arXiv:2301.04740v1 [cs.LG] 11 Jan 2023

Berkelmans-Pries 特征重要性方法：衡量特征信息量的通用方法

Joris Pries[1, *]，Guus Berkelmans[1], Sandjai Bhulai[2], 和 Rob van der Mei[1]

[1]Centrum Wiskunde & Informatica，随机学系，Science Park 123，阿姆斯特丹 1098 XG，荷兰[2]Vrije Universiteit，数学系，De Boelelaan 1111，阿姆斯特丹 1081 HV，荷兰[*]通讯作者：Joris Pries，joris.pries@cwi.nl

2023年1月13日

摘要

在过去几年中，机器学习模型的使用已成为一种通用且强大的预测手段。与此同时，人们对预测模型的可解释性 需求也日益增长。为了确定数据集中的哪些特征对于预测目标变量 Y 很重要，可以使用特征重要性 （FI）方法。通过量化每个特征对于预测 Y 的重要程度，可以识别并移除无关特征，从而提高模型的速度和准确性；此外，还可以发现重要特征，进而获得有价值的洞见。评估特征重要性方法时，一个主要问题是，特征重要性的真实值通常未知。因此，现有的特征重要性方法无法给出完全准确的 FI 值。这也是难以正确解读特征重要性方法结果的诸多原因之一。基于此，我们提出了一种名为 $Berkelmans\text{-}Pries$特征重要性方法的新型全局方法，该方法结合了 ***Shapley*** 值 和$Berkelmans\text{-}Pries$ 依赖函数。我们证明，该方法具有许多有用的性质，并且在若干可以精确推导出特征重要性的真实值的情况下，能够准确预测正确的 FI 值

。我们通过实验表明，对于大量的特征重要性方法集合（468 种），现有方法并不具备同样的有用性质。这说明，Berkelmans-Pries FI 是分析具有复杂相互依赖关系的数据集的极有价值工具。

1 引言

你有多重要？这是研究人员（尤其是数据科学家）多年来一直思考的问题。研究人员需要了解随机变量（RV）有多重要 X ，以确定 Y。哪些特征对于预测天气很重要？能否找到可作为特定疾病症状的指标？能否舍弃冗余变量以提升性能？这类问题几乎与所有研究领域都相关。尤其是在如今，随着机器学习模型的兴起，人们需要揭开预测模型的神秘面纱。
Altmann 等人 [3]指出：“在生命科学中，机器学习模型的可解释性与其预测准确率同等重要。”尽管这句话可能并不适用于所有研究领域，但可解释性非常有用。了解预测是如何得出的以及为何如此，对于在日常生活中调整这些方法至关重要。

确定特征重要性 （FI）是一门发现每个特征在 X_i 预测 Y时重要程度的艺术。以下两种情况尤其有用。（I）寻找不重要的变量：可以使用特征重要性方法发现冗余变量。由于高维度和无关信息，无关特征可能会降低预测模型的性能[26]。因此，消除冗余特征可以同时提高预测模型的速度和准确率。（II）寻找重要的变量：重要的 特征可能揭示潜在结构，从而带来有价值的洞见。观察到变量 X 对于预测 Y 很重要，可能会引导研究工作朝着正确的方向发展。尽管必须牢记，高 FI 并不意味着因果关系。然而，FI 值确实能够，例如，“让麻醉医师通过了解患者和操作过程中哪些属性促成了当前预测的风险，从而更好地制定诊断”[36]。通过这种方式，特征重要性方法能够产生真正有意义的影响。

多年来，人们提出了许多特征重要性方法，导致同一数据集会得到范围很广的特征重要性值。例如，随机方法甚至无法反复预测出相同的特征重要性值。这使得结果难以解读。比如，看看 Fryer 等人[17], 的一项研究结果：一种度量方法将某个变量的特征重要性评为 3.19，而另一种方法给出的

同一变量的特征重要性值为 0。265。这引发了许多疑问：“哪种特征重要性方法才是正确的？”“这个变量是否被视为重要？”以及更广泛地说，“这能告诉我们什么信息？”。要评估一种特征重要性方法的性能，需要知道真实值，但实际中往往并非如此 [1,21, 56, 61]。因此，在开发特征重要性方法时，重点尚未放在预测精确的正确特征重要性值上。此外，许多特征重要性方法并不具备理想性质。例如，包含相同信息量的两个特征应获得相同的特征重要性。我们之后会说明，实际情况往往并非如此。

为提高可解释性，我们提出了一种新的特征重要性方法，称为 *Berkelmans-Pries* 特征重要性方法。该方法基于***Shapley*** 值 [49] 和 *Berkelmans-Pries* 依赖函数[5]。已有多种方法采用 Shapley 值，而 Shapley 值已被证明具有诸多优良性质。然而，进一步 使用*Berkelmans-Pries* 依赖函数，还能获得更多实用性质。值得注意的是，我们证明，在某些特征重要性的真实值可以精确推导的情况下，这种方法能够准确预测特征重要性。将***Shapley*** 值与*Berkelmans-Pries*依赖函数相结合，便构成了一种强大的特征重要性方法。鉴于以下原因，这项研究是特征重要性领域的一项重要进展：

- 我们提出了一种新的特征重要性方法；
- 我们证明了该方法具有多种实用性质；
- 我们给出了一些可以精确推导特征重要性的真实值的情形；
- 我们证明，在这些情形下，我们的特征重要性方法能够准确预测正确的特征重要性；
- 我们收集了现有特征重要性方法中规模最大的集合；
- 我们检验了这些方法是否符合相同的性质，结果表明，没有一种方法接近满足所有有用的性质；
- 我们提供了用于确定 FI 值的 Python 代码 [44]。

2 Berkelmans-Pries FI

Kruskal [27]指出：“关联度量有无数种可能，有时似乎曾经提出过的度量几乎也有这么多种。”这句话虽然是在谈依赖函数，但用来描述特征重要性方法也同样恰当。多年来，人们提出了许多特征重要性

方法，但何时以及为何应使用哪种方法，仍不清楚 [21]。本节中，我们提出另一种新的特征重要性方法，称为*Berkelmans-Pries* 特征重要性方法（BP-FI）。尽管人们对特征重要性方法的期望无疑因人而异，但我们将在第 3 节展示 BP-FI 具备许多实用且直观的性质。BP 特征重要性方法基于两个关键要素：（1）*Shapley* 值和（2）*Berkelmans-Pries* 依赖函数。我们将先讨论这两个组成部分，以阐明 BP-FI 方法的工作原理。

2.1 Shapley 值方法

夏普利值 是一种独特的博弈论方法，它基于四条公理，为参与多人博弈的每位参与者分配价值。这一概念广泛应用于特征重要性方法中，因为它可以自然地用于确定每个[49]特征在预测目标变量时有多重要（价值），其中每个特征对应一名参与者，预测目标变量则对应一场博弈。设 N_{vars} 为特征数量，则特征 i的 Shapley 值定义为

$$\phi_i(v) = \sum_{S \subseteq \{1,\ldots,N_{\text{vars}}\} \setminus \{i\}} \frac{|S|! \cdot (N_{\text{vars}} - |S| - 1)!}{N_{\text{vars}}!} \cdot \left(v(S \cup \{i\}) - v(S)\right), \quad (1)$$

其中， $v(S)$可解释为联盟S [49]的“价值”。这一公式背后的原理也可以这样表述：对于每一种可能的特征排列，在特征 i之前，特征 i 带来的增量价值等于纳入该特征前（即 $v(S)$）与纳入后（即 $v(S \cup \{i\})$）的价值之差。对所有可能的特征排列所产生的增量价值取平均，即得到特征 i的最终 Shapley 值。

SHAP 现有多种采用 Shapley 值的特征重要性方法[14, 17, 35], ，这立即保证了若干有用的性质。其中最著名的方法是 SHAP [35]。该方法广泛用于局部解释（见第 4.1 节）。为衡量特定样本 x在预测模型 f下的局部特征重要性，将 v用作特征函数（即方程（1）中的条件期望）。设 $x = (x_1, x_2, \ldots\ldots, x_{N_{\text{vars}}})$，其中 x_i是特征 i的取值，则可通过以下方式确定 SHAP 特征重要性值：

$$v_x(S) := \mathbb{E}_z \left[f(z) | z_i = x_i \text{ for all } i \in S, \text{ where } z = (z_1, \ldots, z_{N_{\text{vars}}}) \right]. \quad (2)$$

请注意，特征函数 v_x 是针对每个 x在局部定义的。若要获得全局 FI 值，可以对所有局部 FI 值取平均。我们提出的新颖

特征重要性方法采用了不同的特征函数，即*Berkelmans-Pries* 依赖函数。这带来了许多额外的实用性质。此外，本研究关注的不是局部 解释，而是全局FI 值。

2.2 Berkelmans-Pries依赖函数

一种新的依赖度量，称为*Berkelmans-Pries* (BP) 依赖函数，于[5], 中提出，并用于构建 BP-FI 方法。研究表明，BP 依赖函数满足一系列理想性质，而现有的依赖度量则不具备这些性质。它有一种测度论表述，但当所有变量均为离散变量时，该表述可简化为一个更简单、直观的版本 [5]。我们希望重点介绍这一表述，以便直观理解BP 依赖函数。其定义如下：

$$\operatorname{Dep}(Y|X) := \begin{cases} \frac{\operatorname{UD}(X,Y)}{\operatorname{UD}(Y,Y)} & \text{if } Y \text{ is not a.s. constant,} \\ \text{undefined} & \text{if } Y \text{ is a.s. constant,} \end{cases} \tag{3}$$

其中（在离散情形下）有：

$$\operatorname{UD}(X,Y) := \sum_x p_X(x) \cdot \sum_y \left| p_{Y|X=x}(y) - p_Y(y) \right|. \tag{4}$$

BP 依赖度量可以这样理解：分子是 Y 的分布与给定 X 时 Y 的分布之间绝对差值的期望。如果 Y 高度依赖于 X，那么由于了解 X 能够提供关于 Y 的信息，分布就会发生变化；而如果 Y 与 X 相互独立，这两个分布则没有差异。分母是任意变量的 Y 分布可能发生的最大变化，用于对依赖函数进行标准化。请注意，BP 依赖函数是非对称的：依赖度 Dep $(Y|X)$ 表示 Y 对 X 的依赖，而非反过来。由于具有许多理想性质，BP 依赖函数被用于 BP-FI。

2.3 Berkelmans-Pries FI方法

将 SHAP 值的博弈论方法转用于特征重要性领域的一个关键环节，是选择函数 v 在 方程（1）中的形式。

该函数为每个特征集 S 赋予一个值 $v(S)$，用于刻画该集合 S的“价值”。如何定义这一函数，会对最终的特征重要性产生关键影响。我们选择将集合 S 的“价值”定义为 Y对特征集 S的 BP 依赖度，记为 Dep $(Y|S)$ [5]。这里，Dep $(Y|S)$ $=$Dep $(Y|Z_S(\mathcal{D}))$，其中 $\mathcal{D}$ 表示包含全部特征的完整数据集，而 $Z_S(\mathcal{D})$表示将数据集缩减为仅包含特征子集 S后的结果。令 Ω_{feat} 表示所有特征变量构成的集合。现在，对于每个 $S \subseteq \Omega_{\text{feat}}$，我们定义：

$$v(S) := \text{Dep}\,(Y|S)\,. \tag{5}$$

换言之，特征集 S 的值恰好表示目标变量对特征集中的特征Y的依赖程度 S。差值 $v(S \cup \{i\}) - v(S)$ 在方程（1）中现在可视为目标变量 Y 对特征集的依赖程度的增加，其中还已知特征 i 。将 BP依赖函数用作特征函数所得的 Shapley 值定义为 BP 特征重要性结果。对于每个特征 i，我们得到：

$$\begin{aligned}\text{FI}(i) &:= \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{i\}} \frac{|S|! \cdot (N_{\text{vars}} - |S| - 1)!}{N_{\text{vars}}!} \cdot (v(S \cup \{i\}) - v(S)) \\ &= \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{i\}} \frac{|S|! \cdot (N_{\text{vars}} - |S| - 1)!}{N_{\text{vars}}!} \cdot (\text{Dep}\,(Y|S \cup \{i\}) - \text{Dep}\,(Y|S))\,.\end{aligned} \tag{6}$$

简写记号可提高后续推导的可读性，因此我们定义

$$w(S, N_{\text{vars}}) := \frac{|S|! \cdot (N_{\text{vars}} - |S| - 1)!}{N_{\text{vars}}!}, \tag{N1}$$

$$D(X, Y, S) := \text{Dep}\,(Y|S \cup \{X\}) - \text{Dep}\,(Y|S)\,. \tag{N2}$$

请注意，当 Y是几乎必然为常数（即 $\mathbb{P}(Y = y) = 1$），对于任何特征集，依赖度$(Y|S)$都未定义 S （参见方程（3））。我们认为，自然的假设是，特征重要性(i) 也未定义，因为每个特征同时决定一切又什么也不决定。在本文余下部分中，我们假设 Y 并非几乎必然为常数。

3 BP-FI 的性质

请注意，特征重要性方法难以评估，因为特征重要性的真实值往往未知[1, 21, 56, 61]。有鉴于此，我们希望证明 BP-FI 方法具有许多理想性质。我们还给出一些合成案例，在这些案例中，BP-FI 方法得出的结果符合自然预期。BP-FI 方法基于*Shapley* 值，它们是由四条公理[60]确定的唯一解。这些公理已经赋予特征重要性方法许多理想特征。此外，使用BP 依赖函数可确保该方法具有其他理想性质。本节将证明 BP-FI 方法的性质，并讨论这些性质为何相关且有用。

性质 1 （效率）。所有 FI 分数之和等于 Y 对所有特征的总依赖性：

$$\sum_{i\in\Omega_{\text{feat}}} \text{FI}(i) = \text{Dep}\left(Y|\Omega_{\text{feat}}\right).$$

证明。Shapley 值具有效率性，这意味着所有价值都会分配给各参与者。因此，

$$\sum_{i\in\Omega_{\text{feat}}} \text{FI}(i) = v(\Omega_{\text{feat}}) = \text{Dep}\left(Y|\Omega_{\text{feat}}\right). \qquad \square$$

相关性。采用我们的方法，我们试图回答这样一个问题：“每个特征对总依赖性贡献了多少？”在我们的情形中，总“收益”就是总依赖性。因此，将全部收益（但不超过全部收益）分配给所有特征是很自然的做法。

推论 1.1。如果向数据集添加随机变量 X 不会带来任何额外信息（即 $\text{Dep}\left(Y|\Omega_{\text{feat}}\cup X\right) = \text{Dep}\left(Y|\Omega_{\text{feat}}\right)$），那么所有 FI 之和保持不变。

证明。这可直接由性质 1得出。 $\square$

相关性。如果集体知识保持不变，可供分配给各特征的信用总量也保持不变。只有在加入新信息时，总信用才应增加。该推论的一个直接结果是，向数据集中添加某个变量的克隆 （即 $X^{\text{clone}} := X$）X 永远不会增加特征重要性的总和。

性质 2 （对称性）。 如果对于每个 $S \subseteq \Omega_{\text{feat}} \setminus \{i,j\}$ ，都有 Dep ($Y|S \cup \{i\}$) = Dep ($Y|S \cup \{j\}$)，那么 FI(i) = FI(j)。

证明。 Shapley 值具有对称性，这意味着如果对于每个 $S \subseteq \Omega_{\text{feat}} \setminus \{i,j\}$， v ($S \cup \{i\}$) = v($S \cup \{j\}$)，则 FI(i) = FI(j)。因此，BP-FI 也自动具有对称性。 □

相关性。如果两个变量可以互换，也就是说它们始终对依赖关系的贡献相同，那么让它们获得相同的特征重要性才合理。这是特征重要性方法的一项理想性质，因为贡献相同的两个特征应该获得相同的特征重要性。

性质 3 （范围）。 对于任意随机变量 X，均有 FI(X) $\in [0,1]$。

证明。 BP 依赖函数在 X[5], 的函数作用下是 非递增的，这意味着对于任意可测函数 f ，均有

$$\text{Dep}\,(Y|f(X)) \leq \text{Dep}\,(Y|X)\,.$$

取 $f := Z_S$，它是将 $\mathcal{D}$ 缩减为S中的特征子集的函数。利用BP 依赖函数的非递增性质，可得:

$$\begin{aligned} \text{Dep}\,(Y|S) = \text{Dep}\,(Y|Z_S(\mathcal{D})) = \text{Dep}\left(Y|Z_S(Z_{S\cup\{i\}}(\mathcal{D}))\right) \\ \leq \text{Dep}\left(Y|Z_{S\cup\{i\}}(\mathcal{D})\right) = \text{Dep}\,(Y|S\cup\{i\})\,. \end{aligned} \tag{7}$$

考察公式（6），我们可以看到，每个 FI 值都必须大于或等于零，因为依赖度 ($Y|S \cup \{i\}$) − 依赖度 ($Y|S$) ≥ 0.
BP 依赖函数的性质之一是，对于任意 X, Y，都有依赖度 ($Y|X$) $\in [0,1]$ [5]。根据性质 1，所有 FI 值的总和因此必定为 $[0,1]$,，因为 $\sum_{i\in\Omega_{\text{feat}}}$ FI(i) = 依赖度 ($Y|\Omega_{\text{feat}}$) $\in [0,1]$。这给出了 FI 值的一个上界，因此我们现在可以得出结论，对于任意随机变量 X，FI(X) $\in [0,1]$ 。 □

相关性。为了确保可解释性，特征重要性方法必须受已知界限约束。例如，如果上界或下界未知，那么 FI 分数 4.2 就无法得到恰当解释。

性质 4 (界限)。 当 $X \in \Omega_{\text{feat}}$ 时，每个特征重要性(X)均有界

$$\frac{\text{Dep}\,(Y|X)}{N_{\text{vars}}} \leq \text{FI}(X) \leq \text{Dep}\,(Y|\Omega_{\text{feat}})\,.$$

证明。 上界由性质 1 和 3可得:

$$\mathrm{Dep}\left(Y|\Omega_{\mathrm{feat}}\right)=\sum_{i\in\Omega_{\mathrm{feat}}}\mathrm{FI}(i)\geq\mathrm{FI}(X),$$

其中，最后一个不等式成立，是因为对于所有 $i\in\Omega_{\mathrm{feat}}$，$\mathrm{FI}(i)\in[0,1]$。

下界可以利用公式（7）中的公式（6）不等式来证明。由此可得（利用记号（N1））

$$\begin{aligned}\mathrm{FI}(X)&=\sum_{S\subseteq\Omega_{\mathrm{feat}}\setminus\{X\}}w(S,N_{\mathrm{vars}})\cdot\Big(\mathrm{Dep}\left(Y|S\cup\{X\}\right)-\mathrm{Dep}\left(Y|S\right)\Big)\\&\geq w(0,N_{\mathrm{vars}})\cdot\left(\mathrm{Dep}\left(Y|\varnothing\cup\{X\}\right)-\mathrm{Dep}\left(Y|\varnothing\right)\right)\\&=\frac{0!\cdot(N_{\mathrm{vars}}-0-1)!}{N_{\mathrm{vars}}!}\cdot\mathrm{Dep}\left(Y|X\right)\\&=\frac{\mathrm{Dep}\left(Y|X\right)}{N_{\mathrm{vars}}}.\end{aligned}$$

□

相关性。这些界限对后续证明很有用。

性质 5 （零特征重要性）。对于任意随机变量 X，都有

$$\mathrm{FI}(X)=0\Leftrightarrow\mathrm{Dep}\left(Y|S\cup\{X\}\right)=\mathrm{Dep}\left(Y|S\right)\ \text{for all}\ S\in\Omega_{\mathrm{feat}}\setminus\{X\}.$$

证明。 $\Leftarrow$: 当 $\mathrm{Dep}(Y|S\cup\{X\})=$依赖度 $(Y|S)$ 对所有 $S\in\Omega_{\mathrm{feat}}\setminus\{X\}$均成立时，由公式（6）（结合记号（N1））可直接推出

$$\begin{aligned}\mathrm{FI}(X)&=\sum_{S\subseteq\Omega_{\mathrm{feat}}\setminus\{X\}}w(S,N_{\mathrm{vars}})\cdot\Big(\mathrm{Dep}\left(Y|S\cup\{X\}\right)-\mathrm{Dep}\left(Y|S\right)\Big)\\&=\sum_{S\subseteq\Omega_{\mathrm{feat}}\setminus\{X\}}\frac{|S|!\cdot(N_{\mathrm{vars}}-|S|-1)!}{N_{\mathrm{vars}}!}\cdot0\\&=0.\end{aligned}$$

$\Rightarrow$: 假设特征重要性$(X)=0$。由性质 3的证明可知，依赖度$(Y|S\cup\{X\})-$依赖度 $(Y|S)\geq0$ 对每个 $S\subseteq\Omega_{\mathrm{feat}}\setminus\{X\}$都成立。如果

依赖度 $(Y|S^* \cup \{X\})$ – 依赖度$(Y|S^*)$ > 0 对于某个给定的 $S^* \in \Omega_{\text{feat}} \setminus \{X\}$，由公式（6）(其中记号（N1）)可得

$$\begin{aligned}
\text{FI}(X) &= \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{X\}} w(S, N_{\text{vars}}) \cdot \big(\text{Dep}\,(Y|S \cup \{X\}) - \text{Dep}\,(Y|S)\big) \\
&\geq w(S^*, N_{\text{vars}}) \cdot (\text{Dep}\,(Y|S^* \cup \{X\}) - \text{Dep}\,(Y|S^*)) \\
&= \frac{|S^*|! \cdot (N_{\text{vars}} - |S^*| - 1)!}{N_{\text{vars}}!} \cdot (\text{Dep}\,(Y|S^* \cup \{X\}) - \text{Dep}\,(Y|S^*)) \\
&> 0.
\end{aligned}$$

这与 FI(X) $= 0$ 的假设相矛盾，因此这样的 S^* 不可能存在。这意味着，对于所有 $S \in \Omega_{\text{feat}} \setminus \{X\}$，Dep $(Y|S \cup \{X\})$ =Dep $(Y|S)$。 □

相关性。当一个特征从不贡献任何信息时，它不应获得任何特征重要性，这才公平。由于该特征对目标变量没有影响，因此可以将其从数据集中移除。另一方面，如果一个特征的特征重要性为零，但它实际上在某些情况下贡献了信息，那么这对该特征是不公平的。此时应给予它一定的特征重要性，尽管数值较小，但应大于零。

零独立性：当对于所有 $S \in \Omega_{\text{feat}} \setminus \{X\}$，Dep $(Y|S \cup \{X\})$ =Dep $(Y|S)$ 时，特征获得零特征重要性；这一性质与博弈论中的零玩家概念相同。Berkelmans 等人 [5] 证明，当 Y 与 X 独立时，Dep $(Y|X) = 0$。要成为零玩家，需要对独立性作出更严格的定义，我们称之为零独立性。如果对于所有 $S \in \Omega_{\text{feat}} \setminus \{X\}$，$Y$ 在 X 上满足 Dep $(Y|S \cup \{X\})$ =Dep $(Y|S)$，则称 Y 在 X 上是零独立的。换言之，当且仅当 FI(X) $= 0.$ 时，X 是零独立的。

推论 5.1。 独立特征 $\not\Rightarrow$零独立特征。

证明。举例来说，考虑一个包含两个二元特征 $X_1, X_2 \sim \mathcal{U}(\{0, 1\})$ 和一个目标变量 $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$ 的数据集，其中 $+ X_2 \cdot (1 - X_1)$ 是 X_1 和 X_2 的异或。单独来看，这些变量并不能提供有关 Y的任何信息；但合在一起，它们可以完全确定 Y。在性质 15 的证明中，我们表明，这会导致特征重要性 FI(X_1) = FI(X_2) = $\frac{1}{2}$，而依赖度 Dep $(Y|X_1)$ =Dep $(Y|X_2)$ $= 0$。因此，X_1 和 X_2 相互独立，但并非零独立。 □

推论 5.2。 独立特征 $\Leftarrow$零独立特征。

证明。 当 X 是零独立时，特征重要性 $\mathrm{FI}(X) = 0$。根据性质 4，我们得到

$$0 = \mathrm{FI}(X) \geq \frac{\mathrm{Dep}\,(Y|X)}{N_{\mathrm{vars}}} \Leftrightarrow \mathrm{Dep}\,(Y|X) = 0.$$

因此，当 X 是零独立时，它也独立。 □

推论 5.3。 几乎必然为常数的变量具有零特征重要性。

证明。 如果 X 是几乎必然为常数（即 $\mathbb{P}(X = x) = 1$），则可立即推出，对于任意 $S \subseteq \Omega_{\mathrm{feat}} \setminus \{X\}$，$\mathrm{Dep}\,(Y|S \cup \{X\}) = \mathrm{Dep}(Y|S)$，因为 Y 的分布不受 X影响。 □

性质 6 （特征重要性等于 1）。 当 $\mathrm{FI}(X) = 1$时，$\mathrm{Dep}\,(Y|X) = 1$，且所有其他特征均为零独立。

证明。 由于BP 依赖函数有界于 $[0, 1]$ [5]，，根据性质 1可得$\sum_{i \in \Omega_{\mathrm{feat}}}$特征重要性$(i) \leq 1$。注意到，由于$[0, 1]$ 根据性质 3，每个特征重要性都必须属于，我们得到

$$\mathrm{FI}(X) = 1 \Rightarrow \mathrm{FI}(X') = 0 \text{ for all } X' \in \Omega_{\mathrm{feat}} \setminus \{X\}.$$

因此，所有其他特征都是零独立的。接下来，我们证明当 $\mathrm{FI}(X) = 1$ 时，$\mathrm{Dep}\,(Y|X) = 1$ 也必须成立。假设 $\mathrm{Dep}\,(Y|X) < 1$。利用

公式（6）（结合 记号（N1）和（N2）），我们得到

$$
\begin{aligned}
1 = \mathrm{FI}(X) &= \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{X\}} w(S, N_{\text{vars}}) \cdot D(X, Y, S) \\
&= \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{X\} : |S| > 0} (w(S, N_{\text{vars}}) \cdot D(X, Y, S)) + w(\varnothing, N_{\text{vars}}) \cdot D(X, Y, \varnothing) \\
&\leq \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{X\} : |S| > 0} (w(S, N_{\text{vars}}) \cdot (1 - 0)) + w(\varnothing, N_{\text{vars}}) \cdot (\mathrm{Dep}\,(Y|X) - 0) \\
&< \sum_{S \subseteq \Omega_{\text{feat}} \setminus \{X\}} w(S, N_{\text{vars}}) \\
&= \sum_{k=0}^{N_{\text{vars}}-1} \binom{N_{\text{vars}} - 1}{k} \cdot \frac{k! \cdot (N_{\text{vars}} - k - 1)!}{N_{\text{vars}}!} \\
&= \sum_{k=0}^{N_{\text{vars}}-1} \frac{(N_{\text{vars}} - 1)!}{k! \cdot (N_{\text{vars}} - 1 - k)!} \cdot \frac{k! \cdot (N_{\text{vars}} - k - 1)!}{N_{\text{vars}}!} \\
&= \sum_{k=0}^{N_{\text{vars}}-1} \frac{1}{N_{\text{vars}}} \\
&= 1.
\end{aligned}
$$

请注意，这一步不等式是根据BP 依赖函数的范围得出的（即 $[0, 1]$）。可能的最大增量出现在 $\mathrm{Dep}\,(Y|S \cup \{X\})$–$\mathrm{Dep}\,(Y|S) = 1 - 0 = 1$ 时。这个结果造成了矛盾，因为 $1 < 1$ 不可能成立，这意味着 $\mathrm{Dep}\,(Y|X) = 1$。 □

相关性。当一个变量的特征重要性为 1 时，其余变量的特征重要性应为 0。此外，这意味着该变量包含了完全确定 Y 所需的信息，因此 $\mathrm{Dep}\,(Y|X) = 1$ 应当成立。

性质7。依赖度$(Y|X) = 1 \not\Rightarrow$ 特征重要性$(X) = 1$.

证明。作为反例，考察存在多个完全决定 Y 的变量的情况。性质 1 和 3 仍必须成立。因此，如果特征重要性

对于每个完全决定 Y 的变量都为 1，则我们得到

$$\sum_{i\in\Omega_{\text{feat}}} \text{FI}(i) \geq 1+1 \neq 1 = \text{Dep}\left(Y|\Omega_{\text{feat}}\right),$$

这是一个矛盾。 □

相关性。这一性质对于解释 FI 分数很重要。当 $\text{FI}(X) \neq 1$ 时，不能自动得出结论说 Y 并非由 X完全决定。

如果 Y 由 X完全决定，我们就称 X 完全信息，因为它提供了确定 Y所需的全部信息。

性质 8（完全信息时的最大特征重要性）。 如果 X 是完全信息，则对于任意 $i \in \Omega_{\text{feat}}$，都有 $\text{FI}(i) \leq \text{FI}(X)$。

证明。 假设存在一个特征 i，使得 $\text{FI}(i) > \text{FI}(X)$，且当$Y$ 由 X完全决定时。要获得更高的特征重要性，在公式（6）的求和中，某处必须由 i 获得比 X更高的增益。注意，对于任意 $S \subseteq \Omega_{\text{feat}} \setminus \{i, X\}$，都有

$$\begin{aligned}\text{Dep}\left(Y|S\cup\{i\}\right) - \text{Dep}\left(Y|S\right) &\leq 1 - \text{Dep}\left(Y|S\right)\\ &= \text{Dep}\left(Y|S\cup\{X\}\right) - \text{Dep}\left(Y|S\right).\end{aligned}$$

对于任意满足 $X \in S$的 $S \subseteq \Omega_{\text{feat}} \setminus \{i\}$，都有

$$\begin{aligned}\text{Dep}\left(Y|S\cup\{i\}\right) - \text{Dep}\left(Y|S\right) &= \text{Dep}\left(Y|S\cup\{i\}\right) - 1\\ &= 0.\end{aligned}$$

最后一步由公式（7）得出，因为依赖函数是递增的，因此 $\text{Dep}(Y|S\cup\{i\}) = 1$。换言之，在夏普利值中，无法相对于 X 获得任何可能的增益。因此，$\text{FI}(i) > \text{FI}(X)$。 □

相关性。 只要某个变量能够完全确定 Y，它就应获得最高的特征重要性。比这样的分数更高的特征重要性意味着什么？它所提供的信息是否超过了最大信息量？如果不满足这一性质，就会导致解释过程令人困惑且难以进行。

性质9（限制结果空间）。对于任意可测函数 f 和随机变量 X，将 X 替换为 $f(X)$，绝不会提高赋予该变量的特征重要性。

证明。BP 依赖函数 在对 X[5]应用函数后不会增加。这意味着，对于任意可测函数 g，均有

$$\mathrm{Dep}\left(Y|g(X)\right) \leq \mathrm{Dep}\left(Y|X\right).$$

令 g 为这样一个函数：它将任意特征集 S 与原始随机变量 X 的并映射为 S 与替代项 $f(X)$ 的并。换言之，对于任意特征集 S，有 $g(S \cup \{X\}) = S \cup \{f(X)\}$。于是可得：

$$\mathrm{Dep}\left(Y|S \cup \{f(X)\}\right) = \mathrm{Dep}\left(Y|g(S \cup \{X\})\right) \leq \mathrm{Dep}\left(Y|S \cup \{X\}\right),$$

and

$$\mathrm{Dep}\left(Y|S \cup \{f(X)\}\right) - \mathrm{Dep}\left(Y|S\right) \leq \mathrm{Dep}\left(Y|S \cup \{X\}\right) - \mathrm{Dep}\left(Y|S\right)$$

对于任意 $S \subseteq \Omega_{\text{feat}} \setminus \{X\}$。因此，利用 公式（6），我们可以得出结论：将 X 替换为 $f(X)$ 绝不会增加所赋予的特征重要性。 □

相关性。这对于预处理而言是一个重要的观察结果。每当一个变量被分箱时，使用较少分箱所获得的特征重要性会更低（或相等）。如果在将 X 替换为 $f(X)$ 之前已知特征重要性，这也可能提供一个有用的上界。

推论 9.1。对于任意可测函数 f 和随机变量 X，若对于另一个随机变量 X'，有 $X = f(X')$，则用特征 X' 替换特征 X 绝不会降低所赋予的特征重要性。

证明。当 $X = f(X')$ 成立时，再次可得（类似于 性质9）：

$$\mathrm{Dep}\left(Y|S \cup \{X\}\right) = \mathrm{Dep}\left(Y|S \cup \{f(X')\}\right) \leq \mathrm{Dep}\left(Y|S \cup \{X'\}\right)$$

对于任意 $S \subseteq \Omega_{\text{feat}} \setminus \{X\}$。因此，根据 公式（6），注意到将$X$ 替换为 X' 永远不会降低所赋予的特征重要性。 □

当特征函数 v为次可加和/或超可加 [49]时，Shapley 值具有其他性质。我们证明，由公式（5）定义的函数两者皆非。

性质 10（既非次可加也非超可加）。 我们的特征函数 $v(S) =$ 依赖度 $(Y|S)$ 既非次可加，也非超可加。

证明。 考虑以下两个反例。

次可加反例：函数 f 是次可加的，如果对于任意 S， $T \in \Omega_{\text{feat}}$都有

$$f(S \cup T) \leq f(S) + f(T).$$

考察由两个二元特征 $X_1, X_2 \sim \mathcal{U}(\{0,1\})$ 和一个目标变量 $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$ 组成的数据集，其中目标变量是X_1 和 X_2的异或。 X_1 和 X_2 各自都不能提供关于 Y分布的任何新信息，因此 $v(X_1) = v(X_2) = 0$ （参见 BP 依赖函数 [5]的性质）。然而，它们合在一起可以完全确定 Y，因此 $v(X_1 \cup X_2) = 1$。由此我们可以得出结论： v 不是次可加的，因为

$$v(X_1 \cup X_2) = 1 \not\leq 0 + 0 = v(X_1) + v(X_2).$$

超可加反例：函数 f 是超可加的，如果对于任意 S， $T \in \Omega_{\text{feat}}$ 都有

$$f(S \cup T) \geq f(S) + f(T).$$

考虑由两个二元特征 $X \sim \mathcal{U}(\{0,1\})$ 和一个克隆 $X^{\text{clone}} := X$组成的数据集，其中目标变量 Y 定义为 $Y := X$. 注意， X 和 X^{clone} 都可以完全确定 Y，因此 $v(X) = v(X^{\text{clone}}) = 1$ （参见BP 依赖函数 [5]的性质）。结合 X 和 X^{clone} 也可以完全确定 Y，由此得到：

$$v(X \cup X^{\text{clone}}) = 1 \not\geq 1 + 1 = v(X) + v(X^{\text{clone}}).$$

因此， v 也不是超可加的。 □

相关性。如果特征函数 v 是次可加的，则对于任意 $X \in \Omega_{\text{feat}}$，都有 $\text{FI}(X) \leq v(X)$。当 v 是超可加时，对于任意 $X \in \Omega_{\text{feat}}$，可推出 $\text{FI}(X) \geq v(X)$。这有时也称为个体理性，意味着任何参与者获得的收益都不会低于其单独行动时能够获得的收益。在由人类参与者构成的博弈论场景中，这一性质是合理的，因为当某个参与者不合作反而能获得更多收益时，他可以选择不参与

合作。在我们的情况下，特征没有自由意志，因此这一性质并非必要。上述证明表明，在我们的情况下，v 既不是次可加的，也不是超可加的，因此我们无法使用与之对应的界限。

性质 11（添加特征可以提高特征重要性）。向数据集中添加额外特征时，X 的特征重要性可能会提高。

证明。考虑前面提到的异或数据集，其中 X_1，$X_2 \sim \mathcal{U}(\{0,1\})$ 和 $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$。如果一开始 X_2 不在数据集中，那么 X_1 的特征重要性为零，因为依赖度 $(Y|X_1) = 0$。然而，如果将 X_2添加到数据集中，X_1 的特征重要性就会增加到 $_1{}_2$（参见 性质 15）。因此，添加另一个特征可以提高某个特征的特征重要性。

□

性质 12（添加特征可以降低特征重要性）。向数据集中添加额外特征时，X 的特征重要性可能会降低。

证明。考虑由 $X \sim \mathcal{U}(\{0,1\})$ 和 $Y := X$. 给出的数据集。由此立即可得，特征重要性$(X) = 1$。然而，引入克隆后$(X^{\text{clone}} := X)$，由于性质 8，有特征重要性(X) =特征重要性(X^{clone})。此外，由性质 1 可得，特征重要性(X) +特征重要性$(X^{\text{clone}}) = 1$。因此，特征重要性$(X) = \frac{1}{2}$，所以添加另一个变量可能会降低某个变量的特征重要性。 □

相关性。需要注意的是，一个变量的特征重要性取决于其他特征（性质 11 和 12）。添加或删除特征可能会改变特征重要性，因此需要对此有所了解。

性质 13（克隆不会增加特征重要性）。对于任意随机变量 $X \in \Omega_{\text{feat}}$，向数据集添加一个相同的变量 $X^{\text{clone}} := X$（克隆），不会增加 X的特征重要性。

证明。令 $\text{FI}_{\text{with clone}}(X)$ 表示添加克隆 X^{clone} 后 X 的特征重要性。

利用 公式（6）（结合记号（N1）和（N2）），我们得到

$$\begin{aligned}
\mathrm{FI}_{\text{with clone}}(X) &= \sum_{S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S) \\
&\overset{(a)}{=} \sum_{S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}:X^{\text{clone}}\in S} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S) \\
&\quad + \sum_{S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}:X^{\text{clone}}\notin S} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S) \\
&\overset{(b)}{=} \sum_{S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}:X^{\text{clone}}\in S} w(S, N_{\text{vars}}+1)\cdot 0 \\
&\quad + \sum_{S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}:X^{\text{clone}}\notin S} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S) \\
&= \sum_{S\subseteq\Omega_{\text{feat}}\setminus\{X\}} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S).
\end{aligned}$$

等式（a）可通过将对 $\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}$ 所有子集的求和拆分为 X^{clone} 是否属于该子集两种情况得出。向一个已包含克隆 X^{clone} 的子集中添加 X 不会改变 BP 依赖函数，因此可得等式（b）。由此推得，对所有子集 $S\subseteq\Omega_{\text{feat}}\cup\{X^{\text{clone}}\}\setminus\{X\}$ 求和可简化为对 $S\subseteq\Omega_{\text{feat}}\setminus\{X\}$求和。

将新的特征重要性$_{\text{with clone}}(X)$与原始 FI(X)进行比较，可得

$$\begin{aligned}
\mathrm{FI}(X) - \mathrm{FI}_{\text{with clone}}(X) &= \sum_{S\subseteq\Omega_{\text{feat}}\setminus\{X\}} w(S, N_{\text{vars}})\cdot D(X,Y,S) \\
&\quad - \sum_{S\subseteq\Omega_{\text{feat}}\setminus\{X\}} w(S, N_{\text{vars}}+1)\cdot D(X,Y,S).
\end{aligned}$$

使用记号（N1），我们发现

$$\frac{w(S, N_{\text{vars}}+1)}{w(S, N_{\text{vars}})} = \frac{\frac{|S|!\cdot(N_{\text{vars}}+1-|S|-1)!}{(N_{\text{vars}}+1)!}}{\frac{|S|!\cdot(N_{\text{vars}}-|S|-1)!}{N_{\text{vars}}!}} = \frac{N_{\text{vars}}-|S|}{N_{\text{vars}}+1} < 1,$$

因此，特征重要性(X) $-$ 特征重要性$_{\text{with clone}}(X)$ ≥ 0，且仅当特征重要性(X) $= 0$ 时等号成立。因此，我们可以得出结论：克隆一个变量不会增加X 的特征重要性；当 X 为零独立时，特征重要性会降低。 □

相关性。我们认为这是良好特征重要性方法的一项自然性质，因为没有任何逻辑上的理由可以解释，为什么加入完全相同的信息会导致原始变量的特征重要性增加。随着某个变量所包含的信息逐渐成为常识，这些信息的价值只会越来越低。

性质 14 （顺序不会改变特征重要性）。 特征的顺序不会影响分别赋予它们的特征重要性。考虑数据集 $[X_1, X_2, \ldots, X_{N_{\text{vars}}}]$和 $[Z_1, Z_2, \ldots, Z_{N_{\text{vars}}}]$，其中 $Z_{\pi(i)} = X_i$对于某个置换 π。对于任意 $i \in \{1, \ldots, N_{\text{vars}}\}$.，都有 FI$(X_i)$ $=$FI$(Z_{\pi(i)})$

证明。 注意，在 BP-FI 的定义（公式（6））中，特征的顺序从未起到任何作用。 BP 依赖函数也与给定的顺序无关，因此这项性质显然成立。 □

相关性。 这是良好特征重要性的一项非常自然的性质。试想一下，如果特征重要性取决于数据集中的顺序，会发生什么？是否需要评估所有可能的顺序并取平均值，才能得到最终的特征重要性？我们找不到任何理由说明为什么有人会希望特征重要性取决于特征的顺序。

数据集

接下来，我们考虑几个数据集，并推导出 BP-FI 的理论结果。这些数据集也用于第4.3节中测试特征重要性方法。评估特征重要性方法非常困难，因为真实值往往未知。不过，我们认为，在这些数据集上得到的特征重要性结果都很自然且有理可据。然而，什么样的 FI 值才算“正确”，仍然是主观的。

性质 15 （异或数据集）。 考虑以下数据集，其中包含两个二元特征 X_1、 $X_2 \sim \mathcal{U}$（$\{0, 1\}$）和一个目标变量 $Y = X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$，该目标变量是 X_1 和 X_2 的异或。可知

$$\text{FI}(X_1) = \text{FI}(X_2) = \frac{1}{2}.$$

证明。 注意到，依赖度 $(Y|X_1)$ $=$依赖度 $(Y|X_2)$ $= 0$ 以及依赖度 $(Y|X_1 \cup X_2)$ $=$

1。 根据 公式（6），可得

$$\mathrm{FI}(X_1) = \sum_{S\subseteq\{1,2\}\setminus X_1} \frac{|S|!\cdot(1-|S|)!}{2!}\cdot\left(\mathrm{Dep}\left(Y|S\cup X_1\right)-\mathrm{Dep}\left(Y|S\right)\right)$$

$$= \frac{|\{\varnothing\}|!\cdot(1-|\{\varnothing\}|)!}{2!}\cdot\left(\mathrm{Dep}\left(Y|\{\varnothing\}\cup X_1\right)-\mathrm{Dep}\left(Y|\{\varnothing\}\right)\right)$$
$$+\frac{|\{X_2\}|!\cdot(1-|\{X_2\}|)!}{2!}\cdot\left(\mathrm{Dep}\left(Y|X_1\cup X_2\right)-\mathrm{Dep}\left(Y|X_2\right)\right)$$

$$= \frac{1}{2}\cdot\left(\mathrm{Dep}\left(Y|X_1\right)-0\right)+\frac{1}{2}\cdot\left(\mathrm{Dep}\left(Y|X_1\cup X_2\right)-\mathrm{Dep}\left(Y|X_2\right)\right)$$

$$= \frac{1}{2}\cdot 0+\frac{1}{2}\cdot(1-0)$$

$$= \frac{1}{2}.$$

利用性质 1，可得特征重要性(X_2) $=1-$ 特征重要性(X_1) $=\frac{1}{2}$。 □

相关性。此异或公式在[17]中被讨论并用于测试特征重要性方法。不过，他们只测试是否相等（特征重要性(X_1) $=$ 特征重要性(X_2)），而不测试具体数值。鉴于对称性，我们也认为 X_1 和 X_2 应获得相同的特征重要性，因为它们发挥着相同的作用。二者共同完全决定Y，因此总特征重要性应为 1（参见 性质 6）。将其平均分配给这两个变量，每个变量的特征重要性便得到逻辑上理想的结果：$\frac{1}{2}$。

性质 16 (概率数据集)。 考虑以下数据集，它由 $Y=\lfloor X_S/2\rfloor$ 和 $X_i=Z_i+(S-1)$ 组成，其中 $Z_i\sim\mathcal{U}(\{0,2\})$ 对应 $i=1$，2 和 $\mathbb{P}(S=1)=p$, $\mathbb{P}(S=2)=1-p$。由此可得

$$\mathrm{FI}(X_1)=p \text{ and } \mathrm{FI}(X_2)=1-p.$$

证明。 注意，根据 方程（4）

$$\begin{aligned}
\mathrm{UD}\left(X_1, Y\right) &= \sum_{x_1 \in \{0,1,2,3\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - p_Y(y) \right| \\
&= \sum_{x_1 \in \{0,2\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - \frac{1}{2} \right| \\
&\qquad + \sum_{x_1 \in \{1,3\}} p_{X_1}(x_1) \cdot \sum_{y \in \{0,1\}} \left| p_{Y|X_1=x_1}(y) - \frac{1}{2} \right| \\
&= \sum_{x_1 \in \{0,2\}} \frac{p}{2} \cdot \left(\left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| \right) \\
&\qquad + \sum_{x_1 \in \{1,3\}} \frac{1-p}{2} \cdot \sum_{y \in \{0,1\}} \left| p_Y(y) - p_Y(y) \right| \\
&= p.
\end{aligned}$$

同理可得，$\mathrm{UD}\left(X_2, Y\right) = 1 - p$。

$$\begin{aligned}
\mathrm{UD}\left(Y, Y\right) &= \sum_{y' \in \{0,1\}} p_Y(y') \cdot \sum_{y \in \{0,1\}} \left| p_{Y|Y=y'}(y) - p_Y(y) \right| \\
&= \sum_{y' \in \{0,1\}} \frac{1}{2} \cdot \left(\left| 1 - \frac{1}{2} \right| + \left| 0 - \frac{1}{2} \right| \right) \\
&= 1.
\end{aligned}$$

由方程（3）可知，依赖度 $(Y|X_1) = p$ 且依赖度$(Y|X_2) = 1 - p$。此外，注意到已知 X_1 和 X_2 即可完全确定 Y，因此

依赖度$(Y|X_1 \cup X_2) = 1$。根据 公式（6），现在可得

$$\begin{aligned}
\mathrm{FI}(X_1) &= \sum_{S \subseteq \{X_1, X_2\} \setminus X_1} \frac{|S|! \cdot (1 - |S|)!}{2!} \cdot (\mathrm{Dep}\,(Y|S \cup X_1) - \mathrm{Dep}\,(Y|S)) \\
&= \frac{|\{\varnothing\}|! \cdot (1 - |\{\varnothing\}|)!}{2!} \cdot (\mathrm{Dep}\,(Y|\{\varnothing\} \cup X_1) - \mathrm{Dep}\,(Y|\{\varnothing\})) \\
&\quad + \frac{|\{X_2\}|! \cdot (1 - |\{X_2\}|)!}{2!} \cdot (\mathrm{Dep}\,(Y|X_1 \cup X_2) - \mathrm{Dep}\,(Y|X_2)) \\
&= \frac{1}{2} \cdot (\mathrm{Dep}\,(Y|X_1) - 0) + \frac{1}{2} \cdot (\mathrm{Dep}\,(Y|X_1 \cup X_2) - \mathrm{Dep}\,(Y|X_2)) \\
&= \frac{1}{2} \cdot (p - 0) + \frac{1}{2} \cdot (1 - (1 - p)) \\
&= \frac{p}{2} + \frac{p}{2} = p.
\end{aligned}$$

利用 性质 1，可得特征重要性$(X_2) = 1$ –特征重要性$(X_1) = 1 - p$. □

相关性。乍看之下，这些 FI 值为何如此自然并不显而易见，因此我们会更详细地讨论这个数据集。 S 可视为一个选择参数，用于决定是否以概率分别为 p 和 $1-p$ 使用 X_1 或 X_2 来表示 Y。 X_i 的构造方式是：根据 S，从 $\{0,2\}$ 或 $\{1,3\}$ 中均匀抽取。不过，正如 $Y = \lfloor X_S/2 \rfloor$ 所示，$X_S = 0$ 和 $X_S = 1$ 对于 Y会得到相同的结果。 $X_S = 2$ 和 $X_S = 3$也同样如此。因此请注意， Y 的分布与选择参数 S 无关。已知 X_1 可以获得以下信息。首先，可以根据 $X_1.$ 的值推导出 S 。当 $X_1 \in \{0,2\}$ 时，必有 $S = 1$；而如果 $X_1 \in \{1,3\}$ ，则可得出 $S = 2$。其次，当 $S = 1$ 时，这意味着 Y完全由 X_1决定。如果 $S = 2$，那么知道 $X_1 = 1$ 或 $X_1 = 3$ 并不会提供关于 Y 的任何额外信息。以概率 p ，知道 X_1 将完全确定 Y；而以概率 $1-p$，它不会提供关于 Y 分布的任何信息。因此，结果 $\mathrm{FI}(X_1) = p$ 非常自然。同样的论证也适用于 X_2，从而得到 $\mathrm{FI}(X_2) = 1 - p.$

4 与现有方法的比较

在上一节中，我们展示了 BP-FI 具有许多理想的性质。接下来，我们将针对大量特征重要性方法，评估这些性质在若干合成数据集上是否成立。请注意，这些数据集只能用作反例，不能用来证明某项性质。首先，我们讨论第 4.1 节 中考察的特征重要性方法。其次，我们介绍这些数据集（第 4.2 节），并说明如何使用它们来检验这些性质（第4.3节）。结果将在第4.4节中讨论。

4.1 其他特征重要性方法

针对各种情况，人们提出了范围广泛的特征重要性方法。因此，首先有必要讨论这些方法之间主要的类别差异。

全局与局部 特征重要性方法的一个重要区别在于，它们是用于生成局部 还是 全局 解释。 全局 特征重要性方法会针对整个数据集中的每个特征给出一个重要性得分，而局部 特征重要性方法则解释哪些变量对单个示例很重要 [18]。全局得分和局部得分不必一致：“全局重要的特征在局部情境中可能并不重要，反之亦然” [46]。本研究聚焦于全局特征重要性方法，但有时也可以对局部特征重要性方法的结果取平均，以获得全局特征重要性。例如，[34]提出了一种称为$Tree\ SHAP$的局部特征重要性方法。该方法也可用于全局分析，即对局部特征重要性的绝对值取平均。

模型特定与模型无关 特征重要性方法还可分为模型-特定方法和-无关方法。模型特定方法旨在借助神经网络或随机森林等预测模型来确定特征重要性，而模型无关 方法则不使用预测模型。BP-FI 是模型无关的，因此能够揭示数据集本身的信息。使用模型特定方法时，关注重点更多在于获取有关预测模型的信息，而非数据集的信息。在我们的测试中，两类方法均有使用。

分类与回归根据具体的数据集，目标变量要么是类别型，要么是数值型，这正是分类与回归的区别。并非所有现有的特征重要性方法都能处理这两种情况。在本研究中，我们生成合成分类数据集，因此

我们只考察适用于这些情况的特征重要性方法。回归数据集还有一个问题：连续变量需要转换为离散分箱。这种转换可能会大幅改变特征重要性分数，从而使得难以得出公平的结论。

集合我们从各种来源收集了目前已知数量最多的特征重要性方法[2, 4, 6, 8, 11–13,17, 18, 20, 22, 28, 35, 38, 40, 42, 43,45, 47, 48, 57, 58]或者由我们自行实现。我们遵循以下政策：只要某个分类特征重要性方法的代码可在R语言或 Python 中获取，或者其实现相对简单，就会将其纳入集合。由此得到 196 种基础方法，以及总计 468 种方法，因为某些基础方法可以与多种机器学习方法或选择目标相结合，参见表 1。不过请注意，大多数方法还包含其他参数，本研究并未考察这些参数。始终使用这些参数的默认值。

4.2 合成数据集

接下来，我们简要讨论用于检验替代特征重要性方法是否具备第 3 节所述性质的数据集。在附录 A 中，我们将逐一介绍各个数据集，并说明其生成方式。为得出公正的结论，这些数据集并非随机抽取，而是固定的。举例来说，为说明我们如何生成数据集，我们考察数据集 1：二元系统（见附录 A ），其中目标变量 Y 定义为 $Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$，其中 $X_i \sim \mathcal{U}(\{0,1\})$ 对所有 $i \in \{1,2,3\}$ 均成立。为获得可解释结果，我们对 X 和 Y 值的每种组合抽取相同次数。表 2 给出了一个示例。大多数数据集的总样本数为 1,000。不过，数据集 6 和 7 包含 2,000 个样本，以确保零独立性。选择这些数据集是为了降低计算成本，并在数据集数量有限的情况下检验多种性质（见第4.3节）。表 3 汇总了生成数据集及其对应的 BP-FI 结果。附录 A 提供了有关特征和目标变量的更多技术细节。

4.3 性质评估

在第 4.1 节中，我们收集了现有的特征重要性方法。本节将评估这些特征重要性方法是否具备 BP-FI 方法同样的理想且已获验证的性质（参见第 3 节）。鉴于特征重要性方法数量庞大（468 种），逐一证明每种方法都具备每项性质并不可行。

表 1：所有经过评估的特征重要性方法：实验中评估的所有特征重要性方法列表。着色标记的方法可与多种选项组合使用：逻辑回归[I], [II], III, 岭回归[I], II, 线性回归[I], II, Lasso[I], II, 随机梯度下降分类器[I], III, 多层感知机分类器[I], II, K近邻分类器[I], II, 梯度提升分类器[I], II, IV, AdaBoost 分类器[I], [II], 高斯朴素贝叶斯[I], II, 伯努利朴素贝叶斯[I], II, 线性判别分析[I], II, 决策树分类器[I], [II], IV, V, 随机森林分类器[I], [II], IV, V, 支持向量分类器[I], CatBoost分类器[I], II, LightGBM分类器[I], II, IV, XGBoost分类器[I], [II], IV, VII, XGBoost随机森林分类器[I], [II], IV, VII, 极端随机树分类器[IV],[V], 极端随机树分类器[IV],[V], 偏最小二乘判别分析[VI], 稀疏偏最小二乘判别分析[VI], 基尼指数[VIII], 熵[VIII], NN1[IX], NN2[IX]。由此，来自各种来源的特征重要性方法总计达到 468 种 [2, 4, 6, 8, 11–13, 17, 18, 20, 22,28, 35, 38, 40, 42, 43, 45, 47, 48, 57, 58]，其余为自行实现的方法。

特征重要性方法

1. AdaBoost 分类器	2. 随机森林分类器[VIII]	3. 极端随机树分类器[V III]	4. 梯度提升分类器
5. 支持向量回归绝对权重	6. 弹性网络绝对权重	7. 置换重要性分类器[I]	8. 主成分分析求和
9. 主成分分析加权	10. 卡方	11. F检验分类	12. 互信息分类
13. KL散度	14. R互信息	15. Fisher Score	16. 特征向量
17. R变量重要性分类器	18. R PIMP分类器	19. TreeInterpreter分类器[V]	20. DIFFI
21. Tree分类器[IV]	22. 线性分类器[II]	23. 置换分类器[I]	24. 分区分类器[I]
25. 抽样分类器[I]	26. 核分类器[I]	27. 精确分类器[I]	28. RFI分类器[I]
29. CFI分类器[I]	30. 求和分类器[VI]	31. 加权X分类器[VI]	32. 加权Y分类器[VI]
33. 单因素方差分析F检验	34. Alexander-Govern检验	35. Pearson相关系数	36. Spearman秩相关系数
37. 点二列相关系数	38. 肯德尔秩相关系数	39. 加权肯德尔秩相关系数	40. 索默斯D
41. 线性回归	42. 西格尔斜率	43. 泰尔斜率	44. 多尺度图相关
45. 提升器权重[VII]	46. 提升器增益[VII]	47. 提升器覆盖度[VII]	48. 共享近邻
49. K近邻	50. 贝叶斯广义线性模型	51. lssvmRadial	52. 稳健一类分类器
53. 最优加权最近邻	54. 正交随机森林偏最小二乘	55. 随机蕨类分类器	56. 树袋装法
57. 正则化随机森林	58. 径向基支持向量机	59. 条件推断树	60. 进化树
61. 多重判别分析	62. rpart	63. cforest	64. 线性支持向量机
65. xyf	66. C5.0Tree	67. 平均神经网络	68. kknn
69. 径向基核支持向量机（成本参数）	70. gausssprRadial	71. FH.GBML	72. 线性支持向量机2
73. bstSm	74. LogitBoost	75. wsrf	76. 惩罚逻辑回归
77. xgbLinear	78. 随机森林	79. 空值	80. protoclass
81. monmlp	82. Rborist	83. mlpWeightDecay	84. svmRadialWeights
85. mlpML	86. 条件推断树	87. loclda	88. SDWD
89. mlpWeightDecayML	90. 径向基核支持向量机	91. 提升树	92. 深度神经网络
93. 有序随机森林	94. 二次判别分析	95. 提升线性模型	96. 全局正则化随机森林
97. 多层感知机	98. 一标准误差剪枝分类树	99. 主成分神经网络	100. 有序随机森林支持向量机
101. 并行随机森林	102. 二次剪枝回归树	103. 多项式核高斯过程分类器	104. C5.0 规则
105. 正则化判别分析	106. 径向基函数动态衰减	107. 多项式	108. 线性高斯过程
109. 多项式支持向量机	110. K近邻	111. 树袋装法	112. 正则化随机森林
113. 条件推断树	114. 进化树	115. 多重判别分析	116. rpart
117. cforest	118. xyf	119. C5.0Tree	120. kknn
121. gausssprRadial	122. LogitBoost	123. wsrf	124. xgbLinear
125. 随机森林	126. 空	127. monmlp	128. Rborist
129. mlpWeightDecay	130. mlpML	131. 条件推断树	132. mlpWeightDecayML
133. 深度神经网络	134. 二次判别分析	135. 全局正则化随机森林	136. 多层感知机
137. 一标准误差剪枝分类树	138. 并行随机森林	139. 二次剪枝回归树	140. 多项式核高斯过程分类器
141. C5.0 规则	142. 径向基函数动态衰减	143. 多项式	144. 线性高斯过程
145. 二元一致性	146. 卡方	147. cramer	148. gainRatio
149. 基尼指数	150. IE一致性	151. IEP一致性	152. 互信息
153. 粗糙集一致性	154. Relief特征集度量	155. 对称不确定性	156. 迭代估计器[II]
157. 置换估计器[II]	158. 核估计器[II]	159. 符号估计器[II]	160. 夏普和值[I]
161. 斯瑟洛关值	162. 随机森林	163. Garson[IX]	164. VIANN[IX]
165. LOFO[IX]	166. Relief	167. ReliefF	168. RReliefF
169. 拟合准则测度	170. F 比率测度	171. 基尼指数	172. SU 测度
173. 斯皮尔曼相关系数	174. 皮尔逊相关系数	175. 费希纳相关系数	176. 肯德尔相关系数
177. 卡方测度	178. anova	179. laplacian score	180. 信息增益
181. 修正 t 分数	182. 最大互信息	183. 最大相关最小冗余	184. 联合互信息
185. 添加：条件互信息特征选择	186. 条件互信息最大化	187. 交互补偿特征选择	188. DCSF
189. CFR	190. 最大相关性指标	191. IWFS	192. NDFS
193. 鲁棒特征选择	194. SPEC	195. MCFS	196. UDFS
197. R2	198. DC	199. BCDC	200. AIDC
201. HSIC	202. BP-FI		

图例

1-12	sklearn	[42]	13-20	其他方法	[2, 8, 11, 18, 22, 38, 45, 48]	21-27	SHAP 解释器	[35]
28-29	相对特征重要性	[6]	30-32	R vip	[20]	33-44	SciPy 统计	[58]
45-47	提升分类器	[12]	48-109	R caret 分类器	[28]	110-144	R 固定分类器	[20]
145-155	R FSinR Classifier	[4]	156-159	Sage 分类器	[13]	160-161	QII 平均分类器	[57]
162-165	Rebelosa 分类器	[47]	166-168	Relief 分类器	[40]	169-196	ITMO	[43]
197-201	Sunnies	[17]	202	**BP-FI**	-			

表 2： **固定抽样：** 数据集抽取方式示例。我们不是对每种可能的结果进行均匀随机抽样，而是以相同的固定次数抽取每种组合。

结果				# 抽取次数	
X_1	X_2	X_3	Y		
0	0	0	0	125	133
0	0	1	4	125	129
0	1	0	2	125	121
0	1	1	6	125	109
1	0	0	1	125	136
1	0	1	5	125	124
1	1	0	3	125	115
1	1	1	7	125	133

相反，我们设计了一些测试，使用生成数据集寻找这些性质的反例（参见第 4.2 节）。由于测试数量较多（18 项），我们只讨论不那么显而易见的部分，因为大多数测试都直接衡量相应的性质。每项测试的概述见 附录 B。测试概要见 表 4，其中列出了每项测试针对哪些数据集检验哪种性质。

计算误差 为容许计算误差，我们在每项测试中接受 $\epsilon = 0.01$ 的误差范围。例如，如果 FI 值应为零，那么得分为0。01 或 -0.01 仍视为通过，而 FI 值为 0.05 则计为未通过。通常，这对特征重要性方法有利。然而，在测试 9 中，我们评估特征重要性方法是否会为非零独立变量赋予零特征重要性。在这种情况下，我们认为 $|\mathrm{FI}(X)| \leq \epsilon$ 为零，因为数据集的构造方式使得变量要么是零独立的，要么与零独立相去甚远。

运行时间 我们将每个数据集在 i7-12700K 处理器上的运行时间限制为一小时，同时运行四种算法。这些数据集的特征数量较少，结果空间也非常有限，样本数量为 1,000 或 2,000，因此一小时是合理的运行时长。

NaN 或无穷值 在某些情况下，某个特征重要性方法会为某个特征分配 NaN 或 $\pm\infty$。我们如何处理这些值取决于具体测试。例如，我们认为 NaN 不在范围 $[0, 1]$ (测试 4 和 55 ）之外，但在评估 FI 值之和是否保持稳定时（测试 2 ），或评估两个对称

onding BP-FI outcome.
ing FI methods adhere

FI outcome
3, 0.333, 0.333)
2, 0.202, 0.298, 0.298)
8, 0.148, 0.183, 0.183, 0.338)
7, 0.117, 0.136, 0.136, 0.248, 0.248)
6, 0.248, 0.248, 0.117, 0.117, 0.136)
0, 0.000, 0.000)
0, 0.000, 0.000, 0.000)
7, 0.342, 0.361)
9, 0.193, 0.204, 0.208, 0.216)
2, 0.253, 0.223, 0.262)
0, 0.000, 0.000)
0, 0.500, 0.000)
3, 0.333, 0.333)
0, 0.500)
0)
7, 0.167, 0.667)
0, 0.500, 0.000)
$-p$)

roperties as the BP-FI,
t is performed on this
The details of the tests

21	22	23	24	25	26	27	28
✓	✓	✓	✓	✓	✓	✓	✓
		✓					
✓	✓	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	✓	✓
							✓
✓	✓	✓	✓	✓	✓	✓	✓
							✓
↕(25)	↕(24)		✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	✓	✓

特征获得相同的特征重要性（测试 3），我们认为两次 NaN 或两次 $\pm\infty$相同。

性质9（限制结果空间）性质9指出，对随机变量应用任意可测函数 f X 不会增加特征重要性。换句话说，特征重要性(X) $\geq$ 特征重要性$(f(X))$ 成立。我们使用 数据集 8 至 10 (参见表 4)来检验此性质。这些数据集包含一些变量，它们是将目标变量按不同数量的分箱进行分箱后得到的结果。检验性质9时，应满足特征重要性(X_i) $\geq$特征重要性(X_j)，只要 X_i的分箱数多于 X_j。

性质11和12（添加特征可以增加/降低特征重要性）在所有其他测试中，目标是找到该性质的一个反例。然而，测试 13和14 旨在评估添加一个特征后，某个特征的重要性是否会增加或降低。增加或降低的幅度应大于 ϵ。数据集经过特意选择，使得根据 BP-FI，特征重要性既可能增加，也可能降低。仅对于这些测试，如果没有发现反例（增加/降低），我们才认为测试失败。

4.4 评估结果

总体结果概览见表5，其中给出了每项测试中通过和未通过的方法数量。接下来，我们将进一步探讨这些实验结果。

表现最佳的方法 通过测试最多的前 20 种特征重要性方法列于 表 6。在 18 项测试中，BP-FI 通过了所有测试，这正如预期，因为我们已在 第 3 节 证明 BP-FI 确实具备这些性质。来自R语言 *FSinR* 分类器 和 *ITMO* 的分类器占据了前 20 名中的 11 个席位。在 11 种 R语言 FSinR 分类器方法中，有六种进入前 20 名，这相当出色。不过请注意，BP 特征重要性方法与排名第二的方法之间的差距为 $18-11=7$ 项通过的测试。此外，468 种方法中有 424 种未能通过一半以上的测试。 图1 展示了通过测试数量的出现频率。各个前 20 名方法未通过哪些测试的详细概览见 表5。再次注意，在 测试 13 和 14中，如果添加特征后 FI 始终没有增加或减少（分别对应两种情况），就会被判为失败。这些方法或许实际上能够使 FI 增加或减少，只是由于某种原因，在我们的数据集中没有出现这种情况。值得注意的是，这些方法大多在具有期望结果的数据集上表现不佳

表5：结果概览：使用附录 B中概述的测试评估每种特征重要性方法，以检查该方法是否遵循与 BP-FI 相同的属性（参见第 3 节）。此表总结了 468 种特征重要性方法中有多少种通过或未通过测试。表中还特别列出了排名前 20 的通过测试的方法。未通过测试意味着找到了一个反例。请注意，通过测试并不能“证明”该特征重要性方法确实具有该属性。无结果表示无法执行测试，因为特征重要性方法的运行时间过长或发生了错误。

	Test																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
总体																		
# 通过	1	92	45	438	200	97	132	283	97	31	141	241	243	314	365	172	13	5
# 失败	466	369	421	29	267	370	335	184	370	413	326	98	216	145	58	288	421	459
# 无结果	1	7	2	1	1	1	1	1	1	24	1	129	9	9	45	8	34	4
前 20 名																		
# 通过	1	10	15	20	19	7	18	18	2	13	17	20	4	6	20	17	2	4
# 失败	19	10	5	0	1	13	2	2	18	7	3	0	16	14	0	3	17	16
# 无结果	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

（测试 17 和 18）。添加一个不包含额外信息的变量（测试 2），通常也会导致总特征重要性发生变化。

测试 1 此项测试用于评估 FI 值之和是否与 BP-FI 的总和相同。乍看之下，这似乎是一个相当严格的要求。然而，对于所有使用的数据集，依赖度$(Y|\Omega_{\text{feat}})$ 要么为零，要么为一。因此，我们实质上是在评估：当所有变量共同完全决定 Y 时，FI 之和是否等于一；而当所有变量均为零独立时，FI 之和是否为零。测试结果显示，除 BP-FI 外，没有任何特征重要性方法能通过此项测试。以下列举几种表现较为接近的方法：162. Rebelosa 分类器 RF、2. 随机森林分类器（熵）、2. 随机森林分类器（基尼指数） 仅在总和应为零（因为零独立）的数据集上未通过测试；而 1. AdaBoost 分类器仅在基于异或函数的四个数据集中，有三个未通过测试（见 附录 A），在这些数据集中总和应为一，但实际为零。特征重要性方法 *51.lssvmRadial* 表现最接近，仅有两项未通过。对于零独立的数据集（数据集 6 和 7），它为每个特征赋予 0.5 的 FI 值，因此总和大于零。

测试 2 在 图 2中，列出了 FI 值总和不稳定的情况分布。错误最多的是 二元系统 数据集，

表 6： **前 20 名：** 在 468 种特征重要性方法中，这 20 种方法通过附录 B 中列出的 18 项测试的次数最多。这些测试旨在检验某种特征重要性方法是否遵循与 第 3 节中所述 BP-FI 相同的性质。 通过 表示来自附录 A 的数据集未提供反例。当然，这并不意味着已证明该特征重要性方法确实具有此性质。 失败 表示发现了反例。 无结果 表示无法执行测试，因为该特征重要性方法的运行时间过长，或发生了错误。

		综合结果：		
方法		# 通过	# 失败	# 无结果
202.	**BP-FI**	**18**	**0**	**0**
147.	cramer	11	7	0
148.	gainRatio	11	7	0
153.	粗糙集一致性	11	7	0
155.	对称不确定性	11	7	0
172.	SU 测度	11	7	0
88.	sdwd	10	7	1
3.	极端随机树分类器	10	8	0
116.	rpart	10	8	0
126.	null	10	8	0
145.	二元一致性	10	8	0
152.	互信息	10	8	0
161.	班扎夫岭	10	8	0
197.	R2	10	8	0
162.	RF	10	8	0
166.	Relief	10	8	0
173.	斯皮尔曼相关系数	10	8	0
188.	DCSF	10	8	0
189.	CFR	10	8	0
191.	I	10	8	0

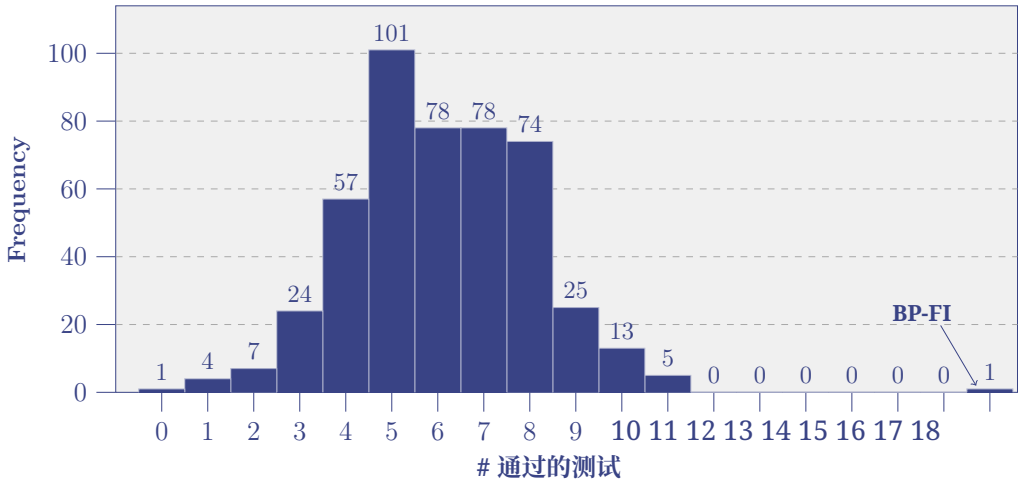


图1： **通过测试的总频次：** 468 种特征重要性方法通过的测试数（共 18 项）的直方图。

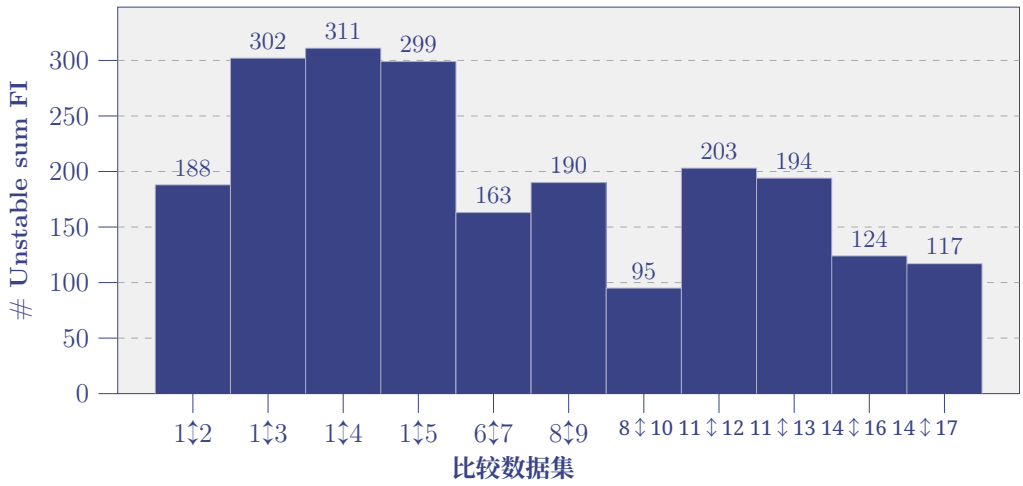


图 2: 不稳定的FI总和: 添加一个不提供任何额外信息的变量时，所有FI之和应保持稳定。对于每次比较，我们统计468种特征重要性方法中有多少次不符合这一点。

添加完全信息特征时，共有92种方法通过了测试，而369种方法未通过。在这369种方法中，有279种至少出现过一次总和增加的情况，另有232种至少出现过一次总和减少的情况。因此，数量惊人的特征重要性方法会在添加不包含任何额外信息的变量时，赋予显著更多或更少的特征重要性。尽管集体知识保持稳定，不足以证明应增加或减少权重，分配的权重却有所增减。此外，如果初始总和与最终总和都包含NaN值，则视为通过。若不采用此规则，92种方法中有3种无法通过。如果只有初始总和或最终总和包含NaN，则视为失败，因为两者的总和并不相同。只有五种方法仅因这一规则而失败：15. Fisher Score、11. F检验分类、178. anova、179. laplacian score和192. NDFS。

测试 11 图3 展示了在数据集中，每个变量的 FI 值与最大 FI 值相差不超过 ϵ 的频率。根据 性质 8，完全信息变量应取得最大的 FI 值。总的来说，我们观察到，与其他变量相比，完全信息变量往往具有最大的 FI 值。然而，仍有许多情况并非如此。326 种特征重要性方法未能通过此测试，因此可以确定它们不具备 性质 8。当一个变量的特征重要性高于一个完全决定目标变量的变量时，便很难进行解释。如果一个变量比能够提供完美信息的变量更重要，这意味着什么?

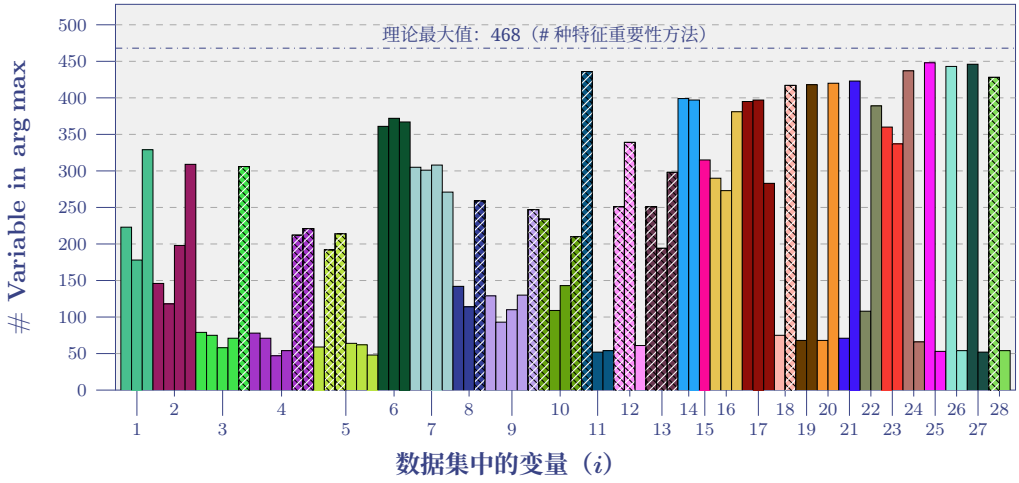


图3： 特征重要性最大值： 对于每个数据集中的每个变量，我们确定它相对于数据集中其他变量获得最大特征重要性的频率（在 ϵ 对 $\epsilon = 0.01$ 的界限内）。完全信息变量应获得最大的特征重要性（参见 性质 8）。图中对所有完全信息变量均进行了阴影标示。

测试 10、17、18 这些测试均用于评估特征重要性方法是否会为某个特征赋予特定值。从表5中可以看到，能够通过这些测试的方法并不多。这并不令人意外，因为这些方法尚未经过充分测试，无法确认它们是否会赋予特定值。这是本研究的重要贡献之一，因此我们希望详细介绍以往研究中所做的尝试。已有许多用于特征重要性的合成数据集 [1–3, 6, 7, 9, 15–17, 19, 21, 23–25, 30, 32–34, 37, 39, 41, 50–52, 55,56, 59, 61, 62], 但并未给出具体且理想的 FI 值。最常见的做法是生成合成数据集，以评估某种特征重要性方法发现噪声特征的能力 [3, 7, 19, 21, 23, 24, 30, 50, 52, 55, 59, 61]。此类数据集的常见总体构想是，目标变量与某些变量相互独立。通常通过比较独立变量与因变量的 FI 值来评估这些 FI 值，目的是确定特征重要性方法能否发现独立变量。至于特征重要性方法是否确实能预测出理想的确切 FI 值，则不在考虑之列。接下来，我们重点介绍对理想 FI 值有所论述的论文。Lundberg 等人 [34]给出了两个相似的数据集，其中一个变量的重要性逐渐增加。他们评估了多种特征重要性方法，以观察这些方法的结果是否体现出相同的行为。这表明，一些常用方法可能会降低某个变量的重要性，尽管该变量的重要性实际上应该逐渐增加。Giles 等人[19] 也设计了多个

用于呈现不同情形的人工数据集，并对哪些变量应具有更高的特征重要性作出评论。Sundararajan 等人 [55] 指出，如果每个特征值都不相同，那么即使函数在变量上并不对称，特征重要性方法（CES）仍会为所有变量分配相同的归因值。如果给每个特征添加少量噪声，所有特征也会获得相同的归因值。不过，他们没有评估这一结果是否合理。Owen 等人 [41] 给出了以下示例。设 $f(x_1, x_2) = 10^6 x_1 + x_2$，其中 $x_1 = 10^6 x_2$；他们认为，尽管 x_1的方差更大，但两个变量同样重要，因为该函数既可以写成仅关于 x_1 的函数，也可以写成仅关于 x_2的函数。虽然我们此前已看到，“写成关于……的函数”并不是一个好的判据（因为存在依赖关系），但我们同意作者的看法，即特征重要性应相等。Owen 等人 [41], 还给出了另一个示例，其中 $\mathbb{P}(x_1 = 0, x_2 = 0, y = y_0) = p_0$, $\mathbb{P}(x_1 = 1, x_2 = 0, y = y_1) = p_1$，而 $\mathbb{P}(x_1 = 0, x_2 = 1, y = y_2) = p_2$是可能的结果。若 $p_0 = 0$，文献 [41]指出， x_1 的 Shapley相对重要性为 1 2，这是“必然如此，因为此时 x_1 与 x_2之间存在双射”。这是一个有趣的观察，因为大多数论文都不会评论某个结果是否合理。此外，当 $y_1 = y_2$ （且 $y_0 \neq y_1$）时，Owen 等人 [41] 认为，最重要的变量是方差最大的变量。Fryer 等人 [17] 也构建了一个二元异或数据集（见数据集 14），并针对该数据集评估了七种特征重要性方法。X_1 和 X_2 的作用具有对称性，因此分配给它们的特征重要性也应相同。结果显示，七种方法中有六种确实给出了对称的结果。不过，具体的特征重要性值差异很大。SHAP 给出的特征重要性为 3.19，而 Shapley DC 分配的特征重要性为 0.265。研究只检验了对称性，并未检验特征重要性方法的准确性。总之，现有研究并未着重预测精确、准确的特征重要性值。因此，特征重要性方法频繁未能通过这些准确性测试，也就不足为奇了。表7更详细地列出了各变量被分配的特征重要性值偏离期望结果 ϵ界限（$\epsilon = 0.01$）的频率。在数据集 11 中，特征重要性方法主要难以为完全信息变量分配 1。总计有 413 种方法未能通过测试10。对于数据集 14 和 17，两个异或变量未通过测试的频率大致相同。比较这两个数据集时，一个有趣的现象是：加入一个零依赖变量后，异或变量更频繁地未通过测试。总计有 421 种方法未能通过测试 17。测试 18 难度较高，因为特征重要性方法必须为所有概率数据集分配正确的值（见附录 A）。只有五种方法通过了该测试：152. 互信息、153. 粗糙集一致性、162. 随机森林、175. 费希纳相关系数和 202. BP-FI。这五种方法也通过了测试 10。不过，除了 BP-FI 之外，只有一种方法也通过了测试 17，即 162. 随机森林。其余三种方法在数据集 14 和 17 中都只分配零，未

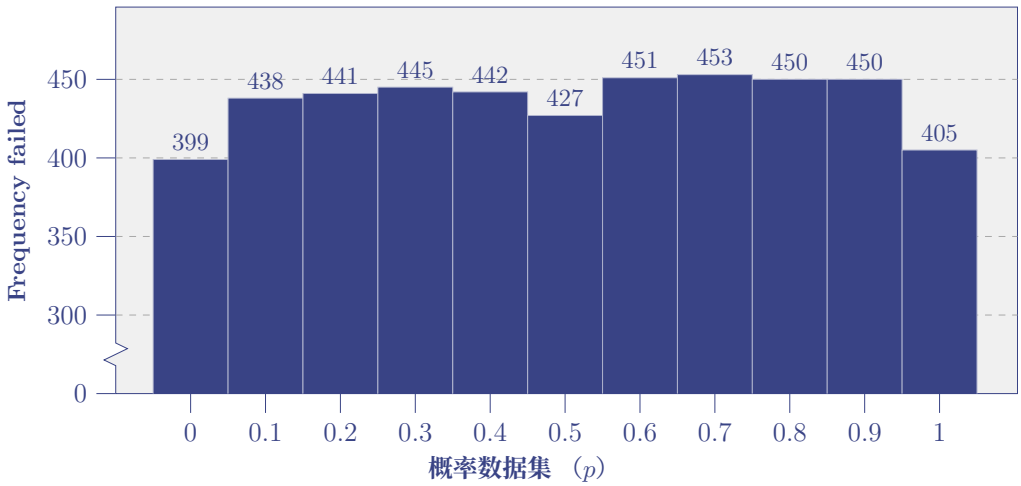


图4： 按数据集划分的测试 18 结果： 在 测试 18 中，特征重要性方法需要为每个概率数据集分配正确的 FI 值（参见 附录 A）。在此图中，我们按数据集细分了特征重要性方法失败的频率。

识别异或变量所取的值，及其信息组合后的结果。在图4中，按概率数据集逐一列出特征重要性方法未通过测试的频率。一个出乎意料的结果是，概率为 $p < \frac{1}{2}$ 的数据集，以及概率为 $1 - p$ 的数据集，未通过测试的频率较低。始终如此的是，$p < \frac{1}{2}$ 未通过测试的频率低于与之对应的 $1 - p$，尽管这两个数据集除特征和样本的顺序不同外完全相同。这种现象在表7中也能看到。

无结果 聚焦于 无结果 行 表5，有一种名为 *158. KernelEstimator* 与 *Lasso* 结合的基础方法，在所有情况下都无法运行或超出运行时间限制。测试 12中大量的无结果主要源于测试所用三个数据集的运行速度过慢。每个数据集至少有 63 种方法运行过慢，这意味着该测试无法执行。

5 讨论与未来研究

尽管我们建议使用新的特征重要性方法，但了解其局限性和潜在陷阱非常重要。下面我们将详细说明所提出方法的不足之处，以及进一步研究所面临的相关挑战。我们首先讨论一些需要

表7： 特定结果： 测试 10、17 和18 均用于评估某种特征重要性方法是否会针对某个数据集给出特定结果。本表概述了这些数据集中各变量被赋予期望结果的一个ϵ界限之外的值（带有 $\epsilon = 0.01$）。

数据集	期望结果	# 非期望结果					
		非 NaN			NaN		
		X_1	X_2	X_3	X_1	X_2	X_3
11	(1, 0, 0)	360	89	88	4	4	4
14	$(\frac{1}{2}, \frac{1}{2})$	353	351	-	5	5	-
17	$(\frac{1}{2}, \frac{1}{2}, 0)$	369	364	90	5	5	5
18	(0, 1)	82	352	-	4	4	-
19	$(\frac{1}{10}, \frac{9}{10})$	412	434	-	3	3	-
20	$(\frac{2}{10}, \frac{8}{10})$	434	438	-	3	3	-
21	$(\frac{3}{10}, \frac{7}{10})$	435	441	-	3	3	-
22	$(\frac{4}{10}, \frac{6}{10})$	439	436	-	3	3	-
23	$(\frac{5}{10}, \frac{5}{10})$	423	422	-	3	3	-
24	$(\frac{6}{10}, \frac{4}{10})$	448	447	-	3	3	-
25	$(\frac{7}{10}, \frac{3}{10})$	449	446	-	3	3	-
26	$(\frac{8}{10}, \frac{2}{10})$	446	444	-	3	3	-
27	$(\frac{9}{10}, \frac{1}{10})$	444	435	-	3	3	-
28	(1, 0)	352	86	-	5	5	-

应用 BP-FI 时需要注意的事项（第5.1节）。接下来，我们讨论第5.2节中实验所采用的一些选择。最后，我们在第5.3节中详细探讨其他可能的研究方向。

5.1 提高认知

分箱Berkelmans 等人[5] 解释说，连续数据的离散化方式会显著影响BP依赖函数，因此我们的研究中使用的所有数据集都是离散的。如果某个特征具有过多不同的取值（由于分箱效果不佳），BP-FI 会赋予它较高的特征重要性，因为这些不同取值可以存储更多信息（见性质9）。另一方面，如果选择的分箱过少，重要特征的特征重要性可能会偏低，因为分箱会导致信息丢失。未来的研究应调查并测试哪些分箱算法最接近底层特征重要性的结果。

样本过少考虑以下数据集：$X_i, Y \sim \mathcal{U}(\{0, 1, \ldots, 9\})$ 对于 $i \in \{1, \ldots, 5\}$ 独立同分布。注意，所有特征都是零独立的，因为 Y 只是均匀抽取的，完全没有考虑任何特征。如果$n_{\text{samples}} = \infty$，那么期望结果应为 (0, 0, 0, 0, 0)。然而，当数据集中的样本不足时，特征会得到非零特征重要性。考虑到所有特征组合起来的不同取值总数为 10^5, ，确实能提供有关 Y 的信息，因为当$n_{\text{samples}} \ll 10^5$时，任何一种可能的特征组合在数据集中出现不止一次的概率都很低。因此，知道所有特征值几乎可以确定 Y 的取值。性质1表明，所有特征重要性的总和应为 1。所有特征变量也都是对称的（性质2），因此期望结果应为 ($\frac{1}{5}$, $\frac{1}{5}$, $\frac{1}{5}$, $\frac{1}{5}$, $\frac{1}{5}$)。这个例子说明，我们应注意样本数量对最终特征重要性的影响。即使变量不影响 Y，在样本不足时仍可能包含信息。因此，如果不警惕这一现象，样本不足可能导致错误结论。

反直觉依赖案例 *Berkelmans-Pries* 依赖度衡量的是，得知 X后， Y 的概率质量有多少发生了偏移。然而，两个相似的概率质量偏移可能会导致不同的预测能力。为了解释这一点，我们考察以下数据集。

$X_1, X_2 \sim \mathcal{U}(\{0,1\})$，其中

$$\mathbb{P}(Y=y|X_1=x_1,X_2=x_2)=\begin{cases}1/4 & \text{if } (x_2,y)=(0,0),\\ 3/4 & \text{if } (x_2,y)=(0,1),\\ 5/8 & \text{if } (x_1,x_2,y)=(0,1,0),\\ 3/8 & \text{if } (x_1,x_2,y)=(0,1,1),\\ 7/8 & \text{if } (x_1,x_2,y)=(1,1,0),\\ 1/8 & \text{if } (x_1,x_2,y)=(1,1,1).\end{cases}$$

知道 X_2 的取值会改变 Y 的分布。起初，Y 分成两类的比例是 50/50；但已知 X_2 的取值后，标签的比例会变为 25/75 或 75/25，具体取决于 X_2 的取值。知道 X_1 能提供更多信息，例如，知道 $X_1 = X_2 = 1$ 后，$Y = 0$ 的可能性会更高。然而，仅知道 X_2 与同时知道 X_1 和 X_2 时，Y 的分布变化相同，因此 Dep $(Y|X_2)$ = Dep $(Y|X_1 \cup X_2)$。这是一个违反直觉的结果。从整体上看，知道 X_2 或 $X_1 \cup X_2$ 会带来相同的分布变化；但在局部层面，如果也知道 X_1，我们就能更准确地预测 Y。我们不确定这会如何影响 BP-FI。在这种情况下，FI$(X_1 \cup X_2)$ > FI(X_2)，这是理想的结果。并非不可能找到一种方法，通过修改依赖函数来为此类情况得出更直观的结果。例如，可以采用不同的距离度量，将给定特征值时的局部准确性纳入考量；也可以采用条件变体，它不仅检验独立性，还检验条件独立性。这些都是亟待深入研究的关键方向。

将 FI 用于特征选择 特征选择 (FS) 是“选择一个较小的特征子集，该子集理想情况下对于描述目标概念既必要又充分的问题” [26]。基本而言，目标是从所有特征中找出一个子集，使给定模型的性能达到最佳，因为较大的特征集可能会降低模型的准确率 [29]。许多 FI 方法实际上源自特征选择过程。不过，重要的是要强调，高 FI 并不意味着该特征应自动被选中。与其他特征共享的信息可能会使该特征不如预期有用。反过来，低 FI 的特征也不应自动舍弃。与其他特征结合使用时，它仍可能提供其他特征无法带来的额外见解。计算 BP-FI 值也有助于了解哪一组 K 特征 Y依赖性最强。为得出 BP-FI 的结果，需要确定所有的依赖关系 Y

在子集 $S \subseteq \Omega_{\text{feat}}$ 上。如果只选择 K 个变量，自然会选择

$$S_K^* \in \underset{S \subseteq \Omega_{\text{feat}}:|S|=K}{\arg\max} \{\text{Dep}(Y|S)\}.$$

这些值作为 BP-FI 的中间步骤存储，因此 S_K^* 之后可以快速推导出来。

结果空间越大，特征重要性越高我们已证明，对于 BP-FI，结果空间增大绝不会导致特征重要性降低。这意味着，可能结果更多的特征更有可能获得较高的特征重要性，具体取决于分布。可能结果很多但几乎从不出现的特征，与许多可能结果都经常出现的特征之间存在差异。我们并不认为这一性质不可取，尽管有些文章提出了这种看法[53, 61]，，因为我们认为，特征可以通过增加结果种类来承载更多信息，从而使特征重要性不降低。

5.2 实验设计选择

回归为避免分箱问题，我们仅考虑了分类模型和数据集。回归特征重要性方法要多得多，也应以类似方式加以考察。然而，要得出清晰且准确的结论，首先需要了解分箱如何影响结果。分箱有时会导致违反直觉的结果，但这些结果未必有误。在这种情况下，至关重要的是不要贬低该特征重要性方法。

运行时间 在实验中，由于运行时间过长或 FI 分数不兼容，某些特征重要性方法可能会出现无结果。每种算法在每个数据集上的最长运行时间设为一小时，测试使用 i7-12700K 处理器，并同时运行 4 种算法。由于特征重要性方法和数据集的数量庞大，因此有必要设置最长运行时间。并行运行四种算法可能会不公平地影响运行时间，因为处理器有时会受到其他算法的限制。有时，其他并行进程已经运行完毕，这可能会缩短某种算法的运行时间。这里存在一种潜在风险：准确但运行缓慢的特征重要性方法可能不会出现在结果中。不过，相对于样本数量和特征数量而言，我们的合成数据集较小，因此我们认为一小时是合理的。

根据具体用途，有时可以花很长时间来确定 FI 值，而在其他情况下，快速确定该值可能至关重要。尤其是对于较大的数据集，运行某些特征重要性方法甚至可能不可行。BP-FI 使用 Shapley 值，随着特征数量增加，其计算难度会呈指数级增长。应开发近似算法，以更快地估计真实的 BP-FI 结果。若运行时间大幅缩短且近似结果足够准确，快速近似方法可能会很有用。已有多篇论文提出了更快地近似 Shapley 值的方法 [1, 10, 24, 31, 54]。这些方法节省了时间，但代价是什么？可以开展一项研究，以确定在给定数据集和允许运行时间的情况下，哪种特征重要性方法最优。

随机性方法 我们未纳入考量的一个因素是，某些特征重要性方法的随机性。有些方法在重复使用时，预测出的 FI 值并不相同。例如，79. 随机森林 对数据集 3 预测的结果为 (12.1, 11.7, 17.9, 15.2, 37.7)，四舍五入到小数点后一位。再次运行该方法会得到不同的结果：(11.4, 12.0, 17.4, 15.6, 37.1)，因为该方法使用了随机性随机森林。原则上，特征重要性方法具有随机性并不可取，因为我们认为，给定一个数据集，其特征重要性应有唯一的分配结果。由于特征重要性方法和数据集的数量较多，我们没有对每种特征重要性方法重复运行并取平均值。不过，这样做可以更全面地了解随机性特征重要性方法的表现。

参数调优 所有特征重要性方法均使用默认参数值。不同的参数值可能导致更多或更少的测试失败。不过，事先并不知道理想的参数设置，因此有必要在较大的参数范围内进行搜索。这并非本研究的重点，但未来的研究可以尝试了解并确定针对给定数据集应选择哪些参数值。

特征重要性方法排名 在 表 6中，突出了通过测试数量最多的 20 种特征重要性方法。不过，必须强调的是，并非每项测试的难度都相同。对不同用户而言，某些属性的重要程度可能有所不同。例如，要准确预测 11 个数据集的具体数值（测试 18），远比始终预测非负值（测试 4）困难。所有测试的权重相同，但这未必能反映准确通过每项测试的难度。不过，我们注意到，175. 费希纳相关系数 是唯一通过了测试 18却最终未进入前 20 名的特征重要性方法。我们强调，我们的重点是查明特征重要性方法是否符合这些属性，而不一定是找出最佳、最公平的排名。

5.3 其他事项

全局与局部 BP-FI 旨在全局确定特征重要性。然而，另一个重要的研究领域关注局部解释。这些解释应说明为什么某个特定样本具有某个目标值，而不是其他值。它们提供了实际应用日益需要的可解释性。这有助于回答诸如“如果我的收入更高，我能贷到更多款吗？”“种族会影响这项预测吗？”以及“对于这项自动化机器学习决策，哪些因素起了关键作用？”等问题。人们已经提出了许多局部特征重要性方法，其中有些甚至采用了 Shapley 值。应对所有已提出的局部方法进行结构化综述，类似于我们针对全局特征重要性方法所采用的做法，以找出哪些局部特征重要性方法能够真正给出准确的解释。

BP-FI 可以修改为提供局部解释。例如，我们可以通过以下方式将特征函数局部化。令 $Y_{S,z}$ 为Y 限制在事件 $X_i = z_i$（对于 $i \notin S$）上的结果，并类似地定义 $X_{S,z}$。然后，我们可以定义一个局部化的特征函数：

$$v_z(S) := \mathrm{Dep}\left(Y_{S,z} | X_{S,z}\right). \tag{8}$$

处理连续数据时，假设相等可能过于严格。在这种情况下，可以使用精度向量参数 ϵ，其中我们将$Y_{S,z,\epsilon}$ 定义为 Y 限制在事件 $|X_i - z_i| \leq \epsilon_i$（对于 $i \notin S$）上的结果，并以同样的方式定义 $X_{S,z,\epsilon}$。由此得到以下局部化的特征函数：

$$v_{z,\epsilon}(S) := \mathrm{Dep}\left(Y_{S,z,\epsilon} | X_{S,z,\epsilon}\right).$$

此外，如果有距离函数 $d(i, j)$ 和参数 δ 可用于判断样本 j 是否足够接近 i、从而被视为“局部”，那么至少有两种方法可以调整 BP-FI，使其适用于局部解释。我们可以 (I) 丢弃所有满足 $d(i, j) > \delta$ 的样本，和/或 (II) 生成样本，使得所有生成的样本均满足 $d(i, j) \leq \delta$。然后，我们可以对剩余样本和/或生成的样本使用 BP-FI，从而得到局部FI。请注意，样本数量仍应足够多，因为如前所述，样本过少可能导致特征重要性结果不同。不过，调整 BP-FI 以用于局部解释的方法还有很多。

特定于模型的特征重要性 BP-FI 原则上是模型无关的，因为特征重要性是根据数据集确定的，而非针对某个预测模型的特征重要性。不过，BP-FI 仍可为任何特定模型提供见解。

将目标变量替换为模型的预测结果后，我们仍可对这个新数据集应用 BP-FI，从而了解哪些特征对预测模型有用。此外，还可以将这些特征重要性结果与原始 FI（即替换目标变量为预测结果之前的结果）进行比较，以了解模型以何种方式改变了特征重要性。

其他性质本研究证明了 BP-FI 的一些性质。不过，深入研究或许能发现更多有用的性质。这既适用于 BP-FI，也适用于其所依据的依赖函数。例如，应用同构并不会改变依赖函数。因此，BP-FI 在同构下也具有稳定性。了解 BP-FI 具备哪些性质是一把双刃剑。发现有用的性质可以彰显 BP-FI 的优势，而发现不理想的行为则可能推动未来的改进。

更多数据集特征重要性的真实值往往未知。本研究给出了两类数据集，其期望结果是显而易见的。不过，如果能为全局和局部特征重要性创建一个包含更多数据集的更大集合，并为其提供精确的真实值，将会很有帮助。我们知道这项工作可能极具挑战，但相信它对进一步改进特征重要性方法至关重要。

人工标注在一些文章中[35, 46]，研究者会借助人类来评估解释。一个值得探究的问题是，人类是否擅长预测特征重要性。BP-FI 可用作基准，以验证参与者给出的数值。人类能否识别特征重要性的正确顺序？更进一步，他们能否预测出接近实际特征重要性值的结果？

6 总结

我们首先介绍了一种名为*Berkelmans-Pries* 特征重要性（BP-FI）的新方法，它结合了*Shapley* 值与*Berkelmans-Pries* 依赖函数[5]。在第 3 节中，我们证明了 BP-FI 的许多有用性质。我们讨论了现有的特征重要性方法，并介绍了一些数据集，用于评估这些方法是否符合相同的性质。在第4.3节中，我们解释了如何测试这些性质。结果表明，与大量特征重要性方法集合（468 种）中的其他任何方法相比，BP-FI 能通过更多测试，这是向前迈出的重要一步。由于缺少真实值，以往尚未对大多数方法进行测试以确认其能否得出精确结果。

在本研究中，我们提供了若干特定数据集，可据此推导出期望的特征重要性。测试结果表明，以往的方法无法准确预测期望的 FI 值。在第5节中，我们深入讨论了本文的不足，以及进一步研究面临的挑战。为了进一步提升数据集和机器学习模型的可解释性与解释性，还有许多值得探索、颇具挑战性的研究机会。

A 数据集

在本附录中，我们讨论实验所用数据集的生成方式。我们使用 固定抽样 而非 均匀随机 来抽取每个数据集，以确保 其严格遵循相应的分布。这样做是为了消除数据集中的随机性，从而获得精确且易于解释的结果。固定抽样与均匀随机抽样之间的差异示例见 表 2。除数据集 6 和 7 外，各数据集均由 1,000 个样本组成；数据集 6 和 7 包含 2,000 个样本，以确保零独立性。这些数据集的设计兼顾计算效率，同时仍能检验多种性质（见第4.3节）。下文将概述用于生成这些数据集的公式，并给出我们提出的新方法 BP-FI 对应的特征重要性值。

数据集 1：二元系统
特征变量：$X_i \sim \mathcal{U}(\{0,1\})$ 对于 $i \in \{1,2,3\}$ 独立同分布。目标变量：$Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$。顺序：$(X_1, X_2, X_3)$。BP-FI：(0.333, 0.333, 0.333)。

数据集 2：带克隆的二元系统
特征变量：$X_i \sim \mathcal{U}(\{0,1\})$，对于 $i \in \{1,2,3\}$ 和 $X_1^{\text{clone}} := X_1$，服从独立同分布。
目标变量：$Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$。顺序：$(X_1^{\text{clone}}, X_1, X_2, X_3)$。BP-FI：(0.202, 0.202, 0.298, 0.298)。

数据集 3：带克隆和一个完全信息变量的二元系统

特征变量：$X_i \sim \mathcal{U}(\{0,1\})$ 对于 $i \in \{1,2,3\}$ 和 $X_1^{\text{clone}} := X_1$ 以及 $X_4^{\text{full}} := Y^2$ 独立同分布。目标变量：$Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$。顺序：$(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}})$。BP-FI：(0.148, 0.148, 0.183, 0.183, 0.338)。

数据集 4：含克隆和两个完全信息变量的二元系统

特征变量：$X_i \sim \mathcal{U}(\{0,1\})$ 对于 $i \in \{1,2,3\}$ 和 $X_1^{\text{clone}} := X_1$ 以及 $X_4^{\text{full}} := Y^2$, $X_5^{\text{full}} := Y^3$ 独立同分布。目标变量：$Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$。顺序：$(X_1^{\text{clone}}, X_1, X_2, X_3, X_4^{\text{full}}, X_5^{\text{full}})$。BP-FI：(0.117, 0.117, 0.136, 0.136, 0.248, 0.248)。

数据集 5：含克隆和两个顺序不同的完全信息变量的二元系统

特征变量：$X_i \sim \mathcal{U}(\{0,1\})$，对于 $i \in \{1,2,3\}$ 和 $X_1^{\text{clone}} := X_1$ 以及 $X_4^{\text{full}} := Y^2$，$X_5^{\text{full}} := Y^3$，均为独立同分布。目标变量：$Y := \sum_{i=1}^{3} 2^{i-1} \cdot X_i$。顺序：$(X_3, X_4^{\text{full}}, X_5^{\text{full}}, X_1^{\text{clone}}, X_1, X_2)$。BP-FI：(0.136, 0.248, 0.248, 0.117, 0.117, 0.136)。

数据集 6：零独立系统

特征变量：$X_i^{\text{null-indep.}} \sim \mathcal{U}(\{0,1\})$，对于 $i \in \{1,2,3\}$，均为独立同分布。目标变量：$Y \sim \mathcal{U}(\{0,1\})$。顺序：$(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$。BP-FI：(0.000, 0.000, 0.000)。

数据集7：含常数变量的零独立系统。特征变量：$X_i^{\text{null-indep.}} \sim \mathcal{U}(\{0,1\})$，对于 $i \in \{1,2,3\}$ 和 $X^{\text{const}} := 1$。目标变量：$Y \sim \mathcal{U}(\{0,1\})$。顺序：$(X_1^{\text{null-indep.}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}}, X^{\text{const}}$, 零独立 4 $)$。BP-FI：(0.000, 0.000, 0.000, 0.000)。

数据集 8：均匀系统递增分箱
特征变量：令 $\mathcal{L}_i := \{0, 1/(i-1) \ldots, 1\}$ 为一个等间距集合。定义：

$$X_1^{\text{bins=10}} := \arg\max_{x_1 \in \mathcal{L}_{10}} \{Y \geq x_1\},$$

$$X_2^{\text{bins=50}} := \arg\max_{x_2 \in \mathcal{L}_{50}} \{Y \geq x_2\},$$

$$X_3^{\text{bins=1,000, full}} := \arg\max_{x_3 \in \mathcal{L}_{1,000}} \{Y \geq x_3\}.$$

目标变量：$Y \sim \mathcal{U}(\mathcal{L}_{1,\ 000})$。顺序：$(X_1^{\text{bins=10}}, X_2^{\text{bins=50}}, X^{\text{bins=1}}{}_{,\ 000,\ \text{完整}\,3})$。BP-FI：（0.297,0.342,0.361）。

数据集 9：均匀系统递增分箱及更多变量
特征变量：令 $\mathcal{L}_i := \{0, 1/(i-1) \ldots, 1\}$ 为一个等间距集合。定义：

$$X_1^{\text{bins=10}} := \arg\max_{x_1 \in \mathcal{L}_{10}} \{Y \geq x_1\},$$

$$X_2^{\text{bins=20}} := \arg\max_{x_2 \in \mathcal{L}_{20}} \{Y \geq x_2\},$$

$$X_3^{\text{bins=50}} := \arg\max_{x_3 \in \mathcal{L}_{50}} \{Y \geq x_3\},$$

$$X_4^{\text{bins=100}} := \arg\max_{x_4 \in \mathcal{L}_{100}} \{Y \geq x_4\},$$

$$X_5^{\text{bins=1,000, full}} := \arg\max_{x_5 \in \mathcal{L}_{1,000}} \{Y \geq x_5\}.$$

目标变量：$Y \sim \mathcal{U}(\mathcal{L}_{1,\ 000})$。顺序：$(X_1^{\text{bins=10}}, X_2^{\text{bins=20}}, X_3^{\text{bins=50}}, X_4^{\text{bins=100}}, X^{\text{bins=1}}{}_{,\ 000,\ \text{完整的}5})$。BP-FI：(0.179,0.193,0.204,0.208,0.216)。

数据集 10：均匀系统递增分箱，克隆顺序不同

特征变量：令 $\mathcal{L}_i := \{0, 1/(i-1) \ldots, 1\}$ 为一个等间距集合。定义：

$$X_1^{\text{bins=10}} := \arg\max_{x_1 \in \mathcal{L}_{10}} \{Y \geq x_1\},$$

$$X_2^{\text{bins=50}} := \arg\max_{x_2 \in \mathcal{L}_{50}} \{Y \geq x_2\},$$

$$X_3^{\text{bins=1,000, full}} := \arg\max_{x_3 \in \mathcal{L}_{1,000}} \{Y \geq x_3\},$$

$$X_3^{\text{clone, full}} := X_3^{\text{bins=1,000, full}}.$$

目标变量：$Y \sim \mathcal{U}(\mathcal{L}_{1,000})$。顺序：$(X^{\text{bins=1}}{}_{,000,\ \text{完全信息 3}}$，$X_2^{\text{bins=50}}, X_1^{\text{bins=10}}, X^{\text{clone}}{}_{,\ \text{完全信息 3}})$。BP-FI：(0.262, 0.253, 0.223, 0.262)。

数据集 11：依赖系统：1 个完全信息变量 特征变量：X_1^{full}，$X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}} \sim \mathcal{U}(\{1, 2\})$。目标变量：$Y := X_1^{\text{full}}$。顺序：$(X_1^{\text{full}}, X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}})$。BP-FI：(1.000,0.000,0.000)。

数据集 12：依赖系统：2 个完全信息变量 特征变量：X_1^{full}，$X_3^{\text{null-indep.}} \sim \mathcal{U}(\{1, 2\})$ 和 $X_2^{\text{full}} := Y^2$。目标变量：$Y := X_1^{\text{full}}$。顺序：$(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{null-indep.}})$。BP-FI：(0.500, 0.500, 0.000)。

数据集 13：依赖系统：3 个完全信息变量。特征变量：$X_1^{\text{full}} \sim \mathcal{U}(\{1, 2\})$ 和 $X_2^{\text{full}} := Y^2$, $X_3^{\text{full}} := Y^3$。目标变量：$Y := X_1^{\text{full}}$。顺序：$(X_1^{\text{full}}, X_2^{\text{full}}, X_3^{\text{full}})$。BP-FI：(0.333, 0.333, 0.333)。

数据集 14：异或数据集
特征变量：$X_1, X_2 \sim \mathcal{U}(\{1, 2\})$。目标变量：$Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$。顺序：$(X_1, X_2)$。
BP-FI：(0.500, 0.500)。

数据集15：单变量异或数据集
特征变量：$X_1^{\text{null-indep.}} \sim \mathcal{U}(\{1,2\})$。目标变量：$Y := X_1^{\text{null-indep.}} \cdot (1 - X_2) + X_2 \cdot (1 - X_1^{\text{null-indep.}})$，其中$X_2 \sim \mathcal{U}(\{1,2\})$。顺序：$(X_1^{\text{null-indep.}})$。BP-FI：(0.000)。

数据集16：带克隆特征变量的异或数据集：$X_1, X_2 \sim \mathcal{U}(\{1,2\})$ 和 $X_1^{\text{clone}} := X_1$。目标变量：$Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$。顺序：$(X_1^{\text{clone}}, X_1, X_2)$。BP-FI：(0.167, 0.167, 0.667)。

数据集17：带零独立特征变量的异或数据集：$X_1, X_2 \sim \mathcal{U}(\{1,2\})$ 和 $X_3^{\text{null-indep.}} \sim \mathcal{U}(\{0,3\})$。目标变量：$Y := X_1 \cdot (1 - X_2) + X_2 \cdot (1 - X_1)$。顺序：$(X_1, X_2, X_3^{\text{null-indep.}})$。BP-FI：(0.500, 0.500, 0.000)。

数据集18-28：概率数据集
特征变量：$X_i = Z_i + S$，其中 $Z_i \sim \mathcal{U}(\{0,2\})$ 对于 $i = 1$、2 和 $\mathbb{P}(S = 1) = p$，$\mathbb{P}(S = 2) = 1 - p$ 独立同分布。目标变量：$Y = \lfloor X_S/2 \rfloor$。顺序：$(X_1, X_2)$。BP-FI：$(p, 1-p)$。

B 测试

本附录概述了用于各特征重要性方法的测试，以评估这些方法是否符合第3节中给出的性质。大多数测试都很简单，但第4.3节中提供了额外的解释。

测试 1：BP-FI 效率总和
评估：性质 1。
说明：我们评估所有 FI 之和是否等于 Y 在所有特征上的 *Berkelmans-Pries* 依赖函数之和。当 FI 值被赋为 NaN 或无穷大时，FI 之和会自动不等于 BP-FI 的总和。

测试 2：效率稳定性
评估：推论 1.1。
说明：每当向数据集添加一个变量时，我们都会检查所有 FI 之和是否发生变化。如果某个变量相较于其他变量没有提供任何额外信息，那么所有 FI 之和应保持不变。

测试 3：对称性
评估：性质 2。
说明：在某些数据集中，存在对称 变量（见性质 2）。我们判断所有对称变量是否获得相同的特征重要性。

测试 4：范围（下限）评估： 性质 3。说明： 我们检查所有特征重要性结果是否大于或等于零。

测试 5：范围（上界）评估：性质 3。
说明：我们检查所有特征重要性结果是否都小于或等于 1。

测试 6：BP-FI 界限（下界）
评估：性质 4。
说明：我们评估性质 4 给出的界限是否也适用于其他特征重要性方法。对于满足 $X \in \Omega_{\text{feat}}$ 的每个 FI(X)，BP-FI 的下界为 $\frac{\text{依赖度}(Y|X)}{N_{\text{vars}}} \leq \text{FI}(X)$。

测试 7：BP-FI 界限（上界）
评估：性质 4。
说明：我们评估性质 4 中给出的界限是否也适用于其他特征重要性方法。对于满足 $X \in \Omega_{\text{feat}}$ 的每个 FI(X)，BP-FI 都可以通过 $X \leq \text{Dep}(Y|\Omega_{\text{feat}})$ 给出上界。

测试 8：零独立意味着零特征重要性
评估：性质 5。说明：在某些数据集中，存在零独立变量。在这些情况下，我们研究它们是否也会获得零特征重要性。

测试 9：零特征重要性意味着零独立
评估：性质 5。
解释：当一个变量的特征重要性为零时，该特征应当与目标变量零独立。

测试10： 一个完全信息特征，两个零独立特征
评估：性质 6。
解释：特征重要性：附录：数据集）由一个完全依赖的目标变量 Y := X_1^{full}和两个零独立变量 $X_2^{\text{null-indep.}}, X_3^{\text{null-indep.}}$ 组成。我们检验 $\text{FI}(X_1^{\text{full}}) = 1$ 和 $\text{FI}(X_2^{\text{null-indep.}}) = \text{FI}(X_3^{\text{null-indep.}}) = 0$.

测试 11： 特征重要性最大值中的完全信息变量
评估：性质 8。
解释：只要数据集中存在完全信息特征，就不应有特征的特征重要性高于它。

测试 12： 限制结果空间
评估：性质9。
说明：为了评估对随机变量X 应用可测函数 f 是否会提高特征重要性，我们考察将同一随机变量使用不同分箱进行分箱的数据集。分箱可以看作应用函数 f。使用的分箱越少，特征重要性就不应提高。

测试 13： 添加特征可以提高特征重要性
评估：性质 11。
说明：每当向数据集添加一个特征时，我们都会考察这是否会提高某个原始变量的特征重要性。如果特征重要性从未提高，我们就认为测试失败。

测试 14： 添加特征可以降低特征重要性
评估：性质 12。
说明：每当向数据集添加一个特征时，我们都会检查这是否会降低某个原始变量的特征重要性。如果特征重要性从未降低，我们就判定测试失败。

测试15： 克隆不会提高特征重要性
评估：性质 13。
说明：我们评估向数据集添加一个克隆是否会提高原始变量的特征重要性。

测试16： 顺序不会改变特征重要性
评估：性质14。
说明：我们检查变量的顺序是否会改变分配的特征重要性。

测试 17： **结果异或**评估：
性质 15。
说明：此测试评估两个数据集的特定结果。数据集 14 的期望结果为 $(1/2, 1/2)$，数据集17 的期望结果为 $(1/2, 1/2, 0)$。当某个 FI 值超出期望结果的 ϵ 界限时，特征重要性方法未通过此测试。

测试 18： **结果概率数据集**
评估：性质 16。
说明：此测试评估所有概率数据集（数据集 18 至 28）的特定结果。概率p 的期望结果为 $(p, 1-p)$。当某个 FI 值超出期望结果的 ϵ 界限时，特征重要性方法未通过此测试。

参考文献

[1] K. Aas、M. Jullum 和 A. Løland。“特征存在依赖时解释个体预测：对 Shapley 值的更准确近似”。载于：人工智能298（2021），第 103502 页。ISSN：0004- 3702。DOI：https://doi.org/10.1016/j.artint.2021.103502。网址：https://www.sciencedirect.com/science/article/pii/S0004370221000539。[2] N. Abe 和 M. Kudo。“用于分类器无关特征选择的熵准则”。载于：基于知识的智能信息与工程系统。R. Khosla、R. J. Howlett 和 L. C. Jain 编。柏林、海德堡：施普林格柏林海德堡，2005 年，第 689–695 页。ISBN：978-3-540-31997-9。[3] A. Altmann、L. Tološi、O. Sander 和 T. Lengauer。“置换重要性：一种经校正的特征重要性度量”。载于：生物信息学26.10（2010 年 4 月），第 1340–1347 页。ISSN：1367-4803。DOI：10. 1093/bioinformatics / btq134。电子预印本：https : // academic.oup.com/bioinformatics/article-pdf/26/10/1340/16892402/btq134.pdf。网址：https://doi.org/10.1093/bioinformatics/btq134。[4] F. Aragón-Royón、A. Jiménez-Vílchez、A. Arauzo-Azofra 和 J. M. Benítez。*FSinR*：用于特征选择的穷举式 *R* 包。2020年。DOI：10.48550/ARXIV.2002.10330。网址：https://arxiv.org/abs/2002.10330。[5] G. Berkelmans、J. Pries、S. Bhulai 和 R. van der Mei。*BP* 依赖函数：一种用于度量随机变量之间依赖关系的通用方法。2022年。DOI：10.48550/ARXIV.2203.12329。网址：https://arxiv.org/abs/2203.12329。[6] A. del Bimbo、R. Cucchiara、S. Sclaroff、G. M. Farinella、T. Mei、M. Bertini、H. J. Escalante 和 R. Vezzani 编。模式识别。

ICPR国际研讨会与挑战赛。Springer International Publishing，2021年。DOI：10.1007/978-3-030-68787-8。网址：https://doi.org/ 10.1007%2F978-3-030-68787-8。[7] M. Carletti、C. Masiero、A. Beghi 和 G. A. Susto。“工业4.0中的可解释机器学习：评估异常检测中的特征重要性以实现根因分析”。载于：2019年 IEEE 系统、人与控制论国际会议（SMC）。2019年，第 21–26 页。DOI：10.1109/SMC.2019.8913901。[8] M. Carletti、M. Terzi 和 G. A. Susto。借助 DIFFI 的可解释异常检测：基于深度的孤立森林特征重要性。2020年。DOI：10.48550/ARXIV.2007.11117。网址：https://arxiv.org/abs/2007.11117。[9] G. Casalicchio、C. Molnar 和 B. Bischl。“可视化黑箱模型的特征重要性”。载于：数据库中的机器学习与知识发现。M. Berlingerio、F. Bonchi、T. Gärtner、N. Hurley 和 G. Ifrim 编。Cham：Springer International Publishing，2019年，第 655–670 页。ISBN：978-3-030-10925-7。[10] J. Castro、D. Gómez 和 J. Tejada。“基于抽样的 Shapley 值多项式计算”。载于：计算机与运筹研究 36.5（2009）。第十届国际选址决策研讨会（ISOLDE X）特选论文，第 1726–1730 页。ISSN：0305-0548。DOI：https://doi.org/10.1016/j.cor.2008.04.004。网址：https://www.sciencedirect.com/science/article/pii/S0305054808000804。[11] E. Celik。vita：变量重要性检验方法。R 包版本 1.0.0。2015年。网址：https://CRAN.R-project.org/package=vita。[12] T. Chen 和 C. Guestrin。“XGBoost：一种可扩展的树提升系统”。载于：第 22 届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集。KDD ’16。美国加利福尼亚州旧金山：美国计算机协会，2016年，第 785–794 页。ISBN：978-1-4503-4232-2。DOI：10.1145/2939672.2939785。网址：http://doi.acm.org/10.1145/2939672.2939785。[13] I. Covert、S. M. Lundberg 和 S.-I. Lee。“利用加性重要性度量理解全局特征贡献”。载于：神经信息处理系统进展。H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin 编。第 33 卷。Curran Associates, Inc.，2020年，第 17212–17223 页。网址：https://proceedings.neurips.cc/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf。[14] A. Datta、S. Sen 和 Y. Zick。“通过定量输入影响实现算法透明性：学习系统的理论与实验”。载于：2016年 IEEE 安全与隐私研讨会（SP）。2016年，第 598–617 页。DOI：10.1109/SP.2016.42。

[15] K. Dhamdhere、A. Agarwal 和 M. Sundararajan。《Shapley-Taylor交互指数》。载于：第*37*届国际机器学习会议论文集。ICML'20。JMLR.org，2020年。[16] C. Frye、C. Rowat 和 I. Feige。《非对称Shapley值：将因果知识纳入模型无关可解释性》。载于：第*34*届神经信息处理系统国际会议论文集。NIPS'20。加拿大不列颠哥伦比亚省温哥华：Curran Associates Inc.，2020年。ISBN：9781713829546。[17] D. V. Fryer、I. Strumke 和 H. Nguyen。《模型无关的特征归因：揭示非线性依赖关系的 Shapley 值》。载于：*PeerJ*计算机科学7（2021年），e582。[18] A. Ghorbani、D. Berenbaum、M. Ivgi、Y. Dafna 和 J. Y. Zou。《超越重要性分数：通过可视化特征语义解读表格机器学习》。载于：*Information* 13.1（2022年）。ISSN：2078-2489。DOI：10.3390/info13010015。URL：https://www.mdpi.com/2078-2489/13/1/15。[19] O. Giles 等人。伪造特征重要性：使用差分隐私合成数据的警示。2022年。DOI：10.48550/ARXIV.2203.01363

网址：https://arxiv.org/abs/2203.01363.[20] B. M. Greenwell 和 B. C. Boehmke。《变量重要性图——vip 软件包简介》。载于：*The R Journal* 12.1（2020），第 343–366 页。网址：https://doi.org/10.32614/RJ-2020-013.[21] S. Hooker、D. Erhan、P.-J. Kindermans 和 B. Kim。《深度神经网络可解释性方法的基准测试》。载于：第*33*届神经信息处理系统国际会议论文集。美国纽约州雷德胡克：Curran Associates Inc.，2019年。[22] T. Hothorn 和 A. Zeileis。《partykit：用于 R 语言递归分区的模块化工具包》。载于：机器学习研究期刊 16（2015），第 3905–3909 页。网址：https://jmlr.org/papers/v16/hothorn15a.html.[23] V. A. Huynh-Thu、Y. Saeys、L. Wehenkel 和 P. Geurts。《用于生物标志物发现的基于机器学习的特征重要性分数的统计解释》。英文。载于：生物信息学 28.13（2012年4月），第 1766–1774 页。[24] P. V. Johnsen、I. Strümke、S. Riemer-Sørensen、A. T. DeWan 和 M. Langaas。在高维数据中推断特征重要性及其不确定性。2021年。doi：10.48550/ARXIV.2109.00855。网址：https://arxiv.org/abs/2109.00855.[25] S. Khodadadian、M. Nafea、A. Ghassami 和 N. Kiyavash。用于公平性感知特征选择的信息论度量。2021年。doi：10.48550/ARXIV.2106.00772。网址：https://arxiv.org/abs/2106.00772.[26] K. Kira 和 L. A. Rendell。《特征选择的一种实用方法》。载于：第九届国际机器学习研讨会论文集——

学习。ML92。阿伯丁，苏格兰，英国：Morgan Kaufmann Publishers Inc.，1992年，第 249–256 页。[27] W. H. Kruskal。《关联的序数测量》。载于：《美国统计学会杂志》53.284（1958），第 814–861 页。ISSN：01621459。网址：http : //www.jstor.org/stable/2281954（访问日期：2022年11月1日）。[28] M. Kuhn。caret：分类与回归训练。R 包，版本 6.0-92。2022年。网址：https://CRAN.R-project.org/package=caret。[29] M. B. Kursa 和 W. R. Rudnicki。《使用 Boruta 包进行特征选择》。载于：《统计软件杂志》36（2010年9月11日），第 1–13 页。ISSN：1548-7660。doi：10 . 18637 / JSS . V036 . I11。网址： https : //www.jstatsoft.org/index.php/jss/article/view/v036i11。[30] X. Li、Y. Wang、S. Basu、K. Kumbier 和 B. Yu。《随机森林的去偏 MDI 特征重要性度量》。载于：第33届神经信息处理系统国际会议论文集。美国纽约州雷德胡克：Curran Associates Inc.，2019年。[31] S. Lipovetsky 和 M. Conklin。《博弈论方法中的回归分析》。载于：《应用随机模型：商业与工业》17.4（2001），第 319–330 页。doi：https://doi.org/10.1002/asmb.446。电子印本：https : //onlinelibrary.wiley.com/doi/pdf/10.1002/asmb.446。网址：https : //onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446。[32] Y. Y. Lu、Y. Fan、J. Lv 和 W. S. Noble。DeepPINK：深度神经网络中的可复现特征选择。2018年。doi：10.48550/ARXIV.1809.01185。网址：https://arxiv.org/abs/1809.01185。[33] S. M. Lundberg、G. Erion、H. Chen、A. DeGrave、J. M. Prutkin、B. Nair、R. Katz、J. Himmelfarb、N. Bansal 和 S.-I. Lee。《借助面向树模型的可解释人工智能，从局部解释走向全局理解》。载于：《自然·机器智能》2.1（2020年1月），第 56–67 页。ISSN：2522-5839。doi：10 . 1038 / s42256 - 019 - 0138 - 9。网址：https : //doi.org/10.1038/ s42256-019-0138-9。[34] S. M. Lundberg、G. G. Erion 和 S.-I. Lee。树集成的一致个体化特征归因。2018年。doi：10.48550/ARXIV.1802.03888。网址：https://arxiv.org/abs/1802.03888。[35] S. M. Lundberg 和 S.-I. Lee。《解释模型预测的统一方法》。载于：第31届神经信息处理系统国际会议论文集。NIPS'17。美国加利福尼亚州长滩：Curran Associates Inc.，2017年，第 476–4777 页。ISBN：9781510860964。[36] S. M. Lundberg 等人。《用于预防手术期间低氧血症的可解释机器学习预测》。英文。载于：Nat Biomed Eng 2.10（2018年10月），第 749–760 页。

[37] L. Merrick 和 A. Taly. “解释游戏：使用 Shapley 值解释机器学习模型”。载于《机器学习与知识提取》。A. Holzinger、P. Kieseberg、A. M. Tjoa 和 E. Weippl 编。Cham：Springer International Publishing，2020年，第 17–38 页。ISBN：978-3-030-57321-8。[38] P. E. Meyer. infotheo：信息论度量。R 包，版本 1.2.0.1。2022年。网址：https://CRAN.R-project.org/package=infotheo。[39] C. Molnar、G. König、B. Bischl 和 G. Casalicchio。《相关特征下模型无关的特征重要性与效应——条件子组方法》。2020年。DOI：10.48550/ARXIV.2006.04628。网址：https:// arxiv.org/abs/2006.04628。[40] A. Mungo. sklearn-relief。Python 软件包，版本 1.0.0b2。2017年12月。网址：https://libraries.io/pypi/sklearn-relief。[41] A. B. Owen 和 C. Prieur. “关于使用 Shapley 值衡量相关输入的重要性”。载于《SIAM/ASA 不确定性量化期刊》5.1（2017年），第 986–1002 页。DOI：10 . 1137 / 16M1097717。电子版：https : //doi.org/10.1137/16M1097717。网址：https : //doi.org/10.1137/16M1097717。[42] F. Pedregosa 等。“Scikit-learn：Python 机器学习”。载于《机器学习研究期刊》12（2011年），第 2825–2830 页。[43] N. Pilnenskiy. ITMO-FS。Python 软件包，版本 0.3.3。2020年8月。网址：https://pypi.org/project/ITMO-FS/。[44]J. Pries. bp 特征重要性软件包。https ://github.com/joris-pries/BP-Feature-Importance。2023年。[45] 用于计算属性重要性的 Fisher Score Python 实现。https : //www.codestudyblog.com/cs2112pyc/1223230432.html。访问日期：2022-09-26。[46] M. T. Ribeiro、S. Singh 和 C. Guestrin。“‘我为什么要相信你？’：解释任意分类器的预测”。载于《第 22 届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集》。KDD ’16。美国加利福尼亚州旧金山：计算机协会，2016年，第 1135–1144 页。ISBN：9781450342322。DOI：10 . 1145 / 2939672 . 2939778。网址：https : //doi.org/10.1145/2939672. 2939778。[47] C. R. de Sá。“神经网络中基于方差的特征重要性”。载于《发现科学》。P. Kralj Novak、T. Šmuc 和 S. Džeroski 编。Cham：Springer International Publishing，2019年，第 306–315 页。ISBN：978-3-030-33778-0。[48] A. Saabas. TreeInterpreter。2021年1月。网址：https : //pypi.org/project/ treeinterpreter/。

[49] L. S. Shapley 和 A. E. Roth 编。《Shapley 值：纪念 Lloyd S. Shapley 的论文集》。剑桥 [剑桥郡] ；纽约：剑桥大学出版社，1988年。ISBN：9780521361774。[50] K. Shin、T. Kuboyama、T. Hashimoto 和 D. Shepard。“sCWC/sLCC：高度可扩展的特征选择算法”。载于《Information》8.4（2017年），第 159 页。[51] E. Song、B. L. Nelson 和 J. Staum。“用于全局敏感性分析的 Shapley 效应：理论与计算”。载于《SIAM/ASA 不确定性量化期刊》4.1（2016年），第 1060–1083 页。DOI：10.1137/15M1048070。电子版：https://doi.org/10.1137/15M1048070。网址：https://doi.org/10.1137/15M1048070。[52] S. Stijven、W. Minnebo、K. Vladislavleva 和 N. e. Krasnogor。“去芜存菁：回归随机森林与符号回归中的特征选择和特征重要性”。英文。载于：（2011年）。网址：http://lib.ugent.be/catalog/pug01:3198359。[53] C. Strobl、A.-L. Boulesteix、A. Zeileis 和 T. Hothorn。“随机森林变量重要性度量中的偏差：示例、来源及解决方案”。英文。载于《BMC 生物信息学》8（2007年1月），第 25 页。[54] E. Štrumbelj 和 I. Kononenko。“通过特征贡献解释预测模型和个体预测”。载于《Knowledge and Information Systems》41.3（2014年12月），第 647–665 页。ISSN：0219-3116。DOI：10.1007/s10115-013-0679-x。网址：https://doi.org/10.1007/s10115-013-0679-x。[55] M. Sundararajan 和 A. Najmi。“用于模型解释的多种 Shapley 值”。载于《第37届国际机器学习会议论文集》。H. D. III 和 A. Singh 编。第 119 卷。《机器学习研究论文集》。PMLR，2020年7月，第 9269–9278 页。网址：https://proceedings.mlr.press/v119/sundararajan20b.html。[56] S. Tonekaboni、S. Joshi、K. Campbell、D. K. Duvenaud 和 A. Goldenberg。“哪里出了问题，何时发生？时间序列黑盒模型的实例级特征重要性”。载于《神经信息处理系统进展》。H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin 编。第 33 卷。Curran Associates, Inc.，2020年，第 799–809 页。网址：https : //proceedings.neurips.cc/paper/2020/file/08fa43588c2571ade19bc0fa5936e028-Paper.pdf。[57] H. X. Vinh. QII 工具。Python 软件包，版本 0.1.3。2019年2月。网址：https://pypi.org/project/qii-tool/。[58] P. Virtanen 等。“SciPy 1.0：用于 Python 科学计算的基础算法”。载于《Nature Methods》17（2020年）。Python 软件包，版本 1.8.0，第 261–272 页。DOI：10.1038/s41592-019-0686-2。

[59] B. D. Williamson 和 J. Feng。“使用 Shapley 值对总体特征重要性进行高效的非参数统计推断”。英文。载于：机器学习研究会议录 119（2020年7月），第 10282–10291 页。[60] E. Winter。“夏普利值”。载于博弈论及其经济应用手册。R. Aumann 和 S. Hart 编。第 1 版。第 3 卷。爱思唯尔，2002年。第 53 章，第 2025–2054 页。网址：https://EconPapers.repec.org/RePEc:eee:gamchp:3-53.[61] Z. Zhou 和 G. Hooker。“基于树的方法中特征重要性的无偏测量”。载于：*ACM*《数据知识发现汇刊》15.2（2021年1月）。ISSN：1556-4681。DOI：10.1145/3429445。网址：https://doi.org/10.1145/3429445.[62] A. Zien、N. Krämer、S. Sonnenburg 和 G. Rätsch。“特征重要性排序度量”。载于：机器学习与数据库知识发现。W. Buntine、M. Grobelnik、D. Mladenić 和 J. Shawe-Taylor 编。柏林、海德堡：施普林格柏林海德堡，2009年，第 694–709 页。ISBN：978-3-642-04174-7。