用，从报告到部署等等。它基于用于可视化编程的图形用户界面，使其非常直观且易于使用，大大减少了学习时间。

此外，它被设计为对不同的数据格式、数据类型、数据源、数据平台以及外部工具（例如R和Python）开放，还包括许多用于分析非结构化数据的扩展，如文本、图像或图形。

KNIME Analytics Platform中的计算单元是小彩色块，名为“节点”。一个接一个地组装管道中的节点，实现数据处理应用程序。管道也被称为“工作流程”。

鉴于所有这些特性，本文选择它来实现上述的四种异常值检测技术。图1中展示了异常值检测技术的工作流程。工作流程：

- 1.读取Read data metanode中的数据样本；
- 2.进行数据预处理并计算Preproc元节点内每个机场的平均到达延迟；
- 3.在下一个名为密度延迟的元节点中，对数据进行标准化，并将标准化平均到达延迟的密度与标准正态分布的密度进行对比；
- 4.使用四种选定的技术检测异常值；
- 5.使用KNIME与Open Street Maps的集成，在MapViz元节点中显示美国地图中的异常值机场。

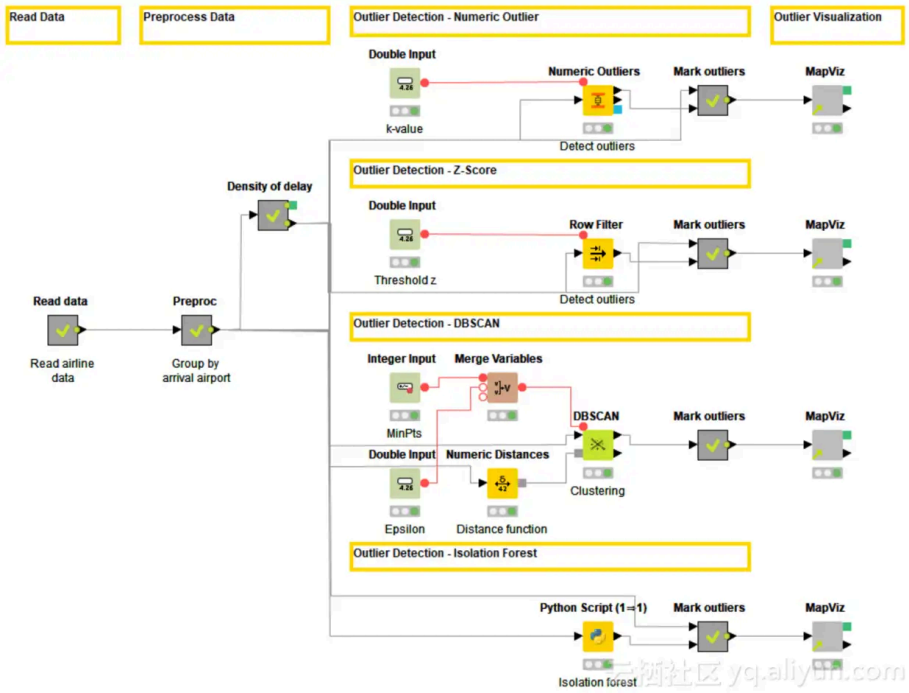

图1：实施四种离群值检测技术的工作流程：数字异常值、Z-score、DBSCAN以及孤立森林

检测到的异常值

在图2-5中，可以看到通过不同技术检测到的异常值机场。其中。蓝色圆圈表示没有异常行为的机场，而红色方块表示具有异常行为的机场。平均到达延迟时间定义的大小了记。

一些机场一直被四种技术确定为异常值：斯波坎国际机场（GEG）、伊利诺伊大学威拉德机场（CMI）和哥伦比亚大都会机场（CAE）。斯波坎国际机场（GEG）具有最大的异常值，平均到达时间非常长（180分钟）。然而，其他一些机场仅能通过一些技术来识别、例如路易斯阿姆斯特朗新奥尔良国际机场（MSY）仅被孤立森林和DBSCAN技术所发现。

对于此特定问题，Z-Score技术仅能识别最少数量的异常值，而DBSCAN技术能够识别最大数量的异常值机场。且只有DBSCAN方法（MinPts = 3/Ɛ= 1.5，欧几里德距离测量）和孤立森林技术（异常值的估计百分比为10%）在早期到达方向发现异常值。

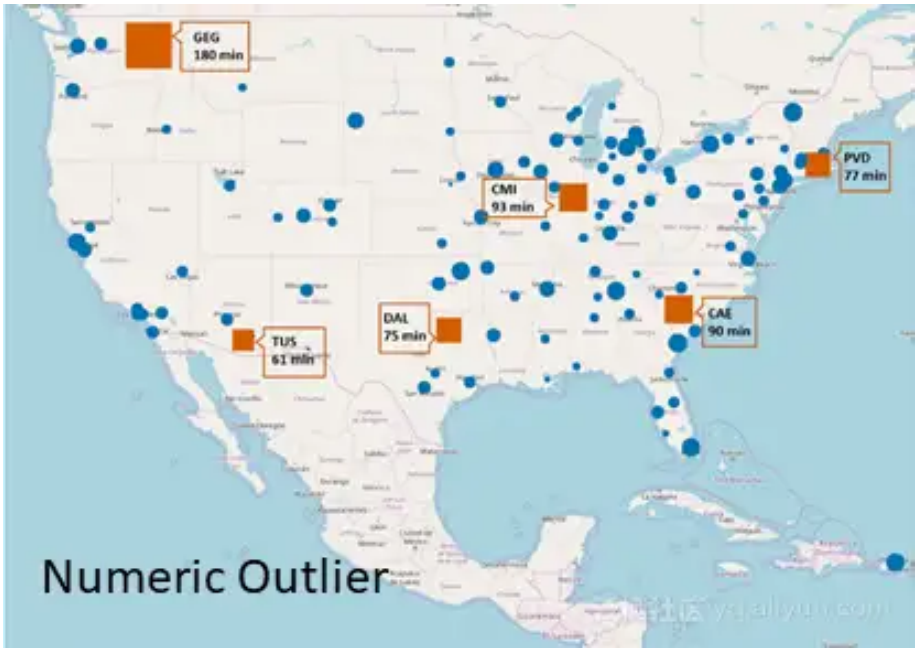

图2：通过数字异常值技术检测到的异常值机场

图3：通过z-score技术检测到的异常机场

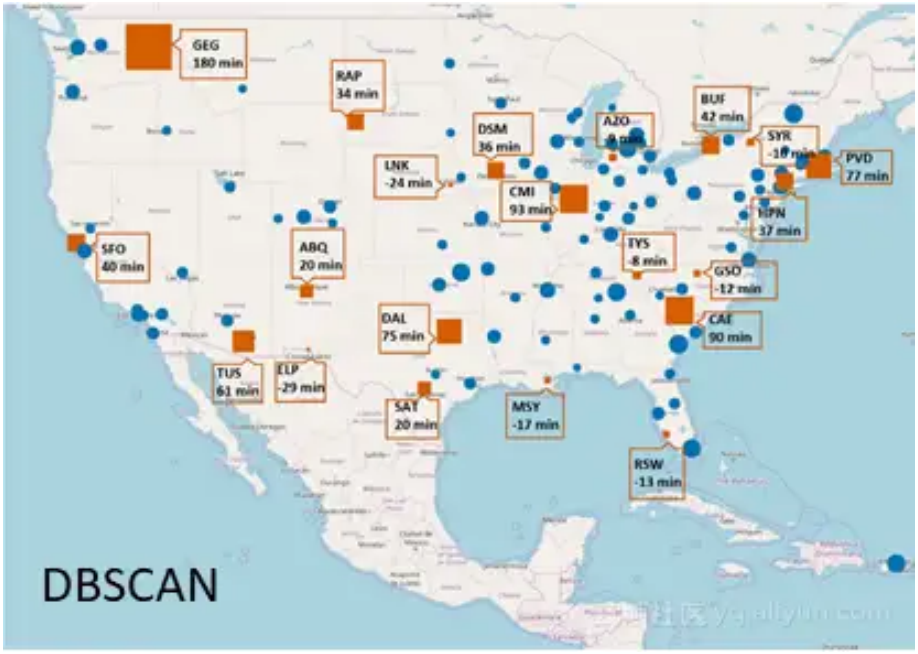

图4：DBSCAN技术检测到的异常机场

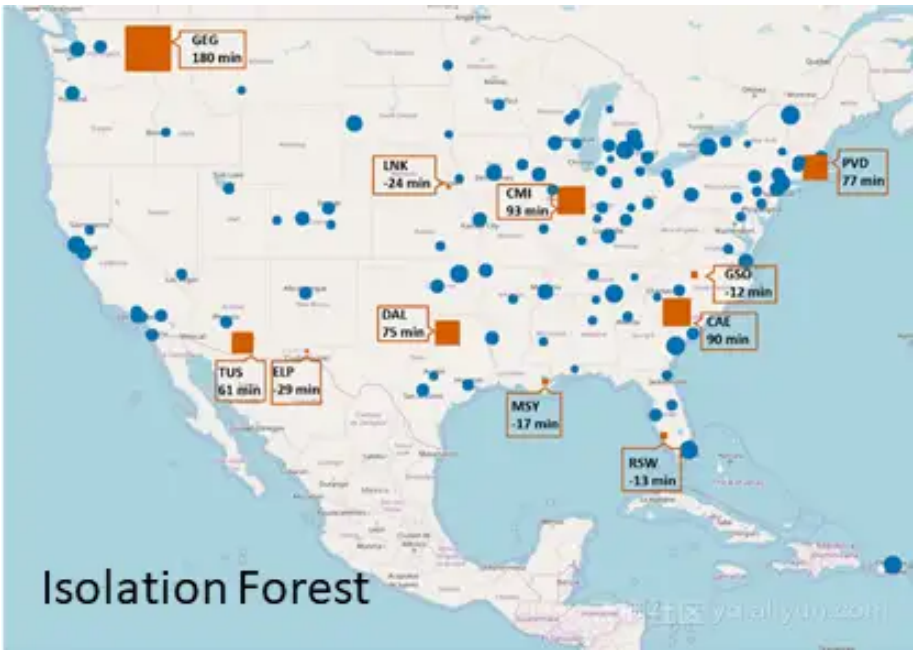

图5：孤立森林技术检测到的异常机场

总结

本文在一维空间中描述并实施了四种不同的离群值检测技术：2007年至2008年间所有美国机场的平均到达延迟。研究的四种技术分别是Numeric Outlier、Z-Score、DBSCAN和Isolation Forest方法。其中一些用于一维特征空间、一些用于低维空间、一些用于高维空间、一些技术需要标准化和检查维度的高斯分布。而有些需要距离测量，有些需要计算平均值和标准偏差。有三个机场，所有异常值检测技术都能将其识别为异常值。但是，只有部分技术（比如，DBSCAN和孤立森林）可以识别分布左尾的异常值，即平均航班早于预定到达时间到达的那些机场。因此，应该根据具体问题选择合适的检测技术。

参考

- Santoyo, Sergio. (2017, September 12). A Brief Overview of Outlier Detection Techniques；

以上为译文，由阿里云云栖社区组织翻译。

译文链接