通过去相关化和随机森林回归检测趋势来衡量特征重要性

Yannick Gerstorfer、Lena Krieg、Max Hahn-Klimroth

Abstract. 在许多研究中，我们希望确定某些特征对因变量的影响。更具体地说，我们关注影响的强度——也就是说，该特征是否相关？如果相关，该特征又如何影响因变量？近年来，数据驱动方法（如 *random forest regression* ）已应用于实际场景（Boulesteix 等人，2012年）。这些模型可以直接得出特征重要性的度量值，而这自然能够反映影响的强度。对于相关特征，通常使用特征与因变量之间的相关性或秩相关来确定影响的性质。较新的方法中，有些也能衡量特征间交互作用，它们基于建模方法。尤其是在使用机器学习模型时，SHAP 值是近年来用于确定这些趋势的一种重要且广受关注的方法（Lundberg 等人，2017年）。

本文介绍了一种基于广为人知的格拉姆-施密特去相关方法的新型特征重要性概念。此外，我们提出了两种利用随机森林回归识别数据中趋势的估计量，分别称为绝对横向率和相对横向率。我们通过多种合成数据集和真实世界数据集，对比了这些估计量与成熟估计量的特性。

arXiv:2303.01156v1 [stat.ML] 2 Mar 2023

1. Introduction

在许多研究中，科研人员需要面对高维但有限的数据，以确定特定特征对因变量的影响。通常，数据同时包含数值特征和类别特征，并且会出现强烈的人为多变量相关性。尤其是，当数据来自对活体动物的观察或在医疗程序中收集时，数据很可能是不平衡的；更糟的是，并非所有特征组合都包含样本。因此，经典统计检验所需的全部假设很可能无法满足。机器学习方法因只需最少的假设便能产生稳健的效应估计，而日益受到研究人员青睐。一个简单而典型的例子是随机森林回归。随着数据科学理念不断发展，以及各研究团队可用计算能力不断增强，这类数据驱动方法正逐渐应用于生命科学研究 [3]。与所有机器学习模型一样，随机森林回归对底层数据的分布只作少量假设，并且对噪声和离群值尤其稳健。最后，它还可以直接得出特征重要性的度量，这些度量是衡量各个特征影响强度的自然指标 [2, 9, 15]。在可以使用方差分析等经典统计工具的情况下，众所周知，方差分析判定为显著的大多数特征也具有较高的特征重要性，反之亦然 [7, 16]。

找到相关特征后，确定特征取值如何影响因变量至关重要。最早的方法可能是测量特征与因变量之间的相关性或秩相关。较新的方法中，有些还可以测量特征间交互作用，它们基于建模方法。训练一个模型（例如多元线性回归模型），并利用其参数确定趋势。尤其是在使用机器学习模型时，SHAP 值 [17] 是用于确定这些趋势的一种较新且颇具影响力的方法。这些方法利用模型而非原始数据。这有助于识别数据中无法直接观察到、却隐藏在噪声背后的趋势。另一方面，要使这些趋势可靠，就需要一个质量过关的模型。

所有作者均获得德国研究基金会资助，项目编号为 DFG FOR 2975。1

本文有两个目标。首先，鉴于特征间的依赖关系已知会影响特征重要性分数，我们基于久经研究的格拉姆-施密特去相关方法，提出一种特征重要性的概念。我们通过实证将这一概念与一种基于残差学习的类似方法，以及经典的基于不纯度的特征重要性和置换重要性进行比较。其次，我们提出两种估计量，利用随机森林回归识别数据中的趋势。我们利用随机森林的结构，即在每个分裂节点处比较左右子树中的平均预测值。由于左子树基于低于某一阈值的数据构建，而右子树包含高于该阈值的数据，因此可以自然地估计特征与预测变量之间某种形式的相关性。

2. Background and Notation

2.1. 特征重要性。 关于随机森林，文献中广为人知的特征重要性分数有两种类型。第一种是基于不纯度的 特征重要性。所谓不纯度，是通过所包含决策树集合的分裂准则进行量化的。因此，如果数据集未经标准化，它很可能会高估大型数值特征的重要性。此外，在过拟合的情况下，一些在未见数据上可能不具备预测能力的特征也可能被发现很重要。基于这些原因，第二种特征重要性，即所谓的置换重要性，已在文献中得到应用，并且更值得优先采用 [4]。它被定义为随机打乱单个特征时模型性能的下降幅度。当然，这种基于置换的方法也存在缺点——尤其是在存在（高度）相关的 [4]特征聚类时。克服这一问题的一种方法，也是特征提取过程中常用的方法，是每个聚类仅保留一个变量 [6, 10, 13]。如果最终目标是设计一个仅使用尽可能少特征的优良预测模型，这就是当前的最佳实践。但在某些情况下，研究人员实际上更关注于估计每个特征的重要性，以确定哪些特征会影响因变量以及影响程度。在这种情况下，以不同方式处理相关性可能更为方便。通常用于数据聚类或设计性能良好的预测模型的去相关技术至少有两种：格拉姆-施密特去相关技术 [19]或基于残差的去相关 [8]。两种方法的主要思想都是从给定特征 F_i 中减去由 F_1、.....、F_{i-1}、F_{i+1}、$\ldots F_d$ 提供的信息，并使用该残差 训练模型。

2.2. 趋势。 我们比较了三种不同的方式来定义数据集中的趋势 。考察特征 Y 与预测变量之间关系时，最简单的做法或许是使用相关系数 $r(X,\ Y) := \mathrm{Cov}(X,\ Y)/(\sigma(X)\sigma(Y))$，它反映的是线性趋势。能够处理任意单调 趋势的更一般的相关系数，是各种类型的秩相关系数，例如斯皮尔曼相关系数$\rho(X,\ Y) = r(R(X)\ R(Y))$，其中 $R(\cdot)$ 表示秩函数。这种寻找趋势的方法已广为人知，且只考虑可观测的原始数据。

另一种方法不考察原始数据，而是拟合一个模型并在该模型中寻找趋势。许多从业者倾向于在多变量任务中通过对数据拟合线性模型，并将特征的系数符号及其对应的 p–值解读为趋势，来识别趋势。我们将这个系数记为 $r_{LM}(F)$。然而，我们将看到，即使对于非常简单的数据集，这种做法也可能产生严重误导。

近年来，数学博弈论中的一个古老概念——称为*Shapley* 值——被用于解释机器学习模型 [14]。具体而言，对于基于树的模型和随机森林，人们已经从数学上充分理解了这些值。相对于某个数据点，某个特征的 Shapley 值衡量该特征值相较于平均预测对预测结果的贡献大小，其定义为：在所有可能的特征组合中，该特征值边际贡献的平均值。有关形式化定义，请参阅 Shapley 的原始论文 [17], ；有关如何在机器学习中使用这一概念的详细讨论，请参阅 [12, 18]。显然，这些 Shapley 得分可用于确定趋势。

2.3. **研究数据集**。为在实践中检验我们的估计量的性能，我们使用了两个非常知名的真实数据集，分别称为*Kaggle* 鱼类市场数据集（FISH） [1]和加利福尼亚住房数据（HOUSING） [5]。此外，我们还创建了三个不同的合成数据集，以探索这些估计量的某些方面。

FISH包含鱼类市场销售中七种常见鱼类物种的记录。每条鱼的特征包括物种、重量、垂直长度、对角线长度、横向长度、高度和宽度。在这些特征中，我们使用重量、高度和宽度来预测垂直长度。加利福尼亚住房数据指加利福尼亚州某个县的房屋数据，以及基于1990年人口普查数据得出的汇总统计数据。每个县的特征包括经度、纬度、房屋中位年龄、房间总数、卧室总数、人口、家庭户数量、收入中位数和距海远近，预测目标变量为房屋价值中位数。我们将距海远近这一特征转换为有序尺度。

第一个合成数据集（SYN1）由一个基础数据集 B 衍生而来，该数据集包含1000个样本和10个特征，其中3个为信息特征。基础数据集通过去除均值并缩放至单位方差进行标准化。随后，将其与一个N 采用相同方式标准化、结构相同但不含信息特征的噪声数据集相结合。由此得到一系列数据集：

$$D_w = (1-w)B + wN \quad (w \in [0.01, 0.02, \ldots, 1.]).$$

SYN1 由基础数据集与 250 个不同的随机噪声数据集组合而成。SYN1 用于比较趋势估计器的稳健性。

第二个合成数据集（SYN2）包含 100 个样本，其特征是独立生成的

$$X_0 \sim 3 \cdot \mathcal{N}(0,1), \qquad X_1 \sim 2 \cdot \mathcal{N}(0,1), \qquad X_2 \sim \mathcal{N}(0,1).$$

此外，给定 X_0，我们定义

$$A_0 = X_0 + \mathcal{N}(0,1), \qquad A_1 = X_0 + 10 \cdot \mathcal{N}(0,1), \qquad A_2 = X_0^2 + 10 \cdot \mathcal{N}(0,1).$$

真实标签定义为

$$Y = 4X_0^{1.5} + 2X_1 + 0.5X_2^2.$$

因此，真实标签取决于 X_0， X_1， X_2和 A_0， A_1， A_2可视为具有不同类型依赖关系的 X_0 的含噪实例。SYN2 用于比较不同的特征重要性概念。

第三个合成数据集（SYN3）包含 100 个样本，且只有一个信息特征 X_0，其定义同前。此外， A_0， A_1， A_2 的定义同上。现在，真实标签定义为 $Y = 4X_0^{1.5}$. SYN3 用于比较相关特征簇 上的特征重要性概念，并与 SYN2 直接比较；SYN2 中存在另外两个不相关的信息特征。

3. Contribution

3.1. **在数据集中发现趋势**。 我们比较常用的回归系数r， ρ、基于线性模型的趋势估计量、基于 Shapley 值的趋势估计量，并提出两种基于随机森林回归的新型估计量，以确定特征的趋势。为此，我们将某个特征的基于 Shapley 值的趋势简单定义为其取值 X 与其 Shapley 值 $s(X)$ 之间的相关性，从而分别得到估计量 $r(X,\ s(X))$ 和 $\rho(X,\ s(X))$。

提出的两个趋势估计器分别是绝对遍历率和相对遍历率。随机森林回归模型使用由互不相关的决策树组成的集成。在每个节点，根据该节点特征的取值，将当前数据集划分为两个分区类。不失一般性，我们假设左 分区类中的数据对应较小的特征值，而右 分区类中的数据对应较大的特征值。为了确定特征 F 的趋势，比较每个节点左右分区类中特征的均值似乎是合理的。如果平均值

如果左子树中预测变量的平均值小于右子树中的平均值，则这对应于与特征 F 的正相关性。更正式地说，令 $\{F_j\}_{j=1\ldots n}$ 表示随机森林中依据特征 F 对数据进行划分的节点集合。相应的分区类称为 $L(F_n)$ 和 $R(F_n)$。如果特征 F 可从上下文中明确，则将这些类简称为 L_n 和 R_n。此外，对于预测变量取值的子集 A，我们将 $AVG(A) = |A|^{-1}\sum a$ 定义为$a\in A$的平均值

集合 A。这使我们能够定义趋势估计量。

定义 1. 给定随机森林 $\mathcal{R}$，令 $\{F_j\}_{j=1\ldots n}$ 表示随机森林中依据特征 F对数据进行分割的节点集合。特征 F 的绝对横向率定义为

$$ATR(\mathcal{R}, F) = n^{-1}\sum_{i=1}^{n}\left(\mathbf{1}\{AVG(L(F_i)) < AVG(R(F_i))\} - \mathbf{1}\{AVG(L(F_i)) > AVG(R(F_i))\}\right).$$

此外，特征 F 的相对横向率定义为

$$RTR(\mathcal{R}, F) = 2\sum_{i=1}^{n}\frac{R(F_i) - L(F_i)}{|L(F_i) + R(F_i)|}.$$

ATR 将这样一种观点形式化：当特征值较高时，模型也会返回较高的值，此时我们认为存在趋势。RTR 还考虑了分区类中平均值之间的相对差异。

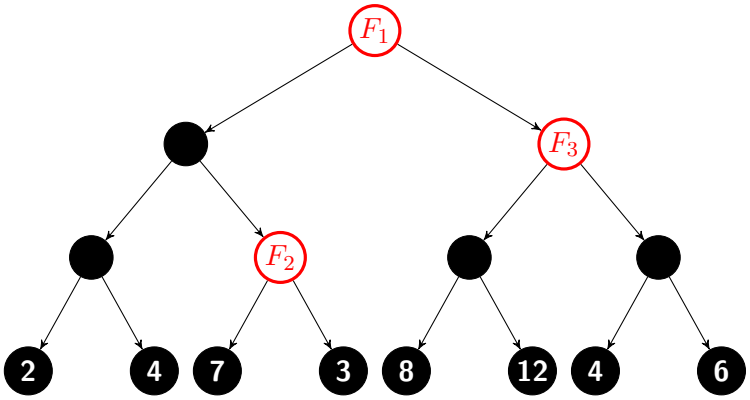


FIGURE 1. 每次出现特征 F 时，都会将数据集分成两部分。在此示例中，F_1 生成分区类 $L_1 = \{2, 4, 7, 3\}$ 和 $R_1 =\{8, 12, 4, 6\}$。在 F_2 处进行的分割生成类 $L_2 = \{7\}$ 和 $R_2 = \{3\}$，而在 F_3处进行的分割则定义了 $L_3 = \{8, 12\}$ 和 $R_3 = \{4, 6\}$。

趋势估计器。为对数据集进行实证分析，我们编写了一个趋势估计模块，该模块在输入数据上训练随机森林回归模型和线性模型。随后，针对每个特征，趋势估计器输出

（1）线性模型的系数；（2）目标变量与以下各项之间的斯皮尔曼等级相关和皮尔逊相关系数：（a）沙普利值，（b）特征；（3）绝对遍历率（ATR）；（4）相对遍历率（RTR）。

3.2. 重要性度量。 我们比较四种不同的特征重要性概念。基于不纯度的特征重要性和基于置换的特征重要性这两种定义，都是经过充分研究的对象 [4]。我们使用*scikit-learn* 的 默认实现来计算这些度量。

此外，我们还介绍两种基于残差学习的新型特征重要性。其思路是，根据特征$F_1,\ldots,F_{i-1},F_{i+1},\ldots F_d$，通过特征 F_i 的残差来确定该特征的重要性。稍微变通一下符号记法，我们将 $F_i \in \mathbb{R}^n$ 视为与特征 F_i 对应的所有取值构成的向量，并用 $Y \in \mathbb{R}^n$ 表示因变量的取值。我们用 $\mathcal{A}_j$ 表示一个任意算法，它以 $F_1,\ldots,F_{j-1}$ 为输入，并输出一个属于 $\mathbb{R}^n$ 的向量。给定 $[d]$ 的一个固定置换 π，我们用 $i_1,\ldots,i_d$ 表示在 π 下的新顺序。为确定 F_i 的重要性，我们计算它在所有满足 $i_d = i$ 条件的置换 π 下的重要性，并根据仅由特征 F_i 构成的模型的性能对其加权。其含义如下：给定其他所有特征，我们能从特征 F_{i_d} 中学到什么？计算该重要性的算法可表示如下。

- 对于所有将 $i \mapsto d$映射的置换 π，执行以下操作– 定义 $W_1^\pi = F_{i_1}$。–将特征 F_{i_j}的取值替换为$W_{i_j}^\pi = F_{i_j} - \mathcal{A}_{\rangle|}(W_{i_1}^\pi, \ldots, W_{i_{j-1}}^\pi)$（对于$j = 2\ldots d$）。– 使用特征 $\{W_{i_j}\}$ 训练一个随机森林。– 确定 $W_{i_d}^\pi$的基于不纯度的特征重要性。

- 将 F_i 的平均特征重要性确定为所有 $W_{i_d}^\pi$上的均值，记为（FI）$_i$。

- 使用特征 F_i 和因变量 Y 训练一个随机森林回归器 $\mathcal{R}_i$，并测量 r（$\mathcal{R}$（F_i） Y）。~

- 返回 $f_i = r$（$\mathcal{R}_i$（F_i） Y）FI_i

~ ~应用该算法后，我们得到 f_1，.. . ， f_d。最后，我们将基于残差算法的特征重要性 $\mathcal{A}$ 定义为上述估计量的标准化版本，即

$$f_j(\mathcal{A}) = \frac{\tilde{f}_j}{\sum_{i=1}^d \tilde{f}_i}.$$

具体而言，该算法如算法 1 所示。我们注意到以下几点。

- $f_j(\mathcal{A})$ 是一个随机量，因为它取决于随机森林回归器 $\mathcal{R}_1$、.....、 $\mathcal{R}_d$ 以及使用特征 $\{W_{i_j}\}$ 的随机森林回归器的训练。
- 在实际应用中，可能无法遍历所有置换 π。因此，通过抽取一些置换来估计平均基于不纯度的特征重要性。
- 该算法高度依赖于残差算法 $\mathcal{A}$。

算法 1 基于残差的特征重要性

要求： d 特征 F_1、.. .、 F_d、残差算法 $\mathcal{A}$、因变量 $YS_d \leftarrow$ 置换集合 $\{1,2,\ldots,d\}$FeatImp $\leftarrow$ (0, ..., 0) $\in \mathbb{R}^d$对于 $\pi \in S_d$ 执行$F_{i_j} \in \mathbb{R}^n \leftarrow j$–置换下的第 $\pi W_{i_1}^\pi = F_{i_1}$个特征向量 $j = 2\ldots d$ 执行$W_{i_j}^\pi \leftarrow F_{i_j} - \mathcal{A}_{i_j}(W_{i_1}^\pi$、 . . .、 $W_{i_{j-1}}^\pi)$ 结束执行 RF $\leftarrow$使用特征 $W_{i_1}^\pi$、 . . .、 $W_{i_d}^\pi$和因变量 Y生成随机森林模型 FI $\leftarrow$ RF 针对特征 $W_{i_d}^\pi$的基于不纯度的特征重要性结果$\mathcal{R}_{\mathrm{i_d}} \leftarrow$ 使用特征 F_{i_d} 和因变量 Y生成随机森林模型 FeatImp$[k] \leftarrow$ FeatImp$[k] + r(\mathcal{R}_{i_d}(F_{i_d})\ Y)$ · FI 结束执行 FeatImp $\leftarrow$ FeatImp $_{||}$FeatImp$_{||_1}$返回 FeatImp

本文通过实证分析，研究了基于两种不同残差算法的特征重要性：通过随机森林回归进行经典残差学习，以及通过格拉姆-施密特方法进行去相关。

基于残差学习的特征重要性。 接下来，[8], 自然可以将残差算法族定义为随机森林回归器族 $\mathcal{A}_2$，……，$\mathcal{A}_d$。更准确地说，给定 $W^\pi_{i_1}$，……，$W^\pi_{i_{j-1}}$，我们使用这些特征和因变量 F_{i_j} 训练一个随机森林回归器 $\mathcal{R}$。因此，

$$\mathcal{A}_j(W^\pi_1,\ldots,W^\pi_{j-1}) = \mathcal{R}(W^\pi_1,\ldots,W^\pi_{j-1}).$$

因此，我们从 F_{i_j} 中减去在 π条件下，随机森林回归器能够从前 $j-1$ 个特征中学到的所有内容。这种方法通常称为残差学习，在机器学习 [11]中有着广泛的应用。

基于格拉姆-施密特去相关的特征重要性。 另一种自然的方法是采用著名的格拉姆-施密特正交化技术。一个多世纪以来，它一直被用于数学领域，以生成向量空间的正交基；直到21世纪初，它才首次被用于在复杂数据集中寻找独立成分 [19]。最重要的观察是，协方差是一种内积，因此，可以将通用的格拉姆-施密特正交化技术应用于协方差，从而生成去相关特征。这里，我们定义

$$\mathcal{A}_j(W^\pi_1,\ldots,W^\pi_{j-1}) = \sum_{i=1}^{j-1} \frac{\mathrm{Cov}(F_j, W^\pi_i)}{\mathrm{Cov}(F_j, F_j)} F_j.$$

一个主要优点可能是，与上述方法不同，这种正交化方法在数学上完全可以理解。不过，它可能容易受到非线性依赖关系的影响。

4. Results

4.1. 发现趋势。 下面，我们报告在 HOUSING、FISH 和 SYN1 数据集上，不同趋势估计量的性能实证结果。

4.1.1. *SYN1*。 该数据集用于测试不同趋势估计量在数据集混入噪声时的稳健性。为此，对 250 个随机噪声数据集中的每一个，趋势估计模块都应用于 D_w，针对每个 $w \in [0.01,\ldots,1.]$。在我们的实验中，汇总结果显示，对于信息特征，ATR 和 RTR 以及 Shapley相关性都比线性模型更加稳健（图 2）。沙普利值最为稳健，其次是 RTR 和 ATR。有趣的是，非信息特征也被赋予了较大的 ρ_S r_S，RTR 和 ATR 值。

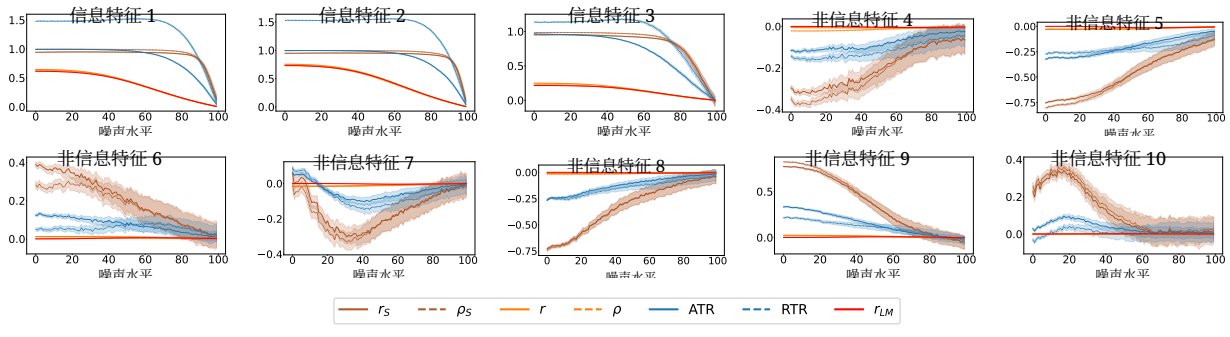


Figure 2. SYN1 上不同趋势估计量在 250 次独立试验中的均值和 95% 置信区间。 x–轴上报告的是噪声比例。特征 1–3 为信息特征，而特征 4–10 为非信息特征。

4.1.2. *FISH*。 我们使用特征重量、高度和宽度对 FISH 数据集进行了三项实验，以预测长度。所选的三个特征均与目标变量呈正相关（图 3）。

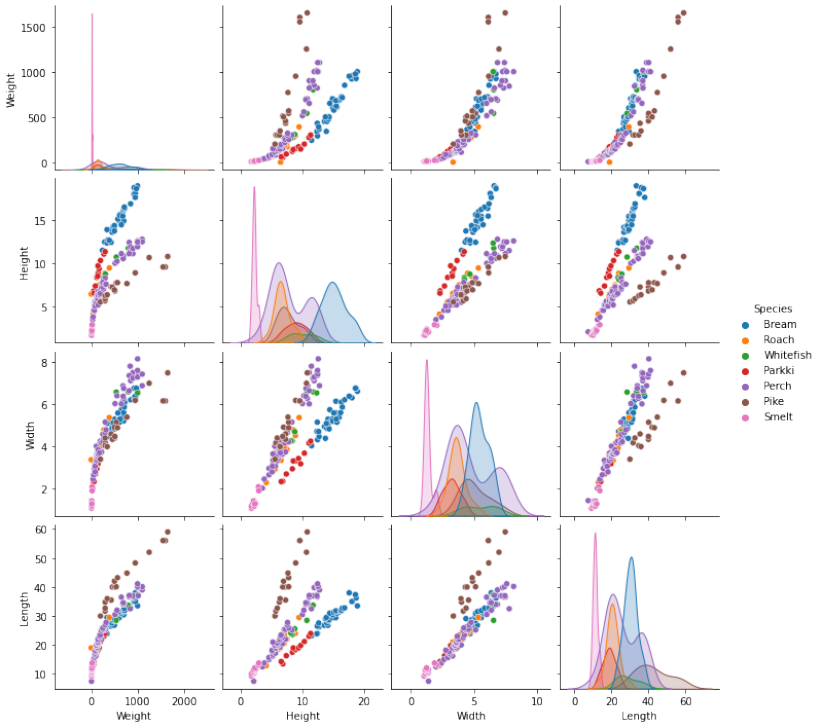


FIGURE 3. 所用鱼市数据集中的特征（重量、高度和宽度）与预测变量（长度）的成对关系图。

首先，我们将趋势估计模块应用于 FISH 数据集。为控制随机效应，我们从一个占 70% 的子集中进行抽样，并进行了 100 次自助迭代。线性回归模型为 高度 特征赋予了负系数，而其他趋势估计器则报告了正趋势（图 4）。

为了评估趋势估计量对噪声的稳健性，我们采用了与创建 SYN1 时类似的随机混合策略。FISH 数据经过标准化后，与比例从 0% 到 99% 的随机噪声混合，再作为趋势估计模块的输入。我们发现，随着特征与目标变量之间的相关性 r 和 ρ 降低，线性模型和 RTR 会变得不稳定；而 ATR 和 Shapley 度量 r_S和 ρ_S 即使在混合比例高得多时，也基本不受影响（图 5）。

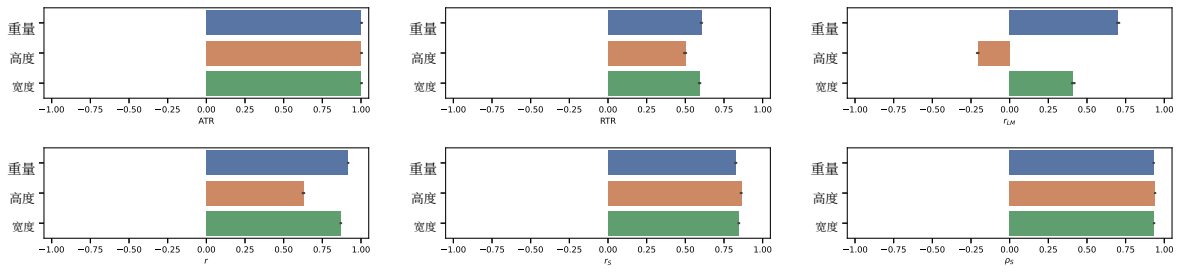


FIGURE 4. FISH 趋势估计量的比较。我们报告了对数据进行 100 次自助抽样迭代（每次包含 70% 的数据）所得不同趋势估计量的均值和标准差。

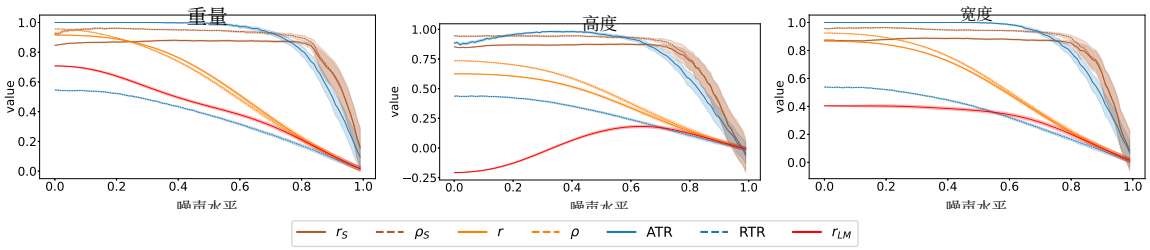


FIGURE 5. 在 FISH 数据上，对噪声进行 100 次独立迭代所得的均值和 95% 置信区间。 x–轴表示混入真实数据中的噪声比例。

4.1.3. *HOUSING*。 随机森林在 HOUSING 数据集上运行了 100 次，不同趋势估计量的均值和标准差如图 6 所示。特征人口 与房价呈非常弱的负相关。然而，所有趋势估计量都报告人口呈显著负趋势。特征房间总数与目标变量呈正相关。然而，线性模型为房间总数赋予了负系数。其他所有趋势估计量都报告了正趋势。

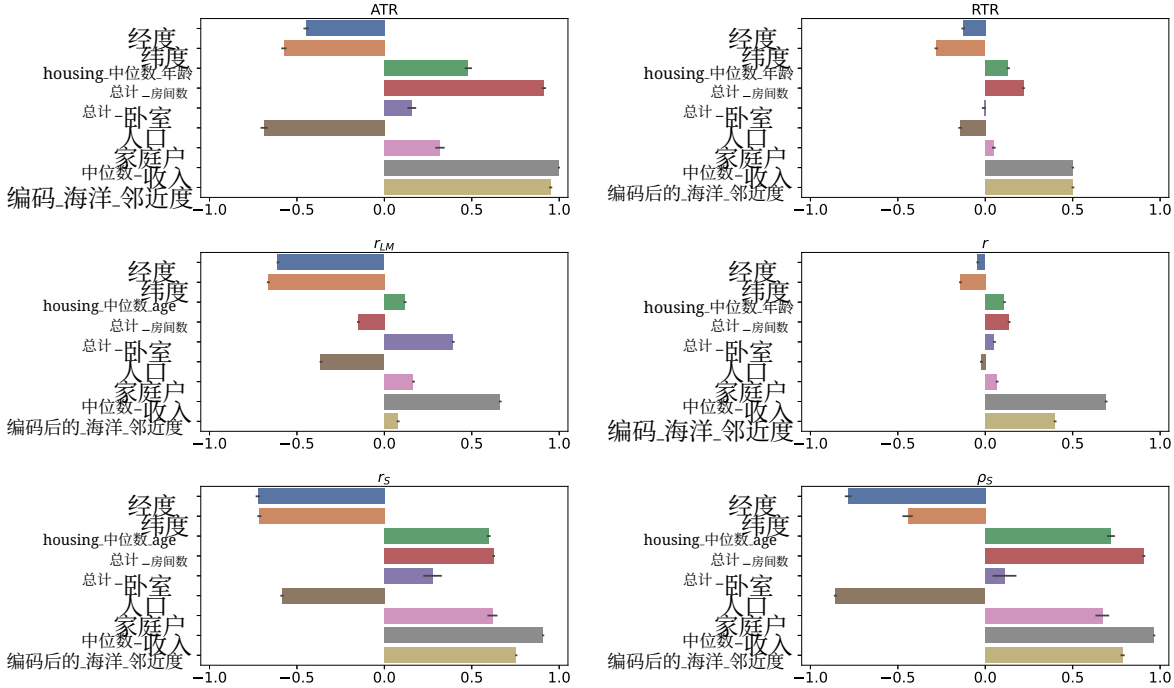


FIGURE 6. HOUSING 数据上趋势估计量的比较。线性模型为房间总数这一特征赋予了负系数，尽管该特征本身与目标变量呈正相关。

4.2. **重要性度量**。我们比较基于不纯度的特征重要性、基于置换的特征重要性，以及由所述两种残差算法（残差学习和格拉姆-施密特去相关）产生的特征重要性。一次运行包括拟合一个随机森林。为确定基于残差的重要性得分，对每个特征分别独立抽取 20 个置换，使该特征位于最后一个位置。

首先，我们比较了合成数据集 SYN2 和 SYN3 上的不同分数（见图 7）。或许最重要的观察结果是，基于不纯度的特征重要性在这两个数据集中都为所有特征赋予了相同的分数。这与

所有其他特征重要性分数形成了鲜明对比。值得注意的是，给定数据集上，两种基于残差的方法得出的分数非常接近。两种基于残差的方法和基于置换的分数，都为 X_0 和带有轻微噪声的变体 A_0赋予了大致相同的分数。然而，受到更多噪声影响的特征 A_1在基于残差的评分下获得了显著更高的分数。尤其是在 SYN3 上，基于残差的方法为所有信息特征都赋予了不算太低的分数。信息特征 X_1 和 X_2 的基于置换的分数相对较低。然而，所有分数都认为，带有噪声的实例 A_0 的重要性高于信息特征 X_1 和 X_2。

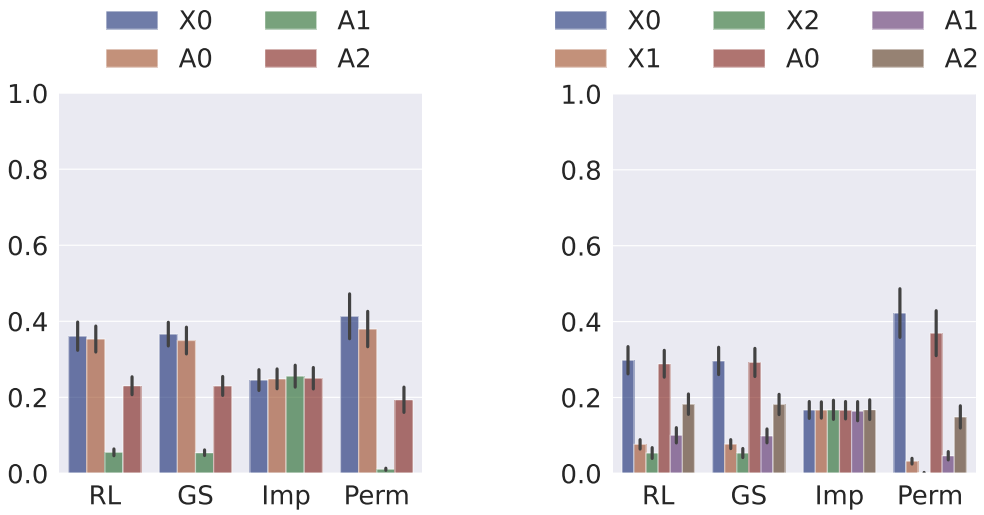


FIGURE 7. 合成数据上四种不同特征重要性概念的比较。左侧展示了 SYN2 的结果。此处，标签生成方式为 $Y = 4 \cdot X_0^{1.5}$，而 $\{A_i\}$ 则由 $X_0 + \mathcal{W}_i$ 针对不同强度的噪声 $\mathcal{W}_i$生成。右侧报告了 SYN3 的结果。此处，标签生成方式为 $Y = 4 \cdot X_0^{1.5} + 2 \cdot X_1 + 0.5 \cdot X_2^2$，因此还给出了另外两个（较弱的）信息特征。

接下来，我们比较真实数据集 HOUSING 和 FISH 上的不同分数（见图 8）。对于 HOUSING，最引人注目的是，残差学习、基于不纯度的分数和基于置换的分数都将最大值赋予收入中位数，其次是距海洋的距离和纬度/经度；而基于格拉姆－施密特的分数仅为收入中位数赋予较大值，其他所有特征的分数都相近。此外，与基于残差学习的方法相比，基于不纯度和基于置换的方法认为人口更为重要。

对于 FISH，所有指标都将最高分赋予重量，并且都为三个变量赋予非零分数。然而，在基于残差学习、基于不纯度和基于置换的方法中，宽度的分数都在一个标准差范围内。只有基于格拉姆－施密特的分数会为重量赋予显著更高的值，并将高度视为第二重要的特征。

5. CONCLUSION

我们提出了两种基于随机森林回归、用于数据集单调趋势的新型估计量。与常被提出的线性模型系数相比，它们的表现可靠得多，而且对噪声具有鲁棒性。不过，SHAP值的表现同样出色，且从理论角度来看，更容易理解。尽管如此，我们认为基于横向速率的方法有其优点。它仅依赖随机森林模型（该模型在某个数据集上训练），模型一旦建立，计算就完全独立于具体数据。另一方面，SHAP值是结合模型和一些数据计算得出的（这或许也有其优点）。

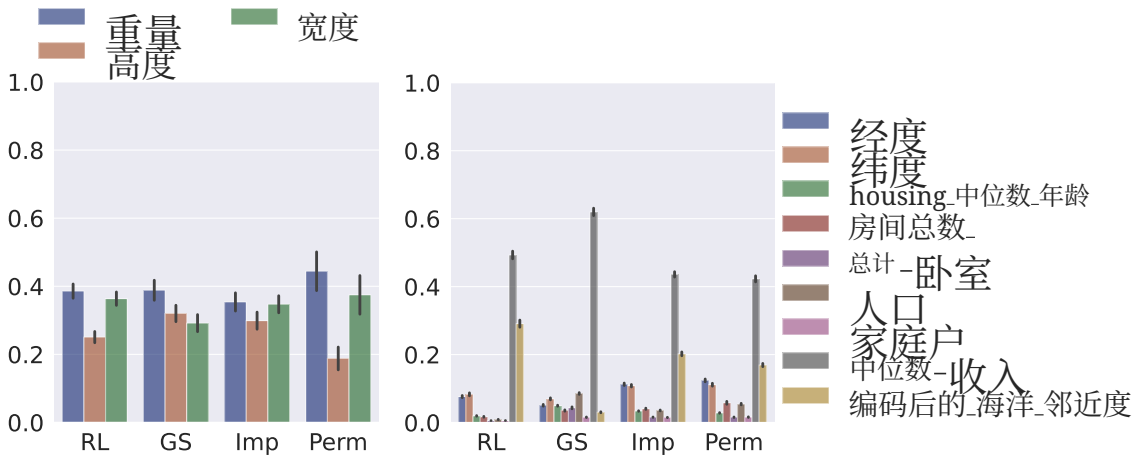


FIGURE 8. 在真实数据实例上比较四种不同的特征重要性概念。左侧展示 FISH 数据集上的特征重要性分数（400 次独立运行的均值和标准差），右侧展示 HOUSING 数据集上的结果（100 次独立运行的均值和标准差）。

关于特征重要性，我们介绍了基于残差的方法。我们比较了合成数据和两个真实实例上的结果。值得注意的是，两种基于残差的方法在合成数据集上得出的结果相近，但这可能是因为噪声是线性添加的。总体而言，与基于不纯度的方法相比，基于残差的方法在处理高度相关的特征时表现好得多。在许多方面，其结果与基于置换的方法相当。不过，我们也发现了显著差异。具体而言，对于对噪声贡献较弱的信息特征，基于残差的方法赋予的值高于基于置换的分数。因此，我们认为，对于具有高度依赖特征的数据集，应优先采用基于残差的特征重要性分数。

References

[1] Aung Pyae。鱼市场数据集，2019年。版本 2，访问日期：2023年2月24日。[2] Mario Beraha、Alberto Maria Metelli、Matteo Papini、Andrea Tirinzoni 和 Marcello Restelli。基于互信息的特征选择：新的理论见解。 *2019 International Joint Conference on Neural Networks (IJCNN)* 2019 ，。[3]

Anne-Laure Boulesteix、Silke Janitza、Jochen Kruppa 和 Inke R. K¨onig。随机森林方法概述及实践指南，重点关注计算生物学与生物信息学。 *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2 6 493 507, 2012 ，（）：–10月。

[4] Leo Breiman。 *Machine Learning*，45(1):5–32，2001年。[5] Cameron Nugent。加利福尼亚州房价，2017年。版本 1，访问日期：2023年2月24日。[6] Rung-Ching Chen、Christine Dewi、Su-Wen Huang 和 Rezzy Eko Caraka。基于机器学习方法选择用于数据分类的关键特征。 *Journal of Big Data*，7(1)，2020年。[7] Davide Chicco 和 Giuseppe J urman。采用集成学习方法，提升肝炎和肝硬化患者的分类效果。 *IEEE Access*，9:24485–24498，2021年。[8] Amir Dezfouli、Hassan Ashtiani、Omar Ghattas、Richard Nock、Peter Dayan 和 Cheng Soon Ong。解缠的行为表征。 *Advances in Neural Information Processing Systems (NeurIPS)*，32 2019，。[9] D. A. S. Fraser。论统计学中的信息。 *The Annals of Mathematical Statistics*，36(3):890–896，1965年6月。[10] Isabelle Guyon、Jason Weston、Stephen Barnhill 和 Vladimir Vapnik。 *Machine Learning*，46(1/3):389–422，2002年。[11] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*，第770–778页，2016年。[12] Dominik Janzing、Lenon Minorics 和 Patrick Bloebaum。可解释人工智能中的特征相关性量化：一个因果问题。载于 Silvia Chiappa 和 Roberto Calandra（编）， *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 108 *Proceedings of Machine Learning Research* 2907 2916 26 28 2020

Proceedings of Machine Learning Research 卷，第 – 页。PMLR，– 8月。[13] Gilles Louppe。理解随机森林：从理论到实践。 *arXiv:1407.7502*，2014年。[14] Scott M. Lundberg 和 Su-In Lee。解释模型预测的统一方法。 *31st International Conference on Neural Information Processing Systems (NeurIPS)* 4768 4777, 2017 ，第 – 页。[15]

A. Nisthana Parveen、H. Hannah Inbarani 和 E. N. Sathish Kumar。无监督特征选择方法的性能分析。 *2012 International Conference on Computing Communication and Applications* 2012 ，。[16] Mirka Saarela 和 Susanne Jauhiainen。比较特征重要性度量作为分类模型的解释。 *SN Applied Sciences*，3(2)，2021年2月。[17] L. S. Shapley。 *17. A Value for n-Person Games*，第307–318页。普林斯顿大学出版社，普林斯顿，1953年。[18] Mukund Sundararajan 和 Amir Najmi。用于模型解释的多种 Shapley 值。载于 Hal Daum´e III 和 Aarti Singh（编）， *Proceedings of the 37th International Conference on Machine Learning*，第119卷， *Proceedings of Machine Learning Research*，第9269–9278页。PMLR，2020年7月13–18日。[19] Kun Zhang 和 Lai-Wan Chan。作为独立成分分析去膨胀方法的降维。 *IEEE Signal Processing Letters* 13 1 45 48, 2006 ，（）：–1月。

Frankfurt Institute for Advanced Studies Goethe University Frankfurt Frankfurt 60325 , , Germany (e-mail: gerstorfer@fias.uni-frankfurt.de)

Faculty of Computer Sciences TU Dortmund University Dortmund 44227 Germay (e-mail: , , lena.krieg@tu-dortmund.de

Faculty of Computer Sciences, TU Dortmund University, Dortmund 44227 Germay (e-mail: maximilian.hahnklimroth@tu-dortmund.de