DQSOps：面向数据驱动应用的数据质量评分运维框架

Firas Bayram
firas.bayram@kau.se 数学与计算机科学系，卡尔斯塔德大学 瑞典卡尔斯塔德

Bestoun S. Ahmed
bestoun@kau.se 数学与计算机科学系，卡尔斯塔德大学 瑞典卡尔斯塔德；计算机科学系，捷克技术大学 捷克共和国布拉格

Erik Hallin Anton Engman
erik.hallin@uddeholm.com
anton.eng
man@uddeholm.com
Uddeholms AB 瑞典韦姆兰省哈格福什

arXiv:2303.15068v1 [cs.DB] 27 Mar 2023

摘要

数据质量评估已成为复杂数据驱动人工智能（AI）软件系统成功运行的重要组成部分。在实践中，实际应用会以极快的速度生成海量数据。这些数据流在永久存储或用于学习任务之前，需要经过分析和预处理。因此，人们高度重视对高质量数据集进行系统化管理和构建。然而，海量、高速数据流通常仍以人工方式（即离线）管理，这使其难以在生产环境中实施。为应对这一挑战，DataOps 应运而生，旨在运用 DevOps 原则实现数据流程的生命周期自动化。然而，在 DataOps 框架内，依据适配度量表确定数据质量是一项复杂的任务。本文提出了一种新颖的数据质量评分运维（DQSOps）框架，可在 DataOps 工作流中为生产数据生成质量评分。该框架采用两种评分方法：一种是基于机器学习预测的方法，用于预测数据质量评分；另一种是基于标准的方法，定期评估多个数据质量维度并生成真实值分数。我们在一个现实工业应用案例中部署了 DQSOps 框架。结果表明，与传统的数据质量评分方法相比，DQSOps 在保持较高预测性能的同时，实现了显著的计算加速比。

CCS 概念

• **计算方法** → **机器学习**；• **软件及其工程** → **软件开发技术**；**软件验证与确认**。

允许免费为个人或课堂用途制作本作品全部或部分内容的数字或纸质副本，但前提是不得为营利或商业利益而制作或传播副本，且副本首页须附上本声明和完整引文。对于本作品中由 ACM 以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，须事先获得特别许可和/或支付费用。请向 permissions@acm.org 申请许可。*EASE 2023*，*2023*年*6*月*14*日至*16*日，芬兰奥卢© 2023 计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/XXXXXXX.XXXXXXX

关键词

自动化数据评分、DataOps、数据评估、数据质量维度、变异测试

ACM 引用格式：
Firas Bayram、Ahmed、Erik Hallin 和 Anton Engman。2023。DQSOps：面向数据驱动应用的数据质量评分运维框架。载于 第*24*届软件工程评价与评估国际会议论文集（*EASE 2023*）。ACM，纽约，纽约州，美国，10页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

在大数据爆发的时代，数据已成为人工智能（AI）软件项目最关键的资产 [41]。主要原因是，在机器学习（ML）软件的整个生命周期中，数据始终相伴而行。值得注意的是，各行业和企业正在以不断加快的速度收集海量数据。数据的规模和流速给生产环境中的机器学习软件系统带来了严峻挑战，因为这些系统需要做出（近乎）实时的决策 [53]。在现实场景中，这一挑战有两方面：评估系统中流入数据的质量，并在规定的时间范围内对其进行处理。

数据质量并非新话题，其根源可追溯至数据库管理系统相关文献 [56]。然而，数据驱动系统的发展近来使这一话题转向人工智能研究 [5]。人们普遍认为，机器学习项目的性能主要取决于其底层数据的质量 [26]。数据质量是一个多面向的概念，它根据问题的相对性质定义了多个质量维度 [42]。通常，数据质量维度依据具体情境和领域来确定，这意味着数据质量维度可能因应用而异。数据质量的每个维度都衡量数据在某一特定方面的状况，通常以质量指数或度量指标加以概括，并用作数据有效性的指示器 [54]。

数据质量评分在数据质量评估中具有重要意义，其定义是：针对数据集 [52]中给定观测集合的数据质量各个维度，获取评分指标的方法。在数据质量评分领域，研究人员探讨如何衡量和量化目标应用中数据质量的不同维度。评分被视为适用于每条数据记录或每个数据块的特定数据质量维度的适配度量表。数据质量分数用于数据存储管理，或通过引入

可接受性阈值 [38]应用于预测性机器学习软件。可接受性阈值会筛除不符合质量标准的数据实例。因此，数据质量评分方法能够通过定量分析，根据数据记录的质量对其进行表征。

传统的数据质量评分流程通过代码执行，人工检查并验证数据在既定数据质量维度 [51]方面的状况。然而，这种传统流程耗费大量资源，尤其是在数据质量维度数量较多时，而且通常需要人工干预 [43]。此外，在部署环境中，这种人工流程并不适用于应自动执行任务的机器学习软件 [33]。近年来，DevOps 原则——持续集成和持续交付（持续集成/持续交付）——已被采用并引入，用于构建可投入生产环境的机器学习软件，这一做法称为机器学习运维 [28]。机器学习运维系统使人工智能应用的整个生命周期中都能以规范化的方式部署机器学习软件 [7]。机器学习运维不可或缺的组成部分是 DataOps，它是一种系统化、持续地管理和处理生产数据的方法。数据运维流水线旨在准备准确、可靠的数据，供机器学习模型使用。然而，对流数据窗口进行评分颇具挑战，在高采样率为特征的生产系统中，如何在 DataOps 的背景下开展此项工作仍有待探索。

本文提出了一种新颖的数据质量评分运维（DQSOps）框架，可融入 DataOps 工作流，从而推动数据运维领域的发展。DQSOps 可视为一种通用的数据评分方法，用于自动评估流数据的质量。DQSOps 的通用性得益于可配置组件，这些组件可根据问题轻松调整和扩展。在实际应用中，无论定义了多少个质量维度，DQSOps 都能加快对采集到的生产数据进行质量评分。它通过采用一种创新的机器学习预测器来确定处理后的数据窗口的质量评分，从而实现加速。与基于标准的数据评分方法相比，使用机器学习预测器可大幅缩短总体处理时间。不过，为了在系统演化过程中持续保持机器学习预测器的高性能，我们实现了一个测试预言机，用于持续评估机器学习准确率。测试预言机使用基于标准的方法定期生成的数据质量真实值评分。此外，我们还受变异测试原则启发，引入了一种新的数据变异组件，用于模拟数据窗口中的数据质量问题，从而简化框架的初始化阶段。

本文其余部分的结构如下：第2节回顾该领域的相关工作，并概述数据质量评分任务。第3节介绍用于开发数据质量评分框架的方法。第4节报告各用例的实验结果。第5节讨论将所提出的框架推广到其他场景时可能面临的威胁。最后，第6节总结全文的主要观点，并展望未来。

2 背景与先前研究

本节概述数据质量维度及相关定义。随后，我们回顾文献中致力于评估数据质量的研究。

2.1 数据质量维度

数据质量维度是为评估数据质量而定义的。每个维度都捕获了特定任务中所评估数据的一项特定特征。Knight [31] 提出了一个关于信息系统质量的综合概念框架。机器学习软件关注信息系统中系统数据与任务属性之间的关系。因此，情境质量与构建机器学习解决方案相关，因为它是在系统数据与任务属性之间的交互空间中确定的 [19]。

情境质量包含五个质量维度：附加值、相关性、及时性、适当数据 和完整性 [31]。然而，由于这些质量维度基于情境，因此必须根据任务和系统的特征进行调整。Wand和Wang [55]指出，数据质量维度具有主观性，应在生成数据的系统应用中加以定义。在一项较新的研究中， [52], 这些维度经过了修订，以用于持续质量管理。修订后的数据质量维度包括：准确性、完整性、一致性、及时性。然而，对于不断演化的机器学习软件系统，在系统演化过程中很可能会发生分布变化 [39]。因此，我们将扩展这些质量度量，并纳入偏斜度 [15] ，以监控传入数据的分布。以下定义对这些数据质量维度进行了说明：

1. **准确性：** 衡量观测到的数据值是否代表实际值。2. **完整性：** 检查观测数据是否包含缺失值。3. **一致性：** 验证观测值是否符合该领域的完整性约束。4. **及时性：** 描述数据对于相应任务是否为最新数据。5. **偏度：** 计算观测数据相对于参考分布的分布偏差。

2.2 相关工作

数据质量评估是早期文献中许多研究的主题 [45]。Chug *et al.* [15] 提出了一种量化给定数据集质量的方法。该方法使用九个数据质量维度来估算数据集的质量。因此，该方法会为数据集的数据质量提供评分、报告和标签。类似地，针对大数据系统，Taleb *etal.* [52]提出了大数据质量管理框架（BDQMF），基于多个数据质量维度解决大数据系统中的数据质量问题。该框架包含多个组件，用于管理、验证和监控数据质量。此外，研究还从数据集的单元实例和模式两个层面考察了数据质量问题。BDQMF 框架还会量化数据质量各方面的分数。然而，作者并未通过实验评估该框架。对于业务流程，

有人提出了数据质量验证方法（DQVM），用于评估数据质量对流程结果的影响 [10]。该方法允许领域专家根据各数据质量维度对业务流程的相对重要性，为其分配相应的权重。随后，观察无故障数据集与注入故障的数据集之间质量分数的偏差，并最终评估其对流程的影响。

其他方法研究了如何量化单个数据质量维度。Heinrich 等 [25] 提出了一种基于概率论的质量度量指标，用于评估语义一致性。该度量指标反映了被评估数据不存在矛盾的程度。类似地，研究者还探讨了最小性或唯一性，以衡量数据在数据级和模式级的冗余 [20]。该方法采用层次聚类和相似度计算技术来计算度量指标。针对物联网系统，Byabazaire 等 [9] 提出了一个实时数据质量评估框架。为衡量数据质量，该框架使用信任作为量化度量指标。研究人员使用皮尔逊相关系数计算信任评分与均方根误差（RMSE）和平均绝对误差（MAE）之间的相关性，以验证信任是否是数据质量的良好指标。鉴于针对生产系统中数据质量评分的研究方法有限，我们设计了一个可在现实问题场景中高效应用的框架。与为离线静态数据集设计的现有方法不同，所提出的框架能够处理采样率较高的流数据。此外，我们提出的框架还能有效支持数量众多的质量维度。

3 数据质量评分运作（DQSOps）框架

本节介绍实现我们所提出的 DQSOps 框架的方法。DQSOps 可通过定量分析，高效评估生产数据流的质量。DQSOps 的输出是一项评分指标，用作衡量所收集数据窗口整体有效性的指标，并根据对多个数据质量维度的评估计算得出。这些分数可用于按质量对系统数据进行排序。因此，可根据不同业务分析任务对数据质量的不同要求，选择相应的数据。在实践中，DQSOps 通过采用机器学习预测器取代传统的数据评分方法，大幅缩短了数据窗口的评分时间。运行时间不受质量维度数量的影响，因此非常适用于采样率较高的实际应用。此外，还会定期执行测试预言机，评估机器学习预测器的性能。

所提出框架的整体流程如图 1 所示。接收到来自数据生成源的数据流后，数据流会根据预先定义的数据窗口大小，被划分为包含数据样本的窗口。受变异测试原则的启发 [27], , 我们引入了一个数据变异体模拟器组件，以全面控制实验条件。该组件根据预先配置的参数，在数据窗口内模拟非等价数据变异体；这些参数从配置文件中加载，该文件存储了每个数据质量维度预先设定的变异比例。非等价变异体用于对问题施加有意义的改变 [50]。数据变异体的典型示例包括缺失或不一致的数据值，以及异常。因此，我们利用数据变异组件模拟错误数据，以便根据指定维度影响数据质量。此外，该组件在初始化阶段尤其有用，后文将在第3.3节进一步讨论。

准备好数据窗口后，将调用方法激活器组件，根据特定决策激活适当的数据评分工作流路径。选择方法的标准将在第3节 3.4 节中进一步详述。候选的数据质量评分方法包括：标准评分方法、回归机器学习模型或重新训练信号。标准评分方法会计算已定义维度的数据质量真实值评分，并将其存储在存储库中；而机器学习模型则预测每个数据窗口的数据质量评分。如果激活重新训练信号，就会使用存储库中保存的真实值分数重新训练机器学习模型。在重新训练过程中，当前的预测方法和基于标准的方法都会用于数据质量评分，直到新模型准备就绪并取代当前模型。此外，还会使用测试预言机持续监控回归模型的性能，该预言机负责评估预测的数据质量分数。接下来的小节将深入探讨 DQSOps 框架主要组件的运作机制。

3.1 数据质量维度评分

基于标准的方法是 DQSOps 的核心要素，用于量化数据质量维度的真实值分数，这些分数用于训练机器学习模型。每个数据质量维度都对应一个质量评分。具体而言，该评分用于量化所收集的数据窗口在特定数据质量维度上的适配程度。相关的数据质量维度包括：准确性、完整性、一致性、及时性和偏度，详见第 2.1 节。如图 1 所示，系统会从配置文件中加载元信息，以帮助计算数据质量维度的分数。例如，元信息可能包括用于表征特定问题完整性约束的值，如最大值和最小值限制。配置文件还可以包含辅助文件的路径，这些文件用于支持某些分数的计算，例如异常检测器或数据分布。计算出数据质量分数后，将所有分数值通过最小-最大归一化标准化至 0 到 1 之间，以确保各数据质量维度的尺度统一，详见下一节。数据质量分数的获取方式如下：

3.1.1 准确性评分。准确性评分通过计算处理后的数据窗口[52]: $Accuracy = \frac{NAV}{N}$中异常数据所占的比例得出，其中 NAV 是数据窗口中检测到的异常值总数， N 是数据窗口的大小。异常可能由多种因素导致，例如恶意活动、硬件故障、数据采集不准确或对抗性攻击 [12]。从数据质量的角度来看，异常数据记录代表不健康数据实例中的异常值，因此被视为低质量数据的一个指标

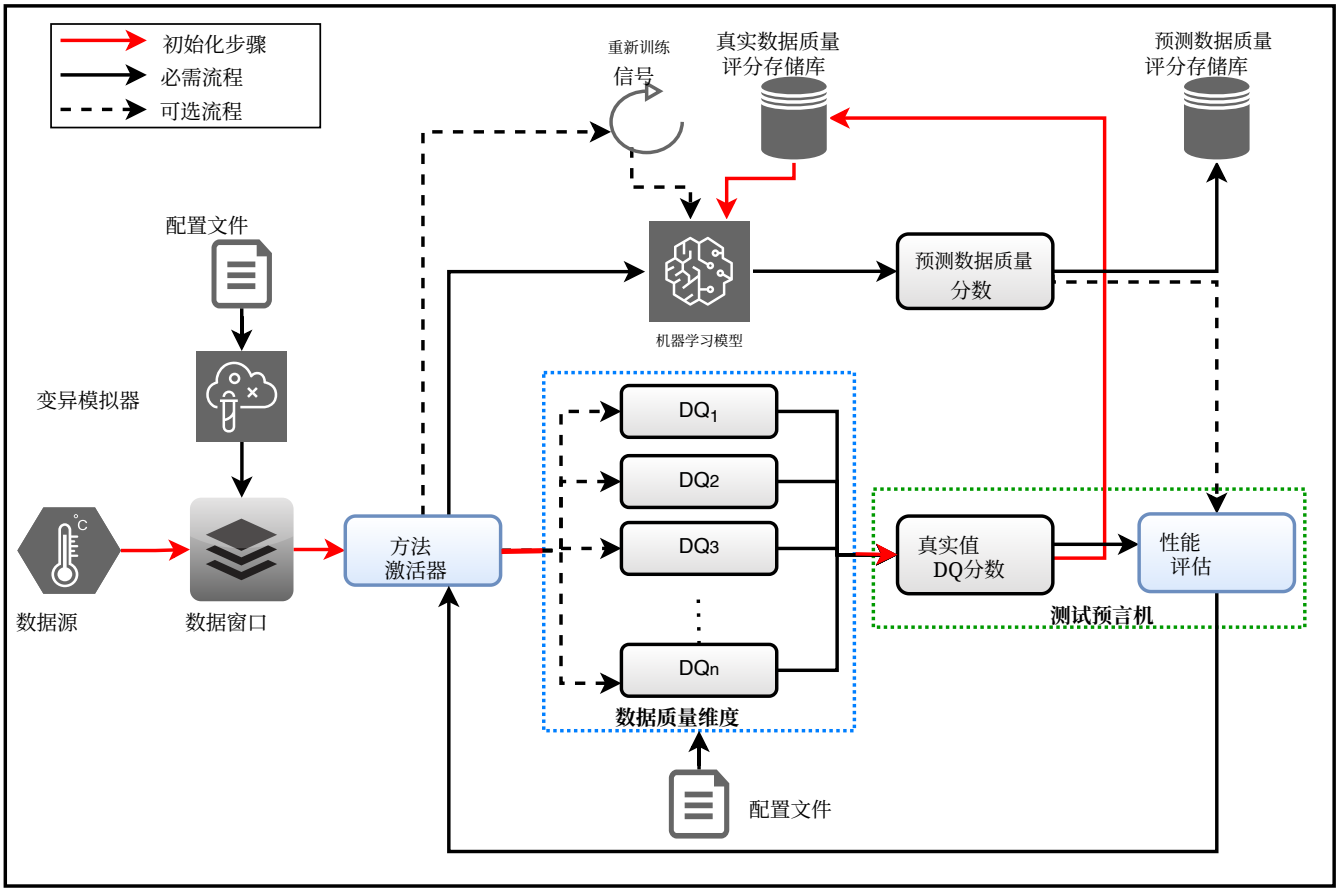


图 1：DQSOps框架，红色路径表示初始化阶段。

[49]。异常检测器可在初始化步骤中获取；参见第3.3节。

3.1.2 完整性评分。数据窗口中观测到的缺失值所占比例，可按如下方式计算 [24]:$Completeness = \frac{NNV}{N}$，其中 NNV是缺失值的数量，例如NA（不可用）或 NULL观测值，而 N是数据窗口的大小。缺失值是数据质量中常见的问题，可能表明数据源 [46]已断开连接或受损。

3.1.3 一致性评分。完整性约束因应用领域和应用条件而异。因此，需根据特定问题设置来定义。例如，在某些领域中，数据观测值不能为负数，或者其取值应处于特定范围内。定义数据值的完整性规则后，可将不满足完整性约束的数据值占比计算为一致性评分[47]: $Consistency = \frac{NCV}{N}$，其中 NCV 是一致的值的数量， N 是数据窗口的大小。数据一致性是数据质量的基本问题，因为它会检查数据冲突，从而检测数据记录过程中的错误 [22]。

3.1.4 及时性评分。及时性（或称数据时效性）是一种语义度量，用于描述数据是否符合应用领域 [29]。为量化数据的及时性，我们使用了拟合优度检验。拟合优度检验用于衡量当前数据是从底层数据生成分布的特定累积分布函数（cdf）或概率密度函数（pdf）中抽样得到的可能性 [2]。根据数据的性质，可以采用多种拟合优度检验技术 [17]。最常用的检验包括 Kolmogorov-Smirnov、Anderson-Darling 和 Cramér-von Mises 统计检验 [21]。每种检验都会计算检验统计量，用以评估数据与所比较分布的拟合程度。

在我们的实验中，我们使用双样本柯尔莫哥洛夫–斯米尔诺夫统计检验来计算拟合优度度量指标。柯尔莫哥洛夫–斯米尔诺夫检验是一种非参数统计工具，用于确定两个样本是否来自同一分布 [48]。采用柯尔莫哥洛夫–斯米尔诺夫检验的主要动机是，它适用于较小的子集，是一种有效的方法 [3],，这与我们的用例设置相符。对于两个独立随机样本 $X = X_1$、……、X_n和 $Y = Y_1$、……、Y_m，其经验累积分布函数分别为 $\hat{} F_1$ 和 $\hat{} F_2$，柯尔莫哥洛夫–斯米尔诺夫检验统计量

计算如下 [32]:

$$KS = \max_{1 \leqslant i \leqslant N} \left| \hat{F}_1(Z_i) - \hat{F}_2(Z_i) \right|, \quad (1)$$

其中， Z是 X和 Y的组合样本， $N = n + m$。

3.1.5 偏度评分。在实际应用中，尤其是物联网（IoT）领域，数据由传感器采集，分布漂移（或偏移）是系统中最常见的数据问题之一 [23]。数据流会根据导致数据分布偏度的分布偏差进行验证 [11]。为了计算分布偏度分数，可以计算散度大小，以衡量当前数据窗口与历史数据的分布之间的差异 [34]。计算散度的方法有很多；詹森-香农（JSD）散度和库尔贝克-莱布勒（KLD）散度是其中最常用的 [44]。

在我们的框架中，我们使用 JSD度量来计算偏度评分。JSD度量是对 KLD度量进行对称化的结果。JSD 的主要特性是其值有界，落在区间 $[0, 1]$, 内，而 KLD 的值可能为无穷大 [37]。根据 JSD，两个概率分布 P和 Q之间的不相似度按如下方式计算 [36]:

$$\mathrm{JSD}(P \| Q) = \mathrm{JSD}(P \| Q) := H\left(\frac{P+Q}{2}\right) - \frac{H(P) + H(Q}{2}, \quad (2)$$

其中，函数 H表示香农熵，其定义如下:

$$H(p) = -\int p(\mathbf{Y}) \log p(\mathbf{Y}) d\mathbf{Y}. \quad (3)$$

本节所定义的数据质量维度还有一点值得说明：可以根据任务引入其他定义。例如，若数据窗口包含相应的数据质量问题，则可将准确性、完整性和一致性评分定义为“1”，否则定义为“0” [6], ，而不是使用错误数据实例所占的比例。在医疗应用和自动驾驶汽车等容差水平较低、对故障数据要求严格的安全关键领域，可以采用这种定义 [13]。在我们的工业用例中，使用错误数据实例的比例更合适，因为我们不希望丢弃数据实例（即将其评分设为“1”）。

3.2 寻找综合数据质量评分

计算出数据质量维度的评分值后，会将这些值汇总为一个综合评分，用以表示数据窗口 [16] 的质量。汇总度量指标最基本的方法是对数据质量维度取算术平均数，或采用其变体，例如加权平均数。然而，有观点认为算术平均数无法进行合理的汇总 [24]。此外，汇总方法对变量的尺度很敏感。因此，聚合度量指标的计算会受大尺度维度主导，而小尺度维度的影响则微乎其微。

在汇总质量分数之前，会对矩阵 DQ_{nm} 进行标准化，该矩阵包含 n 数据实例在 m 维度上的质量分数。汇总前的标准化步骤至关重要，这样才能将不同的数据质量度量指标同等地纳入总体质量评分 [18]。标准化也称为白化，可通过计算矩阵 DQ_{nm}中每个元素 dq_{ij}的 Z 分数来实现：$z_{ij} = (dq_{ij} - \mu_j)\sigma_j$，其中$\mu_j$为均值，$\mu_j$为相应列的标准差。在计算聚合指标时，标准化过程会为不同的数据质量维度赋予相同的权重。

近期的一项调查 [53], 显示，基于主成分分析的方法是检测数据错误最常用的技术。因此，我们遵循相关文献，使用主成分分析（PCA）提取代表数据窗口整体质量的评分。PCA 是一种广泛使用的降维技术，可从数据中提取最关键的信息。PCA 使用正交变换，将d 维空间中的每个数据实例映射到 k 维空间，使得 $k < d$。第一主成分通过寻找方差最大的方向构造而成 [1]。数据集 X 的主成分可通过求解以下特征值问题得到:

$$\Sigma \mathbf{M} = \lambda \mathbf{M}, \quad (4)$$

其中， Σ 是协方差矩阵 $\sum = XX^T$ ，用于衡量原始数据 X中各变量之间的相关性； λ 是特征值向量， M 是特征向量矩阵，其中包含按特征值大小排序的主成分。

3.3 初始化阶段

在部署之前，DQSOps框架需要使用包含真实质量评分的训练数据，对机器学习预测器进行热启动准备。DQSOps的初始化阶段在图 1 中以红色路径表示。经过这一阶段，将生成一个可靠的机器学习预测器，以便在生产环境中投入实践。为了生成该预测器，需要如前文所述找到真实标签，并将其存储在一个存储库中，用于训练机器学习模型。系统会持续监控机器学习模型的预测准确率，直到其达到预先设定的阈值 τ；该阈值根据性能指标进行评估。达到此阈值后，机器学习预测器便会部署到实际问题中。在这一阶段，变异模拟器组件在学习任务中发挥着关键作用。变异体能够为模型提供更全面、更多样的数据质量问题样本，帮助模型学习识别现实环境中可能出现的数据质量问题。此外，与等待现实中可能并不常见的数据质量问题相比，该组件还能加快学习过程，因为等待这些问题出现可能会耗费大量时间。

除了生成机器学习预测器外，此阶段还会准备其他元信息文件，以供实践中使用。这些文件包括用于计算准确性评分的异常检测模型，以及用于计算偏斜评分的参考数据分布。应使用高质量干净数据准备这些文件，以提高真实质量评分的可靠性。随后，这些文件的路径会存储在配置文件中，以便在DQSOps框架最终部署到生产环境时加载。

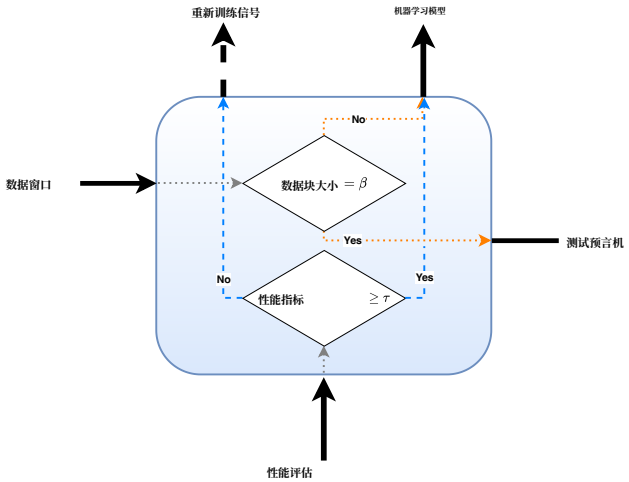


图 2：方法激活器组件流程图

3.4 方法激活器组件

DQSOps 的核心要素之一是方法激活器，它负责协调流水线流程。如图 2 的决策流程图所示，该组件使用一组预设标准，执行适用于数据质量评分的方法。收集完数据窗口后，系统会根据数据块大小作出决策。激活器会反复运行机器学习模型以获取数据质量分数，直到数据块大小达到预配置的阈值 β。图 3 展示了处理数据窗口和数据块的机制。阈值 β 表示我们希望测试机器学习模型性能的频率。收集到 β 个数据窗口后，达到该阈值，激活器便会使用基于标准的方法执行评估方法。随后，标准方法会生成 n 个数据窗口的真实值分数，其中 $1 \leq n < \beta$。这是对机器学习模型性能进行持续监控所必需的。

评估使用指定的测试预言机 [35],，如图 1 所示。指定的测试预言机需要预言机信息和预言机过程 [4]。预言机信息描述预期输出，该输出通过基于标准的方法获得的真实值分数得出。对于预言机过程，则使用相关的性能指标（例如预测误差），依据预言机信息评估机器学习模型的预测质量分数。随后，测试预言机的结果会传递给激活器，由其将结果与容差水平 τ 进行比较。该水平表示达到目标所需的最低准确性，可以根据质量要求设定，也可以通过观察初始化阶段的机器学习性能推断得出；详见第 3.3 节。如果机器学习模型的性能低于预期水平，则会发送重新训练信号，利用新获得的数据质量分数更新机器学习模型。这一策略确保机器学习模型随着时间推移保持高效和可靠。

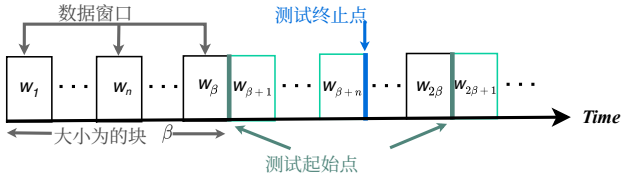


图 3：数据流处理机制

4 实施与评估

该框架应用于一个现实工业应用案例并接受评估。生产环境的结果根据预测性能和时间效率标准进行分析。此外，还分析了变异百分比对初始化阶段的影响。

4.1 工业用例描述

我们实现框架的工业应用案例，是瑞典 Uddeholm钢铁公司的电渣重熔（ESR）真空泵送工业流程[1]。真空泵用于确保生产优质钢。为此，需要从熔炉中抽出氧气。在生产环境中，每次泵送事件最长持续 20 分钟。我们关注的观测量是真空室内产生的压力值。为记录压力值，传感器连接到熔炉，并每毫秒记录一次数值。压力数据窗口通过 Apache Kafka 流处理平台[2] 每秒持续传输一次，以便进行实时分析。在正常的泵送事件中，压力值应逐渐下降，并在 20 分钟内达到所需的最低压力值。然而，有时会发生异常泵送事件，未能达到最低压力值，因此必须重新启动熔炉。因此，区分泵送事件的状态至关重要。本研究的任务是使用所提出的 DQSOps框架，量化每次泵送事件所采集压力数据的质量。针对生产环境，DQSOps 使用 Python 编程语言开发，并通过 Docker容器部署到系统中[3]。无论底层系统如何，Docker生态系统都能实现灵活高效的部署。对于机器学习模型，我们采用两种常用的基于决策树的算法：随机森林（RF） [8] 和极端梯度提升（XGBoost） [14] 模型。RF 和 XGBoost 是两种高效的集成算法，因其简单、高性能且具有可解释性，已广泛应用于各种场景，因此适用于工业应用 [30]。

4.2 预测性能评估

我们框架中的基于预测的方法的性能结果，在我们用例的生产环境中进行评估。这些方法基于回归算法，通过预测数据质量评分，而不是使用

[1]https://www.uddeholm.com/ [2] https://kafka.apache.org/ [3] https://www.docker.com/

表 1：基于随机森林预测的方法的平均性能

DQD	MAE			R^2		
	均值	标准差	CV	均值	标准差	CV
1	0.0148	0.0052	0.3529	0.7133	0.2111	0.2959
2	0.4142	0.0561	0.1353	0.7325	0.2289	0.3125
3	0.5604	0.0928	0.1656	0.7978	0.1353	0.1696
4	0.6411	0.1032	0.1610	0.8435	0.0362	0.0430
5	0.6132	0.0000	0.0000	0.8936	0.0000	0.0000

表 2：基于 XGBoost 预测的方法的平均性能

DQD	MAE			R^2		
	均值	标准差	CV	均值	标准差	CV
1	0.0138	0.0036	0.2617	0.7820	0.1690	0.2161
2	0.4242	0.0619	0.1459	0.7231	0.2341	0.3237
3	0.5483	0.1044	0.1904	0.7846	0.1919	0.2445
4	0.5934	0.1075	0.1812	0.8582	0.0641	0.0746
5	0.6320	0.0000	0.0000	0.8916	0.0000	0.0000

表 3：数据质量分数统计

DQD	数据质量分数范围	数据质量分数标准差
1	[0, 0.9345]	0.0896
2	[-15.173, 4.820]	2.27
3	[-13.819, 5.472]	2.355
4	[-12.204, 8.152]	2.93
5	[-9.899，9.855]	2.846

标准方法计算数据质量评分。性能使用两个性能度量指标进行评估：平均绝对误差（MAE）和决定系数 R^2，其公式如下：

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{6}$$

其中，y_i 是使用基于预言机的方法计算的真实评分，$\hat{y}_i$ 是使用回归算法得到的预测值，而 $\bar{y}$ 表示所有评分的均值。

表 1 展示了基于随机森林预测的方法的性能，表 2 展示了基于 XGBoost 预测的方法的性能。文中记录了所有可能的、规模相同的数据质量维度组合的错误率算术平均数。我们还计算了变异系数（CV），用于衡量每次实验中错误率的相对离散程度。CV 的计算方法是将总体的标准差（std）σ 除以均值 μ，即 $CV = \frac{\sigma}{\mu}$。此外，为了深入了解数据质量评分的分布，并评估错误率相对于数据质量评分统计量的重要性，我们针对不同实验设置下表 3 所示的数据质量评分值计算了汇总统计量。这些统计量概述了数据质量分数范围及其标准差，使我们能够结合具体背景分析机器学习模型的性能。

就预测性能而言，对于两种机器学习算法，随着质量维度数量的增加，平均绝对误差值也会增大，因为算法越来越难以捕获真实的评分函数。在仅使用单个数据质量维度进行的实验中，平均绝对误差度量指标的平均值较低。这是因为大多数数据窗口的质量较高，因此其质量评分接近 0。不过，对于随机森林和 XGBosst，随着数据质量维度数量的增加，该度量指标也会增大，因为算法对数据的拟合效果更好。从表中还可以看出，随机森林与 XGBoost 的性能差异并不显著，后者的平均绝对误差略低，且 R^2 值更高。不过，正如下文分析所示，XGBoost 模型具有显著的运行效率优势，因此是部署在 DQSOps框架中的最优机器学习模型。

4.3 时间加速评估

为评估运行效率，我们比较了执行标准数据质量评分方法和基于预测的数据质量评分方法所需的时间。我们针对不同数量的数据质量维度进行了实验，以分析数据质量维度数量对执行时间的影响。例如，单个数据质量维度总共有五种可能的组合。每种组合分别评估一个数据质量维度。对于两个数据质量维度，可以从框架中定义的五个总体质量维度构造十对数据质量维度，依此类推。表4总结了每种方法处理每个数据窗口的平均计算时间，单位为秒。

与上一节所述的分析类似，我们分析了不同数据质量维度数量下的计算运行时间。我们计算了每项实验计算运行时间的汇总统计量和变异系数（CV），以考察结果的离散程度。可以观察到，基于标准的评分方法的变异系数较高。这是因为计算完整性和一致性评分所需的计算时间少于计算拟合优度或偏度所需的时间。计算完整性和一致性评分只需进行简单的分数运算，而拟合优度和偏度则涉及概率分布估计。因此，在仅涉及单个数据质量维度的实验中，变异性较高。随着数据质量维度数量增加，变异系数逐渐降低，不同实验的运行时间差异也随之减小。然而，与基于标准的方法相比，基于预测的评分方法的标准差较低，变异系数也更为接近。这表明，数据围绕均值的离散程度较低，且基于预测的方法的执行时间相对更为接近。

对于基于标准的方法，平均计算时间与质量维度的数量成正比。相比之下，

表4：使用不同方法为数据窗口评分所需的时间（秒）

DQD	基于标准			基于随机森林预测				基于XGBoost预测			
	mean	std	CV	mean	std	CV	加速比	mean	std	CV	加速比
1	0.19057	0.32218	1.69066	0.01953	0.00142	0.07260	**9.76x**	0.00050	0.00004	0.08774	**381.90x**
2	0.37671	0.39409	1.04614	0.01268	0.00035	0.02740	**29.71x**	0.00103	0.00007	0.06453	**366.45x**
3	0.57191	0.40000	0.69941	0.01259	0.00042	0.03334	**45.44 倍**	0.00109	0.00004	0.03334	**523.72 倍**
4	0.75229	0.32335	0.42983	0.01245	0.00030	0.02405	**60.42 倍**	0.00118	0.00004	0.03472	**638.07 倍**
5	0.94101	0.00000	0.00000	0.01265	0.00000	0.00000	**74.38 倍**	0.00115	0.00000	0.00000	**818.98 倍**

对于基于预测的方法，平均计算时间不受质量维度数量的影响。我们可以看到，在所有实验中，各机器学习模型用于预测数据质量评分时，其汇总统计量都相近。这一发现对生产环境中质量维度数量变化所带来的敏感性具有重要的实际意义。基于预测的评分过程的运行时间不受质量维度数量的影响。相比之下，标准评分过程的运行时间会随着质量维度数量的增加而延长。关于加速比，表4中的结果表明，基于预测的数据质量评分方法具有显著的运行效率。具体而言，随机森林（RF）基于预测的方法在单个质量维度下，相较于基于标准的方法实现了10倍的加速比；使用全部数据质量维度时，加速比约为75倍。同样，基于XGBoost的预测方法在使用一个质量维度时，加速比约为382倍，而使用全部质量维度时则约为819倍。然而，比较不同的机器学习模型后可以发现，XGBoost算法所需的运行时间明显少于随机森林（RF）。这是因为在超参数调优阶段，随机森林（RF）依赖于比XGBoost更复杂的决策树集成来进行预测。

4.4 变异百分比影响

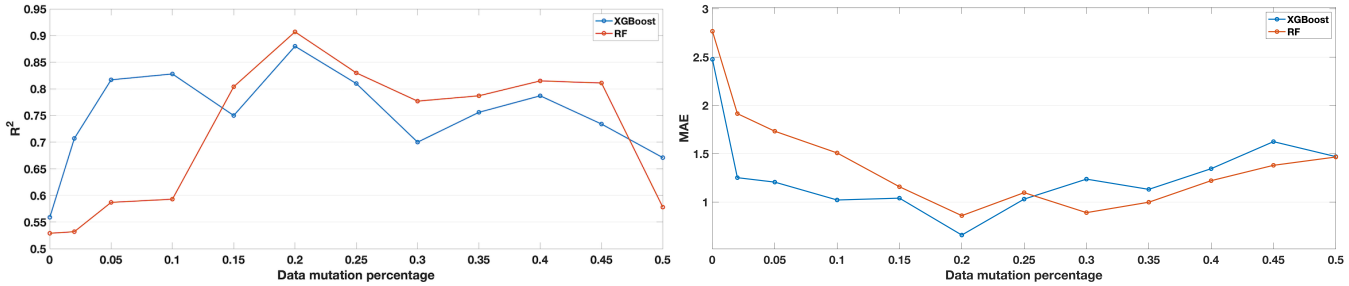
如第3节所述，我们按照指定的变异百分比实施了数据变异方法，以模拟影响处理后的数据窗口质量的问题。具体而言，本实验使用了全部五个数据质量维度。为分析变异百分比在初始化阶段的影响，我们通过使用验证集调整百分比值，观察了测试预言机的结果。针对每种机器学习算法，我们使用 R^2 和平均绝对误差性能指标对测试预言机进行了评估。随机森林和XGBoost算法的结果汇总于图4。

对于这两种算法，我们可以看到评估指标呈现 U 形趋势。具体而言，如图4a所示，对于 R^2，在未模拟数据变异体时，两种算法的数值均略高于 0.5；随后随着数据变异百分比达到 20%，数值提升至约 0.9。达到峰值后，随着数据变异体的百分比增加，性能指标开始下降；当变异百分比达到 50% 时，XGBoost 和随机森林的数值分别约为 0.68 和 0.58。同样，如图4b所示，平均绝对误差度量指标的曲线也呈现相同趋势。训练数据中引入噪声变异体时，误差水平起初最高；当变异体百分比达到 20% 时，模型性能达到最佳，此时 XGBoost 的平均绝对误差为 0.65，随机森林为 0.86。此后，引入更多数据变异体并不能改善训练过程，且随着变异百分比升高，误差持续增加。

本实验结果表明，应谨慎选择诱导突变体的百分比，以获得能实现最佳性能的取值。具体而言，较低的变异百分比可能无法引入足够的数据质量问题，因而无法充分训练机器学习模型，使其正确评估数据质量。然而，若将变异百分比设得过高，则可能在数据中引入过多质量问题，扰乱学习过程并导致性能不佳。因此，在初始化阶段，谨慎选择变异百分比，在引入足够数据质量问题与避免引入过多可能对机器学习性能产生负面影响的错误之间取得平衡，至关重要。

5 外部效度威胁

外部效度关注的是如何将所提出框架的研究发现推广应用。对于实际应用的框架设计而言，可推广性是主要挑战，因为每个用例可能都需要采用不同的方法来解决问题。然而，图 1 所示的 DQSOps框架提出了一种抽象的流水线工作流，其设计具有灵活性，组件没有严格的规范，以促进可复现性。因此，只需对本研究使用的方法进行简单调整，该框架便可用于在各种场景下为不同性质的数据评分。例如，第3.1节介绍的数据质量维度评分旨在处理单变量时间序列数据，这是我们的用例所要求的。然而，也可以采用其他方法对多变量数据 [40, 52] 的质量进行评分，并将其集成到 DQSOps框架中。同样，其他组件也可以根据所处理用例的系统需求进行确定。此外，配置文件的使用增强了框架的灵活性。配置文件存储辅助元信息，例如数据质量维度的参数和变异百分比。这些元信息可根据用例需求进行配置，而无需修改 DQSOps 的底层结构。根据我们的分析，我们认为，即使采用不同于本研究所用的技术，仍会得出与本研究相似的结果和发现。



(a) R^2 (b) 平均绝对误差

图4：变异体百分比对性能的影响

6 结论

本文介绍了数据质量评分运维（DQSOps）框架。该框架能够量化数据记录的质量，并区分高质量数据和低质量数据。该框架集成了两种评分方法：基于预测的方法和基于标准的方法。基于预测的方法使用机器学习模型预测所收集数据窗口的质量评分。相比之下，基于标准的方法会定期调用，以生成真实质量分数。该分数通过组合定义的数据质量维度中的多项评分指标得出，并用于设计测试预言机。测试预言机持续评估机器学习模型，并触发重新训练信号以更新机器学习模型。此外，框架还集成了数据变异体模拟器，以在数据中引入质量问题并促进学习过程。该框架已在一个现实工业应用案例中部署并评估。结果表明，与标准评分方法相比，DQSOps 实现了显著的加速比，同时保持了较高的预测性能。本文还分析了最优变异百分比，以深入了解其对学习过程的影响。

致谢

本研究由瑞典知识基金会（KKS）通过协同项目 AIDA——面向工业物联网的整体式人工智能驱动网络与处理框架（Rek:20200067）资助。

参考文献

[1] Hervé Abdi 和 Lynne J Williams。2010。主成分分析。*Wiley*跨学科评论：计算统计学2，4（2010），433–459。[2] Muhammad Aslam。2021。不确定参数存在时的一种新拟合优度检验。复杂与智能系统7，1（2021），359–365。[3] Claudia Augste 和 Martin Lames。2011。相对年龄效应与德国精英U-17足球队的成功。运动科学杂志29，9（2011），983–987。[4] Earl T Barr、Mark Harman、Phil McMinn、Muzammil Shahbaz 和 Shin Yoo。2014。软件测试中的预言机问题：综述。*IEEE*软件工程汇刊41，5（2014），507–525。[5] Carlo Batini、Anisa Rula、Monica Scannapieco 和 Gianluigi Viscusi。2015。从数据质量到大数据质量。数据库管理杂志（*JDM*）26，1（2015），60–82。[6] Roger Blake 和 Paul Mangiameli。2011。数据质量与问题复杂度对分类的影响及其交互作用。数据与信息质量期刊（*JDIQ*）2，2（2011），1–28。[7] Jan Bosch、Helena Holmström Olsson 和 Ivica Crnkovic。2021。人工智能系统工程：研究议程。面向智能网络物理系统的人工智能范式（2021），1–19。[8] Leo Breiman。2001。随机森林。《机器学习》45卷第1期（2001），5–32。[9] John Byabazaire、Gregory O'Hare 和 Declan T Delaney。2022。使用信任对共享数据的物联网部署进行端到端数据质量评估。《IEEE传感器期刊》（2022）。[10] Cinzia Cappiello、C Cerletti、C Fratto 和 Barbara Pernici。2018。验证评分流程中的数据质量操作。《数据与信息质量期刊》（JDIQ）9卷第2期（2018），1–27。[11] Emily Caveness、Paul Suganthan GC、Zhuo Peng、Neoklis Polyzotis、Sudip Roy 和 Martin Zinkevich。2020。Tensorflow数据验证：持续机器学习流水线中的数据分析与验证。载于《2020年ACM SIGMOD数据管理国际会议论文集》，2793–2796。[12] Raghavendra Chalapathy 和 Sanjay Chawla。2019。用于异常检测的深度学习：综述。arXiv预印本 arXiv:1901.03407（2019）。[13] Abraham Chan、Arpan Gujarati、Karthik Pattabiraman 和 Sathish Gopalakrishnan。2022。数据星辰中的故障：研究机器学习应用中针对有缺陷训练数据的缓解技术。载于《2022年第52届IEEE/IFIP可靠系统与网络国际年会（DSN）》，IEEE，163–171。[14] Tianqi Chen 和 Carlos Guestrin。2016。XGBoost：一种可扩展的树提升系统。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，785–794。[15] Sezal Chug、Priya Kaushal、Ponnurangam Kumaraguru 和 Tavpritesh Sethi。2021。通过统计学习将领域无关的数据质量评分付诸实践。arXiv预印本 arXiv:2108.08905（2021）。[16] Corinna Cichy 和 Stefan Rass。2019。数据质量框架概述。《IEEE Access》7（2019），24634–24648。[17] RalphB D'Agostino。2017。拟合优度检验技术。Routledge。[18] Koen Decancq 和 Maria Ana Lugo。2012。福祉不平等：一种多维方法。《Economica》79卷第316期（2012），721–746。[19] Adenekan Dedeke。2000。为信息系统制定质量度量指标的概念框架。载于《IQ》，126–128。[20] Lisa Ehrlinger 和 Wolfram Wöß。2018。一种用于最小性的新型数据质量度量指标。载于《大数据质量与信任国际研讨会》，Springer，1–15。[21] Diane L Evans、John H Drew 和 Lawrence M Leemis。2008。估计参数的指数总体中 Kolmogorov–Smirnov、Cramer–von Mises 和安德森–达林检验统计量的分布。《统计通讯：模拟与计算》37卷第7期（2008），1396–1421。[22] Wenfei Fan。2015。数据质量：从理论到实践。《ACM SIGMOD记录》44卷第3期（2015），7–18。[23] Harald Foidl 和 Michael Felderer。2019。基于风险的数据验证在基于机器学习的软件系统中的应用。载于《第3届ACM SIGSOFT软件质量评估机器学习技术国际研讨会论文集》，13–18。[24] Bernd Heinrich、Diana Hristova、Mathias Klier、Alexander Schiller 和 Michael Szubartowicz。2018。数据质量度量指标的要求。《数据与信息质量期刊》（JDIQ）9卷第2期（2018），1–32。[25] Bernd Heinrich、Mathias Klier、Alexander Schiller 和 Gerit Wagner。2018。评估数据质量——一种用于语义一致性的基于概率的度量指标。《决策支持系统》110（2018），95–106。[26] Abhinav Jain、Hima Patel、Lokesh Nagalapatti、Nitin Gupta、Sameep Mehta、Shanmukha Guttula、Shashank Mujumdar、Shazia Afzal、Ruhi Sharma Mittal 和 Vitobha Munigala。2020。数据质量对机器学习任务的概述及重要性。载于《第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，3561–3562。[27] Yue Jia 和 Mark Harman。2010。变异测试发展的分析与综述。《IEEE软件工程汇刊》37卷第5期（2010），649–678。[28] Meenu Mary John、Helena Holmström Olsson 和 Jan Bosch。2021。迈向机器学习运维：一种框架与成熟度模型。载于《2021年第47届Euromicro会议

软件工程与高级应用（*SEAA*）。IEEE，1–8。[29] Steven G Johnson、Stuart Speedie、Gyorgy Simon、Vipin Kumar 和 Bonnie L Westra。2016。将本体用于刻画电子健康记录数据二次使用的数据质量。《应用临床信息学》7卷第01期（2016），69–88。[30] Sonia Kahiomba Kiangala 和 Zenghui Wang。2021。在工业4.0环境下，利用极端梯度提升（XGBoost）和随机森林集成学习算法，为小型制造工厂构建有效的自适应定制框架。《机器学习与应用》4（2021），100024。[31] Shirlee-ann Knight。2011。信息质量的综合概念生命周期模型：第1部分，调查框架。《国际信息质量期刊》2卷第3期（2011），205–230。[32] Paul Kvam、Brani Vidakovic 和 Seong-joon Kim。2022。《非参数统计及其在科学与工程中的*R*语言应用》。约翰·威立父子出版公司。[33] Erin LeDell 和 Sebastien Poirier。2020。H2o automl：可扩展的自动机器学习。载于《*ICML AutoML*研讨会论文集》，第2020卷。[34] Lillian Lee。2000。分布相似性的度量。《*arXiv*预印本 *cs/0001012*》（2000）。[35] Nan Li 和 Jeff Offutt。2016。基于模型的测试中的测试预言机策略。《*IEEE*软件工程汇刊》43卷第4期（2016），372–395。[36] Jianhua Lin。1991。基于香农熵的散度度量。《*IEEE*信息论汇刊》37卷第1期（1991），145–151。[37] Antonios Lionis、Konstantinos P Peppas、Hector E Nistazakis 和 Andreas Tsigopoulos。2021。使用 Jensen-Shannon 散度和皮尔逊分布比较 RSSI概率密度函数。《技术》9卷第2期（2021），26。[38] David Loshin。2010。《数据质量改进实践指南》。Elsevier。[39] Jie Lu、Anjin Liu、Fan Dong、Feng Gu、Joao Gama 和 Guangquan Zhang。2018。概念漂移下的学习：综述。《*IEEE*知识与数据工程汇刊》31卷第12期（2018），2346–2363。[40] Harald Martens 和 Magni Martens。2001。《质量多变量分析：导论》。约翰·威立父子出版公司。[41] Xiaofeng Meng 和 Xiang Ci。2013。大数据管理：概念、技术与挑战。《计算机研究与发展》50卷第1期（2013），146–169。[42] Helen-Tadesse Moges、Karel Dejaeger、Wilfried Lemahieu 和 Bart Baesens。2013。信用风险管理中数据质量的多维分析：新见解与挑战。《《信息与管理》》50卷第1期（2013），43–58。[43] Michael Muller、Ingrid Lange、Dakuo Wang、David Piorkowski、Jason Tsay、Q Vera Liao、Casey Dugan 和 Thomas Erickson。2019。数据科学工作者如何工作数据：发现、捕获、整理、设计、创建。载于*2019*年计算机系统中的人为因素*CHI*会议论文集。1–15。[44] Hoang-Vu Nguyen 和 Jilles Vreeken。2015。非参数詹森-香农散度。载于欧洲机器学习与数据库知识发现联合会议。Springer，173–189。[45] Jack E Olson。2003。数据质量：准确性维度。Elsevier。[46] Liu Peng 和 Lei Lei。2005。缺失数据处理方法综述。*Intell. Inf. Manag. Syst. Technol* 1（2005），412–419。[47] Leo L Pipino、Yang W Lee 和 Richard Y Wang。2002。数据质量评估。*Commun. ACM* 45，4（2002），211–218。[48] John Winsor Pratt 和 Jean Dickinson Gibbons。2012。非参数理论概念。Springer Science & Business Media。[49] Laura Rettig、Mourad Khayati、Philippe Cudré-Mauroux 和 Michał Piórkowski。2019。大数据流在线异常检测。载于应用数据科学。Springer，289–312。[50] David Schuler 和 Andreas Zeller。2013。覆盖与揭示等价变异体。软件测试、验证与可靠性 23，5（2013），353–374。[51] Kelly M Sunderland、Derek Beaton、Julia Fraser、Donna Kwan、Paula M McLaughlin、Manuel Montero-Odasso、Alicia J Peltsch、Frederico Pieruccini-Faria、Demetrios J Sahlas、Richard H Swartz 等。2019。多变量离群值检测技术在大型研究数据质量评估中的效用：在ONDRI项目中的应用。*BMC*医学研究方法学 19，1（2019），1–16。[52] Ikbal Taleb、Mohamed Adel Serhani、Chafik Bouhaddioui 和 Rachida Dssouli。2021。大数据质量框架：持续质量管理的整体方法。大数据期刊 8，1（2021），1–41。[53] Hui Yie Teh、Andreas W Kempa-Liehr 和 Kevin I-Kai Wang。2020。传感器数据质量：系统综述。大数据期刊 7，1（2020），1–49。[54] Reza Vaziri、Mehran Mohsenzadeh 和 Jafar Habibi。2019。使用加权指标衡量数据质量。全面质量管理与卓越经营 30，5-6（2019），708–720。[55] Yair Wand 和 Richard Y Wang。1996。以本体论基础为数据质量维度奠定基础。*Commun. ACM* 39，11（1996），86–95。[56] Richard Y Wang、Veda C Storey 和 Christopher P Firth。1995。数据质量研究分析框架。*IEEE*知识与数据工程汇刊 7，4（1995），623–640。