arXiv:2303.08193v1 [cs.DB] 14 Mar 2023

RODD：数据立方体中的稳健离群点检测

Lara Kuhlmann*[2], [3][0009−0006−2934−0203], Daniel Wilmes*[1][0009−0000−6358−7732], Emmanuel M¨uller[1], [4][0000−0002−5409−6875], Markus Pauly[2], [4][0000−0002−0976−7190]和 Daniel Horn[2], [4][0000−0002−5208−0482]

[1] 计算机科学系，多特蒙德工业大学，德国多特蒙德，{daniel.wilmes, emmanuel.mueller}@cs.tu-dortmund.de[2] 统计系，多特蒙德工业大学，德国多特蒙德，lara.kuhlmann@tu-dortmund.de, {pauly, dhorn} @statistik.tu-dortmund.de[3] 物流研究生院，机械工程系，多特蒙德工业大学，德国多特蒙德[4] 可信数据科学与安全研究中心，多特蒙德工业大学，德国多特蒙德

摘要。 数据立方体是多维数据库，通常由多个独立数据库构建而成，为数据分析提供了灵活的基础。令人惊讶的是，针对数据立方体的离群点检测尚未得到广泛研究。在本文中，我们提出首个用于评估数据立方体中稳健离群点检测方法的框架（RODD）。我们介绍一种新颖的基于随机森林的离群点检测方法（RODD-RF），并将其与基于稳健地点估计量的传统方法进行比较。我们提出一种通用的测试数据类型，并通过模拟研究考察所有方法。此外，我们将 RODD-RF 应用于真实世界数据。结果表明，RODD-RF 能够提升离群点检测效果。

关键词： 离群点检测 · 数据立方体 · 分类数据 · 随机森林。

1 引言

过去几年，全球每年创建、采集、复制和消费的数据量迅速增长：从 2010 年的 2 泽字节增至 2021 [15,27] 年的约 79 泽字节。无论是在公共数据中，还是在企业内部数据中，以异常值形式出现的罕见事件都能为客户意见、市场发展和流程提供有价值的见解 [14]。因此，数据通常并非存储在单一数据库中，而是分布于多个数据库，并汇总至一个独立的决策支持数据库，称为数据仓库。与用于记录事务的传统数据库管理系统不同，数据仓库能够通过分析从数据中提取知识。数据仓库通常采用所谓的数据立方体 [11]进行建模。

数据立方体让我们能够从多个维度查看数据。一个示例

* 这些作者对本研究贡献相同

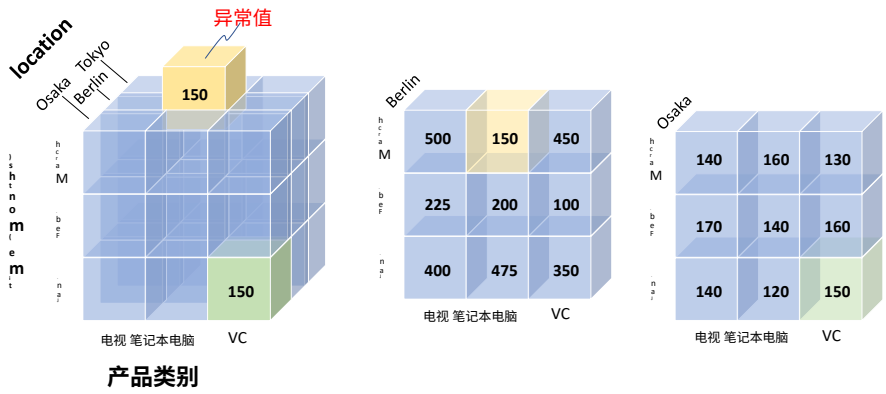


图 1：左侧：数据立方体示例；中间：显示柏林销售额的数据立方体切片；右侧：显示大阪销售额的数据立方体切片。

数据立方体如图 1 所示。考虑一家电子公司的销售数据，这些数据按产品类别、月份和城市等类别维度进行查看。我们构造以下玩具示例：产品类型可以是笔记本电脑、吸尘器或电视；月份可以是（为清晰起见）2022年一月、二月或三月；所考虑的城市为大阪、东京和柏林。类别维度的取值定义了数据立方体中的一个单元格，其中存储着一个数值 y，在本例中即具体的销售额：例如，绿色标记的单元格表示2022年三月大阪所有吸尘器的销售额，本例中为150。在这个示例中，另一个单元格（笔记本电脑、三月、柏林；黄色标记）具有相同的销售额150。我们如何判断这些数据点中是否有一个是异常值？数据立方体稳健离群点检测（RODD）的方法应该是什么样的？

我们没有异常值的真实值，因此只能采用无监督学习方法。一种数据驱动的方法是将每个点与“相关”点进行比较。图 1 中间部分的数据立方体切片展示了柏林不同月份和不同产品的销售情况，其中尤其包含了标记为黄色的单元格。尽管 150 这一销售额并不是该切片中的最低值，但它明显低于所考虑单元格所在列和行中的数值：三月电视和吸尘器的销售额，以及二月和一月笔记本电脑的销售额都更高。因此，预计三月份柏林的笔记本电脑销售额会远高于 150，将此单元格视为异常值是合理的。请注意，对于绿色单元格（图 1 右侧），数据立方体中有关大阪的视图并未显示上述任何异常之处：该单元格的值与视图中的其他值相差不大。因此，值 $y = 150$ 仅对标记为黄色的单元格而言是异常的，对标记为绿色的单元格则不是。

因此，简单的极值分析等统计方法 [38] 在这种情况下会失效：150 并非极值。将柏林城市切片（图 1 中间部分）中的值聚类为 $\{100, 150, 200, 225\}$和 $\{350, 400, 450, 475, 500\}$两个簇，反而会得出 $y = 150$ 属于某个簇的结论。因此，基于距离的异常值检测 [18] 以及局部异常因子（LOF）等基于密度的离群点检测技术 [4] 也会失效。这给数据立方体中的异常值检测带来了挑战。

另一个挑战是，许多离群点检测算法仅适用于存在数值的情况 [39]。然而，在我们的数据立方体场景中，类别值决定了单元格在数据立方体中的位置，而数值只有一个，在我们的例子中就是销售值。

根据这些观察，RODD 的一种直观方法如下：使用任何合理的预测方法来估计每个单元格的销售额。如果某个单元格的估计值 $\hat{y}$ 与实际值 y 相差过大，就将其标记为异常值。基于这一思路， [33] 类似于一种用于 RODD 的方差分析方法 [36] 。采用截尾均值估计 $\hat{y}$ 时，实际销售值与估计销售值之间的偏差 $r := |y - \hat{y}|$ 会根据数据的离散程度 σ 进行归一化。由此得到所谓的 *SelfExp* 值，它用于定义某个事件的离群程度：$\text{SelfExp}(y) := \frac{r}{\sigma}$。所有超过某个阈值 τ 的 SelfExp 值都被视为异常值。 [33] 仅考虑使用特定截尾均值来计算 $\hat{y}$，而该方法实际上也适用于其他合理的估计器。特别是，我们建议使用随机森林（RF）回归器 [3] 来计算 $\hat{y}$ ，因为事实证明，它在许多回归任务中都是一种稳健的方法，参见例如 [7, 9, 16]。我们将此方法称为 RODD-RF。

本文结构如下：第2节介绍相关工作。第3节介绍用于稳健离群点检测的框架。随后，我们讨论估计数据立方体单元的值的不同方法（第4节）。模拟研究评估了不同估计器的性能（第5节）。我们讨论了研究结果及其局限性（第6节），并在真实世界数据集上验证了RODD-RF方法（第7节）。最后，本文给出结论和展望（第8节）。

2 相关工作

有监督 异常值检测方法需要对异常训练数据进行标注。例如， [23, 37] 在带标签的异常值上训练随机森林分类器。相比之下，无监督 方法不需要对异常进行标注。 [20–22, 42]提出了无监督场景下基于随机森林的方法。基于模型的 离群点检测方法会根据给定数据学习模型参数。此类方法假设数据确实由所用模型生成。这类方法的一个子类是高斯混合模型 [41] ，其参数通过期望最大化 [8] 进行学习。第二个子类是基于回归的方法。在这类方法中，会将回归器的预测值与数据点的实际值进行比较 [26, 39]。基于模型的离群点检测方法的一个缺点是，需要拟合模型参数

以符合对数据所作的假设，而这一假设未必成立。因此，非参数 离群点检测方法应运而生。这类方法不会拟合模型参数。其中包括基于距离的异常值检测方法 [18], 、局部异常因子（LOF） [4, 17] 和核密度估计方法（KDE） [19, 28]。

第三种离群点检测范式是使用神经网络。可以将该问题视为监督分类任务，其中异常值有各自的标签 [39]。然而，由于通常无法获得标签，许多方法侧重于无监督学习 [5,31,44]。一种方案是采用自编码器方法[1,6,12,43]。为解决缺少标签的问题，研究人员提出了利用已标记数据点信息的半监督方法 [25, 32]。

在工业环境中应用数据立方体时，许多常见的离群点检测方法都无法使用。企业通常在应用离群点检测方法之前，并不了解其数据中的异常值。因此，所有有监督或半监督方法都可以排除。此外，虽然可以应用基于距离的方法，但这类方法会忽略分类数据及其所包含的宝贵信息。因此，我们侧重于无监督异常值检测方法，具体而言，是那些计算数据立方体单元期望值并将其与实际值进行比较的方法。随后，我们的研究主题是比较并进行估计器的性能评估，例如我们提出的随机森林回归器。

3 数据立方体离群点检测框架

数据立方体是一个 n 维数据库 DB。令 $D := \{d_1, \ldots, d_n\}$ 表示所有维度的集合，其中每个维度都是类别型的。每个数据点 $\mathrm{p} \in DB$ 都会被赋予一个数值 $y := f(\mathrm{p})$，并存储在数据立方体中。例如， $f(\mathrm{p})$ 可以是销售额。下文将根据 [33] 中介绍的技术，抽象出一个用于数据立方体离群点检测的框架：从高层次来看，需要为每个数值 y 计算相应的数值估计 $\hat{y}$ ，并计算数据的离散程度。随后将这些信息纳入异常值分数 $\mathrm{score}(y) = |y - \hat{y}|/\sigma.$ 。如果 $\mathrm{score}(y)$ 超过阈值 τ，则值 y 及其对应的数据点 p 会被视为异常值。下面介绍一种确定 $\hat{y}$ 的通用方法。为此，先引入一些记号：给定数据点 $\mathrm{p} \in DB$，将 p 在维度 $D' \subsetneq D$ 上的投影记作 $\mathrm{p}' := \mathrm{proj}(\mathrm{p}\ D')$。现在，我们采用如下通用的方差分析（ANOVA）方法 [36]：为了描述维度子集 D' 对销售额 $y(p)$ 的效应，我们为每个投影赋予一个模型系数 $\gamma(\mathrm{p}')$。模型系数 $\gamma(\mathrm{p}')$ 的一种典型取法，是对投影 p' 中的所有值求（截尾）均值，示例如下。由于销售额通常具有乘法性质 [33], ，我们考虑其对数值：因此，我们通过对每个投影的效应$\gamma(\mathrm{p}')$求和，来估计 p 的对数销售额，即

$$\log \hat{y}(\mathrm{p}) = \sum_{\mathrm{p}' \in \mathrm{proj_set}(\mathrm{p})} \gamma(\mathrm{p}'), \tag{1}$$

其中，求和遍历 p 的所有可能投影 p′ 。每个模型系数都会考虑维度的一个子集 $D' \subsetneq D$ 及其对应的数据单元格选择。模型系数可以看作是对其选取的数据单元格的数值估计。下面通过一个示例说明：假设图2中的设置与第 1 节的玩具示例相同，并且我们希望计算数据点 p=(笔记本电脑, 三月, 柏林)（以黄色标出）的数值估计（销售值）$\hat{y}$ ，记作 p $=(i, j, k)$。此时，每个模型系数都会基于一个视图计算数值估计，也就是说，基于数据立方体中的一个维度子集和数据点子集进行计算。

对于元组 p $=(i, j, k)$，p 的所有投影构成的集合 proj set 为 proj set(p) $=\{(i, j), (i, k), (j, k), (i), (j), (k), ()\}$。注意，允许空投影 ()。然后，我们将所有模型系数收集到集合 $\{\gamma(\text{p}') \mid \text{p}' \in$proj set(p)$\} = \{\gamma(i, j)\ \gamma(i, k)\ \gamma(j, k)\ \gamma(i)\ \gamma(j)\ \gamma(k)\ \gamma()\}$ 中。现在我们知道

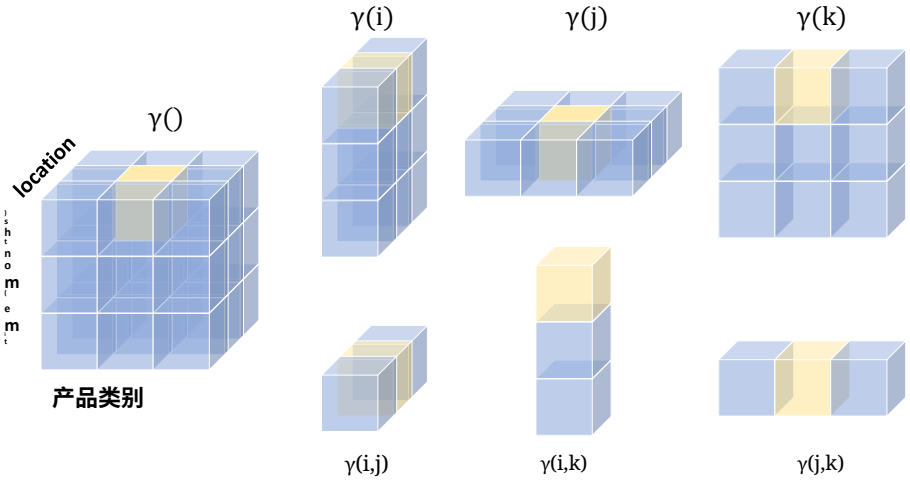


图 2：示例：数据立方体的不同视图。数据立方体的七个视图分别显示了用于计算相应模型系数 γ 的相关单元格。随后，需要全部七个模型系数来确定黄色单元格的估计值 $\hat{y}$ 。

为了确定要计算哪些模型系数，我们必须为每个模型系数确定其对应的计算数据点集合。给定一个维度子集 D'，模型系数 γ(p′) 的视图定义为 view(γ(p′)) : $= \{\text{p} \in DB \mid \text{p}'$是 p$\}$ 的投影（参见图2中的数据点 p $=(i,\ j,\ k)$）。现在，我们通过找出所有充分的模型系数，并在相应的数据点集合上计算每个模型系数，来计算 log $\hat{y}(p)$，进而计算偏差$|y(\text{p}) - \hat{y}(\text{p})|$ 。例如，对于 $\gamma()$，我们将完整玩具数据立方体视为空元组 ()。它是每个数据点的投影

因此类似于全局效应。最后，我们计算：

$$\log \hat{y}(\mathrm{p}) = \gamma(i,j) + \gamma(i,k) + \gamma(j,k) + \gamma(i) + \gamma(j) + \gamma(k) + \gamma(). \tag{2}$$

为获得标准化的异常值分数，我们将所得绝对残差 $|y(\mathrm{p}) - \hat{y}(\mathrm{p})|$ 除以适当的离散度度量 $\sigma = \sigma(\mathrm{p})$，具体选择见第4节。与阈值 τ 比较后，得到离群程度分数

$$\mathrm{SelfExp}(y(\mathrm{p})) = \max\left(\frac{|y(\mathrm{p}) - \hat{y}(\mathrm{p})|}{\sigma(\mathrm{p})} - \tau, 0\right). \tag{3}$$

尽管我们的框架更为通用，但我们仍沿用原论文 [33]中的 SelfExp 概念。如果归一化残差超过某个阈值 τ，即 $\mathrm{SelfExp}(y(\mathrm{p})) > \tau$，则数据点 p 被视为异常值。通常， τ 设为 2.5。这是因为，对于正态分布，几乎 99% 的数据都位于均值上下 2.5 倍标准差的范围内。

4 DataCube 单元格的稳健估计

基于第4节中引入的方程，我们现在可以通过为 $\hat{y}$ 和 σ指定具体的估计器，来定义不同的异常值检测方法。对于 SelfExp [33], ，模型系数通过求平均计算。我们从由 $\gamma()$ 表示的一般效应入手，计算数据立方体中所有值 $y(\mathrm{p})$ 的 12.5% 截尾均值（tm）。使用 tm 是为了提高稳健性，它会舍弃最大和最小的 12.5% 数值。回到 $\mathrm{p} = (i, j, k)$ 的示例，可按如下方式计算投影到一个维度上的模型系数

$$\gamma(i) = \underset{\mathrm{p}'=(j,k)}{\mathrm{tm}} \log y(\mathrm{p}) - \gamma(),$$

以及对于二维系数 $\gamma(i, j)$，其计算方式为

$$\gamma(i,j) = \underset{\mathrm{p}'=(k)}{\mathrm{tm}} \log y(\mathrm{p}) - \gamma(i) - \gamma(j) - \gamma().$$

所有其他系数均以类似方式计算，例如，系数 $\gamma(i, k)$ 是通过对时间维度上所有汇总销售数据的对数求平均，然后分别减去全局系数 $\gamma()$ 和所有依赖于 i 的系数 k 得到的。如前所述，计算 $\hat{y}(\mathrm{p})$ 的方式类似于方差分析；更具体地说，在本例中，它是一个三因素方差分析，包含二阶交互效应 [34]。同样的构造模式也适用于更高维度的系数。

接下来 [33],我们通过以下方式定义 $\sigma(\mathrm{p})$

$$\sigma(\mathrm{p})^2 = (\hat{y}(\mathrm{p}))^{\rho}. \tag{4}$$

这里， ρ 是在假设数据服从均值为 $\hat{y}(\mathrm{p})$ 的正态分布的前提下，依据最大似然原则选取的。由此可推得， ρ 必须满足以下方程 [33]:

$$\sum_{\mathrm{p}} \left(\frac{(y(\mathrm{p}) - \hat{y}(\mathrm{p}))^2}{(\hat{y}(\mathrm{p}))^{\rho}} \cdot \log(\hat{y}(\mathrm{p})) - \log(\hat{y}(\mathrm{p})) \right) = 0, \tag{5}$$

据我们所知，SelfExp 估计量是迄今为止数据立方体中稳健离群点检测的唯一实现。但它从未经过系统评估或检验。截尾均值作为一种估计量虽然直观，但计算稳健估计量或许还有更有效的方法。下文将进一步讨论 RODD 的其他方法。为了将 SelfExp 方法与其他方法区分开来，我们用 $\hat{y}_{\mathrm{S75}}$ 表示 SelfExp（销售值）估计量，因为计算截尾均值时它会忽略 25% 的观测值。

自 1998 年 SelfExp 发表以来，机器学习回归方法有了长足发展。据报道，机器学习模型往往优于线性回归方法（后者是方差分析的基础），并且在研究人员和从业者中日益普及 [2]。一项重要创新是随机森林 [10]的引入。Breiman [3] 提出的随机森林是一种采用装袋法的集成回归器。它由多棵决策树组成，这些树基于从数据集中抽取的样本进行训练。每棵树都会预测一个值，最后一步则是计算所有这些值的平均值，并将其作为最终预测结果。随机森林回归器已应用于多个场景，并取得了令人鼓舞的成果，例如 [7, 9, 16, 24]。针对我们的目的，我们在整个数据立方体上训练随机森林。我们提出将基于随机森林的估计量作为 $\hat{y}_{\mathrm{RF}}$ 的一种实现，并据此实现了 RODD-RF 检测算法。

$\hat{y}_{\mathrm{RF}}$ 的一个缺点是，与基于均值的估计器相比，它的计算量更大。因此，我们还将其他计算量较小的估计器作为主要竞争方法。如前所述，截尾均值非常直观。此外，它易于计算，并且在稳健统计中颇受欢迎[40]。SelfExp 方法采用 12.5% 的截尾比例，因此使用 75% 的数据计算均值。 [33] 并未说明选择这一截尾比例的依据，其他选择也可能同样合理，甚至更为合理。例如，有研究表明，使用 90% 数据的截尾均值在多项研究中取得了良好的结果 [13, 30, 35]。因此，我们提出了 $\hat{y}_{\mathrm{S90}}$，一种截尾比例为 5% 的截尾均值估计器。此外，我们还评估了截尾比例为 20% 的截尾均值估计器 $\hat{y}_{\mathrm{S60}}$ 。最后，我们还考虑将中位数作为稳健位置度量 $\hat{y}_{\mathrm{Median}}$。

5 模拟研究 – 实验设置

为了评估所有已介绍的 RODD 实例的性能，我们开展了一项模拟研究。由于缺少异常值的真实值，无法使用真实世界的数据立方体，因此我们使用了合成数据。研究重点是比较估计器 $\hat{y}_{\mathrm{S75}}$、$\hat{y}_{\mathrm{RF}}$、$\hat{y}_{\mathrm{S90}}$、$\hat{y}_{\mathrm{S60}}$

和 $\hat{}y_{\text{Median}}$ 在预测数据立方体单元格的值时的表现。我们计算了分类指标：灵敏度、特异度和准确率。然而，这三个评估指标的缺点是它们依赖于阈值；在本例中，阈值设为 2.5。为避免依赖该阈值，我们还测量了 ROC 曲线下面积（AUC）。ROC 曲线展示了灵敏度与 (1 - 特异度) 之间的权衡关系，并随阈值参数 τ 的变化而变化。若 AUC 分数为 1，则分类结果完美；若分数为 0.5，则分类效果不比随机猜测好。

此外，我们使用方差分析对离群点检测方法进行统计比较 [34]。我们检验了这些方法的效应，以及其他实验参数对 AUC 分数的效应。方差分析是一种统计方法，用于确定各组均值之间是否存在显著差异 [34]。它与 t 检验类似，但能够比较两个以上的组。

据我们所知，目前没有公开可用且包含异常值真实值的数据立方体，因此我们合成了数据立方体。下文将说明合成过程，并提供一个示例以便更好地理解。我们合成了八个数据立方体，每个都包含三个类别维度。维度数量是根据销售数据中离群点检测的常见应用示例选定的（见第 3 节）。每个维度中不同值的数量，是根据一家德国家用电器销售公司的真实数据集确定的。表 1 分步展示了数据立方体单元格的构建过程。首先，为每个维度的每个值从一个随机数值范围中选取一个期望值（见表 1a）。每个维度以及每个数据立方体的取值范围都各不相同。在示例中，吸尘器这一产品的月销售额期望值为 98，大阪这一城市的期望值为 110，一月这一月份的期望值为 91。由于实际销售数据中存在交互效应，我们合成了两个不同维度的值组合对应的期望值（见表 1b）。随后，计算期望值与交互效应的算术平均值，并将结果四舍五入（见表 1c）。某些组合存在交互效应，而另一些则没有，因为噪声经过四舍五入后通常为零。计算大阪的吸尘器期望值时，将两个维度的期望值 98 和 110 相加，再加上交互效应 4，最后将总和除以 2。四舍五入后的结果为 106。一月吸尘器销售额的期望值则是 98 和 91 的平均值四舍五入后的结果，因为这一特定产品与月份之间不存在交互效应。计算一月大阪的销售额期望值时，将 110、91 和 4 相加，再除以 2，结果为 103。最后，考虑每个类别的期望值和交互效应，计算并四舍五入得到一个单元格的值（见表 1c）。由于该值与单元格的期望值完全一致，因此可称其为无噪声值。根据这一计算过程，期望值近似服从正态分布。

为生成异常值，从数据立方体中抽取了一个小样本。根据剩余单元格，计算了算术平均值和四分位距

维度 类别	值	
产品	VC	98
地点	大阪	110
地点	柏林	87
Time	一月	91
Time	二月	93

(a) 随机选择各维度的预期值。

组合	交互值	
VC ∩大阪	4	106
风险投资 ∩柏林	0	93
风险投资 ∩一月	0	95
风险投资 ∩二月	-3	94
大阪 ∩ 一月	4	103
大阪 ∩ 二月	0	102
柏林 ∩ 一月	0	89
柏林 ∩ 二月	5	93

(b) 交互效应的随机选择。

产品位置		Time	计算	值
VC	大阪	一月	（98 + 110 + 91 + 106 + 95 + 103）/ 6	101
VC	大阪	二月	（98 + 110 + 93 + 106 + 94 + 102）/ 6	101
VC	柏林	一月	（98 + 87 + 91 + 93 + 95 + 89）/ 6	92
VC	柏林	二月	(98 + 87 + 93 + 93 + 94 + 93) / 6	93

(c) 无噪声数据立方体的计算，单元格值经过了四舍五入。

表 1：人工构建无噪声数据立方体的示例。

针对每个分类变量的每个取值分别进行计算。将四分位距的 1.5 倍从相应的算术平均值中减去并加上，由此得到每个维度的每个取值对应的两个“异常值边界”。对于每个单元格，选取与原始无噪声单元格值最接近的四舍五入后的异常值边界，并用其替换原始值。随后，若分别考察各个维度，样本中的值将根据四分位距法被分类为异常值。为使数据更加真实，向异常值和内点都添加了整数噪声。噪声的计算方式为：一个随机数乘以数据立方体中所有值的标准差的一小部分$\in \{0.25\%, 1\% or 5\%\}$。为了在添加噪声时考虑销售值的分布，对每个数据立方体分别计算标准差。然后，标准差分别除以 2.5、5、7.5、10 和 12.5，再乘以一个介于 -10 和 10 之间的随机整数。将数据立方体划分为异常值和内点时，使用了 30 个不同的随机种子。最终得到 3600 个数据立方体，并在其上测试了 RODD方法。

对于我们的 RODD-RF方法，我们使用了 Python 库 sklearn 中 RF 回归器的实现 [29]。出于计算方面的考虑，我们使用 50 次随机搜索，在 120 个数据立方体的小型子集上对 RF 的 6 个参数进行超参数调优（见下文）。所得的最佳参数用于所有其他 RF。这样做符合实际情况，因为在每次应用 RODD-RF之前都重新进行参数调优会耗费过多时间。调优后的 RF 包含 1500 棵树。特征数量被限制为数据集中特征总数的平方根。

RF 的最大深度设为 60，最小样本分割数设为 5。叶节点所需的最少样本数为 1。其他参数均采用默认值。构建树时使用了自助采样样本。由此形成了一个规模庞大的森林，其中包含许多深树。每棵（深）树都容易严重过拟合其训练样本。这样一来，树便能捕捉数据中的许多模式。只有当样本包含异常值时，才可能导致单元格值的估计出现偏差。不过，之后通过对大量树求平均，可以纠正这种过拟合，从而有效检测异常值。

6 模拟研究 – 结果

实验表明，并不存在一种普遍优越的方法，但对于每项指标，都有另一种 RODD 方法在离群点检测方面表现最佳（见表 2）。使用 $\hat{}y_{\mathrm{S60}}$时，灵敏度最高。使用 $\hat{}y_{\mathrm{Median}}$则获得了近乎完美的特异度。$\hat{}y_{\mathrm{S75}}$ 和 $\hat{}y_{\mathrm{S90}}$ 两种方法的准确率同样很高。采用 RODD-RF 时，AUC 分数最佳。$\hat{}y_{\mathrm{S60}}$、$\hat{}y_{\mathrm{S75}}$ 和 $\hat{}y_{\mathrm{S90}}$ 之间的差异通常较小。

估计量	灵敏度	特异度	准确率	AUC 分数
$\hat{y}_{\mathrm{S75}}$	0.2079	0.9900	**0.9817**	0.6985
$\hat{y}_{\mathrm{S60}}$	**0.2116**	0.9984	0.9813	0.6998
$\hat{y}_{\mathrm{S90}}$	0.1899	0.9993	**0.9817**	0.6947
$\hat{y}_{\mathrm{Median}}$	0.0032	**0.9998**	0.9790	0.5848
$\hat{y}_{\mathrm{RF}}$	0.2012	0.9951	0.9778	**0.7222**

表 2：五种不同 RODD方法的评估指标。

为评估实验参数的影响，我们进行了析因方差分析。由于模拟次数很多，所有 p 值在数值上几乎为零。 F 值表明，所选估计量 $\hat{}y$、基准数据集、噪声和离群值百分比均会影响离群值检测性能。根据方差分析的结果，我们针对每个参数进行了 t 检验。表 3 展示了估计量对 AUC 分数以及噪声水平影响的 t 检验结果（此处省略了基准数据集的检验结果）。$\hat{}y_{\mathrm{S60}}$ 和 $\hat{}y_{\mathrm{RF}}$ 的 AUC 分数均高于作为参考模型的 $\hat{}y_{\mathrm{S75}}$。不过，$\hat{}y_{\mathrm{RF}}$ 是正向效应最大的估计量。此外，它的结果甚至与 $\hat{}y_{\mathrm{S75}}$ 的结果存在显著差异，错误水平为 0.1% 。$\hat{}y_{\mathrm{S60}}$ 的 p 值为 0.2376，因此与 $\hat{}y_{\mathrm{S75}}$的差异不显著。

对于噪声对 AUC 分数影响的 t检验，以噪声非常多的子集作为参照。显然，噪声量越少，AUC 分数越高。对于数据集中异常值的数量，

估计量	效应
$\hat{y}_{\text{S60}}$	0.0014
$\hat{y}_{\text{S90}}$	-0.0038^{**}
$\hat{y}_{\text{Median}}$	-0.1136^{***}
$\hat{y}_{\text{RF}}$	0.0256^{***}

噪声	效应
大量噪声	0.0598^{***}
中等噪声	0.1476^{***}
少量噪声	0.2379^{***}
极少量噪声	0.3033^{***}

异常值百分比	效应
1%	-0.0011
5%	-0.0206^{***}

表 3：对所选 RODD 方法在曲线下面积（左）、噪声水平（中）和离群值百分比（右）方面的效应进行 t 检验。*** 表示在 0.1% 的显著性水平上显著，** 表示在 1% 的显著性水平上显著。

以异常值占比为 0.25% 的子集作为参照。正如预期，异常值越多，就越难以找到它们，即 AUC 分数会降低。

为了直观呈现不同的 AUC 分数，我们对其进行了调整。方差分析和 t 检验表明，实验参数会影响 AUC 分数，因此可能使结果产生偏差。对于每种参数组合，我们计算了所有方法和随机种子对应的平均 AUC。随后，从每个单独的 AUC 分数中减去相应的平均值。图3展示了由此得到的比较结果。箱线图直观呈现了估计器 $\hat{y}_{\text{S75}}$、$\hat{y}_{\text{S60}}$ 和 $\hat{y}_{\text{S90}}$ 的相似性，也显示了 RODD-RF（$\hat{y}_{\text{RF}}$）的优越性。估计器 $\hat{y}_{\text{Median}}$ 得到的 AUC 分数明显较低。

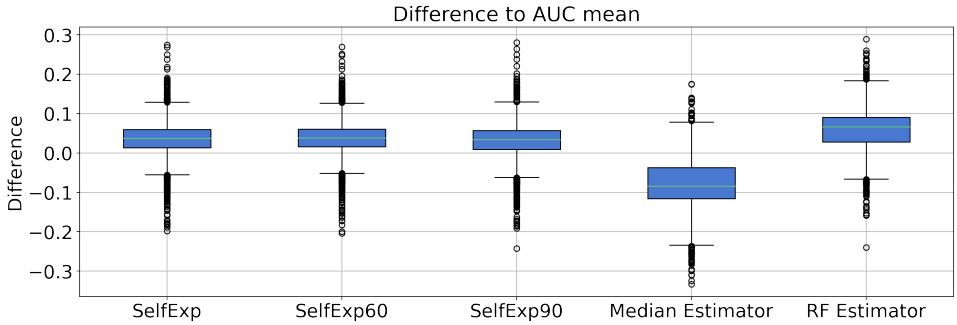


图 3：比较五种不同 RODD方法的 AUC 分数的箱线图。

除了评估RODD方法的性能外，还评估了它们的运行时间。$\hat{y}_{\text{S75}}$、$\hat{y}_{\text{S60}}$、$\hat{y}_{\text{S90}}$ 和$\hat{y}_{\text{Median}}$ 估计器的运行时间非常相近。它们的最短运行时间为0.13秒，平均运行时间介于0.39至0.43秒之间。最长运行时间介于1.00至1.49秒之间。计算$\hat{y}_{\text{RF}}$则需要多得多的时间。其最短运行时间为4.08秒，平均为37.92秒，最长为1592。05秒。随机森林始终需要相当可观的运行时间，但模拟研究中所用随机森林模型的超参数大幅增加了计算时间。例如，森林中的树的默认数量为100，但根据超参数调优结果，我们将其设为1 500。不过，未经参数调优的随机森林性能会显著降低。

真实世界的数据立方体可能远大于测试数据立方体。在这种情况下，建议评估AUC分数的提升是否值得增加计算开销。运行时间可能长达1592秒（26分钟），这可能导致该方法无法使用，尤其是在应用频繁时。

7 真实数据应用

我们还在一家德国家用设备销售公司的真实世界数据集上验证了RODD方法。表4的前五列概述了所提供的部分数据。

Year	月份	产品	分布渠道	实际销售额	预期销售额	SelfExp	偏差
2017年1月		产品3 渠道1		5182	3917	2.8	+32.3%
2017年2月		产品3 渠道1		4288	2547	3.9	+68.4%
2017	三月	产品3 渠道1		5056	3675	3.1	+37.6%
2017年4月		产品2 渠道1		4099	2822	2.8	+45.3%
2019年11月		产品1 渠道1		3287	5069	3.9	-35.2%
2019年11月		产品1 渠道2		4883	2379	5.5	+105.3%

表4：家用设备销售数据立方体的子集，以及通过 RODD-RF方法得出的预期销售额和 SelfExp 值。

数据包含三种不同产品通过三种不同分销渠道（例如本地商店和网店）销售的销售额信息。2017、2018 和 2019 年的销售数据按月汇总，共得到 324 个数据点。由于公司本身没有关于数据中异常值的信息，因此无法假定真实值，我们只能对识别出的异常值进行合理性检查。RODD-RF 方法发现了 17 个异常值，其中 12 次销售额高于预期，5 次低于预期。在调查这些异常值时，公司专家为每个异常值都找到了合理的解释。有两次，营销活动将销售额从一个渠道转移到另一个渠道，导致某一产品在一个渠道的销售额异常偏高，而同一产品在另一渠道的销售额则有所下降。第一种情况下，与期望值相比，渠道 1 的销售额增长了 105%，而渠道 2 的销售额则减少了 35%（见表4第5行和第6行）。此外，针对某种产品的直接折扣以及套装降价都被证明是非常有效的营销措施。直接折扣使销售额增加了 45%（见表4第4行），套装优惠则使销售额增加了 32% 至 68%（见表4第1至3行）。

8 结论

我们提出了 RODD，这是一种用于数据立方体稳健离群点检测的通用框架。RODD 将当前的黄金标准——所谓的 *SelfExp* 方法[33], 纳为一个特例。作为首篇研究成果，我们系统比较了用于离群点检测的 4 种新型稳健估计量。它们在单元格值的统计估计方法上各不相同。 [33]采用 12.5% 的截尾百分比进行截尾均值估计（$\hat{}$ y_{S75}）；我们则分析了基于 5%（$\hat{}y_{\mathrm{S90}}$）和 20% 截尾比例（$\hat{}y_{\mathrm{S60}}$）、中位数（$\hat{}$ y_{Median}）以及随机森林（$\hat{}y_{\mathrm{RF}}$，RF-RODD）的方法。我们在一项模拟研究中，对 3600 个数据立方体比较了全部五种方法。以 AUC 分数作为质量指标，我们发现新提出的 RODD-RF 表现最佳。与传统的 *SelfExp* [33], 相比，这一效应具有高度显著性（p 值 $< 0.1\%$）。其他方法对 AUC 的正向效应并不那么明显（$\hat{}y_{\mathrm{S60}}$），甚至呈负向效应（$\hat{}y_{90}$，$\hat{}y_{\mathrm{Median}}$）。

考虑到这些结果，我们将 ROOD-RF 应用于一个真实世界数据集。该数据集没有异常的标签。然而，结果表明，对于算法发现的所有异常值，人类专家也都能结合领域知识给出解释。

未来，我们计划纳入包含三个以上类别维度的高维数据立方体。此外，还可以考虑采用其他方法来标准化实际值与估计值之间的差异，例如基于分位数的方法。另外，我们认为层次学习器具有很大潜力，但机器学习文献对其探索还不够。

致谢。 本研究得到可信数据科学与安全研究中心（鲁尔大学联盟下属机构）的支持。

参考文献

1. Andrews, J.T., Morton, E.J., Griffin, L.D.: 使用自编码器检测异常数据。《国际机器学习与计算杂志》**6**(1), 21 (2016) 2. Ardabili, S., Mosavi, A., V´arkonyi-K´oczy, A.R.: 机器学习建模领域的进展：混合与集成方法综述。载于：全球研究与教育国际会议。第 215–227 页。施普林格 (2020) 3. Breiman, L.: 随机森林。机器学习 **45**(1), 5–32 (2001) 4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF：识别基于密度的局部离群点。载于：2000年 ACM SIGMOD 数据管理国际会议论文集。第 93–104 页 (2000) 5. Chalapathy, R., Menon, A.K., Chawla, S.: 使用单类神经网络检测异常。arXiv预印本 arXiv:1802.06360 (2018) 6. Chen, J., Sathe, S., Aggarwal, C., Turaga, D.: 使用自编码器集成进行离群点检测。载于：2017年 SIAM 数据挖掘国际会议论文集。第 90–98 页。SIAM (2017)

7. Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: 使用随机森林回归投票进行稳健且准确的形状模型拟合。载于：欧洲计算机视觉会议。第 278–291 页。施普林格 (2012) 8. Dempster, A.P., Laird, N.M., Rubin, D.B.: 通过EM算法从不完整数据中进行最大似然估计。英国皇家统计学会会刊：B辑（方法论）**39**(1), 1–22 (1977) 9. El Mrabet, Z., Sugunaraj, N., Ranganathan, P., Abhyankar, S.: 基于随机森林回归器的方法，用于检测电力系统中的故障位置和持续时间。《传感器》**22**(2), 458 (2022年) 10. Friedman, J.H.: 预测（机器）学习的最新进展。《分类学杂志》**23**(2), 175–197 (2006) 11. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., Pirahesh, H.: 数据立方体：一种关系聚合算子，可泛化分组、交叉表和小计。《数据挖掘与知识发现》**1**(1), 29–53 (1997) 12. Hawkins, S., He, H., Williams, G., Baxter, R.: 使用复制器神经网络进行离群点检测。载于：数据仓库与知识发现国际会议。第 170–180 页。施普林格 (2002) 13. Hill, M., Dixon, W.: 现实生活中的稳健性：临床实验室数据研究。《生物计量学》第 377–396 页 (1982) 14. Hochkamp, F., Rabe, M.: 数据挖掘中的离群点检测：排除错误还是丢失信息？载于：汉堡国际物流会议 (HICL) 2022。《汉堡国际物流会议 (HICL) 论文集》 (2022) 15. Holst, A.: 2010年至2025年全球创建、采集、复制和消费的数据/信息量。Statista，6月 (2021) 16. Huang, H., Pouls, M., Meyer, A., Pauly, M.: 使用基于树的集成方法进行旅行时间预测。载于：国际计算物流会议。第 412–427 页。施普林格 (2020) 17. Jin, W., Tung, A.K., Han, J., Wang, W.: 使用对称邻域关系对异常值进行排序。载于：亚太知识发现与数据挖掘会议。第 577–593 页。施普林格 (2006) 18. Knorr, E.M., Ng, R.T.: 统一的异常值概念：性质与计算。载于：KDD。第 97 卷，第 219–222 页 (1997) 19. Latecki, L.J., Lazarevic, A., Pokrajac, D.: 使用核密度函数进行离群点检测。载于：国际模式识别机器学习与数据挖掘研讨会。第 61–75 页。施普林格 (2007) 20. Liu, F.T., Ting, K.M., Zhou, Z.H.: 孤立森林。载于：2008年第八届 IEEE 国际数据挖掘会议。第 413–422 页。IEEE (2008) 21. Liu, F.T., Ting, K.M., Zhou, Z.H.: 基于孤立的异常检测。《ACM知识发现与数据汇刊》(TKDD) **6**(1), 1–39 (2012) 22. Mensi, A., Cicalese, F., Bicego, M.: 使用随机森林距离进行离群点检测。载于：国际图像分析与处理会议。第 75–86 页。施普林格 (2022) 23. Mohandoss, D.P., Shi, Y., Suo, K.: 使用随机森林分类器预测异常值。载于：2021年 IEEE 第11届年度计算与通信研讨会暨会议 (CCWC)。第 0027–0033 页。IEEE (2021) 24. Nakashima, H., Arai, I., Fujikawa, K.: 基于随机森林回归器的乘客计数器，使用公交车上的行车记录仪和传感器。载于：2019年 IEEE 国际普适计算与通信研讨会 (PerCom Workshops)。第 561–566 页。IEEE (2019)

25. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: 深度半监督学习算法的真实评估。神经信息处理系统进展 **31** (2018) 26. Park, C.M., Jeon, J.: 使用自变量合成对传感器测量值进行基于回归的离群值检测。载于：国际数据科学会议。第 78–86 页。施普林格 (2015) 27. Pauleen, D.J., Wang, W.Y.: 大数据是否意味着大知识？关于大数据与分析的知识管理观点。《知识管理杂志》 (2017) 28. Pavlidou, M., Zioutas, G.: 核密度离群值检测器。载于：《非参数统计专题》，第 241–250 页。施普林格 (2014) 29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn：Python 机器学习。《机器学习研究杂志》**12**, 2825–2830 (2011) 30. Rocke, D.M., Downs, G.W., Rocke, A.J.: 稳健估计量真的有必要吗？《技术计量学》**24**(2), 95–101 (1982) 31. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., M¨uller, E., Kloft, M.: 深度单类分类。载于：国际机器学习会议。第 4393–4402 页。PMLR (2018) 32. Ruff, L., Vandermeulen, R.A., G¨ornitz, N., Binder, A., M¨uller, E., M¨uller, K.R., Kloft, M.: 深度半监督异常检测。arXiv 预印本 arXiv:1906.02694 (2019) 33. Sarawagi, S., Agrawal, R., Megiddo, N.: 联机分析处理数据立方体的发现驱动探索。载于：国际数据库技术扩展会议。第 168–182 页。施普林格 (1998) 34. Searle, S.R., Gruber, M.H.: 线性模型。约翰·威立父子公司 (2016) 35. Spjøtvoll, E., Aastveit, A.H.: 田间试验数据中稳健估计量的比较。《斯堪的纳维亚统计学杂志》第 1–13 页 (1980) 36. St, L., Wold, S., et al.: 方差分析（ANOVA）。《化学计量学与智能实验室系统》**6**(4), 259–272 (1989) 37. Vargaftik, S., Keslassy, I., Orda, A., Ben-Itzhak, Y.: RADE：使用基于决策树的集成方法实现资源高效的监督式异常检测。《机器学习》**110**(10), 2835–2866 (2021) 38. Walfish, S.: 统计离群值方法综述。《制药技术》**30**(11), 82 (2006) 39. Wang, H., Bah, M.J., Hammad, M.: 离群值检测技术的进展：综述。《IEEE 开放获取期刊》**7**, 107964–108000 (2019) 40. Welsh, A.: 线性模型中的截尾均值。《统计年鉴》**15**(1), 20–36 (1987) 41. Yang, X., Latecki, L.J., Pokrajac, D.: 使用全局最优的基于样本的高斯混合模型进行离群点检测。载于：2009 年 SIAM 数据挖掘国际会议论文集。第 145–154 页。SIAM (2009) 42. Zhang, J., Zulkernine, M.: 使用无监督离群点检测的基于异常的网络入侵检测。载于：2006 年 IEEE 国际通信会议。第 5 卷，第 2388–2393 页。IEEE (2006) 43. Zhou, C., Paffenroth, R.C.: 使用稳健深度自编码器进行异常检测。载于：第 23 届 ACM SIGKDD国际知识发现与数据挖掘会议论文集。第 665–674 页 (2017) 44. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: 用于无监督异常检测的深度自编码高斯混合模型。载于：国际学习表征会议 (2018)